DYNAMICAL SYSTEMS DEFINED ON GROUPS:
STRUCTURAL PROPERTIES AND ESTIMATION

by

ALAN STEVEN WILLSKY

B.S., Massachusetts Institute of Technology, 1969

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

April 27, 1973

Signature of Author

Department of Aeronautics and Astronautics
April 27, 1973

Certified by

Thesis Supervisor

Certified by

Thesis Supervisor

Certified by

Thesis Supervisor

Certified by

Thesis Supervisor

Accepted by

Chairman, Departmental Committee on Graduate Students
DYNAMICAL SYSTEMS DEFINED ON GROUPS:
STRUCTURAL PROPERTIES AND ESTIMATION

by

Alan Steven Willsky

Submitted to the Department of Aeronautics and Astronautics on April 27, 1973 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

Several classes of systems evolving on groups are defined and studied. Both continuous and finite groups are considered. In the continuous group case, a general bilinear estimation problem is posed, and the special case in which the group of interest is the circle (S^1) is solved, as is a multidimensional version, corresponding to estimation on an arbitrary abelian Lie group. The form of the optimal estimator -- a nonlinear preprocessor, followed by a linear filter and a nonlinear postprocessor -- is quite appealing, and extensions of this technique to the general matrix Lie group case, including the rigid body rotation group, are considered. Also, the more complicated discrete-time, bilinear S^1 estimation problem is discussed and solved. In addition, a large class of estimation problems on S^1 is resolved with the aid of Fourier series analysis. The various S^1 estimation techniques have applications to frequency and phase demodulation problems, as well as to optical communications and phase tracking systems for navigational purposes (Loran, Omega). Some simulations of the various estimation systems have been carried out, and improvements over standard techniques have been observed.

In the finite group case, a class of systems evolving homomorphically on finite groups is defined, and their structural properties -- algebraic conditions for controllability, observability, realizability, invertibility, etc. -- are analyzed in detail. Comparisons between these results and those for linear systems are made throughout the development. By making the inputs to these systems stochastic in nature, we consider a class of estimation problems on finite groups, and, with the introduction of the concept of group algebra, some computational questions similar to those associated with the fast Fourier transform are considered. Finally, a class of estimation problems on the group Z_n = {0,1,...,n-1} with addition modulo n is resolved using a finite Fourier transform (group character) approach in much the same way as Fourier series was used in the S^1 case. Applications of these finite group methods to coding problems -- including the application of the Z_n results to the phase-shift-keying problem -- are briefly discussed.
Thesis Supervisors:

Professor Wallace E. Vander Velde

Title: Professor of Aeronautics and Astronautics, M.I.T.

Professor Roger W. Brockett

Title: Professor, Division of Engineering and Applied Physics, Harvard University

Professor John J. Deyst, Jr.

Title: Associate Professor of Aeronautics and Astronautics, M.I.T.

Professor Sanjoy K. Mitter

Title: Professor of Electrical Engineering, M.I.T.
ACKNOWLEDGEMENTS

"Certain authors, speaking of their works, say, 'My book,' 'My commentary,' 'My history,' etc. They resemble middle-class people who have a house of their own and always have 'My house' on their tongue. They would do better to say, 'Our book,' 'Our commentary,' 'Our history,' etc., because there is in them usually more of other people's than their own."

- Blaise Pascal [P11]

It is with great pleasure that I write this short part of my thesis, not only because it is one of the last bits of writing I have left before I finish this monster, but also because there have been so many people who I wish to thank for their advice, encouragement and friendship.

First I wish to express my gratitude to my thesis committee: Professor Roger W. Brockett whose constant encouragement and never-ending stream of enthusiastic ideas provided me with the basis and the motivation for my work; Professor John Devst whose criticism and advice have been of great importance in the development of the central ideas of my research; Professor Sanjoy K. Mitter who drew often on his immense storehouse of knowledge to provide me with the advice and perspective I needed; and Professor Wallace E. Vander Velde, committee chairman, whose guidance and carefully weighed criticism, starting when I was a sophomore in the Department of Aeronautics and Astronautics, have been invaluable. In addition, I wish to thank Professor Jan C. Willems for the valuable suggestions he has made during our many discussions and Professor James T. Lo with whom I have collaborated on some of the results presented in this manuscript.
I also wish to express my sincere thanks to the Fannie and John Hertz Foundation, who have supported me throughout my graduate education and whose faith in me has been greatly appreciated.

Thanks are also due to Miss Maureen Stanton for her conscientious efforts in typing this error-free manuscript and for her friendship (which I hope will not end when she reads this sentence).

I am also in the debt of all my friends for making the last few years even more enjoyable, and of the Boston Bruins, New England Whalers, and Boston Red Sox for providing me with an emotional release that turned out to be a sure-fire cure for the "my-proof-contains-a-major-mistake" blahs.

I wish to thank my parents, Beatrice and Max Willsky, for their constant encouragement and support over the years.

Finally I wish to express special and most sincere thanks to my wife Gail for her encouragement and patience throughout this research endeavor and for making my efforts much more meaningful.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>1.1 Background and Motivation</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>1.2 Problem Descriptions</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>1.3 Synopsis</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td><strong>PART I: CONTINUOUS GROUPS</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DYNAMICAL SYSTEMS AND RANDOM PROCESSES ON LIE GROUPS</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>2.1 Introduction</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>2.2 A Certain Class of Bilinear Control Systems</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>2.3 Brownian Motion Defined on Matrix Lie Groups</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>2.4 Some Estimation Problems on the Circle Group</td>
<td>39</td>
</tr>
<tr>
<td>3</td>
<td>PROBABILITY DENSITIES AND OPTIMAL ESTIMATES ON $S^1$</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>3.1 Introduction</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>3.2 The Optimal Estimation Problem and Fourier Series</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.3 Symmetric Criteria, Unimodal Distributions, and the</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Folded Normal Density</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>A BILINEAR ESTIMATION PROBLEM ON $S^1$</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>4.1 Introduction</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>4.2 Signal Processes, Observation Processes, and</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Conditional Probability Distributions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.3 A Continuous Time Bilinear Estimation Problem</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>4.4 A Discrete-Time Estimation Problem on $S^1$</td>
<td>106</td>
</tr>
<tr>
<td>5</td>
<td>THE USE OF FOURIER SERIES FOR PHASE TRACKING AND</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td>SYNCHRONOUS COMMUNICATION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.1 Introduction</td>
<td>119</td>
</tr>
</tbody>
</table>
5.2 A General Discrete-Time $S^1$ Estimation Problem 120

5.3 Fourier Analysis and Continuous-Time Phase Tracking and Demodulation 132

6 SOME APPLICATIONS AND NUMERICAL RESULTS 151

6.1 Introduction 151

6.2 Some Practical Problems Involving Single-Degree-of-Freedom Rotation 151

6.3 A Frequency Demodulation Problem in the Presence of Oscillator Phase Drift 164

6.4 A Phase Tracking Problem in the Presence of Additive Channel Noise 179

7 EXTENSIONS TO MORE GENERAL PROBLEMS INCLUDING RIGID BODY ROTATION 191

7.1 Introduction 191

7.2 Random Rotation of Rigid Bodies 191

7.3 Some Comments on General Lie Group Estimation Problems 213

PART II: FINITE GROUPS 220

8 FINITE GROUP HOMOMORPHIC SEQUENTIAL SYSTEMS 220

8.1 Introduction 220

8.2 Some Results from the Theory of Linear Sequential Circuits 223

8.3 A Class of Systems Evolving on Finite Groups 233

8.4 Number-Theoretic Arguments and Invertibility Conditions for FGHSS's 258

8.5 A Class of Algebraic Coding Systems 287
9 GROUP ALGEBRA FORMULATION OF ESTIMATION PROBLEMS FOR FGHSS'S AND SOME COMPUTATIONAL CONSIDERATIONS 294
   9.1 Introduction 294
   9.2 Conditional Distribution Equations on Groups 298
   9.3 Some Computational Aspects of Group Algebra Multiplications 320
   9.4 Some Comments on Probabilistic Decoding 341

10 ESTIMATION ON CYCLIC GROUPS USING GROUP CHARACTERS 348
   10.1 Introduction 348
   10.2 Group Characters and Random Variables on $\mathbb{Z}_n$ 350
   10.3 Character Equations for Discrete-Time Filtering on $\mathbb{Z}_n$ 358
   10.4 A Continuous-Time Estimation Problem on $\mathbb{Z}_n$ 365

11 SUMMARY, CONCLUSIONS, AND SUGGESTIONS 373

Appendix

A A SUMMARY OF SOME BASIC RESULTS IN MODERN ALGEBRA 378
   A.1 Introduction 378
   A.2 Groups 378
   A.3 Rings and Fields 391
   A.4 Lie Groups and Lie Algebras 394

B SOME TECHNICAL RESULTS RELATED TO ESTIMATION ON THE CIRCLE 399
   B.1 Introduction 399
   B.2 Differentiable Functions on $\mathbb{R}^1$ and Their Projections Onto $\mathbb{S}^1$ 399
   B.3 A Single Stage Estimation Problem 403
   B.4 A Limiting Argument 413
C EQUIVALENCE CLASS OBSERVATIONS AND TRUNCATION OF SERIES OF NORMAL DENSITIES

C.1 Introduction

C.2 Measurement Ambiguity Problems

C.3 Truncation of Series of Normal Densities

D MOMENT TRUNCATION METHODS WITH APPLICATIONS TO PHASE TRACKING PROBLEMS

E SOME RESULTS FROM ALGEBRAIC SYSTEM THEORY

F LIE GROUP HOMOMORPHIC SEQUENTIAL SYSTEMS

G AN EXAMPLE OF THE ERRORS INTRODUCED BY FOURIER SERIES TRUNCATION

Bibliography
LIST OF FIGURES

Figure

2.1 Illustrating the Projection Map $q$ | Page 42
2.2 Brownian Motion on $S^1$ | 44
4.1 Illustrating the Projection Map $J$ | 73
4.2 Block Diagram for Optimal Filtering on $S^1$ | 93
4.3 A Picture of the Torus $(S^1)^2$ | 96
4.4 Block Diagram of the Signal, Observation, and Optimal Estimation Processes of Example 4.1 | 102
4.5 Block Diagram of the Optimal Filter for Example 4.2 | 105
4.6 Conceptual Diagram of the Truncation Method for Sub-optimal Discrete-Time Filtering | 114
4.7 Illustrating the Concept of Using the Continuous-Time Filter to Approximate the Discrete-Time Filter | 115
5.1 The Physical Situation Described in Example 5.3 | 124
5.2 The Observation Geometry of Example 5.3 | 124
5.3 Block Diagram of Discrete-Time Optimal Estimator for the Criterion $\delta(1-\cos(\theta-\delta))$ | 131
5.4 Illustrating the Form of the Infinite Dimensional Optimal Filter of Example 5.6 | 137
5.5 Diagram of the $(a_n, b_n)$ Filter Shown in Figure 5.4 | 138
5.6 Diagram of a Phase-Lock Loop | 138
5.7 Illustrating the Form of the Optimal Phase Demodulator of Example 5.7 | 144
6.1 The Loran C Navigation System | 153
6.2 The Basic PLL Model | 162
6.3 The Baseband PLL Model | 162
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4 Illustrating the Optimal Filter Discussed in Example 6.1</td>
<td>167</td>
</tr>
<tr>
<td>6.5 A Procedure for Computing Total Phase</td>
<td>170</td>
</tr>
<tr>
<td>6.6 Block Diagram of the Optimal Filter Given Differential Phase Information</td>
<td>172</td>
</tr>
<tr>
<td>6.7 Block Diagram of the Optimal Filter Given Total Phase Information</td>
<td>172</td>
</tr>
<tr>
<td>6.8 Block Diagram of the PLL System</td>
<td>173</td>
</tr>
<tr>
<td>6.9 Baseband Model of the PLL System</td>
<td>173</td>
</tr>
<tr>
<td>6.10 A Graph of Some of the FM Simulation Results</td>
<td>181</td>
</tr>
<tr>
<td>6.11 Phase Error Variance Results</td>
<td>188</td>
</tr>
<tr>
<td>7.1 Form of the Optimal SO(3) Filter</td>
<td>206</td>
</tr>
<tr>
<td>7.2 Illustrating the Optimal Filter for x with Measurement Process (7.49)</td>
<td>209</td>
</tr>
<tr>
<td>7.3 Illustrating the Optimal Filter for x with Measurement Process (7.54)</td>
<td>209</td>
</tr>
<tr>
<td>7.4 A Possible Suboptimal SO(3) Estimator</td>
<td>211</td>
</tr>
<tr>
<td>7.5 Form of a General Lie Group Filter</td>
<td>217</td>
</tr>
<tr>
<td>8.1 Illustrating the Realizability Condition</td>
<td>239</td>
</tr>
<tr>
<td>8.2 A Factorization of f</td>
<td>245</td>
</tr>
<tr>
<td>8.3 A Factorization of a FGHSS Input-Output Map</td>
<td>246</td>
</tr>
<tr>
<td>8.4 The Commutative Diagram of Theorem 8.11</td>
<td>248</td>
</tr>
<tr>
<td>9.1 Subdivision of the Circle Into n Equal Size Subintervals</td>
<td>311</td>
</tr>
<tr>
<td>10.1 Illustrating the Form of the Optimal Filter Developed in Section 10.4</td>
<td>371</td>
</tr>
<tr>
<td>8.1 Graph of the Function of Example B.1</td>
<td>402</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table

6.1 Values of $\tau_e$, the Error Time Constant

6.2 Analytical Values of $1/\sigma_e^2$ for Optimal System Either with Perfect Differentiator or Perfect Total Phase Detector

6.3 Simulation Results ($1/\sigma_e^2$) for Optimal System, Perfect Differentiator

6.4 Step Size Variation with $c = 20, \ q = 0.01$

6.5 Simulation Results for Optimal System with Total Phase Detector ($f_c = 10,000$ Hz.)

6.6 Simulation Results for Optimal System with Perfect Total Phase Detector (Baseband)

6.7 Simulation Results for Optimal System with Perfect Total Phase Detector with Nonzero Initial Condition

6.8 Simulation Results for PLL with $K = 1$ (Baseband Model)

6.9 PLL Simulations with $K = 10$

6.10 PLL Simulations with $K = 100$

6.11 Phase Tracking Simulation Performance Summary — RMS Phase Error (Degrees)

6.12 Phase Tracking Simulation Performance Summary — $\bar{\theta}_c[1-\cos(\theta-\hat{\theta})]$

6.13 Summary of Results for FCF Without Wong-Zakai Correction Terms

9.1 The State Transition Table for the Permutation Reset Machine of Example 9.9

9.2 Multiplication Table for the Semigroup of Example 9.10
NOTATION

ψ - "for all"

E - expected value

δ - Dirac delta function

(Ω, ℳ, P) - a probability space

=, ~ - equivalence or isomorphism

< - "is a subgroup of"

≤ - "is a normal subgroup of"

R^1 - the real line

S^n - the n sphere (S^1 = the circle)

<=>, iff - "if and only if"

M(n, R^1) - n x n real-valued matrices

a | a - "a divides a"

μ << ρ - "μ is absolutely continuous with respect to ρ"

dμ/dρ - Radon-Nikodym derivative

A' - A transpose

\lim_{n \to \infty} - \lim_{n \to \infty}

∫ - Ito integral

~f - Nerode equivalence

≡f - Myhill equivalence

a.s. - almost surely

a.e. (P) - almost everywhere with respect to P

|ω| - the length of the string ω

|G| = card G - the cardinality of the set G
\(\mathbb{F}(G)\)  - the group algebra of \(\mathbb{F}\) over \(G\)

\([\cdot]\)  - equivalence class

\([x]\)  - smallest integer \(\geq x\) (ceiling)

\([x]\)  - largest integer \(\leq x\) (floor)

\([x]\)  - integer closest to \(x\)

\([A,B]\)  - the Lie bracket, \(AB-BA\)

\(\mathbb{Z}_n\)  - the cyclic group of order \(n\)

\(\mathbb{N}(x;\eta,\gamma)\)  - normal density with mean \(\eta\) and variance \(\gamma\)

\(\mathbb{F}(\theta;\eta,\gamma)\)  - folded normal density

\(\Pr(A)\)  - probability that \(A\) occurs

\(\dim\)  - dimension

\(\text{Im}\)  - image

\(\text{Ra}\)  - range

\(<S> = \{S\}_G\)  - group generated by \(S\)

\(\{M\}_A\)  - Lie algebra generated by \(M\)

\(O(t)\)  - \(\rightarrow 0\) as \(t \rightarrow 0\)

\(o(t)\)  - \(\frac{o(t)}{t} \rightarrow 0\) as \(t \rightarrow 0\)

\(\text{FGRSS}\)  - finite group homomorphic sequential system

\(\text{FGHE}\)  - finite group homomorphic encoder

\(\text{FSMP}\)  - finite state Markov process

\(\text{LSC}\)  - linear sequential circuit

\(\text{CE}\)  - convolutional encoder

\(\text{FM}\)  - frequency modulation

\(\text{AM}\)  - amplitude modulation

\(\text{PM}\)  - phase modulation
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCO</td>
<td>voltage controlled oscillator</td>
</tr>
<tr>
<td>PLL</td>
<td>phase-lock loop</td>
</tr>
<tr>
<td>SDNF</td>
<td>state dependent noise filter</td>
</tr>
<tr>
<td>FCF</td>
<td>Fourier coefficient filter</td>
</tr>
<tr>
<td>RMS</td>
<td>root mean square</td>
</tr>
<tr>
<td>PSK</td>
<td>phase-shift-keying</td>
</tr>
<tr>
<td>LGHSS</td>
<td>Lie group homomorphic sequential system</td>
</tr>
<tr>
<td>LPF</td>
<td>low pass filter</td>
</tr>
<tr>
<td>FFT</td>
<td>fast Fourier transform</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

"In the beginning, God created the Earth, and he said, 'Let there be mud.' And there was mud. And God said, 'Let Us make living creatures out of mud, so that mud can see what We have done.' And God created every living creature that now moveth, and one was man. Mud—as-man alone could speak.

'What is the purpose of all this?' man asked politely.

'Everything must have a purpose?' asked God.

'Certainly,' said man.

'Then I leave it to you to think of one for all of this,' said God. And He went away."

- Kurt Vonnegut, Jr., [V4]

1.1 Background and Motivation

In recent years, much of the research in system and control theory has involved the use of the tools of vector space theory to analyze and synthesize systems for a variety of estimation and control applications. In particular, the easily implemented and extremely useful linear-quadratic optimal controller [A12], [B2], [B11], and the Kalman-Bucy optimal linear filter [K6], [K7], have spurred a significant amount of effort to be put into using the techniques that were so successful in these problems to study many inherently nonlinear problems.

In many cases vector space theory and the vector state space setting are quite helpful in understanding the underlying structure of the system of interest. For a few classes of nonlinear problems, such as optimal estimation for nonlinear systems with finite dimensional sensor orbits [L17],
these vector space tools have led to exact solutions. In addition, this approach has resulted in a number of suboptimal and approximate techniques motivated by linear theory. Examples of these techniques are the extended Kalman-Bucy filter [J2], local controllability conditions for nonlinear systems [L13], and linear perturbation feedback control about a nominal trajectory [B11].

For many problems, these approximate methods yield excellent or at least satisfactory results. However, there are some nonlinear systems, such as the class of systems that can be described by bilinear differential equations (e.g., \( \dot{x} = ux \), or \( \dot{x}_1 = x_1 x_2, \dot{x}_2 = u \)), which possess a great deal of structure of their own. This structure is usually neglected when we apply techniques motivated by linear theory to these problems. In addition, techniques such as linearization about a nominal trajectory can turn inherently time-invariant problems into time-varying ones, and can lead to substantial computational requirements.

Motivated not only by the success and simplicity of linear theory and the easily implemented and analyzed structure of linear systems, but also by those nonlinear problems for which a vector space is not the natural setting, we ask the following question: are there any other types of mathematical tools that are particularly well-suited to analyze and synthesize particular classes of nonlinear systems (as linear algebra and the theory of linear ordinary differential equations are well-suited to deal with linear systems)?

The answer to this question is yes. The recent results of Brockett [B1], [B3], [B5] and Brockett and Rahimi [B3], [R2] indicate the value of analyzing bilinear and even linear control systems in the setting of
Lie groups and Lie algebras [C1], [W1], [S1]. In addition, the works of Myhill [M12], Nerode [N2], Krohn and Rhodes [K16], and Arbib and Manes [A8], among others, have all been based on utilizing the available algebraic structure of the (usually finite state) system of interest to develop tools for system analysis and synthesis.

The studies of bilinear systems and finite state automata appear to be rather odd bedfellows. It is our feeling that this is not the case, and we mention them together because we feel that the recent research efforts in both these directions stem from a common point of view -- that is, a desire to study the system of interest in its natural setting in the hope that the underlying properties of the system will be exposed and will lead to the development of valuable system-theoretic tools.

It is in this spirit that this manuscript is written. Specifically, we will investigate the properties of certain dynamical systems for which the framework of group theory [R1], [L1], [F1], [M1], [C1], [W1], [S1] provides the natural setting. In the next section we will discuss the particular problems to be considered in this thesis and present some physical and mathematical motivation for them. It should be emphasized that this research effort is aimed at problems of analysis and synthesis of dynamical systems. That is, in addition to considering the problem of analyzing highly structured nonlinear systems of practical importance, we will introduce and discuss new types of system structures that may be of use in realizing certain desired input-output maps.
1.2 Problem Descriptions

This research effort can be naturally divided into two parts. The first deals with problems involving continuous (Lie) groups, and is completely devoted to estimation problems. In the second part we consider a class of finite-state systems evolving on finite groups, and, in addition to some estimation and probability results, we present some purely algebraic results concerning a new class of highly structured finite-state systems.

In the work of Wei and Norman [W3], [W4], Brockett [B1], [B3], and Sussmann and Jurjjevic [S3], [J1], the powerful tools of Lie theory have been used to study the structure and properties of deterministic systems that can be described by bilinear equations of the form

\[ \dot{X}(t) = \left( \sum_{i=1}^{N} u_i(t)A_i + A_0 \right)X(t) \]  

(1.1)

where the \( u_i \) are scalar (control) functions and the \( A_i \) are \( nxn \) matrices. \( X \) can be either an \( n \)-vector or an \( nxn \) matrix. Examples of physical systems that can be described by (1.1) are some chemical process control problems [G13], DC to DC conversion [B1], and the rigid body rotation problem [B1], [W2], [H15]. The reader is referred to [R24] for a discussion of other bilinear systems.

The rigid body rotation problem is of special interest to us. In this case \( X \) is either a unit vector in \( R^3 \) (i.e. a direction in 3-space) or a 3x3 matrix of the direction cosines of one set of (rotating) orthogonal axes with respect to a fixed (inertial) set [W2]. The \( u_i \), in this case, are angular velocities about the three orthogonal rotating
or inertial axes, depending upon the coordinitization (see [W2]), and the $A_i$ are the skew-symmetric "infinitesimal rotations" about these axes (see Appendix A). If $X$ is taken to be a $3 \times 3$ direction cosine matrix, then $X$ is an orthogonal matrix -- i.e., $X'X = I$, $\det X = +1$. The set of all $3 \times 3$ orthogonal matrices with positive determinant is the matrix Lie group $SO(3)$, while the set of $3 \times 3$ skew-symmetric matrices is the Lie algebra associated with $SO(3)$ (see Appendix A). If $X$ is a 3-vector with $X'(0)X(0) = 1$, then $X'(t)X(t) = 1 \ \forall t$ -- i.e., $X(t)$ lies on the unit sphere $S^2$ in $\mathbb{R}^3$.

As mentioned before, in Part I of this thesis we will be concerned with questions related to random processes on continuous groups. McKean [M3], [M4], Ito [I3], Yosida, [Y2], Gorman [G1], Perrin [P1], and others have investigated processes on continuous groups. In many cases, the processes considered in these papers arise from stochastic analogs of (1.1) -- i.e., the $u_i$ are stochastic in nature. Motivated by this work and also by such practical problems as optimal orientation estimation given noisy observations [W2], we ask the following question: can one analyze stochastic dynamical systems as in (1) with some associated measurement process in such a way as to use the structure of the system to yield easily implemented optimal or suboptimal estimation equations?

This question is much more complex and general than the problems considered in this thesis; however, the results we shall present can be viewed as taking a first step toward the solution of the more general problem. In fact, we will formulate our problems in such a manner as to suggest possible generalizations.
In Part I we will mostly be concerned with rotational processes where we allow rotation about one axis only (some of these results have been reported in [L6]); however, at the end of Part I we will briefly consider the full three-dimensional rotation problem and a general Lie group problem. The one dimensional rotation case is extremely important. In addition to rigid body rotation, any process involving sinusoidally time-varying signals can be viewed as a one-degree-of-freedom rotation problem. Thus, any phase tracking or synchronization problem — be it phase or frequency demodulation [V1], [V2], clock synchronization [A13], [B22], [C10], and even some problems in optical communication [H9] — is included in the class of problems that we will consider. Specifically we will use an idea first considered in [O1] to reduce a multiplicative measurement noise process on the Lie group of one dimensional rotation to a linear measurement noise process on the associated Lie algebra. Also, we will use Fourier series techniques to analyze a general class of estimation problems. All of these techniques are introduced in an attempt to take full advantage of the structure of the problem of interest.

We use the deterministic results of Brockett, Wei and Norman, and the others, and the results on Brownian motion on Lie groups of McKean, Ito, et. al., as background for our work in Part I. The bilinear estimation results developed for one dimensional rotation and generalized to more complex problems are new, and are motivated partially by McKean's work. The technique used is related to one developed in [O1]. The results on optimal estimation criteria on the circle also represent an original
contribution. Fourier series techniques were used by Bucy and his associates [M14], [B9] to devise an infinite dimensional phase tracking filter for a specific problem. We rederive this result, consider much more general phase estimation problems, and derive physically appealing finite-dimensional approximate filters. Simulations of these new techniques demonstrate their usefulness.

In Part II of this manuscript we will study the properties of certain finite-state systems evolving on finite groups. Much of the work associated with finite state machines is connected with the problem of inventing classes of systems which can perform prescribed tasks efficiently and which can be built relatively easily -- i.e. which possess desirable structural properties. As noted by Arbib [A14] and Brockett [B23], the work of Nerode, Myhill, Krohn and Rhodes, and others in this area, considered questions quite similar to those posed by people working in control theory. The problem of finding a "minimal" or "most efficient" realization of some desired input-output behavior is of interest to people working in both areas, as is the question of decomposing systems into basic building blocks.

Motivated by this similarity in the problems associated with automata and control systems, we define a class of highly structured finite-state systems, called finite group homomorphic sequential systems (FGHSS's), that have many of the properties of linear systems (some of these results may also be found in [B14]). We will spend a good deal of time analyzing the structure of these systems for several reasons. First of all, this analysis will point out and emphasize the fact that linear systems are not the only ones that possess structure that can be successfully exploited.
Also, by allowing the input to our FGHSS to be random, we will be able to study the properties of a large class of finite-state Markov processes. We shall see that the inherent structure of the FGHSS leads to highly efficient methods for computing conditional probability distributions when we pose the problem in terms of group algebras [N3]. These computational methods, which are related to the fast Fourier transform [N3], [C6], [C7], [C9], are also helpful in the optimal control of certain finite-state Markov processes with incomplete state information [A10] and may prove to be of use for designing coding systems (as we shall see, the class of FGHSS's includes all convolutional encoders [M10], [M11], [M13], [S10], [W4] as a very special subclass).

Finally, throughout Part II, we will point out the similarities between the techniques being used and those that were developed for the continuous group problems. In particular, the group algebra formulation for conditional distribution computation generalizes to the continuous group case. However, the most striking similarity arises when we consider the analog of Fourier series for the cyclic group \(Z_n\) of nonnegative integers less than \(n\) with addition defined modulo \(n\). As we shall see, this group can be thought of as the spatial discretization of the circle (which is the natural state space for one dimensional rotation). Indeed, it was this precise point that, to a large extent, motivated the entire random process study reported in Part II. We will show that the structure of the estimation problem on \(Z_n\) is remarkably similar to that for the problem on the circle, and will briefly discuss how the \(Z_n\) results may prove useful for digital signal processing and coding applications.
The results of linear system theory and the initial result of Arbib [A9] for abelian group systems provide the background for the algebraic theory of FGHSS's developed here. The definitions of this class of systems and all of the results presented are new. In addition, the group algebra formulation developed for probability distribution computations for FGHSS's is new, while the computational questions raised are associated with problems in computational complexity considered by Nicholson [N3] and Depeyrot [D5], [D6]. The observation that the computational considerations related to fast Fourier transforms may be of use in many problems involving finite-state Markov processes is a major contribution of this work. Several examples using a brute force technique suggested by Brockett and Dobkin [B18], [D4] indicate that large computational savings can be achieved. Finally, the finite Fourier transform techniques for $\mathbb{Z}_n$ problems represent an original contribution motivated by the Fourier series results of Part I.

The $\mathbb{Z}_n$-circle analogy is perhaps the most obvious example of the conceptual approach that forms the basis for and theme of this thesis. This approach or point of view is that of formulating a problem in its most natural setting and using all available structure in order to find a solution. The wide variety of problems considered in this thesis indicates the breadth of such an approach — even if we restrict ourselves to a particular structural setting, such as groups. It is one of the major hopes of the author that any success that we have had using this method of attack will lead to future efforts along these lines.
1.3 Synopsis

We now present a brief summary of the dissertation. Part I, continuous groups, consists of Chapters 2 through 7. In Chapter 2 we review some of the important results for deterministic bilinear systems, describe the contraction of Brownian motion on a matrix Lie group, and consider the specific case of the circle, $S^1$, in some detail. Chapter 3 contains a number of results on the estimation of random variables on the circle. Fourier series techniques play a major role in this chapter, as does the circle analog of the normal probability density. In Chapter 4 we study both the continuous and discrete time versions of an optimal filtering problem. The continuous time problem is of interest in that we use a technique similar to that developed in [01] to construct the optimal estimator as the cascade of a nonlinear pre-processor, a linear Kalman-Bucy filter, and a nonlinear post-processor. The discrete time version provides a striking comparison with the continuous time problem, since in this case the discrete time solution is infinite-dimensional and much more complex than the continuous time solution. In Chapter 5 we consider rather general phase tracking problems for both continuous and discrete time applications. Using Fourier series techniques similar to those discussed in [M14], we are able to expose the underlying structure of large classes of $S^1$-estimation problems. Chapter 6 contains a discussion of some of the physical problems that can be analyzed using the tools of Chapters 3-5. Also, simulation results comparing the techniques of Chapters 4 and 5 to standard phase-lock loop techniques [V1], [V2] are presented in Chapter 6. In Chapter 7 we return to the general Lie group estimation problem and make a few comments that are motivated by the
results of the preceding chapters. In addition, the $SO(3)$-rigid body rotation problem is considered in somewhat greater detail, and some suboptimal techniques are discussed.

Chapters 8-10 comprise Part II of the thesis. In Chapter 8 we introduce the class of finite group homomorphic sequential systems and derive a number of algebraic-structural results analogous to some of the standard results in linear system theory [B2], [K11]. A number of results for the case in which the inputs to and initial state of a FGHSS are random are presented in Chapter 9. The introduction of the concept of the real group algebra of a finite group allows us to write conditional probability distribution equations in a particularly appealing manner. This formulation leads naturally to some questions of computational complexity related to the fast Fourier transform [N3], [D6]. In Chapter 10, we consider the cyclic group $\mathbb{Z}_n$. The basic tool used is the group character decomposition of functions defined on $\mathbb{Z}_n$. This technique, which is the $\mathbb{Z}_n$ analog of Fourier series, allows us to derive probability distribution and optimal estimation equations that are quite similar to those derived in Chapters 3 and 5.

In Chapter 11 we summarize the findings and contributions described in this dissertation and discuss a number of future research directions that are motivated by our results. In addition, a number of appendices are included to supplement the discussions presented in the main body of the thesis. The reader who is unfamiliar with some of the algebraic concepts used in the dissertation is referred to Appendix A, in which we summarize some of the basic results from the theory of modern algebra.
PART I: CONTINUOUS GROUPS

CHAPTER 2

DYNAMICAL SYSTEMS AND RANDOM PROCESSES ON LIE GROUPS

"A group that operates this way may seem strange sometimes."
- H.J. Leavitt [L2]

2.1 Introduction

In this chapter we will review some of the recent work on the control of bilinear systems and will also present some of McKean's results [M3], [M4], concerning Brownian motion processes on Lie groups. Motivated by these results, we will discuss the Lie group $S^1$ -- realized as the unit circle in $\mathbb{R}^2$ -- in detail, and several specific problems will be described. The problems described in Sections 2.2 and 2.3 are rather general, while those of Section 2.4, where we discuss $S^1$, are very specific. We provide the general framework in order to use the specific results for $S^1$ (Chapters 3-6) to motivate some specific estimation problems on other matrix Lie groups (see Chapter 7).

2.2 A Certain Class of Bilinear Control Systems

In the very recent past, a great deal of effort has been put into investigating the properties of control systems that can be described by the bilinear equation

$$\dot{X}(t) = (A_0 + \sum_{i=1}^{n} u_i(t)A_i)X(t)$$  \hspace{1cm} (2.1)

where the $u_i$ are scalars, the $A_i$ are nxn matrices and $X$ is either an n-vector or an nxn matrix. As mentioned by Brockett [B1], there are
many papers in the literature that study the controllability of non-linear control systems by using concepts related to the results of Chow [C2]. Among these are the work of Hermann [H3], Kucera [K1], [K2], Haynes and Hermes [H4], Elliott [E1], and Lory [L3]. The point of view taken in these papers is more differential geometric than algebraic in nature.

On the other hand, there are a number of recent system-theoretic papers that analyze the properties of linear and bilinear control systems in a Lie-algebraic manner. Among these are the work of Brockett [B1], [B3], [B5], Sussmann and Jurdjevic [S3], [J1], [B3], and Brockett and Rahimi [B3], [R2]. The motivation for these papers comes at least in part from the success of Magnus [M5] and Wei and Norman [W3], [W4] in using Lie algebraic methods to solve linear differential equations. These papers deal with the construction of representations of the solution of (2.1) by making use of the properties of the Lie algebra generated by the \( A_i \). Of particular importance is the following result, proven by Wei and Norman [W3].

**Theorem 2.1:** Consider the differential equation

\[
X(t) = \left( \sum_{i=1}^{m} a_i(t)A_i \right)X(t) \quad X(0) = I
\]  

(2.2)

where the \( a_i \) are scalar functions of time, the \( A_i \) are constant nxn matrices, and \( X \) is an nxn matrix function of \( t \). Let \( B_1, B_2, \ldots, B_k \) be a basis for \( L = \{A_1, \ldots, A_m\}_A \) (see Appendix A for notation). Then there exists a neighborhood of \( t = 0 \) in which the solution of (2.2) may be expressed in the form

\[
X(t) = \exp(g_1(t)B_1)\exp(g_2(t)B_2)\ldots\exp(g_k(t)B_k)
\]  

(2.3)
where the $g_i$ are scalar functions of time. Moreover, the $g_i$ satisfy a set of different equations which depend only on the Lie algebra $L$ and the $a_i$.

We note that for the case in which $X \in SO(3)$ (see Appendix A) -- i.e. when $X$ is a direction cosine matrix -- Wei and Norman's result simply states that the Euler angle representation is valid over the time period ending when we experience gimbal lock [W2]. We will discuss this example in more detail in Chapter 7.

Theorem 2.1 and several other results provide the foundation for the more recent work of Brockett, et. al. (we note that one can view the results of Brockett [B1] as standing in the same relation to those of Wei and Norman [W3], [W4] as the results of Chow [C2] on distributions of vector fields on a differentiable manifold stand in relation to the results of Frobenius [W1]). Note in particular the last sentence of Theorem 2.1. It indicates that a thorough understanding of the structure of $L$ may lead to efficient methods for computing the solution to (2.2), which can be viewed as a state transition matrix equation [B2]. For example, we include the following important special case where the representation (2.3) is global.

**Definition 2.1:** Let $L$ be a matrix Lie algebra. The derived algebra $L'$ is defined by

$$L' \triangleq [L,L] = \{ [A,B] | A, B \in L \}$$

(2.4)

Define $L^{(r)}$ inductively

$$L^{(1)} = L'$$

$$L^{(r)} = (L^{(r-1)})'$$

(2.5)
The algebra $L$ is called solvable if $L^{(r)} = \{0\}$ for $r$ large enough.

**Theorem 2.2:** Consider (2.2). If $L$ is solvable, then there exists a basis $B_1, \ldots, B_k$ of $L$ and an ordering of this basis, for which (2.3) is global.

**Proof:** See Theorem 2 of [W3].

Such results lead naturally to the consideration of questions such as controllability and observability of bilinear control systems, and the success of these considerations is evident from the many concrete results that have been obtained. An example is the following result from [B1].

**Theorem 2.3:** Consider the dynamical system

$$\dot{X}(t) = \left( \sum_{i=1}^{m} u_i(t)B_i \right)X(t) \quad (2.6)$$

where the $u_i$ are scalars, $X$ is an $n \times n$ matrix, and the $B_i$ are constant $n \times n$ matrices. Given a time $t_a > 0$ and given two nonsingular matrices $X_1$ and $X_2$, there exist piecewise continuous controls which steer the state from $X_1$ at $t = 0$ to $X_2$ at $t = t_a$ if and only if $X_2X_1^{-1} \in \{ \exp\{B_iA\} \}$.  

The proof of this result relies heavily on Theorem 2.1 and some other results developed in [B1].

The power of Theorem 2.3 is indicated by the fact that extensions of this result include the controllability condition for linear time-invariant systems as an extremely special case (see Theorem 7 and Example 4 of [B1]).

Thus, the class of systems described by (2.1), although inherently nonlinear in nature, has a great deal of structure, which can be used
to develop a theory for these systems that is almost as detailed as the theory for linear systems. As stated by Brockett [B1] in referring to the work of Hermann, Kucera, et. al., "This work is relevant here but we are directly interested in controllability only in so far as it contributes to the identification of a framework in which we can study a full range of system theoretic questions, including observability and realization theory." The class of systems (2.1) provides such a framework, and it is the purpose of Part I of this manuscript to investigate within this framework some problems involving stochastic processes.

2.3 Brownian Motion Defined on Matrix Lie Groups

The highly structured matrix Lie Group provides an appropriate space on which to generalize the definition of Brownian motion. In this section we will describe the "injection procedure," defined by McKean [M3]. For a more detailed description of the specific example of Brownian motion on SO(3), the set of 3x3 orthogonal matrices, see [M4]. We also note that this subject has received much attention in the literature. In particular, we cite the work of K. Ito [I1], [I2], [I3], Levy [L4], Perrin [P1], Yosida [Y1], [Y2], and Gorman [G1].

We first introduce the notion of product integral. Let A(t) be a continuous n×n matrix function of time on the interval [0,T]. For 0 < s₁ < s₂ < T, define

\[ E(s₂, s₁) = \int_{s₁}^{s₂} A(t) dt \]  

(2.7)

and for every positive integer n, define the function Xₙ(t) on [0,T]:
\[ X_n(t) \triangleq \exp(B(t, \lfloor 2^n \rfloor 2^{-n})) \exp(B(\lfloor 2^n \rfloor 2^{-n}, (\lfloor 2^n \rfloor -1)2^{-n})) \cdot \cdots \cdot \exp(B(2^{-n}, 0)) \]
\[ \triangleq \exp(B(t, \lfloor 2^n \rfloor 2^{-n})) \prod_{k=1}^{\lfloor 2^n \rfloor} \exp(B(k2^{-n}, (k-1)2^{-n})) \] (2.8)

where \( \lfloor x \rfloor \triangleq \) the largest integer \( \leq x \).

It has been shown (see [M3], [M4], [C2], [B19], and [C3]) that
\[ \lim_{n \to \infty} X_n(t) = X(t) \] exists, is continuously differentiable, and satisfies
\[ \dot{X}(t) = A(t)X(t) \quad X(0) = I \] (2.9)

We denote the product integral solution of (2.9) by
\[ X(t) = \bigcap_{s \leq t} \exp[A(s)ds] \] (2.10)

To see the correctness of the product integral form for the solution to (2.9), consider the case in which \( A \) is a scalar. Then (2.8) becomes
\[ X_n(t) = e^{\int_0^t A(\tau)d\tau} \psi_n \] (2.11)

and this obviously satisfies (2.9). Also, if \( A \) is nxn but we have the special condition
\[ A(t)A(\tau) = A(\tau)A(t) \quad \forall t, \tau \in [0,T] \] (2.12)
(i.e. the Lie algebra \( \{A(t) : t \in [0,T]\}_A \) is abelian), then, from [B2] we again have
\[ X_n(t) = \exp(B(t, \lfloor 2^n t \rfloor 2^{-n})) + \sum_{k=1}^{\lfloor 2^n t \rfloor} B(k2^{-n}, (k-1)2^{-n})) \]

\[ = \exp \int_0^t A(\tau) d\tau \quad (2.13) \]

The general case, as indicated by the results of Wei and Norman [W2], [W4], is not this simple.

We will now follow McKean [M3] in using the product integral to "inject" a Brownian motion into a matrix Lie group. Suppose we are given a matrix Lie algebra \( L \) with basis \( A_1, \ldots, A_n \). Let \( w(t) \) be a standard \( n \)-dimensional Brownian motion process [M3], [J2], [W5]:

\[ E(w(t)) = 0 \quad \forall \ t \geq 0 \quad (2.14) \]

\[ E(w(t)w'(\tau)) = t I_n \quad 0 \leq t \leq \tau \quad (2.15) \]

Let \( Q \) be an \( nxn \) (symmetric) positive semidefinite matrix, and let \( f \) be an \( n \)-vector. Define the \( n \)-dimensional skew Brownian motion

\[ v(t) = Q^{1/2} w(t) + ft \quad (2.16) \]

where \( Q^{1/2} \) is the symmetric positive semidefinite square root of \( Q \).

We now inject \( v(t) \) into the matrix Lie group \( \{ \exp(L) \} \) via the product integral

\[ G(t) = \bigcap_{\text{sft}} \exp\left( \sum_{i=1}^{n} A_i dv_i(s) \right) \quad (2.17) \]

See [M3] for the proof that this stochastic version of the product integral is well defined.

McKean shows that \( G \) satisfies the Ito stochastic differential equation [J2], [W5]

\[ dG = \sum_{i=1}^{n} A_i dv_i(s) \]

\[ + \frac{1}{2} \sum_{i,j=1}^{n} \langle A_i, A_j \rangle dv_i(s) dv_j(s) \quad (2.18) \]
The second term in the parentheses on the right hand side of (2.18) arises from the peculiarities of Ito stochastic calculus. Basically, we have

$$dG(t) = \exp\left( \sum_{i=1}^{n} A_i dv_i(t) \right) G(t) - G(t)$$

(2.19)

Expanding the matrix exponential using the formal rule

$$dw_i dw_j = \delta_{ij} dt$$

(2.20)

where

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

(2.21)

we have

$$dG(t) = \left( \sum_{i=1}^{n} A_i dv_i(t) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij} A_i A_j dt \right) G(t) + o(dt)$$

(2.22)

See [M3] for a rigorous development.

We wish to note two things about (2.18). If we formally divide by $dt$ and define

$$A_o = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij} A_i A_j$$

(2.23)

$$u_i = \dot{v}_i$$

(2.24)

equation (2.18) reduces to (2.1). Thus this injection procedure leads to a stochastic analog of the bilinear deterministic system described by (2.1). We note that this analogy is a formal one in that the process $\dot{v}_i$ is a nonexistent white noise process. We thus must be
careful in drawing too many conclusions from this formal analogy. The most important problem is the following. From the deterministic results of Wei and Norman [W3], [W4] and Brockett [B1] we would expect that the solution $G(t)$ of (2.18) satisfies

$$G(t) \in \{ \exp[A_o, \ldots, A_n] \}_G^A$$

(2.25)

In addition, if $A_o \notin \{ A_1, \ldots, A_n \}_A$, then in general, if the deterministic results are applicable, we should have

$$G(t) \notin \{ \exp[A_1, \ldots, A_n] \}_A$$

(2.26)

(see [B1])—i.e. the solution must have a "component" in the $A_o$ "direction." However, by the manner in which we defined the process $G(t)$ (see (2.17)), we specifically injected the process into $\{ \exp[A_1, \ldots, A_n] \}_G$.

This apparent contradiction is explained by observing that the stochastic differential equation (2.18) should be interpreted in the mean square sense. In fact, $A_o$ defined by (2.23) is precisely the term necessary to keep $G(t)$ in $\{ \exp[A_1, \ldots, A_n] \}_G$ in the mean square sense. Consider the example of $SO(3)$ (see [M4]), where we take

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

(2.27)

In this case

$$A_o = \frac{1}{2} \begin{bmatrix} -(\gamma_{22} + \gamma_{33}) & \gamma_{12} & \gamma_{13} \\ \gamma_{12} & -(\gamma_{33} + \gamma_{11}) & \gamma_{23} \\ \gamma_{13} & \gamma_{23} & -(\gamma_{11} + \gamma_{22}) \end{bmatrix}$$

(2.28)
If we take \( Q = I \), then \( A_0 = -I \), which clearly is not skew-symmetric. Thus the solution to the deterministic equation

\[
\dot{X}(t) = [-I + \sum_{i=1}^{3} A_i u_i(t)]X(t) \quad X(0) = I
\]

(2.29)
is not an element of \( SO(3) \) in general. However, let us consider the stochastic equation

\[
dG(t) = [-Idt + \sum_{i=1}^{3} A_i dv_i(t)]G(t) \quad G(0) = I
\]

(2.30)
where the \( v_i \) are defined by (2.16) with \( Q = I, f = 0 \). Consider the matrix function

\[
K(t) = G'(t)G(t) \quad K(0) = I
\]

(2.31)
and compute its stochastic differential using the differential rule [J2], [I4].

\[
dK(t) = [dG'(t)]G(t) + G'(t)[dG(t)] + P(t)dt
\]

(2.32)
where \( P(t) \) is the second order correction term, which can be shown to be \( 2K(t) \). Then substituting (2.30) into (2.32) and using the fact that \( A_i' = -A_i, i=1,2,3 \), we get \( dK(t) = 0 \). Thus, if \( K(0) = I \), i.e. if we start on \( SO(3) \) -- \( K(t) = I \) (in the mean square sense) -- i.e. we stay on \( SO(3) \) in the mean square sense. Note that if we do not include the correction term in (2.30), then

\[
dK(t) = 2K(t)dt \quad K(0) = I
\]

(2.33)
and \( G(t) \) does not stay on \( SO(3) \).

It is appropriate to comment on the diffusion equation satisfied by the probability "density" for \( G \). This discussion will be informal.
and not rigorous. To do this, we must first define Haar measure on a matrix Lie group. Any matrix Lie group $H$ has a natural topology induced from the standard Euclidean topology on $M(n, \mathbb{R})$ [K3], [D1]. We can then consider the sigma algebra $\mathcal{B}_H$ of Borel subsets of $H$ — the sigma algebra generated by the class of all open sets in $H$ [H5], [R3].

**Definition 2.2:** A **left invariant Haar measure** $\mu$ on the Borel sigma algebra $\mathcal{B}_H$ of a matrix Lie group $H$ is a positive measure on $\mathcal{B}_H$ such that

$$\mu(aB) = \mu(B) \quad \forall a \in H, \ B \in \mathcal{B}_H$$  \hspace{1cm} (2.34)

An example of Haar measure is Lebesgue measure $\lambda$ on the real line — i.e. $\lambda(x+A) = \lambda(A)$, where $x+A = \{x+y | y \in A\}$ (translate the set $A$ $x$ units).

It can be shown [H5], [L5] that for any matrix Lie group $H$, there is such a left invariant Haar measure which is unique up to a scaling constant. Also, if the group is compact, or abelian the Haar measure is also right invariant, and a group is compact if and only if $\mu(H) < \infty$. Thus for compact groups, we can normalize the measure ($\mu(H) = 1$), and in the rest of this section we will assume for simplicity that $H$ is compact and $\mu$ is normalized.

Now let $H = \{\exp L\}_G$ where $L$ is a Lie algebra with basis $\{A_1, \ldots, A_n\}$. Consider the Brownian motion process $G(t)$ satisfying (2.18) and consider the induced probability measure $\nu_G(t)$ on $\mathcal{B}_H$

$$\nu_G(t)(A) = \Pr[\{G(t) \in A\}]$$  \hspace{1cm} (2.35)

(the almost sure continuity of $G$ allows us to define this). McKean
states that, if \( \Omega \) is nonsingular, \( \nu_C(t) \ll \mu \), the Haar measure on \( \mathbb{H} \), and the probability density

\[
p(g,t) = \frac{d\nu_C(t)}{du}(g) \quad g \in C
\]  \hspace{1cm} (2.36)

(i.e. \( \nu_C(t)(A) = \int_A p(g,t)\,du(g) \)) satisfies the forward equation

\[
\frac{\partial p}{\partial t} = \mathcal{L}^* p
\]  \hspace{1cm} (2.37)

where \( \mathcal{L}^* \) is the formal adjoint of the differential operator

\[
\mathcal{L} = \frac{1}{2} D'QD + f'D = \frac{1}{2} \sum_{i,j=1}^{n} \Omega_{ij} \frac{\partial}{\partial i} \frac{\partial}{\partial j} + \sum_{i=1}^{n} f_i \frac{\partial}{\partial i}
\]  \hspace{1cm} (2.38)

The differential operators \( D_i \) are defined as follows: let \( r \) be a function on the group \( \mathbb{H} \); then we define

\[
(D_i r)(h) = \frac{d}{dt} \left. r(\exp(A_i t)) \cdot h \right|_{t=0} \quad h \in \mathbb{H}
\]  \hspace{1cm} (2.39)

(one can show that the \( D_i \) are quite similar to directional derivatives in \( \mathbb{R}^n \)). The operator \( \mathcal{L} \) is called either the infinitesimal generator governing the process \( C(t) \) or the backward operator.

2.4 Some Estimation Problems on the Circle Group

In this section we will apply the injection procedure of Section 2.3 to construct a Brownian motion on the circle, \( S^1 \). The probability density of this process is discussed in detail and a class of synchronization - demodulation problems is presented to motivate the following chapters (the results of the present section can also be found in Lo and Willsky [L6]). Chapters 3, 4, 5, and 6 will deal with
these estimation problems in detail, while in Chapter 7 we return to
the general setting of Sections 2.2 and 2.3, and, motivated by the
results of Chapters 3-6, we will formulate and discuss a general Lie
group problem.

Before describing random processes on $S^1$, it is appropriate
to comment on several distinct representations of the circle. A point
on the unit circle in $\mathbb{R}^2$ can be represented by either the angle
$\theta \in [-\pi, \pi]$ it makes with a fixed reference point on the circle or
by the 2x2 orthogonal matrix

$$
\begin{bmatrix}
\cos \theta & \sin \theta \\
-sin \theta & \cos \theta
\end{bmatrix}
$$

The set of 2x2 orthogonal matrices with determinant +1 is the abelian
Lie group $SO(2)$. For $\theta$ near zero, we have the first order approximation

$$
\begin{bmatrix}
\cos \theta & \sin \theta \\
-sin \theta & \cos \theta
\end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \theta \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}
$$

(2.40)

The matrix

$$
R = \begin{bmatrix}
0 & 1 \\
-1 & 0
\end{bmatrix}
$$

(2.41)

is called the infinitessimal rotation and is a basis for the Lie
algebra of $SO(2)$. The Lie algebra and group are related by the
exponential map

$$
\begin{bmatrix}
\cos \theta & \sin \theta \\
-sin \theta & \cos \theta
\end{bmatrix} = \exp \theta R
$$

(2.42)

and the logarithm map [H10]
\[ \log(B) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{(B-I)^n}{n} \quad B \in SO(2) \quad |B-I| < 1 \quad (2.43) \]

Note that the addition of \( \theta_1 \) and \( \theta_2 \) modulo 2\( \pi \) corresponds to the multiplication of the two matrices representing the points. Another representation of \( S^1 \) is the set of complex numbers of unit modulus. Any such number can be uniquely written as \( e^{i\theta} \) with \( \theta \in [\pi, \pi) \), and the relationship with the above representations is obvious.

Finally, there exists a natural projection from \( \mathbb{R}^1 \) to \( S^1 \), here identified with \( [-\pi, \pi) \):

\[ q(x) = x \mod 2\pi \quad (2.44) \]

As Figure 2.1 indicates, two points \( x_1 \) and \( x_2 \) are projected onto the same point if and only if they differ by an integral multiple of \( 2\pi \). Also since \( q(x_1 + x_2) = [q(x_1) + q(x_2)] \mod 2\pi \), \( q \) is a homomorphism of \( \mathbb{R}^1 \) into \( S^1 \) with kernel

\[ \ker q = \{ 2n\pi | n \in \mathbb{Z} \} \quad (2.45) \]

Clearly the range of \( q \) is all of \( [-\pi, \pi) \). Thus, by the First Isomorphism Theorem (Theorem A.7)

\[ S^1 \cong \mathbb{R}^1/\ker q \triangleq \mathbb{R}^1/2\pi\mathbb{Z} \quad (2.46) \]

and any point \( \theta \in [-\pi, \pi) \) is identified with the equivalence class (coset) \( \{ \theta + 2n\pi | n \in \mathbb{Z} \} \) (we note that since \( \ker q \) is a closed subgroup of \( \mathbb{R}^1 \), \( \mathbb{R}^1/\ker q \) is a Lie group (see [Cl]), and one can show that the isomorphism (2.46) is a Lie group isomorphism -- i.e. it is smooth).

The topology of the circle is obtained by endowing the interval \( [-\pi, \pi] \) with the quotient topology [KJ] obtained by identifying \(-\pi \) and \( \pi \).
The topology of any other $S^1$ representation is obtained from this topology and the relationship between $[-\pi, \pi]$ and the other representation. Throughout this manuscript we will use these various representations interchangeably and will refer to them all as "the circle" or "$S^1$.

![Diagram of $S^1$ with points labeled $-\pi$, $-3\pi$, $-\pi$, $\pi$, $3\pi$, $5\pi$]

**Figure 2.1:** Illustrating the Projection Map $q$

We now wish to see precisely what the McKean injection procedure of Section 2.3 looks like for $S^1$. Let $W(t)$ be a standard one-dimensional Brownian motion process. We will inject this into $S^1$ identified with $SO(2)$. Let

$$W(t) = \bigcap_{s \leq t} \exp(Rdw(s))$$  \hspace{1cm} (2.47)

(refering to (2.16) and (2.17), take $Q = 1$, $f = 0$). Recall that since $SO(2)$ is abelian, we can simplify (2.48) (see (2.13))

$$W(t) = \exp Rw(t)$$

$$= \begin{bmatrix} \cos w(t) & \sin w(t) \\ -\sin w(t) & \cos w(t) \end{bmatrix}$$  \hspace{1cm} (2.48)
Note that if we included the skew term -- i.e., a constant velocity term \( f \) -- it would appear as a "carrier frequency."

Referring to (2.18), we find that \( W(t) \) satisfies the stochastic differential equation

\[
dW(t) = \begin{bmatrix} -\frac{1}{2} \, dt & dw(t) \\ -dw(t) & -\frac{1}{2} \, dt \end{bmatrix} W(t)
\]  

(2.49)

Note that (2.49) appears to be an equation for a damped oscillator, while we know that since \( WW = I \), there is no damping. This apparent contradiction is the same as the one discussed earlier and arises from the peculiarities of Ito calculus and white noise. Referring to [G3], we note that the damping terms can be interpreted as providing the dissipation necessary to compensate for the infinite power of white noise (equation (2.33) indicates what happens if we leave out the correction term). Note also that since \( S^1 \) is abelian, \( W(t) \) is explicitly a function of \( w(t) \) and not the entire time history of \( w \) up to time \( t \). Thus, instead of using (2.18) to derive (2.49), we could have used the Ito differential rule [J2], [I4].

Recalling one of the other \( S^1 \) representations, we see that \( W(t) \) is equivalent to (see [I5], p. 269)

\[
\theta(t) = w(t) \mod 2\pi
\]

(2.50)

and there is a simple physical interpretation of this representation. Consider the situation depicted in Figure 2.2. We have a unit circle in \( \mathbb{R}^2 \) with a straight line of infinite length tangent to it. We allow the line to perform a one-dimensional Brownian motion, fix the center of the circle, and require that there be no slipping at the point of
Figure 2.2: Brownian Motion on $S^1$

tangency. The line induces a rotation of the circle, and if the line moves a distance $w$, the circle rotates $w$ radians, and is thus in a position $\theta = w \mod 2\pi$ from its initial orientation.

Either by simple calculations or by referring to McKean's treatment of infinitesimal operators (note that Haar measure here is just $\frac{1}{2\pi} d\theta$), one can show that the probability density for $\theta$ satisfies the classical heat (forward diffusion, Fokker-Planck) equation on the circle:

$$\frac{\partial p_\theta}{\partial t} - \frac{1}{2} \frac{\partial^2 p_\theta}{\partial \xi^2} = 0 \quad (2.51)$$

with periodicity condition

$$p_\theta(\xi, t) = p_\theta(\xi + 2\pi, t) \quad (2.52)$$

and initial condition

$$p_\theta(\xi, 0) = \delta(\xi - \eta) \quad (2.53)$$
where the initial orientation of the circle is \( \eta \) radians from some reference position. The solution of this is well known \([B6], [P5]\), and is given by the two equivalent expressions

\[
p_\eta(\xi, t) = \frac{1}{\sqrt{2\pi t}} \sum_{n=-\infty}^{+\infty} e^{-\frac{(\xi+2\pi n-\eta)^2}{2t}} \tag{2.54}
\]

\[
= \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} e^{-n^2 t/2} \cos(n(\xi-\eta)) \tag{2.55}
\]

This solution is a special case of the transition density \( p_\theta(\xi, t; \eta, s) \) from time \( s \) to time \( t > s \), given that \( \theta(s) = \eta \). The transition density must satisfy (2.51) - (2.53) plus the backward equation

\[
\frac{\partial p_\theta}{\partial s} + \frac{1}{2} \frac{\partial^2 p_\theta}{\partial \eta^2} = 0 \tag{2.56}
\]

with appropriate periodicity in \( \eta \) and terminal condition

\[
p_\theta(\xi, t; \eta, t) = \delta(\eta-\xi) \tag{2.57}
\]

The solution is

\[
p_\theta(\xi, t; \eta, s) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} e^{-n^2(t-s)/2} \cos(n(\xi-\eta)) \tag{2.58}
\]

and clearly the solution (2.55) is just \( p_\theta(\xi, t; \eta, 0) \).

The density (2.54) or (2.55) will be called the folded normal density. We give it this name since the density form (2.54) clearly indicates that if \( x \) is a normal random variable of mean \( \eta \) and variance \( \gamma \) (such a density is denoted \( N(x; \eta, \gamma) \)), and if we let \( \theta = x \mod 2\pi \), then the density for \( \theta \) is obtained by "folding" \( N(x; \eta, \gamma) \) around the circle.
\[
    p_\theta(\xi) = \frac{1}{\sqrt{2\pi\gamma}} \sum_{n=-\infty}^{\infty} e^{-\frac{(\xi+2n\pi-\eta)^2}{2\gamma}} = F(\xi; \eta,\gamma) \quad (2.59)
\]

Levy [L4] and Perrin [P1] have done extensive work with the folded normal density. Using this folding procedure, we can solve (2.51) - (2.53) by first solving (2.51) on \( R^1 \) and then folding it around the circle. We can then consider the general problem of injecting \( x(t) = \eta^{1/2}w(t) + ft + \eta \) into \( S^1 \) \((\theta(t) = x(t)\text{mod}2\pi)\). The density for \( x(t) \) is \( N(x; \eta + ft, 0) \) and therefore the density for \( \theta(t) \) is \( F(x; \eta + ft, 0) \).

Similarly, one can consider the case in which \( Q \) and \( f \) are functions of time (see Chapter 4).

We note that if \( \theta \) and \( \phi \) are independent random variables on \([-\pi, \pi)\) with densities \( F(\theta; \eta_1, \gamma_1) \) and \( F(\phi; \eta_2, \gamma_2) \), respectively, then the density for \( \psi = (\theta + \phi)\text{mod}2\pi \) is \( F(\psi; \eta_1 + \eta_2, \gamma_1 + \gamma_2) \). As we shall see in the next chapter, the mode of the density \( F(\alpha; \eta, \gamma) \) is \( \eta \). Thus if \( \Theta, \Phi, \) and \( \Psi \) are the SO(2) representations of \( \theta, \phi, \) and \( \psi \), we have \( \Psi = \Theta \Phi \). Also, if we let \( N_\Theta, N_\Phi, \) and \( N_\Psi \) be the SO(2) representations of the modes of the densities for \( \theta, \phi, \) and \( \psi \), we have \( N_\Psi = N_\Theta N_\Phi \).

The concept of "injecting" or "folding" processes onto \( S^1 \) will be used and extended in the subsequent chapters. These chapters might be read with the following problem in mind. Suppose we receive a signal of the form

\[
    r(t) = \sin(\omega_c t + \phi(t) + v(t)) + N(t) \quad (2.60)
\]

where \( \omega_c \) is a carrier frequency, \( \phi \) is some sort of "message," \( v(t) \) is a random phase drift, and \( N(t) \) is additive channel noise. The problem is either to track the phase (synchronization) or to find out what the message is (demodulation).
With this problem as motivation, we proceed. Chapter 3 contains some basic analysis, while Chapter 4 extends McKean's results on $S^1$ and discusses how to filter out $v(t)$ in (2.60) if $N \equiv 0$. Chapter 5 will attack (2.60) as a special case of a general Fourier series problem, and in Chapter 6 we will discuss physical problems and the results of some simulations (the results of Chapters 3 and 4 and some of the results of Chapter 5 can be found in [L6]). Chapter 7 contains some generalizations motivated by these results and a few comments on the $SO(3)$ problem. In addition, some of the finite group work discusses in Chapters 9 and 10 have some relevance to a quantized version of (2.60).

It is one of the major hopes of the author that within the remaining chapters of this manuscript are contained some of the necessary tools for analyzing (2.60) and solving the associated demodulation or phase-tracking problem and the mathematical machinery needed to handle many conceptually similar problems.

To this end, we will formulate the $S^1$ estimation problems in a setting that is somewhat more complicated than that which is needed to solve these specific problems. In particular, the $[-\pi, \pi)$ representation of $S^1$ would suffice for the consideration of the conceptually straightforward problem of estimating the position of a point moving around a circle. However, the other $S^1$ representations suggest generalizations of these results or their application to specific practical problems. The $SO(2)$ representation leads to the formulation of a general matrix Lie group estimation problem that includes the rigid body motion problem (see Chapter 7). The complex number representation is useful for some applications (Chapter 6), and the $R^1/2\pi Z$ representation
suggests a generalization discussed in Appendix C. Also, the results we will derive using Fourier series have generalizations to other Lie groups (e.g. spherical harmonics, [G10], [B20], [W5], [R6]). It is with these considerations in mind that the reader should read the following chapters.
CHAPTER 3

PROBABILITY DENSITIES AND OPTIMAL ESTIMATES ON S^1

"While the square is closely connected with man and his constructions, with architecture, building forms, lettering, etc, the circle is related to the divine: since ancient times a simple circle has represented eternity, for it has neither beginning nor end. An old text says that God is a circle whose center is everywhere, but whose circumference is nowhere."

- Bruno Munari [MB]

3.1 Introduction

In the following chapters, we will study the properties of certain stochastic processes on the circle and will derive equations for probability distributions conditioned on observations. The question of optimal and suboptimal estimation is of central importance in these investigations. Thus it is necessary to study how one uses the knowledge of the probability distribution of the quantity to be estimated to choose an estimate that gives the "best" performance, as measured by some pre-determined figure of merit.

In this chapter we will present a number of results on the optimal estimation of random variables taking values on the circle as a prelude to the study of random processes on the circle. The motivation throughout this chapter is to study cost criteria that lead to simple optimal and suboptimal estimation schemes and to devise methods for analyzing the performance of these schemes. To this end, a number of special cases -- i.e. particular families of probability densities and performance measures -- are considered in detail.
In Section 3.2 we define the optimal estimation problem, discuss the use of Fourier series, and examine several specific criteria in detail. Section 3.3 contains both the analog of an $\mathbb{R}^1$ result for symmetric criteria and unimodal distributions and also some detailed analysis for the folded normal density.

3.2 The Optimal Estimation Problem and Fourier Series

Let $\theta$ be a random variable on $S^1$ (here we will identify $S^1$ with $[-\pi, \pi]$) with probability density $p(\theta)$, which clearly must be periodic with period $2\pi$. Also, we assume that we have an error function $\phi$ that is periodic with period $2\pi$. The estimation problem is to choose $\tilde{\theta} \in [-\pi, \pi]$ to minimize

$$J(\tilde{\theta}) = \mathbb{E}(\phi(\theta-\tilde{\theta})) = \int_{-\pi}^\pi \phi(\theta-\tilde{\theta}) p(\theta) d\theta$$  \hspace{1cm} (3.1)

Our initial analysis will be to determine necessary conditions for $\tilde{\theta}$ to minimize $J(\tilde{\theta})$. We assume that we have $p(\theta)$ and $\phi(\theta)$ in Fourier series form

$$p(\theta) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} [a_n \sin n\theta + b_n \cos n\theta]$$  \hspace{1cm} (3.2)

$$\phi(\theta) = d_0 + \sum_{n=1}^{\infty} [c_n \sin n\theta + d_n \cos n\theta]$$  \hspace{1cm} (3.3)

where

$$a_n = \frac{1}{\pi} \int_{-\pi}^\pi \sin n\theta p(\theta) d\theta = \frac{1}{\pi} \mathbb{E}(\sin n\theta)$$  \hspace{1cm} (3.4)

$$b_n = \frac{1}{\pi} \int_{-\pi}^\pi \cos n\theta p(\theta) d\theta = \frac{1}{\pi} \mathbb{E}(\cos n\theta)$$  \hspace{1cm} (3.5)

and $c_n$ and $d_n$ are defined analogously. We will have more to say about
the Fourier series decomposition of $p$ in Chapter 5.

A simple computation yields

$$J(\tilde{\theta}) = d_0 + \pi \sum_{n=1}^{\infty} [a_n(c_n \cos \tilde{n}\theta + d_n \sin \tilde{n}\theta)$$

$$+ b_n(d_n \cos \tilde{\theta} - c_n \sin \tilde{\theta})]$$

(3.6)

Thus, necessary conditions for a local minimum are

$$\frac{d}{d\tilde{\theta}} J(\tilde{\theta}) = \pi \sum_{n=1}^{\infty} [na_n(d_n \cos \tilde{n}\theta - c_n \sin \tilde{n}\theta)$$

$$- nb_n(d_n \sin \tilde{n}\theta + c_n \cos \tilde{n}\theta)] = 0$$

(3.7)

$$\frac{d^2}{d\tilde{\theta}^2} J(\tilde{\theta}) = \pi \sum_{n=1}^{\infty} [-n^2a_n(d_n \sin \tilde{n}\theta + c_n \cos \tilde{n}\theta)$$

$$+ n^2b_n(c_n \sin \tilde{\theta} - d_n \cos \tilde{\theta})] \geq 0$$

(3.8)

Explicit solution of (3.7) and (3.8) is possible only for certain error functions. Note that the sums in (3.7) and (3.8) are finite if the sum in either (3.2) or (3.3) is. The following two examples indicate how this finiteness can simplify the problem of finding the optimal estimate.

**Example 3.1:** Consider the error function

$$\phi(\theta) = 1 - \cos \theta$$

(3.9)

This criterion has also been considered in [M14] and [B9]. For this function we have

$$J(\tilde{\theta}) = 1 - \pi(a_1 \sin \tilde{\theta} + b_1 \cos \tilde{\theta})$$

(3.10)

and the necessary conditions are
\[ s_1 \cos \tilde{\theta} - b_1 \sin \tilde{\theta} = 0 \]  
\[ s_1 \sin \tilde{\theta} + b_1 \cos \tilde{\theta} \geq 0 \]  

If \( a_1 = b_1 = 0 \), \( J(\tilde{\theta}) \) is independent of \( \tilde{\theta} \). In any other case, there are two inequivalent solutions to (3.11), where two solutions are considered equivalent if they differ by a multiple of \( 2\pi \) (i.e. (3.11) determines two equivalence classes in \( \mathbb{R}/2\pi \mathbb{Z} \)). Representatives of the two solutions are

\[ \tilde{\theta} = \tan^{-1}(a_1/b_1), \quad \tan^{-1}(a_1/b_1) + \pi \]

\[ = \tan^{-1} \left( \frac{\varepsilon(\sin \tilde{\theta})}{\varepsilon(\cos \tilde{\theta})} \right), \quad \tan^{-1} \left( \frac{\varepsilon(\sin \tilde{\theta})}{\varepsilon(\cos \tilde{\theta})} \right) + \pi \]  

where \( \tan^{-1} : [-\infty, \infty] \to [-\pi/2, \pi/2] \). Examination of (3.12) yields a method for choosing the proper solution:

\[ a_1 > 0, \ b_1 > 0 \implies \text{Choose solution in first quadrant} \]
\[ a_1 > 0, \ b_1 < 0 \implies \text{Choose solution in second quadrant} \]
\[ a_1 < 0, \ b_1 < 0 \implies \text{Choose solution in third quadrant} \]
\[ a_1 < 0, \ b_1 > 0 \implies \text{Choose solution in fourth quadrant} \]

With these choices, it is easy to see that if \( \tilde{\theta}_0 \) is the optimal value

\[ \sin \tilde{\theta}_0 = \frac{a_1}{\sqrt{a_1^2 + b_1^2}} \]  
\[ \cos \tilde{\theta}_0 = \frac{b_1}{\sqrt{a_1^2 + b_1^2}} \]  

and

\[ J(\tilde{\theta}_0) = 1 - \pi \sqrt{a_1^2 + b_1^2} \]
We note that this particular error function has some very appealing properties. First of all, since the optimal estimate is an explicit function only of the first mode of the probability distribution (i.e. $a_1$ and $b_1$), the computational procedure to determine $\tilde{\theta}_0$ is quite simple. In addition, there is strong physical motivation for using this criterion.

First note that for small values of $\theta$

$$1 - \cos \theta = \frac{1}{2} \theta^2$$  \hspace{1cm} (3.17)

Thus this is, at least locally, a type of least squares criterion. In fact, suppose that $x_1$ and $x_2$ are real-valued random variables such that

$$x_1^2 + x_2^2 = 1$$  \hspace{1cm} (3.18)

That is, there exists a random variable $\theta$ with

$$x_1 = \sin \theta \quad x_2 = \cos \theta$$  \hspace{1cm} (3.19)

Suppose we wish to choose $\tilde{x}_1$ and $\tilde{x}_2$ to minimize

$$J = \frac{1}{2} \varepsilon \left[ (x_1 - \tilde{x}_1)^2 + (x_2 - \tilde{x}_2)^2 \right]$$  \hspace{1cm} (3.20)

subject to the constraint

$$\tilde{x}_1^2 + \tilde{x}_2^2 = 1$$  \hspace{1cm} (3.21)

That is

$$\tilde{x}_1 = \sin \tilde{\theta} \quad \tilde{x}_2 = \cos \tilde{\theta}$$  \hspace{1cm} (3.22)

and, substituting into (3.20), we have

$$J = \varepsilon (1 - \cos (\theta - \tilde{\theta}))$$  \hspace{1cm} (3.23)
Thus this error function is a constrained least squares criterion.

We note that although the higher modes do not affect the estimate directly, we shall see in Chapter 5 that these coefficients have an indirect effect on \( \theta \). Specifically we shall find that in dealing with random processes and time-varying densities, the time rates of change of \( a_1 \) and \( b_1 \) depend, in general, on the other coefficients.

**Example 3.2:** We wish to consider a criterion that involves some of the higher coefficients directly. Consider

\[
\phi(\theta) = (1 - \cos \theta)^2 = \frac{3}{2} - 2\cos \theta + \frac{1}{2} \cos 2\theta
\]  

(3.24)

In this case, the necessary conditions become

\[
(a_2 \cos 2\tilde{\theta} - b_2 \sin 2\tilde{\theta}) - 2(a_1 \cos \tilde{\theta} - b_1 \sin \tilde{\theta}) = 0
\]  

(3.25)

\[
(a_1 \sin \tilde{\theta} + b_1 \cos \tilde{\theta}) - (a_2 \sin 2\tilde{\theta} + b_2 \cos 2\tilde{\theta}) \geq 0
\]  

(3.26)

We make the substitutions

\[
\cos 2\tilde{\theta} = (1 - 2\sin^2 \tilde{\theta})
\]  

(3.27)

\[
\sin 2\tilde{\theta} = 2\sin \tilde{\theta} \cos \tilde{\theta}
\]  

(3.28)

\[
x = \sin \tilde{\theta}
\]  

(3.29)

\[
\cos \tilde{\theta} = \pm \sqrt{1 - x^2}
\]  

(3.30)

Then (3.25) becomes

\[
a_2 (1 - 2x^2) + 2b_1 x = \pm 2(b_2 x \sqrt{1 - x^2} + a_1 \sqrt{1 - x^2})
\]  

(3.31)

Squaring both sides and dividing by 4, we get
\[(a_2^2+b_2^2)x^4 + 2(a_1b_2-a_2b_1)x^3 + (a_2^2+b_1^2-a_2^2-b_2^2)x^2 + 4a_2b_1-2a_1b_2)x + \left(\frac{a_2^2}{4} - a_2^2\right) = 0\]  

(3.32)

We can then solve for the four roots \(x_1, x_2, x_3, x_4\) of (3.32) and keep only the real roots of magnitude < 1. For each of these \(x_i\) we compute

\[\tilde{\theta}_{11} = \sin^{-1} x_i \quad -\frac{\pi}{2} \leq \tilde{\theta}_{11} \leq \frac{\pi}{2}\]  

(3.33)

\[\tilde{\theta}_{12} = \pi - \tilde{\theta}_{11}\]  

(3.34)

We can then compute \(J(\tilde{\theta}_{ij})\) and \(\tilde{\theta}_o\) is the \(\tilde{\theta}_{ij}\) that yields the smallest value of \(J(\tilde{\theta}_{ij})\). Note that the optimal value \(\tilde{\theta}_o\) must be one of the \(\tilde{\theta}_{ij}\).

The process of choosing \(\tilde{\theta}_o\) can be even further simplified. First, using (3.6), we have

\[J(\tilde{\theta}) = \frac{3}{2} + \pi\left[\frac{1}{2} (a_2 \sin 2\tilde{\theta} + b_2 \cos 2\tilde{\theta}) - 2(a_1 \sin \tilde{\theta} + b_1 \cos \tilde{\theta})\right]\]  

(3.35)

Then some straightforward computation yields the following procedure.

Given \(x_i\) real and of magnitude \(\leq 1\), compute

\[J_o(x_i) = -4a_1x_i + b_2(1-2x_i^2) + (2a_2x_i\sqrt{1-x_i^2} - 4b_1\sqrt{1-x_i^2})\]  

(3.36)

\[J_1(x_i) = -4a_1x_i + b_2(1-2x_i^2) - (2a_2x_i\sqrt{1-x_i^2} - 4b_1\sqrt{1-x_i^2})\]  

(3.37)

choose \(i_o\) and \(k_o\) that minimize \(J_k(x_i)\). Then

\[\tilde{\theta}_o = k_o\pi + (-1)^k \sin^{-1} x_{i_o}\]  

(3.38)

where, as before, \(\sin^{-1} : [0,1] \to [-\pi/2, \pi/2]\). Note that we needn't compute both \(J_o(x_i)\) and \(J_1(x_i)\). Examining (3.36) and (3.37), we see that \(J_o(x_i) \leq J_1(x_i)\) if and only if
\[-(1)^n(2a_2x_1\sqrt{1-x_1^2} - 4b_1\sqrt{1-x_1^2}) \leq 0 \quad (3.39)\]

Thus in example 3.2 we have reduced the problem of finding the optimal estimate to the solution of a quartic polynomial equation and the calculation of several functions — a procedure that can be done easily by computer. However, even when we add in just the second mode, the increase in complexity is such that no closed form for the optimal error in terms of the Fourier coefficients is available.

As can be seen, the error analysis and computational procedures become increasingly more difficult as the number of nonzero Fourier coefficients in \(\phi\) increases. Since quartic polynomial equations are the highest order polynomial equations for which there exist solution formulae (see [F1]), we see that adding any terms beyond the second mode coefficients will lead to greatly increased computational requirements just to perform the first step of the calculations. Thus, any such criteria must be picked carefully, and, the Fourier series for the criteria may have to be truncated (this may not be too severe a requirement, since it is known that the nth Fourier coefficients of a function satisfying certain differentiability conditions go as \(\frac{1}{n^2}\) or \(\frac{1}{n}\) (see [K10]).

Two other criteria deserve mention. Let \(\rho\) be the standard distance function (Riemannian metric) on \(S^1\) — i.e., the distance \(\rho\) between two points on the circle is the arc length of the shortest path (geodesic line) joining them. If we restrict \(\theta_1\) and \(\theta_2\) to take on values in \([\pi, \pi]\), we have

\[\rho(\theta_1, \theta_2) = \min(|\theta_1-\theta_2|, 2\pi-|\theta_1-\theta_2|) \quad (3.40)\]

Then two possible error functions, which appear to be direct analogs of \(|x|\) and \(x^2\) on the real line, are
\[ \phi_1(\theta) = \rho(\theta,0) \]  
\[ \phi_2(\theta) = \rho^2(\theta,0) \]  

The Fourier series decompositions of these functions are

\[ \phi_1(\theta) = \frac{\pi}{2} - \sum_{n=0}^{\infty} \frac{4}{(2n+1)^2} \cos(2n+1)\theta \]  
\[ \phi_2(\theta) = \frac{\pi^2}{3} + \sum_{n=1}^{\infty} \frac{4(-1)^n}{n^2} \cos n\theta \]

For the case of \( \phi_2 \), we can write the necessary conditions in a different form. For \( \phi = \rho^2 \), one can show that the cost function is

\[ J(\tilde{\theta}) = \mathcal{E}(\rho^2(\theta-\tilde{\theta},0)) = \int_{-\pi+\theta}^{\pi+\theta} (\theta-\tilde{\theta})^2 p(\theta)d\theta \]  

Using Leibnitz's rule [H6] and the periodicity of \( p \), we have the following necessary conditions for optimality:

\[ \frac{d}{d\tilde{\theta}} J(\tilde{\theta}) = 2\tilde{\theta} - 2 \int_{-\pi+\theta}^{\pi+\theta} \theta p(\theta)d\theta = 0 \]  

\[ \frac{d^2}{d\tilde{\theta}^2} J(\tilde{\theta}) = 2 - 4\pi p(\tilde{\theta}+\pi) \geq 0 \]

Equations (3.46) and (3.47) offer an alternate method for solving for \( \tilde{\theta}_o \). Note that (3.46) resembles the necessary condition for the least squares estimate on \( R^1 \). In that case the mean value

\[ \tilde{x}_o = \mathcal{E}(x) = \int_{-\infty}^{\infty} xp(x)dx \]

is the optimal estimate. However, in this case, essentially because of the topological difference between \( S^1 \) and \( R^1 \), the integral
\[
\int_{-\pi+\alpha}^{\pi+\alpha} \theta p(\theta) d\theta
\]

is not independent of \(\alpha\) and thus cannot be called \(\mathcal{E}(\theta)\).

3.3 Symmetric Criteria, Unimodal Distributions, and the Folded Normal Density

At the start of this section we wish to consider the class of symmetric, nondecreasing cost functions -- i.e. functions \(\phi : S^1 \rightarrow \mathbb{R}^1\) which satisfy

\[
0 \leq \phi(\theta) = \phi(-\theta)
\]

\[
0 \leq \rho(\theta_1,0) < \rho(\theta_2,0) \Rightarrow \phi(\theta_1) > \phi(\theta_2)
\]

Some examples of such error functions are \(\rho, \rho^2, (1-\cos\theta),\) and \((1-\cos\theta)^2\).

We also wish to consider the special class of unimodal, mode-symmetric probability density functions -- i.e. densities of the form \(p : S^1 \rightarrow [0, \infty)\) with a unique maximum at \(\eta\), such that

\[
p(\eta+\theta) = p(\eta-\theta) \quad \forall \theta
\]

As the following theorem demonstrates, under these conditions the mode of the density is the optimal estimate.

**Theorem 3.1:** Given an error function \(\phi\) that satisfies (3.49) and (3.50) and a unimodal mode-symmetric probability density function \(p\), then

\[
\mathcal{E}(\phi(\theta-\eta)) \leq \mathcal{E}(\phi(\theta-\alpha)) \quad \forall \alpha
\]

where \(p\) has its maximum at \(\eta\).

**Proof:** The theorem follows immediately from results on similarly ordered functions and the rearrangement inequalities. The basic result for real-valued functions defined on \(\mathbb{R}^1\) is contained in Hardy,
Littlewood, and Pólya [H7, Thm. 378] and Szego and Pólya [S4, p. 183].
The result for $S^1$ is obtained by making only minor changes in these
proofs.

We remark that from the symmetry of the problem, $\phi$ has its global
maximum at $\pi$ and $p$ its global minimum at $\eta+\pi$. Thus

$$\phi(\theta-\eta+\pi) > \phi(\theta-\pi) \quad \forall \alpha$$

(3.53)

It should be noted that Theorem 3.1 is the $S^1$ analog of a result
of Sherman [S5], [S6]. Also, the same result is true if a probability
density doesn't exist, but the probability measure is unimodal at and
symmetric about some point $\eta$. Here we define these concepts as follows:
let $\theta$ be a random variable on $S^1$, and define the distribution function
$F : [\pi,\pi] + [0,1]$ by

$$F(\alpha) = \Pr(\theta \in [-\pi,\alpha])$$

(3.54)

Then $F$ is unimodal at and symmetric about $0$ if it is convex for
$\alpha \in [-\pi,0)$, and if

$$F(\alpha) = 1 - F(-\alpha)$$

(3.55)
at each continuity point of $F$ (see [S5]).

The next result shows that Theorem 3.1 holds for the folded
normal density.

**Theorem 3.2:** The folded normal density

$$F(\theta; \eta, \gamma) = \frac{1}{\sqrt{2\pi} \gamma} \sum_{n=-\infty}^{\infty} e^{-\frac{(\theta+2n\pi-\eta)^2}{2\gamma}}$$

$$= \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} e^{-n^2 \gamma/2} \cos(\theta-\eta)$$

(3.56)
is unimodal with mode at $\theta = \eta \mod 2\pi$, and is symmetric about $\eta \mod 2\pi$.

Proof: Since $\cos < 1$, the second form of $F$ in (3.56) yields

$$F(\theta; \eta, \gamma) \leq \frac{1}{2\pi} + \sum_{n=1}^{\infty} e^{-n^2 \gamma/2} = F(\eta; \gamma)$$  (3.57)

Since $F(\theta; \gamma) = F(\theta-\eta; 0, \gamma)$, we need only show that $F(\theta; 0, \gamma)$ is symmetric about 0 and monotone decreasing as $\phi(\theta, 0)$ increases.

Symmetry is obvious ($\cos \theta = \cos(-\theta)$) and monotonicity will follow if we can show

$$\frac{\partial F}{\partial \theta}(\theta; 0, \gamma) < 0 \quad \theta \in (0, \pi)$$  (3.58)

$$\frac{\partial F}{\partial \theta}(\theta; 0, \gamma) > 0 \quad \theta \in (-\pi, 0)$$  (3.59)

We now remark that the properties of $F(\theta; 0, \gamma)$ have been studied extensively, since it is a theta function. See Bellman [B6] and Whittaker and Watson [W6] for discussions of some properties of theta functions.

Using the notation of [B6, pp. 2, 42], we have

$$F(\theta; 0, \gamma) = \frac{1}{2\pi} \Theta_4\left(\frac{\theta+\pi}{2}, \frac{1}{2} \gamma\right)$$

$$= k \prod_{n=1}^{\infty} \left(1+2q^{2n-1}\cos \theta+q^{4n-2}\right)$$  (3.60)

where

$$q = e^{-\gamma/2}$$  (3.61)

$$k = \frac{1}{2\pi} \prod_{n=1}^{\infty} (1-q^{2n})$$  (3.62)

Using the fact that $F > 0$ (this is clear from the first form
of $F$ in (3.56) and (3.60), we have

$$ \frac{\partial F}{\partial \theta}(\theta;0,\gamma) = \left[ \sum_{n=1}^{\infty} \frac{2q^{2n-1}}{1+2q^{2n-1}\cos\theta+q^{4n-2}} \right] \sin\theta $$

(3.63)

It is readily seen that the term in brackets on the right-hand side of (3.63) is positive for all values of $\theta$, and thus (3.58) and (3.59) are correct.

Some work along these lines has been done by Perrin [P1]. See [P1] for a discussion of other relevant properties of theta functions, hypergeometric functions, Legendre polynomials, and Tchebycheff polynomials.

Note that the symmetry requirements of Theorem 3.1 are necessary. For instance, if $\phi$ is not symmetric, the mode of the density need not be the optimal estimate even if all the other hypotheses of Theorem 3.1 are satisfied. As an example, consider the function

$$ \phi(\theta) = \begin{cases} 
\theta & 0 \leq \theta \leq \pi \\
\frac{\theta^2}{\pi} & -\pi \leq \theta \leq 0
\end{cases} $$

(3.64)

Suppose our distribution is $F(\theta;\eta,\gamma)$. Then straightforward computations show that $\tilde{\theta} = \eta$ does not satisfy necessary condition (3.7) and thus is not the optimal estimate.

If we restrict our attention to the folded normal density, we can consider a larger class of error functions than those satisfying (3.49) and (3.50). We shall drop the symmetry requirement (3.49) but shall require $\phi$ to be monotone nonincreasing on $[-\pi,0)$ and monotone non-decreasing on $(0,\pi)$ (that is, (3.50) holds if $\theta_1, \theta_2 \in [-\pi,0)$ or if
\( \theta_1, \theta_2 \in (0, \pi) \). For such a \( \phi \), as mentioned previously, the mode \( \eta \) of \( F(\theta; \eta, \gamma) \) need not be the optimal estimate; however, for this discussion we will take it as our estimate.

In analyzing random variables on \( \mathbb{R}^1 \), the variance plays an important role, in that intuitively, as the variance decreases, our knowledge increases, so our estimate should get "better." For the folded normal, \( F(\theta; \eta, \gamma) \), \( \gamma \) is not really a variance, but it clearly is a measure of how sharply peaked the density is (see Section 2.4, where we show how \( F(\theta; \eta, \gamma) \) is derived from \( N(x; \eta, \gamma) \)). The following theorem reveals that we have an \( S^1 \) analog of the notion that our estimate improves as \( \gamma \) decreases.

**Theorem 3.3:** Let \( \phi \) be monotone nonincreasing on \([-\pi, 0)\) and monotone nondecreasing on \((0, \pi)\). Then, if \( p(\theta) = F(\theta; \eta, \gamma) \), \( \mathcal{E}(\phi(\theta-\eta)) \) is an increasing function of \( \gamma \) -- i.e.,

\[
\frac{d}{d\gamma} \mathcal{E}(\phi(\theta-\eta)) > 0
\]  

(3.65)

**Proof:** Writing

\[
\phi(\theta) = d_o + \sum_{n=1}^{\infty} c_n \sin n\theta + d_n \cos n\theta \]  

(3.66)

and referring to (3.6), we see that

\[
\mathcal{E}(\phi(\theta-\eta)) = d_o + \sum_{n=1}^{\infty} d_n e^{-n^2 \gamma/2}
\]  

(3.67)

Thus we get the same error if we compute \( \mathcal{E}(\psi(\theta-\eta)) \) where \( \psi \) is the function satisfying (3.49) and (3.50) given by

\[
\psi(\theta) = \frac{1}{2} [\phi(\theta) + \phi(-\theta)]
\]  

(3.68)
We note that the property that only the even part of $\phi$ (i.e. the $\cos n\theta$ coefficients) is of importance in computing $E(\phi(\theta-\eta))$ is not peculiar to the folded normal density. In fact, if $p(\theta)$ is any density symmetric about $\eta$, then $E(\phi(\theta-\eta)) = E(\psi(\theta-\eta))$ where $\psi$ is defined by (3.68).

We now see that it is enough to prove the theorem for $\phi$ satisfying (3.49) and (3.50). In this case, $\eta$ is the optimal estimate, and

$$E(\phi(\theta-\eta)) = \int_{-\pi}^{\pi} \phi(\theta-\eta) p(\theta; \eta, \gamma) d\theta$$

$$= \int_{\pi}^{\pi} \phi(\theta) p(\theta; 0, \gamma) d\theta$$

$$= 2 \int_{0}^{\pi} \phi(\theta) p(\theta; 0, \gamma) d\theta$$

(3.69)

Then (3.65) will hold if

$$\int_{0}^{\pi} \phi(\theta) \frac{\partial}{\partial \gamma} p(\theta; 0, \gamma) d\theta \geq 0$$

(3.70)

Suppose we can show that there exists $\xi \in [0, \pi)$ such that

$$\frac{\partial}{\partial \gamma} p(\theta; 0, \gamma) < 0 \quad \theta \in [0, \xi)$$

(3.71)

$$\frac{\partial}{\partial \gamma} p(\xi; 0, \gamma) = 0$$

(3.72)

$$\frac{\partial}{\partial \gamma} p(\theta; 0, \gamma) > 0 \quad \theta \in (\xi, \pi]$$

(3.73)

Then, since

$$\phi(\theta) \leq \phi(\xi) \quad \theta \in [0, \xi]$$

(3.74)

$$\phi(\theta) \geq \phi(\xi) \quad \theta \in [\xi, \pi]$$

(3.75)
we have

\[ \int_0^\pi \phi(\theta) \frac{\partial}{\partial \gamma} F(\theta;0,\gamma) d\theta \geq \phi(\xi) \frac{d}{d\gamma} \int_0^\pi F(\theta;0,\gamma) d\theta \]

\[ = \phi(\xi) \frac{d}{d\gamma} \left( \frac{1}{2} \right) = 0 \quad (3.76) \]

and we get strict inequality if \( \phi \) is not a constant.

Now, from Section 2.4 we know that

\[ \frac{\partial}{\partial \gamma} F(\theta;0,\gamma) = \frac{1}{2} \frac{\partial^2}{\partial \theta^2} F(\theta;0,\gamma) \quad (3.77) \]

and the theorem will be proved once we prove the following lemma, which yields more information about the shape of the folded normal density (and thus points out another analogy with the \( R^1 \)-normal density case).

**Lemma 3.1:** For an arbitrary but fixed value of \( \gamma > 0 \), there exists \( \xi \in [0,\pi) \) such that

\[ \frac{\partial^2}{\partial \theta^2} F(\theta;0,\gamma) < 0 \quad \theta \in (0,\xi) \quad (3.78) \]

\[ \frac{\partial^2}{\partial \theta^2} F(\theta;0,\gamma) = 0 \quad (3.79) \]

\[ \frac{\partial^2}{\partial \theta^2} F(\theta;0,\gamma) > 0 \quad \theta \in (\xi,\pi] \quad (3.80) \]

That is, \( F \) has a unique inflection point (at \( \xi \)) on \([0,\pi)\).

**Proof:** We use the form of \( F \) given in (3.60). We compute

\[ \frac{\partial^2}{\partial \theta^2} \frac{F(\theta;0,\gamma)}{F(\theta;0,\gamma)} = -A \cos \theta + B \sin^2 \theta \quad (3.81) \]
\[ A = \sum_{n=1}^{\infty} \frac{2q^{2n-1}}{(1+2q^{2n-1}\cos \theta + q^{4n-2})} \]  

\[ B = \sum_{n \neq m} \frac{4q^{2(n+m-1)}}{(1+2q^{2n-1}\cos \theta + q^{4n-2})(1+2q^{2m-2}\cos \theta + q^{4m-2})} \]

and a simple computation yields

\[ \frac{\partial^2}{\partial \theta^2} F(0; 0, \gamma) < 0 \]  

(3.84)

\[ \frac{\partial^2}{\partial \theta^2} F(\theta; 0, \gamma) > 0 \quad \forall \theta \in [\frac{\pi}{2}, \pi) \]  

(3.85)

\[ \frac{\partial^2}{\partial \theta^2} \frac{\partial^2}{\partial \theta^2} F(\theta; 0, \gamma) > 0 \quad \forall \theta \in (0, \frac{\pi}{2}) \]  

(3.86)

These inequalities imply that there is a \( \xi \in (0, \frac{\pi}{2}) \) such that

\[ \frac{\partial^2}{\partial \theta^2} F(\theta; 0, \gamma) < 0 \quad \theta \in [0, \xi) \]  

(3.87)

\[ \frac{\partial^2}{\partial \theta^2} F(\xi; 0, \gamma) = 0 \]  

(3.88)

Then for \( \alpha \in (\xi, \frac{\pi}{2}) \), we have from (3.86)

\[ \frac{\partial^2}{\partial \theta^2} \frac{F(\alpha; 0, \gamma)}{F(\alpha; 0, \gamma)} > \frac{\partial^2}{\partial \theta^2} \frac{F(\xi; 0, \gamma)}{F(\xi; 0, \gamma)} \]  

(3.89)

or

\[ \frac{\partial^2}{\partial \theta^2} F(\alpha; 0, \gamma) > 0 \]  

(3.90)

and the lemma and theorem are proved.
Note that by symmetry $F$ has a unique inflection point at $-\xi$ on the interval $(-\pi, 0]$, and in fact $\xi \in (-\pi/2, 0)$. We also note that Theorem 3.3 is the $S^1$ analog of an $R^1$ result proved by J.L. Brown [B7].

Another family of densities of importance in the subsequent chapters is the class of infinite linear combinations of folded normal densities

$$p(\theta) = \sum_{n=1}^{\infty} c_n F(\theta; \eta_n, \gamma_n)$$  \hspace{1cm} (3.91)

$$\sum_{n=1}^{\infty} c_n = 1 \hspace{1cm} \gamma > 0$$  \hspace{1cm} (3.92)

It should be noted that it has been shown ([L7], [A2]) that the set of densities given by (3.91) with only finitely many nonzero $c_n$'s is dense in $L^1(-\pi, \pi)$, and this is still true if all the $\gamma_n$'s are the same. We do not make the assumption $c_n = 0$ $\forall n > N$, and the reason for this can be seen in Chapter 4.

When $p(\theta)$ is given by (3.91), we do not have neat, simple estimation results, as in the case when $p(\theta) = F(\theta; \eta, \gamma)$. We will examine the optimal estimation problem for two error functions, $1 - \cos \theta$ and $p^2(\theta, 0)$.

As discussed in Section 3.2, to minimize $\mathcal{E}(1 - \cos(\theta - \tilde{\theta}))$ with respect to $\tilde{\theta}$ we need only know the lowest mode Fourier coefficients, $a_1$ and $b_1$. We have

$$a_1 = \frac{1}{\pi} \sum_{n=1}^{\infty} c_n e^{-\gamma_n/2} \sin \eta_n$$  \hspace{1cm} (3.93)
\[ b_1 = \frac{1}{\pi} \sum_{n=1}^{\infty} c_n e^{-\gamma_n/2 \cos \eta_n} \]  

(3.94)

and (assuming \( a_1 \) and \( b_1 \) are not both zero) the optimal estimate \( \tilde{\theta}_o \) is given by (3.13) and the optimal cost is

\[ \mathcal{E}(1-\cos(\theta-\tilde{\theta}_o)) = 1 - [\left( \sum_{n=1}^{\infty} c_n e^{-\gamma_n/2 \sin \eta_n} \right)^2 + \left( \sum_{n=1}^{\infty} c_n e^{-\gamma_n/2 \cos \eta_n} \right)^2]^{1/2} \]  

(3.95)

In general, this optimal error is not an increasing function of each of the \( \gamma_n \) individually. However, if all of the \( \gamma_n \) are equal to some number \( \gamma \) (as is the case in the problem discussed in Chapter 4), it is easy to see that the optimal estimate is an increasing function of \( \gamma \).

In the case of \( \rho^2 \), we recall from Section 3.2 that it was necessary to evaluate

\[ \int_{-\pi+\alpha}^{\pi+\alpha} \theta p(\theta) d\theta \]

as a function of \( \alpha \). For a folded normal density, \( P(\theta; \eta, \gamma) \), we have

\[ \int_{-\pi+\alpha}^{\pi+\alpha} \theta p(\theta) d\theta = \eta - \sum_{k=-\infty}^{+\infty} \frac{(2k+1)\pi}{2k\pi} \int_{2k-1}^{2k+1} N(\theta; \eta-\alpha, \gamma) d\theta \]  

(3.96)

The second term on the right-hand side of (3.96) involves various values of the error function, \( \text{erf} \), \( [H6] \), and can be tabulated as a function of \( \eta-\alpha \) and \( \gamma \). Then, if we call this term \( g(\eta-\alpha, \gamma) \), in the case where \( p(\theta) \) is given by (3.91), necessary conditions for the optimal estimate are
\[ \tilde{\theta}_o = \sum_{n=1}^{\infty} c_n [\eta_n - g(\eta_n - \tilde{\theta}_o, \gamma_n)] \quad (3.97) \]

\[ 1-2\pi \sum_{n=1}^{\infty} c_n F(\tilde{\theta}_o + \pi; \eta_n, \gamma_n) \geq 0 \quad (3.98) \]

There does not appear to be a simple formula for the optimal cost, nor is it clear whether or not the optimal cost is a monotone increasing function of the \( \gamma_n \) or of \( \gamma \), in the case in which all of the \( \gamma_n \) equal \( \gamma \).

One final note on computing (3.96). We have two infinite sum expressions (3.56) for \( F(\theta; \eta, \gamma) \), and it is clear that one converges more quickly for large values of \( \gamma \), while the other for small values of \( \gamma \). Thus, using this fact, we can speed up computations, such as those required in (3.98), by choosing the proper expression for \( F(\theta; \eta, \gamma) \).
CHAPTER 4

A BILINEAR ESTIMATION PROBLEM ON $S^1$

"And the seasons they go round and round
And the painted ponies go up and down
We're captive on the carousel of time
We can't return we can only look behind
From where we came
And go round and round and round
In the circle game."

- "The Circle Game," Joni Mitchell [M2]

4.1 Introduction

In this chapter we will extend the injection mapping discussed in Section 2.4 and use the extended mapping to define a class of signal and observation processes on the circle. In Section 4.2 we define a general class of continuous-time estimation problems on $S^1$ and discuss the propagation of conditional probability densities. We remark that the approach of this section is measure-theoretic in nature, and the major results are a Bucy-type [B7] representation theorem (Theorem 4.1) and a Kushner-type [K5] stochastic partial differential equation (equations (4.39) - (4.42)) for a conditional density. In Section 4.3 we apply these general results to a special class of problems, for which we obtain explicit equations for the optimal estimator. The form of the estimator is especially appealing (see Figure 4.2). As mentioned at the end of Chapter 2, the $S^1$ estimation problem is conceptually quite simple, and we do not need all of the elaborate framework that we have developed. However, this framework -- in particular the SO(2) formulation of the problem -- immediately suggests a generalization of this approach.
This generalization includes the very important problem of rigid body rotation, which involves the matrix Lie group SO(3).

In the first part of Section 4.2, we will describe the dynamical system to be considered in this chapter and will discuss some questions of practical importance. The reader who is not interested in the mathematical details is then referred to Section 4.3, in which we discuss a special case, a multidimensional version of the problem, and several examples. Also, the reader is referred to Chapter 7, in which we take a first step toward generalizing the results of the present chapter.

A discrete time analog of the problems examined in Sections 4.2 and 4.3 is discussed in Section 4.4, and some of the technical difficulties encountered are detailed in Appendices B and C. We remark that Section 4.4 is of particular interest, since it reveals a striking example of a class of estimation problems for which the continuous time problem is far simpler than the discrete time problem.

Some applications of the results of this chapter are contained in Chapter 6. Also, in Chapter 6 we will present some numerical results.

4.2 Signal Processes, Observation Processes, and Conditional Probability Distributions

As mentioned in Section 4.1, the methods and results of this section are quite abstract. The motivation for this is not only to be mathematically precise, but also to develop mathematical machinery that may eventually be applied to a more general class of estimation problems.
In this chapter, we will use Itô differential equations to describe the signal and observation processes. We do this in order to make use of some extremely powerful results concerning conditional Markov processes (see, for instance, [J2], [B7], [K4], [K5], and [S8]). However, as we shall see, this formulation leads to a situation in which we may have to differentiate the observation. This problem will be discussed later in this chapter and in more detail in Chapter 6.

Before defining the signal and observation processes to be analyzed, we introduce the following notation:

\[ C^T_1 = \text{The family of continuous real valued functions, } a, \text{ on } [0,T] \text{ such that } a(0) = 0 \]

\[ B^T_1 = \text{The Borel } \sigma\text{-algebra of } C^T_1 \]

\[ C^T_2 = \text{The family of continuous } 2\times2 \text{ orthogonal-matrix-valued functions, } A, \text{ on } [0,T] \text{ such that } A(0) = I \]

\[ B^T_2 = \text{The Borel } \sigma\text{-algebra of } C^T_2 \]

The definition of \( B^T_1 \) is given in Wong [W5, p. 53]. The idea is essentially the following: let \( C^T \) be the set of all real-valued continuous functions on \([0,T]\), and define for any \( t \in [0,T] \) the function \( X_t : C^T \to \mathbb{R}^1 \)

\[ X_t(f) = f(t) \quad (4.1) \]

Then the appropriate Borel \( \sigma \)-algebra is the smallest \( \sigma \)-algebra on \( C^T \) for which \( X_t \) is measurable for all \( t \in [0,T] \) (the \( \sigma \)-algebra on \( \mathbb{R}^1 \) is the usual Borel \( \sigma \)-algebra generated by the open intervals \((a,b))\).
Extensions of this definition to $C^1_T$ and vector- and matrix-valued functions (e.g. $C^T_2$) are obvious. Throughout this section we will use lower case letters to denote elements of $C^1_T$ and upper case letters for elements of $C^T_2$.

We now define a map $J : C^1_T \rightarrow C^T_2$ that is clearly related to the product integral on $S^1$ (see eqns. (2.47) and (2.48)).

$$(J(a))(t) = \exp(R_a(t)) = \begin{bmatrix} \cos(a(t)) & \sin(a(t)) \\ -\sin(a(t)) & \cos(a(t)) \end{bmatrix} \quad t \in [0,T] \quad (4.2)$$

It is clear that $J$ is $\mathcal{B}^T_1$-measurable. It is claimed that $J$ is also bijective. To show this, we first note that $J$ is the same as the map

$$a(t) \mapsto a(t) \mod 2\pi \quad \Delta \theta_a(t) \quad (4.3)$$

where we identify $SO(2)$ with $[-\pi, \pi)$. Consider Figure 4.1 which illustrates the representation of $J$ given in (4.3). The processes of breaking up $a(t)$ or piecing $\theta_a(t)$ together are clearly inverses of one another. Intuitively what this says is the following: if we continuously watch a rotating object, we know not only its present orientation but also the number of rotations it has performed. Here $a(t)$ is the total angle swept and $\theta_a(t)$, also represented by $Ja(t) \in SO(2)$, is the present orientation. The inverse of $J$ can be seen to be equal to

$$a(t) = \int_0^t [A'(s)dA(s)]_{12} \quad (4.4)$$

where $B_{ij}$ is the $ij$th element of the matrix $B$. Therefore, given a
Figure 4.1 Illustrating the Projection Map $J$
probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and any continuous \(\mathcal{F}\)-measurable random process \(Y: \Omega \to C^T_2\), there corresponds an \(\mathcal{F}\)-measurable random process \(y: \Omega \to C^T_1\) such that

\[
Y(t) = (Jy)(t) \quad t \in [0, T]
\] (4.5)

We now wish to specify our signal process pair \((x,X)\) and observation pair \((z,Z)\), where \(x\) and \(z\) are real valued, and \(X\) and \(Z\), are obtained from them via (4.5). We wish to allow \(x\) to have a nonzero (and perhaps random) initial condition. If we do not specify this initial condition, the map \(J\) is not invertible. That is, one can easily check that if we inject a continuous real-valued function \(a(t)\) into \(SO(2)\) via \(J\), we have

\[
(Ja)(t) = \begin{bmatrix}
\cos[a(t)-a(0)] & \sin[a(t)-a(0)] \\
-sin[a(t)-a(0)] & \cos[a(t)-a(0)]
\end{bmatrix} (Ja)(0)
\] (4.6)

or

\[
[(Ja)(t)][(Ja)(0)]^{-1} = (Ja)(t)
\] (4.7)

where

\[
\tilde{a}(t) = a(t) - a(0)
\] (4.8)

Since \(\tilde{a}(0) = 0\), \((J\tilde{a})(0) = I\), and the previous arguments show that we can recover \(\tilde{a}\) from \(J(\tilde{a})\). However, the only information \(J(a)\) yields about \(a(0)\) is contained in \((Ja)(0)\), which determines \(a(0) \mod 2\pi\). Thus, if \(b\) is a real valued function and \(n\) is any integer such that

\[
b(t) - a(t) = 2n\pi \quad \forall \ t
\] (4.9)
then (and only then)

\[(Ja)(t) = (Jb)(t) \quad \forall t\]  \hspace{1cm} (4.10)

Thus, if we allow nonzero initial conditions, we lose the bijectivity of \(J\). However, as long as we assume that we are always given the process \(a\), the process \(J(a)\) is uniquely defined. For our signal process pair \((x, x)\), we can usually assume that we start with \(x\) and obtain \(X\) via (4.5), or else, if we are given \(X\), we can make some assumption about \(x(0)\). We then can assume that we are given the process \(x\), with a possibly random initial condition of known distribution.

We can now define our signal and observation processes. Let \(x : \Omega \rightarrow \mathcal{C}_r^T\) be the continuous random process that is the solution of the Ito stochastic differential equation

\[dx(t) = a(x(t), t)dt + b^{1/2}(x(t), t)dv(t)\]  \hspace{1cm} (4.11)

where \(v : \Omega \rightarrow \mathcal{C}_r^T\) is the standard one-dimensional Brownian motion independent of the random initial condition \(x(0)\). We assume that \(a\) and \(b^{1/2}\) satisfy the necessary conditions to guarantee the existence of a solution to (4.11) (see, for example, Jazwinski [J2]). See [L6] for a more general class of signal processes. Note that by the Ito differential rule, \(X\) satisfies

\[dX(t) = \begin{bmatrix} -\frac{1}{2} b(x(t), t)dt & dx(t) \\ -dx(t) & -\frac{1}{2} b(x(t), t)dt \end{bmatrix} X(t)\]  \hspace{1cm} (4.12)

We remark that although \(x\) is always a Markov process, \(X\) need not be. In fact, we can show that \(X\) is a Markov process if and only if the
right-hand side of (4.11) is periodic in \( x \) with period \( 2\pi \) (i.e., if and only if it depends on \( x(t) \mod 2\pi \), which is equivalent to \( X(t) \)). For instance, \( X \) is a Markov process if \( x \) is a Brownian motion process.

Depending on the particular problem, we will regard either \( x \) or \( X \), given by (4.5), as the signal process. We remark that the probability distributions of \( x(t) \) and \( X(t) \) are related by "folding" the distribution for \( x(t) \) around the circle. Letting \( \theta_x(t) \) be the representation of \( X(t) \) as an element of \([-\pi, \pi] \) and, choosing any Borel subset \( A \) of \([-\pi, \pi] \), we have

\[
\Pr(\theta_x(t) \in A) = \Pr(x(t) \in \bigcup_{n=-\infty}^{+\infty} (A+2n\pi))
\]

\[
= \bigcup_{n=-\infty}^{+\infty} \Pr(x \in A+2n\pi) \tag{4.13}
\]

We define a random process \( z : \Omega \rightarrow C_T^1 \) satisfying the Ito differential equation

\[
dz(t) = h(x(t), t)dt + \frac{1}{2}r(t)dw(t) \quad z(0) = 0 \tag{4.14}
\]

where \( w : \Omega \rightarrow C_T^1 \) is a standard \( R^1 \) Brownian motion process independent of \( v \) and \( x(0) \), and \( h \) and \( \frac{1}{2}r \) satisfy the same type of conditions as \( a \) and \( b\frac{1}{2} \), respectively. Let \( Z : \Omega \rightarrow C_T^2 \) be defined by

\[
Z(t) = (Jz)(t) \tag{4.15}
\]

Applying the Ito differential rule, we obtain the following matrix Ito differential equation:

\[
dZ(t) = \\
\begin{bmatrix}
- \frac{r(t)}{2} dt & dz(t) \\
- dz(t) & - \frac{r(t)}{2} dt
\end{bmatrix} Z(t) \quad Z(0) = I \tag{4.16}
\]
where the diagonal terms, \(-r(t)dt/2\), are the second order correction terms described in Section 2.4. We note that \(Z(t)\) can be written in the form

\[
Z(t) = \begin{bmatrix}
\cos\left(\int_0^t h(x(s),s)ds + \int_0^t r^{1/2}(s)dw(s)\right) & \sin\left(\int_0^t h(x(s),s)ds + \int_0^t r^{1/2}(s)dw(s)\right) \\
-sin\left(\int_0^t h(x(s),s)ds + \int_0^t r^{1/2}(s)dw(s)\right) & \cos\left(\int_0^t h(x(s),s)ds + \int_0^t r^{1/2}(s)dw(s)\right)
\end{bmatrix}
\]

\[
= \begin{bmatrix}
cos\int_0^t r^{1/2}(s)dw(s) & \sin\int_0^t r^{1/2}(s)dw(s) \\
-sin\int_0^t r^{1/2}(s)dw(s) & \cos\int_0^t r^{1/2}(s)dw(s)
\end{bmatrix}
\begin{bmatrix}
cos\int_0^t h(x(s),s)ds & \sin\int_0^t h(x(s),s)ds \\
-sin\int_0^t h(x(s),s)ds & \cos\int_0^t h(x(s),s)ds
\end{bmatrix}
\]

(4.17)

We wish to point out the relationship between the first form of (4.17) and (2.60). This relationship will be discussed in Section 4.3 and in the simulation results in Chapter 6. In the second line of (4.17) we see the inherently multiplicative nature of the observation noise. This will be referred to in Chapter 6 in relation to lognormal noise in optical communications and will be extended in Chapter 7.

We assume that our sensor observes either \(Z(t)\) or \(dZ(t)\). In this case, equation (4.16) is the Ito differential equation that represents the sensor dynamics. This model deserves some further comment. We first note that we shall find that the differential equations for estimation -- either the Kushner type partial differential equations (4.39) - (4.42), or the optimal estimation equations in Section 4.3 -- explicitly involve \(dZ\). Thus, if we observe \(Z\), we will have to differentiate it to obtain \(dZ\).
Also, in (4.14) and (4.16) we have assumed that \( z(0) = 0 \) and \( Z(0) = I \).

However, if \( dZ \) is taken as the observation, the initial condition is unimportant. The reader is referred to Chapter 6, in which we discuss a scheme that avoids differentiating \( Z \) but retains the property of being independent of the initial condition. In this case we can drop the assumption that \( z(0) = 0 \) and \( Z(0) = I \).

We note that the input to the sensor is not \( X(t) \) but \( x(t) \), and that, from (4.14), it is \( dz(t) \), not \( Z(t) \) or \( z(t) \), that is an explicit function of \( x(t) \), and if we interpret \( X \) as the orientation of some object, then \( x \) is the total angle swept. Thus, taking \( x \) as the input to the sensor reflects the bijectivity of \( J \) -- i.e., in observing a rotational process \( X(t) \), our observation yields information concerning the total rotation \( x(t) \). Referring to (4.14), if we assume that \( h(x,s) = x \) and \( r = 1 \), we have

\[
z(t) = \int_0^t x(s)ds + w(t)
\]

(4.18)

If, as with an integrating gyro or some other angle-detecting sensor, \( z \) and \( Z \) are measures of angular orientation, then \( x \) should be an angular velocity, as opposed to an angle. Thus, it would be physically appealing to consider a problem of the following type: let \( x_1 = \) total angle swept and \( x_2 = \) angular velocity; in this case, the dynamical rotational state is \( X = J(x_1) \). If we take noisy measurements of the total angle swept, our measurement process can be written as

\[
z(t) = x_1(t) + w(t) = \int_0^t x_2(s)ds + w(t)
\]

(4.19)

As we shall see, the solution of the scalar estimation problem with
state and sensor dynamics given by (4.11), (4.12), (4.14), and (4.16) will lead directly to the solution of multidimensional problems, including the one just described. Also, the scalar results are directly applicable to some practical problems such as frequency demodulation (see Chapter 6). For the multidimensional results, the reader is referred to the part of Section 4.3 on the general abelian Lie group problem and to the examples in that section (in particular, see Example 4.1).

Finally, we note that an observation equivalent to \( Z(t) \) is \( y(t) = (\cos z(t), \sin z(t)) \) -- i.e. we do not need the full 2x2 matrix. In fact, we will use this measurement pair formulation in the examples of the next section. However, in the mathematical development of this chapter, we will continue to write the SO(2) equations in order to suggest the generalizations previously mentioned.

The problem considered in this section is to determine the conditional probability measures \( P(X(\lambda) \in A|Z(\tau), \tau \in [0,t]) \) and \( P(X(\lambda) \in B|Z(\tau), \tau \in [0,t]) \), where \( A \) is a Borel subset of \( \mathbb{R}^1 \) and \( B \) a Borel subset of SO(2). We remark that the physical motivation for determining the distribution for \( X(\lambda) \) comes from such problems as the frequency demodulation problem [S7], [V1], [V2], [M6]. The physical motivation for finding the distribution for \( X(\lambda) \) is related to the problems of orientation estimation and phase synchronization and tracking (see Example 4.1).

In the rest of this section we will use \( z^t \) to denote \( \{z(\tau), \tau \in [0,t]\} \) and \( Z^t \) for \( \{Z(\tau), \tau \in [0,t]\} \). In addition, we will
sometimes write \( x(t) \) as \( x_t \) and \( X(t) \) as \( X_t \). We now note that we assume that all of the random processes are defined on a given underlying probability space \((\Omega, \mathcal{F}, P)\) (e.g. \( x : \Omega \to C^t_1, \ X : \Omega \to C^t_2 \)). Let \( \mathcal{F}_1 \) be the \( \sigma \)-algebra of Borel subsets of \( R^1 \) and \( \mathcal{F}_2 \) the Borel subsets of \( SO(2) \). The various processes induce probability measures on either \( (R^1, \mathcal{F}_1) \) or \( (SO(2), \mathcal{F}_2) \). For instance

\[
\nu_{x_t}(A) \triangleq P(x^{-1}_t(A)) \quad A \in \mathcal{F}_1 \tag{4.20}
\]

\[
\nu_{X_t}(B) \triangleq P(X^{-1}_t(B)) \quad B \in \mathcal{F}_2 \tag{4.21}
\]

where

\[
x^{-1}_t(A) = \{ \omega \in \Omega | x_t(\omega) \in A \} \tag{4.22}
\]

\[
x^{-1}_t(B) = \{ \omega \in \Omega | X_t(\omega) \in B \} \tag{4.23}
\]

Since \( J \) is bijective, the information contained in \( z^t \) is the same as that in \( z^\lambda \). Mathematically, this means that the \( \sigma \)-subalgebra of \( \mathcal{A} \) generated by \( z^t \) --- i.e. the \( \sigma \)-algebra generated by \( z^{-1}_s(A), A \in \mathcal{F}_1, s \in [0, T] \) --- is the same as that generated by \( z^\lambda \), and it will be denoted by \( \mathcal{A}_z^t \). The \( \sigma \)-subalgebra of \( \mathcal{A} \) generated by \( x_\lambda \) will be denoted by \( \mathcal{A}_x^\lambda \), and \( \mathcal{A}_x^\lambda \) denotes that generated by \( X_\lambda \).

For the present, we assume that the times \( \lambda \) and \( t \) are fixed, and for simplicity we drop them as subscripts and superscripts (we simply note that \( \lambda < t \) corresponds to a smoothing problem, \( \lambda = t \) is the filtering problem, and \( \lambda > t \) is the prediction problem). Let \( P_{xz} \) be the conditional probability measure on \((\Omega, \mathcal{A}_x)\) given \( \mathcal{A}_z \).
\[ P_{xz}(A, \omega) = P(A \mid \mathcal{A}_x)(\omega) \quad A \in \mathcal{A}_x, \ \omega \in \Omega \] (4.24)

and we define \( P_{xz} \) analogously. These measures induce the following measures on \((R^1, \mathcal{P}_1)\) and \((SO(2), \mathcal{P}_2)\):

\[ \nu_{xz}(A, \omega) = P_{xz}(x^{-1}(A), \omega) \quad A \in \mathcal{P}_1 \] (4.25)

\[ \nu_{xz}(B, \omega) = P_{xz}(x^{-1}(B), \omega) \quad B \in \mathcal{P}_2 \] (4.26)

We also note that since these measures are \( \mathcal{A}_z \)-measurable, they are functions of \( Z^t(\omega) \) -- e.g. we can write

\[ \nu_{xz}(A, \omega) = \nu^*(A, Z^t(\omega)) \] (4.27)

From now on we omit the * and the measures \( \nu_{xz} \) and \( \nu_{xz} \) should be thought of as explicit functions of \( Z^t \).

The problem is to determine equations for \( \nu_{xz}(A, Z^t) \) and \( \nu_{xz}(B, Z^t) \). Because of the equivalence of \( z^t \) and \( Z^t \), the equations can be obtained from standard vector space results (see details in [L6], [J2], [L8], [K4], [B7], [B8], and [M7]). The measure \( \nu_{xz} \) can be shown, [K4], [L8], [L6], to be absolutely continuous with respect to \( \nu_x \) (eqn. (4.20)), and the Radon-Nikodym derivative, [H5], [R3], is given by

\[ \frac{d\nu_{xz}(x, Z^t)}{d\nu_x} = \frac{\delta(0^t \mid x_A = x)}{\delta(0^t)} \] (4.28)

where

\[ \theta^t = \exp \left( -\frac{1}{2} \int_0^t \frac{h^2(x(\tau), \tau)}{r(\tau)} \, d\tau \right) + \int_0^t \frac{h(x(\tau), \tau)}{r(\tau)} \left[ Z'(\tau) dZ_1(\tau) \right] d\tau \] (4.29)
where \( \int \) denotes an Ito stochastic integral [J2]. Here \( dz(\tau) = [Z'(\tau)dZ(\tau)]_{12} \) and the expectation is over \( x^T \) with \( z^t \) fixed (in the numerator of (4.28) we also keep \( x_\lambda \) fixed).

Suppose that \( \nu_x < \mu \), where \( \mu \) is Lebesgue measure on \( \mathbb{R}^1 \) — i.e., \( d\nu_x / d\mu = p(x, \lambda) \) exists and satisfies

\[
\nu_x(A) = \int_A p(x, \lambda) dx
\]

(4.30)

where we write \( dx \) for \( d\mu(x) \). In this case, it is clear that \( \nu_{xz} \) is also absolutely continuous with respect to \( \mu \) and, denoting this density by \( p_x(x, \lambda | z^t) \), we have

\[
p_x(x, \lambda | z^t) = \frac{\mathcal{L}(\Theta^t | x_\lambda = x)}{\mathcal{L}(\Theta^t)} p_x(x, \lambda)
\]

(4.31)

We note that since \( x \) satisfies an Ito stochastic equation (4.11), if \( p_x(x, t) \) exists, is once continuously differentiable in \( t \) and four times in \( x \), and if \( b(x, t) \) is twice continuously differentiable in \( x \) and \( a(x, t) \) is once in \( x \), then \( p_x \) must satisfy the forward equation

\[
\frac{\partial p}{\partial t} = \Lambda^a p_x
\]

(4.32)

where

\[
\Lambda^a(g) = -\frac{\partial(a^g)}{\partial x} + \frac{1}{2} \frac{\partial^2(b^g)}{\partial x^2}
\]

(4.33)

See [J2], [G4], [M3], and [W5]. In particular, McKean [M3, p. 61] discusses additional conditions on \( a \) and \( b \) that guarantee the existence of \( p_x \) that satisfies (4.32). Basically, he requires \( a \) and \( b \) to be infinitely differentiable in \( x \). Also, Wong [W5, p. 173] uses the backward equation to yield the result that if \( a, b \) satisfy the usual
conditions to guarantee a solution to (4.11) (see [W5, eqn (7.25)], [J2, p. 105]) and also
\[ b(x,t) \geq b_0 > 0 \quad \forall x \in \mathbb{R}^1, \ t \in [0,T] \]
(4.34)
then the density \( p_x(x,t) \) exists (although it may not satisfy (4.32)).
We finally remark that all of these qualifications will be satisfied in
the following section, in which the coefficients are as nice as one
could wish.

We now must consider the measure \( \nu_{xz} \). We consider instead the
measure \( \nu_{\theta z} \), where \( \theta \) is the representation of \( X \) on \([-\pi, \pi)\). Then, from
(4.13)
\[ \nu_{\theta z}(A) = \sum_{n=-\infty}^{+\infty} \nu_{xz}(A+2n\pi) \]
(4.35)
and also for the unconditional measure
\[ \nu_{\theta}(A) = \sum_{n=-\infty}^{+\infty} \nu_{x}(A+2n\pi) \]
(4.36)
for any Borel subset of \([-\pi, \pi)\). Then, if (4.31) holds, we have the
following result

**Theorem 4.1:** Consider a signal process pair \((x,X)\) defined by (4.5),
(4.11), and (4.12) and a measurement process defined by (4.14) and
(4.15). If the unconditional density for \( x, p_x(x,\lambda) \), exists, then the
conditional densities \( p_x(x,\lambda|z^t), p_{\theta}(\theta,\lambda|z^t) \) for \( x, \lambda \), respectively,
exist and are given by
\[ p_x(x,\lambda|z^t) = \frac{\mathcal{E}(\theta^t|x,\lambda=x)}{\mathcal{E}(\theta^t)} p_x(x,\lambda) \]
(4.37)
where \( \Theta^t \) is defined by (4.29).

We can now consider various values of \( \lambda \). The one with which we shall mostly concern ourselves is \( \lambda = t \) -- the filtering problem. In this case, following Kushner [K5], Bucy [B8], or Jazwinski [J2], we can derive the stochastic partial differential equations for \( p_x(\Theta, \lambda | Z^t) \) and \( p_\Theta(\Theta, \lambda | Z^t) \), assuming the densities are sufficiently smooth (see also Stratonovich [S8]):

\[
dp_x(x,t | Z^t) = (A^* p_x)(x,t | Z^t) dt + \frac{(h(x,t) - \hat{h}_t)}{r(t)} \left( [Z'(t) dZ(t)]_{L2} - \hat{h}_t dt \right) p_x(x,t | Z^t)
\]

(4.39)

where

\[
\hat{h}_t = \mathcal{E}(h(x,t) | Z^t) = \int_{-\infty}^{+\infty} h(x,t) p_x(x,t | Z^t) dx
\]

(4.40)

\[
\dp_\Theta(\Theta,t | Z^t) = \sum_{n=-\infty}^{+\infty} \dp_x(\Theta+2\pi n, t | Z^t)
\]

(4.41)

In general, we cannot simplify (4.41) any further -- i.e. \( \dp_\Theta \) depends upon \( p_x \) and not just \( p_\Theta \). However, suppose \( a, b, \) and \( h \) are periodic in \( x \) with period \( 2\pi \). Then, we have the equation

\[
dp_\Theta(\Theta,t | Z^t) = (B^* p_\Theta)(\Theta,t | Z^t) dt + \frac{(h(\Theta,t) - \hat{h}_t)}{r(t)} \left( [Z'(t) dZ(t)]_{L2} - \hat{h}_t dt \right) p_\Theta(\Theta,t | Z^t)
\]

(4.42)

where

\[
(B^* g) = - \frac{\partial (ag)}{\partial \Theta} + \frac{1}{2} \frac{\partial^2 (bg)}{\partial \Theta^2}
\]

(4.43)

We have used the relation...
\[
\frac{\partial}{\partial \theta} p_{\theta}(\theta, t | Z^t) = \frac{\partial}{\partial \theta} \sum_{n=-\infty}^{+\infty} p_x(\theta + 2n\pi | Z^t)
= \sum_{n=-\infty}^{+\infty} \frac{\partial}{\partial \theta} p_x(\theta + 2n\pi | Z^t)
= \sum_{n=-\infty}^{+\infty} \frac{\partial}{\partial \theta} p_x \bigg|_{x=\theta + 2n\pi}
\]

(4.44)

and an analogous expression for \( \frac{\partial^2}{\partial \theta^2} p_{\theta} \). These equations need not hold in general (since \( p_{\theta} \) might be infinite at a given point in \([-\pi, \pi])\), however, under certain restrictions on \( p_x \), (4.44) holds.

See Appendix B (Section B.2) for some further comment on this.

Also, we note that we have encountered the periodicity conditions on \( a \) and \( b \) earlier in this chapter. That is, we remarked that \( X \) is a Markov process if and only if \( a \) and \( b \) are periodic in \( x \) with period \( 2\pi \). In addition, the requirement that \( h \) satisfy the same periodicity condition is the same as requiring that \( h \) be a function of \( X(t) \) as opposed to \( x(t) \).

For \( \lambda > t \), we have the prediction problem. Let \( p_x(x, \lambda | y, t) \) be the transition density for \( X_\lambda \) given \( X_t = y \). Then one has (see [J2], [W5])

\[
p_x(x, \lambda | Z^t) = \int_{-\infty}^{+\infty} p_x(x, \lambda | y, t)p_x(y, t | Z^t) dy
\]

(4.45)

\[
p_{\theta}(\theta, \lambda | Z^t) = \int_{-\infty}^{+\infty} \left[ \sum_{n=-\infty}^{+\infty} p_x(\theta + 2n\pi, \lambda | y, t) \right] p_x(y, t | Z^t) dy
\]

(4.46)

where we have used the monotone convergence theorem [H5], [R3] to derive (4.46). Thus the prediction result can be expressed explicitly in terms of the filtering solution.
For the smoothing problem, \( \lambda < t \), we can also express the densities in terms of filtering (see [L6], [L9], and [F2]).

\[
 p_x(x, \lambda | Z^t) = p_x(x, \lambda | Z^\lambda) \exp\left\{ - \frac{1}{2} \int_{\lambda}^{t} \frac{\alpha_s^2}{r(s)} \, ds - \frac{1}{2} \int_{\lambda}^{t} \alpha_s^2 \, ds \right\} \tag{4.47}
\]

where

\[
 \alpha_s = \mathcal{E}(h(x(s), s) | Z^s, x_\lambda = x) - \mathcal{E}(h(x(s), s) | Z^s) \tag{4.48}
\]

\[
 dy(s) = [Z'(s) dZ(s)]_{12} - \mathcal{E}(h(x(s), s) | Z^s) \tag{4.49}
\]

We can write the equation \( p_\theta(\theta, \lambda | Z^t) \) by using the identity

\[
 p_\theta(\theta, \lambda | Z^t) = \int_{n=-\infty}^{\infty} p_x(\theta + 2\pi n, \lambda | Z^t) \tag{4.50}
\]

We see that the filtering density is of great importance, and we shall concentrate on the filtering problem in the next section, in which we specialize these results to a very important special case.

4.3 A Continuous Time Bilinear Estimation Problem

In this section, we will consider a subclass of the class of systems described in the previous section and will use the tools of linear estimation theory to solve an optimal estimation problem. Specifically, we consider the signal process pair \((x, X)\) given by (4.5), (4.11), and (4.12) with the restriction that (4.11) be a linear Ito equation

\[
 dx(t) = a(t)x(t)dt + b^{1/2}(t)dv(t) \tag{4.51}
\]

with initial condition \( x(0) \) being a random variable with density \( N(x; \xi_0, P_0) \). We assume that \( b(t) > 0 \ \forall \ t \in [0, T] \). In this case, we can write the bilinear Ito equation satisfied by \( X = Jx \) in two ways:
\[
dX(t) = \begin{bmatrix}
-\frac{1}{2} b(t) dt & dx(t) \\
-dx(t) & -\frac{1}{2} b(t) dt
\end{bmatrix} X(t)
\]

or

\[
dX(t) = -\frac{1}{2} b(t) X(t) dt + \{ a(t) \left[ \int_0^t e^{s} b^{1/2}(s) dv(s) \right] dt + b^{1/2}(t) dv(t) \} RX(t)
\]

where we note that \( R \) is given by (2.41), that

\[
x(t) = \int_0^t e^s b^{1/2}(s) dv(s) + x(0)
\]

and also that

\[
X(0) = \exp[Rx(0)]
\]

Equation (4.52) is of interest if we regard \( x(t) \) as an angular velocity (see Example 4.1), while (4.53) shows the inherent nonlinearity of the \( X \)-process (see Example 4.2).

The observation process to be used is

\[
dZ(t) = \begin{bmatrix}
-r(t)/2 & c(t)x(t) \\
-c(t)x(t) & -r(t)/2
\end{bmatrix} Z(t) dt + \begin{bmatrix}
0 & r^{1/2}(t) dw(t) \\
-r^{1/2}(t) dw(t) & 0
\end{bmatrix} Z(t)
\]

\[
Z(0) = I
\]

where \( r(t) > 0 \) \( \forall t \in [0,T] \). As shown in Section 4.2,
\[ Z(t) = \begin{bmatrix} \cos\left(\int_0^t c(s)x(s)\,ds + \int_0^t r^{1/2}(s)\,d\omega(s)\right) & \sin\left(\int_0^t c(s)x(s)\,ds + \int_0^t r^{1/2}(s)\,d\omega(s)\right) \\ -\sin\left(\int_0^t c(s)x(s)\,ds + \int_0^t r^{1/2}(s)\,d\omega(s)\right) & \cos\left(\int_0^t c(s)x(s)\,ds + \int_0^t r^{1/2}(s)\,d\omega(s)\right) \end{bmatrix} \]

and \( z = J^{-1}(Z) \) satisfies

\[ dz(t) = c(t)x(t)\,dt + r^{1/2}(t)\,d\omega(t) \quad z(0) = 0 \]  

(4.59)

Note that like (4.52), (4.56) is bilinear.

The problem now is to determine the equations for the optimal filtering estimate of \( x(t) \) and \( X(t) \) given \( Z^t \). (We will consider smoothing and prediction later). Our choice of criteria will be the following: let \( \phi_1 : S^1 \rightarrow R^1 \) be any error function satisfying (3.49) and (3.50), and let \( \phi_2 : R^1 \rightarrow R^1 \) satisfy analogous conditions -- i.e.

\[ 0 \leq \phi_2(x) = \phi_2(-x) \]

\[ |x_1| < |x_2| \Rightarrow \phi_2(x_1) < \phi_2(x_2) \]  

(4.60)

(4.61)

Choose random functions \( \hat{x}(t|t) \) and \( \hat{X}(t|t) \) to be respectively \( B_1^t \)- and \( B_2^t \)-measurable functions of \( Z^t \) such that

\[ \mathcal{E}(\phi_1(X(t), \hat{x}(t|t))|Z^t) \leq \mathcal{E}(\phi_1(X(t), M)|Z^t) \]  

(4.62)

\[ \mathcal{E}(\phi_2(x(t)-\hat{x}(t|t))|Z^t) \leq \mathcal{E}(\phi_2(x(t)-u)|Z^t) \]  

(4.63)

for all \( \mathcal{A}_Z^t \)-measurable random 2x2 orthogonal matrices \( M \) and all \( \mathcal{A}_Z^t \)-measurable real random variables \( u \), where
\[ \phi_1(X_1, X_2) = \phi_1(X_2, X_1) = \phi_1(X_1X_2^{-1}, I) = \phi_1(\theta) \]  

(4.64)

and

\[ X_1^{-1}X_2 = \exp R\theta \quad \theta \in [-\pi, \pi) \]  

(4.65)

(i.e. \( \theta \) is the \([-\pi, \pi)\) representation of \( X_1^{-1}X_2 \)).

We first solve a well-known problem. Since \( z^t \) and \( z^\tau \) generate the same \( \sigma \)-algebra \( \mathcal{A}_z \), \( \mathcal{E}(x(t) | \mathcal{A}_z^t) \) is both a \( \mathcal{B}_1 \)-measurable function \( f_1 \) of \( z^t \) and a \( \mathcal{B}_2 \)-measurable function \( f_2 \) of \( z^\tau \), and

\[ f_2(z^\tau) = f_1(J^{-1}(z^\tau)) \]  

(4.66)

In terms of \( z^t \), the solution to (4.63), for \( x(t) \) and \( z(t) \) given by (4.51) and (4.59) is well known [K6], [K7], [J2]. The conditional density \( p_x(x, t | z^t) \) is a normal density

\[ p_x(x, t | z^t) = \frac{1}{\sqrt{2\pi \text{P}(t)}} \exp \left[- \frac{(x-\hat{x}_t)^2}{2\text{P}(t)} \right] \]  

(4.67)

where Kalman-Bucy linear filtering theory yields the equations

\[ d\hat{x}_t = a(t)\hat{x}_t dt + \text{P}(t)c(t)r^{-1}(t)(dz(t) - c(t)\hat{x}_t|t) dt \]  

(4.68)

\[ \hat{x}_0|0 = \hat{x}_0 \]  

(4.69)

\[ \text{P}(t) = 2a(t)\text{P}(t)-c^2(t)r^{-1}(t)\text{P}^2(t)+b(t) \]  

(4.70)

\[ \text{P}(0) = \text{P}_0 \]  

(4.71)

Then by the \( \text{R}^1 \) version of Theorem 3.1, since the normal density is unimodal and symmetric,
\[ \mathcal{E}(\phi_2(x(t) - \hat{x}_t | t) | z^t) \leq \mathcal{E}(\phi_2(x(t) - u | z^t) \tag{4.72} \]

for all \( \mathcal{A}_z \)-measurable \( u \). Thus, the optimal estimate in terms of \( z^t \) is

\[ \hat{x}_t | t = f_1(z^t) \tag{4.73} \]

so the optimal estimate as a function of \( Z^t \) is just

\[ \hat{x}(t | t) = f_1(J^{-1}(Z^t)) \tag{4.74} \]

and we have proven the following:

**Theorem 4.2**: Let \( x \), the signal process, be given by (4.51) and let (4.56) be the observation process. Then the optimal filtering equations are

\[ d\hat{x}(t | t) = \sigma(t)\hat{x}(t | t) dt + P(t)c(t)r^{-1}(t)(\mathcal{Z}'(t)dz(t))_{12} - c(t)\hat{x}(t | t) dt \tag{4.75} \]

\[ \hat{x}(0 | 0) = \hat{x}_0 \tag{4.76} \]

\[ \dot{P}(t) = 2\sigma(t)P(t) - c^2(t)r^{-1}(t)P^2(t) + b(t) \tag{4.77} \]

\[ p(0) = P_0 \tag{4.78} \]

We note that these equations will be of importance in some of the frequency demodulation problems discussed in Chapter 6. Also note that as random processes on \( (\Omega, \mathcal{A}, P, \hat{x}_t | t \) and \( \hat{x}(t | t) \) are identical. These are simply alternate representations (\( z \) or \( Z \)) of the same process.

We now turn to the problem of estimating \( X(t) \) using the criterion (4.62). An alternative representation for \( X(t) \) is

\[ \theta_x(t) = x(t) \mod 2\pi \tag{4.79} \]
and, using (4.67), we have that the conditional density \( p_\theta(\theta, t | Z^t) \)
for \( \theta (t) \) given \( Z^t \) is the folded normal density

\[
p_\theta(\theta, t | Z^t) = P(\theta; \hat{\varphi}(t | t), P(t))
\]

(4.80)

where \( \hat{\varphi}(t | t) \) and \( P(t) \) are computed via (4.75) - (4.78). Then, from

Theorems 3.1 and 3.2, we have that the estimate \( \hat{\theta}_x (t | t) \) that min-

mizes \( \delta (\phi_1 (\theta_x (t | t), \hat{\theta}_x (t | t) | Z^t) \) (over the class of all \( \mathcal{X}_z \)-measurable

functions) is

\[
\hat{\theta}_x (t | t) = \hat{\varphi}(t | t) \mod 2\pi
\]

(4.81)

and the estimate \( \hat{X}(t | t) \) that is optimal with respect to (4.62) is

\[
\hat{X}(t | t) = \exp[\hat{\varphi}(t | t) R]
\]

(4.82)

We will call both \( \hat{X}(t | t) \) and \( \hat{\varphi}(t | t) \mod 2\pi \) the mode of the folded normal
density, depending upon which representation is more convenient. Note

that

\[
\delta (\exp[\varphi(t | t) R] | Z^t) = e^{-P(t)/2} \exp[\hat{\varphi}(t | t) R]
\]

\[
= e^{-P(t)/2} \exp[ \delta (\varphi(t | t) | Z^t) R]
\]

(4.83)

so the optimal estimate in the sense of (4.62) is not the conditional

expectation unless \( P^2(t) = 0 \) -- i.e. unless we know exactly where

we are on the circle and the line.

We can write a stochastic differential equation for \( \hat{\varphi}(t | t) \), and

we include this in the following theorem which has just been proven.

**Theorem 4.3:** Let \( X \), the signal process, be given by (4.53) and

(4.55), and let (4.56) be the measurement process. Then, with respect
to (4.62), the optimal filtering equations are
\[ d\hat{x}(t|t) = -\frac{1}{2} p^2(t)c^2(t)r^{-1}(t)\hat{x}(t|t)dt + \]
\[ +\int_0^t (a(\tau)-P(t)c^2(\tau)r^{-1}(\tau))d\tau \]
\[ +P(t)c(t)r^{-1}(t)[Z'(t)dZ(t)]_{12} \hat{x}(t|t) \]
\[ \hat{x}(0|0) = \exp(R\hat{x}_0) \]

(4.84)

(4.85)

where \( P(t) \) satisfies (4.77) and (4.78).

The optimal cost \( \mathcal{E}(\phi_1(X(t),\hat{x}(t|t))|Z^t) \) for various functions \( \phi_1 \) can be computed using the Fourier series results of Chapter 3.

For example, if \( \phi_1(\theta) = 1-\cos\theta \), then

\[ \mathcal{E}(\phi_1(X(t),\hat{x}(t|t))|Z^t) = 1 - e^{-P(t)/2} \]

(4.86)

The filter equation (4.84) is quite complex; however, using its relationship with \( \hat{x}(t|t) \) (eqn. (4.82)), we have the easily implemented optimal filter depicted by the block diagram of Figure 4.2. The measurement process \( dZ \) is processed by a nonlinear transformer that yields

\[ dz = [Z'dZ]_{12} \]

(4.87)

as its output. This process then goes through a Kalman-Bucy filter that computes \( \hat{x}(t|t) \), which is then injected into \( S^1 \) via the map \( J \) to produce the desired estimate \( \hat{x}(t|t) \).

The form of the optimal estimator deserves some further comment.
Figure 4.2  Block Diagram for Optimal Filtering on $S'$
First, the concept of pre-processing and post-processing a signal with a linear filter in the middle has also been discussed by Oppenheim, Schafer, and Stockham [01]. Also, as mentioned in the previous section, the differential nature of the observation means that the results are unchanged if \( Z(0) \neq I \). This is of importance, for instance, in the frequency demodulation problem where we observe only \( \sin z(t) \) and must construct \( \cos z(t) \) from it. This will be discussed in detail in Chapter 6. In addition, in Chapter 6 we will discuss an alternative nonlinear pre-processor that avoids differentiating \( Z(t) \).

The smoothing and prediction problems can be solved in a manner analogous to the solution of the filtering problem -- i.e. we can sandwich the solution to the linear smoothing or prediction problem between the nonlinear pre- and post-processors of Figure 4.2. The optimal smoothing solution \( \hat{x}(\lambda | t) \) with \( \lambda < t \) is given by ([L6], [K8])

\[
\hat{x}(\lambda | t) = \hat{x}(\lambda | \lambda) + \alpha(\lambda, t)
\]  

(4.88)

where

\[
\alpha(\lambda, t) = P(\lambda) \int_{\lambda}^{t} e^{\lambda \tau} \int_{\lambda}^{\tau} \left( s(\tau) - P(\tau) c^2(\tau) r^{-1}(\tau) \right) d\tau
\]

and

\[
c(\lambda) r^{-1}(\lambda) \left[ \int_{\lambda}^{t} \left( [Z(s) dZ(s)]_{12} - c(s) \hat{x}(s | s) ds \right) \right]
\]

(4.89)

and, since the relevant conditional densities are again normal and folded normal, the optimal \( S^1 \) estimate is
\[ \hat{x}(\lambda|t) = \exp(\hat{x}(\lambda|t)R) \]

\[ = \exp(\hat{x}(\lambda|\lambda)R)\exp(\alpha(\lambda,t)R) \]

\[ = \hat{x}(\lambda|\lambda)\exp(\alpha(\lambda,t)R) \]

Thus the smoothing solution can be expressed in terms of the filtering solution. The prediction estimates \( \hat{x}(\lambda|t) \), \( \hat{x}(\lambda|t) \) for \( \lambda > t \) are given by

\[ \hat{x}(\lambda|t) = e^{\int_{t}^{\lambda} \alpha(\tau) d\tau} x(\lambda|\lambda) \]

\[ \hat{x}(\lambda|t) = \exp[\hat{x}(\lambda|t)R] \]

(4.91)

(4.92)

The results presented so far in this section can be extended to a larger class of problems -- those involving processes evolving on abelian Lie groups. It is known (see [W1]) that a given abelian Lie group \( G \) is isomorphic to the direct product of a number of copies of the circle and a number of copies of \( \mathbb{R} \)

\[ G = \mathbb{R}^n \times (\mathbb{S}^1)^m \]  

(4.93)

where \( (\mathbb{S}^1)^m \) is called a "torus" (see Figure 4.3 for a picture of the torus \( (\mathbb{S}^1)^2 \)). This type of space has been considered by Ku and Sheporsitis [K9] and by Bucy, Hecht, and Senne [B9]. We note that this setting allows us to consider the multidimensional version of the problem just considered, and in this setting we can study problems like the physically more appealing two dimensional problem mentioned in Section 4.2 (see Example 4.1).
Figure 4.3: A Picture of the Torus \((S^1)^2\)

It is clear that we can now construct multidimensional versions of the problems just considered. Analogous to (4.2), we define the bijective map \(J_{nm} : (C^T_1)^{n+m} \rightarrow (C^T_1)^n \times (C^T_2)^m\) by

\[
(J_{nm}(a_1, \ldots, a_{n+m}))(t) = [a_1(t), \ldots, a_n(t), (J(a_{n+1}))(t), \ldots, (J(a_{n+m}))(t)]
\]  

(4.94)

Then a continuous \(\mathcal{F}\)-measurable random process \(X : \Omega \rightarrow (C^T_1)^n \times (C^T_2)^m\) on \(G\) corresponds to a unique continuous \(\mathcal{F}\)-measurable random process \(x : \Omega \rightarrow (C^T_1)^{n+m}\) on \(R^{n+m}\) via the identification

\[
x(t) = (J_{nm} x)(t)
\]  

(4.95)

Let \(x = (x_1, \ldots, x_{n+m})'\) be a normally distributed \((n+m)\)-dimensional random variable with density \(N_{n+m}(x; \eta, P)\) where \(\eta = \mathcal{E}(x)\) and \(P = \mathcal{E}[(x-\eta)(x-\eta)']\). Then the density \(P_{nm}(y; \eta, P)\) for \(y\) which is defined by
\[ y = (x_1, \ldots, x_n, x_{n+1} \mod 2\pi, \ldots, x_{n+m} \mod 2\pi)' \]

(4.96)

is given by

\[
F_{nm}((y_1, \ldots, y_{n+m})', \eta, \rho) =
\sum_{k_1=\infty}^{+\infty} \cdots \sum_{k_m=\infty}^{+\infty} N((y_1, \ldots, y_n, y_{n+1} + 2k_1 \pi, \ldots, y_{n+m} + 2k_m \pi)' \eta, \rho)
\]

(4.97)

and is called the \((n, m)\) folded normal density. Note that \(n\) of the marginal densities (those for \(x_1\) through \(x_n\)) are normal and the other \(m\) are folded normal.

We can now consider the vector analog of the processes described by (4.5) and (4.11). We can model our observation process analogously. Let \(z\) be the \((\ell+k)\)-dimensional solution of the vector Ito equation

\[ dz(t) = h(x(t), t)dt + Q^{1/2}(t)dv(t) \]

(4.98)

\[ z(0) = 0 \]

(4.99)

and define our observation process \(Z \in (C_T^T)^{\ell} \times (C_T^T)^k\) by

\[ Z(t) = (J_{\ell k}z)(t) \]

(4.100)

We can write stochastic differential equations for \(X\) and \(Z\) analogous to (4.12) and (4.15) by means of the Ito differential rule. The calculation is straightforward and will not be displayed. By the bijectiveness of \(J_{nm}\), it is clear that the preceding analysis extends directly to the multidimensional case. Furthermore, if the Ito equations for \(x\) and \(z\) are linear, the block diagram of Figure 4.2 is conceptually
correct. The nonlinear preprocessor becomes \((J_{\mathbf{q}_k})^{-1}\), the postprocessor is \(J_{nm}\), and we use the vector version of the Kalman-Bucy filter.

Thus, we have shown that the domain of the Kalman-Bucy filter includes estimation on abelian Lie groups. Also, since \(\mathbb{R}^{n+m}\) can be identified with the Lie algebra of \(\mathbb{R}^n \times (S^1)^m\) we see that the filtering technique involves the processing of the observations so that the actual filtering is done in the Lie algebra — a vector space.

Note that this multidimensional technique allows us to treat colored measurement noise by the technique of differentiating the observations to obtain reduced filter equations \([J2]\), \([B11]\), \([B12]\). Also note that the isomorphism \((4.93)\) may be such that the actual noise processes are quite interesting. For example, the Lie group of positive real numbers under multiplication is isomorphic to \(\mathbb{R}^1\) under addition by the map \(a \mapsto e^a\). Thus multiplicative amplitude changes can be looked at as additive noises. We will have more to say about this type of problem in Chapter 6, where we will discuss lognormal processes \([H9]\), \([A1]\).

The analysis of this section is best illustrated by example.

**Example 4.1:** Consider a cylindrical shaft of unit radius being spun about its longitudinal axis by an electric motor. We assume that the total rotation of the shaft, \(x_1\), is related to the driving force \(u\) by the differential equation

\[
\ddot{x}_1 + \dot{x}_1 + x_1 = u
\]  
(4.101)
with both \( x_1(0) \) and \( \dot{x}_1(0) \) equal to zero. The last term on the left-hand side of (4.101) can be thought of as a torsional spring effect that helps to stabilize the servo loop that drives the shaft. The driving force consists of a known force and a disturbance. The known force adds neither difficulty to the analysis nor complexity to the solution. Thus, for simplicity, we assume that the known force is zero and that the disturbance is white Gaussian noise, \( \dot{v} \), with unit strength. Setting \( x_2 = \dot{x}_1 \) we obtain the vector differential equation

\[
\frac{dx(t)}{dt} = Ax(t) + Bv(t) \quad x(0) = 0
\]  

(4.102)

where

\[
x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

(4.103)

Suppose we wish to determine the orientation of the shaft — i.e., \( x_1(t) = (J x_1)(t) \). Then our estimation state space is \( \mathbb{R}^1 \times S^1 \) with

\[
(J_{1,1}(x_1,x_2))(t) = (x_2(t), x_1(t))
\]

(4.104)

The orientation is determined by

\[
[x_1(t)]_{11} = \cos x_1(t) = \cos \int_0^t x_2(\tau) d\tau
\]

(4.105)

\[
[x_2(t)]_{12} = \sin x_1(t) = \sin \int_0^t x_2(\tau) d\tau
\]

(4.106)

Suppose we measure these quantities but that noise corrupts our observations, so we actually observe
\[ z_1(t) = \cos(\int_0^t x_2(\tau)d\tau + w(t)) \]  
\[ z_2(t) = \sin(\int_0^t x_2(\tau)d\tau + w(t)) \]

where \( w \) is a Brownian motion of unit strength independent of \( v \). Now, \( z_1 \) and \( z_2 \) are the \((1,1)\) and \((1,2)\) components of

\[ Z(t) = \begin{bmatrix} z_1 & z_2 \\ -z_2 & z_1 \end{bmatrix} \]

and satisfy the equations

\[ dz_1(t) = -\frac{1}{2}z_2(t)dt - x_2(t)z_2(t)dt - z_2(t)dw(t) \quad z_1(0) = 1 \]
\[ dz_2(t) = -\frac{1}{2}z_1(t)dt + x_2(t)z_1(t)dt + z_1(t)dw(t) \quad z_2(0) = 0 \]

Using the results of this section, we have the following optimal filter equations

\[ d\hat{x}(t|t) = A\hat{x}(t|t)dt + P(t)c'[z_1(t)dz_2(t) - z_2(t)dz_1(t) - c\hat{x}(t|t)dt] \]
\[ \hat{x}(0|0) = 0 \]

where

\[ c = [0,1] \]

and \( P \) is the solution of the Riccati equation

\[ \dot{P}(t) = AP(t) + P(t)A' - P(t)c'cP(t) + BB' \]
\[ P(0) = 0 \]

The optimal estimate of the orientation is

\[ \hat{\alpha}_1(t|t) = \exp(\hat{\alpha}_1(t|t)R) \]
The steady state filter has the same form as the time-varying filter, but \( P(t) \) is replaced by the positive definite solution, \( P_\infty \), of the algebraic equation

\[
A P_\infty + P_\infty A' - P c' c P_\infty + BB' = 0
\]

(4.116)

The solution is

\[
P_\infty = \begin{bmatrix} \sqrt{2} -1 & 0 \\ 0 & \sqrt{2} -1 \end{bmatrix}
\]

(4.117)

If we formally divide \( dz_1 \) and \( dz_2 \) by \( dt \) and take \( \dot{z}_1 \) and \( \dot{z}_2 \) as our measurements we get the block diagram of Figure 4.4 for the signal, observations, nonlinear processors, and the linear filter. Note that regarding \( \dot{z}_1 \) and \( \dot{z}_2 \) as the observations means differentiating \( z_1 \) and \( z_2 \). See Chapter 6 for a discussion of this problem and how to get around it, and, in particular, see Example 6.1, in which we re-solve the present example using a pre-processor that avoids differentiating \( z_1 \) and \( z_2 \).

**Example 4.2:** In this example, we will take a set of nonlinear signal and observation equations and show that by interpreting them as signals on \( S^1 \times R^2 \) and \( S^1 \times R^1 \), respectively, we can use the results of this section to solve the optimal filtering problem. The signal process is specified by the following equations:

\[
dx_1 = -\frac{1}{2} x_1 dt - x_2 x_4 dt - x_2 dv
\]

(4.118)

\[
dx_2 = -\frac{1}{2} x_2 dt + x_1 x_4 dt + x_1 dv
\]

(4.119)
\[ dx_3 = x_4 dt + dv \]
\[ \dot{x}_3 = -x_3 \]  \hspace{1cm} (4.120)
\[ \dot{x}_4 = x_4 + x_5 \]  \hspace{1cm} (4.121)
\[ x_1(0) = 1 \hspace{1cm} x_2(0) = x_3(0) = x_4(0) = x_5(0) = 0 \]  \hspace{1cm} (4.122)

where \( v \) is a unit strength Brownian motion process. The measurement equations are

\[ dz_1 = - \frac{1}{2} w_1 dt - (2x_5 + x_4 - \int_0^t x_1(s)dx_2(s) + \int_0^t x_2(s)dx_1(s))z_2(t) dt - z_2 dw_1 \]  \hspace{1cm} (4.124)
\[ dz_2 = - \frac{1}{2} w_2 dt + (2x_5 + x_4 - \int_0^t x_1(s)dx_2(s) + \int_0^t x_2(s)dx_1(s))z_1 dt + z_1 dw_1 \]  \hspace{1cm} (4.125)
\[ dz_3 = (\int_0^t x_1(s)dx_2(s) - \int_0^t x_2(s)dx_1(s))dt + dw_2 \]  \hspace{1cm} (4.126)

where \( w_1 \) and \( w_2 \) are standard Brownian motions, independent of each other and of \( v \). The problem is to find the least squares estimates, \( \hat{x}_1 \) and \( \hat{x}_2 \) for \( x_1 \) and \( x_2 \) under the constraint \( \hat{x}_1^2 + \hat{x}_2^2 = 1 \).

Rearranging (4.118) and (4.119), we have

\[ \begin{bmatrix} dx_1 \\ dx_2 \\ -dx_2 \\ dx_1 \end{bmatrix} = - \frac{1}{2} \begin{bmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{bmatrix} dt + \begin{bmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{bmatrix} R(x_4 dt + dv) \]  \hspace{1cm} (4.128)

and it is clear that

\[ x_1 = \cos x_3 \hspace{1cm} x_2 = \sin x_3 \]  \hspace{1cm} (4.129)

and therefore our optimal filter need only estimate \( x_3 \), \( x_4 \), and \( x_5 \).
to get estimates of $x_1$ and $x_2$. Also, (4.124), (4.125), and (4.129) yield

$$\begin{bmatrix} \frac{dz_1}{dt} \\ \frac{dz_2}{dt} \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} z_1 & z_2 \\ -z_2 & z_1 \end{bmatrix} \, dt + \begin{bmatrix} z_1 & z_2 \\ -z_2 & z_1 \end{bmatrix} R \left[ (2x_5 + x_4 - x_3) dt + dw_1 \right] \quad (4.130)$$

Note that the system is not observable with just the $S^1$ observation pair $(z_1, z_2)$ or with just the $R^1$ observation $z_3$, but the system is observable with both observations (see (4.133)).

Following the approach developed in this section, we have the optimal filter equations

$$\begin{bmatrix} \frac{d\hat{x}_3}{dt} \\ \frac{d\hat{x}_4}{dt} \\ \frac{d\hat{x}_5}{dt} \end{bmatrix} = A \begin{bmatrix} \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{bmatrix} + \begin{bmatrix} \frac{z_1 dz_2 - z_2 dz_1}{dz_3} \end{bmatrix} - C \begin{bmatrix} \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{bmatrix} \quad (4.131)$$

$$\begin{bmatrix} \hat{x}_3(0), \hat{x}_4(0), \hat{x}_5(0) \end{bmatrix} = 0$$

where

$$\dot{P} = AP + PA' - PC'CP + BB' \quad P(0) = 0 \quad (4.132)$$

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad C = \begin{bmatrix} -1 & 1 & 2 \end{bmatrix} \quad (4.133)$$

Then, from Example 3.1, we have

$$\begin{aligned} \hat{x}_1 &= \cos \hat{x}_3 \\ \hat{x}_2 &= \sin \hat{x}_3 \end{aligned} \quad (4.134)$$

(Here $\hat{x}_1 \neq \mathcal{E}(x_1)$, $\hat{x}_2 \neq \mathcal{E}(x_2)$ - see (4.83)). The block diagram of the optimal estimator is given in Figure 4.5.

We note that equations (4.118) and (4.119) involve products of state variables. The fact that we can handle a product nonlinearity,
Figure 4.5 Block Diagram of the Optimal Filter for Example 4.2
at least in this special case, is significant. It is hoped that generalizations of these techniques will allow us to handle other examples of product nonlinearities.

4.4 A Discrete-Time Estimation Problem on $S^1$

We wish to examine the problem of estimating a discrete-time random process on $S^1$ given a series of discrete measurements. A natural model for the signal and measurement processes is a discrete approximation to the continuous signal and measurement processes discussed in the preceding section. We first approximate the measurement equations

$$dz(t) = h(x(t),t)dt + r^{1/2}(t)dw(t)$$  \hspace{1cm} (4.135)

$$Z(t) = (Jz)(t)$$  \hspace{1cm} (4.136)

by the discrete equations

$$\Delta z_k = z_k - z_{k-1} = m_k(x_k, \Delta t + r_k^{1/2} w_k)$$  \hspace{1cm} (4.137)

$$Z_k = \exp(Z_k R)$$  \hspace{1cm} (4.138)

where $\Delta t$ is the inter-measurement time, $x_k = x(k\Delta t)$, $q_k = q(k\Delta t)$, $m_k(\cdot) = m(\cdot, k\Delta t)$, and $\Delta w_k = w(k\Delta t) - w((k-1)\Delta t)$.

We can rewrite the $Z_k$ equation as

$$Z_k = Z_{k-1} \exp(\Delta z_k R)$$  \hspace{1cm} (4.139)

and we see that, given $Z_1, \ldots, Z_{k-1}$, the new information contained in $Z_k$ is equivalent to the new information in $Z_{k-1}^{-1} Z_k$. In addition, this information is equivalent to knowledge of
\[ \Delta z_k = \Delta z_k \mod 2\pi \quad \Delta z_k \in (-\pi, \pi) \] (4.140)

It is here that we see a marked difference between the discrete and continuous time problems. In the continuous time case, the continuity of the random processes involved results in our knowing \( dz(t) \), not just \( dz(t) \mod 2\pi \). However, in the discrete time problem, the ambiguity associated with our lack of knowledge of the number of rotations that occur in the \( \Delta t \) between measurements, is reflected in the fact that our information is \( \Delta z_k \mod 2\pi \) and not \( \Delta z_k \). In some sense, this makes the discrete-time problem more interesting, since this ambiguity makes the solution much more difficult than the analogous continuous time problem.

Motivated by this discussion, we can state precisely a discrete-time analog of the continuous-time problem of Section 4.2. Let \( x_k \) be a discrete-time real-valued signal process satisfying

\[ x_{k+1} = a(x_{k},k) + b^{1/2}(x_{k},k)w_k \quad x_0 = 0 \] (4.141)

where \( \{w_k\} \) is a unit variance white Gaussian sequence. Consider an observation process \( \{\tilde{y}_k\} \) defined by

\[ y_k = h(x_{k},k) + r_k^{1/2}v_k \] (4.142)

\[ \tilde{y}_k = y_k \mod 2\pi \quad \tilde{y}_k \in [-\pi, \pi) \] (4.143)

where \( \{v_k\} \) is a unit variance Gaussian sequence independent of the \( w_k \). Note that choosing \( \tilde{y}_k \in [-\pi, \pi) \) as opposed to \( \tilde{y}_k \in [a, a+2\pi) \), \( a \in \mathbb{R}^1 \), is a convention we have chosen but which does not affect the results. The problem is to determine the conditional probability distributions
and to determine optimal estimates, where optimality is defined as in (4.62) and (4.63). We note that one can interpret $\theta_k$ as angular orientation and $x_k$ as total angle swept.

Instead of this problem, we will consider a general one-stage problem. An intuitive argument, involving delta functions, that leads to an equation for the conditional probability density, is given here, and the reader is referred to Section B.3 for the rigorous but tedious details. We then will consider the linear case (a and h linear, b a function of k only), will comment on the multistage problem in this case, will discuss the optimal estimation problem, and will discuss approximations and suboptimal schemes. For the relationship between the continuous-time and discrete-time problems and the details of some suboptimal schemes, the reader is referred to Section B.4 and Appendix C.

Let $x$ and $v$ be independent real-valued random variables with a priori densities $p_x(\alpha)$ and $p_v(\nu)$. We wish to compute the conditional density $p_x|\tilde{y}(\alpha|\beta)$ where

$$y = h(x) + v$$

$$\tilde{y} = y \mod 2\pi$$

and $h : \mathbb{R}^1 \to \mathbb{R}^1$ is a measurable function. By a property of conditional densities [P2],
\[ p_x|\tilde{y}(\alpha|\beta) = \int_{-\infty}^{+\infty} p_x|\tilde{y},y(\alpha|\beta,\xi)p_y|\tilde{y}(\xi|\beta) d\xi \]  \hspace{1cm} (4.148)

Now \( \tilde{y} \) is a deterministic function of \( y \), so the "density" \( p_y|\tilde{y} \) is given by

\[ p_y|\tilde{y}(\beta|\xi) = \frac{p_y(\beta|\xi)p_y(\xi)}{p_y(\beta)} \delta(\beta-\xi \text{ mod } 2\pi)p_y(\xi) \]  \hspace{1cm} (4.149)

where \( \delta \) is the Dirac delta function. Also, since \( y \) determines \( \tilde{y} \) uniquely

\[ p_x|y,\tilde{y} = p_x|y \]  \hspace{1cm} (4.150)

Thus (4.148) becomes

\[ p_x|\tilde{y}(\alpha|\beta) = \frac{1}{p_y(\beta)} \int_{-\infty}^{+\infty} p_x|y(\alpha|\xi)p_y(\xi)\delta(\beta-\xi \text{ mod } 2\pi) d\xi \]  \hspace{1cm} (4.151)

We now must interpret \( \delta(\beta-\xi \text{ mod } 2\pi) \) as a function of \( \xi \). Clearly what we mean by this is

\[ \delta(\beta-\xi \text{ mod } 2\pi) = \sum_{n=-\infty}^{+\infty} \delta(\beta+2n\pi-\xi) \]  \hspace{1cm} (4.152)

Substituting (4.152) into (4.151) we have

\[ p_x|\tilde{y}(\alpha|\beta) = \sum_{n=-\infty}^{+\infty} \frac{p_y(\beta+2n\pi)}{p_y(\beta)} p_x|y(\alpha|\beta+2n\pi) \]  \hspace{1cm} (4.153)

\[ = \sum_{n=-\infty}^{+\infty} \frac{p_y|\tilde{y}(\beta+2n\pi|\alpha)p_x(\alpha)}{p_y(\beta)} \]  \hspace{1cm} (4.154)

where (4.154) comes from applying Bayes' rule to \( p_x|y \). Also

\[ p_y|\tilde{y}(\beta+2n\pi|\alpha) = p_y(\beta+2n\pi-h(\alpha)) \]  \hspace{1cm} (4.155)

\[ p_y(\beta+2n\pi) = \int_{-\infty}^{+\infty} p_y|\tilde{y}(\beta+2n\pi|u)p_x(u) du \]  \hspace{1cm} (4.156)

\[ p_y(\beta) = \sum_{n=-\infty}^{+\infty} p_y(\beta+2n\pi) \]  \hspace{1cm} (4.157)
These results have been derived very informally. The rigorous derivation is contained in the proofs of Theorems B.2 and B.3. See Corollary 1 of Theorem B.3 for the derivation of (4.153) and (4.154).

For the rest of this section we will limit ourselves to a discussion of the linear-Gaussian case — i.e. where

$$a(x,k) = a_k x$$  \hspace{1cm} (4.158)

$$h(x,k) = h_k x$$  \hspace{1cm} (4.159)

$$b^{1/2}(x,k) = b_k^{1/2}$$  \hspace{1cm} (4.160)

and all a priori densities are normal. If $p_x(x) = N(\alpha;\eta, \gamma_1)$, $p_v(v) = N(v;0,\gamma_2)$, and

$$y = hx + v$$  \hspace{1cm} (4.161)

then plugging into (4.153)-(4.157) and using known results about linear measurements (or referring to Corollary 2 of Theorem B.3) we have

$$p_x|y(\alpha|\beta) = \sum_{n=-\infty}^{+\infty} c_n(\beta)N(\alpha;\eta_n, \gamma_3)$$  \hspace{1cm} (4.162)

where

$$N(\alpha;\eta_n, \gamma_3) = p_x|y(\alpha|\beta + 2n\pi)$$  \hspace{1cm} (4.163)

$$\eta_n = \frac{\gamma_2 \eta + \gamma_1 h(\beta + 2n\pi)}{h^2 \gamma_1 + \gamma_2}$$  \hspace{1cm} (4.164)

$$\gamma_3 = \frac{\gamma_1 \gamma_2}{h^2 \gamma_1 + \gamma_2}$$  \hspace{1cm} (4.165)
Thus the nth term is evaluated by an optimal linear estimator which takes as its measurement $\beta + 2n\pi$.

We now consider the multistage linear problem. It will suffice to consider only two stages, since this will indicate the type of recursions necessary in building the discrete filter. Thus, consider (4.141)-(4.145) with $a_k^1$, $h_k^1$, and $b_k^1/2$ given by (4.158)-(4.160) and with $p_{\nu_1}(\nu) = N(\nu; 0, r_1)$ and $p_{\nu_2} = N(\nu; 0, r_2)$. Also, the a priori density for $x_1$ is taken to be

$$p_{x_1}(\alpha) = N(\alpha; 0, b_1)$$  \hspace{1cm} (4.167)

The preceding analysis shows that the density $p_{x_1|\tilde{y}_1}(\alpha|\beta_1)$ is given by (4.162)-(4.166) with $n = 0$, $\gamma_1 = b_1$, $\gamma_2 = r_1$, $\beta = \beta_1$, and $h = h_1$. Then, it is easy to see that the density $p_{x_2|\tilde{y}_1}(\alpha|\beta_1)$ is given by

$$p_{x_2|\tilde{y}_1}(\alpha|\beta_1) = \sum_{n=-\infty}^{\infty} c_n(\beta)N(\alpha; a_n^1, a_n^2\gamma_3+b_2)$$  \hspace{1cm} (4.168)

where the $c_n$, $\eta_n$, and $\gamma_3$ are computed from (4.164)-(4.166) with the above substitutions. This result comes from the fact that the density for $z_1+z_2$ is the convolution of the densities for $z_1$ and $z_2$ and that convolution is linear (i.e. $(a+b)*c = a*c + b*c$).

It remains to include the effect of the additional measurement $\tilde{y}_2$. To do this, we refer to (4.153)-(4.157) and assume that $p_{\nu}(\nu) = N(\nu; 0, \gamma_2)$ and $h(x) = hx$. In this case, the a priori density for $x_2$ --
the density for \( x_2 \) just before we process \( \tilde{y}_2 \) — is \( p_{x_2 | \tilde{y}_1} \), which is given by (4.168). This density is of the form

\[
p_x(\alpha) = \sum_{k=-\infty}^{+\infty} d_k N(\alpha; \eta_k, \gamma_1)
\]

where

\[
\sum_{k=-\infty}^{+\infty} d_k = 1 \quad d_k > 0
\]

In this case (4.153) becomes

\[
p_{x|\tilde{y}}(\alpha|\beta) = \sum_{n=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} e_n k N(\alpha; \eta_{nk}, \gamma_3)
\]

where \( \gamma_3 \) is given by (4.165) and

\[
\eta_{nk} = \frac{\gamma_2 \eta_k + \gamma_1 h(\beta + 2n\pi)}{h^2 \gamma_1 + \gamma_2}
\]

\[
e_n k = \frac{d_k N(\beta + 2n\pi; \eta_k; h^2 \gamma_1 + \gamma_2)}{\sum_{k=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} d_m N(\beta + 2l\pi; \eta_m; h^2 \gamma_1 + \gamma_2)}
\]

These equations clearly apply to the problem of computing \( p_{x_2 | \tilde{y}_1, \tilde{y}_2} \) from \( p_{x_2 | \tilde{y}_1} \) if we identify \( p_x(\alpha) \) in (4.169) with \( p_{x_2 | \tilde{y}_1}(\alpha|\beta_1) \) in (4.168), and from this computation we can see a pattern. After \( m \) measurements, the density \( p_x(x, m | \tilde{y}_m) \) is an \( m \)-times infinite sum of normal densities, all with the same variance — the one given by linear theory — and the mean of the \( (i_1, i_2, \ldots, i_m) \)th term is the linear result if

\[
\tilde{y}_1 = \tilde{y}_1 + 2i_1\pi, \ldots, \tilde{y}_m = \tilde{y}_m + 2i_m\pi
\]
The coefficient of this term is the probability that (4.174) holds. These terms are nonlinear functions of the measurements.

It is clear that any practical implementation of these results requires some approximate scheme. We will make some general comments here about this problem and will comment on it further in Appendix C. First note that the coefficients of the terms in the infinite series are probabilities for the various $y_i$. Thus, it seems likely that workable scheme would keep the $N$ terms corresponding to the $N$ largest coefficients. This type of approach and several others were first considered by Buxbaum and Haddad [810] and are discussed in Appendix C. Figure 4.6 illustrates the basic concept behind such a truncation procedure.

We remark that Section B.4 contains a result relating the discrete and continuous time problems. This result states that if our discrete-time problem is a discretization of the continuous one, then, as the time $\Delta t$ between measurements becomes small, the term in the conditional density corresponding to $n$ rotations between measurements ($\Delta y = \bar{y} + 2n\pi$) goes to zero as $e^{-n^2/\Delta t}$. Thus for small $\Delta t$ a rather crude truncation procedure may provide adequate accuracy. In fact, it may be appropriate to approximate the discrete filter by the continuous filter preceded by a sample and hold (see Figure 4.7).

Note that these "$\Delta t$ approximations" become even better if we know that our noises are amplitude bounded, since we can then bound the measurement size, and we immediately have that the number of nonzero terms in the various summations is finite. In fact, if we choose $\Delta t$
Figure 4.6 Conceptual Diagram of the Truncation Method for Suboptimal Discrete-Time Filtering. (Each linear filter computes $r$ estimates, using as data the given initial estimate $\eta_i$, the precomputed variance, and the $r$ chosen inputs $\tilde{y}_k + 2n_i \sigma$, $i = 1, \ldots, r$.)
small enough, we can be certain that there are no rotations in the
\( \Delta t \) between measurements, and thus \( \Delta y \mod 2\pi = \Delta y \). However, the
analytical convenience of the normal density, which is nonzero on the
whole real line, has motivated our use of it in this section, and
thus we cannot assume our processes are amplitude limited (although for
suboptimal design we may use the results assuming normal densities and
still truncate the series as if the processes were amplitude limited).

An additional comment can be made if we are interested in estimating
\( \theta_k = x_k \mod 2\pi \) as opposed to \( x_k \). In the linear case, the conditional
density \( p_\theta(\theta_k|\tilde{y}_k) \) is an infinite sum of folded normal densities. For
instance, if we fold (4.162), we have

\[
p_\theta|\tilde{y}(\alpha|\beta) = \sum_{n=-\infty}^{+\infty} c_n(\beta) F(\alpha;\eta_n, \gamma_3)
\]

(4.175)
where \( \eta_n \) is given by (4.164). Since

\[
F(\alpha; \eta, \gamma) = F(\alpha; \eta + 2\pi n, \gamma)
\]  
(4.176)

We need only keep those \( \eta_n \) that are inequivalent mod \( 2\pi \). For example, suppose \( \gamma_1 \) and \( \gamma_2 \) are rationally related -- i.e. \( \gamma_1 / \gamma_2 \) is a rational number -- and \( h = 1 \). Then let \( \eta_0 \) be the smallest positive integer such that

\[
\frac{\gamma_1 \eta_0}{\gamma_1 + \gamma_2} = \text{an integer}
\]  
(4.177)

Referring to (4.164), we have that

\[
\eta_{n_0 + k} = \eta_k \text{ mod } 2\pi \quad k
\]  
(4.178)

and thus there are only \( \eta_0 \) distinct folded normal densities in (4.175), with "means" \( \eta_0, \ldots, \eta_{\eta_0 - 1} \). In this case, we can write (4.175) as a sum of a finite number of folded normal densities, where the coefficient of each of the terms is computed by summing the \( c_n(\theta) \) corresponding to all those \( \eta_n \) that are equivalent mod \( 2\pi \) to a particular \( \eta_j \). Thus, if we approximate \( \gamma_1 \) and \( \gamma_2 \) so that \( \eta_0 \) is small, \( p_{\theta | \parallel} \) is the sum of only a few terms.

We now comment on the optimal estimation problem. We assume that

\[
p_x(\alpha) = \sum_{n=-\infty}^{\infty} c_n N(\alpha; \eta_n, \gamma)
\]  
(4.179)

As discussed by Jazwinski [J2], the mean \( \delta(x) \) is the minimum variance estimate and in this case

\[
\delta(x) = \sum_{n=-\infty}^{\infty} c_n \eta_n
\]  
(4.180)
Then in the linear-Gaussian case depicted in Figure 4.6, the optimal estimate is a linear combination of the outputs of the various linear filters, and the coefficients are the same as the coefficients computed to decide which terms to keep.

If we wish to estimate $\hat{\theta} = x \mod 2\pi$, we use the results of Section 3.3. In particular, if we wish to minimize $\mathcal{E}(1 - \cos(\theta - \hat{\theta}))$, the optimal estimate $\Theta_o$ is given by

$$\Theta_o = \tan^{-1} \left( \frac{\sum_{n=-\infty}^{\infty} c_n \sin \eta_n}{\sum_{n=-\infty}^{\infty} c_n \cos \eta_n} \right) \tag{4.181}$$

so, again referring to Figure 4.6, the optimal estimate is a nonlinear function of the outputs of the linear filters and the coefficients $c_n$.

As in the continuous time case, we can consider the multidimensional analog of the results of this section. For instance, let $x$ be an $n$-dimensional normally distributed random variable, $w$ a $k$-dimensional normal random variable independent of $x$, and $C$ a $k \times n$ matrix. Define the measurement $\tilde{y}$

$$y = Cx + w \tag{4.182}$$

$$\tilde{y}_i = \begin{cases} y_i & 1 \leq i \leq k_1 \\ y_i \mod 2 & k_1 < i \leq k \end{cases} \tag{4.183}$$

Then the conditional density $p_{x|\tilde{y}}$ can be written as a $(k-k_1)$-times countably infinite sum of normal distributions, the $(r_1, \ldots, r_{k-k_1})$th of which is the conditional density assuming
\[ y_{k_1+1} = \bar{y}_{k_1+1} + 2r_{i_1} \quad i=1, \ldots, k-k_1 \] (4.184)

and the coefficient of this term is the conditional probability that (4.184) holds, given \( \bar{y} \).
CHAPTER 5

THE USE OF FOURIER SERIES FOR PHASE TRACKING AND SYNCHRONOUS COMMUNICATION

"No pleasure is fully delightful
without communication"
- Montaigne [M19]

5.1 Introduction

In Section 3.2 we saw how one could express the optimal estimate
of a random variable on $S^1$ in terms of the Fourier coefficients of
the probability density of the random variable. In view of this, in
considering the estimation of a random process on $S^1$, it is natural
to seek methods for tracking the Fourier coefficients. In this chapter,
we will consider some aspects of this type of problem. Section 5.2
deals with a very general discrete-time problem, and in Section 5.3 we
consider some specific continuous-time phase-tracking and demodulation
problems. Some computational aspects of these problems will be considered,
and reference will be made to more detailed discussions contained in
Appendix D.

The purpose of this chapter is to point out the usefulness of Fourier
series analysis as a tool in estimation problems. The treatment here
does not include all possible cases but rather is designed to indicate the
range of problems that can be handled by the use of Fourier techniques.
Some of these ideas have been discussed previously in [B9] and [M14]. Also,
the reader is referred to Chapter 6, in which we discuss some applications
of these methods and present some numerical results.
5.2 A General Discrete-Time $S^1$ Estimation Problem

In this section we will first consider a general single stage $S^1$ estimation problem. Extensions to multistage problems and computational considerations will be discussed later in the section.

Let $\theta$ be a random variable on the circle with a priori density

$$p_\theta(\xi) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} [a_n(0)\sin n\xi + b_n(0)\cos n\xi] \quad (5.1)$$

Suppose that we take a single (possibly nonlinear) measurement $y$ of $\theta$ and that the noise density $p_y|\theta(y|\xi)$ exists. Considering this as a function of $\xi$ for fixed $\nu$, we must have

$$p_y|\theta(y|\xi + 2\pi) = p_y|\theta(y|\xi) \quad (5.2)$$

Thus, we can write $p_y|\theta(y|\xi)$ in Fourier series form in $\xi$

$$p_y|\theta(y|\xi) = d_0(\nu) + \sum_{n=1}^{\infty} [c_n(\nu)\sin n\xi + d_n(\nu)\cos n\xi] \quad (5.3)$$

where the $c_n$'s and $d_n$'s are functions of $\nu$. Applying Bayes' rule

$$p_\theta|y(\xi|\nu) = \frac{p_y|\theta(y|\xi)p_\theta(\xi)}{\int_{-\infty}^{\infty} p_y|\theta(y|u)p_\theta(u)du} \quad (5.4)$$

we can compute the Fourier series form for the conditional density $p_\theta|y(\xi|\nu)$

$$p_\theta|y(\xi|\nu) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} [a_n(1)\sin n\xi + b_n(1)\cos n\xi] \quad (5.5)$$

where

$$a_n(1) = \frac{\alpha_n(\nu)}{2\pi c(\nu)} \quad b_n(1) = \frac{\beta_n(\nu)}{2\pi c(\nu)} \quad (5.6)$$
\[ c(\nu) = \frac{1}{2\pi} p_y (\nu) = \frac{d_0 (\nu)}{2\pi} + \frac{1}{2} \sum_{n=1}^{\infty} \left[ a_n (0) c_n (\nu) + b_n (0) d_n (\nu) \right] \] (5.7)

\[ \alpha_k (\nu) = a_k (0) d_0 (\nu) + \frac{c_k (\nu)}{2\pi} + \frac{1}{2} \sum_{n=1}^{k-1} \left[ a_n (0) d_{k-n} (\nu) + b_n (0) c_{k-n} (\nu) \right] \]

\[ + \frac{1}{2} \sum_{n=1}^{\infty} \left\{ \left[ a_n (0) d_n (\nu) + b_n (0) c_n (\nu) \right] - \left[ a_n (0) d_{n+k} (\nu) + b_n (0) c_{n+k} (\nu) \right] \right\} \] (5.8)

\[ \beta_k (\nu) = b_k (0) d_0 (\nu) + \frac{d_k (\nu)}{2\pi} + \frac{1}{2} \sum_{n=1}^{k-1} \left[ b_n (0) d_{k-n} (\nu) - a_n (0) c_{k-n} (\nu) \right] \]

\[ + \frac{1}{2} \sum_{n=1}^{\infty} \left\{ \left[ a_n (0) c_n (\nu) + b_n (0) d_n (\nu) \right] + \left[ a_n (0) c_{n+k} (\nu) + b_n (0) d_{n+k} (\nu) \right] \right\} \] (5.9)

Note that the equations for \( c, \alpha_k, \) and \( \beta_k \) are bilinear in the Fourier coefficients of \( p_\theta (\xi) \) and \( p_y |_y (\nu | \xi) \). Thus the computation of \( p_\theta |_y \) involves the (in general nonlinear) computation of the coefficients \( \{ c_n (\nu) \} \) and \( \{ d_n (\nu) \} \), the evaluation of the bilinear equations (5.7)-(5.9), and the normalization (5.6). Thus, although the equations (5.7)-(5.9) look complicated, they are highly structured. The reader is referred to the comments at the end of this section on some of the computational aspects of these equations. Also, see Example 9.5, where we consider the efficient computation of an extremely simplified version of this problem.

The following examples indicate the wide variety of problems that can be handled in this manner.

**Example 5.1:** Let \( \theta \) be an \( S^1 \) random variable with density given by (5.1). Let \( v \) be a real-valued random variable, independent of \( \theta \), with density \( p_v (v) = N(v; 0, \gamma) \). Consider the observation

\[ y = \sin \theta + v \] (5.10)

In this case we see that
\[ p_{y \mid \theta}(v \mid \xi) = N(\nu - \sin \xi; 0, \gamma) \]  
\hspace{1cm} (5.11)

We now wish to write \( p_{y \mid \theta}(v \mid \xi) \) in a Fourier series expansion in \( \xi \).

Using some formulae related to Bessel functions of integral order [All], [H6], we find that

\[ p_{y \mid \theta}(v \mid \xi) = d_0(v) + \sum_{n=1}^{\infty} [c_n(v) \sin n \xi + d_n(v) \cos n \xi] \]  
\hspace{1cm} (5.12)

where

\[ d_0(v) = \eta(v) \sum_{k=-\infty}^{\infty} (i)^k J_{-k}(-\frac{i v}{4 \gamma}) J_k(-\frac{i}{4 \gamma}) \]  
\hspace{1cm} (5.13)

\[ c_n(v) = \eta(v) \sum_{k=-\infty}^{\infty} (i)^k J_k(-\frac{i}{4 \gamma}) [J_{-(n+k)}(-\frac{i v}{4 \gamma}) - J_{(n-k)}(-\frac{i v}{4 \gamma})] \]  
\hspace{1cm} (5.14)

\[ d_n(v) = \eta(v) \sum_{k=-\infty}^{\infty} (i)^k J_k(-\frac{i}{4 \gamma}) [J_{(n-k)}(-\frac{i v}{4 \gamma}) + J_{-(n+k)}(-\frac{i v}{4 \gamma})] \]  
\hspace{1cm} (5.15)

where \( i = \sqrt{-1} \) and

\[ \eta(v) = \frac{1}{\sqrt{2\pi\gamma}} e^{-\frac{(2v^2+1)}{4\gamma}} \]  
\hspace{1cm} (5.16)

Also, \( J_k \) is the Bessel function of the first kind of order \( k \), given by

\[ J_k(u) = (\frac{i}{2})^k u^n \sum_{n=0}^{\infty} \frac{(-\frac{1}{4} u^2)^n}{n!(k+n)!} \]  
\hspace{1cm} (5.17)

We note that the \( c_n \) and \( d_n \) can also be expressed in terms of modified Bessel functions [All], [H6].

Thus, given an observation value \( \nu \), we can use (5.13)-(5.17) together with (5.5)-(5.9) to compute the conditional density. Of course these computations involve infinite series, and some approximations must be made. We will discuss this question further later in this section.
Example 5.2: We now consider a discrete-time estimation problem of the type discussed in Section 4.4. Let $\theta$ and $\nu$ be as in the previous example, and suppose our observation is

$$y = (\theta + \nu) \mod 2\pi$$  \hspace{1cm} (5.18)

In this case,

$$p_{y|\theta}(\nu|\xi) = \mathcal{F}(\nu - \xi; 0, \gamma)$$  \hspace{1cm} (5.19)

and

$$p_{y|\theta}(\nu|\xi) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} e^{-n^2\gamma/2}[\cos n\nu \cos n\xi + \sin n\nu \sin n\xi]$$  \hspace{1cm} (5.20)

We can then compute the Fourier series form for $p_{\theta|y}$ from (5.5)-(5.9).

Example 5.3: Consider the situation depicted in Figures 5.1 and 5.2. A hollow sphere is sitting inside an infinite cylinder. Consider the coordinate systems depicted in Figure 5.1. The $xyz$ system is fixed, and the sphere is allowed to rotate freely about the $z$ axis. This rotation is denoted by the angle $\phi$ and is not of importance. The axes $x'$ and $y'$ are in the plane perpendicular to $z$ and are rotated an amount $\phi$ from the $x$ and $y$ axes (i.e., the $x'$ and $y'$ axes rotate with the sphere about the $z$ axis). We are interested in determining the value of the rotation $\theta$ that the sphere performs about the $x'$ axis.

Our measurement process can be described as follows: a point light source is positioned at the center of the sphere, and there is a pinhole in the sphere and the $y'z$ plane. As shown in Figures 5.1 and 5.2, the rotation $\theta$ is determined by the angle the line from the center to the pinhole makes with the $y'$ axis. To measure this angle,
Figure 5.1: The Physical Situation Described in Example 5.3

Figure 5.2: The Observation Geometry of Example 5.3
the inside surface of the cylinder is lined with photodetectors. Consequently, by measuring where the light from the point source hits the cylinder, we can measure the distance \( d \). However, as depicted in Figure 5.2, noise enters the system because the light source can move randomly in the \( z \) direction (although it is restrained in the \( y' \) and \( x' \) directions). The amount of this deviation is a random variable \( \nu \) with uniform density on the interval \([-\varepsilon, \varepsilon]\).

The measured value of \( d \) is \( d_m \), which is given by (if we ignore the singular points \( \theta = \pm \pi/2 \))

\[
d_m = 2\tan \theta - \nu \left( \frac{2}{\cos \theta} - 1 \right) \quad \theta \in \left( -\frac{\pi}{2}, \frac{\pi}{2} \right) \tag{5.21}
\]

\[
d_m = -2\tan \theta + \nu \left( \frac{2}{\cos \theta} + 1 \right) \quad \theta \in \left( \frac{\pi}{2}, \frac{3\pi}{2} \right) \tag{5.22}
\]

We can compute \( p_{d_m|\theta(\nu|\xi)} \)

\[
p_{d_m|\theta(\nu|\xi)} = \frac{1}{2\varepsilon \left( \frac{2}{\cos \xi} - 1 \right)} \quad \xi \in \left( -\frac{\pi}{2}, \frac{\pi}{2} \right)
\]

\[
\nu \in [2\tan \xi - \varepsilon \left( \frac{2}{\cos \xi} - 1 \right), 2\tan \xi + \varepsilon \left( \frac{2}{\cos \xi} - 1 \right)]
\]

\[
(5.23)
\]

\[
p_{d_m|\theta(\nu|\xi)} = -\frac{1}{2\varepsilon \left( \frac{2}{\cos \xi} - 1 \right)} \quad \xi \in \left( \frac{\pi}{2}, \frac{3\pi}{2} \right)
\]

\[
\nu \in [-2\tan \xi + \varepsilon \left( \frac{2}{\cos \xi} + 1 \right), -2\tan \xi - \varepsilon \left( \frac{2}{\cos \xi} + 1 \right)]
\]

(5.24)

The computation of the Fourier series decomposition of \( p_{d_m|\theta} \) in \( \xi \) for fixed \( \nu \) is complicated and will not be computed here. For practical application, we can, of course, approximate the \( c_n(\nu) \) and \( d_n(\nu) \) by polynomials, rational functions, or some other class of functions of \( \nu \).
These examples indicate the wide variety of problems that can be handled by means of Fourier series methods. We can also handle the problem of computing conditional densities of random processes given a series of discrete measurements. Assuming that the measurement noises of the various measurements are independent of each other and of the process $\theta$ that is to be estimated, we can process each measurement as in (5.5)-(5.9) and can propagate the density between measurements via some sort of "diffusion" equation. Of course, if we are using the Fourier series representation for probability densities, our diffusion update equations should be in terms of the Fourier coefficients. We will consider two examples of random processes on $S^1$ to show how the Fourier series approach leads to simple diffusion update equations.

**Example 5.4:** Suppose our random process $\theta(t)$ is a standard Brownian motion process on $S^1$ and that we take possibly nonlinear measurements $y_k$ at the time instants $t = k\Delta$, $k = 1, 2, \ldots$. The measurement updates are handled by (5.5)-(5.9), and thus we need only consider how to compute $p(\theta((k+1)\Delta) = \theta | y_1, \ldots, y_k)$ from $p(\theta(k\Delta) = \theta | y_1, \ldots, y_k)$. Because of the independence assumptions, the value at $t = (k+1)\Delta$ of $p(\theta, t)$ satisfying the diffusion equation

$$
\frac{3p}{3t} - \frac{1}{2} \frac{3^2 p}{3\theta^2} = 0
$$

(5.25)

with initial condition

$$
p(\theta, k\Delta) = p(\theta(k\Delta) = \theta | y_1, \ldots, y_k)
$$

(5.26)

is $p(\theta((k+1)\Delta) = \theta | y_1, \ldots, y_k)$. Writing

$$
p(\theta(k\Delta) = \theta | y_1, \ldots, y_k) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} a_n(k|k)\sin n\theta + b_n(k|k)\cos n\theta
$$

(5.27)
\[ p(\theta((k+1)\Delta) = \theta | y_1, \ldots, y_k) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} a_n(k+1|k) \sin n\theta + b_n(k+1|k) \cos n\theta \]

we have

\[ a_n(k+1|k) = e^{-n^2\Delta/2} a_n(k|k) \quad b_n(k+1|k) = e^{-n^2\Delta/2} b_n(k|k) \] (5.29)

Thus, the Fourier series diffusion update equations are quite simple in this case. The next example contains a more general result.

Example 5.5: Consider a discrete-time random process \( \theta_k \) that satisfies

\[ \theta_{k+1} = (\theta_k + \nu_k) \mod 2\pi \] (5.30)

where the \( \nu_k \) are independent random variables on \( S^1 \). Note that writing \( \theta(k\Delta) = \theta_k \), the previous example is of this type, since a Brownian motion process has independent increments. Again, we will consider only the diffusion update part of the computation. We wish to compute

\[ p(\theta((k+1)\Delta) = \theta | y_1, \ldots, y_k) \]

from \( p(\theta(k\Delta) = \theta | y_1, \ldots, y_k) \), where these densities are given by (5.27) and (5.28). Assuming that the density for \( \nu_k \) can be written as

\[ p(\nu_k = \nu) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} c_n \sin n\nu + d_n \cos n\nu \] (5.31)

we have the convolution form of the diffusion update

\[ p(\theta((k+1)\Delta) = \theta | y_1, \ldots, y_k) = \int_{0}^{2\pi} p(\theta(k\Delta) = \theta - \xi | y_1, \ldots, y_k) p(\nu_k = \xi) d\xi \] (5.32)

which leads to the Fourier series update equations

\[ a_n(k+1|k) = \pi [d_n a_n(k|k) + c_n b_n(k|k)] \] (5.33)
\[ b_n(k+1|k) = \pi [d_n b_n(k|k) - c_n a_n(k|k)] \] (5.34)

which are also quite simple to use, since the nth coefficients of the updated density depend only on the nth coefficients of the prior density.

We also note that we could use the complex Fourier series form for the densities -- i.e.

\[ p(\theta) = \sum_{n=-\infty}^{+\infty} f_n e^{i n \theta} \] (5.35)

For example, for the density \( p(\theta(k\Delta)=\theta|y_1,\ldots,y_k) \), we have

\[ f_n(k|k) = \begin{cases} 
\frac{1}{2} [a_n(k|k)+ib_n(k|k)] & n>0 \\
\frac{1}{2\pi} & n=0 \\
\frac{1}{2} [a_n(k|k)-ib_n(k|k)] & n<0 
\end{cases} \] (5.36)

In this context, (5.33) and (5.34) become

\[ f_n(k+1|k) = 2\pi g_n f_n(k|k) \] (5.37)

where the \( g_n \) are the complex coefficients for \( p(w_k=\nu) \). Also, the complex Fourier coefficients in (5.34) can be written as

\[ f_n = \frac{1}{2\pi} \Phi(e^{-i n \theta}) \] (5.38)

This, plus the fact that the convolution of densities (5.32) corresponds to the multiplication of Fourier coefficients (5.37) indicates the close relationship between the Fourier coefficients and the characteristic function of a real random variable [P2]. The reader is referred to the related discussion of group characters of cyclic groups in Chapter 10 and the generalization of Fourier series in Chapter 7 (the reader is
also referred to \([K17]\) and \([G8]\)).

We also note that the last two examples show that we do not get the cross-coupling effect of Fourier coefficients when computing diffusion updates of the types described in these examples -- i.e. \(a_n(k+1|k)\) and \(b_n(k+1|k)\) do not depend on \(a_j(k|k)\) and \(b_j(k|k)\) except for \(j=n\). Thus, for the prediction problem our computational load is quite simple. For instance if we wish to choose an estimate to minimize \(\mathcal{E}(1-\cos(\theta-\hat{\theta}))\), and if \(\theta(t)\) is of the type discussed in Example 5.5, we need only retain \(a_1\) and \(b_1\), and, if we are doing pure prediction, \(a_1(t)\) and \(b_1(t)\) do not depend on \(a_k(0)\) and \(b_k(0)\) except for \(k = 1\). Thus, we can drop all other Fourier coefficients. As we will now see, the complexity of the problem is much greater if we wish to process measurements.

We now make some remarks concerning the computational aspects of the Fourier series measurement update equations \((5.5)-(5.9)\). First of all, since we can only store a finite number of terms, we must truncate the various series. As discussed in Section 3.2, this is not terribly serious since the coefficients fall off rapidly in size. Also, if we are using an estimation criterion such as \(\mathcal{E}(1-\cos(\theta-\hat{\theta}))\) (see Example 3.1), at any stage we actually use only \(a_1\) and \(b_1\). Of course, equations \((5.5)-(5.9)\) show that, in general, all of the a priori coefficients directly influence each of the updated coefficients.

The problem of truncating the probability densities is itself an interesting problem. Suppose we keep the first \(N\) modes of \(p_{\theta}\) and the first \(M\) modes of \(p_{y|\theta}\). It is easy to check that, in general, \(p_{\theta|y}\) will then have nonzero terms up to the \((N+M)\)th mode. Thus, if we are considering
a sequence of measurements, we must devise techniques for sequentially truncating the conditional density. In addition, we must truncate the equations (5.5)-(5.9). As discussed in Appendix D, if we keep N modes of \( p_\theta | y \), and if we make some assumption about the shape of \( p_\theta | y \) (this is called "assumed density" approximation [L11]), we can use \( \{a_n, b_n\}_{n=1}^N \) to approximate the higher coefficients, which can then be used in computing the first N updated coefficients. Such a procedure may provide better accuracy than a straight truncation of the equations. An example that shows the type of errors that enter when we truncate the various Fourier series is presented in Appendix G (where, for simplicity, we use the straightforward truncation approach).

In general, just keeping the first N modes of \( p_\theta | y \) may not be acceptable, since the truncated density may not be strictly non-negative. However, if the coefficients fall off as \( \frac{1}{n^2} \) (for instance, if the density is continuous), if we keep enough coefficients, this may not be a problem. A better approximation might be the following: if \( \tilde{p} \) represents the truncated version of \( p_\theta | y \) obtained by keeping only the first N modes, and we define

\[
\tilde{p} = \max(0, \tilde{p})
\]  

we can take the Fourier coefficients of \( \tilde{p} \) as the coefficients of our approximation. It should be noted that the calculation of the Fourier coefficients of \( \tilde{p} \) is rather involved, and it is not clear that this procedure is computationally feasible. The reader is referred to further discussions of the truncation problem in Appendix D and to Section 6.4, in which a particular assumed density approximation is
used to design a continuous-time phase tracker.

If we truncate the Fourier series and update equations in the straightforward manner, equations (5.7)-(5.9) can be written in the form

$$g(v) = A(v)h$$  \hspace{1cm} (5.40)

where $h$ is the vector whose elements are the Fourier coefficients of $p_\theta$, $g(v)$ contains $c(v)$ and the $a_k(v)$ and $b_k(v)$, and $A(v)$ is a $(2N+1) \times 2N$ matrix (assuming we keep $N$ modes of $p_\theta$ and $p_\theta|\theta$) whose elements are the Fourier coefficients of $p_y|\theta$. The structure of (5.7)-(5.9) is reflected in (5.40) and may lead to efficient methods for evaluating (5.40). In fact, the computational efficiency results of Chapter 9 may be applicable here (see Example 9.5, where we discuss

![Recursive Computation of Conditioned Fourier Coefficients-Measurement and Diffusion Updates](image)

**Figure 5.3:** Block Diagram of Discrete-Time Optimal Estimator for the Criterion $\mathcal{E}(1-\cos(\theta-\hat{\theta}))$.  


the very simple case of N=2).

Finally, assuming that we somehow implement equations (5.5)-(5.9), we can then use the Fourier analysis results of Section 3.2 to obtain an optimal estimate. For instance, Figure 5.3 depicts how such an estimator would work if the estimate were chosen to minimize $\mathcal{E}(1-\cos(\theta-\hat{\theta}))$.

5.3 Fourier Analysis and Continuous-Time Phase Tracking and Demodulation

In this section we wish to consider a class of estimation problems of importance in a number of communication applications. We wish to consider the processing of a signal of the form

$$ r(t) = \sin(\omega_c t + \phi(t) + v(t)) + \dot{\hat{\phi}}(t) $$

(5.41)

where $\omega_c$ is a carrier frequency, $\phi$ is some type of modulating information, $v$ is a phase drift, and $\dot{\hat{\phi}}$ is additive channel noise. With this model, we can consider phase tracking, phase demodulation, and frequency demodulation. Standard techniques for such problems involve a system called a phase-lock loop (PLL) [S7], [V1], [V2], [M6], [B21]. The reader is referred to Example 5.6 and to Chapter 6, where we review the basic ideas behind the PLL and present some results comparing the performance of PLL systems with the performance of systems designed using the techniques developed here and in Chapter 4. The reader is also referred to Chapter 10, in which we discuss a quantized version of the phase-tracking problem.

In the following discussion we utilize the tool of Fourier series analysis to design systems for handling (5.41). If we desire only to track the phase
\[ \theta(t) = (\omega_c t + \phi(t) + v(t)) \mod 2\pi \] (5.42)

of the signal — e.g. in Loran systems [P4] — the system we will design is simply a phase tracker (PT). Such a tracking problem using Fourier series has been previously considered in [M14], and our result is essentially the same as theirs. If \( x(t) \) is the signal to be recovered and \( \phi(t) = cx(t) \) with \( c \) a constant (we could consider it to be time varying), the signal \( r(t) \) is called phase modulated, and the system we will design to recover \( x \) is called a phase demodulator (PD). If \( \phi(t) = c \int_0^t x(s)ds \), \( r(t) \) is frequency modulated and our system is a frequency demodulator (FD).

The design techniques we will use to construct PD's and FD's are new. However, they can be viewed as extensions of the PT design technique described in [M14] and discussed later in this section. We note that a phase demodulation problem is considered in [B9], and two techniques, one involving discrete-time Fourier series equations similar to those of Section 5.2, are discussed. It is worth noting that the approach of this section and the approaches used in [M14] and [B9] are similar in the sense that the goal of all these approaches is to devise methods for designing filters that utilize the inherent structure of the problem and that are of practical value.

In addition to the fact that [M14] and [B9] use Fourier series as we do, there are two other points that indicate the connection between our work and the results in [M14] and [B9]. First of all, we shall see that the optimal solutions to the problems discussed in the rest of this section are infinite dimensional in nature, and in Appendix D we will
consider some finite dimensional approximation techniques including the Fourier series truncating techniques mentioned in Section 5.2. In addition to the techniques discussed in Appendix D, it is possible that the finite dimensional approximation techniques discussed in [B9] may be of some help in designing suboptimal filters for our problems. Also, the phase demodulation and general nonlinear filtering technique of most importance in [B9] requires a quantization of the state space into a grid of points. In Chapter 10 we will use a different type of quantization to design phase tracking and demodulation systems that are appealing because of the similarity between the techniques used in designing these systems and the Fourier series analysis techniques used in the present chapter.

We now consider the phase tracking, phase demodulation, and frequency demodulation problems. Instead of describing general techniques, we will present an example of each type of problem in order to illustrate the basic concepts behind these Fourier series techniques.

**Example 5.6**: We now consider a phase tracking problem that is essentially the same as that considered in [M14]. This problem is the continuous time analog of Example 5.1, and one should note the simpler form of the Fourier coefficient equations (although the optimal solution is still infinite dimensional). Suppose we receive the signal

\[ \dot{x}(t) = \sin \theta(t) + r^{1/2}(t) \dot{w}(t) \]  
(5.43)

where

\[ \theta(t) = \omega_c t + \int_0^t q^{1/2}(s) dv(s) + \theta_0 \]  
(5.44)

and \( v \) and \( w \) are independent Brownian motions, \( q(t) \geq 0, r(t) > 0 \), and
\( \omega_c > 0 \). Also \( \theta_0 \) is a random initial condition independent of \( v \) and \( w \).

Suppose we wish to estimate \( \theta(t) \mod 2\pi \) given \( \{ z(s) | 0 \leq s \leq t \} \) — i.e. we wish to filter out \( r^{1/2} \) and track the phase. Equation (5.43) is, of course, only formal, since \( \hat{\phi} \) is white noise. The Ito differential forms of (5.43) and (5.44) are

\[
d\theta(t) = \omega_c dt + q^{1/2}(t) dv(t) \quad \theta(0) = \theta_0	ag{5.45}
\]

\[
dz(t) = \sin \theta(t) dt + r^{1/2}(t) dw(t)	ag{5.46}
\]

A natural optimality criterion for our estimate \( \hat{\theta}(t) \) is to choose \( \hat{\theta}(t) \) to minimize \( \mathcal{E}[(1-\cos(\theta(t)-\hat{\theta}(t))) | z(s), 0 \leq s \leq t] \). As discussed previously, it then makes sense to compute the Fourier coefficients

\[
a_n(t) = \frac{1}{\pi} \mathcal{E}[\sin n\theta(t) | z(s), 0 \leq s \leq t] \tag{5.47}
\]

\[
b_n(t) = \frac{1}{\pi} \mathcal{E}[^{n}\cos \theta(t) | z(s), 0 \leq s \leq t] \tag{5.48}
\]

(note that \( b_0(t) = \frac{1}{\pi} \) is not the zeroth Fourier coefficient, which is \( \frac{1}{2\pi} \)). We then have the optimal estimate equation

\[
\hat{\theta}(t) = \tan^{-1} \left[ \frac{a_1(t)}{b_1(t)} \right]	ag{5.49}
\]

Stochastic differential equations for \( a_n \) and \( b_n \) can be obtained using the results of Kushner [K5], [K18], [J2, Lemma 6.3]:

\[
da_n(t) = [n\omega_c b_n(t) - \frac{n^2 q(t)}{2} a_n(t)] dt + \left[ \frac{1}{2} (b_{n-1}(t) - b_{n+1}(t)) - n\pi a_n(t) a_1(t) \right] \frac{(dz(t) - \pi a_1(t) dt)}{r(t)}
\]

\[
rb_n(t) = -[n\omega_c a_n(t) + \frac{n^2 q(t)}{2} b_n(t)] dt + \left[ \frac{1}{2} (a_{n+1}(t) - a_{n-1}(t)) - n\pi b_n(t) a_1(t) \right] \frac{(dz(t) - \pi a_1(t) dt)}{r(t)}
\]
where
\[ a_0(t) = 0 \quad b_0(t) = \frac{1}{\pi} \] (5.52)

Note that these equations are simpler than the discrete-time equations of Example 5.1, since the equations for the nth mode coefficients depend only on the first, (n-1)st, nth, and (n+1)st mode coefficients, instead of all of the coefficients (also we don't have to compute the Bessel functions of equations (5.13)-(5.17)). Thus, as in Chapter 4, in this case a reasonable approximation to the discrete time solution might be the continuous time solution preceded by a sample and hold, or we might feed the discrete-time observations into a digital filter that numerically integrates the continuous filter equations.

The structure of the optimal filter, which is illustrated in Figures 5.4 and 5.5, deserves some further comment. The filter consists of an infinite bank of filters. From Figure 5.5, we see that each of these filters essentially is a damped oscillator, with oscillator frequency \[ \frac{n\omega C}{2\pi} \] Hz. It can be shown that the damping term is the second order term that arises in Itô calculus in precisely the same way that the damping terms entered in the analysis of Sections 4.2 and 4.3 (see [G3], where it is shown that filter performance is degraded if we do not include the damping terms in the oscillator; the filter considered in [G3] essentially is a linear approximation to (5.49)-(5.51)). Also, from Figure 5.5, we see that one of the feedback terms in the \( (a_n, b_n) \) filter involves the multiplication of \( a_n \) and \( b_n \) by \( \dot{z} \).

Consider now the phase-lock loop depicted in Figure 5.6. We will discuss this in somewhat greater detail in Chapter 6. Basically the
Figure 5.4: Illustrating the Form of the Infinite Dimensional Optimal Filter of Example 5.6
Figure 5.5: Diagram of the \((a_n, b_n)\) Filter
Shown in Figure 5.4

Figure 5.6: Diagram of a Phase-Lock Loop
PLL has a voltage controlled oscillator (VCO), operating at the frequency $\omega_c$, in the feedback loop. The output of the oscillator looks something like an estimate of $\sqrt{2\pi b_1}$, and this output multiplies the received signal $\tilde{z}$. This VCO-multiplication part of the PLL looks very similar to the form of the $(a_n, b_n)$ filter of Figure 5.5, and thus the optimal system of Figure 5.4 looks like an infinite bank of PLL's, the $n$th of which operates at an oscillator frequency of $n\omega_c$. We also note that our oscillator (Figure 5.5) is a band-pass filter, much like the one used in [G3] for $n = 1$.

For practical application of the optimal filter results developed here, we must devise techniques for truncating the infinite bank of filters. Of course, we can perform the straightforward truncation — assuming that all higher mode coefficients are zero. Other possible numerical approximation techniques will be discussed in Appendix D and in Chapter 6. One physically appealing truncation method is suggested by the PLL. In this system, the product of the received signal and the output of the VCO contains a term with a carrier frequency $2\omega_c$ and one with no carrier frequency ($\sin(\omega_c t+\alpha)\cos(\omega_c t+\beta) = \frac{1}{2}[\sin(2\omega_c t+\alpha+\beta)+\sin(\alpha-\beta)]$). The low pass filter is used to filter out the $2\omega_c$ term and to perform some other filtering (see Chapter 6). The PLL looks like a truncation of the infinite bank of filters where we keep only the lowest mode, and the coupling of this mode to the higher modes is removed by the filtering out of the $2\omega_c$ term. This suggests the following idea: suppose we keep the first $n$ filters in Figure 5.4. We then must account for the effects of $a_{n+1}$ and $b_{n+1}$ in the $(a_n, b_n)$ filter. Following the lead of the PLL,
we can ask if it is possible to include a LPF in the \((a_n, b_n)\) filter that removes this higher frequency \((n+1)\omega_c\) effect (see Appendix D).

We note that using the Ito differential rule [J2] and equations \((5.49)-(5.53)\), we can write out the Ito differential equation satisfied by \(\hat{\theta}\). This equation is complicated and will not be displayed here.

Finally, we note one very appealing feature of our filter -- it is time-invariant (see Figures 5.4 and 5.5) if \(q\) and \(r\) are constant and is equipped to handle any initial conditions. If we are interested in the pure synchronization problem after we have acquired the signal, we might impose the initial conditions \(a_n(0) = 0, b_n(0) = \frac{1}{n} \) \(\forall n \geq 1\), which correspond to our knowing that the initial phase is 0. On the other hand, setting \(a_n(0) = b_n(0) = 0 \) \(\forall n \geq 1\) (note that \(b_0 = \frac{1}{n}\)) corresponds to the assumption that the initial phase is uniformly distributed -- i.e., this is the acquisition problem. Thus, suppose we can build a finite-dimensional approximation to the infinite dimensional filter given by \((5.49)-(5.52)\). Then, assuming we retain the time-invariant property of the original system, we see that the same filter can handle either synchronization or acquisition merely by the proper choice of initial conditions! The reader is referred to Chapter 6 where we present and discuss some simulation results for the problem discussed in Example 5.6.

**Example 5.7:** In this example we will consider a phase demodulation problem. Consider the \(R^1\) signal process \(x(t)\) satisfying the Ito differential equation

\[
dx(t) = a(t)x(t)dt + q^{1/2}(t)dv(t)
\]  
\((5.53)\)
where \( v \) is a standard Brownian motion process independent of the initial condition \( x(0) \). We wish to estimate \( x(t) \) given the phase-modulated observation process

\[
\dot{x}(t) = \sin(\omega_c t + x(t)) + r^{1/2}(t) \dot{w}(t)
\]

(5.54)

where \( w \) is a Brownian motion independent of \( v \) and \( x(0) \). The Ito form of the formal equation (5.54) is

\[
dz(t) = \sin(\omega_c t + x(t)) dt + r^{1/2}(t) dw(t)
\]

(5.55)

Suppose we want the minimum variance estimate of \( x(t) \) given \( \{z(s)|0 \leq s \leq t\} \). It is known [J2] that the desired estimate is the conditional expectation \( \mathbb{E}[x(t)|z(s), 0 \leq s \leq t] = \hat{x}(t|t) \). As in Example 5.6, we shall write the optimal demodulator equations in terms of some very special functions. The form of the optimal demodulator, which is infinite dimensional, is appealing physically and suggests some new suboptimal demodulation techniques. The time correlation of \( x(t) \) (the presence of the term \( a(t)x(t)dt \) on the right-hand side of (5.53)) makes this problem somewhat more complicated than the preceding example. In fact, this time correlation is precisely the reason that \( \theta = x \mod 2\pi \) and \( X = J(x) \) are not Markov processes (see Chapter 4).

We now write down the stochastic differential equations satisfied by

\[
a_{nm}(t) = \frac{1}{\pi} \mathbb{E}[x_n(t)\sin m(\omega_c t + x(t))|z(s), 0 \leq s \leq t]
\]

(5.56)

\[
b_{nm}(t) = \frac{1}{\pi} \mathbb{E}[x_n(t)\cos m(\omega_c t + x(t))|z(s), 0 \leq s \leq t]
\]

(5.57)

The motivation for computing these particular expectations is as follows:
as mentioned above, the quantity $\mathcal{E}[x(t)\mid z(s), \ 0 \leq s \leq t]$ is to be used as our estimate of $x(t)$; if we then compute the stochastic differential equation satisfied by $\mathcal{E}[x(t)\mid z(s), \ 0 \leq s \leq t]$, we find that it involves $\mathcal{E}[x(t)\sin(\omega_c t+x(t))\mid z(s), \ 0 \leq s \leq t]$, and, as equations (5.58) and (5.59) indicate, the stochastic differential equation for this last quantity contains other of the $a_{nm}$ and $b_{nm}$.

In fact, we can show that the exact computation of $\mathcal{E}[x(t)\mid z(s), \ 0 \leq s \leq t]$ requires knowledge of all of the $a_{nm}$'s and $b_{nm}$'s and only these quantities.

Again using Kushner's result [K18], we find that

$$
\begin{align*}
da_{nm}(t) &= \{a(t)[na_{nm}(t)+mb_{n+1,m}(t)]+\omega_c b_{nm}(t) \\
&\quad \quad + \frac{\kappa(t)}{2} [n(n-1)a_{n-2,m}(t)+2nma_{n-1,m}(t)-m^2 a_{nm}(t)]\}dt \\
&\quad \quad + \left[ \frac{1}{2} (b_{n,m-1}(t)-b_{n,m+1}(t))-\pi a_{nm}(t)a_{01}(t) \right] \frac{(dz(t)-\pi a_{01}(t)dt)}{r(t)}
\end{align*}
$$

(5.58)

$$
\begin{align*}
\db_{nm}(t) &= \{a(t)[nb_{nm}(t)-ma_{n+1,m}(t)]-\omega_c a_{nm}(t) \\
&\quad \quad + \frac{\kappa(t)}{2} [n(n-1)b_{n-2,m}(t)-2nma_{n-1,m}(t)-m^2 b_{nm}(t)]\}dt \\
&\quad \quad + \left[ \frac{1}{2} (a_{n,m+1}(t)-a_{n,m-1}(t))-\pi b_{nm}(t)a_{01}(t) \right] \frac{(dz(t)-\pi a_{01}(t)dt)}{r(t)}
\end{align*}
$$

(5.59)

where

$$
\begin{align*}
a_{nk}(t) &\triangleq 0 \quad k \leq 0 \\
b_{nk}(t) &\triangleq 0 \quad k < 0
\end{align*}
$$

(5.60)

(5.61)
Note that
\[ b_{n0}(t) = \frac{1}{\pi} \delta[x^n(t) | z(s), \ 0 \leq s \leq t] \]  
(5.62)

and thus the optimal estimate is just \( \pi b_{10}(t) \).

We make a few comments about this filter, which is illustrated in Figure 5.7. The demodulator consists of a doubly infinite bank of filters. The \( (a_{nm}, b_{nm}) \) filter is directly connected only to the \( (a_{01}, b_{01}), (a_{n+1,m}, b_{n+1,m}), (a_{n-1,m}, b_{n-1,m}), (a_{n-2,m}, b_{n-2,m}), (a_{n,m-1}, b_{n,m-1}), \) and \( (a_{n,m+1}, b_{n,m+1}) \) filters — i.e. referring to Figure 5.7, we get vertical connections to the filters directly above and below, and horizontal connections to the filters directly to the left and right and to the filter two to the left. Referring to (5.58) and (5.59), we see that the \( (a_{nm}, b_{nm}) \) filter for \( m > 0 \) looks something like a PLL with a VCO operating at the frequency \( m\omega_c \) (note the multiplicative terms in (5.58) and (5.59), the \( m\omega_c b_{nm} \) term in (5.58) and the \( -m\omega_c a_{nm} \) term in (5.59)). The \( m = 0 \) filters are much simpler in form (see (5.59) with \( m = 0 \), and use (5.60) and (5.61)).

The reader is referred to Appendix D, in which we discuss a number of systematic methods for truncating this doubly infinite bank of filters. We also note that in this same framework, one can consider phase modulating more complicated signals and the filtering out of phase drift noise or a random initial phase (i.e. the acquisition problem), in addition to additive channel noise. The next example, in which we consider an FM problem, illustrates the versatility of this conceptual approach.

**Example 5.8:** Consider frequency modulating the signal process \( x(t) \)
**Figure 5.7**: Illustrating the Form of the Optimal Phase Demodulator of Example 5.7
satisfying (5.53). Our observation process (in Itô form) is

\[ dz(t) = \sin(\omega_c t + c) \int_0^t x(s) ds + \int_0^t e^{1/2(s)} df(s) dt + \tau^{1/2}(t) dw(t) \]  

(5.63)

where \( f \) and \( w \) are independent standard one-dimensional Brownian motion processes, both independent of \( x \). The term \( \int_0^t e^{1/2(s)} df(s) \) represents random phase drift and the \( dw(t) \) term is additive channel noise. We note that \( x(t) \) can be considered to be an error in our knowledge of \( \omega_c \) (e.g. \( dx \equiv 0 \) with \( x(0) \) unknown corresponds to a constant offset in the carrier frequency or a constant Doppler shift [K21]). Let us define the two dimensional signal proce
ces

\[ y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ c \int_0^t x(s) ds + \int_0^t e^{1/2(s)} df(s) \end{bmatrix} \]  

(5.64)

We then have

\[ dy(t) = A(t) y(t) dt + Q^{1/2}(t) du(t) \]  

(5.65)

where

\[ u(t) = \begin{bmatrix} v(t) \\ f(t) \end{bmatrix} \quad A(t) = \begin{bmatrix} a(t) & 0 \\ c & 0 \end{bmatrix} \quad Q^{1/2}(t) = \begin{bmatrix} q^{1/2}(t) & 0 \\ 0 & e^{1/2}(t) \end{bmatrix} \]  

(5.66)

Also

\[ dz(t) = \sin(\omega_c t + h'y(t)) dt + \tau^{1/2}(t) dw(t) \]  

(5.67)

\[ h' = [0, 1] \]  

(5.68)

As in the last example, suppose we want the minimum variance estimate of \( x(t) \) -- i.e. we wish to compute \( \mathbb{E}[x(t) | z(s), 0 \leq s \leq t]\).
To do this, we write the stochastic differential equations for

\[
a_{nm}(t) = \frac{1}{\pi} \mathbb{E}[y_1^n(t) \sin m(\omega_c t + \gamma_2(t))] z(s), \quad 0 \leq s \leq t \]

\[
b_{nm}(t) = \frac{1}{\pi} \mathbb{E}[y_1^n(t) \cos m(\omega_c t + \gamma_2(t))] z(s), \quad 0 \leq s \leq t \]

We find that our optimal estimate is \( \pi b_{10}(t) \), and the filter equations are

\[
da_{nm}(t) = (n a(t) a_{nm}(t) + m c b_{n+1,m}(t) + m \omega c b_{nm}(t) \\
+ \frac{1}{2} [n(n-1)q(t) a_{n-2,m}(t) - m^2 e(t) a_{nm}(t)] \) dt \\
\]

\[
(\frac{dz(t) - \pi a_{01}(t) dt}{r(t)}) \\
+ [ \frac{1}{2} (b_{n,m-1}(t) - b_{n,m+1}(t)) - \pi a_{nm}(t) a_{01}(t)] \\
\]

\[
db_{nm}(t) = (n a(t) b_{nm}(t) - m c b_{n+1,m}(t) - m \omega c b_{nm}(t) \\
+ \frac{1}{2} [n(n-1)q(t) b_{n-2,m}(t) - m^2 e(t) b_{nm}(t)] \) dt \\
\]

\[
+ [ \frac{1}{2} (a_{n,m+1}(t) - a_{n,m-1}(t)) - \pi b_{nm}(t) a_{01}(t)] \\
\frac{(dz(t) - \pi a_{01}(t) dt)}{r(t)} \\
\]

Comparing these equations to those derived in the two preceding examples we see that we can make the same sort of comments here that have been made previously. We also note that we can consider the general phase problem

\[
dy(t) = A(t) y(t) dt + Q^{1/2}(t) dv(t) \]

\[
dz(t) = \sin(\omega_c t + h' y(t)) + r^{1/2} dw(t) \]

where \( y \) is an \( n \) vector, \( v \) is an \( m \) vector, and the functions \( v_1(t), \ldots, v_m(t), w(t) \)
are independent standard Brownian motion processes all independent of y(0). Given (5.74) we can consider finding the minimum variance estimate (i.e. the conditional mean) of C\(y(t)\) given \(\{z(s) | 0 \leq s \leq t\}\) (note that if we allow y(0) to be random, we can consider a random initial phase problem). Examples 5.6, 5.7, and 5.8 indicate that this model includes the phase tracking, phase demodulation, and frequency demodulation problems in the presence of both additive channel noise and phase drift noise (in the phase tracking problem our estimation criterion may be something other than the minimum variance estimate -- e.g., \(1 - \cos(\hat{\theta} - \theta)\), which is used in Example 5.6).

We also note that the general problem, (5.73), (5.74), is conceptually no more difficult than the problems considered in the examples and, as in these cases, yields a time-invariant solution if \(A, Q, \) and \(r\) are constant. However, the bookkeeping becomes more complicated. Some straightforward calculations indicate that in general we must compute
\[
E[y_1 (t)y_2 (t) \ldots y_n (t) \sin m(\omega_c t + h'y(t))] | z(s), \; 0 \leq s \leq t \] and
\[
E[y_1 (t) \ldots y_n (t) \cos m(\omega_c t + h'y(t))] | z(s), \; 0 \leq s \leq t \] for all nonnegative integers \(k_1, \ldots, k_n, \) and \(m.\) Thus the general demodulator-tracker is a multidimensional version of the demodulators considered in the examples. Finally, in solving a particular multidimensional problem of the form of (5.73) and (5.74), one may find that the equations become more manageable if we adopt some notation from [B4]. We note that there are
\[
N(n, p) = \binom{n+p-1}{p} = \frac{(n+p-1)!}{p!(n-1)!} \tag{5.75}
\]
different products \(y_1 y_2 \ldots y_n\) with \(\sum_{j=1}^{n} k_j = p.\) For \(n \geq 1\) we can
define the $N(n,p)$-vector

$$y[p] = \begin{bmatrix} y_1^p \\ y_1^{p-1} y_2 \\ \vdots \\ y_n^p \end{bmatrix}$$

(5.76)

and, if we adopt the convention $y[0] = 1$, our filter must, in general, compute

$$a_{nm}(t) = \mathbb{E}[y[n](t) \sin m(\omega_c t + y(t))|z(s), \ 0 \leq s \leq t]$$

(5.77)

$$b_{nm}(t) = \mathbb{E}[y[n](t) \cos m(\omega_c t + y(t))|z(s), \ 0 \leq s \leq t]$$

(5.78)

The stochastic differential equations satisfied by these functions can be derived as in the examples.

As a final indication of how these techniques can be used, we consider a phase tracking problem in which the amplitude of the sinusoidal signal is unknown.

**Example 5.9:** Suppose we receive the signal

$$dz(t) = A \sin \theta(t) dt + r_1^{1/2}(t) dw(t)$$

(5.79)

where

$$\theta(t) = \omega_c t + \int_0^t \eta^{1/2}(s) dv(s) + \theta_0$$

(5.80)

We assume that the signal amplitude is constant but unknown with a priori probability distribution $p_A(\alpha)$. Also, we let $v$ and $w$ be independent Brownian motions which are both independent of $A$. We wish to devise a technique for tracking $\theta(t)$. We will use the criterion
\[ \min \mathcal{E}[1 - \cos(\theta(t) - \hat{\theta}(t)) | z(s), s \leq t]. \]
Adapting the techniques used in the preceding examples, we will write down the differential equations satisfied by
\[ a_{nm} = \frac{1}{\pi} \mathcal{E}(A^n \sin m\theta(t) | z(s), s \leq t) \] (5.81)
\[ b_{nm} = \frac{1}{\pi} \mathcal{E}(A^n \cos m\theta(t) | z(s), s \leq t) \] (5.82)

One can show that
\[
d a_{nm}(t) = [mw_c b_{nm}(t) - \frac{m^2 q(t)}{2} a_{nm}(t)] dt \\
+ [\frac{1}{2} (b_{n+1,m-1}(t) - b_{n+1,m+1}(t)) - \pi a_{nm}(t) a_{11}(t)] \frac{(dz(t) - \pi a_{11}(t) dt)}{r(t)}
\] (5.83)
\[
d b_{nm}(t) = [-mw_c a_{nm}(t) + \frac{m^2 q(t)}{2} b_{nm}(t)] dt \\
+ [\frac{1}{2} (a_{n+1,m+1}(t) - a_{n+1,m-1}(t)) - \pi b_{nm}(t) a_{11}(t)] \frac{(dz(t) - \pi a_{11}(t) dt)}{r(t)}
\] (5.84)

and our estimate is
\[ \hat{\theta}(t) = \tan^{-1} \left[ \frac{a_{01}(t)}{b_{01}(t)} \right] \] (5.85)

We note that in acquiring a sinusoidal signal such as in (5.79), we often do not know the signal power -- i.e. A can be taken as a random variable. Thus, we might use a filter of the type described here to determine A (\( \pi b_{10} \)). Once we have "determined" A -- i.e. reduced the variance in our estimate of A to an acceptable level -- we can simplify the tracking filter. That is, we can assume A = \( \pi b_{10} \), and the problem reduces to the type considered in Example 5.6. In relation to (5.83)
and (5.84), we then need only compute \( a_{0m} \) and \( b_{0m} \). We note that
the only \( a_{nm} \) and \( b_{nm} \) with \( n \geq 1 \) that directly enter the equations for \( da_{0m} \) and \( db_{0m} \) are \( a_{1m} \) and \( b_{1m} \), and we can approximate these by

\[
a_{1m} = \pi b_{10} a_{0m} \quad b_{1m} = \pi b_{10} b_{0m}
\]  \hspace{1cm} (5.86)

We can also consider the problem in which \( A \) is time-varying.
One often-used model for such a problem is the tracking of a sinusoidal signal that is transmitted through a Rayleigh channel [V2], [V5]. The received signal for such a channel is

\[
dz(t) = A(t) \sin (\theta(t) + \phi(t)) dt + r^{1/2}(t) dw(t)
\]  \hspace{1cm} (5.87)

where \( A(t) \) has a Rayleigh density [V5], [D8], and \( \phi(t) \) has a uniform density on \([-\pi, \pi]\) and is independent of \( A \). One can show, [V2], [D8],
that \( dz \) can also be written as

\[
dz(t) = x_1(t) \sin \theta(t) dt + x_2(t) \cos \theta(t) dt + r^{1/2}(t) dw(t)
\]  \hspace{1cm} (5.88)

where \( x_1 \) and \( x_2 \) are zero mean, independent, identically distributed Gaussian random processes. Assuming \( x_1 \) and \( x_2 \) satisfy a particular linear Ito differential equation, we can use the techniques developed in this section to track \( \theta(t) \). In this case, the quantities to be estimated are \( \frac{1}{\pi} x_1 x_2^m \sin n\theta \) and \( \frac{1}{\pi} x_1 x_2^m \cos n\theta \).
CHAPTER 6

SOME APPLICATIONS AND NUMERICAL RESULTS

"No, a thousand times no; there does not exist a category of science to which one can give the name applied science. There are science and the applications of science, bound together as the fruit to the tree which bears it."

- Louis Pasteur [P8]

6.1 Introduction

In the preceding chapters we have considered a number of nonlinear estimation problems and have devised a number of optimal and suboptimal estimation techniques. Many of the problems that have been formulated have important practical applications, and in Section 6.2 we will discuss a number of these applications. Also, some simulations comparing standard phase-lock loop estimation techniques with the techniques developed in this research have been carried out, and these results are presented in Sections 6.3 and 6.4.

6.2 Some Practical Problems Involving Single-Degree-of-Freedom Rotation

There is a wide variety of physically important problems that involve sinusoidally time-varying signals or rotations about a fixed axis. In this section we will describe a number of these problems, and then in the following sections, motivated by these possible applications, we will present some simulation results for two specific problems.

The most important area of application is the phase tracking and demodulation problem (see Chapters 4 and 5). A typical model for the
received signal is

\[ \dot{z}(t) = A(t)\cos(\omega_c t + \phi(t) + v(t)) + \dot{\omega}(t) \]  \hspace{1cm} (6.1)

where \( A \) is the signal amplitude (possibly time-varying and random), \( \omega_c \) is the carrier frequency, \( \phi \) is the modulated information, and \( v \) is a noise process, as is \( \dot{\omega} \). In our model we assume that \( \omega_c \) is a known nominal frequency, and we lump all uncertainties -- unknown Doppler shifts, oscillator instabilities -- into \( v \) (or \( \phi \) if, for instance, the Doppler shift is to be determined). There are a great many physical problems that involve tracking or demodulating signals having forms similar to (6.1). We will mention a number of these and will also discuss a number of noise sources.

There are a number of ship and airplane navigation systems that are based on phase (time) and frequency comparisons [K21], [P4], [H18], [E2], [R9], [V7], [V8], [R10], [S23], [P6]. One of the most important navigational aids is Doppler Radar [K21], [E2], [H18]. The basic idea behind this system is the Doppler effect -- the frequency (of an electromagnetic wave) observed by a receiver is shifted an amount approximately proportional to the relative velocity between transmitter and receiver. Thus, if a signal is transmitted at the frequency \( \omega_c \), the received signal would be of the form (excluding all noises)

\[ \dot{z}(t) = \sin(\omega_c t + \Delta \omega_D t) \]  \hspace{1cm} (6.2)

or, if we allowed time-varying velocities

\[ \dot{z}(t) = \sin(\omega_c t + \int_0^t x(s)ds) \]  \hspace{1cm} (6.3)
In this case, the problem is to track the received phase and recover (demodulate) $\Delta \omega_D$ or $x(t)$. We note that a standard technique [L16], [E3] consists of phase tracking using a phase-lock loop (PLL) and a clock that counts cycles and transforms this into frequency-velocity information. Often these clock-counters either neglect partial cycles or only interpolate to some nearest fraction of a cycle. The techniques developed in Chapters 4 and 5 and discussed again in Section 6.3 and 6.4 allow for continuous frequency determination (i.e. no actual "counting" is going on).

Another navigational aid is the hyperbolic position-determining system Loran C [P4], [H18], [R9], [V7], [V8], [R10], [S23]. In this system the receiver (airplane) determines its position as the intersection of two hyperbolas each specified by the time difference between the reception of a pair of pulses. The basic concept behind Loran C is illustrated in Figure 6.1. The master station $M$ sends out a pulsed signal at some known

![Figure 6.1: The Loran C Navigation System](image-url)
frequency \( \omega_c \) -- i.e. the transmitted signal is (excluding oscillator instabilities)

\[
\dot{z}_1 = A(t) \sin \omega_c t \quad t \in [0, \Delta]
\] (6.4)

where \( A(t) \) increases from 0 at \( t = 0 \) to a maximum value and then drops to 0 at \( t = \Delta \). The slave station \( S \) waits a specified time (the coding delay) and then transmits a similar pulse

\[
\dot{z}_2 = A(t-\tau) \sin \omega_c (t-\tau) \quad t \in [\tau, \tau+\Delta]
\] (6.5)

where \( \tau \) is the specified delay (note that the motivation for calling \( S \) a slave is that in some systems the slave waits to receive the master's signal before transmitting). The receiver's problem in this case is to acquire and track both pulses, measure the time delay between the signals and use this to determine a hyperbolic line of position (LOP -- see Figure 6.1). Thus, the accuracy of position determination is quite dependent upon the accuracy of the signal acquisition and tracking systems.

A third navigation system is the very low frequency (VLF) hyperbolic navigation system Omega [P6], [K21], [V7]. In this system eight VLF stations around the world transmit for approximately one second at exactly the same frequency. The receiver (which can be used by ships, aircraft, and submarines at moderate depths) must then acquire two pairs of signals and use a phase comparison to determine its position at the intersection of the two resulting hyperbolic lines of position. The basic signal acquisition and tracking problem is similar to that for the Loran C system, and we will not discuss Omega in any more detail. The reader is referred to [P6] for a complete system description.
A related signal processing problem occurs in radar or sonar detection systems. In these problems a radar or sonar wave is transmitted, and if a target is present, the signal that is returned is of the form

$$\hat{z}(t) = A(t)\sin\omega_c t + \hat{w}(t) \tag{6.6}$$

where $\hat{w}$ is a white noise process. The problem is to devise a detection system to determine whether or not the target -- i.e. the $A(t)\sin\omega_c t$ term -- is present. The detection problem is closely associated with the estimation problem, which is what we are concentrating on in this chapter. Extensions of our techniques to the detection problem can be made, and the reader is referred to [V6] for a detailed description of the radar-sonar problem formulation and standard solution techniques.

Another important class of systems are angle modulation systems [V1], [V2] -- both frequency modulation, FM, and phase modulation, PM. Let $x(t)$ be the information to be transmitted. Then the noise-free FM signal is

$$\hat{z}(t) = A \sin(\omega_c t + c \int_0^t x(s)ds) \tag{6.7}$$

and the analogous PM signal is given by

$$\hat{z}(t) = A \sin(\omega_c t + cx(t)) \tag{6.8}$$

where $c$ is a known positive constant. If we include a phase variation noise, these problems can be analyzed using the techniques of Chapter 4 (see Section 6.3), while the additive noise problem can be studied as in Chapter 5 (see Examples 5.7 and 5.8).

Several interesting extensions of these ideas are motivated by the
generalization of the $S^1$ results of Chapter 4 to a general abelian
Lie group problem. Clearly we can consider a multichannel angle
modulation system as a problem on $(S^1)^n$ and can either use the Chapter 4
results or multidimensional Fourier series analogs of the techniques
of Chapter 5. In addition, we can consider certain AM (amplitude modu-
lation) and joint AM-FM or AM-PM problems [M6] by examining problems
defined on the abelian Lie group $R^1 \times S^1$.

One can show that $R^1 \times S^1$ is isomorphic (as a Lie group) to
$C-\{0\}$ -- the set of nonzero complex numbers with complex multiplica-
tion as the group operation. The associated Lie algebra can be identified
with $R^2$, and we have the map $\exp: R^2 \rightarrow C-\{0\}$, relating the Lie algebra
and group, given by

$$\exp(x_1, x_2) = e^{x_1 + ix_2}$$  \hspace{1cm} (6.9)

Note that an isomorphism between $C-\{0\}$ and $R^1 \times S^1$ is given by

$$(r, \theta) \mapsto \exp(r, \theta) \quad r \in R^1 \quad \theta \in [-\pi, \pi)$$  \hspace{1cm} (6.10)

Thus $S^1$ is the subgroup of $C-\{0\}$ consisting of all complex numbers of
length one, and its Lie algebra is the subalgebra of $R^2$ obtained by
setting $x_1 = 0$. Also we see from (6.9) that $x_1$ controls the amplitude
of $\exp(x_1, x_2)$, while $x_2$ controls the phase.

Now suppose that we have a continuous $R^2$ signal process
$x'(t) = (x_1(t), x_2(t))$. Define the measurement process $z(t)$ as follows:

$$dy(t) = x(t)dt + dv(t) \quad y(0) = 0$$  \hspace{1cm} (6.11)

$$z(t) = \exp(y_1(t), y_2(t))$$  \hspace{1cm} (6.12)
where \( \nu \) is a 2-dimensional Brownian motion process independent of \( \nu \).

This problem clearly fits into the framework discussed in Sections 4.2 and 4.3 -- i.e. knowledge of \( z(s), s \leq t \) is equivalent to knowledge of \( y(s), s \leq t \). In fact, we can express \( dy(t) \) in terms of \( z(t) \) and \( dz(t) \) with the aid of the Ito differential rule.

\[
dy_1(t) = \Re \left[ \frac{dz(t)}{z(t)} \right] dt - \frac{q_{11}(t)q_{22}(t)}{2} dt \quad (6.13)
\]

\[
dy_2(t) = \Im \left[ \frac{dz(t)}{z(t)} \right] dt - q_{12}(t) dt \quad (6.14)
\]

where

\[
E(d\nu(t)d\nu'(t)) = Q(t) dt = \begin{bmatrix} q_{11}(t) & q_{12}(t) \\ q_{12}(t) & q_{22}(t) \end{bmatrix} dt \quad (6.15)
\]

Note that

\[
z(t) = [e^{\int_0^t \nu_1(s) + i \nu_2(s) ds}] [e^{i \int_0^t \nu_1(s) + i \nu_2(s) ds}] \quad (6.16)
\]

Thus \( z \) is both amplitude and angle modulated and the noise is a multiplicative lognormal process [H9], [Al]. Thus (6.16) yields a message model for a joint AM-FM modulation system for which there is a simple optimal estimator. Note that if we set \( \nu_2 = v_2 = 0 \), we have a pure AM problem. Actually, \( x_1(t) \) is more like an amplitude rate modulating signal. However, if we let \( x_1(t) = \frac{d}{dt} \tilde{x}_1(t) \), where \( \tilde{x}_1 \) is the actual signal to be transmitted, we have that the amplitude modulation is

\[
\int_0^t x_1(s) ds \quad \tilde{x}_1(t) = e \quad (6.17)
\]
or, if we let

\[ x_1 = \frac{d}{dt} \tilde{x}_1 \]

then (assuming \( \tilde{x}_1 > 0 \))

\[ \int_0^t x_1(s)ds = \frac{\tilde{x}_1(t)}{\tilde{x}_1(0)} \]  

(6.19)

Also note that in the AM case we can include \( v_2 \) as a random phase, and in the FM case (\( x_1 = C \)) we can include \( v_1 \) as a random amplitude.

We note that the multiplicative lognormal noise process in (6.16) is an important model in some optical communication problems [H9].

In many cases, changes in the transmission medium -- e.g. turbulence in the atmosphere -- cause variations in the refractive index of the air. This disturbance can be modeled [H9] as a multiplicative lognormal noise process. In this case the preceding analysis may prove to be helpful in the design of good receivers. In particular, these results may prove useful in the case of spatially uniform noise, and, in addition, we can treat the problem with real and imaginary parts of the noise process dependent on each other (\( q_{12}(t) \neq 0; \) see [H9]). The reader is referred to [A1] for other problems involving the lognormal distribution.

It is appropriate to comment on the noise sources and models associated with the problems we have discussed. The additive noise process in phase tracking and demodulation problems -- usually modeled as a Gaussian, and perhaps white, process -- is used to account for any interfering signals
that reach or originate in the receiver. The reader is referred to \[V1\] and \[V2\] for further justification of this model. For many applications the additive noise is sufficiently wideband to justify the assumption that it is white. There are some special cases, however, in which the noise characteristics are known, and this information should be used to improve system performance (although we will not consider this problem here). An interesting example of this occurs in Loran C \[V8\], \[S23\] in which a second signal, a skywave — the reflection of the master or slave signal off of the ionosphere — interferes with the original signal.

The other type of noise of interest to us is the type that appears as a frequency or phase shift. We first note that in the standard FM system involving a limiter—discriminator, additive channel noise is processed in such a manner as to yield frequency and phase deviations directly (see \[W9\], \[C11\]). Also, refractive index changes can affect electromagnetic signals as well as optical ones (see \[F8\] and \[W11\]). For instance, there are marked attenuation and phase-shifting effects in deep space communication and radar astronomy applications when the transmitted signal passes through the solar corona \[E4\]. Another source of phase—frequency errors is an error in knowledge of the carrier frequency or a Doppler shift that must be accounted for. Such a problem clearly arises in the Loran C or Omega systems if the receiver, aboard an aircraft, is moving relative to the transmitting stations. A final cause of phase drifting is an instability in the oscillator or clock being used. Edson \[E5\] and Develet \[D11\] have modeled such instabilities
as Brownian motion processes, and such a model clearly fits into the framework of Chapter 4 (see also [S18]).

The problem of frequency stability, standards, and measurement of frequency drifts [A13], [B22], [C10], [S19] is another area of application of the results of Chapters 4 and 5. One model for this problem is

\[ \ddot{z}(t) = \sin(\omega_c t + \int_0^t x(s) ds) + \dot{\omega}(t) \]  

(6.20)

where \( x(t) \) is the instantaneous deviation of the actual frequency from the nominal (note \( x \equiv \) constant can be interpreted as a Doppler shift). In this case, the techniques of Chapter 5 can be used to estimate \( x \).

A second model is the following: suppose we have the oscillator signal

\[ s(t) = e^{i(\omega_c t + \int_0^t x(s) ds)} \]  

(6.21)

One method [C10] for determining \( x \) involves the multiplication of \( s(t) \) by the output of a second oscillator. That is, if the signal from the second oscillator is

\[ r(t) = e^{i(\omega_c t - v(t))} \]  

(6.22)

where \( v \) is a random phase drift, our measurement (after filtering out the \( 2\omega_c \) term) essentially is

\[ m(t) = e^{i[\int_0^t x(s) ds + v(t)]} \]  

(6.23)

If we assume \( x \) is a linear diffusion process and \( v \) is Brownian motion, (6.23) is a special case of (6.12) with \( \gamma_1 = 0 \) and can be handled as discussed previously.
A final area of application is the estimation of the angular position of a body spinning about a given axis. If we consider the single-degree-of-freedom integrating gyroscope [W2], we note that the output of this device is an angle -- essentially the shift in orientation of the body from some reference position. The orientation of the gyro is determined by the integral of the angular velocity. Noise in the system is modeled as gyro drift, analogous to the phase drift in an oscillator. The estimation results of Chapter 4 can be used to design a system to estimate the angular velocity. The reader is referred to the next chapter in which we discuss some extensions of the techniques of the preceding chapters to the study of 3-dimensional rotation.

Before discussing the simulations that have been performed, it is appropriate to comment on the phase-lock loop, which is the system whose performance we compare to that of systems designed using the techniques described in the previous chapters. The basic phase-lock loop model [V1], [V2] is illustrated in Figure 6.2. The received signal is of the form

\[ s(t) = \sqrt{2}P \sin(\omega_c t + \theta(t)) + \dot{\omega}(t) \]  \hspace{1cm} (6.24)

where \( \theta(t) \) is usually taken to be some type of linear diffusion process and \( \dot{\omega} \) is a white noise. The part of the loop in Figure 6.2 below the dashed line essentially performs the function of tracking the time-varying phase (which is all that is needed for some applications, such as Doppler Radar, Loran C, and Omega), while the filtering above the dashed line performs the desired demodulation -- e.g. if \( \theta(t) = c x(t) \) or \( \theta(t) = c \int_0^t x(s) ds \).
Figure 6.2: The Basic PLL Model

Figure 6.3: The Baseband PLL Model
Following Van Tress [V2], we can write the product of $s(t)$ and the output of the voltage controlled oscillator (VCO) as

$$\sqrt{2}s(t)\cos(\omega_c t + \hat{\theta}(t)) = \sqrt{2} \dot{\theta}(t)\cos(\omega_c t + \hat{\theta}(t))$$

$$+ \sqrt{P} \sin(\theta(t) - \hat{\theta}(t)) + \sqrt{P} \sin(2\omega_c t + \theta(t) + \hat{\theta}(t))$$

(6.25)

The reader is referred to [V2], in which it is argued that $n(t) = \sqrt{2} \dot{\theta}(t)\cos(\omega_c t + \hat{\theta}(t))$ is essentially a white noise process of strength equal to that of $\dot{\theta}$. Then, if we assume that the phase tracking linear filter has bandwidth much smaller than $2\omega_c$, we can ignore the double frequency term in (6.25). In this case, we obtain the baseband model of the PLL depicted in Figure 6.3.

The simplest PLL system is the first order loop, in which the phase tracking linear filter is taken to be a constant gain. More complicated loops can be obtained by using standard Kalman filtering or Wiener-Hopf techniques [V5]. In this case, the assumption that the loop is "below threshold" [V2] — i.e. that the approximation

$$\sin(\theta(t) - \hat{\theta}(t)) = \theta(t) - \hat{\theta}(t)$$

(6.26)

is valid — is used to linearize the PLL model in Figure 6.3. Then, standard techniques can be used to determine the optimum linear filter (given the statistical properties of $\theta(t)$).

We close this section by noting a similarity between the PLL and the Fourier coefficient estimation loops of Chapter 5. We found that, for the various tracking and demodulation problems, the Fourier coefficient systems consisted of infinite banks of filters. Thus, for any practical implementation, some method for truncating the bank of filters is needed.
From Figure 6.2, we see that the output of the VCO resembles an estimate of the first Fourier coefficient, and therefore we can regard the PLL as a "lowest order truncation" of the Fourier series estimator (as mentioned in Chapter 5, the truncation is performed by filtering out the double frequency term in (6.25)). See Section 6.4, in which we compare the PLL with a different lowest order truncation system that more closely resembles the systems of Chapter 5.

6.3 A Frequency Demodulation Problem in the Presence of Oscillator Phase Drift

In this section we will present the results of a series of simulations involving a system designed using the techniques developed in Section 4.2 and 4.3. Before discussing these results, it is appropriate to make several comments regarding this type of problem and the implementation of solution methods we have developed. This can best be done via an example.

Example 6.1: Consider the problem discussed in Example 4.1. Let \( v \) and \( w \) be independent standard Brownian motion processes. The two-dimensional state \( x \) satisfies the stochastic equation

\[
dx(t) = Ax(t)dt + Bdv(t) \quad x(0) = 0
\]

(6.27)

with

\[
A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

(6.28)

Suppose we observe

\[
z_1(t) = \cos(\int_0^t x_2(\tau)d\tau + w(t))
\]

(6.29)

\[
z_2(t) = \sin(\int_0^t x_2(\tau)d\tau + w(t))
\]
and wish to estimate $X_1 = J(x_1)$ (orientation estimation) or $x$ (demodulation).

In Example 4.1 we derived optimal estimation equations that directly involved the differential observations $dz_1$ and $dz_2$. Although differentiations such as these are in fact performed by limiter-discriminators in some FM systems, it is desirable to consider other estimation systems that avoid these explicit differentiations. To do this, we define

$$x_3(t) = \int_0^t x_2(\tau) d\tau + w(t) = x_1(t) + w(t) \quad (6.30)$$

and note that

$$x_3(t) = \int_0^t [z_1(\tau)dz_2(\tau)-z_2(\tau)dz_1(\tau)] \quad (6.31)$$

Oppenheim, Schafer and Stockham [01] have devised methods for the practical implementation of (6.31) and have reported on the application of these techniques to a number of filtering problems.

Thus, suppose we can compute $x_3(t)$. Since, as discussed in Section 4.2, knowledge of $z_1(s)$ and $z_2(s)$ for $s \leq t$ is equivalent to knowledge of $x_3(s)$, $s \leq t$, we can consider the equivalent problem of determining estimation equations for $X_1, x_1$ and $x_2$ that depend explicitly only on $x_3$. One possible way to determine such equations is to consider the augmented state

$$v = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (6.32)$$

with a perfect observation of $x_3$. In this case, the "colored noise"
estimation results of Bryson and Johansen [B12] are applicable. We note that their results can be used for time-varying systems -- e.g., we can allow $A$ and $B$ of (6.27) to be time varying and can consider the observation processes

$$z_1(t) = \cos \left( \int_0^t c(\tau)x_2(\tau)d\tau + \omega(t) \right)$$

(6.33)

$$z_2(t) = \sin \left( \int_0^t c(\tau)x_2(\tau)d\tau + \omega(t) \right)$$

(6.34)

In the time-invariant case, we can use Wiener-Hopf techniques [V5] to determine the optimal time-invariant estimator. Suppose, as in Example 4.1, we wish to estimate $X_1$ where $A$ and $B$ are given by (6.28). As discussed in Chapter 4, the optimal steady-state estimate $\hat{x}_1(t|t)$ is given by

$$\hat{x}_1(t|t) = \exp(\hat{x}_1(t|t)R)$$

(6.35)

where $\hat{x}_1(t|t)$ is the optimal steady-state estimate of $x_1(t)$. Computing the power spectral densities of $x_1$ and $x_3$, we can use the Wiener-Hopf equations to determine the filter that computes $\hat{x}_1$ from $x_3$. The optimal filtering procedure is illustrated in Figure 6.4 (compare to Figure 4.4).

We now describe the estimation problem that has been simulated. Let $v$ and $w$ be independent Brownian motions, and define $x$ to be the scalar process satisfying

$$dx(t) = ax(t)dt + b^{1/2}dv(t)$$

(6.36)

and suppose we observe

$$z_1(t) = \sin(\omega_c t + c) \int_0^t x(s)ds + q^{1/2}w(t)$$

(6.37)

where $\omega_c$ is a known carrier frequency. Suppose we wish to demodulate
Fig. 6.4 Illustrating the Optimal Filter Discussed in Example 6.1
$z_1$ to obtain an estimate for $x$. In order to apply the results of Chapter 4 directly -- or even to use the type of solution procedure suggested by (6.31) and developed on [01] -- we need to compute

$$z_2(t) = \cos(\omega_c t + c \int_0^t x(s)ds + q^{1/2}w(t))$$  \hspace{1cm} (6.38)

One possible method for this is to process $z_1$ through a filter that has unity gain and $90^\circ$ of phase shift at $\omega_c$. We note that for such a system to work accurately, we must require $\omega_c$ to be much larger than the other frequency terms, so that the effective signal frequency is approximately $\omega_c$. Otherwise, we may not obtain the desired amount of phase shift and the desired unity gain. A second procedure also requires the assumption that $\omega_c$ is sufficiently large so that the total phase

$$\phi(t) = \omega_c t + \int_0^t x(s)ds + q^{1/2}w(t)$$

increases with time. Of course, since $|x|$ and $|\dot{w}|$ can take on arbitrarily large values with nonzero probability, this can never be achieved theoretically. However, the assumption, for instance, that $\dot{w}$ is a white noise is made for theoretical convenience (in addition, it works quite well), and in any real application, we can be certain that $\dot{\phi} > 0$ (i.e. in real systems the maximum frequency deviation from $\omega_c$ is small relative to $\omega_c$). Assuming this holds, we have

$$\dot{z}_1 = \dot{\phi}z_2$$  \hspace{1cm} (6.39)

and thus we can compute

$$z_2 = (\text{sgn}(\dot{z}_1))(1-z_1^2)^{1/2}$$  \hspace{1cm} (6.40)

We also note that if we assume an initial value for $z_2$ (i.e. $\pm(1-z_1^2(0))^{1/2}$), we can compute $|z_2| = (1-z_1^2)^{1/2}$ and can determine
sgn(z₂) consistently by assuming \( \phi > 0 \). This method avoids computing sgn(\( \dot{z}_1 \)) but has the problem of introducing an ambiguity in initial conditions. We recall that, as discussed in Chapter 4, if we use the differential observations dz₁ and dz₂, initial conditions will not affect the estimation procedure. Also, if we compute

\[
y(t) = \int_0^t [z₂(\tau)dz₁(\tau) - z₁(\tau)dz₂(\tau)] = \phi(t) - \phi(0)
\]

(6.41)

and use the Wiener-Hopf filtering technique discussed in Example 6.1, the steady-state mean square estimation error is independent of the initial conditions of z₁ and z₂.

For this problem we can devise a technique that computes y(t) (modulo an initial condition) without calculating z₂(t). Assuming that \( \dot{\phi} > 0 \), the output of a cycle counter, is equal to \( \frac{\phi(t) - \phi(0)}{2\pi} \). We then have

\[
y(t) = \phi(t) - \phi(0) = 2\pi \left\lfloor \frac{\phi(t)}{2\pi} \right\rfloor + \psi(t)
\]

(6.42)

where

\[
\psi(t) = \phi(t) \mod 2\pi \quad \psi \in [0, 2\pi)
\]

(6.43)

We compute \( \psi(t) \) as follows:

\[
\psi(t) = \left\lfloor \sin^{-1}(z₁(t)) \right\rfloor \mod 2\pi \quad \text{if} \quad \text{sgn}(\dot{z}_1(t)) > 0
\]

(6.44)

\[
\psi(t) = \left\lfloor \pi - \sin^{-1}(z₁(t)) \right\rfloor \mod 2\pi \quad \text{if} \quad \text{sgn}(\dot{z}_1(t)) < 0
\]

(6.45)

where \( \sin^{-1}[-1,1] \rightarrow [-\frac{\pi}{2}, \frac{\pi}{2}] \). This procedure is depicted in Figure 6.5. Note that the \( \psi(t) \) computation is similar to the interpolator used in Doppler shift determination, as mentioned in the preceding section.
Figure 6.5: A Procedure for Computed Total Phase

Also we remark that an alternative procedure for calculating $\psi(t)$ is the following:

1. Initially set $N = 0$
2. Compute $\theta(t) = \sin^{-1}(z_1(t))$
3. If $N = 0$, $\psi(t) = \theta(t) \mod 2\pi$
   
   If $N = 1$, $\psi(t) = [\pi - \theta(t)] \mod 2\pi$
4. When $|z_1|$ goes through the value 1, let $N = (N+1) \mod 2$

(note that initially setting $N = 0$ corresponds to assuming $z_2(0) = +(1-z_1^2(0))^{1/2}$).

Suppose we have implemented a system that produces $dy(t)$ (or $\dot{y}(t)$) given the input $z_1$. In this case, equations (4.68) - (4.71) apply directly (after we subtract out the carrier frequency), and the optimal steady-state filter equations are
\[ d\hat{z}(t|t) = a\hat{z}(t|t)dt + \frac{p_{\infty}c}{q} (dy(t) - c\hat{z}(t|t)dt - \omega_c dt) \]  
\[ p_{\infty} = \frac{aq + \sqrt{a^2q^2 + bc^2}}{c^2} \]  

Here \( p_{\infty} \) is the optimum steady-state error variance.

If we can produce \( y(t) \) as opposed to \( dy(t) \), we obtain the optimal steady-state filtering equations

\[ \dot{\hat{x}}(t) = -\alpha r(t) + y(t) - \omega_c t \]  
\[ \hat{x}(t|t) = M(\alpha r(t) + y(t) - \omega_c t) \]

where

\[ \alpha = \sqrt{a^2 + bc^2/q} \quad M = \frac{bc}{q(\alpha + a)} \]

The block diagrams of these two different filtering systems are given in Figures 6.6 and 6.7, where we have implemented (6.46) in a form that more closely resembles (6.48) and (6.49). Note that the input in Figure 6.6 is the derivative of the input in Figure 6.7 and this is compensated for by picking out the signal before the integrator in Figure 6.7. In addition, one can show that the steady state error variance for the filter of Figure 6.7 is also given by (6.47).

In addition to simulating systems of the type just discussed, we have also obtained results for a system involving a first order PLL. The system simulated is depicted in Figure 6.8. Given a particular value of the loop gain \( K \), the transfer function \( G(s) \) is determined in the manner described at the end of the previous section — i.e. we write down the baseband model (Figure 6.9), linearize, and apply Wiener-Hopf
Figure 6.6: Block Diagram of the Optimal Filter
Given Differential Phase Information

Figure 6.7: Block Diagram of the Optimal Filter
Given Total Phase Information
**Figure 6.8: Block Diagram of the PLL System**

**Figure 6.9: Baseband Model of the PLL System**
techniques. In fact, if one looks at the overall block diagram of the linearized PLL, the transfer function from \( v(t) - \omega_c t \) to \( x(t|t) \) is identical to that in Figure 6.7.

Nominal values \( a = -1 \), \( b = 1 \) were chosen for the simulation and the values of \( c \) and \( q \) were varied to test system performance. Note that \( c \) is what is called \([V2]\) the frequency deviation and the larger \( c \) is, the better we would expect our system to perform. Also \( 1/q \) is sometimes called \([V2]\) the oscillator coherence time, and the quantity

\[
\Lambda = \frac{c^2}{q}
\]  

(6.51)

plays the role of a signal to noise ratio. Also

\[
\rho_{\infty} = \frac{a + \sqrt{a^2 + b\Lambda}}{\Lambda}
\]  

(6.52)

The figure of merit used to compare various system performances was the inverse of the empirically computed sample error variance. This is computed as follows: let \( N \) be the total number of iterations and compute

\[
m_1 = \frac{1}{N} \sum_{i=1}^{N} e_i \quad m_2 = \frac{1}{N} \sum_{i=1}^{N} e_i^2
\]  

(6.53)

where

\[
e_i = x_i - \hat{x}_i
\]  

(6.54)

and the \( x_i \) and \( \hat{x}_i \) are the values of \( x \) and \( \hat{x} \) obtained in the digital simulation. Then the (unbiased) sampled variance is

\[
\sigma_e^2 = \frac{n}{n-1} (m_2 - m_1^2)
\]  

(6.55)

where \( n \) is the number of degrees of freedom.
As discussed in [G3] and [B25], n, the effective number of degrees of freedom is usually much less than N, since the $e_i$ are correlated. For instance, for the filter described by (6.46), the differential equation for $e(t) = x(t) - \hat{x}(t|t)$ is

$$d(e) = -\alpha e d(t) + b^{1/2} dv(t) - \frac{p_{\infty}}{q^{1/2}} dw(t)$$  \hspace{1cm} (6.56)

and, referring to [G3] and [B25], if $\Delta$ is the step size of the simulation, the effective number of degrees of freedom is

$$n = \frac{N \Delta}{\tau_e}$$  \hspace{1cm} (6.57)

where $\tau_e = 1/\alpha$. The values of $\tau_e$ for the various choices of $c$ and $q$ used are given in Table 6.1. We note that the effective number of degrees of freedom for the filter defined by (6.48) and (6.49) and the approximate number for the PLL (using the linearized model) are also given by (6.57).

For our simulations, we chose $n = 500$, and thus

$$N = \frac{500 \tau_e}{\Delta}$$  \hspace{1cm} (6.58)

Therefore, the smaller the step size, the larger the number of iterations. On the other hand, from [G14], we must choose $\Delta$ small enough to ensure reasonable accuracy. In order for the PLL or the total phase detector of Figure 6.5 to work, the carrier frequency must be quite large, and thus, for reasonable accuracy, $\Delta$ must be on the order of $\frac{1}{10}$ or $\frac{1}{50}$ of the period $\frac{2\pi}{\omega_c}$. The number of iterations necessary for even modest size carrier frequencies (e.g. 1000 Hz) is quite high, and thus many
Table 6.1: Values of $\tau_e$, the Error Time Constant

<table>
<thead>
<tr>
<th>$q$</th>
<th>0.01</th>
<th>0.25</th>
<th>1.0</th>
<th>25.0</th>
<th>400.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.707</td>
<td>0.981</td>
<td>0.995</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>0.25</td>
<td>0.142</td>
<td>0.581</td>
<td>0.819</td>
<td>0.990</td>
<td>0.999</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0449</td>
<td>0.242</td>
<td>0.447</td>
<td>0.928</td>
<td>0.995</td>
</tr>
<tr>
<td>25.0</td>
<td>0.0100</td>
<td>0.0499</td>
<td>0.0996</td>
<td>0.447</td>
<td>0.894</td>
</tr>
<tr>
<td>400.0</td>
<td>0.0050</td>
<td>0.0250</td>
<td>0.0499</td>
<td>0.242</td>
<td>0.707</td>
</tr>
</tbody>
</table>

Table 6.2: Analytical Values of $1/\sigma_e^2$ for Optimal System Either with Perfect Differentiator or Perfect Total Phase Detector

<table>
<thead>
<tr>
<th>$q$</th>
<th>0.01</th>
<th>0.25</th>
<th>1.0</th>
<th>25.0</th>
<th>400.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2.42</td>
<td>8.07</td>
<td>21.00</td>
<td>101.00</td>
<td>201.00</td>
</tr>
<tr>
<td>0.25</td>
<td>2.02</td>
<td>2.72</td>
<td>5.13</td>
<td>21.00</td>
<td>41.00</td>
</tr>
<tr>
<td>1.0</td>
<td>2.00</td>
<td>2.22</td>
<td>3.24</td>
<td>11.05</td>
<td>21.00</td>
</tr>
<tr>
<td>25.0</td>
<td>2.00</td>
<td>2.01</td>
<td>2.08</td>
<td>3.24</td>
<td>5.13</td>
</tr>
<tr>
<td>400.0</td>
<td>2.00</td>
<td>2.00</td>
<td>2.01</td>
<td>2.12</td>
<td>2.42</td>
</tr>
</tbody>
</table>

Table 6.3: Simulation Results ($1/\sigma_e^2$) for Optimal System, Perfect Differentiator
of the simulations had to be performed using baseband models. At
baseband, the step size $\Delta = \frac{1}{10} \tau_e$ was used.

Although the steady-state error variance for the system of Figure 6.6 — i.e. the optimal filter given a perfect differentiator — is given analytically by (6.52), this system was simulated (with $\omega_c = 0$) as a check on the validity of the simulation. The analytical results for the various values of $c$ and $q$ are given in Table 6.2, while those obtained from the simulation are given in Table 6.3. Also, a step size variation was performed and the results are contained in Table 6.4.

In addition to this simulation, the cycle counter - total phase determination system of Figure 6.5 followed by the filter of Figure 6.7 was simulated. In order to test out the total phase detector, a number of runs were made with a carrier frequency $f_c = \frac{\omega_c}{2\pi} = 10,000$ Hz and a step size of $10^{-5} = \frac{1}{10} f_c$. These results are given in Table 6.5.

Comparing these values with the corresponding analytical results, we see that the cycle-counter-demodulator system works quite well. As a further test of the simulation — and in order to provide more data to compare to that obtained for the PLL — a baseband model of this system was simulated. That is, it was assumed that the cycle counter-phase detector system worked perfectly, and consequently for simulation purposes the input to the filter of Figure 6.7 could be taken to be $c \int_0^t x(s)ds + q^{1/2} w(t)$ (where we have already subtracted off $\omega_c t$). The results of these tests are summarized in Table 6.6. It should be noted that all of these simulations were done assuming the initial total phase was zero. To show that the steady-state performance is independent of initial error, several simulations of the optimal filter with perfect
\[
\Delta \left\{ \begin{array}{c}
\tau_e/2 \\
\tau_e/5 \\
\tau_e/10 \\
\tau_e/50 \\
\tau_e/100
\end{array} \right. \\
\begin{array}{c}
232 \\
227 \\
212 \\
204 \\
199 \\
201
\end{array}
\begin{array}{c}
15.4 \\
12.9 \\
5.8 \\
1.5 \\
0.95 \\
0.0
\end{array}
\]

<table>
<thead>
<tr>
<th>(1/\sigma^2)</th>
<th>% Error in Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tau_e/2)</td>
<td>232</td>
</tr>
<tr>
<td>(\tau_e/5)</td>
<td>227</td>
</tr>
<tr>
<td>(\tau_e/10)</td>
<td>212</td>
</tr>
<tr>
<td>(\tau_e/50)</td>
<td>204</td>
</tr>
<tr>
<td>(\tau_e/100)</td>
<td>199</td>
</tr>
</tbody>
</table>

**Table 6.4:** Step Size Variation with \(c = 20\), \(q = 0.01\)

<table>
<thead>
<tr>
<th>(c)</th>
<th>0.7</th>
<th>10.0</th>
<th>100.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>(q)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>8.40</td>
<td>100.72</td>
<td>1032.0</td>
</tr>
<tr>
<td>25.0</td>
<td>1.84</td>
<td>3.36</td>
<td>20.94</td>
</tr>
</tbody>
</table>

**Table 6.5:** Simulation Results for Optimal System with Total Phase Detector (\(f_c = 10,000\) Hz.)

<table>
<thead>
<tr>
<th>(c)</th>
<th>0.1</th>
<th>0.7</th>
<th>2.0</th>
<th>10.0</th>
<th>20.0</th>
<th>100.0</th>
<th>500.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>(q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>2.05</td>
<td>8.16</td>
<td>19.32</td>
<td>101.46</td>
<td>213.43</td>
<td>1007.0</td>
<td>5015.0</td>
</tr>
<tr>
<td>0.25</td>
<td>2.05</td>
<td>2.58</td>
<td>5.16</td>
<td>21.14</td>
<td>40.47</td>
<td>230.63</td>
<td>1044.0</td>
</tr>
<tr>
<td>1.0</td>
<td>1.82</td>
<td>2.08</td>
<td>3.56</td>
<td>11.16</td>
<td>18.41</td>
<td>109.17</td>
<td>529.0</td>
</tr>
<tr>
<td>25.0</td>
<td>2.07</td>
<td>2.03</td>
<td>1.92</td>
<td>3.21</td>
<td>5.07</td>
<td>20.94</td>
<td>113.25</td>
</tr>
<tr>
<td>400.0</td>
<td>2.11</td>
<td>1.84</td>
<td>2.18</td>
<td>2.16</td>
<td>2.51</td>
<td>6.23</td>
<td>25.75</td>
</tr>
</tbody>
</table>

**Table 6.6:** Simulation Results for Optimal System with Perfect Total Phase Detector (Baseband)

<table>
<thead>
<tr>
<th>(c)</th>
<th>0.7</th>
<th>10.0</th>
<th>100.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>(q)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>8.01</td>
<td>104.36</td>
<td>972.1</td>
</tr>
<tr>
<td>25.0</td>
<td>2.03</td>
<td>3.47</td>
<td>19.43</td>
</tr>
</tbody>
</table>

**Table 6.7:** Simulation Results for Optimal System with Perfect Total Phase Detector with Nonzero Initial Condition
total phase detector were done assuming a nonzero initial phase of 1 radian -- i.e. the input to the filter was taken to be
\[ c \int_0^t x(s) ds + q^{1/2} w(t) + 1. \]
The results of these simulations are given in Table 6.7.

Finally, the baseband model of the PLL (Figure 6.9) was simulated using three values for \( K = 1, 10, \) and 100. These results are given in Tables 6.8 - 6.10. We remark that the step size criterion for the PLL simulations is somewhat different than that for the other systems. Referring to Figure 6.9, we see that, if we linearize the system (i.e. remove the sine function), the overall system is a second order linear system. Following [G14], we must choose \( \Lambda \) to be on the order of \( 1/10 \) of the smallest time constant, which, for the PLL, need not be \( T_e \) for \( K \) large enough. Note that for small values of \( c \), the PLL results are quite close to the optimal values in Table 6.2, but for large values of \( c \) the performance falls off drastically. The reason for this is that the assumption that allowed us to linearize the loop is violated for \( c \) large enough, and thus the linear analysis used to design the loop is not valid. However, note that as \( K \) increases, the range of values of \( c \) over which PLL performance is near optimal also increases. In fact, one can show that as \( K \to \infty \), PLL performance approaches the optimal. The reader is referred to Figure 6.10, in which some of the data obtained are displayed graphically.

6.4 A Phase Tracking Problem in the Presence of Additive Channel Noise

In this section we will discuss the results of a series of simulations of several different types of phase tracking systems. The tracking problem
### Table 6.8: Simulation Results for PLL with K = 1 (Baseband Model)

<table>
<thead>
<tr>
<th>q</th>
<th>0.1</th>
<th>0.7</th>
<th>2.0</th>
<th>10.0</th>
<th>20.0</th>
<th>100.0</th>
<th>500.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2.20</td>
<td>7.85</td>
<td>2.42</td>
<td>1.17</td>
<td>1.24</td>
<td>5.83</td>
<td>3.50</td>
</tr>
<tr>
<td>0.25</td>
<td>2.16</td>
<td>2.80</td>
<td>2.61</td>
<td>1.91</td>
<td>1.51</td>
<td>2.93</td>
<td>3.75</td>
</tr>
<tr>
<td>1.0</td>
<td>2.13</td>
<td>2.20</td>
<td>2.05</td>
<td>1.49</td>
<td>1.32</td>
<td>1.51</td>
<td>1.00</td>
</tr>
<tr>
<td>25.0</td>
<td>1.96</td>
<td>2.01</td>
<td>2.17</td>
<td>2.11</td>
<td>1.90</td>
<td>1.66</td>
<td>2.00</td>
</tr>
<tr>
<td>400.0</td>
<td>2.11</td>
<td>2.09</td>
<td>2.10</td>
<td>1.62</td>
<td>2.07</td>
<td>1.84</td>
<td>1.75</td>
</tr>
</tbody>
</table>

### Table 6.9: PLL Simulations with K = 10

<table>
<thead>
<tr>
<th>q</th>
<th>0.1</th>
<th>0.7</th>
<th>2.0</th>
<th>10.0</th>
<th>20.0</th>
<th>100.0</th>
<th>500.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2.38</td>
<td>8.15</td>
<td>22.16</td>
<td>57.55</td>
<td>0.72</td>
<td>5.25</td>
<td>3.75</td>
</tr>
<tr>
<td>0.25</td>
<td>2.16</td>
<td>2.62</td>
<td>4.76</td>
<td>1.64</td>
<td>1.52</td>
<td>0.96</td>
<td>2.75</td>
</tr>
<tr>
<td>1.0</td>
<td>1.97</td>
<td>2.32</td>
<td>2.93</td>
<td>2.02</td>
<td>2.38</td>
<td>1.52</td>
<td>1.00</td>
</tr>
<tr>
<td>25.0</td>
<td>2.08</td>
<td>2.07</td>
<td>2.22</td>
<td>2.05</td>
<td>1.84</td>
<td>1.68</td>
<td>2.00</td>
</tr>
<tr>
<td>400.0</td>
<td>2.01</td>
<td>1.81</td>
<td>2.05</td>
<td>1.94</td>
<td>2.07</td>
<td>1.85</td>
<td>1.75</td>
</tr>
</tbody>
</table>

### Table 6.10: PLL Simulations with K = 100

<table>
<thead>
<tr>
<th>q</th>
<th>0.1</th>
<th>0.7</th>
<th>2.0</th>
<th>10.0</th>
<th>20.0</th>
<th>100.0</th>
<th>500.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2.34</td>
<td>8.20</td>
<td>21.91</td>
<td>101.76</td>
<td>210.24</td>
<td>2.65</td>
<td>5.50</td>
</tr>
<tr>
<td>0.25</td>
<td>1.81</td>
<td>2.51</td>
<td>5.24</td>
<td>21.49</td>
<td>40.06</td>
<td>6.22</td>
<td>1.25</td>
</tr>
<tr>
<td>1.0</td>
<td>2.12</td>
<td>2.16</td>
<td>3.10</td>
<td>11.13</td>
<td>18.63</td>
<td>2.90</td>
<td>5.75</td>
</tr>
<tr>
<td>25.0</td>
<td>1.78</td>
<td>2.01</td>
<td>2.03</td>
<td>3.41</td>
<td>4.54</td>
<td>1.74</td>
<td>2.75</td>
</tr>
<tr>
<td>400.0</td>
<td>2.07</td>
<td>2.04</td>
<td>2.05</td>
<td>1.94</td>
<td>2.16</td>
<td>1.54</td>
<td>2.25</td>
</tr>
</tbody>
</table>
Figure 6.10: A Graph of Some of the FM Simulation Results
used is the one discussed in Example 5.6. We wish to track the phase $\theta(t) \mod 2\pi$ where

$$\theta(t) = \omega_c t + q^{1/2} v(t)$$  \hspace{1cm} (6.59)

($v(t)$ is a standard Brownian motion process) and we observe

$$\dot{z}(t) = \sin \theta(t) + r^{1/2} \dot{w}(t)$$  \hspace{1cm} (6.60)

($w$ is a standard Brownian motion independent of $v$). We note that as before, $1/q$ is called the oscillator coherence time.

The first tracking method we discuss is a PLL system. The reader is referred to [V2, pp. 37-41] for the development of the optimal PLL phase tracker. Referring to Figure 6.2, the optimal tracking (steady state) filter is a constant

$$k = \left( \frac{q}{r} \right)^{1/2}$$  \hspace{1cm} (6.61)

Also, the analysis in [V2] yields the result that, if the linear assumption used to aid in the PLL analysis is valid (i.e. if the system is below "threshold"), the phase error variance (in radians) is

$$p_{\theta e} = \sqrt{2rq}$$  \hspace{1cm} (6.62)

A second phase tracking system has been proposed by Gustafson and Speyer [G3]. Essentially their system is the optimal linear filter (in the sense of minimizing error variance). The reader is referred to [G3] for the development of the filter equations.

The other two systems were motivated by the Fourier series results discussed in Example 5.6. As discussed there and in Appendix D, we must consider suboptimal filtering techniques that involve a truncation of
the infinite Fourier series. The first method is the straightforward truncation procedure -- i.e. assume all coefficients $a_n$ and $b_n$ are 0 for $n > N$. We will not present any simulation results for such a method for the following reason. Several runs were made with $N = 3$; however the performance was extremely poor. An intuitive explanation for this is that the higher coefficients need not be negligible. For instance, suppose we know the phase perfectly; then the probability "density" is an impulse at the known value $n$, and the formal Fourier series expansion for this is

$$n(\theta) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} \cos n(\theta - n)$$

(6.63)

In this case, the various coefficients are of the same order. Thus, if we are tracking well (i.e. the density is nearly an impulse), the assumption that the higher coefficients are negligible is a poor one.

Thus, it was necessary to devise an alternative truncation procedure. The one adopted is discussed in Appendix D. Suppose we have computed $(a_n, b_n)_{n=1}^{N}$. We assume that the probability density is a folded normal density

$$p(\theta) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} e^{-n^2\gamma/2} [\cos n\theta \cos n\eta + \sin n\theta \sin n\eta]$$

(6.64)

Then, one can use this assumption to compute the higher coefficients in terms of $a_1$ and $b_1$. For instance, our approximations for $a_{N+1}$ and $b_{N+1}$ can be determined from the equations

$$\frac{1}{\pi} e^{-(N+1)^2\gamma/2} = \pi (N+1)^2 - 1 \left(\frac{a_1^2 + b_1^2}{2}\right)^2$$

(6.65)
\[
\cos \eta = T_n(\cos \eta) \\
(6.66)
\]

\[
\sin \eta = T_n(\cos \frac{\pi}{2n}) \cos \eta + \sin \frac{\pi}{2n} \sin \eta \\
(6.67)
\]

\[
\sin \eta = \frac{a_1}{(a_1^2 + b_1^2)^{1/2}} \quad \cos \eta = \frac{b_1}{(a_1^2 + b_1^2)^{1/2}} \\
(6.68)
\]

where \(T_n\) is the \(n\)th Tchebycheff polynomial \([H21]\). The system that has been simulated is the simplest of this type -- i.e., we only compute \(a_1\) and \(b_1\) and approximate \(a_2\) and \(b_2\) using (6.65) -- (6.68). Referring to the Fourier series equations (5.50), (5.51), our suboptimal filter equations are

\[
\dot{a}_1 = (\omega c b_1 - \frac{\pi}{2} a_1) + \frac{(\dot{z} - \pi a_1)}{r} \left[ \frac{1}{2} \frac{1}{\pi} - \pi^3 (b_1^2 - a_1^2) - \pi a_1^2 \right] \\
(6.69)
\]

\[
\dot{b}_1 = (-\omega c a_1 - \frac{\pi}{2} b_1) + \frac{(\dot{z} - \pi a_1)}{r} \left[ \pi^3 a_1 b_1 (a_1^2 + b_1^2) - \pi a_1 b_1 \right] \\
(6.70)
\]

\[
\hat{\theta} = \tan^{-1} \left[ \frac{a_1}{b_1} \right] \\
(6.71)
\]

Note that the right-hand sides of (6.69) and (6.70) are polynomials in \(a_1\) and \(b_1\) and can be computed easily.

We note that if we are tracking the phase perfectly, equations (6.65) -- (6.68) give precisely the correct values for \(a_2\) and \(b_2\). Then, since the differential equations for \(a_1\) and \(b_1\) do not explicitly depend on \(a_n\) and \(b_n\) for \(n > 2\), the finite dimensional filter (6.65) -- (6.68) performs optimally -- i.e., the \(a_1\) and \(b_1\) values are exactly the same as the values obtained from the optimal infinite dimensional filter, and thus our estimate \(\hat{\theta} = \tan^{-1} \left[ \frac{a_1}{b_1} \right] \) is the optimal one. Thus, for small noise variances one would expect the highly nonlinear filter
(6.69) - (6.71) to operate nearly optimally, where the "optimal" performance is that attained by the linearized PLL -- i.e. P_{0L} (note that above threshold P_{0L} is not actually achieved by the PLL).

Finally, we make one more remark before discussing the simulation results. The right-hand sides of (6.69) and (6.70) are highly non-linear, and this leads to two complications. The first of these concerns the existence of solutions to the equations, since the right-hand sides do not satisfy the Lipschitz conditions that are used in the standard proof of the existence of solutions to Itô differential equations [J2]. We do not prove the existence of a solution, but rather point out that the actual Fourier coefficients are bounded in magnitude by \( \frac{1}{\pi} \), so we can replace the various terms on the right-hand sides of (6.69) and (6.70) by "saturated" versions -- e.g. \( a_1^4 \) can be replaced by the function

\[
f(a_1) = \begin{cases} 
  a_1^4 & |a_1| \leq \frac{1}{\pi} \\
  \frac{1}{\pi^4} & |a_1| > \frac{1}{\pi}
\end{cases}
\]  

(6.72)

If we do this, we obtain an equation that does satisfy the necessary Lipschitz conditions. We remark that this discussion is academic, since our simulations indicate that the performance of this filter is quite good and the values of \( a_1 \) and \( b_1 \) in the simulations never exceeded \( 1/\pi \).

A second complication caused by the nonlinearities in the filter equations arises in considering the numerical integration of the filter equations. Wong and Zakai [W12], [W13] have shown that in numerically integrating stochastic differential equations driven by white noise,
one must include correction terms (which are nonzero only if the equations are nonlinear) in the equations to be integrated in order to obtain a numerical solution that faithfully approximates the solution to the original stochastic equations. Following [W12] and [W13], we obtain the following equations that have been used in the numerical simulation of (6.69) and (6.70).

\[ \dot{c}_1 = (\omega_c d_1 - \frac{q}{2} c_1) - \frac{1}{2} \left( \sigma_1 \frac{\partial \sigma_1}{\partial c_1} + \sigma_2 \frac{\partial \sigma_1}{\partial d_1} \right) + \frac{(\dot{z} - \pi c_1)}{r} \left[ \frac{1}{2} \left( \frac{1}{\pi} - \pi^3 (d_1^4 - c_1^4) \right) - \pi c_1^2 \right] \]  
\[ \dot{d}_1 = (-\omega c_1 - \frac{q}{2} d_1) - \frac{1}{2} \left( \sigma_2 \frac{\partial \sigma_2}{\partial c_1} + \sigma_2 \frac{\partial \sigma_2}{\partial d_1} \right) + \frac{(\dot{z} - \pi c_1)}{r} \left[ \pi^3 c_1 d_1 (c_1^2 + d_1^2) - \pi c_1 d_1 \right] \]  
\[ \hat{\theta} = \tan^{-1} \left[ \frac{c_1}{d_1} \right] \]

where

\[ \sigma_1 = \frac{1}{r^{1/2}} \left\{ \frac{1}{2} \left[ \frac{1}{\pi} - \pi^3 (d_1^4 - c_1^4) \right] - \pi c_1^2 \right\} \]  
\[ \sigma_2 = \frac{1}{r^{1/2}} \left[ \pi^3 c_1 d_1 (c_1^2 + d_1^2) - \pi c_1 d_1 \right] \]

We note that Wong and Zakai's results also require Lipschitz conditions on the right-hand side, but the same type of arguments as above can be used here.

As in [V2] and [G3], we use \( \Phi \) as the variable to be varied in the simulations. We will use the phase error variance and its square root, the RMS phase error, as our performance criteria. As discussed
in the previous section, one must be careful in choosing step size and number of iterations. The step size chosen was 1/50 of the basic cycle time \( 1/f_c \), where we chose \( f_c = 10,000 \) Hz. As in the preceding section, the number of iterations is closely tied in with the choice of number of degrees of freedom. The Fourier coefficient filter (FCF) (6.69), (6.70) is sufficiently nonlinear so that no estimate of degrees of freedom could be directly determined. Thus, the degree of freedom computation used was that used in [G3] for the PLL system and the Gustafson-Spever "state-dependent noise" filter (SDNF). Referring to [G3], the number of degrees of freedom for the PLL is

\[
n = \frac{N\Delta}{\tau} \tag{6.78}
\]

where

\[
\tau = \left(\frac{2\tau}{q}\right)^{1/2} \tag{6.79}
\]

and \( N = \text{number of iterations} \), \( \Delta = \text{step size} \). As before, the value \( n = 500 \) was used, and, in order to keep \( N \) from becoming excessively large, the value \( q = 400 \) was chosen. The value of \( \tau \) was allowed to vary in order to achieve the desired values for \( P_{\theta_2} \) (see (6.62)).

The PLL, SDNF, and FCF were all simulated using identical noise sequences to allow direct comparison. Table 6.11 contains the performances -- as measured by RMS Phase Error (RMSPE) -- of the various filters. Figure 6.11 graphically displays the phase error variance for the PLL and FCF, and filter performance as measured by the criterion \( \mathcal{E}(1-\cos(\theta-\tilde{\theta})) \) is reported in Table 6.12. We note that the phase error density for the PLL tracking system has been analytically determined (using
Figure 6.11: Phase Error Variance Results
<table>
<thead>
<tr>
<th>$P_\theta\theta$ (rad.$^2$)</th>
<th>Linear Predicted RMSPE</th>
<th>Predicted PLL RMSPE</th>
<th>Actual PLL RMSPE</th>
<th>SDNF RMSPE</th>
<th>FCF RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.041</td>
<td>11.6</td>
<td>11.6</td>
<td>12.0</td>
<td>11.9</td>
<td>11.6</td>
</tr>
<tr>
<td>0.130</td>
<td>20.6</td>
<td>20.6</td>
<td>20.3</td>
<td>20.1</td>
<td>19.8</td>
</tr>
<tr>
<td>0.225</td>
<td>27.1</td>
<td>29.1</td>
<td>28.8</td>
<td>28.8</td>
<td>28.1</td>
</tr>
<tr>
<td>0.4</td>
<td>36.6</td>
<td>44.0</td>
<td>43.4</td>
<td>42.0</td>
<td>41.2</td>
</tr>
<tr>
<td>0.7</td>
<td>48.2</td>
<td>62.5</td>
<td>61.2</td>
<td>57.3</td>
<td>57.0</td>
</tr>
<tr>
<td>1.0</td>
<td>57.3</td>
<td>72.5</td>
<td>71.9</td>
<td>69.1</td>
<td>68.9</td>
</tr>
<tr>
<td>1.3</td>
<td>65.1</td>
<td>78.2</td>
<td>76.2</td>
<td>73.1</td>
<td>73.0</td>
</tr>
</tbody>
</table>

**Table 6.11:** Phase Tracking Simulation Performance Summary — RMS Phase Error (Degrees)

<table>
<thead>
<tr>
<th>$P_\theta\theta$</th>
<th>PLL RMSPE</th>
<th>SDNF RMSPE</th>
<th>FCF RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.041</td>
<td>0.0217</td>
<td>0.0212</td>
<td>0.0202</td>
</tr>
<tr>
<td>0.130</td>
<td>0.0608</td>
<td>0.0601</td>
<td>0.0581</td>
</tr>
<tr>
<td>0.225</td>
<td>0.1171</td>
<td>0.1178</td>
<td>0.1120</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2447</td>
<td>0.2335</td>
<td>0.2253</td>
</tr>
<tr>
<td>0.7</td>
<td>0.4098</td>
<td>0.3728</td>
<td>0.3687</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5312</td>
<td>0.5040</td>
<td>0.5002</td>
</tr>
<tr>
<td>1.3</td>
<td>0.6158</td>
<td>0.5719</td>
<td>0.5692</td>
</tr>
</tbody>
</table>

**Table 6.12:** Phase Tracking Simulation Performance Summary — $\mathcal{E}[1-\cos(\theta-\hat{\theta})]$

<table>
<thead>
<tr>
<th>$P_\theta\theta$</th>
<th>RMSPE</th>
<th>$\mathcal{E}[1-\cos(\theta-\hat{\theta})]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.041</td>
<td>16.4</td>
<td>0.0396</td>
</tr>
<tr>
<td>0.4</td>
<td>44.9</td>
<td>0.2534</td>
</tr>
<tr>
<td>1.3</td>
<td>73.6</td>
<td>0.5774</td>
</tr>
</tbody>
</table>

**Table 6.13:** Summary of Results for the FCF Without Wong-Zakai Correction Terms
the baseband PLL model) [V2], and thus our simulation of the PLL system provides a check on the validity of the overall simulation.

We make several comments on the simulation results. First of all, we note that the FCF performed consistently better than the other systems, although the SDNF performance is quite close. One interesting point involves the below threshold performance. As mentioned earlier, the PLL is optimal -- with respect to the minimum variance criterion -- below threshold (see [V2]), while, as discussed in [G3], the SDNF is not optimal below threshold. The simulation results obtained indicate that the FCF may be optimal below threshold. The proof of this is an open question; however, an intuitive argument can be made that below threshold the phase density looks like a folded normal, and thus the approximation used in the FCF is in fact very nearly correct. If this is so, the FCF should perform optimally or very nearly optimally (we note that the FCF is designed using the criterion \( \mathcal{E}[1-\cos(\Theta-\hat{\Theta})] \), but for small phase errors and \( \Theta \) a folded normal random variable, \( \hat{\Theta} \) for the FCF will also minimize \( \mathcal{E}[(\Theta-\hat{\Theta})^2] \). A related idea is that as \( P_{\alpha\beta} \) increases, the phase density looks less and less like a normal density, and thus the FCF approximation is not quite as good. Therefore, filter performance above threshold may be improved by including more Fourier coefficients in the FCF. Finally, we refer the reader to Table 6.13, in which we present the results of several simulation runs of the FCF without the Wong-Zakai correction terms. Note that the performance is somewhat worse, moreso below threshold where the filter works best.
CHAPTER 7

EXTENSIONS TO MORE GENERAL PROBLEMS INCLUDING RIGID BODY ROTATION

"Any general statement is like a cheque drawn on a bank. Its value depends on what is there to meet it."
- Ezra Pound [P10]

7.1 Introduction

In this chapter we briefly discuss some generalizations of the problems and results of the preceding chapters. As mentioned earlier, the $S^1 - SO(2)$ problems are conceptually quite straightforward, and the only reason for placing them in a setting as general as the one we have is to motivate new methods of attack for a wide variety of problems. In Section 7.2 we examine several rigid body rotation problems, while Section 7.3 contains some comments concerning further generalizations. We note that some of the theoretical developments discussed here have been derived by Lo and Willsky (unpublished work).

7.2 Random Rotation of Rigid Bodies

The orientation of a rigid body can be determined by specifying the direction cosines [W2], [W10] between two sets of orthogonal axes -- one rotating with the body and one an inertial reference frame. The direction cosines are usually given as a $3 \times 3$ orthogonal matrix $X$ of positive determinant -- i.e.

$$X'X = I \quad \text{det } X = +1 \quad (7.1)$$

Referring to Chapter 1 and Appendix A, we note that the set of all
direction cosine matrices forms the matrix Lie group SO(3). Before discussing random processes on SO(3), we will mention some of its important properties.

The matrix Lie algebra so(3) associated with SO(3) is three-dimensional and has the basis

\[
R_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \quad R_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad R_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}
\] (7.2)

Note that we have the Lie bracket relations (see Appendix A)

\[
[R_1, R_2] = R_3 \quad [R_2, R_3] = R_1 \quad [R_3, R_1] = R_2
\] (7.3)

Referring to Definition 2.1, we see that so(3) is not a solvable Lie algebra (in fact, it is a simple Lie algebra [S1]). Referring to the results of Wei and Norman [W3], [W4] discussed in Section 2.2, we can make the following comments. Consider the differential equation

\[
\dot{X}(t) = \left( \sum_{i=1}^{3} u_i(t)R_i \right)X(t) \quad X(0) = I
\] (7.4)

We note that \(X(t) \in SO(3)\) and that the \(u_i\) can be interpreted as angular velocities. From Theorem 2.1, we see that there exists a time interval \([0,T]\) and time functions \(g_1(t), g_2(t), g_3(t)\) such that

\[
X(t) = \exp[g_1(t)R_1]\exp[g_2(t)R_2]\exp[g_3(t)R_3]
\] (7.5)
Since so(3) is not solvable, Theorem 2.2 does not apply, and thus (7.5) may not be a global representation (i.e. \( T \neq \infty \)). In fact, referring to [H15] and [W2], we see that (7.5) is an Euler angle representation, and, in general, this is not global. The reader is referred to [W2] for a discussion of the difficulties with Euler angles and the related phenomenon of gimbal-lock. We will discuss a stochastic analog of this later in this section.

In addition to the direction cosine and Euler angle representations, there are several other mathematical concepts related to rigid body rotation. Two mathematical tools often used in studying rotational processes are the unit quaternions and the Cayley-Klein parameters [W10], [S15]. The unit quaternions \( Q \) are defined by

\[
Q = \left\{ q = r_0 + r_1 i + r_2 j + r_3 k \mid \sum_{i=0}^{3} r_i^2 = 1 \right\}
\]  

(7.6)

where we define group multiplication on \( Q \) by the relations

\[
\begin{align*}
i^2 &= j^2 = k^2 = -1 & & i j &= -j i = k \\
j k &= -k j = i & & k i &= -i k = j
\end{align*}
\]

(7.7)

The Cayley-Klein parameters are closely related to the matrix Lie group \( SU(2) \)

\[
SU(2) = \left\{ A \in M(2, C) \mid A A^\dagger = I, \det A = +1 \right\}
\]

(7.8)

We note that these two groups are isomorphic as Lie groups [W10], [W1], [C1], where we have the isomorphism \( f : Q \to SU(2) \) given by
\[ f(x+ui+vj+yk) = \begin{bmatrix} x+iv & u+iv \\ -u+iv & x-iy \end{bmatrix} \]  

(7.9)

In addition the 3-sphere

\[ S^3 \triangleq \{ x' = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4 | x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1 \} \]  

(7.10)

can be made into a Lie group by identifying it with \( Q \)

\[ (x_1, x_2, x_3, x_4) \leftrightarrow x_1 + x_2 i + x_3 j + x_4 k \]  

(7.11)

For those familiar with Lie theory, \( SU(2) \) and \( Q \) are two representations of the simply connected covering group, \([W1], [C1]\), of \( SO(3) \), and we have the isomorphism

\[ SO(3) \cong SU(2)/\{i \} = Q/\{i \} \]  

(7.12)

where, for instance, \( \{i \} \) is the subgroup of \( SU(2) \) containing only these two elements. The isomorphism is illustrated by considering the homomorphism \( g : Q \rightarrow SO(3) \) given by

\[ g(x_1 + x_2 i + x_3 j + x_4 k) = \begin{bmatrix} x_2^2 - x_3^2 + x_4^2 + x_1^2 \\ 2(x_2 x_3 + x_1 x_4) \\ -x_2^2 - x_3^2 + x_4^2 + x_1^2 \\ 2(x_2 x_4 - x_1 x_3) \\ 2(x_3 x_4 - x_1 x_2) \\ 2(x_3 x_4 - x_1 x_2) \end{bmatrix} \]  

(7.13)

Note that

\[ g(q) = g(-q) \quad \forall \ q \in Q \]  

(7.14)

and, in fact, we can show

\[ g(Q) = SO(3) \quad ker \ g = \{i \} \]  

(7.15)
Then, by the First Isomorphism Theorem (Theorem A.7)

\[ SO(3) = \mathbb{Q}/\{\pm 1\} \]  

(7.16)

where the isomorphism is a purely algebraic one. The isomorphism can be shown to be a Lie group isomorphism (i.e., it is smooth), but we will not discuss this here.

We will find it useful to introduce spherical coordinates on \( S^3 \). Any point on \( S^3 \) can be uniquely expressed as

\[ q(\theta_1, \theta_2, \theta_3) \triangleq (\cos \theta_1, \sin \theta_1 \cos \theta_2, \sin \theta_1 \sin \theta_2 \cos \theta_3, \sin \theta_1 \sin \theta_2 \sin \theta_3) \]

\[ \theta_1 \in [0, \pi], \quad \theta_2 \in [0, \pi], \quad \theta_3 \in [0, 2\pi) \]  

(7.17)

Referring to [W10], we note that one can interpret \((\theta_1, \theta_2, \theta_3)\) in the following sense: any rotation or set of rotations in \( \mathbb{R}^3 \) is equivalent to a single rotation about some axis; the angles \( \theta_2, \theta_3 \) determine the axis and \( 2\theta_1 \) is the amount of the rotation. Note that a rotation of \( \theta \) about the axis \( v \) is the same as a rotation of \(-\theta\) about the axis \(-v\). Thus, the points \( q(\theta_1, \theta_2, \theta_3) \) and \( q(\pi-\theta_1, \pi-\theta_2, (\pi+\theta_3) \mod 2\pi) \) lead to the same rotation (one can check that \( q(\theta_1, \theta_2, \theta_3) = -q(\pi-\theta_1, \pi-\theta_2, (\pi+\theta_3) \mod 2\pi) \) and we then use (7.14)). We remark that we will refer to \( q(\theta_1, \theta_2, \theta_3) \) both as an element of \( \mathbb{Q} \) and as an element of \( S^3 \), where we implicitly employ (7.11).

By studying the properties of random processes on \( S^3 \), we can also determine the properties of random processes on \( SO(3) \) and \( SU(2) \). In addition, such a study would yield information about processes on \( S^2 \). The type of equation of interest to us in studying processes on \( S^2 \) is
\[ \dot{x}(t) = \left( \sum_{i=1}^{3} u_i(t)R_i \right)x(t) \]  
(7.18)

Note that \( x'(t)x(t) = \) constant, and thus if \( x(0) \in S^2 \) (i.e. \( x'(0)x(0) = 1 \)), \( x(t) \in S^2 \) \( \forall t \). Also, referring to (7.4),

\[ x(t) = X(t)x(0) \]  
(7.19)

It is precisely this fact that allows us to determine the properties of processes on \( S^2 \) by studying processes on \( SO(3) \). For those familiar with Lie theory and the theory of homogeneous and coset spaces [W1], [H16], we note that \( S^2 \) is diffeomorphic [W1] to the space of cosets \( SO(3)/SO(2) \) (i.e. there is a smooth one-to-one correspondence between points in \( S^2 \) and cosets of \( SO(3)/SO(2) \)). We here identify \( SO(2) \) as a subgroup of \( SO(3) \), as in Appendix A — e.g. \( SO(2) \) is isomorphic to \( \{ \exp R_2 \} \).

We now wish to study random processes in the settings we have just described. To do this, we will apply the Brownian motion injection procedure discussed in Section 2.3. The situation here is much more complicated than that in the \( S^1 \) or the general abelian Lie group case. Recall that if \( w \) is a Brownian motion process on \( R^1 \) and if we inject it into \( S^1 \), we have (see (2.48))

\[ W(t) = \bigcap_{s \leq t} \exp(Rdw(s)) = \begin{bmatrix} \cos w(t) & \sin w(t) \\ -\sin w(t) & \cos w(t) \end{bmatrix} \]  
(7.20)

That is, \( W(t) \) is an explicit function only of \( w(t) \), and this is directly attributable to the commutativity of rotations about a fixed axis. Since general rotations in \( R^3 \) do not commute (i.e. \( SO(3) \) and \( O \) are not abelian), we cannot expect such a simplification. That is, the process formed by
the injection procedure will depend explicitly on the entire past history of the process used in the injection. We will find that this complication has a major effect in considering certain estimation problems.

We will now define the stochastic analog of (7.4). This was first considered in [M4]. Let \( w \) be a standard 3-dimensional Brownian motion process, \( f \) a fixed element of \( \mathbb{R}^3 \), and \( P \) a given nonnegative definite 3x3 matrix. Define the processes

\[
v(t) = Pw(t) + ft \tag{7.21}
\]

\[
X(t) = \bigwedge_{s \leq t} \exp\left( \sum_{i=1}^{3} R_i dv_i(s) \right) \tag{7.22}
\]

As discussed in Section 2.3, \( X \) satisfies the stochastic differential equation

\[
dX(t) = \left( \sum_{i=1}^{3} R_i dv_i(t) + \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} P_{ij} R_i R_j dt \right) X(t) \tag{7.23}
\]

As in the \( S^1 \) case, we wish to examine the probability density for \( X(t) \). In the \( S^1 \) case we did not use the SO(2) representation to find the density, but rather the representation \( \theta \in [-\pi, \pi) \) -- i.e., a coordinatization (see [W1]). We will do the same here -- that is, we will study the \( S^3 \) (or \( O \))-analog of (7.21) - (7.23) and will use the coordinatization given in (7.17). To do this, we consider the Lie algebra with basis
\[
S_1 = \begin{bmatrix}
0 & 1 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 \\
0 & 0 & 1 & 0
\end{bmatrix} \quad S_2 = \begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
-1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0
\end{bmatrix} \quad S_3 = \begin{bmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & -1 & 0 \\
0 & 1 & 0 & 0 \\
-1 & 0 & 0 & 0
\end{bmatrix}
\] (7.24)

This Lie algebra can be shown to be isomorphic to so(3). Consider the process \( v \) given in (7.21) and define the injected process \( s(t) \in S^3 \)

\[
s(t) = \bigcap_{s \in t} \exp\left( \frac{1}{2} \sum_{i=1}^{3} S_i dv_i(s) \right) \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\] (7.25)

One can show that this injection into \( S^3 \) or \( Q \) is a precise analog of the injection into SO(3).

Referring to Section 2.3, or to [M3], we note that the probability density \( p \) for \( s(t) \) satisfies

\[
\frac{\partial p}{\partial t} = \mathcal{L}^* p
\] (7.26)

where \( \mathcal{L}^* \) is the formal adjoint of the differential operator

\[
\mathcal{L} = \frac{1}{2} D' PD + f'D = \frac{1}{2} \sum_{i,j=1}^{3} p_{ij} D_i D_j + \sum_{i=1}^{3} f_i D_i
\] (7.27)

(see Section 2.3). Here \( p \) is the density with respect to the standard surface area element on \( S^3 \), \( \sin^2 \theta_1 \sin \theta_2 \ d\theta_1 d\theta_2 d\theta_3 \), and the differential operators, \( D_i, i=1,2,3 \) are defined as follows. Let \( q(\theta_1, \theta_2, \theta_3) \) be the element of \( S^3 \) given by (7.17) and let \( h : S^3 \to R \) be a given function, where we regard \( h \) as an explicit function of the \( \theta_i \). Then the \( D_i \) are defined by

\[
(D_i h)(\theta_1, \theta_2, \theta_3) = \frac{d}{dt} h[\exp(\frac{1}{2} S_i t) q(\theta_1, \theta_2, \theta_3)] \bigg|_{t=0}
\] (7.28)

Using (7.17), (7.24), and (7.28) we compute...
\[ D_1 = \frac{1}{2} \left\{ -\cos \theta_2 \frac{\partial}{\partial \theta_1} + \frac{\cos \theta_1 \sin \theta_2}{\sin \theta_1} \frac{\partial}{\partial \theta_2} - \frac{\partial}{\partial \theta_3} \right\} \quad (7.29) \]

\[ D_2 = \frac{1}{2} \left\{ -\sin \theta_2 \cos \theta_3 \frac{\partial}{\partial \theta_1} + (\sin \theta_3 - \frac{\cos \theta_1 \cos \theta_2 \cos \theta_3}{\sin \theta_1}) \frac{\partial}{\partial \theta_2} 
+ \left( \frac{\cos \theta_1 \sin \theta_3}{\sin \theta_1 \sin \theta_2} + \frac{\cos \theta_2 \cos \theta_3}{\sin \theta_2} \right) \frac{\partial}{\partial \theta_3} \right\} \quad (7.30) \]

\[ D_3 = \frac{1}{2} \left\{ -\sin \theta_2 \sin \theta_3 \frac{\partial}{\partial \theta_1} - (\cos \theta_3 + \frac{\cos \theta_1 \cos \theta_2 \sin \theta_3}{\sin \theta_1}) \frac{\partial}{\partial \theta_2} 
+ \left( -\frac{\cos \theta_1 \cos \theta_3}{\sin \theta_1 \sin \theta_2} + \frac{\cos \theta_2 \sin \theta_3}{\sin \theta_2} \right) \frac{\partial}{\partial \theta_3} \right\} \quad (7.31) \]

The solution of (7.26) in the general case has not been found. However, we will present the solution for the particular example \( P = I \), \( f = 0 \). In this case, one can show that \( D^* = D \) and (7.26) becomes

\[ \frac{\partial p}{\partial t} = \frac{1}{2} \left( D_1^2 + D_2^2 + D_3^2 \right) = \frac{1}{8} \Delta_3 \quad (7.32) \]

where \( \Delta_3 \) is the Laplacian on \( S^3 \) \([W3], [H6]\), and is given by

\[ \Delta_3 = \frac{\partial^2}{\partial \theta_1^2} + 2 \frac{\cos \theta_1}{\sin \theta_1} \frac{\partial}{\partial \theta_1} + \frac{\cos \theta_2}{\sin^2 \theta_1 \sin \theta_2} \frac{\partial}{\partial \theta_2} + \frac{1}{\sin^2 \theta_1} \frac{\partial^2}{\partial \theta_1^2} + \frac{1}{\sin^2 \theta_2} \frac{\partial^2}{\partial \theta_2^2} + \frac{\partial^2}{\partial \theta_3^2} \quad (7.33) \]

Referring to Perrin \([P1]\), we have the Green's function for (7.32)

\[ p(\theta_1, \theta_2, \theta_3, t) = \frac{e^{t/8}}{\pi^2 \sin^2 \theta_1} \sum_{n=0}^{\infty} \frac{(2n+1)^2 t}{(2n+1) \sin(2n+1) \theta_1} \quad (7.34) \]

This density form is the \( S^3 \)-analog of the normal density, as the folded
normal is the $S^1$ analog. Note that it is an explicit function only of $\theta_1$. The physical meaning of this is quite clear. With $P = I$ and $f = 0$, we are injecting a standard 3-dimensional Brownian motion process. Recalling that $\theta_2$ and $\theta_3$ determine an axis of rotation, any dependence of $p$ on these variables would imply that rotations about particular axes are more likely than rotations about others, which intuitively is not the case. Thus, $p$ depends only on the total rotation which is determined by $\theta_1$.

Solution of more general problems may be aided by the introduction of a tool that is the analog of Fourier series -- spherical harmonics [B20], [W5], [M16]. We will not discuss this here except to mention that spherical harmonics may prove to be as useful a tool here as Fourier series was in the $S^1$ case (see the work of Yosida [Y2]) -- both in defining optimal estimation criteria and in deriving optimal filter forms that have desirable structural properties. We will have more to say about this in the next section. In addition, we note that a different type of series expansion -- one directly involving moments -- has been proposed by Brockett [B4]. One can show that there is a direct relationship between spherical harmonics and moments (i.e. they convey essentially the same information), and in some cases the moments may be easier to work with. The reader is referred to [B4] for details.

Another possible technique for analyzing the properties of the solution to (7.23) is to determine product forms for the solution to this random equation, analogous to the deterministic expressions developed by Wei and Norman [W3], [W4]. In particular, we might ask if there is a stochastic analog of Theorem 2.1 (we note that since $SO(3)$ is not a
solvable Lie group, the global result of Theorem 2.2 does not apply). The answer to this question is that there is no stochastic analog of Theorem 2.1 for the SO(3) case. The reason is the following: Theorem 2.1 states that over some time interval \([0,T]\) the solution to (7.4) can be expressed as in (7.5); however the size of the interval \([0,T]\) depends on the particular functions \(u_i\) (i.e. when we reach gimbal-lock is determined by the angular velocities). Thus, if the \(u_i\) are stochastic, as in (7.23), each sample path leads to a different interval, and one can show that if the \(u_i\) are Brownian, this interval can be made as small as possible by choosing the proper sample path. Thus, there exists no such interval that one can use for all (or even almost all) sample paths, and the stochastic analog fails. We note that one can follow the Wei-Norman construction and determine formal stochastic differential equations for \(g_i\), \(i=1,2,3\), assuming the solution to (7.23) can be represented as in (7.5). However, these stochastic equations do not have solutions (the right-hand sides of these equations contain terms such as \(\frac{1}{\cos \theta_2}\), which is not globally Lipschitz or bounded; this problem is again closely related to the phenomenon of gimbal-lock).

We will now briefly consider some estimation problems on SO(3). As in the \(S^1\) case, we need to study various estimation criteria. To do this, we need a measure of "distance" on SO(3) or \(S^3\) — i.e. a Riemannian metric [L12], [W1], [K20]. An example of such a metric can be derived from the distance element on \(S^3\)

\[
 ds = \left(d\theta_1^2 + \sin^2 \theta_1 d\theta_2^2 + \sin^2 \theta_1 \sin^2 \theta_2 d\theta_3^2 \right)^{1/2} 
\]  
(7.35)
Given such a metric \( \rho : SO(3) \times SO(3) \rightarrow [0, \infty) \), we wish to consider error criteria \( \phi : SO(3) \rightarrow [0, \infty) \) that are nondecreasing with respect to \( \rho \) — i.e.

\[
\rho(I, X_1) \geq \rho(I, X_2) \implies \phi(X_1) \geq \phi(X_2) \quad \forall \, X_1, X_2 \in SO(3)
\]  

(7.36)

Clearly \( \phi(X) = \rho(I, X) \) or \( \rho^2(I, X) \) satisfy (7.36). Another reasonable criterion is

\[
\phi(X) = \frac{1}{2} (3 - \text{tr } X)
\]  

(7.37)

Recalling that any rotation can be regarded as a single rotation of an amount \( \theta_x \) about some axis, and referring to [H15], we can find a similarity transformation such that

\[
X = P \begin{bmatrix}
\cos \theta_x & \sin \theta_x & 0 \\
-sin \theta_x & \cos \theta_x & 0 \\
0 & 0 & 1
\end{bmatrix} P'
\]  

(7.38)

and

\[
\text{tr } X = 1 + 2 \cos \theta_x
\]  

(7.39)

Substituting (7.39) into (7.37), we have

\[
\phi(X) = 1 - \cos \theta_x
\]  

(7.40)

and we see that this particular criterion is a generalization of the \( S^1 - SO(2) \) criterion. Also, recalling the relationship between \( SO(3) \) and the spherical coordinates on \( S^3 \) given by (7.17), the \( S^3 \) analog of (7.37) is

\[
\phi(\theta_1, \theta_2, \theta_3) = 1 - \cos 2\theta_1
\]  

(7.41)
We will now formulate and discuss an analog of the continuous-time bilinear estimation problem considered in Chapter 4. Consider the 3-dimensional stochastic process $x(t)$ that satisfies the linear Ito differential equation

$$dx(t) = A(t)x(t)dt + Q^{1/2}(t)dw(t) \quad (7.42)$$

where $w$ is a standard 3-dimensional Brownian motion process. We can then consider injecting $x$ into $SO(3)$ in one of two ways

$$X_1(t) = \bigcap_{s \leq t} \exp\left( \sum_{i=1}^{3} R_i \, dx_i(s) \right) \quad (7.43)$$

$$X_2(t) = \bigcap_{s \leq t} \exp\left( \sum_{i=1}^{3} R_i x_i(s) \, ds \right) \quad (7.44)$$

(one can verify, as in [M3], that the product integrals are well defined). We can derive differential equations for $X_1$ and $X_2$ much as McKean [M4] does for Brownian motion processes

$$dX_1(t) = \left( \sum_{i=1}^{3} R_i \, dx_i(t) \right) + \frac{1}{2} \sum_{i,j=1}^{3} \mathbb{Q}_{ij}(t) R_i R_j dt X_1(t) \quad (7.45)$$

$$dX_2(t) = \left( \sum_{i=1}^{3} R_i x_i(t) \right) X_2(t) \, dt \quad (7.46)$$

Equation (7.45) is an Ito differential equation, but (7.46) is an ordinary differential equation. We note that $x$ in (7.45) can be interpreted as the total angles of rotation about three orthogonal axes (for instance, as measured by integrating gyroscopes). In (7.46) $x$ plays the role of an angular velocity vector. Note that, as in the $SO(2)$ case discussed in Chapter 4, $X$, given by (7.43) or (7.44), is
not, in general, a Markov process. We remark that $X_t$ is a Markov process if and only if $A = 0$.

We now consider measurement processes. One natural way of the process considered in Chapter 4 is the following. The 3-dimensional process satisfying the Ito equation

$$dz(t) = C(t)x(t)dt + S^{1/2}(t)dv(t) \quad z(0) = 0$$

where $v$ is a standard 3-dimensional Brownian motion process adapted to the process $w$. Define the process

$$Z(t) = \bigcap_{s \leq t} \exp\left( \sum_{i=1}^{3} R_i dz_i(s) \right)$$

We note that $Z$ satisfies

$$dZ(t) = \left( \sum_{i=1}^{3} R_i dz_i(t) + \frac{1}{2} \sum_{i,j=1}^{3} S_{ij} R_i R_j dt \right) Z(t)$$

and that we can formally recover $z$ from $Z$ from the relation

$$dE(t) = [dZ(t)]Z'(t) - \frac{1}{2} \sum_{i,j=1}^{3} S_{ij} R_i R_j dt$$

$$dz(t) = \begin{bmatrix} dE(t)_{32} \\ dE(t)_{13} \\ dE(t)_{21} \end{bmatrix}$$

We note that Lo and Willsky (unpublished work) have rigorously established the one-to-one relationship between $Z$ and $z$, and thus we have knowledge of $\{Z(s) | s \leq t\}$ is equivalent to knowledge of $\{z(s) | s \leq t\}$. Also, if we interpret $x(t)$ as an angular velocity, we can interpret $\dot{z}(t)$ as the output of rate gyroscopes, $z(t)$ as the input...
by integrating gyros, and \( Z(t) \) as the orientation information provided either by the measurement of gimbal angles or as the output of a direction cosine computer that takes \( Z \) as its input (as in a strap-down system [W2], in which quantization errors can be modeled as additive noise in the measurement of angular velocities).

We can also consider an alternative measurement process that is conceptually quite similar to the above process. Suppose we directly measure the orientation \( X(t) \) but that our measurement \( M(t) \in \text{SO}(3) \) is corrupted by noise that can be modeled by

\[
M(t) = X(t) V(t)
\]  

(7.52)

where \( V \) is an \( \text{SO}(3) \) Brownian motion process satisfying

\[
dV(t) = V(t) \left( \sum_{i=1}^{3} R_i dV_i(t) + R_0(t) dt \right)
\]  

(7.53)

the \( V_i \) are Brownian motions independent of \( X \), and \( R_0 dt \) is the Ito correction term. We note that such a measurement model is related to some inertial navigation systems. For instance (see [W2]), suppose an airplane is equipped with an inertial platform that is instrumenting a particular reference coordinate system, \( r \), and, by reading off gimbal angles, we can determine the orientation \( M(t) \) of the airplane with respect to the platform. If the platform drifts from the desired frame \( r \), \( M(t) \) does not give an error free measurement of \( X(t) \), the orientation of the airplane with respect to the reference. In fact, the discrepancy is precisely the drift of the platform from \( r \). Referring to the algebraic manipulations of direction cosine matrices discussed in [W2], we find that (7.52) is the proper observation model.
If we assume \( X \) satisfies (7.46), we have

\[
dM(t) = \left( \sum_{i=1}^{3} R_i x_1(t) \right) M(t) dt + M(t) \left( \sum_{i=1}^{3} R_i v_i(t) + R_0(t) \right) dt
\]  

(7.54)

Since, at time \( t \), \( M(t) \) is known, (7.54) is a "linear" measurement of \( x_1(t) \) if we feed in \( M(t) \) as a gain matrix.

Our problem is to estimate either \( x(t) \) or \( X(t) \) -- satisfying (7.45) or (7.46) -- given one of the observation processes that have been discussed. Since the \( X \) process is formed from the \( x \) process, we would expect the optimal estimator to have the form depicted in Figure 7.1. Since the entire past history of \( x \) directly affects \( X(t) \)

![Diagram](image)

**Figure 7.1:** Form of the Optimal SO(3) Filter

(since the product integral does not reduce to a simple form as it does in the abelian group case), the nonlinear post-processor and the filtering information needed as an input to it will probably be infinite-dimensional in nature. However, the actual optimal solution for \( X \) is unknown to the author, and instead we will discuss the optimal
estimation of $x$ and some intuitively appealing suboptimal approaches for estimating $X$.

The estimation of the $R^3$ variable $x$ is just as straightforward as the estimation of the $R^1$ variable $x$ discussed in Chapter 4 (see Theorem 4.2). We take the conditional mean, $\mathcal{E}(x(t) | \text{observation up to time } t)$, as the optimal estimate. If $dz$ is taken as the observation process, we have a standard linear filtering problem, and the optimal filter is the Kalman-Bucy filter. If we take $z$ as the measurement, we can use the techniques discussed in Chapter 6 for the $S^1$ case (assuming that differentiation is not possible). That is, we consider the augmented state vector

$$y(t) = \begin{bmatrix} x(t) \\ z(t) \end{bmatrix}$$

which satisfies

$$dy(t) = F(t)y(t)dt + G^{1/2}(t)du(t)$$

where

$$F(t) = \begin{bmatrix} A(t) & 0 \\ C(t) & 0 \end{bmatrix}, \quad G^{1/2}(t) = \begin{bmatrix} Q^{1/2}(t) & 0 \\ 0 & S^{1/2}(t) \end{bmatrix}, \quad u(t) = \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}$$

(7.57)

In this case, we are taking a perfect measurement of the last three components of $y$, and we can use standard techniques, [J2], [B12], to determine the optimal estimate of $x$. We also note that if $A$, $C$, $Q$, and $S$ are constant, we can use a vector version of Wiener-Hopf theory [V5], [D7] to solve for the optimum steady-state filter.

The problem of estimating $x$ given $dZ$ is quite similar to the
analogous problem discussed in Chapter 4. Using the equivalence of
Z and z and equations (7.50) and (7.51), we construct the optimal
filter form illustrated in Figure 7.2 (compare to Figure 4.2). A
nonlinear preprocessor is used to determine dz, and this signal is
fed into a Kalman-Bucy linear filter. If Z is taken as the observation,
we can either differentiate it to determine dz or we can use the
technique developed in [O1] and discussed in Chapter 6. That is, we
compute z(t) from Z using the equation

\[ z(t) = \int_0^t Z'(s)dZ(s) - \frac{1}{2} \sum_{i,j=1}^3 R_i R_j \int_0^t S_{ij}(s)ds \]  

(7.58)

Having z(t), we can use one of the techniques mentioned previously.

Finally, suppose we take dM, given by (7.54), as our observation.
At time t, M(t) is known, and our observation is of the form

\[ dM(t) - M(t)R_0(t)dt = \sum_{i=1}^3 H_i(t)x_i(t)dt + \sum_{i=1}^3 D_i(t)dv_i(t) \]  

(7.59)

which clearly is linear. Note however that the matrices H_i and D_i
are functions of M -- i.e. they are not known a priori. On the other
hand, in linear theory knowledge of these matrices is necessary for
the off-line computation of the error covariance (i.e. the solution
of the Riccati equation) and the filter gain. Thus, in this case we
must compute these gains on-line and feed them into the linear filter.

The form of the filter is illustrated in Figure 7.3, and we can show
that it is indeed the optimal filter; however, we will not present a
rigorous derivation of the filter equations. In addition, the problem
if we take M as the observation appears to be more difficult, and we
will not consider it here except to mention that, as in [O1], some sort
Figure 7.2: Illustrating the Optimal Filter for $x$ with Measurement Process (7.49)

Figure 7.3: Illustrating the Optimal Filter for $x$ with Measurement Process (7.54)
of logarithmic preprocessing followed by a linear filter might be optimal (note that for scalars, \( \log xy = \log x + \log y \) -- i.e. multiplicative noise becomes additive; however, it is not clear what we should do in the matrix case given in (7.52)).

We now turn our attention to the estimation of \( X(t) \). Since the entire past history of \( x \) enters into \( X \), it would appear that to incorporate a new observation, say \( dZ \), into our estimate, we will have to smooth the entire estimated past history of \( x \) -- i.e. we will have to consider the effect of \( dZ(t) \) on our estimate of \( x(s) \) for all \( s \leq t \). The precise form of the optimal estimator is not known, but it is the author's feeling that it will turn out to be infinite dimensional. We will now suggest two ad hoc estimation schemes that may prove to be useful. For the first scheme, we assume that \( X \) satisfies (7.46). In this case, if we let

\[
\hat{x}(t|t) = \mathcal{E}[x(t)|\text{observations up to time } t] \tag{7.60}
\]

we might take the solution of

\[
\frac{d}{dt} \hat{x}(t) = \left( \sum_{i=1}^{3} R_i \hat{x}_i(t|t) \right) \hat{x}(t) \tag{7.61}
\]
as our estimate. This filter is illustrated in Figure 7.4 (where we assume that \( dZ \) is the observation). An obvious extension of this would be to incorporate a time delay into the system -- i.e. provide an estimate of \( X(t) \) at time \( t+\Delta \). In this case we could replace the filtering estimate \( \hat{x}(t|t) \) with the smoothing estimate \( \hat{x}(t|t+\Delta) \).

Now suppose \( X \) is given by (7.43). From Section 2.3, we have
Figure 7.4: A Possible Suboptimal SO(3) Estimator

the following approximation to $X$

$$X(k\Delta) = \exp \left[ \sum_{i=1}^{3} R_i (x_i(k\Delta) - x_i((k-1)\Delta)) \right] X((k-1)\Delta)$$

(7.62)

As discussed in Section 2.3, the process defined in (7.62) converges to $X$ as $\Delta \to 0$ (and $k \to \infty$ in an appropriate manner). Thus assume that we have chosen $\Delta$ sufficiently small for the desired application. Then, to estimate $X(k\Delta)$, we need only estimate $x(j\Delta)$, $j=0,1,\ldots,k$. This suggests an approximation technique. Choose a nonnegative integer $N$ and let $x(k\Delta|t,N)$ be defined by

$$\hat{x}(k\Delta|t,N) = \begin{cases} \mathbb{E}[x(k\Delta)|\text{observations up to } t] & t \leq (k+N)\Delta \\ \mathbb{E}[x(k\Delta)|\text{observations up to } (k+N)\Delta] & t > (k+N)\Delta \end{cases}$$

(7.63)
That is, we smooth our estimate of \( x(k\Delta) \) only up to time \((k+N)\Delta\), and from that time on we leave our estimate fixed. Thus at any time \( t \) the only values of \( k\Delta \) for which we are updating our estimates are the \((N+1)\) largest values of \( k\Delta \) that are less than or equal to \( t \).

In other words, we can visualize a bank of \((N+1)\) filters. The first one computes the filtering solution \( \hat{x}(t|t) \) (note that \( \hat{x}(k\Delta|k\Delta,N) = \hat{x}(k\Delta|k\Delta) \)). Letting \( M\Delta < t < (M+1)\Delta \), each of the other \( k \) filters computes \( \hat{x}(k\Delta|t,N) \) with \( k \) fixed for each filter and \( k \in \{M-N+1, M-N+2, \ldots, M\} \). When \( t \) reaches \((M+1)\Delta\), the filter that was computing an estimate of \( x((M-N+1)\Delta) \) begins to compute an estimate of \( x((M+1)\Delta) \) with initial condition \( \hat{x}(M+1)\Delta|(M+1)\Delta,N) = \hat{x}((M+1)\Delta|(M+1)\Delta) \), the output of the first filter.

We note that for all of the measurement processes with the exception of \( M \) and \( dM \), filtering and smoothing error covariances can be computed a priori, and we can choose \( N \) to be large enough so that the smoothed error covariance for the estimate \( \hat{x}(k\Delta|(k+N)\Delta) \) is as small as desired. Having computed the \( \hat{x}(k\Delta|t,N) \) we can compute an estimate, \( \hat{X}(k\Delta|t,N) \), of \( X(t) \) from (7.62) by replacing \( x_1(k\Delta) \) with \( \hat{x}_1(k\Delta|t,N) \). Note that only the last \((N+1)\) terms in the product expression for \( \hat{x}(k\Delta|t,N) \) are affected by the incoming measurement -- i.e. the estimate of \( X(k\Delta) \) is kept fixed from time \((k+N)\Delta \) on. It is not clear how well an estimator of this type will work or even if it will work at all, and some experimental effort is needed in order to determine estimator performance characteristics and values of the parameters \( \Delta \) and \( N \) that provide a reasonable tradeoff between accuracy and computational complexity.
We close this section by mentioning a novel SO(3) estimation technique introduced by Stuck [S15]. The technique relies on the analogy between one-dimensional (SO(2)) and three-dimensional (SO(3)) rotation, and essentially consists of designing a SO(3)-analog of a phase-lock loop. The results in [S15] for stochastic attitude estimation are just a first step in this type of analysis (i.e. Stuck assumes \( x \) is fixed and that we are taking continuous noisy measurements of the orientation), and some further work along this line may yield valuable results. The reader is referred to [S15] for details.

7.3 Some Comments on General Lie Group Estimation Problems

Motivated by the preceding discussion, we will now present and comment on a general class of estimation problems on an arbitrary matrix Lie group. Let \( G \) be such a group with associated matrix Lie algebra \( L \), and let \( \{ A_1, \ldots, A_n \} \) be a basis for \( L \). Suppose we have an \( n \)-dimensional process \( x \) satisfying

\[
dx(t) = A(t)x(t)dt + Q^{1/2}(t)dw(t)
\]

(7.64)

where \( w \) is a standard \( n \)-dimensional Brownian motion. We then inject \( x \) into \( G \) in one of two ways

\[
X_1(t) = \bigcap_{s \leq t} \exp(\sum_{i=1}^{n} A_i dx_i(s))
\]

(7.65)

\[
X_2(t) = \bigcap_{s \leq t} \exp(\sum_{i=1}^{n} A_i x_i(s)ds)
\]

(7.66)

and we have
\[ dX_1(t) = \left( \sum_{i=1}^{n} A_i dX_i(t) + \frac{1}{2} \sum_{i,j=1}^{n} Q_{ij}(t) A_i A_j dt \right) X_1(t) \]  \hspace{1cm} (7.67)

\[ dX_2(t) = \left( \sum_{i=1}^{n} A_i X_i(t) \right) X_2 dt \]  \hspace{1cm} (7.68)

We note the proofs that (7.65) and (7.66) are well defined and that the solutions satisfy (7.67) and (7.68) are analogous to the proofs presented by McKean [M3] for Brownian motion on Lie groups. Also \( X_1 \) is a Markov process if and only if \( A \equiv 0 \).

We also note that if we let \( X_1 \) and \( X_2 \) in (7.67) and (7.68) be \( n \)-vectors we can consider the analog of the \( S^2 \) problem mentioned in the previous section. Such a study would lead to the investigation of the properties of stochastic processes on homogeneous spaces. We will not consider that here except to mention that as with the \( SO(3)-S^2 \) case, many of the properties of the processes on a homogeneous space may be discovered by studying processes on a related Lie group.

We now define an observation process. Let \( H \) be another (possibly the same) matrix Lie group with associated matrix Lie algebra \( M \) having the basis \( B_1, \ldots, B_m \). Let \( z \) be the \( m \)-dimensional stochastic process satisfying

\[ dz(t) = C(t)x(t)dt + R^{1/2} dv(t) \quad z(0) = 0 \]  \hspace{1cm} (7.69)

where \( v \) is a standard \( m \)-dimensional Brownian motion independent of \( w \).

Our measurement process \( Z \) is the \( H \)-valued function given by

\[ Z(t) = \bigcap_{s \leq t} \exp( \sum_{i=1}^{m} B_i dz_i(s) ) \]  \hspace{1cm} (7.70)

and \( Z \) satisfies
\[ dZ(t) = \left( \sum_{i=1}^{m} B_i dz_i(t) \right) + \frac{1}{2} \sum_{i,j=1}^{m} R_{ij} B_i B_j dt \] 
\[ Z(0) = I \quad (7.71) \]

As in the SO(2) and SO(3) cases, it can be shown that the \( z \) and \( Z \) processes are equivalent — i.e. knowledge of one is equivalent to knowledge of the other. In fact, we have

\[ \sum_{i=1}^{m} B_i dz_i(t) = [dZ(t)]Z^{-1}(t) - \frac{1}{2} \sum_{i,j=1}^{m} R_{ij} B_i B_j dt \quad (7.72) \]

Since the \( B_i \) are linearly independent, (7.72) uniquely determines \( dz(t) \).

We can now pose analogs of some of the questions we have considered in the SO(2) and SO(3) cases. The first of these is to determine if and when the solution of (7.67) can be represented à la Wei and Norman. As seen in the earlier chapters, such a representation can be found if the underlying Lie group is abelian (see (2.48)); however, the SO(3) example discussed in the preceding section shows that such a representation need not exist in the general case. The problem in that particular case was related to the fact that the Wei-Norman representation result for SO(3) is only local. An interesting open question is to determine if a stochastic analog of the Wei-Norman representation exists when the deterministic version is global — e.g. if the underlying group is solvable (see Theorem 2.2 and [W3]).

In studying processes on \( S^1 \), we found Fourier series analysis to be a very useful tool. Also, in the preceding section we mentioned that an analog of Fourier series — spherical harmonics — may prove to be a useful tool in the analysis of processes on \( S^2, S^3 \), and SO(3). This analogy can be extended to devise analytical tools for arbitrary compact Lie groups. The basis for this is the Peter-Weyl Theorem. [L5], [W1], [C1],...
which essentially provides a decomposition of $L_2$ functions on a compact Lie group into a (possibly infinite) sum of very special functions. Fourier series on $S^1$ is the very simplest example of this decomposition. We will not go into the general decomposition procedure here except to make several comments. First of all, this type of function decomposition can also be defined for homogeneous spaces [D10]. This, in fact, is precisely what is done to define spherical harmonics on $S^2$, which can be thought of as the homogeneous space $SO(3)/SO(2)$. Also, this type of function decomposition can also be done on finite groups, and the reader is referred to Chapter 10 in which we consider such a decomposition for the group $Z_n$. It is the hope of the author that the results reported in this manuscript and the work of others, such as Kawada and Ito [K17] and Perrin [P1], will lead to the development of new tools to aid in the study of random processes on groups.

We now turn to the estimation problem. The form of the estimator, illustrated in Figure 7.5, is similar to the filter forms discussed for the $SO(2)$ and $SO(3)$ problems. The nonlinear preprocessor is used if $Z$ or $dZ$ is taken as the observation process. Note that if we make the identification

$$x \leftrightarrow \sum_{i=1}^{n} R_i x_i$$

(7.73)

we can view the Lie algebra $L$ as $R^n$, and thus, as indicated in Figure 7.5, the filtering procedure can be viewed as consisting of a preprocessor mapping processes from the Lie group into the Lie algebra, a filter operating in the Lie algebra, and a postprocessor mapping processes from the algebra into the group.
Fig. 7.5 Form of a General Lie Group Filter
If we wish to estimate \( x(t) \), the optimal filtering process is finite dimensional and can be determined in a manner analogous to that for the SO(2) and SO(3) problems. The filter form is essentially the same as that given in Figure 7.2, except that we replace \( Z' \) by \( Z^{-1} \). In addition, as in the SO(3) case, we can consider a process of the form

\[
M(t) = X(t)V(t)
\]  

(7.74)

If we take \( dM \) as the observation process, we are led to an optimal estimation procedure for \( x \) that is analogous to that depicted in Figure 7.3.

As the SO(3) problem indicated, the problem of estimating \( X \) given by (7.65) or (7.66) is quite difficult. The two ad hoc techniques discussed in the previous section can be directly applied to the general problem; however, as mentioned previously, it is not clear how well these techniques will work and simulations are needed to determine their efficacy. Another possible estimation technique is suggested by the Wei-Norman representation results. Suppose there exists a stochastic analog of the Wei-Norman result for the particular Lie group of interest -- i.e. assume that \( X \) can be written as

\[
X(t) = \exp(g_1(t)A_1)\exp(g_2(t)A_2)\ldots\exp(g_n(t)A_n)
\]  

(7.75)

In this case, one could consider designing a filter to estimate the \( g_i(t) \) directly. We note that in general the relationship between the \( g_i \) and \( x \) is quite nonlinear, and it is not clear that this approach will actually lead to simple estimation equations. However in certain
cases we may obtain easily implemented and perhaps optimal estimation schemes by using such a representation. A particular example that may yield such results is the class of systems discussed in Theorem 7 of [81]. This class appears to be a first step up in complexity from the abelian Lie group system and linear system cases.

Finally, in evaluating the performance of such estimation systems, we need to define estimation criteria. As in the $S^1$ and $SO(3)$ cases, we can define a Riemannian metric on the group of interest and can consider error criteria that are nondecreasing with respect to the metric. Of course, any positive power of the metric itself can serve as an estimation criterion. Also, the function decomposition provided by the Peter-Weyl Theorem may prove to be as useful a tool in such analysis for the general Lie group problem as Fourier series is for the $S^1$ case.
PART II: FINITE GROUPS

CHAPTER 8

FINITE GROUP HOMOMORPHIC SEQUENTIAL SYSTEMS

"The machine yes the machine
never wastes anybody's time
never watches the foreman
never talks back"

- Carl Sandburg [S17]

8.1 Introduction

In Part I of this manuscript we described a class of highly
structured dynamical systems evolving on continuous groups and
used the structure of these systems as an aid in analysing in detail
some of the properties (particularly the stochastic ones) of these
systems. In Part II we shall consider a class of systems on finite
groups and will examine their properties. Unlike the systems of Part I,
the systems described here have not been extensively studied, and we
will first undertake the problem of studying the algebraic structure of
these systems.

Thus, the present chapter will deal strictly with some algebraic
problems. It is hoped that the reader will note the similarity between
the problems treated here and those connected with the algebraic theory
of linear systems [K11], [K12], [B2] and bilinear systems [B1], [W3],
[W4] and also the analogy between the methods of solution. To develop
and exploit the analogous methods of studying systems defined in a variety
of algebraic settings is one of the major motivations for this work.
In that sense, one may view Chapter 8 as a contribution to the study of dynamical systems in the context of universal algebras [C4] and category theory [M1]. For instance, a number of the results and proofs that follow have direct analogs in linear theory, and the proofs are presented to emphasize the universality of these arguments. That is, one should read these results keeping the following in mind. In the theory of algebra there are a few basic isomorphism theorems for groups, rings, vector spaces, etc., and one often obtains the results in one setting from those in another simply by replacing key words with their analogs -- e.g. group for ring and normal subgroup for ideal. The results of this chapter indicate that the same type of universal structure and isomorphism results will hold in a system-theoretic framework. The reader is referred to Goguen [G7] and Arbib and Manes [A7], [A8] for some important results along these lines.

In addition to the desire to understand the key results of linear system theory in a more general setting, another major motivation for this work comes jointly from the theory of finite state machines and from control theory. A central problem in automata theory and control theory is the construction of an "efficient" realization of a given input-output function. The desired realization might be an algorithm that performs a desired computational task or a linear system that realizes a given weighting pattern. In any case, we usually have a notion of "minimality" associated with the realization problem, and we are often asked not just to find any realization but a minimal one. The class of systems developed here is useful in that we can find
simple methods for determining realizations and for "reducing" a
given realization to obtain a minimal one. Also, in Chapter 9
we discuss some tools that may prove useful in determining
efficient (i.e. minimal or near minimal) computational schemes for
certain problems associated with finite state Markov processes.

There is some similarity between the problems of Parts I and II.
The systems of Part II are in some sense "bilinear," since the
control multiplies the state. Also, we note that this "bilinearity"
leads to probability distribution evolution equations that are also
bilinear. In addition to this similarity, the strongest connection
between Parts I and II is in the relationship between \( S^1 \) and \( Z_n \), the
group of integers \( \{0,1,\ldots,n-1\} \) under addition modulo \( n \). Consider
the subset of \( S^1 \) consisting of the elements \( \frac{2\pi k}{n} \) \( n=0,1,\ldots,n-1 \).
This is a subgroup of \( S^1 \) (recall that addition is modulo \( 2\pi \)) and
is clearly isomorphic to \( Z_n \).

The relationship between \( S^1 \) and \( Z_n \) will be explored in Chapter 10,
in which we will describe a "demodulation" problem motivated by and
very similar to the standard \( S^1 \) demodulation problem. In Chapter 9
we will develop the properties of stochastic processes on finite groups,
and as mentioned above, we will consider some computational problems
associated with the stochastic processes. A tie-in with Part I is
established in considering the computational problems of the discrete-
time Fourier series update equations of Section 5.2.

The present chapter provides the foundation on which all of the
subsequent analysis is based. Section 8.2 contains a review of some
of the important results from the theory of linear sequential
circuits -- linear systems defined over finite fields [G5]. In
Section 8.3 we define the class of systems to be studied and discuss
some of the algebraic properties of the systems. The reader is re-
ferred to Appendix E, in which we summarize some of the basic concepts
of realization theory and the theory of finite state machines, to
reference [Bl4], which contains many of the results of Section 8.3,
and to Appendix F, in which we present a result that serves to tie
this class of systems together with a similar class of continuous
group systems and with linear systems. In Section 8.4 we study
some of the more complicated questions related to this class of
systems. For instance we analyze in detail the problem of system
invertibility in such a way as to parallel some of the work of
Massey and Sain [M10], [S10], [M11]. Section 8.5 contains a dis-
cussion of possible application of this type of system to coding
problems. In addition the examples in Chapters 9 and 10 are concerned
with potential applications for the concepts developed in this chapter.

8.2 Some Results from the Theory of Linear Sequential Circuits

In this section we will review some of the important results from
the theory of linear sequential circuits. This discussion is included
for several reasons. First of all, linear theory provides a point of
departure motivating some of the results discussed in Sections 8.3 and
8.4, and it is suggested that the reader keep the concepts introduced
in the present section in mind while reading the following sections.
Also, many of the results of the present section can be obtained by the methods of Sections 8.3 and 8.4, since, as will be seen, linear sequential circuits are a very special subclass of the systems we will study. It is this fact that motivates the coding theory discussion in Section 8.5, since the class of convolutional encoders [M10], [S10], [M11], [M13], [H12], [F4], [V3], [K14], is just a particular subclass of linear sequential circuits.

By a linear sequential circuit (LSC) we mean a linear discrete-time dynamical system over a finite field $K$:

$$x(k+1) = Ax(k) + Bu(k)$$  \hspace{1cm} (8.1)

$$y(k) = Cx(k)$$  \hspace{1cm} (8.2)

where the state $x(k) \in K^n$, the input (control) $u(k) \in K^m$, and the output $y(k) \in K^p$ (here $K^r$ is the $r$-dimensional vector space of $r$-tuples of elements of $K$). Also, $A$, $B$, and $C$ are constant $K$-valued matrices of appropriate dimensions. We could have generalized the definition somewhat if we allowed $A$, $B$, and $C$ to be time-varying, but we will not need to study that case. We remark that the results that we will present also hold for $K = R^1$ and for the continuous-time linear system problem. The reader is referred to [K11], [K12], [B2], [M10], [M11], [M13], and [B16] for proofs of the following results and further discussions of linear system theory.

We first introduce the concepts of controllability and observability. We will do this in a general setting, since we will use these concepts for systems other than linear ones.

**Definition 8.1:** Consider a dynamical system described by the equations
\[ x(k+1) = \lambda(x(k), u(k)) \]  
\[ y(k) = \delta(x(k)) \]

where \( x \in X \), the state set, \( u \in U \), the input (control) set, and \( y \in Y \), the output set. We say that the system is controllable from \( x_1 \in X \) if for any \( x_2 \in X \) there exists a sequence of inputs such that the state is driven from \( x_1 \) to \( x_2 \) by this sequence. The system is said to be controllable if it is controllable from all \( x \in X \). Two states \( x_1, x_2 \in X \) are said to be indistinguishable if, given any input sequence, the corresponding output sequences from the initial states \( x_1 \) and \( x_2 \) are identical. Otherwise \( x_1 \) and \( x_2 \) are said to be distinguishable, and an input sequence that yields different output sequences from \( x_1 \) and \( x_2 \) is said to distinguish between \( x_1 \) and \( x_2 \). We call the system observable if any distinct pair of states is distinguishable.

**Theorem 8.1:** Consider the LSC described by (8.1) and (8.2). The system is controllable if and only if it is controllable from the 0 state, which is equivalent to the condition

\[ \text{rank}[B, AB, \ldots, A^{n-1}B] = n \]  
\[ (8.5) \]

Two states are distinguishable if and only if the 0 input sequence distinguishes between them, and \( x_1 \) is distinguishable from \( x_2 \) if and only if \( x_1 - x_2 \) is distinguishable from 0. Furthermore, the system is observable if and only if

\[ \text{rank}[C', A'C', \ldots, (A^{n-1})'C'] = n \]  
\[ (8.6) \]
For proof of Theorem 8.1 the reader is referred to Kalman [K11], [K12] and Kalman, Ho, and Narandra [K13].

Examining (8.1) and (8.2), we see that we can write

\[
x(k) = \sum_{i=0}^{k-1} A^{k-i-1}Bu(i) + A^{k}x(0)
\]

\[
y(k) = \sum_{i=0}^{k-1} CA^{k-i-1}Bu(i) + CA^{k}x(0) \quad k \geq 1
\]

Thus, if we assume that the initial state is zero, we have that the matrices \( \{CA^iB \mid i \geq 0\} \) completely characterize the input-output map of the system, and, by the Cayley-Hamilton Theorem [H2], we need only consider \( \{CA^iB \mid 0 \leq i < n-1\} \). We may then ask when an input-output map given by

\[
y(k) = \sum_{i=0}^{k-1} T_{k-i-1}u(i)
\]

is of the form (8.8) (with \( x(0) = 0 \)) -- i.e. when does (8.9) have a finite dimensional realization of the form (8.1), (8.2). The matrices \( \{T_i\} \) are called the weighting pattern of the input-output map.

**Theorem 8.2:** An input-output map of the form (8.9) has a realization of the type given by (8.1) and (8.2) if and only if there exists an \( r < \infty \) and a linear map \( p : K^r \rightarrow K \)

\[
p(k_0, k_1, \ldots, k_{r-1}) = a_0k_0 + a_1k_1 + \ldots + a_{r-1}k_{r-1} \quad a_i \in K
\]

such that

\[
p(T_i, T_{i+1}, \ldots, T_{i+r-1}) = a_0T_i + a_1T_{i+1} + \ldots + a_{r-1}T_{i+r-1} = T_{i+r} \quad \forall i \geq 0
\]
Also, if (8.9) has such a realization it has a controllable and observable one, which can be seen to be minimal in the sense of Nerode (see Appendix E).

Any two controllable and observable realizations have state spaces of the same dimension \( n \), and if \((A,B,C)\) and \((F,G,H)\) are the triples that characterize the two minimal realizations, then there exists a nonsingular \( nxn \) matrix \( P \) such that

\[
F = PAP^{-1} \quad G = PB \quad H = CP^{-1}
\]  

(8.12)

We refer the reader to the proof of the existence of a realization given in [B2], where two important realizations are discussed — the standard controllable realization and the standard observable realization. The reader is also referred to the work of Gilbert [G6] and Kalman [K11], [K12].

We now consider the concept of invertibility. Again we will consider the general system described by equations (8.3) and (8.4), and we assume that we are given a particular initial state \( x(0) = x_0 \).

Define the input and output vectors

\[
U_k = \begin{bmatrix} u(0) \\ \vdots \\ u(k) \end{bmatrix} \quad Y_k = \begin{bmatrix} y(1) \\ \vdots \\ y(k) \end{bmatrix}
\]  

(8.13)

**Definition 8.2:** We say that the system (8.3), (8.4) is **invertible with delay** \( L \) or has an inverse of delay \( L \) if \( U_k \) can be recovered from \( Y_{k+L+1} \) \( k > 0 \) — i.e. if \( Y_{k+L+1} \) uniquely specifies \( U_k \).

Note that our definition differs from that of Massey and Sain [M10] in that they define \( L \)-invertibility using \( Y_{k+L} \) as opposed to
and include \( y(0) \) in \( Y_k \); however we note that in our case the system equations (8.3) and (8.4) indicate that the first output influenced by the inputs is \( y(1) \), so we at least need \( y(1) = Y_1 \) to determine \( u(0) = U_0 \). Since determining \( U_k \) from \( Y_{k+1} \) is the best that can be hoped for, we choose to call this a zero delay as opposed to a one delay. Also note that if a system has an \( L \)-delay inverse, it has an \( M \)-delay inverse \( \forall M > L \).

Returning to the linear sequential circuit case, we see that if we assume \( x(0) = 0 \), (8.8) implies

\[
Y_k = M_k U_{k-1}
\]

where

\[
M_k = \\
\begin{bmatrix}
T_0 & 0 & 0 & \ldots & 0 \\
T_1 & T_0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_{k-1} & T_{k-2} & T_{k-3} & \ldots & T_0 \\
\end{bmatrix}
\]

(8.15)

\[
T_1 = CA^1 B
\]

(8.16)

**Theorem 8.3:** Given an LSC described by (8.1) and (8.2) with \( \dim u = m \). Then there exists an inverse with delay \( L \) if and only if

\[
\text{rank}(M_{L+1}) = \text{rank}(M_L) + m
\]

(8.17)

This result is proved by Massey and Sain [M10]. We remark that for a linear system to have any chance of being invertible, the number of outputs must be at least as large as the number of inputs (the reader can check that (8.17) can hold only if \( \dim y = p > m = \dim u \)).
In the general case of (8.3) and (8.4) this translates into the statement that there must be "at least as many" output values as input values -- e.g. if the input set $U$ and output set $Y$ are finite, we must have

$$\text{card } Y \geq \text{card } U$$  \hspace{1cm} (8.18)

where card = cardinality = number of elements. One can regard the case where $Y$ is larger than $U$ in dimension or cardinality as a situation in which we provide the observer with redundant information or parity checks [817].

The next result provides another method for checking the invertibility of a linear system.

**Theorem 8.4:** Consider the LSC of (8.1) and (8.2) with $\text{dim } x = n$. Then the system is invertible with some delay $L$ if and only if it has an inverse with delay $n$ -- i.e. if and only if

$$\text{rank}(M_{n+1}) = \text{rank}(M_n) + n$$  \hspace{1cm} (8.19)

Also, (8.19) holds if and only if

$$\text{rank } N = (n+1)m$$  \hspace{1cm} (8.20)

where

$$N = \begin{bmatrix}
T_0 & 0 & 0 & \cdots & 0 \\
T_1 & T_0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_n & T_{n-1} & T_{n-2} & \cdots & T_0 \\
T_{n+1} & T_n & T_{n-1} & \cdots & T_1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_{2n-1} & T_{2n-2} & T_{2n-3} & \cdots & T_{n-1}
\end{bmatrix}$$  \hspace{1cm} (8.21)
For proof of this result the reader is referred to Brockaet and Mesarovic [B16] and Sain and Massey [S10].

The preceding result provides a bound on how much computation we must perform to find out if a LSC is invertible — i.e. if a LSC is not n-invertible, where n is the dimension of the state space, it is not invertible. We will now present a result that tightens this bound. In addition to its importance in the theory of linear systems, this result indicates the type of counting arguments that will be so important in the next few sections. We use an argument from [S10] as the motivation. Suppose the linear system (8.1), (8.2) has no inverse with delay L. Then it has no inverse with delay < L. Thus

\[ \text{rank } M_{L+1} < \text{rank } M_L + (m-1) < \text{rank } M_{L-1} + 2(m-1) < \ldots \]
\[ < (L+1)(m-1) \]  \hspace{1cm} (8.22)

(where we have used (8.17) recursively, with the convention \( \text{rank } M_0 = 0 \). What this means is the following: consider the input vector space \( U = K^m \); the map \( M_{L+1} \) maps \( U^{L+1} \) into \( Y^{L+1} \) and (8.22) states that the nullspace \( \mathcal{N}_{L+1} \subset U^{L+1} \) of \( M_{L+1} \) has dimension \( d_{L+1} \), which we can bound:

\[ d_{L+1} = \text{dim } U^{L+1} - \text{rank } M_{L+1} > (L+1)m - (L+1)(m-1) = L+1 \]  \hspace{1cm} (8.23)

Now we consider the application of another input \( u(L+1) \) while we restrict the choices of \( u(0), \ldots, u(L) \) to lie in \( \mathcal{N}_{L+1} \). The dimension of this restricted input space is \( d_{L+1} + m \). Consider the map that sends \( u(0), \ldots, u(L+1) \) into \( x(L+2) \) (see (8.7) and recall that we require
\( x(0) = 0 \). The range space of this map is \( X = \mathbb{R}^n \), and, therefore, if we have the condition

\[
d_{L+1+m} \geq n+1
\]  

(8.24)

this map must have a nullspace of dimension at least one. Thus there exist \( u(0), \ldots, u(L+1) \) not all zero such that (since \( u(0), \ldots, u(L) \) is \( N_{L+1} \)) \( y(1) = y(2) = \ldots = y(L+1) = 0 \) and \( x(L+2) = 0 \). Clearly if this is so, then the input sequence \( u(0), u(1), \ldots, u(L+1), 0, 0, \ldots \) cannot be distinguished from the all 0 sequence and our system is not invertible with any delay. Examining (8.23) and (8.24), we see that (8.24) will hold -- i.e. our system will not be invertible -- if

\[
L \geq n-m
\]  

(8.25)

and we have proven

**Theorem 8.5:** Consider the linear system (8.1), (8.2) with \( \dim u = m \), \( \dim x = n \). The system is invertible if and only if it is invertible with delay \( L \leq n-m \) -- i.e. if and only if

\[
\text{rank } M_{n-m+1} = \text{rank } M_{n-m} + m
\]  

(8.26)

which is true if and only if we have the following condition: given \( u(0), \ldots, u(n-m+1) \), such that the output of the system (started at \( x(0) = 0 \)) in response to this string followed by all 0's is identically 0, then \( u(0) = \ldots = u(n-m+1) = 0 \).

We remark that if one considers an output equation with direct feedthrough -- i.e. \( y(k) = Cx(k) + Du(k) \) -- and if we let \( q = \text{dimension of the nullspace of } D \), we can show that the system must have an inverse
with delay \(s^{n-q}\). Theorem 8.5 is a special case of this result
\((D = 0 \text{ implies } q = m)\).

**Corollary 1:** The linear system (8.1), (8.2) is invertible if and
only if

\[
\text{rank } P = (n-m+2)m
\]  

(8.27)

where

\[
\begin{bmatrix}
CB & 0 & \ldots & 0 \\
CAB & CB & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
CA^{n-m+1} & CA^{n-m} & \ldots & CB \\
CA^{n-m+2} & CA^{n-m+1} & \ldots & CAB \\
\vdots & \vdots & \ddots & \vdots \\
CA^{2n-m} & CA^{2n-m-1} & \ldots & CA^{n-1}B
\end{bmatrix}
\]  

(8.28)

**Proof:** The proof is a direct analog of the proof of Theorem 3 in [S10].

The next corollary produces a condition that may be easier to
check.

**Corollary 2:** The linear system is invertible if and only if

\[
\text{rank } M_{n-m+1} \geq (n-m+1)(m-1) + 1
\]  

(8.29)

**Proof:** This follows from Theorem 8.3, Theorem 8.5, and an argument
analogous to the one used to derive equation (8.22).

An example will indicate why the result stated in Theorem 8.5 makes
sense. Consider the case \(n=m\). Theorem 8.5 states that an inverse
exists if and only if a zero-delay inverse exists. There are two subcases
to consider, depending on the invertibility of \(B\). If \(B\) is not invertible,
there exists a \( u \neq 0 \) such that \( Bu = 0 \), and the system is not invertible. If \( B \) is invertible and the system is not invertible with delay 0, this means there exists \( u(0) \neq 0 \) such that

\[
CBu(0) = 0 
\]  
(8.30)

If we then choose

\[
u(1) = -B^{-1}ABu(0)
\]  
(8.31)

we see that

\[
x(2) = ABu(0) + Bu(1) = 0
\]  
(8.32)

and the system is not invertible.

Note that Theorem 8.5 requires that \( n > m \), which makes sense since \( n < m \) implies the existence of \( u \neq 0 \) such that \( Bu = 0 \).

8.3 A Class of Systems Evolving on Finite Groups

One of the most important problems in system theory is the synthesis problem: given a desired input-output function, design a sequential system that realizes it. Often we seek not just any realization but one that has certain "desirable" properties (see Appendix E).

Thus it becomes important to determine what system structures have "desirable" properties. In the next two sections we define a class of finite state systems with more structure than arbitrary automata but less than linear systems, discuss the types of input-output maps that can arise in this manner, and analyze the structure of these systems in great detail. In this section we present some basic results analogous
to those of linear system theory, and in the next we use some group-
number-theoretic arguments to tighten some of the numerical bounds
that we will discuss, in addition to considering the invertibility
of these systems. It is hoped that the reader will note the analogies
between the results of this section and those of Section 8.2. We
have not worked for the weakest hypothesis for each individual result,
but rather have sought to place all the results in a common framework —
one motivated by linear theory.

**Definition 8.3:** Let \( \mathcal{X} = (X, \ast) \), \( \mathcal{U} = (U, \circ) \), and \( \mathcal{Y} = (Y, \cdot) \) be finite
state, input, and output groups, respectively. Consider the dynamical
system

\[
x(k+1) = b[u(k)] \ast a[x(k)] \\
y(k) = c[x(k)]
\]

where \( a : X \ast X \), \( b : U \ast X \), and \( c : X \ast Y \) are group homomorphisms.

Invoking an analogy with LSC's, we call this system a **finite group homomorphic sequential system** (FGHSS).

From now on, except where ambiguous, we will refer to \( X \), \( U \), and
\( Y \) as groups (as opposed to \( \mathcal{X} \), \( \mathcal{U} \), and \( \mathcal{Y} \)), and we will not write the
group operations explicitly (e.g. we will write \( x_1 x_2 \) for \( x_1 \ast x_2 \)).

Note that a LSC is a FGHSS. Consider (8.1) and (8.2). The sets
\( K^m \), \( K^n \), and \( K^p \) are finite groups with vector addition as the group
operation, and the linear maps \( A \), \( B \), and \( C \) are group homomorphisms —
i.e. a LSC can be regarded as a FGHSS by forgetting the operation of
scalar multiplication on \( K^m \), \( K^n \), and \( K^p \). Thus, the results we present
here imply results for LSC’s, but they may not be quite as strong, since
we throw away some structure in regarding a LSC as a FCHSS.

However the larger class of FCHSS's has much in common with LSC's.

The most obvious result is the analog of (8.7) and (8.8).

**Theorem 8.6:** The input, state, and output of the FCHSS (8.33), (8.34) are related by

\[
x(k) = b[u(k-1)] \ast a[b[u(k-2)]] \ast \cdots \ast a^{k-1}[b[u(0)]] \ast a^k[x(0)]
\]

\[
\triangleq \left\{ \prod_{i=0}^{k-1} a^{k-1-i}b[u(i)] \ast a^k[x(0)] \right\}
\]

(8.35)

\[
y(k) = c[b[u(k-1)]] \cdot c[a[b[u(k-2)]]] \cdot \cdots \cdot c[a^{k-1}[b[u(0)]]] \cdot a^k[x(0)]
\]

\[
\triangleq \left\{ \prod_{i=0}^{k-1} c[a^{k-1-i}b[u(i)]] \cdot c^k[x(0)] \right\}
\]

(8.36)

where \( a^i \) denotes \( i \) compositions of the function \( a \) with itself and

\( ca^i b(u) = c[a^i[b(u)]] \).

**Proof:** The result follows directly from the system equations and the fact that \( a \) and \( c \) are homomorphisms.

The product forms (8.35) and (8.36) will be called "convolution products," in analogy with the convolution sums (8.7) and (8.8). Note that the composition of homomorphisms is itself a homomorphism. Thus \( ca^i b \) is a homomorphism, and, if \( x(0) = e \) (the identities of \( U, X \) and \( Y \) are all called \( e \)), the homomorphisms \( \{ ca^i b | i \geq 0 \} \) (where \( a^0 \triangleq \) identity map) completely determine the input-output map. Then one may ask when a sequence of homomorphisms \( T_i : U \rightarrow Y \) is of the form \( ca^i b \), for some finite group \( X \). The next result answers this question much as Theorem 8.2 answered the linear analog. Here we replace the linear recursion (8.10), (8.11) with a homomorphic recursion and use
the finiteness of $X$, rather than the Cayley-Hamilton Theorem, to show that \( \{ca^1b^i| i \geq 0\} \) is determined by \( \{ca^ib^i| 0 < i < N\} \) for some $N$.

Before presenting this result, we must introduce another concept. Let $U$ and $Y$ be finite groups and let $F(U,Y)$ be the finite set of maps of $U$ into $Y$. We make $F(U,Y)$ a semigroup by defining the operation

\[
(f \circ g)(u) \triangleq f(u)g(u) 
\]

(8.37)

where the multiplication on the right takes place in $Y$ (note this is not the function composition operation; there will be no ambiguity in the rest of the section, since we will use (8.37) only in the next theorem). Let $Y^r = Y \times \ldots \times Y$ ($r$ factors) and put the direct product group structure on it. Suppose $\pi : Y^r \to Y$ is a homomorphism. Then $\pi$ induces a homomorphism $\hat{\pi}$ of $(F(U,Y))^r$ (direct product semigroup) into $F(U,Y)$:

\[
\hat{\pi}(f_1, \ldots, f_r)(u) \triangleq \pi(f_1(u), f_2(u), \ldots, f_r(u)) \quad u \in U \quad f_i \in F(U,Y)
\]

(8.38)

**Theorem 8.7**: Let $U$ and $Y$ be finite groups. Given a sequence of homomorphisms $T_i : U \to Y$, $i = 0, 1, 2, \ldots$, there exists a finite group $X$ and group homomorphisms $a : X \to X$, $b : U \to X$ and $c : X \to Y$ such that

\[
T_i = ca^ib
\]

(8.39)

if and only if there is an integer $r > 0$ and a homomorphism $\hat{\pi} : Y^r \to Y$ such that

\[
\hat{\pi}(T_i, \ldots, T_{i+r-1}) = T_{i+r} \quad \forall \ i \geq 0
\]

(8.40)
**Proof** (Sufficiency): Suppose such an $r$ and $p$ exist. We construct the analog of what has, in the context of linear system theory, been called the standard observable realization [B2]. We take $X = Y^r$ (direct product structure) and consider the map $a : Y^r \to Y^r$ defined by

$$a(x_1, x_2, \ldots, x_r) = (x_2, x_3, \ldots, x_r, p(x_1, x_2, \ldots, x_r))$$  \hspace{1cm} (8.41)

This is a homomorphism if $p$ is. Also define $b : U \to Y^r$ by

$$b(u) = (T_0(u), T_1(u), \ldots, T_{r-1}(u))$$  \hspace{1cm} (8.42)

This is a homomorphism because the $T_i$'s are. Define $c : Y^r \to Y$ by

$$c(y_1, \ldots, y_r) = y_1$$  \hspace{1cm} (8.43)

and this also is a homomorphism. We claim that (8.39) holds. This is true because of the recursion $\hat{p}$:

$$cb(\cdot) = c(T_0(\cdot), T_1(\cdot), \ldots, T_{r-1}(\cdot)) = T_0(\cdot)$$  \hspace{1cm} (8.44)

$$cab(\cdot) = c(T_1(\cdot), T_2(\cdot), \ldots, T_{r-1}(\cdot), \hat{p}(T_0, T_1, \ldots, T_{r-1})(\cdot))$$

$$= c(T_1(\cdot), T_2(\cdot), \ldots, T_r(\cdot)) = T_1(\cdot)$$  \hspace{1cm} (8.45)

and the rest of the relations follow in a similar manner by applying (8.40).

**Proof** (Necessity): Suppose $T_1 = ca^i b$ for some set of homomorphisms $(a, b, c)$, with $a : X \to X$ being defined on a finite group. Since the set of all maps of a finite set into itself is itself a finite set, we must have $a^r = a^k$ for some $r > k > 0$. Then $a^{r+i} = a^{k+i}$ if $i > 0$, and, defining $p$ as the projection onto the $(k+1)$st component of an $r$-tuple
\[ p(y_0, \ldots, y_{r-1}) = y_k \]  

we see that

\[ \hat{p}(T_0, T_1+1, \ldots, T_{i+r-1}) = T_{i+k} = ca^{i+k}b = ca^{i+r}b = T_{i+r} \]  

(8.47)

We note that this result, in the case in which all groups are abelian, has been proven in [A9]. We also remark that the proof of Theorem 8.7 shows that the only sequences \( \{T_i\} \) that can be realized by a FGHSS are those that are periodic after a finite number of terms (see Figure 8.1). The next result shows a \( a : X \to X \) is an automorphism if and only if there is no "tail" (i.e. \( k \) from the necessity proof above can be taken to be zero; see Fig. 8.1).

**Corollary:** Under the hypotheses of Theorem 8.7, there exists a realization with \( a : X \to X \) an automorphism if and only if \( T_{i+r} = T_i \) for some \( r \) and \( \forall \ i > 0 \).

**Proof:** This follows from the fact that \( a \) is an automorphism if and only if \( a^r \) is the identity map for some \( r > 0 \).

In automata theory (see Appendix E) one usually considers systems described by maps of the form \( f : U^* \to Y \) where \( U^* \) is the set of all finite strings of elements of \( U \) and \( f(u_0, \ldots, u_{n-1}) \) is the output of the system at time \( n \) following the application of the input string \( u_0, \ldots, u_{n-1} \) (in this order). One can then ask which \( f \)'s come from FGHSS's.

**Theorem 8.8:** Given finite groups \( U \) and \( Y \), an input-output map \( f : U^* \to Y \) can be realized as a FGHSS if and only if \( T_i : U \to Y \) defined by
Figure 8.1 Illustrating the Realizability Condition
\[ T_1(u) = f(u, e^i) \]  \hspace{1cm} (8.48)

where \( e^i \) is the string if \( i \) identity inputs, are homomorphisms satisfying the conditions of Theorem 8.7, and

\[ f(u_o, \ldots, u_{n-1}) = T_0(u_{n-1})T_1(u_{n-2}) \cdots T_{n-1}(u_o) \]  \hspace{1cm} (8.49)

**Proof:** The result follows from a straightforward calculation.  

Note that the second condition in Theorem 8.8 is equivalent to the following: if \( w_1, w_2 \in U^* \) and the length of \( w_2 \) is \( k \), then

\[ f(w_1, w_2) = f(w_2) \cdot f(w_1, e^k) \]  \hspace{1cm} (8.50)

Also, one should think of the map from \( U^r \) into \( Y^r \) given by

\[ y_r = f(u_o, \ldots, u_{r-1}) = T_0(u_{r-1})T_1(u_{r-2}) \cdots T_{r-1}(u_o) \]
\[ y_{r+1} = f(u_o, \ldots, u_{r-1}, e) = T_1(u_{r-1})T_2(u_{r-2}) \cdots T_r(u_o) \]
\[ \vdots \]
\[ y_{2r-1} = f(u_o, \ldots, u_{r-1}, e^{r-1}) = T_{r-1}(u_{r-1})T_r(u_{r-2}) \cdots T_{2r-2}(u_o) \]  \hspace{1cm} (8.51)

as being the analog of the map corresponding to the Hankel matrix. As will be shown, the number of elements in the image space of this map equals the number of states in the "minimal realization," just as the rank of the Hankel matrix determines the dimension of the state space of a minimal linear realization (see [B2], [K11]).

It is also appropriate to make a remark about superposition. Given \( f : U^* \to Y \) that satisfies Theorem 8.8, we may ask the following: if we have two input sequences \( (u_o, \ldots, u_k) \) and \( (v_o^i, \ldots, v_k^i) \), is the output corresponding to the product input sequence \( (u_o v_o, u_1 v_1, \ldots, u_k v_k) \)
the product of the outputs corresponding to the two sequences? That
is, does the following "superposition equation" hold?

\[ f(u_o v_o, \ldots, u_k v_k) = f(u_o, \ldots, u_k) f(v_o, \ldots, v_k) \] (8.52)

As will become clear, the answer is, in general, no because the input-
output map (8.51) need not be a homomorphism. However, if \( Y \) is an
abelian group, and if we write the \( Y \) group operation as \( \cdot \), we have
from (8.49) and the fact that the \( T_i \) are homomorphisms

\[
\begin{align*}
  f(u_o v_o, \ldots, u_k v_k) &= T_0(u_k v_k) \cdots T_k(u_o v_o) \\
    &= (T_0(u_k) \cdots T_k(u_o)) + (T_0(v_k) \cdots T_k(v_o)) \\
    &= f(u_o, \ldots, u_k) + f(v_o, \ldots, v_k) 
\end{align*}
\] (8.53)

Therefore the superposition equation holds if \( Y \) is abelian. The reader
is referred to [02] for other comments on generalizing the notion of
superposition.

One of the crucial results of linear system theory is Theorem 8.2. We
will now prove an analog of part of this result -- the relationship
between minimality and controllability and observability and the isomor-
phism of minimal realizations -- and will show that the analog of the
other part -- the existence of a minimal realization as a FGHSS -- is not
true. We then will characterize those systems which do have minimal
FGHSS realizations. (Theorem 8.13).

We refer the reader to the definitions of controllability and
observability given in Definition 8.1. The next two theorems are
analogs of the results summarized in Theorem 8.1.
Theorem 8.9: Consider the FGSS (8.33), (8.34) with state group X. The system is controllable if and only if it is controllable from the identity state e. States \( x_1 \) and \( x_2 \) are distinguishable if and only if the identity control sequence distinguishes between them. Also, \( x_1 \) is indistinguishable from \( x_2 \) if and only if \( x_1^{-1}x_2 \) is indistinguishable from e.

Proof: These results are obtained by straightforward computation. We include only the least straightforward proof -- the proof that if the system is controllable from e, it is controllable. Assuming that the system is controllable from e, we will have system controllability if we can show that we can drive the system to e from any state x. Therefore, choose any \( x \in X \). We know that there exist \( N > M > 0 \) such that

\[
a^{N+k} = a^{M+k} \quad \forall \quad k > 0
\]

(8.54)

Then, one can readily see that

\[
a^{N+k(N-M)} = a^M \quad \forall \quad k > 0
\]

(8.55)

Suppose the input sequence \( v_1, \ldots, v_p \) drives the system from e to \( [a^M(x)]^{-1} \). Let

\[
r = p \mod (N-M) \quad 0 \leq r < N-M
\]

(8.56)

and consider the input sequence

\[
u(i) = e \quad 0 < i < 2N-M-r-1
\]

(8.57)

\[
u(2N-M-r+i-1) = v_i \quad 1 \leq i \leq p
\]

If we let \( x(0) = x \), we compute

\[
x(2N-M-r+p) = e
\]

(8.58)
Theorem 8.10: If the FCHSS (8.33), (8.34) evolves in a state group with \( n \) elements, then the set of states reachable from the identity is

\[
R = \{ b(u_{n-1}) a(b(u_{n-2}) \cdots a^{n-1} b(u_{o}) | u_{o}, \ldots, u_{n-1} \in U \} = b(U) a(U) \cdots a^{n-1} b(U) \tag{8.59}
\]

The set of states indistinguishable from the identity is

\[
K = \ker c \cap \ker ca \cap \cdots \cap \ker ca^{n-1} \tag{8.60}
\]

The set \( R \) is not necessarily a group, but \( K \triangleleft X \).

Proof: With respect to \( R \), the result is immediate from the formula (8.35) and the observation that, because of the stationarity of the system, any state reachable from the identity is reachable along a trajectory that contains no state more than once, and thus is of length \( \leq n \).

Turning to \( K \) and using Theorem 8.9, we see that for a state \( x \) to be indistinguishable from \( e \) the response of the system, started in \( x \), to the identity input string must be all identities. That is,

\[
c(x) = ca(x) = \ldots = ca^{n-1}(x) = e \tag{8.61}
\]

Can it happen that this set of equalities holds but \( ca^p(x) \neq e \) for some \( p \geq n \)? Clearly not, since for any \( x \), \( a_i^j(x) = a_i^k(x) \) for some \( n \geq i > j \geq 0 \) because there are only \( n \) elements in \( X \). This means that \( a^p(x) = a^k(x) = e \) where \( 0 \leq k \leq n-1 \), and (8.60) is correct.

To see that \( K \) is a normal subgroup, we need only observe that the map \( C : X \to Y^n \) (direct product)

\[
C(x) = (c(x), ca(x), \ldots, ca^{n-1}(x)) \tag{8.62}
\]
is a homomorphism and \( \ker C = K \). That \( R \) need not be a subgroup of \( X \) will be shown in Example 8.1.

**Corollary:** Under the hypothesis of Theorem 8.10, \( R \) is a subgroup of \( X \) if \( X \) is abelian.

**Proof:** We need only note that \( m \geq 0, a^m b(U) \) is a subgroup, and the product of two subgroups of an abelian group is itself a subgroup.

We note that these results are of the same type as those in Theorem 8.1. For instance, for vector spaces, the kernel of a linear map is its nullspace. Thus (8.60) is saying that the linear system is observable if and only if for some \( p \), the intersections of the nullspaces of \( C, CA, CA^2, \ldots, CA^p \) is \( \{0\} \) which is the same type of statement as (8.6). We also remark that the bounds in (8.59) and (8.60) — i.e. checking out to \( a^{n-1}b \) or \( ca^{n-1} \) where \( |X| = n \) — are very weak.

Note that (8.6) states that we need only check up to \( p-1 \) where \( p \) is the dimension of the state space \( X \). If the base field \( K \) has \( q \) elements, then \( n = |X| = q^p \), so we have that the bounds of (8.59) and (8.60) are exponentially larger than those of (8.5) and (8.6). In the next section we will see how to improve these bounds.

We now recall some of the concepts of abstract realization theory [K11]. If \( A \) and \( B \) are sets and \( f : A \to B \), a **factorization of \( f \) through a set** \( C \) is a pair of maps \( \alpha : A \to C \) and \( \beta : C \to B \) such that \( f = \beta \circ \alpha \), i.e., the diagram of Figure 8.2 commutes. The factorization is **canonical** if \( \alpha \) is onto and \( \beta \) is one-to-one. In this case, the "size" of \( C \) is minimal in some sense. For instance, if \( A, B, \) and \( C \) are vector
spaces and $f$, $\alpha$, and $\beta$ are linear maps, then for any other not necessarily canonical factorization $\hat{C}$, $\hat{A}$, $\hat{B}$, $\dim C \leq \dim \hat{C}$. Also if $A$, $B$, and $C$ are finite, with $C$ corresponding to a canonical factorization, then, if $\hat{C}$ corresponds to any other factorization, card $C \leq \text{card } \hat{C}$.

Suppose we have finite input and output groups $U$ and $Y$ and an input output map $f : U^* \rightarrow Y$ that has a realization as a FCHSS (8.33), (8.34) with a state group of $n$ elements. Define $F : U^* \rightarrow Y^n$ by

$$F(u_0, \ldots, u_k) = (f(u_0, \ldots, u_k), f(u_0, \ldots, u_k, e), \ldots, f(u_0, \ldots, u_k, e^{n-1}))$$

(8.63)

Then we have the factorization of $F$ given in Figure 8.3,
Figure 8.3: A Factorization of a FGHSS Input-Output Map

where we define $C$ as in (8.62) and

$$B(u_0, \ldots, u_k) = b(u_k)a^2(u_{k-1}) \cdots a^k b(u_0)$$  \hspace{1cm} (8.64)$$

We immediately see that the above factorization is canonical if and only if the FGHSS is controllable and observable, and it is clear that, referring to Appendix E, this is equivalent to minimality in the sense of Nerode. In this case we say that the triple of homomorphisms $(a, b, c)$ defines a minimal realization of $f$.

Another result from abstract realization theory is the following: given $f : A \rightarrow B$ and two canonical factorizations -- that is, two sets $C$ and $\hat{C}$ and corresponding maps $\alpha : A \rightarrow C$, $\hat{\alpha} : A \rightarrow \hat{C}$, both onto and $\beta : C \rightarrow B$, $\hat{\beta} : \hat{C} \rightarrow B$, both one-to-one, such that $f = \beta \circ \alpha = \hat{\beta} \circ \hat{\alpha}$ -- then the two are equivalent, in that there exists a unique one-to-one and onto map $\gamma : C \rightarrow \hat{C}$ such that $\hat{\alpha} = \gamma \circ \alpha$ and $\beta = \hat{\beta} \circ \gamma$. When we apply this result to the class of FGHSS's, we obtain stronger results, as in
linear theory (see Theorem 8.2), because of the structure of the systems.

Theorem 8.11: Suppose U and Y are finite groups and \( f: U^* \rightarrow Y \) is an input-output map that has two controllable and observable FGHSS realizations

\[
\begin{align*}
  x(k+1) &= b[u(k)]a[x(k)] & y(k) &= c[x(k)] \\
  z(k+1) &= g[u(k)]d[z(k)] & y(k) &= h[z(k)]
\end{align*}
\]  

(8.65) (8.66)

where system (8.65) evolves in a finite state group \( X \) and system (8.66) in a finite state group \( Z \). Then there exists a group isomorphism \( p: X \rightarrow Z \) such that \( d = p a p^{-1} \), \( g = p b \), and \( h = c p^{-1} \). The two realizations are said to be conjugate.

Proof: Suppose the cardinality of \( X \) is \( n \). Then the same is true of \( Z \) by the comments preceding the theorem. Let \( F: U^* \rightarrow Y^n \), \( B: U^* \rightarrow X \), and \( C: X \times Y^n \) (direct product group) be defined by (8.62), (8.64), and (8.62), respectively, and define \( G: U^* \rightarrow Z \) and \( H: Z \rightarrow Y^n \) by

\[
G(u_0, \ldots, u_k) = g(u_k)dg(u_{k-1}) \ldots d^kg(u_0) \quad (8.67)
\]

\[
H(z) = (h(z), hd(z), \ldots, h d^{n-1}(z)) \quad (8.68)
\]

Then, by controllability and observability, we have two canonical factorizations of \( F \), and, by the comments preceding the theorem (see also [Z1], [K11]), we have the commutative diagram of Figure 8.4 where \( p \) is the unique one-to-one and onto map such that the diagram remains commutative. Let \( x_1, x_2 \in X \). Then we have

\[
H[p(x_1x_2)] = C(x_1x_2) = C(x_1)C(x_2) = H[p(x_1)]H[p(x_2)] = H[p(x_1)p(x_2)]
\]

(8.69)
Since $H$ is one-to-one, $p(x_1 x_2) = p(x_1)p(x_2)$ -- i.e. $p$ is an isomorphism. It is then a simple computation to arrive at the desired relations between $(a, b, c)$ and $(d, g, h)$.

We note that finite dimensional vector spaces over the same field are isomorphic if and only if they are of the same dimension, whereas finite groups can have the same number of elements and not be isomorphic. Thus this state space isomorphism theorem is decidedly more interesting. Note that in the proof of Theorem 8.11, the group structure of $U$ is never used; however the group structure of $Y$ and the fact that $C$ and $H$ are one-to-one homomorphisms are used to show that $p$ is an isomorphism. This lack of symmetry in the arguments will be discussed later in this section.

As was mentioned in Theorem 8.10, the set $R$ of states reachable from the identity need not be a subgroup. Thus, given a FGHSS, there
need not exist another FGHSS with the same input-output map for which every state is reachable from the identity. One then could consider the problem of finding the FGHSS realization with the "smallest number" of states that can't be reached from e. As shown by Arbib [A5], this system can be obtained from the original realization by taking as the state group \(<R>\) — the smallest subgroup of X that contains R.

One might expect that a FGHSS has a minimal realization as a FGHSS if and only if in any realization the set R is a subgroup. Example 8.1 shows that this need not be the case. If R is a subgroup in any realization, we can restrict our homomorphisms to R modulo the kernel of \(C : X \rightarrow Y^N\), and thus construct a controllable and observable FGHSS (a simple check shows that one can redefine the homomorphisms in a well-defined manner after "extracting" ker C — therefore there always exists an observable FGHSS realization). Thus, for example, if there exists a FGHSS realization with an abelian state group, there exists a minimal FGHSS realization. Also, it is easy to see that there exists a minimal FGHSS realization if and only if R is a subgroup in any observable realization. Thus, as Example 8.1 indicates, one should first find an observable realization before drawing any conclusions about controllability and minimality.

**Example 8.1**: Consider the dihedral group \(D_4 = \{e, x, x^2, x^3, xy, x^2y, x^3y\}\) with the relations

\[
x^4 = e \quad y^2 = e \quad xyx = y
\]

(8.70)

Also consider the group \(Z_2 = \{0, 1\}\) with addition mod 2 (see Appendix A for more details on these groups). Consider the FGHSS (8.33), (8.34)
with $U = Y = Z_2$, $X = D_4$, and $a$, $b$, and $c$ being the homomorphisms uniquely determined by

$$b(1) = y$$ (8.71)

$$a(x) = e \quad a(y) = xy$$ (8.72)

$$c(x) = 0 \quad c(y) = 1$$ (8.73)

The set of states reachable from $e$ may be shown to be

$$R = \{e, y, xy, x^3\}$$ (8.74)

which is not a subgroup.

However, if we compute the input-output homomorphisms $T_i = ca^i b$, we find that

$$T_i = \text{identity map} \quad \forall \quad k \geq 0$$ (8.75)

and we see that although the above nonminimal realization has an identity-reachable set that is not a group, there still exists a minimal FGHSS realization. In fact, such a realization is obtained by taking $U = X = Y = Z_2$ and $a = b = c = \text{identity map}$. The reason we can find such a realization is that our original system is not observable. An example of an observable system for which $R$ is not a group is found by taking $U = Z_2$, $X = Y = D_4$, letting $a$ and $b$ be as in (8.72) and (8.71), and taking $c = \text{identity (i.e. full state output)}$. This is observable and $R$ is given by (8.74).

There are conditions under which $R$ is a subgroup, in which case we do have a minimal FGHSS realization. The following theorem indicates one such condition (see Theorems 8.14 and 8.15 and [A5] for others).
**Theorem 8.12**: Under the hypotheses of Theorem 8.10, the set $R$ of states reachable from $e$ is a subgroup if $a$ is an automorphism.

**Proof**: The set of automorphisms of a finite group is itself a finite group with function composition as the group operation. Thus there exists a $k > 0$ such that

$$a^k = \text{identity map} \quad (8.76)$$

From Theorem 8.6 we see that $R$ can be written in the form

$$R = \bigcup_{m \geq 0} \prod_{i=0}^{m} a^{m-i}b(U)$$

$$= \bigcup_{m \geq 1} [b(U)ab(U)\ldots a^{k-1}b(U)]^m \quad (8.77)$$

where for $H \subseteq X$

$$H^m \triangleq \{h_1h_2\ldots h_m | h_i \in H\} \quad (8.78)$$

and from Theorem A.3 and equation (A.1) we see that

$$R = \langle b(U)ab(U)\ldots a^{k-1}b(U) \rangle \quad (8.79)$$

i.e. it is the subgroup generated by $b(U)ab(U)\ldots a^{k-1}b(U)$.

The next theorem completely characterizes those sequences of input-output homomorphisms that have controllable and observable FCHSS realizations. To do this we must define what we mean by a free response of a system. If a system is given in recursive form (as (8.33), (8.34)), a free response is the identity input response of the system from some initial state. If the system is given in input-output form, it is the response to an input sequence that consists of the identity only from some point onward, where the response is observed from the point in
time where the nonidentity inputs stop. Thus we apply a (possibly) nonidentity input up to some time k and record the output from time k+1 on. Note that the set of free responses of an input-output map corresponds to the set of free responses of a FGHSS realization started in a state reachable from the identity state. In what follows, free responses refer to the input-output system description. Note that we can consider the set of free responses as a subset of the infinite direct product group $Y \times Y \times \ldots \times Y \times \ldots$.

**Theorem 8.13:** Let the sequence of homomorphisms $T_i : U \to Y$, $i=0,1,2,\ldots$, with U and Y finite groups, satisfy the hypotheses of Theorem 8.7. Then there exists a minimal FGHSS realization if and only if the set of free responses form a subgroup of the infinite direct product group.

**Proof (Sufficiency):** Let $F$ be the group of all free responses. Let $F_n$ be defined by

$$F_n = \begin{cases} (y_0, \ldots, y_{n-1}) \in Y^n \text{ (direct product)} & \text{the } v_i \text{ are the first } n \\ \text{elements of a free response } \in F & \end{cases}$$

(8.80)

Obviously $F_n < Y^n$ if $F$ is a subgroup of the infinite direct product.

Consider the standard observable realization given in the proof of Theorem 8.7. In that realization, the state space is $Y^r$, and one can check that the set R of states reachable from $e$ is just $F_r$. Then, restricting all the homomorphisms to $F_r$, we have a minimal homomorphic realization.

**Proof (Necessity):** Suppose we have a minimal FGHSS realization of the $T_i$. Since every state is reachable from $e$, the set of free responses in the
input-output sense is identical to the set of free responses in the state space sense. Consider the map \( C_\infty : X \rightarrow Y \times Y \times \ldots \times Y \times \ldots \) given by

\[
C_\infty(x) = (c(x), ca(x), \ldots, ca^k(x), \ldots)
\]  

(8.81)

This is a homomorphism, and its image is \( F \), which therefore must be a group.

**Corollary:** Under the hypotheses of Theorem 8.13, if \( F \) is a group, \( F \) is isomorphic to \( F_k \) for some \( k \).

**Proof:** Suppose \( a \) is the state transition homomorphism for a minimal realizations. Then there exist \( k > p > 0 \) such that \( a^k = a^p \), and then

\[
(c(x), ca(x), \ldots, ca^l(x), \ldots) = (c(x), ca(x), \ldots, ca^{k-1}(x), ca^p(x), \ldots, ca^{k-1}(x), ca^p(x), \ldots)
\]  

(8.82)

and the isomorphism is obvious. Note that even if \( F \) is not a group, there exists a \( k \) such that the elements of \( F \) and \( F_k \) are in one-to-one correspondence.

We have raised a number of questions in this section. We have derived the standard observable realization; what about a standard controllable realization? The set of states indistinguishable from \( e \) is a (normal) subgroup; why isn't the set of states reachable from the identity necessarily a subgroup? In Theorem 8.11 we used the fact that \( C \) and \( H \) are homomorphisms; what about \( B \) and \( G \)? We have seen that \( R \) need not be a group, and, for similar reasons, \( B \) and \( G \) are not homomorphisms, and there is no standard controllable realization.

Note that these difficulties arise from the following consideration.
Suppose we have a set of homomorphisms \( c_i, i=1, \ldots, n \), mapping a finite group \( X \) into a finite group \( Y \). Then the "fan-out" map taking \( X \) into \( Y^n \)

\[
\begin{align*}
\begin{array}{c}
x \\
\end{array} & \longrightarrow (c_1(x), c_2(x), \ldots, c_n(x)) \\
\end{align*}
\]

(8.83)

is always a homomorphism, but the "fan-in" map taking \( X^n \) into \( Y \)

\[
\begin{align*}
(x_1, x_2, \ldots, x_n) & \longmapsto c_1(x_1)c_2(x_2)\cdots c_n(x_n) \\
\end{align*}
\]

(8.84)

need not be a homomorphism (for example, the map of \( X \times X \to X \) defined by group multiplication is typically not a homomorphism). In the rest of this section we will discuss these problems in some depth. We will also present some additional conditions which enable us to circumvent some of the difficulties.

Even if \( R \) is a subgroup, we cannot be sure that the map

\[
\begin{align*}
B_k : U^k & \to R \\
B_k(u_0, \ldots, u_{k-1}) & = b(u_{k-1})a_{b(u_{k-2})} \cdots a_{k-1}b(u_0) \\
\end{align*}
\]

(8.85)

is a homomorphism. Note that assuming that \( \text{card } X = n \), then \( B_n \) is onto \( R \). We would like to investigate putting a semidirect product structure (see Appendix A) on \( U^n \) such that \( B_n \) is a homomorphism. We have the following necessary condition.

**Theorem 8.14:** Consider the FGHSS (8.33), (8.34). If there exists a semidirect product structure on \( U^n \) such that \( B_n : U^n_{\text{sd}} \to X \) is a homomorphism, then the set of states reachable from the identity in \( k \) steps is a group \( Y \) \( k > 0 \).

**Proof:** Choose \( k \in \{1, \ldots, n\} \). Consider the set of input strings

\[
\begin{align*}
W_k = \{(e^{n-k}, u_0, \ldots, u_{k-1}) | u_i \in U\} \\
\end{align*}
\]

(8.86)
For any semidirect product structure on $U^n$, this is a subgroup.

Thus $B_n(W_k)$ is a subgroup if $B_n$ is a homomorphism, and $B_n(W_k)$ is precisely the set of states reachable from $e$ in $k$ steps. For $k > n$ use Theorem 8.10.

Example 8.2: We modify Example 8.1 and consider only the input-state part of the system. Let $U = Z_2$, $X = D_4$, let $b$ be defined by (8.71) and define $a$ by

$$a(y) = xy, \quad a(xv) = y$$  \hfill (8.87)

It is easy to check that $a$ is an automorphism of $D_4$, and by Theorem 8.12, $R$ is a subgroup. However

$$B_n(W_2) = \{e, y, xy, x^3\}$$  \hfill (8.88)

which is not a subgroup, and therefore $B_n$ is not a homomorphism for any semidirect product structure on $U^n$.

Examples 8.1 and 8.2 illustrate an asymmetry in the theory. Unlike linear system theory -- or even the abelian group case here, where it is clear that none of these difficulties appear -- we do not have a naive duality theory without additional assumptions. An assumption that avoids some of these difficulties is that of requiring $a$ to be a normal endomorphism.

Definition 8.4: A homomorphism $f$ of a group $G$ into itself is called a normal endomorphism if for all $x, y \in G$

$$xf(y)x^{-1} = f(xyx^{-1})$$  \hfill (8.89)

Note that if $G$ is abelian, all endomorphisms are normal.
Theorem 8.15: Consider the FGHSS (8.33), (8.34) evolving in a finite group $X$ of order $n$. Suppose $a$ is a normal endomorphism. Then there exists a semidirect product structure on $U^n$ such that the input-state map $B_n$ is a homomorphism, and thus the identity-reachable set $R$ is a subgroup.

Proof: Define the binary operation on $U^n$

$$ (u^o_0, u^1_1, \ldots, u^n_{n-1})(v^o_0, v^1_1, \ldots, v^n_{n-1}) = $$

$$ (v^{-1}_1 v^1_2 \cdots v^{-1}_{n-1} u^0_0 v^{-1}_1 v^1_2 \cdots v^{-1}_{n-1} u^1_1 v^{-1}_2 v^2_3 \cdots v^{-1}_{n-1} u^n_{n-1} \cdots v^n_{n-1} v^n_{n-1}) \cdots (8.90) $$

Direct computation verifies that this does define a semidirect product structure on $U^n$, and another computation, using the fact that $a$ is a normal endomorphism, verifies that $B_n$ is a homomorphism.

Thus, in this case, we can reduce our system to a minimal FGHSS realization by first restricting the homomorphisms to $R$ and then taking $R/\text{ker } C$, where $C$ is defined by (8.62). Note that the condition that $a$ be normal is sufficient to guarantee that $R$ be a group and $B_n$ be a homomorphism; however, as the next example shows, it is not necessary.

Example 8.3: Let $G$ be a non-abelian group, and consider the FGHSS (8.33), (8.34) with $U = Y = G$, $X = G \times G \times G$ (direct product) and $a$, $b$, and $c$ defined by

$$ a(g_1, g_2, g_3) = (e, g_1, g_2) \quad (8.91) $$

$$ b(g) = (g, g, g) \quad (8.92) $$

$$ c(g_1, g_2, g_3) = g_3 \quad (8.93) $$
It is easy to see that these are homomorphisms and that

$$\ker c \cap \ker ca \cap \ker ca^2 = \{e\}$$  \hspace{1cm} (8.94)

so the system is observable. Also, given any \((g_1, g_2, g_3) \in X\), let

\[ u_0 = g_2^{-1}g_3, \quad u_1 = g_1^{-1}g_2, \quad u_2 = g_1. \]

Then

$$b(u_2)ab(u_1)a^2b(u_0) = (g_1, g_2, g_3)$$  \hspace{1cm} (8.95)

and thus the system is controllable. However, letting \((g_1, g_2, g_3)\)

and \((h_1, h_2, h_3)\) be arbitrary elements of \(X\), we have

$$a[(h_1, h_2, h_3)](g_1^{-1}, g_2^{-1}, g_3^{-1}) = (e, g_2 h_1 g_2^{-1}, g_3 h_2 g_3^{-1})$$  \hspace{1cm} (8.96)

while

$$a[(g_1, g_2, g_3), (h_1, h_2, h_3)](g_1^{-1}, g_2^{-1}, g_3^{-1}) = (e, g_1 h_1 g_1^{-1}, g_2 h_2 g_2)$$  \hspace{1cm} (8.97)

and the two quantities are, in general, not equal. Thus \(a\) is not

a normal endomorphism.

However, we will demonstrate that a semidirect product structure
can still be put on \(U^k\) to make \(B_k\) a homomorphism. Consider \(k = 3\) and
define the product on \(U^3\)

$$\begin{align*}
(u_1, u_2, u_3)(v_1, v_2, v_3) &= (v_2^{-1}u_1v_1, v_2^{-1}u_2v_2, v_3^{-1}u_3v_3) \\
&= (v_2^{-1}u_1v_1, v_2^{-1}u_2v_2, v_3^{-1}u_3v_3)
\end{align*}$$  \hspace{1cm} (8.98)

Then one can check that

$$B_3(u_1, u_2, u_3)B_3(v_1, v_2, v_3) = B_3[(u_1, u_2, u_3)(v_1, v_2, v_3)]$$

$$= (u_3v_3, u_3u_2v_3, u_3u_2u_1v_3v_2)$$  \hspace{1cm} (8.99)

Another question arises in the case in which \(R\) is not a group.

When this happens, we have \(x_1, x_2 \in R\) such that \(x_1x_2 \notin R\). Thus this
particular group multiplication never occurs in the operation of
the system and is irrelevant information. One can then ask whether
or not we can redefine these irrelevant multiplications in such a
manner as to make $R$ a group, while at the same time requiring that
$a$, $b$, and $c$ remain homomorphisms when restricted to $R$. The system
defined in Example 8.1 shows that, at least in some cases, this can be
done. Again let $U = Y = Z_2$, $X = D_4$ and define $a$, $b$, and $c$ by (8.72),
(8.71), and (8.73) respectively. We saw that

$$R = \{e, y, xy, x^3\} \quad (8.100)$$

The superfluous multiplications are $(xy)(y)$, $(xy)(x^3)$, $(x^3)(y)$,
and $(x^3)(x^3)$. If we define these as follows:

$$
(xy)(y) = x^3 \quad (xy)(x^3) = y \\
(x^3)(y) = xy \quad (x^3)(x^3) = e
$$

(8.101)

then $R$ is the Klein-4 group (see [Fl]), and it is easy to check
that $a$, $b$, and $c$ are still homomorphisms. This realization is not
observable, and, reducing it modulo the kernel of the observatvility
map $C$

$$\ker C = \{e, x^3\} \quad (8.102)$$

we obtain a minimal realization, which by Theorem 8.11 is equivalent
to the one given in Example 8.1.

8.4 Number-Theoretic Arguments and Invertibility Conditions for FGHSS's

In this section we will consider some other problems related to
FGHSS's. Recall that after the proofs of Theorem 8.10 and its corollary,
we noted that the bounds in (8.59) and (8.60) are exponentially weaker
than those of linear theory. We now present a result that corrects this problem and a second result, motivated by the proof of the first, that tightens the bounds even more than the first. A very important tool is Theorem A.5, which states that if \( G \) is a group and \( S < G \), then \( |S| \leq |G| \).

We adopt the following notation let \( n \) and \( m \) be positive integers, and define

\[
\begin{align*}
    s_m &= \text{smallest integer} > 1 \text{ that divides } m \\
n(m,n) &= \lfloor \log_{s_m}(n) \rfloor
\end{align*}
\]  

(8.103) \hspace{1cm} (8.104)

where \( \lfloor x \rfloor \) is the smallest integer \( \geq x \). We can now prove the following.

**Lemma 8.1:** Consider the FGHSS (8.33), (8.34) evolving in a state group with \( n \) elements. Then the system is observable if and only if

\[
    \ker c \cap \ker ca \cap \ldots \cap \ker ca^{n(n,n)-1} = \{e\}
\]  

(8.105)

**Proof:** By the definition of observability, if (8.105) holds, the system is observable. Thus, we need only show that if the system is observable, (8.105) must hold. We first let

\[
    \begin{align*}
    K_p &= \ker c \cap \ldots \cap \ker ca^{p-1} \quad p \geq 1 \\
    K_0 &= x
    \end{align*}
\]  

(8.106)

and note that

\[
    K_p \supseteq K_{p+1}
\]  

(8.107)

We claim that if \( K_p = K_{p+1} \) for some \( p \), then

\[
    K_q = K_p \quad \forall \quad q \geq p
\]  

(8.108)

Suppose \( x \in K_p \). Then since \( x \in K_{p+1} \), we have that
c[a(x)] = ca[a(x)] = ... = ca^{p-1}[a(x)] = e \quad (8.109)

but this implies that \( a(x) \in K_p \). Repeating this argument, we see that \( a^{l}(x) \in K_p \quad \forall \quad l \geq 0 \) if \( x \in K_p \), which implies that \( x \in K_q \quad \forall \quad q \geq p \), and \( K_p \subseteq K_q \quad \forall \quad q \geq p \). Thus \( K_p \subseteq K_q \quad \forall \quad q \geq p \), and this and (8.107) yield (8.108).

Thus if our system is observable, the chain \( X = K_0 \supseteq K_1 \supseteq K_2 \supseteq ... \) must be strictly decreasing in size until for some \( r \), \( K_r = \{ e \} \).

As proved in Theorem 8.10, \( K_p \) is a subgroup of \( K \), and, since \( K_{p+1} \subseteq K_p \), \( K_{p+1} \subset K_p \). Then by Theorem A.5

\[
\text{card } K_{p+1} \setminus \text{card } K_p \quad (8.110)
\]

Given \( s_n \) defined by (8.103), and assuming the system is observable but \( K_p \neq \{ e \} \), we must have that

\[
\frac{\text{card } K_p}{\text{card } K_{p+1}} \geq s_n \quad (8.111)
\]

Thus

\[
\text{card } K_q(n,n) \leq \frac{1}{s_n} \text{card } K_q(n,n)-1 \leq \ldots \leq \left( \frac{1}{s_n} \right)^{q(n,n)} \text{card } K_0
\]

\[
= n \left( \frac{1}{s_n} \right)^{q(n,n)} \quad (8.112)
\]

and by the definition of \( q(n,n) \) we see that we must have

\[
K_q(n,n) = \{ e \} \quad (8.113)
\]

This result can be strengthened in the following manner.

**Theorem 8.16:** Consider a system satisfying the hypotheses of Lemma 8.1.

Let us write

\[
n = (p_1)^{k_1}(p_2)^{k_2}\ldots(p_s)^{k_s} \quad (8.114)
\]
where the $p_i$ are distinct primes and the $k_i$ are positive integers.

Let

$$k(n) = \sum_{i=1}^{s} k_i \quad (8.115)$$

Then the system is observable if and only if

$$\ker c \cap \ldots \cap \ker ca^{k(n)-1} = \{e\} \quad (8.116)$$

**Proof:** The proof is similar to the proof of the preceding theorem.

Here we use the fact that the order of $K_p$ must be of the form

$$\text{card } K_p = (p_1)^{l_1} \ldots (p_s)^{l_s} \quad l_i \leq k_i \quad (8.117)$$

We now present an example showing that this bound is the best possible.

**Example 8.4:** Let $r > 2$ and define $a : Z_2^r \rightarrow Z_2^r$ and $c : Z_2^r \rightarrow Z_2^r$ by

$$a(x_1, \ldots, x_r) = (x_2, x_3, \ldots, x_r, x_1) \quad (8.118)$$

$$c(x_1, \ldots, x_r) = x_1 \quad (8.119)$$

For $1 \leq s \leq r$, let

$$G_s = Z_2^{s-1} \times \{0\} \times Z_2^r \quad (8.120)$$

Then

$$\ker ca^k = G_{k+1} \quad (8.121)$$

and

$$\text{card } K_p = \text{card}(G_1 \cap \ldots \cap G_p) > 1 \quad p < r \quad (8.122)$$

$$\text{card } K_r = 1 \quad (8.123)$$

Also

$$\text{card } Z_2^r = 2^r \quad (8.124)$$
and

\[ q(2^r, 2^r) = r = k(2^r) \]  \hspace{1cm} (8.125)

so the previous results are as good as could be hoped for.

One may ask whether one can make analogous statements about the concept of controllability. Taking into account the discussions of the previous section, where we saw that the identity-reachable set \( R \) need not be a subgroup, one would suppose that in general one would not be able to say as much about the controllability sets. The next example indicates that our skepticism is well-founded.

**Example 8.5:** We present an example that indicates that

\( R \neq b(U)ab(U)\ldots a^r b(U) \) for \( r = q(n, n) - 1 \) or \( r = k(n) - 1 \). Let \( U = Z_2, X = D_4 \).

Then \( n = 8 \) and \( q(n, n) = k(n) = 3 \). Let \( a \) and \( b \) be defined by

\[ a(y) = xy, \quad a(xy) = y \]  \hspace{1cm} (8.126)

\[ b(1) = y \]  \hspace{1cm} (8.127)

Then, from Example 8.2, \( R = D_4 \), but

\[ b(U)ab(U)a^2 b(U) = b(U)ab(U)a^{q(8, 8) - 1} b(U) \]

\[ = b(U)ab(U)a^{k(8) - 1} b(U) = \{e, v, xy, x^3, x, x^3 v\} \neq R \]  \hspace{1cm} (8.128)

Note also that

\[ \text{card } b(U)ab(U)a^2 b(U) = 6 \]  \hspace{1cm} (8.129)

which does not divide \( \text{card } D_4 = 8 \). One can also find examples where

\( \text{card } R \not\subseteq \text{card } X \). Let \( U = Z_2, X = D_3 \) and define \( a \) and \( b \) by

\[ a(x) = e \quad a(y) = xy \quad b(1) = y \]  \hspace{1cm} (8.130)

Then

\[ R = \{e, y, xy, x^2\} \]  \hspace{1cm} (8.131)
and 4 does not divide 6 = card \( D_3 \).

Let us define the set of states reachable from \( e \) in \( k \) steps

\[
R_k = b(U)ab(U) \ldots a^{k-1}b(U)
\]  \hspace{1cm} (8.132)

\[
R_0 = \{e\}
\]  \hspace{1cm} (8.133)

If this is a subgroup for all \( k \) (as is the case if \( a \) is a normal endomorphism), we can say something, as the following theorem indicates.

**Theorem 8.17:** Consider a FGHSS with card \( X = n \). Then, if \( R_k \) is a subgroup for all \( k \),

\[
R = R_k(n)
\]  \hspace{1cm} (8.134)

where \( k(n) \) is given by (8.115). Thus the system is controllable if and only if

\[
R_k(n) = X
\]  \hspace{1cm} (8.135)

**Proof:** The proof is the "dual" of the proofs of Lemma 1 and Theorem 8.16, and we will just sketch the proof. Clearly

\[
R_k \subseteq R_{k+1}
\]  \hspace{1cm} (8.136)

and, if for some \( k \)

\[
R_k = R_{k+1}
\]  \hspace{1cm} (8.137)

then, using the formula

\[
R_{j+1} = b(U)a(R_j)
\]  \hspace{1cm} (8.138)

we can show that

\[
R_q = R_k \hspace{0.5cm} \forall \hspace{0.3cm} q \geq k
\]  \hspace{1cm} (8.139)

and thus

\[
R = R_k
\]  \hspace{1cm} (8.140)
Suppose

\[ R \neq R_k \quad k \leq k(n) - 1 \quad (8.141) \]

That is,

\[ \text{card } R_{k+1} > \text{card } R_k \quad k \leq k(n) - 1 \quad (8.142) \]

Since \( R_k < R_{k+1} \quad \forall k \), we have

\[ \text{card } R_k \setminus \text{card } R_{k+1} \quad (8.143) \]

and, then, referring to the counting arguments of Theorem 8.16 and Lemma 1, we see that (8.141) implies

\[ R = R_{k(n)} = X \quad (8.144) \]

and if \( R = R_k \) for \( k < k(n) \), we clearly have \( R = R_{k(n)} \).

Clearly a weaker result, the controllability analog of Lemma 8.1, is also true.

**Example 8.6:** This example is the dual of Example 8.4. Let

\[ U = \mathbb{Z}_2, \quad X = \mathbb{Z}_2^r \quad (r \geq 2), \]

\[ b(x) = (x, 0^{r-1}) \quad (8.145) \]

\[ a(x_1, \ldots, x_r) = (x_r, x_1, \ldots, x_{r-1}) \quad (8.146) \]

Then for \( s < r \),

\[ R_s = b(U) + \ldots + a^{s-1}b(U) = \mathbb{Z}_2^s \times \{0\}^{r-s} \quad (8.147) \]

and

\[ R \neq R_s \quad s \leq r-1 \quad (8.148) \]

\[ R = R_r \quad (8.149) \]

and

\[ r = q(2^r, 2^r) = k(2^r) \quad (8.150) \]
We wish to relate these results to those in the theory of LSC's. We have succeeded in Theorems 8.16 and 8.17 in getting rid of the exponential discrepancy between the results for FCHSS's and those for LSC's. In fact, suppose we regard the LSC (8.1), (8.2) as a FCHSS as discussed earlier. The underlying field is $\mathbb{GF}(p^m)$ for some prime $p$ and some integer $m > 1$. Then if the dimension of the state space is $N$, the cardinality $n$ of the state space is

$$n = p^{mN} \tag{8.151}$$

and our results say we need only check

$$B, AB, \ldots, A^{mN-1}B \tag{8.152}$$

for controllability and

$$C, CA, \ldots, CA^{mN-1} \tag{8.153}$$

for observability. Thus, if $m=1$ -- i.e. if the underlying field has prime cardinality -- we get precisely the LSC results. Why is our result weaker for $m > 1$? The reason is related to the following ideas: (1) the field $\mathbb{GF}(p^m)$ can be regarded [M1], [L1], as a vector space of dimension $m$ over $\mathbb{GF}(p)$, so a vector space of dimension $N$ over $\mathbb{GF}(p^m)$ is of dimension $mN$ over $\mathbb{GF}(p)$; (2) in the class of FCHSS we allow homomorphisms over $\mathbb{GF}(p^m)$ which are not linear maps -- e.g., the additive group of $\mathbb{GF}(4)$ is $\mathbb{Z}_2 \times \mathbb{Z}_2$ (see Appendix A) and one could consider the homomorphism of $\mathbb{Z}_2 \times \mathbb{Z}_2 \to \mathbb{Z}_2 \times \mathbb{Z}_2$

$$(a, b) \mapsto (a, 0) \tag{8.154}$$

which is a linear map over $\mathbb{GF}(2) = \mathbb{Z}_2$ but not over $\mathbb{GF}(4)$, since $\{0\} \times \mathbb{Z}_2$ is not a subspace of the one-dimensional vector space $\mathbb{GF}(4)$. 
These are precisely the facts that make our results weaker for LSC's -- i.e. in enlarging the class of systems considered, we must allow for the slowest rate of decrease in card $K_2$ and the slowest rate of increase in card $R_2$. In the LSC case we have the added information that dim $K_2$ and dim $R_2$ must change by at least one, so the cardinalities of $R_2$ and $K_2$ must change by at least a factor of the cardinality of the underlying field $GF(p^m)$ -- i.e. at least a factor of $p^m$. Thus if we use this knowledge to replace $s_n = p$ with $s_n = p^m$, we get precisely the LSC results.

In addition to the upper bounds of Theorems 8.16 and 8.17, we can derive some lower bounds for the controllability and observability conditions. These are related to the following: suppose $B : K^m \rightarrow K^n$, $A : K^n \rightarrow K^n$, where $K$ is a field; then

$$\text{rank } [B, AB, \ldots, A^{r-1}B] = n$$ \hspace{1cm} (8.155)

only if the number of columns is $\geq n$ -- i.e. only if

$$r \geq \left\lceil \frac{n}{m} \right\rceil$$ \hspace{1cm} (8.156)

**Theorem 8.18:** Consider the FGHSS (8.33), (8.34) with card $U = m$, card $X = n$, card $Y = r$. Write $n$ in the prime decomposition

$$n = (p_1)^{k_1} \ldots (p_s)^{k_s}$$ \hspace{1cm} (8.157)

Then the system is observable only if $p_i \nmid r \forall i$, and writing

$$r = M_{1}^{t_1} \ldots (p_s)^{t_s}$$ \hspace{1cm} (8.158)

where $M_{1}$ is relatively prime with respect to all the $p_i$, we have the smallest possible $\ell$ such that $K_2 = \ker c \cap \ker ca \cap \ldots \cap \ker ca^{\ell-1} = \{e\}$


is
\[ l_0 = \max \left( \left\lfloor \frac{k_1}{c_1} \right\rfloor \mid i=1,\ldots,s \right) = \min \left( i \mid n \setminus r^i \right) \] (8.159)

Also, if \( B_k : U^k \to X \) (see (8.85)) is a homomorphism for all \( k \) (e.g. if \( a \) is a normal endomorphism), then the system is controllable only if \( p_i \setminus m \forall i \), and writing
\[ m = M_2(p_1)^{q_1} \cdots (p_s)^{q_s} \] (8.160)

where \( M_2 \) is relatively prime with respect to all the \( p_i \), we have that the smallest possible \( i \) such that \( R_i = b(U)ab(U) \cdots a^{i-1}b(U) = X \) is
\[ i_0 = \max \left( \left\lfloor \frac{k_1}{q_1} \right\rfloor \mid i=1,\ldots,s \right) \] (8.161)

Proof: The map \( C_\ell : X \to Y \) defined by
\[ C_\ell(x) = (c(x),\ldots,ca^{\ell-1}(x)) \] (8.162)

is a homomorphism. Thus by the First Isomorphism Theorem for groups (Theorem A.7),
\[ X/\ker C_\ell = X/K_\ell = C_\ell(x) < Y^\ell \] (8.163)

and Theorem A.5 tells us that
\[ \text{card } C_\ell(x) \setminus \text{card } Y^\ell \] (8.164)

Also
\[ \text{card } C_\ell(x) = \text{card } X/K_\ell = n/\text{card } K_\ell \] (8.165)
\[ \text{card } Y^\ell = r^\ell \] (8.166)

Thus
\[ n \setminus (r^\ell \text{card } K_\ell) \] (8.167)

and, if the system is observable, for some \( \ell \)
\[ \text{card } K_2 = 1 \]  
(8.168)

which implies that \( p_1 \setminus r \not\subset i \) and also that \( n \setminus r \not\subset \), which is clearly true if and only if

\[ \begin{align*}  
p_1 & \setminus (p_1) \not\subset i \quad \forall i 
\end{align*} \]  
(8.169)

i.e., if and only if

\[ l \geq \max \left( \left\lceil \frac{k_i}{t_1} \right\rceil \mid i = 1, \ldots, s \right) = l_0 \]  
(8.170)

The arguments for controllability are analogous, where we use the fact that, since there exists a semidirect product structure on \( U^k \) such that \( B_k : U^k_{sd} \to X \) is a homomorphism, the First Isomorphism Theorem states that

\[ U^k_{sd}/\ker B_k = B_k(U^k_{sd}) = R_k < X \]  
(8.171)

We now will consider the question of invertibility of FCHSS's. The reader is referred to the general definition of L-delay invertibility given in Definition 8.2. Our work will parallel the work of Massey and Sain [M10], [S10]. The reader is also referred to [B16], [F4], [O3], and [O4] for other work on the invertibility of dynamical systems. Recall the invertibility results for LSC's (Theorems 8.3 - 8.5). In particular, Theorem 8.3 states that an inverse of delay L exists if and only if

\[ \text{rank } M_{L+1} = \text{rank } M_L + s \]  
(8.172)

(note the change in notation from (8.17)), where \( M_k \) is defined by (8.15) and \( s \) is the dimension of the input space. If the underlying field has \( q \) elements in it, then the cardinality \( m \) of the input space is \( s^q \), and (8.172)
is equivalent to

\[ \text{card } Ra M_{L+1} = m \text{ card } M_L \quad (8.173) \]

It is this result that we shall generalize first, and the reader is referred to the parallel LSC arguments in [M10].

Recalling the general invertibility problem, our problem here is the following: consider the FGHSS (8.33), (8.34) with \( x(0) = e \), card \( X = n \), card \( U = m < \text{card } Y = p \) (we need this to have any chance of inverting the system); we wish to find conditions on the system such that an \( L \)-delay inverse exists -- i.e. such that \( U_k = (u(0), \ldots, u(k)) \) can be recovered from \( Y_k+L+1 = (v(1), \ldots, v(k+L+1)) \) (as we will discuss, the inverse system need not be a FGHSS). Defining the input-output homomorphisms \( T_1 = ca^b \), we see that we have the relation

\[ Y_k = M_k U_{k-1} \quad (8.174) \]

where \( M_k : U^k \rightarrow Y^k \) is defined by

\[
\begin{align*}
y(1) &= T_0[u(0)] \\
y(2) &= T_0[u(1)]T_1[u(0)] \\
& \vdots \\
y(k) &= T_0[u(k-1)]T_1[u(k-2)] \cdots T_{k-1}[u(0)] 
\end{align*}
\quad (8.175)
\]

Note that unlike \( M_k \) in the LSC case and much like the input-state map \( B_k \) for FGHSS's, \( M_k \) for a FGHSS need not be a homomorphism (for any direct or semidirect product structures on the \( U^k \)). However, it always will be a homomorphism if \( Y \) is abelian (we can then put the direct product structure on \( U^k \) and \( Y^k \)).
Theorem 8.19: A FGHSS has an inverse with delay $L$ if and only if $U_0$ can be determined from $Y_{L+1}$.

Proof: We need only show sufficiency, since it is necessary by definition. Thus suppose $u(0) = U_0$ can be determined from $Y_{L+1}$. We can then "divide out" the effects of $u(0)$ on $Y_k$. The modified outputs $z(k)$, given by

$$z(k) = y(k)[T_{k-1}[u(0)]]^{-1}$$  

(8.176)

if we omit $z(1)(=e)$, are the same as they would be if $u(1)$ were the first input to the system (stationarity is important here). Thus $u(1)$ can be determined from $Y_{L+2}$ \textit{i.e.}, $U_1 = (u(0), u(1))$ can be recovered from $Y_{L+2}$. Continuing this procedure, we see that $U_k$ can be recovered from $Y_{k+L+1}$.

Define the maps $A_k : U \rightarrow Y^k$, $D_k : U^{k-1} \rightarrow Y^k$

$$A_k(u) = (T_0(u), T_1(u), \ldots, T_{k-1}(u))  \quad (8.177)$$

$$D_k(u_1, \ldots, u_{k-1}) = (e, M_{k-1}(u_1, \ldots, u_{k-1}))  \quad (8.178)$$

and endow $Y^k$ with the direct product group structure. Note that $A_k$ is a homomorphism, $D_k$ need not be, and

$$M_k[u(0), \ldots, u(k-1)] = D_k[u(1), \ldots, u(k-1)]A_k[u(0)]  \quad (8.179)$$

Lemma 8.2: A FGHSS has an inverse with delay $L$ if and only if

(1) $A_{L+1}$ is one-to-one

(2) If $x \in \text{Ra } A_{L+1}$, $y \in \text{Ra } D_{L+1}$, then

$$yx \in \text{Ra } A_{L+1} \Rightarrow y = (e, \ldots, e)  \quad (8.180)$$

$$yx \in \text{Ra } D_{L+1} \Rightarrow x = (e, \ldots, e)  \quad (8.181)$$
Proof (Necessity): Suppose the FCHSS has an L-delay inverse. Clearly $A_{L+1}$ must be one-to-one, or else there exists an input $u(0) \neq e$ such that $A_{L+1}(u(0)) = (e, \ldots, e)$. Then, from (8.179)

$$M_{L+1}(u(0), e, \ldots, e) = (e, \ldots, e) = Y_{L+1} = M_{L+1}(e, \ldots, e)$$

so $u(0)$ cannot be recovered from $Y_{L+1}$. Now suppose we have $x \in Ra A_{L+1}$, $y \in Ra D_{L+1}$ such that $yx \in Ra A_{L+1}$. This implies that there exists an input sequence $u(0), \ldots, u(L)$ and another input $u'(0)$, such that

$$yx = D_{L+1}(u(1), \ldots, u(L)) A_{L+1}(u(0)) = M_{L+1}(u(0), u(1), \ldots, u(L))$$

$$= M_{L+1}(u'(0), 0, \ldots, 0) = A_{L+1}(u'(0))$$

Since we are assuming L-delay invertibility, we must have $u(0) = u'(0)$, but then we must have $y = (e, \ldots, e)$. Now assume $x \in Ra A_{L+1}$, $y \in Ra D_{L+1}$ such that $yx \in Ra D_{L+1}$. Then there is an input sequence $u(0), \ldots, u(L)$ such that the response to this is the same as the response to $e$, $u'(1), \ldots, u'(L)$. Thus, by L-invertibility, $u(0) = e$, and $x = A_{L+1}(u(0)) = (e, \ldots, e)$.

Proof (Sufficiency): We now assume conditions (1) and (2) of the theorem statement hold. Suppose we have two different input sequences $u(0), \ldots, u(L)$ and $u'(0), \ldots, u'(L)$ such that

$$Y_{L+1} = M_{L+1}(u(0), \ldots, u(L)) = M_{L+1}(u'(0), \ldots, u'(L)) = Y'_{L+1}$$

Referring to the definition of $M_k$ in (8.175), we see that (8.184) is equivalent to
\[ T_0[u(0)u'(0)^{-1}] = e \]
\[ T_0[u(1)]T_1[u(0)u'(0)] = T_0[u'(1)] \]
\[ \vdots \]
\[ T_0[u(L)] \cdots T_L[u(0)u'(0)^{-1}] = T_0[u'(L)] \cdots T_{L-1}[u'(1)] \]

That is

\[ D_{L+1}[u(1), \ldots, u(L)]A_{L+1}[u(0)u'(0)^{-1}] = D_{L+1}[u'(1), \ldots, u'(L)] \]  

(8.186)

By condition (2), we must have \( A_{L+1}[u(0)u'(0)^{-1}] = (e, \ldots, e) \), and by (1), \( u(0)u'(0)^{-1} = e - i.e. u(0) = u'(0) \). Thus we recover \( u(0) \) uniquely from \( Y_{L+1} \) and our system is L-delay invertible.

We refer the reader to the proof of Theorem 4 in [M10]. Lemma 8.2 is the analog of the part of that proof concerning the linear independence of certain columns of the matrix \( M_L \) in the LSC case. We can now prove the analog of Theorem 8.3.

**Theorem 8.20:** A FGHSS has an inverse with delay \( L \) if and only if

\[ \text{card } Ra M_{L+1} = m \text{ card } Ra M_L \]  

(8.187)

where card \( U = m \).

**Proof:** Note that

\[ \text{card } Ra D_{L+1} = \text{card } Ra M_L \]  

(8.188)

We also see that condition (1) of Lemma 8.2 is equivalent to

\[ \text{card } Ra A_{L+1} = m, \text{ and we claim that (2) is equivalent to the following:} \]

if \( x_1, x_2 \in \text{Ra } A_{L+1}, y_1, y_2 \in \text{Ra } D_{L+1}, \) then

\[ y_1x_1 = y_2x_2 \Rightarrow y_1 = y_2, \quad x_1 = x_2 \]  

(8.189)
Suppose (2) holds. Then if \( y_1 x_1 = y_2 x_2 \)

\[ y_1 x_1^{-1} x_2^{-1} = y_2 \]  \hspace{1cm} (8.190)

but \( Ra_{A_{L+1}} \) is a group, so \( x_1 x_2^{-1} \in Ra_{A_{L+1}} \). Then by (2), \( x_1 x_2^{-1} = e \), which implies \( x_1 = x_2, y_1 = y_2 \). Now suppose (8.189) holds. Then, if \( y_1 x_1 = x_2 \in Ra_{A_{L+1}}, y_1 x_1 = e \) (\( e \in Ra_{D_{L+1}} \)), which implies \( y_1 = e \) by (8.189). Similarly, \( y_1 x_1 = y_2 \in Ra_{D_{L+1}} \) and \( e \in Ra_{A_{L+1}} \) imply \( x_1 = e \) by (8.189).

Now note that (8.189) holds if and only if

\[
\text{card}(Ra_{D_{L+1}} Ra_{A_{L+1}}) = \text{card} Ra_{B_{L+1}} \text{card} Ra_{A_{L+1}}
\]  \hspace{1cm} (8.191)

Then, referring to (8.179), (8.188), and (8.191), we see that (8.187) holds.

We now define the concept of pointwise observability as in [M10].

**Definition 8.5:** A FGHSS is **pointwise input observable** if any \( u(0) \in U \) in the input sequence \( u(0), e, e, e, \ldots \), is uniquely determined from the output sequence \( y(1), y(2), \ldots \).

**Theorem 8.21:** A FGHSS with card \( X = n \) is pointwise input observable if and only if \( A_k(n) \) is one-to-one.

**Proof:** It is easy to see that the system is pointwise input observable if and only if

\[ \ker cb \cap \ker cab \cap \ldots \cap \ker ca^r b = \{e\} \]  \hspace{1cm} (8.192)

for some \( r \). Now if \( b \) is not one-to-one, we cannot satisfy (8.192).

Thus assume \( b \) is one-to-one, and, since

\[ b[\ker cb \cap \ldots \cap \ker ca^r b] = \ker c \cap \ldots \cap \ker ca^r \cap b(U) \]  \hspace{1cm} (8.193)
we have that (8.192) holds if

$$\ker c \cap \ldots \cap \ker ca^r = \{e\}$$  \hspace{1cm} (8.194)$$

By Theorem 8.16, if (8.194) holds, it holds for $r = k(n)-1$, and, referring to (8.177), we see that we have $A_k$ is one-to-one for some $k$ (i.e. the system is pointwise input observable) if and only if $A_k(n)$ is.

We note that much as Theorem 8.21 is the analog of Theorem 8.16, the next result is the analog of Theorem 8.18.

**Theorem 8.22:** Consider a FGHSS with card $U = m$, card $Y = r$. Then $A_k$ cannot be one-to-one for

$$k < k_o = \min (\ell/m \backslash r^\ell)$$  \hspace{1cm} (8.195)$$

and thus $(k_o - 1)$ is a lower bound on the delay in any inverse of the system.

**Proof:** Consider $A_k : U \rightarrow Y^k$. Then

$$U/\ker A_k = A_k(U) < Y^k$$  \hspace{1cm} (8.196)$$

so

$$m \backslash r^k \text{ card } \ker A_k$$  \hspace{1cm} (8.197)$$

and $\ker A_k = \{e\} \iff m \backslash r^k$.

Note that this result has no analog for LSC's -- i.e. there is no lower bound on the delay. The reason is that in the LSC case $m = q^{s_1}$ and $r = q^{s_2}$, where $q$ is the cardinality of the underlying field and the $s_i$ are the dimensions of the input and output space. Since we must have $m \leq r$, we have $m \backslash r$, so $k_o = 1$.

We now wish to consider analogs of other LSC results. However,
we run into the same type of problem that confronted us when we considered the question of controllability. We first wish to consider an alternative characterization of invertibility analogous to Lemma 5 of [M10]. The idea for LSC's is the following: since the $M_k$ are linear maps, a system is L-invertible if and only if it is "kernel-free" -- i.e. if and only if $Y_{L+1} = 0$ implies $u(0) = 0$.

In the FGHSS case, the $M_k$ are not homomorphisms (it appears that they need not be homomorphisms even if $a$ is a normal endomorphism), and thus in general one does not have the guarantee that $Y_{L+1} = e \Rightarrow u(0) = e$ is the same as L-invertibility. However, we do have the following:

**Theorem 8.23:** Consider a FGHSS with $a$ a normal endomorphism. Then the system is L-invertible if and only if $Y_{L+1} = e$ implies $u(0) = e$.

**Proof:** As mentioned above, $M_k$ need not be a homomorphism, and we will have to work around this. The computations we will make are similar to the type needed in the proof of Theorem 8.15.

We need only show that if there exist two input sequences $u_0, \ldots, u_L$ and $v_0, \ldots, v_L$ such that $u_o \neq v_0$ but

$$M_{L+1}(u_0, \ldots, u_L) = M_{L+1}(v_0, \ldots, v_L) \quad (8.198)$$

then there exists a third sequence $w_0, \ldots, w_L$ such that $w_0 \neq e$ but

$$M_{L+1}(w_0, \ldots, w_L) = e \quad (8.199)$$

Using the semidirect product structure on $U^{L+1}$ given in Theorem 8.15, let

$$(w_0, \ldots, w_L) = (u_0, \ldots, u_L)(v_0, \ldots, v_L)^{-1} \quad (8.200)$$

Now using the semidirect product structure of Theorem 8.15 on $U^{k+1}$ for
$0 \leq k \leq L$ (which is not the same as that on $U_L^{L+1}$) we can show that

$$(w_0, \ldots, w_k) = (e^k, v_L \ldots v_{k+1})(z_0, \ldots, z_k)(e^k, v_{k+1} \ldots v_L)$$

(8.201)

where

$$z_0 = v_k \ldots v_1 u o v_1 \ldots v_k$$
$$z_1 = v_k \ldots v_2 u v_1 v_2 \ldots v_k$$
$$\vdots$$
$$z_{k-1} = v_k \ldots v_{k-1} v_k$$
$$z_k = u_{k-1} v_k$$

(8.202)

Because $a$ is a normal endomorphism we can use the Theorem 8.15 result to compute

$$T_0(w_k) \ldots T_k(w_0) = T_0(v_L \ldots v_{k+1})T_0(z_k) \ldots T_k(z_0)T_0(v_{k+1} \ldots v_L)$$

(8.203)

Thus we will have $T_0(w_k) \ldots T_k(w_0) = e$ if $T_0(z_k) \ldots T_k(z_0) = e$, and this follows by a straightforward computation. We illustrate this computation for $k = 2$.

$$T_0(z_2)T_1(z_1)T_2(z_0) = T_0(u_2)T_0(v_2^{-1})T_1(v_2^{-1}v_1^{-1}v_2^{-1})T_2(v_2v_1u_o v_1^{-1}v_2^{-1})$$

(8.204)

One can check the defining property of a normal endomorphism, equation (8.89), to show that

$$T_0(v_2^{-1})T_1(v_2^{-1}v_1^{-1}v_2^{-1}) = T_1(u_1)T_1(v_1^{-1})T_0(v_2^{-1})$$

(8.205)

and

$$T_1(v_1^{-1})T_0(v_2^{-1})T_2(v_2v_1u_o v_1^{-1}v_2^{-1}) = T_2(u_0)T_2(v_0^{-1})T_1(v_1^{-1})T_0(v_2^{-1})$$

(8.206)
Thus

$$T_0(z_2)T_1(z_1)T_2(z_o) = T_0(u_2)T_1(u_1)T_2(u_o)[T_0(v_2)T_1(v_1)T_2(v_o)]^{-1} \quad (8.207)$$

Equation (8.198) then implies

$$T_0(z_2)T_1(z_1)T_2(z_o) = e \quad (8.208)$$

which is the desired result.

Although the assumption that $a$ is a normal endomorphism allows us to prove the preceding result, we cannot derive analogs of the invertibility conditions of Theorems 8.4 and 8.5 without even stronger assumptions. Thus, we now assume that $Y$ is an abelian group. Let $U_c$ be the \textbf{commutator subgroup} of $U$

$$U_c = \text{subgroup generated by } \{aba^{-1}b | a, b \in U\} \quad (8.209)$$

Note that $U_c$ is abelian if and only if $U_c = \{e\}$, and for any homomorphism $\gamma : U \to Y$

$$\gamma(U_c) = \{e\} \quad (8.210)$$

Thus for our system to have any chance of being invertible, we must also assume that $U$ is abelian. In this case, $M_k : U^k$ (direct product) $\to Y^k$ (direct product) is a homomorphism $\Psi_k$ (the sum of homomorphisms on abelian groups is itself a homomorphism), as is $D_k : U^{k-1} \to Y^k$, defined by (8.178).

\textbf{Theorem 8.24:} Consider a FGHSS with $U$ and $Y$ abelian, and consider $A_k : U \to Y^k$ and $D_k : U^{k-1} \to Y^k$ defined by (8.177) and (8.178) respectively. Then
\[
\frac{\text{card } Ra_{M_{L+1}}}{\text{card } Ra_{M_L}} = \frac{\text{card } Ra_{A_{L+1}}}{\text{card}(Ra_{A_{L+1}} \cap Ra_{D_{L+1}})} \tag{8.211}
\]

**Proof:** We first note that \( Ra_{A_{L+1}} \) and \( Ra_{D_{L+1}} \) are groups and from (8.179)

\[
Ra_{M_{L+1}} = Ra_{A_{L+1}} + Ra_{D_{L+1}} \tag{8.212}
\]

where we use "+" to denote the commutative group operation on \( Y \).

Let

\[
Ra_{A_{L+1}} \cap Ra_{D_{L+1}} = \{z_1, \ldots, z_r\} \tag{8.213}
\]

Choose \( w \in Ra_{M_{L+1}} \) and \( w_A \in Ra_{A_{L+1}}, w_B \in Ra_{D_{L+1}} \) such that

\[
w = w_A + w_B \tag{8.214}
\]

Then

\[
w = (w_A + z_1) + (w_B - z_1), \quad \forall z_1 \in Ra_{A_{L+1}} \cap Ra_{D_{L+1}} \tag{8.215}
\]

and, if

\[
w = v_A + v_B \quad v_A \in Ra_{A_{L+1}}, v_B \in Ra_{D_{L+1}} \tag{8.216}
\]

then

\[
v_A = w_A + z_1 \quad v_B = w_B - z_1 \tag{8.217}
\]

for any \( z_1 \in Ra_{A_{L+1}} \cap Ra_{D_{L+1}} \). Thus, there are precisely

\[
r = \text{card}(Ra_{A_{L+1}} \cap Ra_{D_{L+1}}) \quad \text{ways to write each } w \in Ra_{M_{L+1}} \text{ as a sum of an element of } Ra_{A_{L+1}} \text{ and an element of } Ra_{D_{L+1}}.
\]

Then

\[
\text{card } Ra_{M_{L+1}} = \frac{[\text{card } Ra_{A_{L+1}}][\text{card } Ra_{D_{L+1}}]}{\text{card}(Ra_{A_{L+1}} \cap Ra_{D_{L+1}})} \tag{8.218}
\]

and, we observe that \( \text{card } Ra_{D_{L+1}} = \text{card } Ra_{M_L} \), which proves the theorem.
Note that

\[ \text{card } \text{Ra } A_{L+1} \leq \text{card } U = m \]  
(8.219)

so that the invertibility condition (8.187) denotes the maximum possible value for \( \text{card } \text{Ra } M_{L+1} \) given \( \text{card } \text{Ra } M_L \). Also

\[ \text{card } \text{Ra } A_{L+1} \cap \text{Ra } D_{L+1} < \text{card } \text{Ra } A_{L+1} \]  
(8.220)

so

\[ \frac{\text{card } M_{L+1}}{\text{card } M_L} \text{ is an integer} \]  
(8.221)

We now prove that the ratio (8.211) is a non-decreasing function of \( L \)

(we know that once the ratio equals \( m \), it stays equal to \( m \) because \( L \)-invertibility implies \( I \)-invertibility \( \forall I \succ L \)).

**Theorem 8.25:** Consider a FGHSS under the same hypotheses as in Theorem 8.24. Then the ratio (8.211) is a nondecreasing function of \( L \).

**Proof:** We will show that

\[ \frac{\text{card } \text{Ra } M_{L+1}}{\text{card } \text{Ra } M_L} \geq \frac{\text{card } \text{Ra } M_L}{\text{card } \text{Ra } M_{L-1}} \]  
(8.222)

For any \( \alpha \in \text{Ra } A_{L+1} \cap \text{Ra } D_{L+1} \) there exist \( u_0, u_1, \ldots, u_L \) such that

\[ \alpha = \begin{bmatrix}
T_0(u_0) \\
T_1(u_0) \\
\vdots \\
T_L(u_0)
\end{bmatrix} \begin{bmatrix}
e \\
T_0(u_1) \\
\vdots \\
T_0(u_L) + \cdots + T_{L-1}(u_1)
\end{bmatrix} \]  
(8.223)

We define \( \alpha^c \in \text{Ra } A_L \cap \text{Ra } D_L \) as the first \( L \) "components" of \( \alpha \)
\[ \alpha^c = \begin{bmatrix} T_0(u_0) \\ \vdots \\ T_{L-1}(u_0) \end{bmatrix} = \begin{bmatrix} e \\ \vdots \\ T_0(u_{L-1}) \cdots T_{L-2}(u_1) \end{bmatrix} \quad (8.224) \]

We wish to find an upper bound on the number of \( \alpha \in \text{Ra} \ A_{L+1} \cap \text{Ra} \ D_{L+1} \) that yield the same \( \alpha^c \in \text{Ra} \ A_L \cap \text{Ra} \ D_L \). Thus suppose \( \alpha \neq \beta \), but \( \alpha^c = \beta^c \). This is the same as saying there exist \( u \) and \( v \) such that

\[ \alpha = A_{L+1}(u) \neq A_{L+1}(v) = \beta \quad (8.225) \]

but

\[ \alpha = A_L(u) = A_L(v) = \beta^c \quad (8.226) \]

That is

\[ u - v \in \ker A_L \quad u - v \notin \ker A_{L+1} \quad (8.227) \]

It is clear that \( \ker A_{L+1} < \ker A_L \), and in fact is a normal subgroup because \( U \) is abelian. Thus \( \ker A_L / \ker A_{L+1} \) is a group with cardinality

\[ \text{card} \frac{\ker A_L}{\ker A_{L+1}} = \frac{\text{card} \ker A_L}{\text{card} \ker A_{L+1}} \triangleq q_L \quad (8.228) \]

Then, given \( u \) and \( v \) satisfying (8.227), we see that they must be elements of distinct cosets in \( \ker A_L / \ker A_{L+1} \), and therefore for any \( \gamma \in \text{Ra} \ A_L \cap \text{Ra} \ D_L \) there are at most \( q_L \) \( \alpha \in \text{Ra} \ A_{L+1} \cap \text{Ra} \ D_{L+1} \) such that \( q = \alpha^c \). Thus

\[ \text{card} [\text{Ra} \ A_{L+1} \cap \text{Ra} \ D_{L+1}] \leq q_L \text{ card} [\text{Ra} \ A_L \cap \text{Ra} \ D_L] \quad (8.229) \]

Also, by the First Isomorphism Theorem

\[ \text{Ra} \ A_k = U / \ker A_k \quad (8.230) \]
and (8.228) implies \( \text{card } A_k \backslash \text{card } A_j \) if \( k > j \), so

\[
\frac{\text{card } Ra A_{L+1}}{\text{card } Ra A_L} = \frac{\text{card } \ker A_{L+1}}{\text{card } \ker A_L} = q_L \tag{8.231}
\]

(i.e. the cardinality of the range increases at the same rate as the cardinality of the kernel decreases — this is the analog of the linear algebra result \( \text{dim } \text{Range } + \text{dim } \text{Nullspace } = \text{dim } \text{Domain} \) [H1], [H2]).

Combining (8.211), (8.229), and (8.231) we have

\[
\frac{\text{card } Ra M_{L+1}}{\text{card } Ra M_L} = \frac{\text{card } Ra A_{L+1}}{\text{card } [Ra A_{L+1} \cap Ra D_{L+1}]} \geq \frac{q_L \text{card } Ra A_L}{q_L \text{card } [Ra A_L \cap Ra D_L]}
\]

\[
= \frac{\text{card } Ra M_L}{\text{card } Ra M_{L-1}} \tag{8.232}
\]

We now wish to bound the maximum possible value for the minimum delay in any inverse of an abelian FGHSS. We put the direct product structure on \( U^k \) and \( Y^k \). Using (8.230), (8.211) and the fact that \( Ra A_k \cap Ra D_k < Ra A_k \), we can show that

\[
\frac{\text{card } Ra M_{k+1}}{\text{card } Ra M_k} \backslash m = \text{card } U \tag{8.233}
\]

Suppose no inverse with delay \( L \) exists. Then no inverse with delay less than \( L \) exists, and therefore

\[
\frac{\text{card } Ra M_{k+1}}{\text{card } Ra M_k} < m \quad k=0, \ldots, L \tag{8.234}
\]

where

\[
\text{card } M_0 \Delta = 1 \tag{8.235}
\]
Define
\[ r_m = \text{the largest integer less than } m \text{ that divides } m \left( \frac{m}{s_m} \right) \] (8.236)
where \( s_m \) is defined by (8.103). Combining (8.233) - (8.235), we have
\[ \text{card } \text{Ra} M_{L+1} \leq r_m \text{ card } \text{Ra} M_L \leq r_m^2 \text{ card } \text{Ra} M_{L-1} \leq \cdots \leq r_m^{L+1} \] (8.237)
(we note that one can show that (8.237) holds if and only if the system is not L-invertible). Since \( M_{L+1} : U^{L+1} \rightarrow Y^{L+1} \) and \( \text{card } U^{L+1} = m^{L+1} \), we have
\[ \text{card ker } M_{L+1} \geq s_m^{L+1} \] (8.238)
We now restrict our attention to \( \text{ker } M_{L+1} \triangleleft \eta_{L+1} \). As in the proof of Theorem 8.5, we now assume that \( u(0), \ldots, u(L) \in \eta_{L+1} \) have been applied and we apply another input \( u(L+1) \in U \). The set \( \eta_{L+1} \times U \) of such input strings is a subgroup of \( U^{L+2} \) and, referring to (8.238),
\[ \text{card} (\eta_{L+1} \times U) \geq m s_m^{L+1} \] (8.239)
Consider the input-state map \( E_{L+1} : \eta_{L+1} \times U \rightarrow X \)
\[ E_{L+1}(u(0), \ldots, u(L+1)) = b[u(L+1)] + ab[u(L)] + \cdots + a^{L+1} b[u(0)] \] (8.240)
This is a homomorphism, and, if \( \text{card} (\eta_{L+1} \times U) > n = \text{card } X \), \( E_{L+1} \)
has a nontrivial kernel — i.e. there exists a string \( u(0), \ldots, u(L+1) \)
not identically zero such that \( y(1) = y(2) = \ldots = y(L+1) = e \)
\( ((u(0), \ldots, u(L)) \in \eta_{L+1}) \) and \( x(L+2) = e \). Thus the output response to the string \( u(0), \ldots, u(L+1), e, e, \ldots, \) is the same as that to the all identity sequence. Thus the system is not invertible. Referring to (8.239), we see that the hypothesis \( \text{card} (\eta_{L+1} \times U) > n \) holds if and only if
Thus, the smallest \( L \) such that this holds is

\[
L_0 = \log_m \left( \frac{n+1}{m} \right) - 1 = q(m, \frac{n+1}{m}) - 1
\]  

(8.242)

where \( q \) is defined by (8.104). We have proven the following:

**Theorem 8.26:** Consider an abelian FGHSS with card \( U = m \), card \( X = n \).

Then if the system does not have an inverse with delay \( L_0 \), where \( L_0 \) is given by (8.242), it is not invertible. Also the system is \( L_0 \) invertible if and only if we have the following condition: given \( u(0), \ldots, u(L_0 + 1) \), such that the output of the system (started at \( x(0) = e \)) in response to this string followed by all \( e \)'s is identically \( e \), then \( u(0) = \ldots = u(L_0 + 1) = e \).

This result is the analog of Theorem 8.5 (the strengthened version of the Brockett-Mesarovic-Massey-Sain result). We also have the following results which are the analogs of the corollaries to Theorem 8.5.

**Corollary 1:** An abelian FGHSS is invertible if and only if

\[
k_e = e
\]

(8.243)

\[
P : U^{L_0 + 1} + L_0 + k(n) + 1 \to Y^{L_0 + 1} + L_0 + k(n) + 1
\]

where \( P \) is defined by

\[
P[u(0), \ldots, u(L_0 + 1)] =
\begin{bmatrix}
\bar{c}b[u(0)] \\
\bar{c}a[b[u(0)] + \bar{c}b[u(1)] + \ldots + c[b[u(L_0 + 1)]], \\
\vdots \\
\bar{c}a^2[b[u(0)] + \ldots + \bar{c}b[u(L_0 + 1)]], \\
\vdots \\
\bar{c}a^{L_0 + k(n)}[b[u(0)] + \ldots + c^{k(n) - 1}b[u(L_0 + 1)]]
\end{bmatrix}
\]

(8.244)
and \( k(n) \) is given by (8.115). That is, the system is invertible if and only if

\[
\text{card Ra } P = mq(m, \frac{n+1}{m}) \tag{8.245}
\]

**Proof:** By Theorem 8.26, our system is invertible if and only if the response to \( u(0), \ldots, u(L_0^k+1) \), \( e, e, \ldots \) is not all \( e \)'s if any of the \( u(i) \neq e \). The map \( P \) gives the first \( L_0^k+k(n)+1 \) outputs in response to this sequence. Let

\[
x = a_{L_0^k+1} b[u(0)] + a_{L_0^1} b[u(1)] + \ldots + b[u(L_0^k+1)] \tag{8.246}
\]

Then the last \( k(n) \) outputs in (8.244) are \( c(x), ca(x), \ldots, ca^{k(n)-1}(x) \). If these are all equal to \( e \), then, by Theorem 8.16, \( ca^k(x) = e \) \( \forall k \geq 0 \). Thus, we need only check the kernel of \( P \) to see if the system is invertible.

**Corollary 2:** An abelian FGHSS is invertible if and only if

\[
\text{card Ra } M_{L_0^k+1} \geq r_m + 1 \tag{8.247}
\]

**Proof:** This follows from (8.237).

We note that one can make comments relating these results to the LSC results much like those comments made following Example 8.6.

In addition, one could consider an output equation of the form

\[ y(k) = d[u(k)]c[x(k)] \]

and derive analogous results to those of Section 8.2 and reference [M10]. We also have the following example that shows that the bound in Theorem 8.26 can be realized. Other examples will be given in Section 8.5.

**Example 8.7:** Let \( U = Y = Z_2, X = Z_2^r \) with \( r \geq 2 \) and define \( a : X \to X, b : U \to X, c : X \to Y \) as follows:
\[ a(x_1, \ldots, x_r) = (0, x_1, \ldots, x_{r-1}) \quad (8.248) \]
\[ b(x) = (x, 0, \ldots, 0) \quad (8.249) \]
\[ c(x_1, \ldots, x_r) = x_r \quad (8.250) \]

Then
\[ cb = cab = \ldots = ca^{r-2} = 0 \quad (8.251) \]
\[ ca^{r-1}b = \text{identity} \quad (8.252) \]

Thus the system is \((r-1)\)-invertible but not \(k\)-invertible for \(k < (r-1)\), and
\[ L_o = q(2, 2^{r-1} + \frac{1}{2}) - 1 = r-1 \quad (8.253) \]

We now briefly consider the construction of inverses for \(\text{FGHSS}'s\). We will explicitly consider the analog of Theorem 5 in [M10]. Suppose \(T_0, \ldots, T_{i-1}\) are the trivial homomorphism (map everything onto \(e\)), and \(T_i\) is one-to-one (\(\text{card } Ra T_i = m\)). We then have
\[ y(1) = y(2) = \ldots = y(i) = e \quad (8.254) \]
\[ y(k+i+1) = T_i[u(k)] ca^{i+1}[x(k)] \quad (8.255) \]

We now must make an assumption that is not necessary for \(\text{LSC}'s\) but is for \(\text{FGHSS}'s\) (even in general for abelian \(\text{FGHSS}'s\)). We assume that there is a normal subgroup \(N\) of \(Y\) and a homomorphism \(\theta : T_i(U) \rightarrow \text{Aut}(N)\) such that
\[ Y = N \times_\theta T_i(U) \quad (8.256) \]

(see the definition of semidirect product in Appendix A). The First Isomorphism Theorem tells us that \(U = T_i(U)\), and (8.256) implies that there exists a homomorphism \(M : Y \rightarrow U\) such that \(M \circ T_i \in \text{Aut}(U)\). Then
from (8.255)

\[ u_k = (M \circ T_1)^{-1} [M[y(k+i+1)]Mca^{i+1}[x(k)^{-1}]] \]  

(8.257)

\[ x_{k+1} = \{b(M \circ T_1)^{-1} M[y(k+i+1)]\} \{b(M \circ T_1)^{-1} Mca^{i+1}[x(k)^{-1}]\} a(x_k) \]

These equations define an inverse of our FGHSS, but it may not be homomorphic. Indeed, although it is homomorphic if the system is abelian, it is not homomorphic in any other case because the map

\[ x \rightarrow x^{-1} \]  

(8.258)

is not a homomorphism (it is an anti-homomorphism -- i.e. \( f(xy) = f(y)f(x) \)).

**Example 8.8:** We present two examples that show that (8.256) need not hold -- i.e. we will choose \( U, Y \) and \( f : U \rightarrow Y \) a homomorphism with \( \ker f = \{e\} \) such that there exists no homomorphism \( g : Y \rightarrow U \) such that \( g \circ f \in \text{Aut}(U) \). Let \( U = Z_4, Y = D_4 \) and

\[ f(n) = x^n \quad n = 0, 1, 2, 3 \]  

(8.259)

The reader can check that there is no \( g \) in this case. A second example indicates that we still have problems even in the abelian case. Let \( U = Z_2, Y = Z_4 \) and

\[ f(0) = 0 \quad f(1) = 2 \]  

(8.260)

The reader can check that again there is no \( g \).

We refer the reader to [ML] and [RL] to see that there exists a homomorphism \( M : Y \rightarrow U \) such that \( M \circ T_1 \in \text{Aut}(U) \) if and only if (8.256) holds.
The construction of inverses in the general FGHSS case and the
determination of conditions under which the inverse is a FGHSS appear
to be very difficult problems. This is one of a number of unanswered
questions concerning FGHSS's. Another of these concerns feedback.
Note that a linear system with linear feedback is itself a linear
system. However consider the FGHSS (8.33), (8.34) and suppose we have
"homomorphic" feedback

\[ u(k) = v(k)g[x(k)] \]  \hspace{1cm} (8.261)

where \( g : X \to U \) is a homomorphism. Then the FGHSS becomes

\[ x(k+1) = b[v(k)]bg[x(k)]a[x(k)] \]  \hspace{1cm} (8.262)

and the map

\[ x \mapsto bg(x)a(x) \]  \hspace{1cm} (8.263)

need not be a homomorphism unless the system is abelian.

Other problems include the partial realization problem (see [K11],
[R8]), bounding maximum inverse delays and \( k \) such that \( R = R_k \) in the
general FGHSS case, and a number of other questions that will be dis-
cussed in Chapter 11.

8.5 A Class of Algebraic Coding Systems

In this section we will discuss briefly an extension of a particular
class of algebraic coding systems known as convolutional coding systems.
We will first review some of the basic facts about convolutional codes
(see [M11], [M13], [F4], [V3], and [K14]). A convolutional encoder (CE)
is a LSC (over a finite field), started in the 0 state, with \( m \) inputs
and p outputs. The strings of inputs can be regarded as "data", and the output strings as "codewords." The basic job of a coding system is to transmit the data as a codeword to some point at which we recover the data from the received codeword. Thus, in order for there to be any chance of doing this properly, even if there is no noise in the system, the input-output map (8.8) must be one-to-one -- i.e., the system must be invertible. This places the first restriction on those LSC's to be used as CE's. Some of the other restrictions are related to coding - theoretic concepts which we will not go into in detail since this discussion is meant only as a brief description of a possible application for FCHSS's. See [F4] and [B17] for details.

We mention two of these coding-theoretic concepts in order to indicate the type of questions one must ask when designing coding systems. The first is the concept of equivalent encoders. Suppose we assume that all data strings are equally likely. Then all codewords are equally likely and what is important is how "different" or "far apart" the various codewords are (e.g. the Hamming distance [B17], [P3] between two codewords is the number of points in time at which the codewords differ) as opposed to which codeword corresponds to which data sequence. That is, we are concerned with the range (called the code) of the input sequence - output sequence map rather than the map itself.

The second concept is that of catastrophic error propagation. Basically what this refers to is a LSC such that there exists an infinite input sequence with infinitely many nonzero elements such that the output of the LSC in response to this string is nonzero in only a finite
number of places. In this case a finite number of transmission errors will cause the output to appear to be the response to the all zero input sequence — i.e. we will make an infinite number of decoding errors because of a finite number of transmission errors. Consider the following example [M10] of a one-input two-output LSC defined over the binary field GF(2).

\[ x(k+1) = Ax(k) + bu(k) \]  \hspace{1cm} (8.264)  

\[ y(k) = Cx(k) \]  \hspace{1cm} (8.265)  

where \( x \in (Z_2)^3 \), \( u \in Z_2 \), \( y \in (Z_2)^2 \) and

\[ A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \]  \hspace{1cm} (8.266)  

One can check that this is controllable, observable, and invertible.

Consider the input string that is identically 1. Then, assuming \( x(0) = 0 \),

\[ y(1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad v(2) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad y(k) = 0 \quad k > 2 \]  \hspace{1cm} (8.267)  

and, in this case, if three errors are made (\( v_1(1) = v_2(1) = v_2(2) = 0 \)) the output sequence will be identically 0.

Several comments are in order. First of all, the assumption that all codewords are equally likely may not always be a good one.

Assuming that the actual information to be transmitted has been reduced to a string of input bits, the assumption that the individual bits are independent is sometimes a good one. Also one can argue that a coding system is built not knowing the information source, so the assumption of equally likely input strings seems reasonable. However, in any specific
application, different input strings correspond to different information which may have different probabilities, and such information might prove useful for decoding purposes. It is this type of consideration that will lead us to consider some probabilistic decoding questions in Chapter 9.

Also, we note that the catastrophic error propagation problem arises only if the input string has infinitely many nonzero inputs. If we know that only a finite number N of bits are to be coded, then the system that was "catastrophic" before may not be so bad. However, the argument that N may be very large and also may not be known at all is quite strong.

The purpose of the remainder of this section is the following: since the class of FGHSS's is larger than the class of LSC's, can we find a class of encoding systems that is substantially larger than that of convolutional encoders? The purpose for doing such a study would be to find "good" codes (large distances between codewords) that can be implemented easily (i.e. as FGHSS's) but are not convolutional codes. To this end we define a finite group homomorphic encoder (FGHE) as an invertible FGHSS. An interesting problem is to determine the analogs of the coding theory concepts used to study CE's, as in [F4] and [M10]. We have not done this, but include two examples that indicate that the class of FGHE's contains encoders that are not CE's. The reader is also referred to the work of Johnston [J3], in which a number of codes generated by abelian FGHSS's are considered.
Example 8.9: We first consider an abelian example that is not a LSC.

Let \( U = \mathbb{Z}_3 \times \mathbb{Z}_2, \ X = \mathbb{Z}_6 \times \mathbb{Z}_6, \ Y = \mathbb{Z}_6 \) and define the homomorphisms
\( a : X \to X, \ b : U \to X, \ c : X \to Y \) by

\[
a(x,y) = (x+y,x) \quad \quad (8.268)
\]
\[
b(1,1) = (2,3) \quad \quad (8.269)
\]
\[
c(x,y) = x \quad \quad (8.270)
\]

We can check that this system is controllable and observable. Also, consider the definitions of \( A_k \) and \( D_k \) in (8.177) and (8.178). We can compute

\[
\text{Ra } A_1 = \{0, 2, 4\} < Y \quad \quad (8.271)
\]
\[
\text{Ra } A_2 = \{(0,0), (2,2), (4,4), (0,3), (2,5), (4,1)\} < Y^2 \quad \quad (8.272)
\]
\[
\text{Ra } D_1 = \{0\} \quad \quad (8.273)
\]
\[
\text{Ra } D_2 = \{(0,0), (0,2), (0,4)\} \quad \quad (8.274)
\]

Then using (8.211)

\[
\frac{\text{card Ra } M_1}{\text{card Ra } M_0} = \frac{\text{card Ra } A_1}{\text{card} [\text{Ra } A_1 \cap \text{Ra } D_1]} = 3 \quad \quad (8.275)
\]
\[
\frac{\text{card Ra } M_2}{\text{card Ra } M_1} = 6 = \text{card } U \quad \quad (8.276)
\]

so no inverse of delay zero exists, but an inverse of delay one does.

Thus this system is a FGHE.

Example 8.10: We now give a non-abelian example. We define the quaternion group \( Q \) (not to be confused with the quaternions discussed in Chapter 7) as the group of order 8 having two generators \( p \) and \( t \) satisfying the relations
\[ p^4 = e \quad t^2 = p^2 \quad pt = tp^3 \]  
\[ (8.277) \]

This group is isomorphic to the group of 2x2 complex matrices (under matrix multiplication) generated by

\[ P = \begin{bmatrix} 0 & 1 \\ i & 0 \end{bmatrix} \quad T = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \]  
\[ (8.278) \]

The quaternion group is also isomorphic to the set \( \{ \pm 1, \pm i, \pm j, \pm k \} \) under the operation

\[ i^2 = j^2 = k^2 = -1 \quad ij = -ji = k \quad jk = -jk = i \quad ki = -ik = j \]  
\[ (8.279) \]

(note the relationship between this representation of \( Q \) and the quaternions of Chapter 7). One should note that \( Q \) is a Hamiltonian group [R1] -- i.e. all its subgroups are normal. Thus, since the product of normal subgroups is a subgroup, the reachable set for any FGHSS with \( Q \) or any other Hamiltonian group as the state group is a group.

Consider the following FGHSS: \( U = Z_4 \), \( X = Q \times O \), \( Y = Q \), and define \( b : U \to X \), \( c : X \to Y \), and \( a : X \to X \) to be the homomorphisms uniquely defined by

\[ b(1) = (e, p) \]  
\[ (8.280) \]

\[ c(q_1, q_2) = q_1 \quad \forall q_1, q_2 \in Q \]  
\[ (8.281) \]

\[ a(q_1, q_2) = (f(q_2), q_1) \quad \forall q_1, q_2 \in Q \]  
\[ (8.282) \]

where \( f : Q \to Q \) is the homomorphism uniquely defined by

\[ f(p) = p^3 t \quad f(t) = t \]  
\[ (8.283) \]

One can check that this system is controllable and observable. Also
\[ Ra A_1 = \{e\} < Y \]  \hspace{1cm} (8.284)

\[ Ra A_2 = \{(e,e),(e,p^3 t),(e,p^2),(e,pt)\} < Y \]  \hspace{1cm} (8.285)

\[ Ra D_2 = \{(e,e)\} < Y \]  \hspace{1cm} (8.286)

so from (8.179) (we cannot use (8.211) because \( Y \) is not abelian)

\[ \text{card } Ra M_1 = 1 \]  \hspace{1cm} (8.287)

\[ \text{card } Ra M_2 = 4 = \text{card } U \]  \hspace{1cm} (8.288)

Thus no zero delay inverse exists but a one-delay inverse does.

We note that these examples are relatively simple; however the theory of groups is so rich that it is hoped that the class of FCHE's will include many new types of encoders. We also note that the theory of FCHSS's may be useful in providing methods for optimizing the basic recursions that occur in certain elementary numerical processes. We will not comment on this general problem, but instead, in the next two chapters we will present some extensions and applications of this finite group theory.
CHAPTER 9

GROUP ALGEBRA FORMULATION OF ESTIMATION PROBLEMS FOR FGHSS'S AND SOME COMPUTATIONAL CONSIDERATIONS

"'Just watch this, Linus,' he whispered. 'I'll whip through this problem like a human computer and really impress that little red-haired girl!'"

- Charles M. Schultz [S11]

9.1 Introduction

In the next two chapters we will consider some problems associated with the study of random processes evolving on finite groups (and, in some cases, semigroups). We seek to devise methods for analyzing the interaction of the probabilistic and algebraic properties of certain dynamical systems (namely FGHSS's). Our methods are motivated by the very important work of Grenander [G8] in this field. Our tools - a generalized notion of convolution (Chapter 9) and Fourier-like analysis (Chapter 10) - are quite similar to the techniques used by Grenander.

The original motivation for the analysis of Chapters 9 and 10 came from an attempt to learn something about random processes on $S^1$ by studying processes defined on $Z_n$. This special case, which will be considered in Chapter 10, provides the basic motivation for the general approach considered in the present chapter. In Section 9.2 we introduce another algebraic structure -- the group algebra. The motivation for the introduction of this concept comes from a desire to generalize and exploit the relationship between the convolution of two discrete time
sequences and the multiplication of two polynomials. The following example illustrates this relationship.

**Example 9.1:** Consider the ring of polynomials \( R^1[s]/(s^n-1) \) -- i.e. the set of polynomials \( \{ a_0 + a_1 s + \ldots + a_{n-1} s^{n-1} | a_i \in R^1 \} \) with polynomial multiplication performed with the additional rule that \( s^n \equiv 1 \) (see [Fl] for more on rings of polynomials). If we multiply two such polynomials

\[
c_0 + c_1 s + \ldots + c_{n-1} s^{n-1} = (a_0 + a_1 s + \ldots + a_{n-1} s^{n-1})(b_0 + b_1 s + \ldots + b_{n-1} s^{n-1})
\]

we have that

\[
c_k = \sum_{i=0}^{k} a_i b_{k-i} + \sum_{i=k+1}^{n-1} a_i b_{n+i-k-1}, \quad 0 \leq k \leq n-1
\]

(9.2)

Note that the right-hand side of (9.2) looks like a convolution. Now suppose we consider two independent random variables \( y \) and \( z \) taking values in \( Z_n \). Let their probability distributions be

\[
\eta_y = \begin{bmatrix}
\eta_0 \\
\eta_1 \\
\vdots \\
\eta_{n-1}
\end{bmatrix}, \quad \eta_z = \begin{bmatrix}
\eta_0 \\
\eta_1 \\
\vdots \\
\eta_{n-1}
\end{bmatrix}, \quad \eta_z = \begin{bmatrix}
\Pr(z=0) \\
\Pr(z=1) \\
\vdots \\
\Pr(z=n-1)
\end{bmatrix}
\]

(9.3)

Let \( w \) be the group product of \( y \) and \( z \)

\[
w = (y+z) \mod n
\]

(9.4)

with distribution

\[
\xi_w = \begin{bmatrix}
\xi_0 \\
\vdots \\
\xi_{n-1}
\end{bmatrix}, \quad \xi_w = \begin{bmatrix}
\Pr(w=0) \\
\vdots \\
\Pr(w=n-1)
\end{bmatrix}
\]

(9.5)
since \( y \) and \( z \) are independent, one can compute

\[
\xi_k = \sum_{i=0}^{k} \alpha_i \eta_{k-i} + \sum_{i=k+1}^{n-1} \alpha_i \eta_{n+k-i}
\]  

(9.6)

Comparing (9.2) and (9.6), we see that if we regard \( \rho_y, \eta_z, \) and \( \xi_w \) as polynomials in \( \mathbb{R}^1[s]/(s^n-1) \), then the probability distribution for the group product of \( y \) and \( z \) is just the polynomial product of \( \rho_y \) and \( \eta_z \). As we shall see in the next section, \( \mathbb{R}^1[s]/(s^n-1) \) is the real group algebra of \( Z_n \).

In Section 9.2 we consider a possibly time-varying FGHSS with random inputs and noise observations. By identifying probability distributions on a finite group with elements of a group algebra, we will be able to write conditional probability evolution equations that have a structure that is quite similar to the structure of the FGHSS.

Section 9.3 contains one of the basic benefits to be reaped by viewing the probability distributions in this manner. The basic problem in this section is to devise efficient methods for computing the conditional density. Again, the motivation for this work comes from the \( Z_n \) problem. In that case, we are interested in multiplying two polynomials \( \mod(s^n-1) \). We will find that techniques based on the highly efficient fast Fourier transform (FFT) are applicable to the \( Z_n \) problem (see [C6], [N3], [C7], [C9], [D5], [D6], [K15], [S12]). We will then consider extensions of this to the general FGHSS case.

The problem of computing group algebra products efficiently is stated, but is not (and, as far as the author knows, has not been) solved in general. Instead, we will give several examples using a brute force
technique of Brockett and Dobkin [B18], [D4] to devise efficient computational schemes. The savings achieved even in these simple examples indicates that a theoretical study of efficient group algebra multiplication may reap great computational benefits.

The possibility of applying and extending these computational considerations to the simplification of the computations in other problems (including some of the discrete-time Fourier series computations of Chapter 5) will be considered, as well as the extension of the group algebra approach to include some problems related to stochastic automata [K11], [L10], [R4]. We will study a fairly general class of finite state Markov processes (FSMP) and will consider using Myhill's semigroup equation (see Appendix E, [K11], and [M12]) plus our group algebra techniques to find methods for performing the necessary computations. The reader is referred to the work of Astrom [A10] on the optimal control of FSMP's with noisy observations to see the type of computational demands we are dealing with.

In Section 9.4, we briefly consider some aspects of probabilistic decoding for CE's and FGHE's. This discussion, although only superficial, indicates another possible area in which our results can be applied.

We note that we do not specifically consider the optimal estimation problem in this chapter. This is clearly a major gap in the theory, since, for example, optimal estimation is of great importance in probabilistic decoding. However, our results are potentially valuable for several reasons: (1) the necessary on-line computations in Astrom's work essentially consist of the computation of conditional distributions;
(2) it is hoped that the structure of these conditional distribution
equations will eventually lead to recursive estimation equations that
can be computed efficiently. The reader is referred to Chapter 10,
in which a concept related to the group algebra -- that of group characters --
is used much like Fourier series to formulate and solve some optimal
estimation problems on $\mathbb{Z}_n$.

9.2 Conditional Distribution Equations on Groups

We will now develop the mathematical machinery we will need to study
random processes on FGHSS's. To do this, we need the concept of a group
algebra. Our definition is somewhat specialized. The reader is referred
to [N3], [C8], [F5] for further discussions of this topic and its use in
the theory of group representations.

Definition 9.1: Let $G$ be a finite group and $F$ any field. The group algebra
$F(G)$ of $G$ over $F$ is the algebra of all formal sums

$$
\sum_{g \in G} \alpha_g \cdot g \quad \forall \alpha_g \in F \quad \forall g \in G
$$

(9.7)

with addition, scalar multiplication, and algebra multiplication defined
as follows:

$$
\sum_{g \in G} \alpha_g \cdot g + \sum_{g \in G} \beta_g \cdot g = \sum_{g \in G} (\alpha_g + \beta_g) \cdot g
$$

(9.8)

$$
\alpha( \sum_{g \in G} \alpha_g \cdot g) = \sum_{g \in G} (\alpha \alpha_g) \cdot g \quad \forall \alpha \in F
$$

(9.9)

$$
( \sum_{g \in G} \alpha_g \cdot g) \cdot ( \sum_{h \in G} \beta_h \cdot h) = \sum_{g, h \in G} (\alpha_g \beta_h) \cdot (gh) = \sum_{t \in G} \gamma_t \cdot t
$$

(9.10)
where
\[ \gamma_t = \sum_{g \in G} \alpha_g \beta_{g^{-1} t} = \sum_{g \in G} \alpha_{tg^{-1}} \beta_g \]  
(9.11)

One can regard (9.11) as a generalized convolution. We remark that the group algebra multiplication operation will be written with "·" omitted. Also we can regard \( G \) as a subset of \( \mathbb{F}(G) \) with the obvious identification
\[ g \rightarrow l \cdot g \]  
(9.12)
where \( l \) is the multiplicative identity in \( \mathbb{F} \). We will usually suppress the \( l \). Note that if we write \( Z_n \) as the set \{1, s, \ldots, s^{n-1}\} with the operation
\[ s^j s^k = s^{(j+k) \mod n} \]  
(9.13)
it is clear that \( \mathbb{R}^1[s]/(s^n - 1) \) is the real group algebra \( \mathbb{R}^1[Z_n] \). We will concern ourselves only with real group algebras.

Another binary operation can be defined on group algebras -- the \textit{Hadamard product}, [H1]:
\[ (\sum_{g \in G} \alpha_g \cdot g) \ast (\sum_{g \in G} \beta_g \cdot g) = \sum_{g \in G} (\alpha_g \cdot \beta_g) \cdot g \]  
(9.14)
With this operation replacing group algebra multiplication, \( \mathbb{F}(G) \) becomes a commutative algebra, independent of the structure of \( G \).

Also note that as a vector space over \( \mathbb{F} \), \( \mathbb{F}(G) \) has dimension \( n = \text{card } G \).

We define the class of systems to be studied. Let \( X, U, \) and \( Y \) be finite groups and consider the time-varying random FCHSS
\[ x(k+1) = b_k[u(k)]a_k[x(k)] \]  
(9.15)
\[ v(k) = v(k)c_k[x(k)] \quad k \geq 0 \]  
(9.16)
where we make the assumptions: (1) \{u(k)\} and \{v(k)\} are sequences of independent random variables in \(U\) and \(V\), respectively, independent of each other and of \(x(0)\); (2) \(a_k : X \to X\), \(b_k : U \to X\), and \(c_k : X \to Y\) are group homomorphisms which are known, as are the distributions for \(u(k)\), \(v(k)^{-1}\), and \(x(0)\), denoted \(\eta(k)\), \(\xi(k)\), and \(\rho(0)\), respectively.

Let \(\rho(k)\) be the unconditional distribution for \(x(k)\), and \(\rho(k|l)\) the distribution for \(x(k)\) given \(v(0), \ldots, v(l)\). By convention, we set \(\rho(0) = \rho(0|0)\), and we regard all of these distributions on \(X\) as elements of the group algebra \(R^1(X)\). For example,
\[
\rho(k) = \sum_{g \in G} \rho(k)_g \cdot g \tag{9.17}
\]
\[
\rho(k)_g = \Pr(x(k) = g) \tag{9.18}
\]
Similarly we regard \(\eta(k) \in R^1(U)\) and \(\xi(k) \in R^1(Y)\). Our problem is to compute \(\rho(k|m)\) for various values of \(m\) and for all \(k \geq 0\). Here \(m > k\) is the smoothing problem, \(m = k\) is the filtering problem, and \(m < k\) is the prediction problem. As we shall see, the filtering distributions will be important in devising methods for computing the other distributions. Examples will be given at the end of this section and in the next.

As one might expect, to solve the prediction problem we look for a diffusion-type equation for the propagation of the unconditional distribution. Because of the independence of the \(\{u(k)\}\) and \(\{v(k)\}\), such an equation can be used to compute \(\rho(k|m)\), \(k > m\), if the filtering solution \(\rho(m|m)\) is known.
Theorem 9.1 (Diffusion Equation): Consider the input-state part of a random FGHSS, (9.15). The recursive equation for the unconditional distribution \( \rho(k) \) is given by the following group algebra-homomorphic recursion in \( R^1(X) \):

\[
\rho(k+1) = \hat{b}_k[\eta(k)] \hat{a}_k[\rho(k)] \tag{9.19}
\]

where \( \hat{b}_k : R^1(U) \rightarrow R^1(X) \) and \( \hat{a}_k : R^1(X) \rightarrow R^1(X) \) are defined by

\[
\hat{b}_k \left( \sum_{g \in U} \alpha \cdot g \right) \triangleq \sum_{g \in U} \alpha \cdot \hat{b}_k(g) \tag{9.20}
\]

\[
\hat{a}_k \left( \sum_{h \in X} \beta \cdot h \right) \triangleq \sum_{h \in X} \beta \cdot \hat{a}_k(h) \tag{9.21}
\]

Proof: The independence assumption on the \( \{u(k)\} \) assures the validity of the following equation:

\[
\rho(k+1) = \text{Pr}(x(k+1)=g) = \sum_{h \in X} \text{Pr}(b_k[u(k)]=h) \text{Pr}(a_k[x(k)]=h^{-1}g) \]

\[
= \sum_{h \in X} \left\{ \sum_{\mu \in b_k^{-1}(h)} \text{Pr}(u(k)=\mu) \right\} \left\{ \sum_{\nu \in a_k^{-1}(h^{-1}g)} \text{Pr}(x(k)=\nu) \right\} \]

\[
= \sum_{h \in X} \left\{ \sum_{\mu \in b_k^{-1}(h)} \eta(k)_{\mu} \right\} \left\{ \sum_{\nu \in a_k^{-1}(h^{-1}g)} \rho(k)_{\nu} \right\} \tag{9.22}
\]

One can check (9.20) and (9.21) to see that

\[
(\hat{b}_k[\eta(k)])_h = \sum_{\mu \in b_k^{-1}(h)} \eta(k)_{\mu} \quad \forall \ h \in X \tag{9.23}
\]

\[
(\hat{a}_k[\rho(k)])_h = \sum_{\nu \in a_k^{-1}(h)} \rho(k)_{\nu} \quad \forall \ h \in X \tag{9.24}
\]

Combining (9.22), (9.23), and (9.24), we have

\[
\rho(k+1) = \sum_{h \in X} (\hat{b}_k[\eta(k)])_h (\hat{a}_k[\rho(k)])_{h^{-1}g} \tag{9.25}
\]
and referring to (9.11), we see that (9.19) is correct. The fact
that \( \hat{a}_k \) and \( \hat{b}_k \) are group algebra homomorphisms -- i.e., that

\[
\hat{b}_k(\alpha n_1 + \beta n_2) = \alpha \hat{b}_k(n_1) + \beta \hat{b}_k(n_2)
\]

(9.26)

\[
\hat{b}_k(n_1 n_2) = \hat{b}_k(n_1) \hat{b}_k(n_2)
\]

(9.27)

hold for all \( \alpha, \beta \in \mathbb{R}^1 \) and \( n_1, n_2 \in \mathbb{R}^1(U) \), and analogous equations
hold for \( \hat{a}_k \) -- is proven by straightforward calculations, using the
fact that \( a_k \) and \( b_k \) are group homomorphisms.

Notice that the diffusion equation (9.19) still holds even if
\( a_k \) and \( b_k \) aren't homomorphisms, but in that case \( \hat{a}_k \) and \( \hat{b}_k \) are not
homomorphisms.

\[ \star \]

**Theorem 9.2 (Filtering Distribution Equation):** Consider the random
FCHSS (9.15), (9.16). We can compute \( \rho(k+1|k) \) from \( \rho(k|k) \) and
\( \rho(k|k) \) from \( \rho(k|k-1) \) as follows:

**Diffusion Update**

\[
\rho(k+1|k) = \hat{b}_k[n(k)] \hat{a}_k[\rho(k|k)]
\]

(9.28)

**Measurement Update**

\[
\rho(k|k) = \frac{\gamma(k|k)}{n(k|k)}
\]

(9.29)

where \( \gamma(k|k) \) is the unnormalized density and \( n(k|k) \) the normalizing
constant given by

\[
\gamma(k|k) = \hat{c}_k[\xi(k) v(k)] \ast \rho(k|k-1)
\]

(9.30)

\[
n(k|k) = ev[\gamma(k|k)]
\]

(9.31)

Here \( \hat{c}_k : \mathbb{R}^1(Y) \to \mathbb{R}^1(X) \) and \( ev : \mathbb{R}^1(X) \to \mathbb{R}^1 \) are defined by
\[ \hat{c}_k \left[ \sum_{g \in Y} \alpha_g \cdot g \right] \Delta \sum_{g \in Y} \alpha_g \cdot \left( \sum_{h \in C_k^{-1}(g)} g \right) \] (9.32)

\[ \text{ev} \left( \sum_{g \in X} \alpha_g \cdot g \right) \Delta \sum_{g \in X} \alpha_g \cdot g \] (9.33)

We note that in (9.30) we regard \( y(k) \) as an element of \( R^1(Y) \), and in (9.32) we regard \( h \) as an element of \( R^1(X) \).

**Proof:** The diffusion update follows from Theorem 9.1 and the independence assumptions. Thus we need only check the \( \gamma(k|k) \) equation, since the validity of the \( \eta(k|k) \) equation (assuming \( \gamma(k|k) \) is correct) is obvious. If (9.29) is correct, Bayes' rule yields

\[
\frac{\gamma(k|k)_g}{\sum_{h \in X} \gamma(k|k)_h} = \frac{\rho(k|k) g \Pr[x(k) = g | y(0), \ldots, v(k)]}{\Pr[y(k) | x(k) = g, y(0), \ldots, y(k-1)] \Pr[x(k) = g | y(0), \ldots, y(k-1)]}
\]

\[
= \frac{\Pr[y(k) | x(k) = g, y(0), \ldots, y(k-1)] \Pr[x(k) = g | y(0), \ldots, y(k-1)]}{\sum_{h \in X} \Pr[y(k) | x(k) = h, y(0), \ldots, y(k-1)] \Pr[x(k) = h | y(0), \ldots, y(k-1)]}
\]

\[
= \frac{\Pr[y(k) | x(k) = g] \rho(k|k-1)_g}{\sum_{h \in X} \Pr[y(k) | x(k) = h] \rho(k|k-1)_h}
\] (9.34)

The last equality follows from the definition of \( \rho(k|k-1) \) and the fact that \( v(k) \) is independent of \( v(0), \ldots, v(k-1) \) and \( \{u(m)\} \).

Examining (9.34), (9.29), and (9.30), we see that (9.29) and (9.30) will be correct if we can show that

\[
\hat{c}_k [\xi(k) y(k)]_g = \Pr[y(k) | x(k) = g]
\] (9.35)

First, we rewrite the right-hand side of (9.35):
\[ Pr[y(k)|x(k) = g] = Pr[v(k) = v(k)c_k(g)^{-1}] \]
\[ = Pr[v(k)^{-1} = c_k(g)y(k)^{-1}] \]
\[ = \xi(k)c_k(g)y(k)^{-1} \] (9.36)

We also write
\[ \hat{\xi}_k[\xi(k)v(k)] = \hat{\xi}_k[\sum_{h \in Ye} \xi(k)_h \cdot (hv(k))] \]
\[ = \sum_{h \in Ye} \xi(k)_h \cdot \sum_{t \in e c_k^{-1}(hy(k))} t \] (9.37)

We wish to compute \( \hat{c}_k[\xi(k)y(k)]_g \) -- i.e., the g-component of \( \hat{c}_k[\xi(k)v(k)] \).

Examining (9.37), we see that
\[ g \in e c_k^{-1}(hy(k)) \iff c_k(g) = hy(k) \iff h = c_k(g)v(k)^{-1} \] (9.38)

Thus
\[ \hat{\xi}_k[\xi(k)v(k)]_g = \xi(k)c_k(g)y(k)^{-1} \] (9.39)

and (9.35) and therefore (9.29) and (9.30) are correct.

It remains to check that \( \hat{c}_k \) is a Hadamard homomorphism. Linearity is simple to check. Also
\[ \hat{c}_k((\sum_{g \in Ye} \alpha \cdot g) \ast (\sum_{g \in Ye} \beta \cdot g)) = \sum_{g \in Ye} \alpha \ast g \cdot (\sum_{h \in ec_k^{-1}(g)} h) \]
\[ = \{ \sum_{g \in Ye} \alpha \cdot h \} \ast \{ \sum_{g \in Ye} \beta \cdot h \} \]
\[ = \{ \hat{c}_k(\sum_{g \in Ye} \alpha \cdot g) \} \ast \{ \hat{c}_k(\sum_{g \in Ye} \beta \cdot g) \} \] (9.40)

We remark that \( \hat{c}_k \) is a "pullback" type of map -- i.e., it is uniquely determined by \( c_k : X \to Y \), but \( \hat{c}_k : R^1(Y) \to R^1(X) \) (i.e., it
reverses direction). Note that analogously \( \hat{a}_k \) and \( \hat{b}_k \) could be called "pushforward" maps, since, for example, \( b_k : U \to X \) and \( \hat{b}_k : R^1(U) \to R^1(X) \) (preserves direction). Note that \( \hat{c}_k \) being a Hadamard homomorphism has nothing to do with \( c_k \) being a homomorphism. The fact that \( c_k \) is a homomorphism is useful in computing \( c_k^{-1}(g) \), since we know from the First Isomorphism Theorem of group theory that

\[
c_k(X) = X/\ker c_k
\]  

(9.41)

Thus any element of \( Y \) either has no pre-image points or a fixed number of them, which can be computed by knowing any one pre-image point and the kernel of \( c_k \). Note also that in the measurement update equation the measurement \( y(k) \) enters multiplicatively. Recall that \( \xi(k) \) is the distribution for \( \nu(k)^{-1} \). If we used the distribution for \( \nu(k) \) instead, we would find that \( y(k)^{-1} \) would enter the equation for \( \gamma(k|k) \) (see (9.36)).

The computations involved in using the filtering equations deserve some comment. First note that the multiplication of \( \xi(k) \) by \( y(k) \) results simply in a permutation of the elements of \( \xi(k) \), and the operation \( \hat{c}_k(\eta) \) only involves additions of elements of \( \eta \). As we shall discuss in the next section, multiplications and divisions require the most computation time, and the amount of time needed to compute \( \hat{c}_k[\xi(k)\nu(k)] \) is not significant. However, the division of \( \gamma(k|k) \) by \( n(k|k) \) in (9.29) is time consuming (\( n \) divisions if \( |X| = n \)). These divisions can be avoided if we don't normalize the probability distribution. For instance, since all of the terms in \( \gamma(k|k) \) are divided by the same term, the relative sizes of the terms in \( \rho(k|k) \) is the same as in \( \gamma(k|k) \). Thus, if we are
interested in estimating $x(k)$ by choosing the most probable value 
(conditioned on the observations), we need only work with $\gamma(k|k)$ 
and needn't divide by $n(k|k)$. In this manner, large computational 
savings can be made. The reader is referred to the next section 
in which we consider several other computational questions.

The motivation for the use of the symbol "ev" for the normalizing 
constant computation comes from an examination of the relationship 
between group algebras and polynomials. One can define a finite group 
$G$ by specifying a set of $r$ (in general noncommuting) generators 
$\{g_1, \ldots, g_r\}$ and the set $R$ of relations they satisfy. Let $\{g_1, \ldots, g_r\}^*$ 
be the free monoid generated by $g_1, \ldots, g_r$ -- i.e. the set of elements 
of the form

$$g_{i_1}^{k_1} g_{i_2}^{k_2} \cdots g_{i_m}^{k_m} \quad m \geq 0 \quad k_m \geq 0 \quad (9.42)$$


where the $g_{i_j}$'s can be repeated, except in the case in which $i_j = i_{j+1}$.

In this case we write

$$g_{i_j}^{k_j} g_{i_{j+1}}^{k_{j+1}} = g_{i_j}^{(k_j + k_{j+1})} \quad (9.43)$$

The monoid operation on $\{g_1, \ldots, g_r\}^*$ is concatenation -- i.e. the 
product of $g_{i_1}^{l_1} \cdots g_{i_m}^{l_m}$ with $g_{j_1}^{k_1} \cdots g_{j_n}^{k_n}$ is $g_{i_1}^{l_1} \cdots g_{i_m}^{l_m} g_{j_1}^{k_1} \cdots g_{j_n}^{k_n}$.

We include $\Lambda$, the empty string, in $\{g_1, \ldots, g_r\}^*$ and note that $\Lambda$ is 
the identity element -- i.e. $\Lambda a = a \Lambda = a \forall a \in \{g_1, \ldots, g_r\}^*$.

We then can consider $\{g_1, \ldots, g_r\}^*$ modulo the relations $R$ -- i.e. we 
use the relations to simplify expressions like (9.42). Then we have

$$G = \{g_1, \ldots, g_r\}^*/R \quad (9.44)$$
As an example, let $G = D_4$, which is generated by the elements $x$ and $y$ and the set of relations

$$ R = \{ x^4 = 1, y^2 = 1, xyx = y \} \quad (9.45) $$

Let us reduce an element of $\{x,y\}^*$. Let

$$ \alpha = x^3y^3xy $$

Then using elements of $R$, we have

$$ \alpha = x^3yxyx^3 = x^3yxyx = x^5 = x \quad (9.47) $$

The reader can check that indeed, $D_4 = \{x,y\}^*/R$.

We now let $R^1[g_1, \ldots, g_r]$ be the ring of polynomials in $g_1, \ldots, g_r$ with coefficients in $R^1$, and let $R$ be as above. We then can consider the ring of polynomials

$$ R^1[g_1, \ldots, g_r]/R \cong S \quad (9.48) $$

and one can check that

$$ S = R^1(G) \quad (9.49) $$

In this setting, the reason for the choice of $ev$ to denote the map given by (9.33) is clear — it is just the map defined by evaluating the polynomial with all indeterminates set equal to 1.

We will now derive recursive equations for computing $\sigma(k|m)$ in terms of $\sigma(k|m-1)$ for $m > k$, with the initial condition $\sigma(k|k)$. We must first introduce another group algebra and a number of maps. Let $G_1$ and $G_2$ be finite groups and consider the group algebras $R^1(G_1)$, $R^1(G_2)$, and $R^1(G_1 \times G_2)$. We define the "marginal maps" $M_1: R^1(G_1) \times G_2 \to R^1(G_1)$
and \( M_2 : R^1(G_1 \times G_2) \to R^1(G_2) \) as follows:

\[
M_1 \left[ \sum_{g \in G_1} \sum_{h \in G_2} \alpha_{gh} \cdot (g, h) \right] \triangleq \sum_{g \in G_1} \left( \sum_{h \in G_2} \alpha_{gh} \right) \cdot g 
\]

\[
M_2 \left[ \sum_{g \in G_1} \sum_{h \in G_2} \alpha_{gh} \cdot (g, h) \right] \triangleq \sum_{h \in G_2} \left( \sum_{g \in G_1} \alpha_{gh} \right) \cdot h
\]

One can check that \( M_1 \) and \( M_2 \) are group algebra homomorphisms. Also, one should note the similarity between these maps and the operation of computing marginal probability distributions from the joint distribution.

We can also define the Hadamard homomorphisms \( p_{h_0} : R^1(G_1 \times G_2) \to R^1(G_1) \)

for any \( h_0 \in G_2 \) and \( q_{g_0} : R^1(G_1 \times G_2) \to R^1(G_2) \) for any \( g_0 \in G_1 \):

\[
p_{h_0} \left[ \sum_{g \in G_1} \sum_{h \in G_2} \alpha_{gh} \cdot (g, h) \right] \triangleq \sum_{g \in G_1} \alpha_{g h_0} \cdot g
\]

\[
q_{g_0} \left[ \sum_{g \in G_1} \sum_{h \in G_2} \alpha_{gh} \cdot (g, h) \right] \triangleq \sum_{h \in G_2} \alpha_{g_0 h} \cdot h
\]

Note the similarity between these maps and the operation of computing unnormalized conditional distributions from the joint distribution. We also define a Hadamard-like product "\( \odot \)" between elements of \( R^1(G_1) \) and \( R^1(G_1 \times G_2) \):

\[
\left[ \sum_{g \in G_1} \beta_g \cdot g \right] \odot \left[ \sum_{g \in G_1} \sum_{h \in G_2} \alpha_{gh} \cdot (g, h) \right] \triangleq \sum_{g \in G_1} \sum_{h \in G_2} (\beta_g \alpha_{gh}) \cdot (g, h)
\]

Theorem 9.3 (Smoothing Equation): Consider the random FGHSS (9.15), (9.16). Then for \( m > k \), we compute \( p(k|m) \) from \( p(k|m-1) \) as follows:

\[
p(k|m) = \frac{\gamma(k|m)}{n(k|m)}
\]
where
\[ \gamma(k|m) = \rho(k|m-1)^{m_2} \mathcal{C}_m [\xi(m)y(m)] \circ \phi(m,k) \] (9.56)
\[ n(k|m) = \text{ev}[\gamma(k|m)] \] (9.57)

Here \( \phi(m,k) \in \mathbb{R}^{1}(X \times X) \), and its \((h,g)\) coefficient is
\[ \phi(m,k)_{(h,g)} = \Pr[x(m) = h | x(k) = g, y(k+1), \ldots, y(m-1)] \] (9.58)

**Proof:** We have that
\[ \Pr[x(k)g|y(0), \ldots, y(m)] = \]
\[ \frac{\Pr[x(k) = g|y(0), \ldots, y(m-1)] \Pr[y(m)|x(k) = g, y(0), \ldots, y(m-1)]}{\sum_{h \in X} \Pr[x(k) = h|y(0), \ldots, y(m-1)] \Pr[y(m)|x(k) = h, y(0), \ldots, y(m-1)]} \] (9.59)

As in the proof of Theorem 9.2, we will drop the normalizing constant and will work with the numerator of equation (9.59). Using the various independence assumptions, we have
\[ \Pr[x(k) = g|y(0), \ldots, y(m-1)] \Pr[y(m)|x(k) = g, y(0), \ldots, y(m-1)] \]
\[ = \rho(k|m-1) \{ \sum_{g \in \mathbb{R}} \Pr[y(m)|x(m) = h, y(0), \ldots, y(m-1)] \Pr[x(m) = h|x(k) = g, y(0), \ldots, y(m-1)] \} \]
\[ = \rho(k|m-1) \{ \sum_{g \in \mathbb{R}} \Pr[y(m)|x(m) = h, y(0), \ldots, y(m-1)] \Pr[x(m) = h|\phi(m,k)_{(h,g)}] \} \] (9.60)

As shown in the proof of Theorem 9.2,
\[ \Pr[y(m)|x(m) = h] = \mathcal{C}_m [\xi(m)y(m)]_h \] (9.61)

Now one needs only refer to the definitions preceding the theorem.
statement to see that (9.55) - (9.57) are correct.

The quantity $\phi(m,k)$ is computed "column by column" -- i.e. we compute $p_g[\phi(m,k)]$ for each $g \in X$. But $p_g[\phi(m,k)]$ is just the (unnormalized) distribution for $x(m)$ given that we know $x(k) = g$ and the values of $y(k+1), \ldots, y(m-1)$. This is just a filtering problem (together with a final prediction step to determine the predicted distribution for $x(m)$ from the (filtering) distribution for $x(m-1)$ given $y(k+1), \ldots, y(m-1)$) with initial condition an impulse

$$\phi(k) = g$$

Thus, $\phi(m,k)$ is, in some sense, a weighting pattern or impulse response matrix.

Some of the computational aspects of the results of Theorems 9.1 - 9.3 will be considered in the next section. We now present two examples to illustrate these results.

**Example 9.2:** In this example we will introduce a problem that is a special case of a larger class of problems that will be considered in Chapter 10. Consider a discrete time random process $\{x(k)\}$ on the circle. Suppose $x(k)$ satisfies the following discrete-time random difference equation

$$x(k+1) = \lfloor x(k) + u(k) \rfloor \mod 2\pi \quad k \geq 0$$

where $u(k)$ are independent real-valued random variables, independent of $x(0)$, with probability densities $p_u(k)$. Also suppose we take noisy measurements of the form

$$y(k) = \lfloor x(k) + v(k) \rfloor \mod 2\pi \quad k \geq 0$$
where the \( \{v(k)\} \) are independent real-valued random variables, independent of \( \{u(k)\} \) and \( x_o \), with distributions \( p_v(k) \).

The problem of computing the conditional density for \( x(k) \) given \( y(0), \ldots, y(k) \) is considered in Chapters 4 and 5 and in Appendix B. Here we will concern ourselves with a modification of this problem. We assume that the various random processes are discrete in space as well as in time. That is, we assume that the interval \([-\pi, \pi]\), or equivalently the unit circle \( S^1 \), is subdivided into \( n \) subintervals (see Figure 9.1) and that we only want to estimate into which of these subintervals \( x(k) \) falls. We also assume that the

![Figure 9.1: Subdivision of the Circle Into n Equal Size Subintervals](image)

resolution of our measurement process is such that our measurement only tells us which of these intervals \( y(k) \) is in. Note that for large values of \( n \) the subdivision of the circle is quite fine, and this quantized problem is a good approximation to the original problem. Also, one can imagine a modulation system where we are transmitting
digital information, and, in this case, $x$ takes on only a finite number of values (see Chapter 10).

Thus, we will consider $x, u, v,$ and $y$ taking on values from the set $\{0,1,\ldots,n-1\}$, where one can think of the number as indicating in which subinterval the value of the particular random process lies (see Chapter 10 for a more detailed discussion of this type of quantization; also, note that we could consider $u$ and $v$ as taking on any integer value, but since the effect of $u$ and $u+kn$ is the same modulo $n$, we can, in effect, "wrap" $u$ and $v$ around $Z_n$ as we wrapped $\mathbb{R}$ around $S^1$ again, see Chapter 10). We then have dynamical and measurement equations evolving homomorphically on $Z_n$.

$$x(k+1) = [x(k)+u(k)] \mod n$$ (9.65)

$$y(k) = [x(k)+v(k)] \mod n$$ (9.66)

We note that if we consider the $\{u(k)\}$ as digital data, one could view (9.65) and (9.66) as an encoding process. Also, we can treat the case of correlated $\{u(k)\}$ if we postulate a relationship of the form

$$u(k+1) = [g_1[u(k)]+g_2[u(k-1)]+\ldots+g_\ell[u(k-\ell+1)]+w(k)] \mod n$$ (9.67)

where the $\{w(k)\}$ are independent $Z_n$-valued random variables and the $g_i$ are $Z_n$-homomorphisms. In this case, we can augment the state by defining

$$r(k) = (u(k),\ldots,u(k-\ell+1)) \in (Z_n)^\ell$$ (9.68)

considering our state to be $(x(k),r(k)) \in (Z_n)^{\ell+1}$, and using the homomorphic state evolution equation in $(Z_n)^{\ell+1}$.
\[(x(k+1), r(k+1)) = G[x(k), r(k)] + W(k) \tag{9.69}\]

where \(G : (\mathbb{Z}_n)^{2+1} \rightarrow (\mathbb{Z}_n)^{2+1}\) is the homomorphism and \(W(k)\) the \((\mathbb{Z}_n)^{2+1}\)-valued random variable given by

\[G(a_1, \ldots, a_{2+1}) = (a_1 + a_2, g_1(a_2) + \ldots + g_2(a_{2+1}), a_3, \ldots, a_{2+1}) \tag{9.70}\]

\[W(k) = (0, w(k), 0, \ldots, 0) \tag{9.71}\]

where all additions are mod \(n\).

For simplicity, we will assume that the \(\{u(k)\}\) are independent, and thus will use (9.65) and (9.66). We can now apply the group algebra results to this example. We will concern ourselves solely with the filtering problem, but we note that similar results can be obtained for the prediction and smoothing problems. As discussed earlier, \(R^1(\mathbb{Z}_n)\) can be identified with \(R^1[s]/(s^n-1)\) via the identification

\[\rho(k) = \rho(k)_0 + \rho(k)_1s + \ldots + \rho(k)_{n-1}s^{n-1} \tag{9.72}\]

Then, referring to Theorem 2, we have the filtering equations

\[\rho(k+1|k) = \eta(k)\rho(k|k) \tag{9.73}\]

\[\gamma(k|k) = [\xi(k)\nu(k)]*\rho(k|k-1) \tag{9.74}\]

\[n(k|k) = ev[\gamma(k|k)] \tag{9.75}\]

The right-hand side of (9.73) is the product modulo \((s^n-1)\) of two polynomials of degree \((n-1)\). We can write (9.73) in matrix-vector notation.
\[
\begin{bmatrix}
\rho(k+1|k) & \eta(k) & n(k)_{n-1} & \ldots & \ldots & n(k)_{1} \\
\rho(k+1|k) & n(k)_{1} & \eta(k) & \ldots & \ldots & n(k)_{2} \\
\vdots & \vdots & \vdots & \ddots & \ldots & \vdots \\
\rho(k+1|k) & n(k)_{n-1} & n(k)_{n-2} & \ldots & \ldots & n(k)_{0}
\end{bmatrix}
\begin{bmatrix}
\rho(k|k) \\
\rho(k|k) \\
\vdots \\
\rho(k|k)
\end{bmatrix}
\]

(9.76)

Note that the rows of the matrix on the right-hand side of (9.76) can be obtained from the first row by cyclic shifts. This type of matrix is called **circulant** [H13]. Note that a naive approach to evaluating (9.76) requires \(n^2\) multiplications (in effect, every \(\rho(k|k)_i\) multiplies every \(n(k)_j\)). Reductions in the number of multiplications that arise in problems of this type are considered in Section 9.3.

In vector notation, the Hadamard product has a very simple form

\[
\begin{bmatrix}
\alpha_o \\
\alpha_1 \\
\vdots \\
\alpha_{n-1}
\end{bmatrix}
\begin{bmatrix}
\beta_o \\
\beta_1 \\
\vdots \\
\beta_{n-1}
\end{bmatrix} =
\begin{bmatrix}
\alpha_o \beta_o \\
\alpha_1 \beta_1 \\
\vdots \\
\alpha_{n-1} \beta_{n-1}
\end{bmatrix}
\]

(9.77)

Recall that \(v(k) \in \{0,1,\ldots,n-1\}\) and that, considering \(\mathbb{Z}_n \subset \mathbb{R}^1(\mathbb{Z}_n)\), \(y(k)\) corresponds to \(s^v(k)\) and thus

\[
\xi(k)y(k) = (\xi(k)_0 + \xi(k)_1 s + \ldots + \xi(k)_{n-1} s^{n-1}) s^v(k)
\]

\[
= \xi(k) - y(k) + \xi(k)_1 y(k) s + \ldots + \xi(k)_{n-1-v(k)} s^{n-1}
\]

(9.78)

where

\[
\xi(k)_i \triangleq \xi(k)_i \mod n
\]

(9.79)

Thus, referring to (9.74) and (9.77), we see that the operation
\( \xi(k)y(k) \) shifts \( \xi(k) \), considered as an n-vector, cyclically \( y(k) \) units -- i.e. taking the vector

\[
\begin{bmatrix}
\xi(k)_0 \\
\xi(k)_1 \\
\vdots \\
\xi(k)_{n-1}
\end{bmatrix}
\]

\( \xi(k) \) (9.80)

We shift each element precisely \( y(k) \) units down and wrap those that "fall cut the bottom" around to the top. The rest of the evaluation of (9.74) is simply the Hadamard product of (9.77). We then see that implementation of (9.74) requires a cyclic shifting procedure and \( n \) multiplications. Finally, (9.75) requires \( (n-1) \) additions (note that, as mentioned before, in many instances -- e.g. if we are only interested in determining the most probable estimate -- normalization of distributions is of no importance, except occasionally to keep the numbers from becoming too small).

**Example 9.3:** We now give a nonabelian example. Let \( U = Z_2 \times Z_2 \), \( X = Y = D_4 = \{ e, p, p^2, p^3, t, pt, p^2t, p^3t \mid p^4 = e, t^2 = e, pt = tn \} \) let \( a : X \rightarrow X \) and \( b : U \rightarrow X \) be the homomorphism uniquely determined by

\[
a(t) = pt \quad a(p) = p^3 \quad b(1,0) = p^2 \quad b(0,1) = t
\]

(9.81)

and let \( c : X \rightarrow Y \) be the identity map (state output). Consider the random FGHSS (9.15), (9.16) so defined, and again let us look at the filtering problem. Equation (9.28) can be written in matrix form (from now on we will regard all distributions both as elements of group algebras and as vectors):
\[
\begin{pmatrix}
\rho(k+1) | k_e \\
\rho(k+1) | k_p \\
\vdots \\
\rho(k+1) | p^3 \\
\rho(k+1) | k_t \\
\vdots \\
\rho(k+1) | p^3 t
\end{pmatrix}
= B[\eta(k)]A
\]

(9.82)

where \( B[\eta(k)] \) is the matrix representing \( \hat{b}[\eta(k)] \) and is given by

\[
B[\eta(k)] =
\begin{pmatrix}
\eta(k)(0,0) & 0 & \eta(k)(1,0) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) & 0 \\
0 & \eta(k)(0,0) & 0 & \eta(k)(0,0) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) \\
\eta(k)(1,0) & 0 & \eta(k)(0,0) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) & 0 \\
0 & \eta(k)(1,0) & 0 & \eta(k)(0,0) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) \\
\eta(k)(0,1) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,0) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,1) & 0 \\
0 & \eta(k)(1,1) & 0 & \eta(k)(0,1) & 0 & \eta(k)(1,0) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) \\
\eta(k)(1,1) & 0 & \eta(k)(0,1) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,0) & 0 & \eta(k)(0,1) & 0 \\
0 & \eta(k)(0,1) & 0 & \eta(k)(1,1) & 0 & \eta(k)(0,0) & 0 & \eta(k)(0,1) & 0 & \eta(k)(0,1) \\
\end{pmatrix}
\]

(9.83)

Also \( A \) is the matrix representing \( \hat{a} \)

\[
A =
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}
\]

(9.84)
Thus (9.82) becomes

\[ p(k+1|k) = D_k p(k|k) \]  \hspace{1cm} (9.85)

where

\[ D_k = \begin{bmatrix}
  n(k)(0,0) & 0 & n(k)(1,0) & 0 & 0 & n(k)(0,1) & 0 & n(k)(1,1) \\
  0 & n(k)(1,0) & 0 & n(k)(0,0) & n(k)(1,1) & 0 & n(k)(0,1) & 0 \\
  n(k)(1,0) & 0 & n(k)(0,0) & 0 & 0 & n(k)(1,1) & 0 & n(k)(0,1) \\
  0 & n(k)(0,0) & 0 & n(k)(1,0) & n(k)(0,1) & 0 & n(k)(1,1) & 0 \\
  n(k)(0,1) & 0 & n(k)(1,1) & 0 & 0 & n(k)(0,0) & 0 & n(k)(1,0) \\
  0 & n(k)(1,1) & 0 & n(k)(0,1) & n(k)(0,0) & 0 & n(k)(1,0) & 0 \\
  n(k)(1,1) & 0 & n(k)(0,1) & 0 & 0 & n(k)(1,0) & 0 & n(k)(0,0) \\
  0 & n(k)(1,1) & 0 & n(k)(0,1) & n(k)(1,0) & 0 & n(k)(0,0) & 0
\]  \hspace{1cm} (9.86)

Some comments on these equations are in order. First note that \( A \)
is a permutation matrix. One can check that this will be the case if andonly if \( A \) is an automorphism. Also, note the form of (9.83) and (9.86).
As before, the rows are all permutations of the first row, although notcyclic permutations (as was true in the \( \mathbb{Z}_n \) case), and one can check thatmatrices of these types do not necessarily commute. However, as weshall see in the next section, we can still use the structure of theseequations to design efficient computational schemes. As a final note on the prediction update type of equation, suppose our state dynamic equation is

\[ x(k+1) = u(k)x(k) \]  \hspace{1cm} (9.87)
where all variables are defined on $D_4$. In this case, the group algebra equation

$$\rho(k+1\mid k) = \eta(k)\rho(k\mid k) \quad (9.88)$$

which is the direct analog of the $Z_n$ equation (9.73), can be written in the matrix form

$$\rho(k+1\mid k) = E_k\rho(k\mid k) \quad (9.89)$$

where

$$E_k =
\begin{array}{cccccccc}
\eta(k)e & \eta(k)p^3 & \eta(k)p^2 & \eta(k)p & \eta(k)t & \eta(k)nt & \eta(k)p^2t & \eta(k)p^3t \\
\eta(k)p & \eta(k)e & \eta(k)p^3 & \eta(k)p^2 & \eta(k)pt & \eta(k)p^2t & \eta(k)p^3t & \eta(k)t \\
\eta(k)p^2 & \eta(k)p & \eta(k)e & \eta(k)p^3 & \eta(k)p^2t & \eta(k)p^3t & \eta(k)t & \eta(k)pt \\
\eta(k)p^3 & \eta(k)p^2 & \eta(k)e & \eta(k)p & \eta(k)p^2t & \eta(k)p^3t & \eta(k)t & \eta(k)pt \\
\end{array}
\quad (9.90)

Note the symmetry in this matrix. The upper left and lower right 4x4 submatrices are the same and, in fact, are circulant (basically because $(e, p, p^2, p^3) = Z_4$). Also the upper right and lower left 4x4 submatrices are the same, but the rows are permuted by shifting elements in the opposite direction (because of the nature of the noncommutativity of $D_4$ -- i.e. $pt = tp^3$; incidentally, this extends
to the general dihedral group case). The reader can also check that
the group of permutations that transforms the first row into each of
the others is isomorphic to $D_4$. All this beautiful structure should
be of some help. In the next section we will at least get an
indication of how we can use it.

We now consider the measurement update equation with $Y = X = D_4$
and $c = \text{identity}$. Since the Hadamard product ignores the underlying
group structure, this part of the equation is no different than the
$Z_n$ case -- i.e. in vector form

$$
\begin{bmatrix}
\alpha_e \\
\alpha_p \\
\vdots \\
\alpha_{p^3t}
\end{bmatrix} 
\times 
\begin{bmatrix}
\delta_e \\
\delta_p \\
\vdots \\
\delta_{p^3t}
\end{bmatrix} 
= 
\begin{bmatrix}
\alpha_e \delta_e \\
\alpha_p \delta_p \\
\vdots \\
\alpha_{p^3t} \delta_{p^3t}
\end{bmatrix}
$$

(9.91)

As in the $Z_n$ case, our problem is to see how $y(k)$ permutes the
vector $\xi(k)$ when we perform the group algebra multiplication $\xi(k)y(k)$.
As one can compute

$$
\begin{bmatrix}
\xi(k) \\
\gamma(k)^{-1} \\
\xi(k) \\
\gamma y(k)^{-1} \\
\vdots \\
\xi(k) \\
\gamma^3 y(k)^{-1}
\end{bmatrix}
$$

(9.92)

and again, the measurement update part of the filtering computations
consists of a shifting process plus 8 multiplications (plus the normalization).
We close this section by remarking on the extension of these results to infinite group cases. For finitely generated groups (such as \( \mathbb{Z} \)), elements of the group algebra are power series (e.g. for \( \mathbb{Z} \) the typical group algebra element is \( \rho = \sum_{n=-\infty}^{+\infty} c_n z^n \), and if \( \rho \) is the distribution for a \( \mathbb{Z} \)-valued random variable \( x \), we have \( c_n = \Pr(x=n) \)). For Lie groups the group algebra is a function space. In the former case, convolution of power series is defined in the natural way. For Lie groups, the elements of the group algebra correspond to probability densities as opposed to probability distributions, and convolution is defined as the generalization of the convolution of two functions on the real line. For instance if \( G \) is a compact Lie group with normalized Haar measure \( \mu \) and if \( f_1 \) and \( f_2 \) are integrable functions on \( G \), then

\[
(f_1 * f_2)(g) = \int_G f_1(h)f_2(h^{-1}g)\,d\mu(h) \quad (9.93)
\]

The reader should compare (9.93) and (9.11) and is referred to a treatment of this subject in [L5].

9.3 Some Computational Aspects of Group Algebra Multiplications

As mentioned in the preceding section, it appears that the structure of group algebra multiplications can be used to devise efficient computational schemes. This is of importance to us, since the prediction update equation (9.28) involves a group algebra multiplication. As seen in Examples 9.2 and 9.3, the measurement update involves a Hadamard product, which can be done in \( n \) multiplications where \( n = \text{card } X \). In this section we will discuss computational procedures for the group algebra
multiplication. This multiplication is of more concern than the Hadamard product, since one can check that a naive approach to group algebra multiplication requires $n^2$ multiplications, as well as $n(n-1)$ additions.

In this section we will discuss this problem at some length. We remark that the results of Nicholson [N3] and others point out how the structure of $R^1(Z_n)$ can be used to determine efficient methods for computing $R^1(Z_n)$ multiplication (as well as finite Fourier transforms). The problem of computing general group algebra multiplications is more difficult and it is not the purpose of this section to determine a general procedure for such multiplications, but rather to point out what could be gained from the solution of this problem (for some relevant work the reader is referred to [N3], the work of Depeyrot [D5], [D6], and to the Wedderburn Structure theorem for group algebras [C8]).

To this end, we will discuss the $R^1(Z_n)$ problem as motivation and will then state the general group algebra problem, which we do not solve. Instead, we consider several examples (including $R^1(Z_n)$) using a general multiplication formulation of Brickett and Dobkin [B18], [D4]. This problem formulation leads to a brute force search procedure that we apply to our examples. It should be noted that this procedure does not directly make use of the underlying group algebra structure, and hence cannot uncover general efficient methods; however the results we have obtained using this approach indicate that a detailed analysis of group algebra structure will lead to efficient multiplication algorithms analogous to the FFT and $R^1(Z_n)$ algorithms. At the close of this section, we will enlarge the class of problems under consideration. Specifically,
we will examine a fairly general class of finite state Markov processes and will briefly discuss the associated computation of conditional distributions.

Before starting our discussion, it is appropriate to comment on what we mean by an "efficient computational scheme." In this section we are concerned with the operations of addition, subtraction, and multiplication. It is known that multiplications take longer (a reasonable measure is 5 times as long (see [Kl5])) than subtractions or additions (which take the same amount of time). Consequently, we wish to minimize some weighted sum of the number of operations required. We will take as our figure of merit for any scheme C

\[ J(C) = 5(#M) + #A + #S \] (9.94)

where, \#M, \#A, and \#S are the numbers of multiplications, additions, and subtractions, respectively. Note that if we have a number \( x \) and must compute \( 2x \), we will regard this not as a multiplication, but rather as an addition of \( x \) to itself or as a shift, which is faster than any of the arithmetic operations and can be neglected (as can multiplications by \( \frac{1}{2}, 4, \frac{1}{4}, \) etc.).

We now turn to the problem of multiplying two polynomials of degree \( < n \) modulo \( s^n - 1 \) (i.e. \( R^1(Z_n) \) multiplication). The reader is referred to [B18], [D4], [N3], [Kl5], [S12], [F6], [T2], and [C9] for discussions of this type of problem. In particular, the work of Nicholson [N3], Knuth [Kl5], and Schonhage and Strassen [S12] show how the theory of FFT's is useful in this problem. Although this approach would seem to be of great value in the general group algebra
case, we will not discuss it in detail here. Instead we will formulate the problem à la Brockett and Dobkin.

Thus suppose we have two elements of $\mathbb{R}^1(\mathbb{Z}_n)$

$$\eta(s) = \eta_0 \eta_1 s + ... + \eta_{n-1} s^{n-1} \quad (9.95)$$

$$\rho(s) = \rho_0 \rho_1 s + ... + \rho_{n-1} s^{n-1} \quad (9.96)$$

and wish to compute the product (mod($s^n - 1$))

$$\mu(s) = \eta(s)\rho(s) = \mu_0 + \mu_1 s + ... + \mu_{n-1} s^{n-1} \quad (9.97)$$

As mentioned in Example 9.2, we can represent elements of $\mathbb{R}^1(\mathbb{Z}_n)$ by $n \times n$ circulant matrices -- e.g.

$$\mu = \begin{bmatrix}
\mu_0 & \mu_1 & \cdots & \mu_{n-1} \\
\mu_1 & \mu_0 & & \\
\vdots & \vdots & \ddots & \\
\mu_{n-1} & \mu_{n-2} & \cdots & \mu_0
\end{bmatrix} \quad (9.98)$$

and in this form we still have $\mu = \eta \rho = \eta \rho$ -- i.e. circulant matrices commute, which is related to the fact that $\mathbb{Z}_n$ is abelian). We note from Example 9.1 that we can write

$$\mu_k = \sum_{i=0}^{k} \eta_i \rho_{k-i} + \sum_{i=k+1}^{n-1} \eta_i \rho_{n+k-i} \quad (9.99)$$

and this can be translated into the Brockett-Dobkin form

$$\mu_k = \eta^I C_k \rho \quad k = 0, ..., n-1 \quad (9.100)$$

where we now consider $\eta$ and $\rho$ as $n$-vectors (see Example 9.2) and $C_k$ is an $n \times n$ matrix. The matrices $C_k$ can be easily computed. For instance
\[
G_o = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 1 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0
\end{bmatrix}
\quad
G_1 = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 1 & \cdots & 0
\end{bmatrix}
\quad (9.101)
\]

The structure of these matrices is closely related to the properties of \(Z_n\) (note that if we write \(n' = (\eta_0, \ldots, \eta_{n-1})\) but reversed the direction in \(-\text{i.e.} o' = (o_0, o_{n-1}, \ldots, o_1)\) = then the \(G_k\)'s that would arise form a group isomorphic to \(Z_n\).

Our problem is to perform the sequence of operations given by (9.97), (9.99), or (9.100). As shown in [D4], [K15], [N3], and [S12], one can compute the group algebra product in \(R_1(Z_n)\) using on the order of \(n \log n\) multiplications, which is much better than \(n^2\) for large values of \(n\). A brute force method for finding such schemes consists of considering the group algebra product as a simultaneous computation of \(n\) bilinear forms -- those of equation (9.100) (see [B18] and [D4]). The basic idea is as follows: given the \(G_k, k = 0, 1, \ldots, n-1\), suppose we can find \(M\) dyads \(D_1, \ldots, D_M\) -- i.e. \(M\) matrices of the form

\[
D_i = \alpha_i \beta_i^t
\quad (9.102)
\]

where the \(\alpha_i\) and \(\beta_i\) are \(n\)-vectors (i.e. the \(D_i\) are rank one matrices) -- such that

\[
G_k = \sum_{i=1}^{M} c_{ik} D_i 
\quad k = 0, \ldots, n-1
\quad (9.103)
\]

Then from (9.100),

\[
\omega_k = \sum_{i=1}^{M} c_{ik} (n' \alpha_i)(\beta_i^t o)
\quad (9.104)
\]

where \(n'\alpha_i\) and \(\beta_i^t o\) are just real numbers. Suppose also that the \(c_{ik}\)
and the elements of $\alpha_1$ and $\beta_1$ consist mostly of 0's, ±1's, and ±2's
(as they do in the following examples)-- i.e. the operations in (9.104)
involve these quantities are mostly additions and subtractions.

Then, we can compute the group algebra product using $M$ multiplications
$(\eta^n \alpha_1 \cdot (\beta_1 \rho))$, $i=1, \ldots, N$. The following example illustrates this idea.

**Example 9.4:** We consider $n = 6$. Then the naive group algebra multiplication scheme requires 36 multiplications and 30 addition, and from (9.94)

$$J_{\text{naive}} = 210$$  \hspace{1cm} (9.105)

We can compute the $G_k$ as follows: we have

$$G_0 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}$$  \hspace{1cm} (9.106)

Then $G_1$ is computed from $G_0$ by putting the bottom row of $G_0$ on top and pushing the other rows down one. We then compute $G_k$ from $G_{k-1}$ for $k=2,3,4,5$ in this same manner. Note that the $G_k$ are linearly independent and nonsingular.

We define the following matrices

$$M_1 = G_0 + G_1 + G_2 + G_3 + G_4 + G_5$$
$$M_2 = G_0 + G_2 + G_4$$
$$M_3 = G_0 + G_3$$
$$M_4 = G_1 + G_4$$
$$M_5 = G_0 - G_3$$
$$M_6 = G_2 - G_5$$  \hspace{1cm} (9.107)

Then we have
\[ G_0 = \frac{M_3 + M_5}{2} \quad G_1 = \frac{M_1 + M_4 + M_5 + M_6 - 2M_2}{2} \quad G_2 = \frac{M_1 + M_6 - M_3 - M_4}{2} \]
\[ G_3 = \frac{M_3 - M_5}{2} \quad G_4 = \frac{2M_2 + M_4 - M_1 - M_5 - M_6}{2} \quad G_5 = \frac{M_1 - M_3 - M_4 - M_6}{2} \]

and

\[ \text{rank } M_1 = 1 \quad \text{rank } M_2 = 2 \]
\[ \text{rank } M_3 = \text{rank } M_4 = \text{rank } M_5 = \text{rank } M_6 = 3 \]

Thus we can find 15 dyads \( \alpha_1 \beta_1^i \) and 90 integers \( c_{1k} \) such that

\[ U_k = \frac{1}{2} \sum_{i=1}^{15} c_{1k}(\eta' \alpha_1)(\beta_1^i \rho) \quad k=0, \ldots, 5 \]

for any \( \rho, \eta \in R^1(Z_6) \). Working out the details, we find that (9.110) can be computed using 57 additions or subtractions. Also, we have the 15 multiplications \( (\eta' \alpha_1)(\beta_1^i \rho) \) plus the 6 multiplications by 1/2. Thus the figure of merit for this scheme is

\[ J = 162 \]

which is substantially less than that for the naive method. Note that often the scaling factor is unimportant (e.g., if we use unnormalized distributions as mentioned in the preceding section), and thus the 6 multiplications by 1/2 can be neglected, as they can if we regard them as shifts (for our particular problem, the distribution \( \eta \) of the process noise is known a priori and is stored on the computer; thus, if we store 1/2 \( \eta \) instead of \( \eta \), we can avoid these multiplications). In this case

\[ J = 132 \]

We note that the scheme discussed here is not the optimal one but
illustrates the kind of computational savings that can be realized. The reader is also referred to the high-speed convolution and FFT results presented in [G12].

Before considering the general group algebra problem, we give another example that illustrates the usefulness of this technique.

Example 9.5: We consider the discrete-time Fourier series measurement update equations of Section 5.2. In particular, suppose we have the a priori coefficients \( \{a_n(0), b_n(0)\}_{n=1}^{\infty} \) and the measurement coefficients \( d_0, \{c_n, d_n\}_{n=1}^{\infty} \). We assume that we have stored only \( \{a_n(0), b_n(0)\}_{n=1}^{2}, d_0, \{c_n, d_n\}_{n=1}^{2} \) and wish to compute the unnormalized coefficients \( \{2\alpha_n(1), 2\beta_n(1)\}_{n=1}^{2} \). We will take as our approximations to these coefficients the values obtained by a straightforward truncation of equations (5.8) and (5.9). As before, we can compute these coefficients in a straightforward, naive manner. If we do this, we find

\[
J_{\text{naive}} = 116 \quad (9.113)
\]

We can also apply the Brockett-Dobkin technique; we write

\[
\eta = \begin{bmatrix} 2d_0 \\ c_1 \\ d_1 \\ c_2 \\ d_2 \end{bmatrix}, \quad \rho = \begin{bmatrix} \frac{1}{\pi} \\ a_1(0) \\ b_1(0) \\ a_2(0) \\ b_2(0) \end{bmatrix}, \quad \mu = \begin{bmatrix} 2\alpha_1(1) \\ 2\beta_1(1) \\ 2\alpha_2(1) \\ 2\beta_2(1) \end{bmatrix} \quad (9.114)
\]

\[
\mu_k = \eta^t C_k, \quad k = 1, 2, 3, 4 \quad (9.115)
\]

where the \( C_k \) can be easily computed. We can also check that the method of finding a set of dyads that generate the \( C_k \) leads to a method with
cost

\[ J = 94 \]  \hspace{1cm} (9.116)

The amount of computation time saved increases with the number of Fourier coefficients used.

We will now consider the general group algebra multiplication problem. Let \( G = \{g_1, g_2, \ldots, g_n\} \) be a group, and let \( \eta, \rho \in R^1(G) \). We will write

\[ \eta = \sum_{i=1}^{n} \eta_i^* g_i \quad \rho = \sum_{i=1}^{n} \rho_i^* g_i \]  \hspace{1cm} (9.117)

and will also represent them as \( n \)-vectors

\[ \eta' = [\eta_1, \ldots, \eta_n] \quad \rho' = [\rho_1, \ldots, \rho_n] \]  \hspace{1cm} (9.118)

We can also represent \( \eta \) and \( \rho \) as \( nxn \) matrices, à la the circulant matrix representation of elements of \( R^1(Z_n) \). In the general case, unlike the circulant case, these matrices need not commute. However, like the circulant case, a matrix of this type is determined by the first row and a group of permutations of this row that yields the other rows. As in the \( Z_n \) case, this group of permutations is isomorphic to \( G \) (see Ex. 9.2, 9.3, 9.5, and 9.6). This matrix representation is of interest in that it indicates that any computational results for group algebra multiplication yield results for matrix multiplication in special cases. We remark that Depeyrot [D5], [D6] discusses possible computational gains for \( R^1(G) \) multiplication with \( G \) a nonabelian group, and his results plus those of Nicholson [N3] indicate that the FFT leads directly to efficient computational schemes for \( R^1(G) \) multiplication where \( G \) is
any finite abelian group (not necessarily cyclic).

We will now briefly discuss some of the computational aspects of group algebra multiplication in the setting of [B18] and [D4]. We wish to compute

\[ u = \eta \circ \rho \]  \hspace{1cm} (9.119)

where \( \eta \) and \( \rho \) are given in (9.117). The generalized convolution equation, (9.11) implies that there exist \( n \) matrices, \( G_1, \ldots, G_n \) such that

\[ u_k = \eta^t G_k \rho \hspace{1cm} k=1, \ldots, n \]  \hspace{1cm} (9.120)

The \( G_k \) are closely related to the group \( G \). In particular, let \( g_1 \) be the identity in \( G \), and consider the group \( H \) of permutations of the rows of \( G_1 \) that produce the matrices \( G_k \). We can write the elements of \( H \) as \( n \times n \) permutation matrices \( \{ P_k \} \) such that \( P_k G_1 = G_k \). Thus \( P_k = G_k G_1^{-1} \). The group \( H \) can be shown to be isomorphic to \( G \) and is called the **regular representation** of \( G \) (see [R1], [F1]).

Our problem is to compute (9.119) efficiently. As mentioned earlier, we will not consider this general problem. The fact that the fast Fourier transform and the structure theory for \( R^1(Z_n) \) lead directly to a general technique for \( R^1(Z_n) \) multiplication (see [D4] and [S12]) leads us to believe that the general structure theory for \( R^1(G) \) will lead to highly efficient computational schemes. We present several examples using the method of Brockett and Dobkin. These examples indicate that large savings in computation time -- savings on the order of those obtained in the \( R^1(Z_n) \) case -- can be realized.

**Example 9.6:** We consider \( R^1(D_3) \), where, for example, we write
\[ n = n_e + n_x + n_{x^2} + n_y + n_{xy} + n_{x^2y} \]  \hspace{1cm} (9.121)

or

\[ n' = [n_e, \ldots, n_{x^2y}] \]  \hspace{1cm} (9.122)

Also, we can represent \( n \) by a 6x6 matrix

\[
\begin{bmatrix}
n_e & n_x & n_{x^2} & n_y & n_{xy} & n_{x^2y} \\
n_x & n_e & n_x & n_{xy} & n_{x^2} & n_y \\
n_{x^2} & n_x & n_e & n_{x^2y} & n_y & n_{xy} \\
n_y & n_{xy} & n_{x^2} & n_e & n_x & n_{x^2} \\
n_{xy} & n_{x^2y} & n_y & n_x & n_e & n_{x^2} \\
n_{x^2y} & n_y & n_{xy} & n_x & n_{x^2} & n_e
\end{bmatrix}
\]  \hspace{1cm} (9.123)

One can check that: (1) the group of permutations of the first row that yields the other rows is isomorphic to \( D_3 \); and (2) the product of two matrices of the form (9.123) is itself of this form.

We can compute the matrices \( G_e, \ldots, G_{x^2y} \) (\( n_h = n'(G_h)' \) for \( h \in D_3 \)).

For example

\[
G_y = \begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0
\end{bmatrix}
\]  \hspace{1cm} (9.124)

We then compute

\[
M_1 = G_e + G_x + G_{x^2} + G_y + G_{xy} + G_{x^2y} \quad M_2 = G_x + G_{xy} + G_{x^2} \\
M_3 = G_e - G_y \quad M_4 = G_e + G_{xy} \\
M_5 = G_e + G_{x^2} \quad M_6 = G_x + G_y
\]  \hspace{1cm} (9.125)
and we have

\[ \begin{align*}
G_e &= -M_2 - M_3 + M_4 + M_5 \\
G_x &= M_2 + 2M_3 - M_3 - M_5 + M_6 \\
G_{x^2} &= M_1 - M_6 - M_3 - M_5 \\
G_{x^2y} &= M_2 + M_3 - M_5 \\
G_x &= M_2 - 2M_3 + M_4 + M_5 \\
G_{x^2} &= N_2 + M_4 - M_5
\end{align*} \] (9.126)

\[ \begin{align*}
\text{rank } M_1 &= 1 \\
\text{rank } M_2 &= 2 \\
\text{rank } M_3 &= \text{rank } M_4 = \text{rank } M_5 = 3
\end{align*} \] (9.127)

Thus, just as the \( Z_6 \) case, we can find 15 dyads \( \alpha_i \beta_i \) and 90 integers \( \{c_{ig} | i = 1, \ldots, 15; g \in D_3\} \) such that

\[ u_g = \frac{15}{i=1} c_{ig}(n'^i \alpha_i)(\beta_i n) \quad g \in D_3 \] (9.128)

Note that in this case, unlike the \( Z_6 \) case, there is no factor of \( \frac{1}{2} \) in (9.128). Also, from (9.125) we see that \( G_e \neq G_v, G_e \neq G_{xy} \), and \( G_e \neq G_{x^2y} \) are all rank 3 matrices, while in the \( Z_6 \) case, only \( G_o \leq G_3 \) is of rank 3 (i.e. \( G_o \leq G_k \) is not of rank 3 for any other \( k \)). Thus, in the present case, we had more freedom in choosing which combinations to include in (9.125), and therefore we were able to avoid the \( 1/2 \) factor that arose in the \( Z_n \) case essentially from having \( M_3 = G_o + G_3 \) and \( M_5 = G_o - G_3 \). The reason for this increased freedom is essentially the following: \( Z_6 \) has only one subgroup of cardinality 2, namely \( \{0, 3\} \), while \( D_3 \) has \( \text{three} \) such subgroups -- \( \{e, y\}, \{e, xy\}, \text{and } \{e, x^2y\} \).

Returning to the computation problem, we note that the naive method yields the same cost as the naive method for \( Z_6 \):

\[ J_{\text{naive}} = 210 \] (9.129)
The method suggested by (9.125) - (9.128) yields

\[ J = 133 \quad (9.130) \]

**Example 9.7:** Consider the quaternion group \( Q \) introduced in Example 8.10. The group algebra \( R^1(Q) \) is what is commonly called the ring of quaternions (see Chapter 7, [F1], and [L1]). We now consider the problem of \( R^1(Q) \) multiplication. This example is also discussed in [D4]. Since \( \text{card}(Q) = 8 \), the naive multiplication scheme yields

\[ J_{\text{naive}} = 376 \quad (9.131) \]

One can show that we can realize \( R^1(Q) \) multiplication with 23 dyads. In this case, much like the \( R^1(\mathbb{Z}_n) \) case, we have a factor of \( 1/4 \) in all the terms -- i.e.,

\[ \mu_g = \frac{1}{4} \sum_{i=1}^{23} c_{ig}(\eta' \alpha_i \Phi_i \rho) \quad G \in Q \quad (9.132) \]

If we include the multiplications by \( 1/4 \) we can find an algorithm with cost

\[ J = 266 \quad (9.133) \]

and, if we ignore the \( 1/4 \) multiplications (as in Example 9.4), we have

\[ J = 226 \quad (9.134) \]

**Example 9.8:** We consider the system defined in Example 9.3. That is, \( U = \mathbb{Z}_2 \times \mathbb{Z}_2 \) and \( X = D_4 \) with \( a \) and \( b \) defined by (9.31). The group algebra multiplication to be performed is \( \mu = \hat{b}_k(\eta) \hat{a}_k(\rho) \) where \( \eta \in R^1(U) \) and \( \rho \in R^1(X) \). Then, regarding \( \eta \) as a 4-vector and \( \rho \) as an 8-vector, we have

\[ \mu_h = \eta' G_h \rho \quad h \in D_4 \quad (9.135) \]
where the $G_h$ can be computed from (9.85) and (9.86). The naive evaluation of $u$ yields a cost of

$$J_{\text{naive}} = 184$$  \hspace{1cm} (9.136)

(32 multiplications and 24 additions). Using the Brockett-Dobkin method, we can find 14 dyads that generate the $G_n$ of (9.135). As in the $Z_6$ case a factor of $1/4$ enters each $u_h$ calculation. If we count these as multiplications we can find a method with cost (22 multiplications and 54 additions):

$$J = 164$$  \hspace{1cm} (9.137)

and there is little saving; however, if we neglect the $\frac{1}{4}$'s, we have

$$J = 124$$  \hspace{1cm} (9.138)

Thus, we have a strong indication that we can substantially reduce the time needed to compute group algebra multiplications. Therefore, the group algebra form for the conditional distribution equations considered in Section 9.2 will lead directly to efficient implementation if a general $R^1(G)$ multiplication procedure, à la the FFT, can be determined.

We now consider the extension of the concepts of this and the preceding section to a larger class of finite state Markov processes (FSMP's) than those that can be described by random FCHSS's. The motivation for this study is again basically a computational one. For instance consider the work of Astrom [A10] on the optimal control of FSMP's with noisy observations. In this work, Astrom derives an optimal control law that is explicitly a function only of time and the conditional
distribution of the state. Thus, we are again presented with the problem of computing conditional distributions.

Specifically, we consider the following situation. We are given a FSMP consisting of $n$ states $X = \{s_1, \ldots, s_n\}$, an initial distribution $\rho(0)$ on $X$, and the transition probabilities

$$p_{ij}(k) = \Pr\{x(k+1) = s_i \mid x(k) = s_j\} \quad k \geq 0 \quad (9.139)$$

Often, such a FSMP comes from a dynamical system of the form

$$x(k+1) = \lambda[x(k), u(k), k] \quad k \geq 0 \quad (9.140)$$

where $x \in X$, $u \in U$, a finite set, and $\lambda : X \times U \times Z \to X$ is the "next-state function" (see Appendix E). The transition probabilities (9.139) arise by assuming $u(k)$ is a random variable with distribution $\eta(k)$ on $U$. We note that if all of the $p_{ij}$'s of (9.139) are rational numbers, one can construct a system of the form of (9.140) with $r$ inputs, where $r$ is the least common multiple of the denominators of the $p_{ij}$. The construction is tedious and hence, is omitted. We do note that, since we can approximate the $p_{ij}$'s by rational numbers as closely as we like, we will assume that all FSMP's considered here can be written as in (9.140). We also consider an output function of the form

$$v(k) = \delta(x(k), k) \quad (9.141)$$

where $Y$ is a finite output set and $\delta : X \times Z \to Y$.

As discussed in Appendix E, there is a class of finite state machines that have state sets that are semigroups and input-state dynamics of the form

$$x(k+1) = b_k[u(k)] \cdot x(k) \quad (9.142)$$
where "·" is semigroup multiplication. Thus, we can consider a large class of FSMP's that include random FGHSS's as a special case. Let $X$ be a finite semigroup, $U$, $Y$, and $V$ finite sets, and consider the random dynamical system

$$x(k+1) = b_k[u(k)] \cdot a_k[x(k)] \tag{9.143}$$

$$y(k) = \delta(x(k), v(k), k) \tag{9.144}$$

where $a_k : X \times X$, $b_k : U \rightarrow X$, and $\delta : X \times V \times Z \rightarrow Y$. Also $u(k)$ is a random process on $U$ with distribution $\eta(k)$, and $v(k)$ is a random process on $V$ with distribution $\xi(k)$. We wish to consider some computational aspects of the evaluation $\rho(k|\theta)$, the distribution for $x(k)$ given $y(l), l < n$. As before, the measurement update part of the computation is relatively straightforward (see Section 9.2 and also [A10]), and we will concentrate on the prediction update equation.

Note that we can consider a semigroup algebra $F(S)$ where $F$ is a field and $S$ a semigroup. The general element of $F(S)$ is

$$\rho = \sum_{s \in S} \alpha_s \cdot s \quad \alpha_s \in F \tag{9.145}$$

Addition and scalar multiplication are the same as for group algebras. Multiplication in $F(S)$ is also the same as in group algebras, although the convolution equation (9.11) doesn't hold, since some elements of $S$ need not have inverses. As in the FGHSS case, we can consider the distribution $\rho(k)$ of $x(k)$ in (9.143) as an element of $R^1(X)$ and we define $\hat{b}_k[\eta(k)], \hat{a}_k[\rho(k)] \in R^1(X)$ by

$$\hat{b}_k[\eta(k)] = \sum_{u \in U} \eta(k) b_k(u) \tag{9.146}$$
\[ \hat{a}_k[\rho(k)] = \sum_{x \in X} \rho(k) \cdot a_k(x) \] (9.147)

Then \( \rho(k+1) \) is the semigroup algebra product

\[ \rho(k+1) = \hat{b}_k[\eta(k)] \cdot \hat{a}_k[\rho(k)] \] (9.148)

(similarly \( \rho(k+1|k) \) is the semigroup algebra product of \( \hat{b}_k[\eta(k)] \) and \( \hat{a}_k[\rho(k|k)] \)).

We now present two examples that indicate that we can find methods for semigroup algebra multiplication that use less computation time than that required by the naive method.

**Example 9.9:** Consider the FSMP evolving on the permutation-reset machine (see [K11] and [H11]) described in Table 9.1.

<table>
<thead>
<tr>
<th>( u_1 )</th>
<th>( s_1 )</th>
<th>( s_2 )</th>
<th>( s_3 )</th>
<th>( s_4 )</th>
<th>( s_5 )</th>
<th>( s_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_2 )</td>
<td>( s_2 )</td>
<td>( s_3 )</td>
<td>( s_1 )</td>
<td>( s_5 )</td>
<td>( s_6 )</td>
<td>( s_4 )</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>( s_3 )</td>
<td>( s_1 )</td>
<td>( s_2 )</td>
<td>( s_6 )</td>
<td>( s_4 )</td>
<td>( s_5 )</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>( s_1 )</td>
<td>( s_2 )</td>
<td>( s_3 )</td>
<td>( s_4 )</td>
<td>( s_5 )</td>
<td>( s_6 )</td>
</tr>
<tr>
<td>( u_5 )</td>
<td>( s_4 )</td>
<td>( s_4 )</td>
<td>( s_4 )</td>
<td>( s_4 )</td>
<td>( s_4 )</td>
<td>( s_4 )</td>
</tr>
<tr>
<td>( u_6 )</td>
<td>( s_5 )</td>
<td>( s_5 )</td>
<td>( s_5 )</td>
<td>( s_5 )</td>
<td>( s_5 )</td>
<td>( s_5 )</td>
</tr>
</tbody>
</table>

Table 9.1: The State Transition Table for the Permutation Reset Machine of Example 9.9

The inputs are \( \{u_1, \ldots, u_6\} \), and the states are \( \{s_1, \ldots, s_6\} \). The element in the \( i \)th row and \( j \)th column of the table indicates the state of the system if the previous state is \( s_i \) and the input is \( u_j \). Note that \( u_1, u_2, \) and \( u_3 \) cause permutations of the states, while \( u_4, u_5, \) and \( u_6 \)
are resets. In fact, Table 9.1 is the multiplication table for a semigroup, $S$. That is, our state dynamic equation is

$$x(k+1) = u(k)x(k) \quad (9.149)$$

Thus, if $\eta(k)$ is the distribution of $u(k)$ and $\rho(k)$ the distribution for $x(k)$, then, if we regard them as elements of $R^1(S)$, we have

$$\rho(k+1) = \eta(k)\rho(k) \quad (9.150)$$

A straightforward multiplication scheme requires 18 multiplications and 15 additions. Thus

$$J_{naive} = 105 \quad (9.151)$$

Using the Brockett-Dobkin approach, we can find a scheme that requires 14 multiplications and 18 additions for a cost of 88. We note that for a general permutation-reset machine, part of the semigroup $S$ is a permutation group $G$, and we can expect a computational savings for $R^1(S)$ of the same order as the savings for $R^1(G)$.

**Example 9.10:** Consider a FSMP that evolves on the following semigroup. Let $G_1$ and $G_2$ be finite groups, such that, as sets, they are disjoint. Then let

$$S = G_1 \cup G_2 \cup \{a\} \quad (9.152)$$

where $a \notin G_1 \cup G_2$. Thus

$$\text{card } S = \text{card } G_1 + \text{card } G_2 + 1 \quad (9.153)$$

and we define the semigroup product on $S$ as follows: let $s_1, s_2 \in S$; then
The semigroup multiplication table is given in Table 9.2.

<table>
<thead>
<tr>
<th>( G_1 )</th>
<th>( G_2 )</th>
<th>( a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>( a )</td>
<td>( a )</td>
</tr>
<tr>
<td>( \cdots )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( \cdots )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( \cdots )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( a )</td>
<td>( a )</td>
<td>( a )</td>
</tr>
</tbody>
</table>

Table 9.2: Multiplication Table for the Semigroup of Example 9.10

Note that \( a \) is a sort of "trap" state - i.e. once reached, it cannot be left.

Now suppose our FSMP evolves on \( S \) according to

\[
x(k+1) = u(k) \cdot x(k)
\]  
(9.155)

Then we have that the unconditional probability density evolves in \( R^1(S) \) according to

\[
\sigma(k+1) = \eta(k) \cdot \sigma(k)
\]  
(9.156)

Also note that as a vector space over \( R^1 \)

\[
R^1(S) = R^1(G_1) + R^1(G_2) + R^1a
\]  
(9.157)

That is, given \( \sigma \in R^1(S) \), we have

\[
\sigma = \sum_{g \in G_1} \sigma \cdot g + \sum_{h \in G_2} \sigma \cdot h + \sigma \cdot a
\]  
(9.158)
Also, the semigroup algebra multiplication of an element of \( R^1(G_1) \) with an element of \( R^1(G_2) \) or of an element of \( R^1(G_1) + R^1(G_2) \) with an element of \( R^1a \) is an element of \( R^1a \). Thus we see that except for the evaluation of these cross terms -- which requires very little computation time -- the rest of the computation involved in \( R^1(S) \) multiplication consists of an \( R^1(G_1) \) multiplication and an \( R^1(G_2) \) multiplication. Therefore, the theory of efficient group algebra multiplication will lead directly to efficient \( R^1(S) \) multiplication in this case.

These examples illustrate what kind of structure we can expect in finite semigroups. That is, part of the semigroup may look like a group, while the other part involves some type of "reset" elements (i.e. the part whose elements are not invertible). The reset part creates no problem in \( R^1(S) \) multiplication, and thus any of the real complexity in \( R^1(S) \) multiplication is similar to the complexity encountered in group algebra multiplications. Thus, as in the group algebra case, we have a strong indication that a study of semigroup algebra multiplication will yield major savings in computation time. The reader is referred to [Kl1], [A3], [A4], [H11], and the work of Krohn and Rhodes [Kl6] for more on the structure of finite semigroups.

We now comment on the more general class of FSMP's that have arbitrary finite state sets and state evolution equations of the form (9.140). Given such a realization of a FSMP, we can consider enlarging the state set by considering the state transition semigroup equation
discussed in Appendix E. In this case, our input-state dynamics are of the form \((9.142)\), and the computation of probability distributions again involve semigroup algebra multiplications.

Suppose we can develop a theory of efficient semigroup algebra multiplication. We then have the following tradeoff: is the computational efficiency we gain by using the semigroup algebra formulation more than the computational loss incurred by (possibly) increasing the number of states to obtain a semigroup realization as in \((9.142)\) or \((9.143)\)? We will not consider this question here except to mention one point.

Suppose we have a realization of a FSMP of the type given in \((9.140)\) with a state set containing \(n\) elements. Then, as discussed in Appendix E, the state transition semigroup can have as many as \(n^n\) elements. Also suppose that for this particular semigroup, whose cardinality we denote by \(n'\), we can perform semigroup algebra multiplications with the number of real number multiplications being of the same order as in the \(R^1(Z_n)\) case -- i.e. \(O(n'\log n')\). Then, if \(n' = n^n\), this method is of no value, since the naive approach for computing the state probability distribution \(O(k+1)\) from \(O(k)\) for the \((9.140)\)-type realization requires \(O(n^2)\) multiplications, which is decidedly less than \(O(n^n \log n^n) = O(n^{n+1} \log n)\).

However, if the number of elements of the semigroup goes only linearly with \(n\) (i.e. \(n' \sim \alpha n\) for some \(\alpha\)), then, for large \(n\) we will reap a great savings by going to the semigroup equation.

We note that the success of the FFT for \(R^1(Z_n)\) depends strongly on \(n\) being a highly composite number -- i.e. that there exist integers \(n_1, \ldots, n_k\) greater than 1 such that

\[
n = n_1 n_2 \cdots n_k \quad (9.159)
\]
(see, for instance [N3] for a group algebra interpretation of this phenomenon). If \( n \) is not highly composite, there are techniques for embedding this computation in larger computations (i.e. increasing the cardinality of the state set) that decrease total computation time (the reader is referred to [R5]). Thus, this example indicates that, at least in some cases, the enlargement of the state set will lead to computational savings. The concepts discussed in this section have obviously not been fully developed, and further work is necessary here to determine optimal methods for computing conditional distributions for FSMP's.

9.4 Some Comments on Probabilistic Decoding

In Section 8.5 we discussed the class of convolutional encoders (CE's) — finite dimensional linear, invertible systems over finite fields (with several other restrictions) -- and introduced a larger class of coders -- finite group homomorphic encoders (FGHE's). A FGHE is simply an invertible FGHSS. We note that the class of FGHE's may be shrunk somewhat by the imposition of restrictions analogous to those for CE's. However, for the present, we will consider a FGHE in the unrestricted sense.

Also, in Section 8.5 we discussed the reasons usually given to defend the assumption that all random inputs to a CE (or a FGHE) are equally likely and sequentially independent. As mentioned in that section, there may be applications in which our a priori knowledge tells us that all data bit values are not equally likely or that the bits are sequentially
correlated. Thus, in this section we will briefly consider a probabilistic decoding problem.

Consider the FGHE

\[
\begin{align*}
x(k+1) &= b[u(k)]a[x(k)] & k & \geq 0 \\
r(k) &= c[x(k)] & k & \geq 1 \\
x(0) &= e
\end{align*}
\] (9.160)

with finite input, state, and output groups \(U, X, \) and \(Y,\) respectively.

Let its inverse system be

\[
\begin{align*}
z(k+1) &= \lambda[z(k), r(k)] & k & \geq 1 \\
w(k) &= \delta[z(k)] & k & \geq 2
\end{align*}
\] (9.163)

with \(z(1)\) given and \(z \in Z,\) a finite set. Also \(\lambda : Z \times Y \rightarrow Z\)
and \(\delta : Z \rightarrow U.\) This inverse deserves some comment. First, since
\(x(0) = e,\) the first output containing information is \(r(1);\) therefore
(9.161) and (9.163) start with \(k = 1.\) Also, the inverse system has a
starting state \(z(1),\) so the first \(w\) of any importance is \(w(2).\) Also,
as mentioned in Section 8.4, the inverse of a FGHSS need not be a FGHSS,
and thus, in general we can say no more about \(Z, \lambda,\) and \(\delta\) (or at least
this problem has not been thoroughly investigated). Finally, we assume
that (9.163), (9.164) is an \(L\)-delay inverse -- i.e., that

\[
u(k) = w(k+L+2)
\] (9.165)

(see the definition of \(L\)-invertibility in Section 8.2), and that this
is the least possible delay.

We will now add noise to our system. We assume that the \(\{u(k)\}\)
are independent with probability distributions \(\{\eta(k)\}.\) We can treat
certain types of sequentially correlated processes (and thus can consider the codewords or parts of codewords to have different probabilities, rather than the individual code letters) by using a "shaping filter," as we discussed for the \(Z_n\) case in Example 9.2, but for this brief discussion we assume the \(u(k)\) are independent. We also assume that transmission noise is present, and we do not observe \(r(k)\) but rather

\[
y(k) = v(k) \cdot r(k) = v(k) \cdot c[x(k)]
\]

where the \(\{v(k)\}\) are independent \(Y\)-valued random variables, independent of the \(\{u(k)\}\). We denote the probability distribution for \(v(k)^{-1}\) by \(\xi(k)\). Our problem is the following: given an \(M \geq 0\), devise a sequential scheme for computing a "good" estimate of \(u(k)\) from the observations \(y(1), ..., y(k+M+1)\). We first note that (9.165) tells us that it only makes sense to choose \(M \geq L\), or else we won't be able to determine \(u(k)\) from \(y(1), ..., y(k+M+1)\) even if there is no observation noise. Note that choosing \(M > L\) allows for some amount of smoothing.

We first present a straightforward approach to this probabilistic decoding problem. The basic idea is the following: given the observations, determine "good" estimates of the noise-free outputs \(r(k)\) and use these estimates as inputs to the FCHE inverse. The output of the inverse will be taken as the estimate of the input string. Thus, let \(\rho(k|l)\) be as in the previous sections. We can compute \(\rho(k|l)\) using the group algebra equations of Section 9.2 and a computational scheme devised using the techniques discussed in Section 9.3. Given a value of \(M \geq L\), we compute \(\rho(k|k+M-L)\), \(k \geq 1\), with initial condition
\[ \rho(0|-1) = \rho(0) = e \]  

(i.e. an impulse at the identity). The amount of storage space and computation time needed increases with \((M-L)\) -- i.e., they increase with the amount of smoothing that is done.

Let \(\mu(k|l)\) be the probability distribution for \(r(k)\) given \(y(1), \ldots, y(l)\). It is easy to check that

\[ \mu(k|l) = \mathcal{E}[\sigma(k|l)] = \sum_{g \in X} \sigma(k|l) \cdot c(g) \]  

Given \(\mu(k|k+M-L)\) choose an estimate \(\hat{r}(k)\) of \(r(k)\). One possibility is the most probable estimate

\[ \hat{r}(k) = \arg \max_{g \in Y} [\mu(k|k+M-L)]_g \]  

The string \(\{\hat{r}(k)\}\) is used as the input string to the inverse system. Note that \(\hat{r}(k)\) is a function of \(y(1), \ldots, y(k+M-L)\), and thus, sequentially \(\hat{r}(k)\) is produced at the time instant \(k+M-L+1\) (assume a unit of time for each distribution calculation), and thus the output \(\hat{u}(k)\) of the inverse system that is to be taken as the estimate of \(u(k)\) appears at the time instant \(k+M-L+3\) (comparing this to (9.165), we see that there is an extra unit delay, caused by the computation of the distributions). We note that these time estimates are quite conservative. For instance, if the time rate at which we transmit the coded information is substantially slower than the speed at which we can perform the necessary computations, we will be able to perform all the calculations -- i.e. the distribution and most probable estimate computations, as well as the evaluation of the inverse system functions \(\lambda\) and \(\delta\) -- in one time step and thus can produce
the estimate \( \hat{u}(k) \) at the time instant \( k+M-L+2 \).

Some comments on this procedure are in order. First of all, it is not at all clear that the necessary probability distribution computations can be performed in real time. If not, we are really talking about an off-line decoder, and, in this case, we can let \( M = \) total length of received output string (i.e. the full smoothing problem). Also, we must be careful in choosing the particular FCHE used. In particular, we wish to avoid a situation in which "small" changes in the sequence \( \{r(k)\} \) yield "large" changes in the output \( \{w(k)\} \) of the inverse system ("small" and "large" can be interpreted as referring to Hamming distance [B17]). If we have such a situation, any errors in our estimates, \( \{\hat{r}(k)\} \), of \( \{r(k)\} \) may cause large errors in the resulting estimates of the \( u(k) \).

Once the first few observations are processed -- i.e. once we have reached a "steady state" in the computational procedure -- the necessary computations can be described in the following manner. Just before the processing of the new observation \( y(m) \), we have stored in our computer the \((M-L+1)\) distributions \( \rho(m-M+L|m-1), \rho(m-M+L+1|m-1), \ldots, \rho(m-1|m-1), \rho(m|m-1) \). Using the smoothing equations (Theorem 9.3) and \( y(m) \), we compute the \((M-L)\) distributions \( \rho(m-M+L|m), \ldots, \rho(m-1|m) \). The first of these is used to compute \( \hat{u}(m-M+L|m) \) via (9.168), which yields our estimate \( \hat{r}(m-M+L) \). We discard \( \rho(m-M+1|m) \) and push the other densities down one in the list. We use the filtering equation (Theorem 9.2) and \( y(m) \) to compute \( \rho(m|m) \) and the prediction equation (Theorem 9.1 or 9.2) to compute \( \rho(m+1|m) \). At the end of this stage we have stored the \((M-L+1)\) distributions \( \rho(m-M+L+1|m), \ldots, \rho(m+1|m) \).
As an example of the type of computational load that this process produces, consider the case M=L and X=Z_n. In this case, we have stored ρ(m|m-1) and must compute ρ(m|m) and ρ(m+1|m). The first requires a Hadamard product and the second a polynomial product (see Theorem 9.2). As discussed in the previous sections, the Hadamard product requires n multiplications and the R⁻¹(Z_n) multiplication O(n log n) multiplications (where we have ignored the scaling factors, such as n(m|m)). Thus the computation time of this example goes as n log n.

In addition to this computational load, there is the problem of computing ˆP(k). Methods such as that suggested by (9.169) require quite exhaustive searches. A way of reducing this search procedure and, at the same time reducing the computational burden of the distribution calculation just discussed is as follows: scan each of the presently stored densities ρ(m-M+L|m-1),...,ρ(m|m-1), and set to zero the terms in each that fall below some threshold probability and scale up the remaining values. In this way, as we process more measurements, the density ˆP(k|m) for fixed k gets shifted to the left in the storage stack, and more and more of its entries are zero. Also, given ˆP(k|m), if we wish to compute ˆP(k|m+1), the smoothing equations of Theorem 9.3 involved a matrix ˆφ(m+1,k). If ˆP(k|m) = 0, the proof of Theorem 9.3, in particular equation (9.60), indicates that we need not compute the gth column of ˆφ(m+1,k) -- i.e. ˆP_g[ˆφ(m+1,k)]. This "tree pruning" sort of process can lead to large computational savings.

We close this section with a few final comments on probabilistic decoding. In estimating the u(k) given the observations y(1),...,y(k+M+1), we actually are considering the properties of the conditional distribution
\( \eta(k|k+M+1) \) for \( u(k) \) given these observations. This density is somewhat complicated, and note that, conditioned on the observations, \( u(k) \) and \( u(j) \) need not be independent. In any case, a method that computes the distribution \( \eta(k|k+M+1) \) directly and efficiently would be of great value, since a most probable estimate search over the elements of \( \eta(k|k+M+1) \) would be at most as complex as the search for \( \hat{v}(k) \) over \( \sigma(k|k+M-L) \), since we must have card \( U \leq \text{card } Y \) for invertibility of the FGHE (9.160), (9.161) (see Section 8.4). Also, the presently existing and quite effective maximum likelihood decoding schemes, such as the one developed by Viterbi [V3], may lead directly to efficient decoding algorithms for our problem.
CHAPTER 10

ESTIMATION ON CYCLIC GROUPS USING GROUP CHARACTERS

"The hell to be endured hereafter, of which theology tells, is no worse than the hell
we make for ourselves in this world by habitually fashioning our characters in
the wrong way."

- William James [J4]

10.1 Introduction

In this chapter, we will study at some length the properties
of random variables and random processes that evolve on the cyclic
group \( Z_n = \{0, 1, \ldots, n-1\} \) with addition defined modulo \( n \). Specifically,
we will develop tools similar to the Fourier series techniques used
in Chapters 3 and 5. These methods involve the use of group characters.
As we mention in the next section, the concept of group characters
can be used when dealing with arbitrary finite groups or with compact
Lie groups. In fact, for \( S^1 \) the group character approach is the same
as Fourier series analysis, and for \( S^n \), group characters lead to
spherical harmonics. In this sense, the results of the present
chapter can be viewed as extensions of the Fourier series results much
as spherical harmonics can (see Chapter 7), and, in turn, the techniques
we develop in this chapter can be extended to arbitrary finite groups
(although the details of such an extension have not been worked out
and are more difficult than in the \( Z_n \) case). We will discuss such an
extension in the next section, and the reader is also referred to
Chapter 7 and to the work of Kawada and Ito [K17], Rudin [R6], and
Grenander [G8] (see also Curtis and Reiner [C8], Feit [F5], and
Nicholson [N3]).
In Section 10.2 we define group characters for \( Z_n \) and examine properties of random variables, probability distributions, and estimation criteria. Several discrete-time estimation problems are considered in Section 10.3, and in Section 10.4 we discuss continuous-time problems.

Before beginning these discussions, it is appropriate to discuss the connection between the problems we will consider here and the single-degree-of-freedom rotation problems considered in Part I. Consider the circle \( S^1 \) identified with the interval \([0, 2\pi)\) with addition defined modulo \( 2\pi \). Then, for any integer \( n > 0 \), we have that the set \( C_n = \{ \frac{2\pi k}{n} \mid k = 0, \ldots, n-1 \} \) is a subgroup of \( S^1 \), and it is easy to see that \( C_n \) is isomorphic to \( Z_n \) when we make the identification

\[
k \leftrightarrow \frac{2\pi k}{n} \quad k = 0, \ldots, n-1
\]

(see Figure 9.1). Thus, we can regard \( Z_n \) as a discretization of the circle, and, by making \( n \) large enough, we can make the distance on the circle between \( \frac{2\pi k}{n} \) and \( \frac{2\pi (k+1)}{n} \) as small as we like.

With this in mind we can ask the following type of question: suppose we are trying to track a point moving on the circle but are only interested in some quantized measure of where the point is -- i.e. suppose we only wish to know which point in \( C_n \) is closest to the point we are tracking. Such a problem arises in some coding and digital signal applications [R7], [05], [G11], [H14], [W9]. Also, we may use such a quantized model for general phase tracking problems, since we can make the quantization error as small as we like by choosing \( n \) sufficiently large. Quantizations are particularly useful in digital
phase tracking; see [M14] and [B9] where quantizations are used to compute probability distributions in a manner much different from the approach we will take here.

Having made such a quantization, one is then interested in estimating processes on \( C_n \). The results of the next few sections are directly applicable to and, to a great extent, were motivated by this \( S^1 \) problem.

10.2 Group Characters and Random Variables on \( Z_n \)

In this section we define group characters on \( Z_n \), discuss their relationship to Fourier series, and present a number of results analogous to those in Chapter 3. We first wish to consider the set of all homomorphisms of \( Z_n \) into \( S^1 \), identified with the multiplicative group of complex numbers of unit modulus. One can show, [C8], [N3], that there are precisely \( n \) such homomorphisms, \( \{\chi^j\}_{j=0}^{n-1} \), where

\[
\chi^j(m) = e^{2\pi i jm/n}, \quad j, m \in \{0, 1, \ldots, n-1\}
\]  

(\( i = \sqrt{-1} \)). The functions \( \{\chi^j\}_{j=0}^{n-1} \) are called the **group characters** of \( Z_n \). We note one very important property of the group characters, [N3], is

\[
\sum_{j=0}^{n-1} \chi^r(j)\overline{\chi^s(j)} = n\delta_{rs}
\]  

where "\( - \)" denotes complex conjugate and \( \delta_{rs} \) is the Kronecker delta

\[
\delta_{rs} = \begin{cases} 0 & r \neq s \\ 1 & r = s \end{cases}
\]
Note the relationship between these ideas and Fourier series for $S^1$. Specifically, the set of all homomorphisms of $S^1$ (identified with $[-\pi, \pi]$) into $S^1$ (identified with the unit modulus complex numbers) is $\{\xi_j\}_{j=1}^{\infty}$ where

$$\xi_j(\theta) = e^{i\theta}$$

(10.5)

These functions are the group characters of $S^1$. Also, we have

$$\int_{-\pi}^{\pi} \xi_r(\theta)\overline{\xi_s}(\theta)d\theta = 2\pi\delta_{rs}$$

(10.6)

It is precisely the relationship between (10.2), (10.3) and (10.5), (10.6) that will be exploited in this chapter. The first result we will derive is the analog of the Fourier series decomposition of functions on $S^1$. Let $f$ be any complex-valued function defined on $Z_n$.

Define the character coefficients $\{\hat{f}_k\}$ of $f$

$$\hat{f}_k = \frac{1}{n}\sum_{j=0}^{n-1} f(j)\overline{\chi_k(j)} \quad k \in \{0, 1, \ldots, n-1\}$$

(10.7)

A simple calculation using (10.3) yields the character decomposition of $f$

$$f(j) = \sum_{k=0}^{n-1} \hat{f}_k \chi_k(j) = \hat{f}_0 + \sum_{k=1}^{n-1} \hat{f}_k e^{i2\pi jk/n}$$

(10.8)

Now suppose that $x$ is a random variable on $Z_n$ with probability distribution $p_x(j)$. We can then write

$$p_x(j) = \frac{1}{n} + \sum_{k=1}^{n-1} c_k e^{i2\pi kj/n}$$

(10.9)

where

$$c_k = \frac{1}{n} \sum_{j=0}^{n-1} (e - \frac{12\pi kx}{n})^{j} = \frac{1}{n} \sum_{j=0}^{n-1} p(j)e^{\frac{i2\pi jk}{n}}$$

(10.10)
As mentioned above, this decomposition is analogous to the Fourier decomposition of probability densities on $S^1$. We can also derive a number of other analogous results. Suppose $x$ and $y$ are independent random variables on $Z_n$ with distributions $p_x(j)$ and $p_y(j)$, respectively. Let (10.9) be the character decomposition for $p_x$, and let $p_y$ be given by

$$p_y(j) = \frac{1}{n} + \sum_{k=1}^{n=1} d_k e^{i2\pi kj/n}$$  \hspace{1cm} (10.11)

If $z$ is the random variable defined by

$$z = (x+y) \mod n$$  \hspace{1cm} (10.12)

and if its probability distribution is

$$p_z(j) = \frac{1}{n} + \sum_{k=1}^{n=1} g_k e^{i2\pi kj/n}$$  \hspace{1cm} (10.13)

then we have the equation

$$g_k = n c_k d_k$$  \hspace{1cm} k=0,\ldots,n-1$$  \hspace{1cm} (10.14)

which is the analog of (5.37). We note that the results of Chapter 9 could have been used here. That is, if we considered $p_x$, $p_y$, and $p_z$ to be elements of $R^1(Z_n)$, then

$$p_z = p_x p_y$$  \hspace{1cm} (10.15)

Recalling that (10.15) is a type of convolution product, we see that, analogous to the $S^1$ case, convolution of distributions corresponds to the multiplication of their character coefficients.

There is a class of probability densities on the real line that are called stable [F3]. That is, given two stable distributions of
the same type, their convolution is again of the same type. Clearly
the normal distribution is stable on \( \mathbb{R}^1 \), and the folded normal density
is stable on \( S^1 \). We now wish to find some stable distributions on
\( \mathbb{Z}_n \). To do this, we assume that there is a parameter \( \gamma \) that characterizes
some particular stable distribution on \( \mathbb{Z}_n \), and we assume that the
random variables \( x, y, \) and \( z \), defined earlier, have distributions of
this form for different values of \( \gamma \). Specifically, we assume

\[
c_k = \frac{1}{n} f_k(\gamma_1) \quad d_k = \frac{1}{n} f_k(\gamma_2) \quad g_k = \frac{1}{n} f_k(\gamma_1 + \gamma_2)
\]  \hspace{1cm} (10.16)

In this case, (10.14) implies the functional equations

\[
f_k(\gamma_1 + \gamma_2) = f_k(\gamma_1) f_k(\gamma_2) \quad k=0,\ldots,n-1
\]  \hspace{1cm} (10.17)

which implies (assuming the \( f_k \) are continuous [H8])

\[
f_k(\gamma) = e^{\gamma \beta(k)} \quad k=0,\ldots,n-1
\]  \hspace{1cm} (10.18)

for some function \( \beta \). Of course any set of \( f_k \) satisfying (10.18)
won't do, since we must have a set that yields a probability distribution --
i.e., we must have that \( p(j) \), given by

\[
p(j) = \frac{1}{n} + \frac{1}{n} \sum_{k=1}^{n-1} f_k(\gamma) e^{i 2 \pi j k / n}
\]  \hspace{1cm} (10.19)

is real and positive for \( j = 0,\ldots,n-1 \).

An example that satisfies all these requirements is

\[
\beta(k) = - \binom{n}{k} = - \frac{n!}{k!(n-k)!}
\]  \hspace{1cm} (10.20)

In this case, (10.19) becomes
\[ p(j) = \frac{1}{n} + \frac{1}{n} \sum_{k=1}^{n-1} e^{-\gamma \left( \frac{n}{k} \right) \cos \frac{2\pi k j}{n}} \]  

(10.21)

This distribution has zero as its mode, since

\[ p(0) \ll p(j) \quad \forall j \quad (10.22) \]

\[ p(j) = p(n-j) \quad \forall j \quad (10.23) \]

A more general stable density is the non-zero mode version of this

\[ S(j; \eta, \gamma) = \frac{1}{n} + \frac{1}{n} \sum_{k=1}^{n-1} e^{-\gamma \left( \frac{n}{k} \right) \cos \frac{2\pi k (j-\eta)}{n}} \]  

(10.24)

where \( \eta \in \mathbb{Z}_n \).

We can also discuss estimation criteria and optimal estimation in this framework. Suppose \( x \) is a \( \mathbb{Z}_n \)-valued random variable with distribution \( p_x \) given by (10.9). Let \( \phi \) be a real-valued function on \( \mathbb{Z}_n \) with the character decomposition

\[ \phi(j) = d_0 + \sum_{k=1}^{n-1} d_k e^{-\frac{12\pi i k}{n}} \]  

(10.25)

(note the minus sign in (10.25)). Suppose we wish to choose the estimate \( \hat{x} \) of \( x \) that minimizes \( \mathcal{E}(\phi(x-\hat{x})) \). As in Chapter 3 we compute

\[ \mathcal{E}(\phi(x-\hat{x})) = \sum_{j=0}^{n-1} \phi(j-\hat{x}) p_x(j) = d_0 + n \sum_{k=1}^{n-1} c_k d_k x_k(\hat{x}) \]  

(10.26)

Since \( \mathbb{Z}_n \) is a discrete set, we cannot differentiate (10.26) with respect to \( \hat{x} \) to obtain necessary conditions.

We now consider an important example -- the \( \mathbb{Z}_n \) analog of the \( S^1 \) criterion \( \mathcal{E}(1-\cos(\theta-\hat{\theta})) \). Let \( \phi(j) = 1-\cos \frac{2\pi j}{n} \). The character decomposition of \( \phi \) yields the coefficients

\[ d_0 = 1 \quad d_1 = d_{n-1} = -\frac{1}{2} \]  

(10.27)
with all other \( d_k = 0 \). In this case (10.26) becomes

\[
\mathcal{E}(\phi(x-\hat{x})) = 1 - \frac{n}{2} \left( c_1 e^{\frac{12\pi^2 i}{n}} + c_{n-1} e^{\frac{12\pi(n-1)i}{n}} \right)
\] (10.28)

Since \( p_x \) is real-valued, we can show that

\[
\text{Re}(c_1) = \text{Re}(c_{n-1}) \quad \text{Im}(c_1) = -\text{Im}(c_{n-1})
\] (10.29)

and thus

\[
\mathcal{E}(\phi(x-\hat{x})) = 1 - n \left[ \text{Re}(c_1) \cos \frac{2\pi^2 i}{n} - \text{Im}(c_1) \sin \frac{2\pi^2 i}{n} \right]
\] (10.30)

We note that we can write

\[
\text{Re}(c_1) = A \cos \eta \quad \text{Im}(c_1) = -A \sin \eta
\] (10.31)

with \( A > 0 \). In this case, we have

\[
\frac{\mathcal{E}(\phi(x-\hat{x})) - 1}{nA} = -\cos(\eta - \frac{2\pi^2 i}{n})
\] (10.32)

writing \( \hat{\theta} = \frac{2\pi^2 i}{n} \), we see that we wish to maximize \( \cos(\hat{\theta} - \eta) \) subject to the restriction \( \hat{\theta} \in \{0, \frac{2\pi}{n}, \ldots, \frac{2(n-1)\pi}{n}\} \). A little thought yields the solution

\[
\hat{\theta} = \text{element of } \{0, \frac{2\pi}{n}, \ldots, \frac{2(n-1)\pi}{n}\} \text{ closest to } \eta \text{(mod } 2\pi) \] (10.33)

Thus, we have the estimation solution: compute

\[
\eta = \tan^{-1} \left( -\frac{\text{Im}(c_1)}{\text{Re}(c_1)} \right) \quad \eta \in [0,2\pi)
\] (10.34)

Then the optimal estimate is

\[
\hat{x} = \left[ \frac{n}{2\pi} \eta \right]_n
\] (10.35)

where \( [\alpha]_n \) = the integer closest to \( \alpha \) (modulo \( n \)) -- i.e.
Thus, this solution is quite similar to the $S^1$ solution. Note that $\eta$ in (10.34) is the optimal estimate in the $S^1$ problem, and, therefore, in this case, the optimal estimate for the quantized problem is the quantization of the optimal estimate for the $S^1$ problem.

It is appropriate to mention a possible generalization of this approach to random variables and processes on arbitrary finite groups. In Chapter 9, we saw how to generalize polynomial multiplication by introducing the concept of the real group algebra of a finite group. Analogous to this, one can define group characters for arbitrary finite groups and can extend the ideas developed in this and the following sections. The reader is referred to [C8], [F5], [K17], [R6], and [G8] for more on group characters and their applications.

We note that the extension of our results to the arbitrary finite abelian group case is quite simple. It is well known [F1], [L1] that any finite abelian group $G$ is isomorphic to a direct product of cyclic groups

$$G = \mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z}$$

(10.37)

$$\mathbb{Z}^{r_1} \times \mathbb{Z}^{r_2} \times \cdots \times \mathbb{Z}^{r_n}$$

where the $p_i$ are not necessarily distinct primes (see [F1] for an alternative direct product decomposition). Identifying $G$ with such a direct product as in (10.37), we can define group characters on $G$ in
precisely the same way that we define multidimensional Fourier series.

Let

$$G = \prod_{i=1}^{r} Z_{n_i}$$  \hspace{1cm} (10.38)

In this case $|G| = n = \prod_{i=1}^{r} n_i$, and there are precisely $n$ homomorphisms of $G$ into the unit modulus complex numbers. We see this as follows:

let $\chi_{nm}$ be the homomorphism of $Z_{n}$ into $S^1$ defined by

$$\chi_{nm}(j) = e^{\frac{2\pi \imath j}{n}} \hspace{1cm} j = 0, \ldots, n-1$$  \hspace{1cm} (10.39)

Then the function $\chi_{m_1, \ldots, m_r} : G \to S^1$ defined by

$$\chi_{m_1, \ldots, m_r}(j_1, \ldots, j_r) = \chi_{n_1 m_1}(j_1) \chi_{n_2 m_2}(j_2) \cdots \chi_{n_r m_r}(j_r)$$  \hspace{1cm} (10.40)

is a homomorphism (here $(m_1, \ldots, m_r) \in G$), and there are precisely $n$ of these. That is, the characters of $G$ are precisely all possible products of characters of the $Z_{n_i}$. To simplify notation, we will denote elements of $G$ such as $(m_1, \ldots, m_r)$ and $(j_1, \ldots, j_r)$ by single letters, such as $g$.

In this case we can define the character coefficients of any complex-valued function $f$

$$\hat{f}_h = \frac{1}{n} \sum_{g \in G} f(g) \overline{\chi_h(g)}$$  \hspace{1cm} (10.41)

We can show, [N3], that for $h, t \in G$

$$\sum_{g \in G} \chi_h(g) \overline{\chi_t(g)} = n \delta_{ht}$$  \hspace{1cm} (10.42)

Using (10.41) and (10.42), we can write the character decomposition of $f$
\[ f(g) = \sum_{h \in G} \hat{f}_h \chi_h(g) \quad (10.43) \]

The results of this and the subsequent sections can be readily extended to the finite abelian group case by utilizing an identification such as (10.37) and the character equations developed above (the general case is essentially just a multidimensional version of the \( Z_n \) problem).

We close this section by pointing out another analogy between \( S^1 \) and \( Z\). In Part I we often found it useful to map random variables and random processes from \( R^1 \) onto \( S^1 \) by the transformation

\[ x \mapsto x \mod 2\pi \quad (10.44) \]

It is then natural to consider mapping variables and processes from \( Z \) into \( Z_n \) via

\[ m \mapsto m \mod n \quad (10.45) \]

If \( x \) is an integer-valued random variable with distribution \( p_x \), then \( y = x \mod n \) has the folded distribution

\[ n_y(k) = \sum_{j=-\infty}^{\infty} p_x(k+jn) \quad (10.46) \]

10.3 Character Equations for Discrete-Time Filtering on \( Z_n \)

In this section we will consider a general class of estimation problems on \( Z_n \) in a manner analogous to the approach used in Section 5.2.

We note that the results we present here are applicable to the \( Z_n \) problems considered in Chapter 9 (see Examples 9.2 and 10.1). The comparison of the two methods may be of some interest, since we have the interpretation
that the group algebra approach deals with the probability distributions, while the approach we will develop here deals with the character transforms of the distributions. We will have something to say about the duality of these approaches at the end of this section. We also note that just as with the Fourier series equations of Section 5.2 (Example 9.5), the computational techniques discussed in Chapter 9 will simplify the computational load associated with the estimation problems we will consider. In fact, the intimate relationship between the FFT and the characters of \( Z_n \), [N3], may lead to very efficient algorithms. We will not compare the relative computational merits of the group algebra and group character approaches except to comment that the group character approach has the advantage of providing simple direct methods for computing optimal estimates for criteria such as \( \mathcal{E}[1 - \cos \frac{2\pi(x - \bar{x})}{n}] \) (see (10.34), (10.35)).

We now consider the following general problem. Let \( x \) be a random variable on \( Z_n \) with a priori distribution
\[
p_x(j) = \frac{1}{n} + \sum_{k=1}^{n-1} c_k(0)e^{\frac{i2\pi kj}{n}}
\]
(10.47)

Suppose we take a noisy measurement, \( y \), of \( x \) (here \( y \) may be \( Z_n \)-valued, \( \mathbb{R}^1 \)-valued, etc.). Suppose the noise distribution or density (if, for instance \( y \) is real-valued) \( p_y|x(\nu|\xi) \) exists. From Bayes' rule, we have
\[
p_x|y(\xi|\nu) = \frac{p_y|x(\nu|\xi)p_x(\xi)}{p_y(\nu)}
\]
(10.48)

where
\[
p_y(\nu) = \sum_{j=0}^{n-1} p_y|x(\nu|j)p_x(j)
\]
(10.49)
We note that (10.48) and (10.49) hold either if \( y \) is discrete-valued and \( p_{y|x} \) is a probability distribution or if \( y \) can take on a continuum of values and \( p_{y|x} \) is a density (see [P2] and [L8]).

Considering \( p_{y|x}(\nu|\xi) \) as a function of \( \xi \) for fixed \( \nu \), \( p_{y|x}(\nu|\cdot) : \mathbb{Z}_n \to \mathbb{R}^1 \), we can write

\[
p_{y|x}(\nu|\xi) = h_0(\nu) + \sum_{k=1}^{n-1} h_k(\nu)e^{\frac{12\pi k\xi}{n}}
\]

(10.50)

Using (10.48) - (10.50), we find that

\[
p_{x|y}(\xi|\nu) = \frac{1}{n} + \sum_{k=1}^{n-1} c_k(1)e^{\frac{12\pi k\xi}{n}}
\]

(10.51)

where

\[
c_k(1) = \frac{\gamma_k(\nu)}{n\alpha(\nu)}
\]

(10.52)

\[
\alpha(\nu) = \frac{1}{n} p_y(\nu) = \frac{h_0(\nu)}{n} + \sum_{k=1}^{n-1} c_k(0)h_{n-k}(\nu)
\]

(10.53)

\[
\gamma_k(\nu) = \sum_{j=0}^{n-1} c_j(0)h_{k-j}(\nu)
\]

(10.54)

Here we have used the following notation

\[
c_0(0) = \frac{1}{n}
\]

(10.55)

\[
h_{k-j}(\nu) \triangleq h_{n+k-j}(\nu) \quad \text{if} \quad k-j < 0
\]

(10.56)

Equations (10.51) - (10.54) are the \( \mathbb{Z}_n \) analogs of (5.5) - (5.9). Note that in the present case all summations are finite, although, in general, all \( c_k(0) \) and \( h_j(\nu) \) directly enter the equation for each \( c_k(1) \). We now present two examples to illustrate this approach.
Example 10.1: In this example we will consider an observation that was previously considered in Example 9.2. This observation form is the $Z_n$ analog of the $S^1$ observation process $y = (\theta + \nu) \mod 2\pi$ considered in Example 5.2.

Let $x$ be a $Z_n$-valued random variable with probability distribution

$$p_x(j) = \frac{1}{n} + \sum_{k=1}^{n-1} \frac{12\pi k j}{n}$$ (10.57)

and suppose we observe the $Z_n$-valued variable

$$y = (x + \nu) \mod n$$ (10.58)

where $\nu$ is a $Z_n$ random variable independent of $x$ with distribution

$$p_\nu(j) = \frac{1}{n} + \sum_{k=1}^{n-1} \frac{12\pi k j}{n}$$ (10.59)

In this case

$$p_{y|x}(\nu|\xi) = p_\nu((\nu-\xi) \mod n)$$

$$= \frac{1}{n} + \sum_{k=1}^{n-1} d_k(0) e \frac{12\pi k(\nu-\xi)}{n}$$

$$= \frac{1}{n} + \sum_{k=1}^{n-1} [d_{n-k}(0) e \frac{12\pi k \nu}{n} + \frac{12\pi k \xi}{n}] e \frac{12\pi k \xi}{n}$$ (10.60)

and, in terms of (10.50),

$$h_k(\nu) = \begin{cases} d_{n-k}(0) e \frac{12\pi k \nu}{n} & k = 1, \ldots, n-1 \\ \frac{1}{n} & k = 0 \end{cases}$$ (10.61)

Using (10.51)-(10.54) and (10.61), we can compute $p_{x|y}(\xi|\nu)$. 

Note that a possible choice for $p_y$ is

$$p_y(\|) = S(\|; 0, y)$$

(10.62)

where $S$ is defined in (10.24). This choice is appealing, since $S$ is a stable distribution. What we would really like to have is the precise analog of a normal distribution on $Z$ and the associated folded normal on $Z_n$. One of the most important properties of the normal density (along with its being stable) is that if $x$ and $v$ are normal, independent, real-valued random variables, and if

$$y = x + v$$

(10.63)

then the conditional density $p_{x|y}$ is also normal. In Chapter 4 and Appendix B we used an equivalence class formulation to show that if we observe

$$\tilde{v} = (x + v) \bmod 2\pi$$

(10.64)

the density $p_{x|\tilde{v}}$ is an infinite sum of normals, and thus, if

$$\theta = x \bmod 2\pi, \tilde{v} = (\theta + v) \bmod 2\pi, p_{\theta|\tilde{v}}$$

is an infinite sum of folded normals.

It is not difficult to show that we have the same sort of equivalence class estimation problem here. That is, suppose $x$ and $v$ are independent integer-valued random variables and let

$$y = x + v$$

(10.65)

$$\tilde{v} = y \bmod n$$

(10.66)

Then one can show that

$$p_{x|\tilde{v}}(\alpha|\beta) = \sum_{k=-\infty}^{+\infty} \frac{p_{\tilde{v}}(\beta + kn)}{p_{\tilde{v}}(\beta)} p_{x|\tilde{v}}(\alpha|\beta + kn)$$

(10.67)
\[ p_y(\beta) = \sum_{k=-\infty}^{+\infty} p_y(\beta + kn) \quad (10.68) \]

It remains to determine the analog of the normal distribution -- i.e., a class of distributions on \( Z \) such that if the distributions for \( x \) and \( v \) are in this class, so is \( p_x|v \). In this case, \( p_x|v \) is an infinite sum of these "normal" distributions, and, if \( \theta = x \mod n \) and \( \tilde{v} = (\theta + v) \mod n \), \( p_{\theta|\tilde{v}} \) is an infinite sum of "folded normals." Finding this normal distribution is still an open problem.

Example 10.2: We now consider the analog of Example 5.1. This example provides the basis for considering the quantized, discrete-time version of the phase tracking problem discussed in Chapter 5.

Let \( x \) be a \( Z_n \)-valued random variable, and let \( v \) be a real-valued random variable, independent of \( x \), with density \( p_v(v) = N(v; 0, \gamma) \).

Suppose we observe
\[ y = \sin \frac{2\pi x}{n} + v \quad (10.69) \]

In this case
\[ p_y|x(v|\xi) = N(v - \sin \frac{2\pi x}{n}; 0, \gamma) \quad (10.70) \]

Using the character coefficient equation (10.7), we can write the character decomposition of \( p_y|x \) as in (10.50), and then we can use (10.51) - (10.54) to compute \( p_x|y \). We will not present the details but mention only that the character coefficients of \( p_y|x \) are complex functions of \( v \) and, as mentioned in Example 5.3, for practical application we can approximate the coefficients by polynomials in \( v \).

We note that just as in the \( S^1 \) case (see Examples 5.4 and 5.5), we can consider \( x \) to be a random process on \( Z_n \) and can let \( y \) be a
sequence of measurements. For example let \( \{w(k)\} \) be a sequence of
independent \( \mathbb{Z} \)-valued random variables, let \( f \) be a known positive integer,
and let \( x \) be the random process on \( \mathbb{Z}^n \) defined by
\[
x(k+1) = [x(k) + f + w(k)] \mod n \quad x(0) = 0 \quad (10.71)
\]
Here, if we interpret \( f \) as a "carrier phase," (10.71) is the discrete-
time, quantized version of the \( S^1 \) phase
\[
\theta(t) = [\omega_c t + w(t)] \mod 2\pi \quad (10.72)
\]
where \( \omega_c \) is a known carrier frequency and \( w \) is a Brownian motion.
Suppose we observe
\[
y(k) = \sin \frac{2\pi x(k)}{n} + v(k) \quad k=1,\ldots \quad (10.73)
\]
where the \( v(k) \) are normally distributed. We assume that the \( v(k) \) and
\( \omega(j) \) are all independent. Then, as in Chapter 5, if we let \( p(k|\ell) \) be
the distribution for \( x(k) \) conditioned on \( y(1),\ldots,y(\ell) \), we can compute
\( p(k|k) \) from \( p(k|k-1) \) as in Example 10.2. To compute \( p(k+1|k) \) from
\( p(k|k) \), we must devise a diffusion update equation. We note that
since \( f \) is known, it causes no problem since it just cyclically shifts
the distributions for \( x \) (i.e. if \( z = (x+f) \mod n \), \( \Pr(z=k) = \Pr(x=(k-f) \mod n) \)).
Thus, for simplicity, we assume that \( f = 0 \). Then, because of the independence
assumption, we can use (10.14). In this case, the kth character of
\( p(k+1|k) \) is simply \( n \) times the product of the kth character of \( p(k|k) \)
with the kth character of the distribution for \( w(k) \). We note that we
could also consider \( x \) to be a continuous-time jump process and use con-
tinuous-time equations to propagate the distribution between measurements,
but we will defer that discussion until the next section, where we will also consider continuous-time observation processes.

We close this section by making several comments on the relationship between the character equations and the group algebra equations for the following problem: let \( \{w(k)\} \) and \( \{v(k)\} \) be independent sequences of independent \( \mathbb{Z}_n \)-valued random variables, and let

\[
x(k+1) = [x(k) + w(k)] \mod n
\]

\[
y(k) = [x(k) + v(k)] \mod n
\]

To compute \( p(k|k) \) recursively, we have both diffusion and measurement updates. Referring to Chapter 9, we see that for the group algebra formulation, the diffusion update is a convolution, while the measurement update consists of pointwise (Hadamard) multiplication plus a normalization. For the group character formulation, equation (10.14) indicates that the diffusion update is a pointwise multiplication, while, referring to (10.54), we see that the measurement update equation consists of a convolution and a normalizing constant. This duality is conceptually the same as the duality between convolution and multiplication of functions on \( \mathbb{R}^1 \) and their Fourier transforms [R3].

10.4 A Continuous-Time Estimation Problem on \( \mathbb{Z}_n \)

In this section we will formulate and solve a phase tracking problem on \( \mathbb{Z}_n \). The problem we will consider is the \( \mathbb{Z}_n \) analog of Example 5.6, and we remark that we can use similar techniques to devise \( \mathbb{Z}_n \) analogs of the demodulation problems of Examples 5.7 and 5.8 (in those cases, we would take our signal to be integer valued in keeping with the
Let $x(t)$ be a continuous time jump process, $[C4]$, defined on $Z_n$. Let $p_x(t)$ be the $n$-vector whose $i$th component is $Pr(x(t) = i-1)$. We assume that $p_x$ satisfies the differential equation

$$\dot{p}_x(t) = A(t)p_x(t)$$

(10.76)

Such an equation might arise as the limit of a discrete-time equation such as (10.71). That is, given a $\Delta > 0$, consider the discrete-time process

$$x(t+\Delta) = [x(t)+w(t+\Delta)-w(t)] \mod n$$

(10.77)

where we assume

$$Pr(w(t+\Delta)-w(t)=k) = \beta_k(t+\Delta,t)$$

(10.78)

$$\lim_{\Delta \to 0} \beta_k(t+\Delta,t) = \delta_{k0}$$

(10.79)

$$\lim_{\Delta \to 0} \frac{\beta_k(t+\Delta,t)-\delta_{k0}}{\Delta} = \alpha_k(t)$$

(10.80)

where $\delta_{ij}$ is the Kronecker delta. In this case, we can write

$$p_x(t+\Delta) = B(t+\Delta,t)p_x(t)$$

(10.81)

where $B(t+\Delta,t)$ is a circulant matrix whose elements are the $\beta_k(t+\Delta,t)$. Then, we can take the limit

$$\dot{p}_x(t) = \lim_{\Delta \to 0} \frac{p_x(t+\Delta)-p_x(t)}{\Delta} = \lim_{\Delta \to 0} \frac{1}{\Delta} [B(t+\Delta,t)-I]p_x(t) = A(t)p_x(t)$$

(10.82)

where $A(t)$ is a circulant matrix whose elements are the $\alpha_k(t)$. 
In any event, we assume that we have a random process \( x(t) \) and associated unconditional distribution equation (10.76). We now consider a continuous-time observation process

\[
\dot{x}(t) = \sin \frac{2\pi x(t)}{n} + r^{1/2}(t)\dot{v}(t)
\]

(10.83)

where \( r(t) > 0 \) and \( v(t) \) is a Brownian motion process independent of \( x(t) \). Equation (10.83) is, of course, formal, since \( \dot{v} \) doesn’t exist, and thus we will use instead the Ito differential equation version of (10.83)

\[
dz(t) = \sin \left( \frac{2\pi x(t)}{n} \right) dt + r^{1/2}(t)dv(t)
\]

(10.84)

We wish to compute the conditional probability distribution

\[
P_{x|\cdot}(j,s,t) = \Pr(x(t)=j|z(s), 0 \leq s \leq t)
\]

(10.85)

This type of result has been derived by Wonham, [W7] (see also [F7]), and we will only sketch the derivation. Our problem is, in some sense, a general random telegraph signal problem (see [W7] and [R7]).

We refer to the general representation theorem presented in [L8]. Using this result, we can write

\[
P_{x|\cdot}(j,s,t) = \frac{\mathcal{G}_x(\Theta(t)|j)}{\mathcal{G}_x(\Theta(t))} P_{x}(j,t)
\]

(10.86)

where \( P_{x}(j,t) \) is the unconditional probability that \( x(t) = j \) and

\[
\Theta(t) = \exp(-\frac{1}{2} \int_0^t \left( \frac{\sin \frac{2\pi x(s)}{n}}{r(s)} \right)^2 ds + \int_0^t \left( \frac{\sin \frac{2\pi x(s)}{n}}{r(s)} \right) dz(s))
\]

(10.87)
Also, $E_x(\Theta(t) | j)$ is the expectation of $\Theta(t)$ over $x(s)$, $0 \leq s \leq t$ with $x(t) = j$ and $z(s)$, $0 \leq s \leq t$ fixed. The expectation $E_x(\Theta(t))$ is the expectation over $x(s)$, $0 \leq s \leq t$ with $z(s)$, $0 \leq s \leq t$ fixed (see [B8]). That is,

$$E_x(\Theta(t)) = \sum_{j=0}^{n-1} E_x(\Theta(t) | j)p_x(j, t) \tag{10.88}$$

Now we can derive a stochastic differential equation for $p_{x|z}(j, t)$ either in a manner precisely analogous to the derivations of Kushner's equation, [K5], given in [B8] and [J2] or via the approach used in [W7]. In either case, we obtain the result

$$dp_{x|z}(j, t) = [A(t)p_{x|z}(t)]_j dt$$

$$+ \left[ \sin\left(\frac{2\pi j}{n}\right) - \hat{h}(t) \right] \frac{dz(t) - \hat{h}(t) dt}{r(t)} p_{x|z}(j, t) \tag{10.89}$$

where $p_{x|z}(t)$ is the $n$-vector whose $j$th component is $p_{x|z}(j-1, t)$, and $[A(t)p_{x|z}(t)]_j$ is the $(j+1)$st component of the $n$-vector $A(t)p_{x|z}(t)$. Also

$$\hat{h}(t) = E\left(\sin\left(\frac{2\pi j x(t)}{n}\right) | z(s), 0 \leq s \leq t\right)$$

$$= \sum_{k=0}^{n-1} \sin\left(\frac{2\pi k}{n}\right) p_{x|z}(k, t) \tag{10.90}$$

We now consider the special case where $A(t)$ is a circulant matrix (see equations (10.77) - (10.82)). This essentially corresponds to the case in which $x(t)$ is an independent increment process. Referring to (10.77) - (10.82), we see that

$$[A(t)p(t)]_j = \sum_{k=0}^{n-1} \alpha_k(t)p_{(j-k)\mod n}(t) \tag{10.91}$$
where \( p_i(t) \) is the \((i+1)\)st component of \( p(t) \). Let us take the character transform

\[
q_k(t) = \frac{1}{n} \sum_{j=0}^{n-1} \alpha_j(t) e^{-\frac{12\pi j k}{n}}
\]  

(10.92)

Then from (10.8) we have

\[
\alpha_j(t) = \sum_{k=1}^{n-1} \frac{i 12\pi j k}{n} q_k(t) e^{-\frac{12\pi j k}{n}}
\]  

(10.93)

(it is easy to show that \( q_0(t) = 0 \)). Then suppose that we write the character decomposition of \( p_x|z(j,t) \)

\[
p_x|z(j,t) = \frac{1}{n} + \sum_{k=1}^{n-1} c_k(t) e^{-\frac{12\pi j k}{n}}
\]  

(10.94)

where

\[
c_k(t) = \frac{1}{n} \mathcal{E}[e^{-\frac{ni 2\pi k x(t)}{n}} | z(s), 0 \leq s \leq t]
\]  

(10.95)

Suppose we also write the character decomposition

\[
[A(t)p_x|z(t)]_j = g_o(t) + \sum_{k=1}^{n-1} g_k(t) e^{-\frac{12\pi j k}{n}}
\]  

(10.96)

One can then show that

\[
g_k(t) = n q_k(t) c_k(t)
\]  

(10.97)

(note that \( g_o(t) = 0 \)).

Having these results, we can use (10.89) to derive stochastic differential equations for \( c_k(t), k=1,\ldots,n-1 \). First, comparing (10.90) and (10.95), we have

\[
\hat{h}(t) = \frac{n}{21} (c_{n-1} - c_1)
\]  

(10.98)
where we have used

\[
\sin \frac{2\pi j}{n} = \frac{1}{2i} \left( e^{-\frac{2\pi j}{n}(n-1)} - e^{-\frac{2\pi j}{n}} \right)
\]  

Then, writing

\[
dc_k(t) = \frac{1}{n} \sum_{j=0}^{n-1} e^{-\frac{i2\pi kj}{n}} dp_x|z(j,t)
\]

we find that for \( k = 1, \ldots, n-1 \)

\[
dc_k(t) = nq_k(t)c_k(t)dt
\]

\[
\frac{[2idz(t)-n(c_{n-1}(t)-c_1(t))]dt}{r(t)} \frac{\{[c_{k-1}(t)-c_{k+1}(t)]-nc_k(t)[c_{n-1}(t)-c_1(t)]\}}
\]

where

\[
c_0(t) \equiv 1
\]

Recalling that

\[
\text{Re}[c_k(t)] = \frac{1}{n} \mathcal{E}[\cos \frac{2\pi kx(t)}{n} | z(s), 0 \leq s \leq t]
\]

\[
\text{Im}[c_k(t)] = -\frac{1}{n} \mathcal{E}[\sin \frac{2\pi kx(t)}{n} | z(s), 0 \leq s \leq t]
\]

we see a remarkable similarity between (10.101) and the Fourier series phase tracking equations (5.50) and (5.51) of Example 5.6.

Also, from (10.102) and (10.103) we see that

\[
c_k(t) = \overline{c}_{n-k}(t)
\]

and thus we need only compute \( c_k(t) \) for \( k = 1, \ldots, \left\lfloor \frac{n-1}{2} \right\rfloor \). In addition, we see that the \( dc_k \) equation is directly affected only by \( c_k, c_1, c_{n-1}, c_{k-1}, \) and \( c_{k+1} \), and we have the filter form given in Figure 10.1, where we use
Figure 10.1: Illustrating the Form of the Optimal Filter Developed in Section 10.4
the estimation criterion \[ \varepsilon[1 - \cos \frac{2\pi(x(t) - z(t))}{n}] | z(s), 0 \leq s \leq t]. \]

Note that the filter form is much like that of Example 5.6 (see Figure 5.4), except that in this case the filter is finite dimensional. We also note that the inclusion of a carrier frequency in the preceding analysis causes no difficulties.

Filtering techniques and problem formulations of this type are quite important for certain applications. In addition to the use of these techniques in cases where we discretize the signal space for digital signal processing [05], [G11], there are applications to some coding and Doppler Radar problems. As mentioned in Chapter 6, one type of Doppler-shift-detecting system [L16], [E3], [H20] consists of a cycle counter and an interpolator that measures a fraction of a cycle. The interpolator essentially measures a \(^Z_n\)-valued random variable -- i.e. it measures \(k \in Z_n\), where the fraction of the cycle to be estimated is nearer to \(\frac{k}{n}\) than any other \(\frac{j}{n}, j \in Z_n\).

One coding problem that can be analyzed using the methods of this chapter is the phase-shift-keying (PSK) coding technique [S10], [L17], [H14], [W9]. Essentially a signal of the form

\[ s(t) = \sin(\omega_c t + \frac{2\pi}{n} x(t)) \]  \hspace{1cm} (10.106)

is transmitted, where \(x(t)\) is a \(^Z_n\)-valued random process that is to be recovered from the received signal, which is usually modeled as \(s(t) + \text{noise}\). It is clear that the problems we have considered in this chapter -- both discrete and continuous time -- include the PSK decoding problem. The interested reader is referred to [S10], [L17], and [W9] for detailed descriptions of PSK systems.
CHAPTER 11
SUMMARY, CONCLUSIONS, AND SUGGESTIONS

"I'll let you be in my dream if I can be in yours."
- Bob Dylan [D9]

In this dissertation, we have developed a number of analytical tools and have used these techniques to study a number of nonlinear estimation and control problems. We have been motivated by the general conceptual approach of considering a problem in its most natural setting in order to uncover useful techniques and results suggested by the inherent structure of the problem. Specifically, we have considered a number of problems that can be analyzed using some techniques from the theory of groups -- both continuous and finite. The major contributions of this research are:

(1) The development of various optimal estimation criteria on the circle, $S^1$, using Fourier series analysis, the properties of the folded normal density on $S^1$, and the rearrangement inequality for unimodal densities and symmetric criteria.

(2) The development of easily implemented optimal estimation equations for a class of continuous-time bilinear estimation problems on the circle and on arbitrary abelian Lie groups.

(3) The formulation and solution of a discrete-time $S^1$ estimation problem that is remarkably more complicated than its continuous-time counterpart; the discussion of several suboptimal estimation schemes motivated by the physical interpretation of the exact conditional density equations.
(4) The development of phase tracking and demodulation techniques using Fourier series analysis.

(5) The application of the various $S^1$ estimation techniques to a number of physically important problems, such as phase acquisition, synchronization, and demodulation in the presence of additive channel noise, random phase drift, and amplitude fading.

(6) The development of some new techniques, motivated by the $S^1$ results, for the estimation of the angular velocity and orientation of a rigid body; a generalization of these ideas to problems on arbitrary matrix Lie groups.

(7) The formulation of a class of systems evolving on finite groups and the development of a detailed structure theory analogous to that for discrete-time linear systems.

(8) The study of a class of finite-state Markov processes on finite groups and the introduction of an algebraic concept (group algebra) that results in a form for the conditional distribution equations that leads to efficient computational methods and is related to a generalization of fast Fourier transform techniques.

(9) The development of a finite group analog of a class of $S^1$ estimation problems and the use of finite Fourier transform methods to determine optimal estimation equations that have direct application to some phase coding and digital signal processing problems.

It is the author's feeling that the most important point made in this manuscript is the value of the conceptual approach of choosing one's tools to match the problem. In this research we have used tools drawn only from the theories of groups and vector spaces, and it is hoped that future work using other techniques, such as those found in
ring theory and in the theory of differentiable manifolds, will yield additional valuable results.

The following is a partial list of future research topics related to the problems and approaches discussed in this thesis.

(1) The use of spherical harmonics to study random processes of rotation and to develop orientation estimation techniques; development of general estimation techniques using Fourier analysis on groups -- connections with the group characters of finite groups.

(2) Use of the techniques developed in Part I to design systems for application to some of the phase tracking problems associated with Loran and Omega navigation systems, Doppler shift determination, and deep space communications.

(3) Development of new classes of finite-state algebraic coding systems evolving on groups. Consideration of both deterministic and probabilistic aspects of encoder and decoder design.

(4) A study of group algebra structure in order to uncover computational tools that generalize the fast Fourier transform and reduce the computational load associated with certain Markov process control problems.

(5) The use of spherical harmonics to study the problem of predicting the position of a vehicle orbiting a sphere possessing gravitational anomalies (see [P9]).

(6) The development of techniques for the control of rotational processes and processes evolving on arbitrary Lie groups (see [L14] for some results on the stochastic control of rotational processes with one degree of freedom). Study of such control problems as orientation control
and attitude stability using control moment gyros and dual spin systems [L18], [N4], [H19], [M18], [S21].

(7) Further study and simulation of the angular velocity and orientation estimation techniques motivated by the $S^1$ results and discussed in Chapter 7.

(8) Consideration of more complex communication problems, such as joint AM-FM modulation [M6] and the multipath problem [V2] that arises in many ionospheric communication [P4] and underwater wave propagation problems [V6].

(9) Algebraic decomposition theory for finite group homomorphic sequential systems and the associated group algebra systems.

(10) The formulation and study of systems evolving homomorphically in other algebraic settings, such as rings (see [J3]).

(11) Numerical work for the series truncation techniques discussed in Appendices C and D and associated with the problems discussed in Chapters 4 and 5, respectively.

(12) Simulations and development of other techniques for the phase coding applications discussed in Chapter 10, such as the phase-shift-keying problem.

(13) The use of the estimation techniques discussed in this thesis to devise methods for the maximum likelihood identification of bilinear systems and other systems evolving on groups; also, the study of such identification concepts as parametrizations [A15].

(14) The design of distributed circuits [J5] that realize the infinite dimensional optimal phase trackers studied in Chapter 5.
(15) Extension of the group character estimation techniques, discussed in Chapter 10, to problems on arbitrary finite groups (see [H17]).

(16) The study of random vector fields and the associated integral curves on a differentiable manifold [Wl].

(17) Investigation of the use of the finite group systems of Chapter 8 in the optimization of basic numerical recursions, such as integer programming [S22].

(18) Further simulations of the techniques developed in Chapter 5; consideration of such problems as demodulation and phase tracking when the carrier frequency is not precisely known or Doppler shifted. Extensions of the simulations discussed in Section 6.4, including the investigation of other truncation techniques and the effects of adding another term to the Fourier series density approximations. Theoretical development to determine if the phase tracking system, whose simulation is discussed in Section 6.4, is optimal below threshold.

(19) The study of stable distributions on finite groups and their use in probabilistic decoding theory.

(20) The study of some of the open questions associated with FGHSS's, including an investigation of the Krohn-Rhodes decomposition of a FGHSS, as well as the questions of feedback, partial realizations, and the construction of inverses.
APPENDIX A

A SUMMARY OF SOME BASIC RESULTS IN MODERN ALGEBRA

"'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

'Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!'"

- "Jabberwocky," Lewis Carrol [T1]

A.1 Introduction

It is the purpose of this appendix to describe some of the basic concepts of modern algebra. In particular, we will explore those basic concepts from the theories of groups, fields, and Lie groups that are relevant to the results presented in this thesis. For more thorough treatments of these subjects, see Rotman [R1], Lang [L1], Fraleigh [F1], and MacLane and Birkhoff [M1] (group and field theory), and Chevalley [C1], Warner [W1], and Samelson [S1] (Lie group and Lie algebra theory).

A.2 Groups

Let us lay the foundation with some basic definitions.

**Definition A.1:** A binary operation, "*", on a set S is a rule which assigns to each ordered pair (a,b) of elements of S some element, denoted by a*b, of S.
Examples of binary operations are addition of real numbers and multiplication of nxn matrices.

**Definition A.2:** A binary operation \( \ast \) on a set \( S \) is **commutative** if \( a \ast b = b \ast a \) \( \forall a, b \in S \). The operation is **associative** if \( (a \ast b) \ast c = a \ast (b \ast c) \) \( \forall a, b, c \in S \).

Note that addition of numbers and multiplication of nxn matrices are both associative, but matrix multiplication is not commutative while addition is. Also, the operation of subtraction of real numbers \( (a \ast b \triangleq a - b) \) is neither commutative \( (a - b \neq b - a) \) nor associative \( (a - (b - c) \neq (a - b) - c) \). Finally, we remark that for associative operations \( a \ast b \ast c \) is unambiguously defined, and such notation will be used in this manuscript.

**Definition A.3:** A **semigroup** \( (S, \ast) \) is a set \( S \), together with an associative binary operation \( \ast \).

**Definition A.4:** A **monoid** \( (M, \ast) \) is a semigroup with an **identity element** \( e \in M \) — i.e. \( e \ast m = m \ast e = m \) \( \forall m \in M \).

**Definition A.5:** A **group** \( (G, \ast) \) is a monoid with the property that for each \( g \in G \) there exists an element \( g^{-1} \in G \) such that \( g \ast g^{-1} = g^{-1} \ast g = e \), where \( e \) is the identity element. The element \( g^{-1} \) is an **inverse** of \( g \) with respect to \( \ast \).

Some examples of semigroups are: the set \( \{1, 2, 3, \ldots\} \) with addition as the operation, the set of nxn matrices of determinant \( \neq 2 \) under matrix multiplication, and the set of all nxn matrices with matrix multiplication. This last example is also a monoid, with the identity matrix as the identity. Other monoids are: the set \( \{0, 1, 2, 3, \ldots\} \) under addition and the set \( \{0, 1, \ldots, n-1\} \) under multiplication modulo \( n \). In the former, the number
"0" is the identity, while "1" serves that purpose in the latter. Note that none of the above examples are groups. Some groups are: the set of all integers (positive or negative) under addition, the set of all real numbers under addition, the set of all nxn invertible matrices under matrix multiplication, and the set \{0, 1, \ldots, n-1\} under addition modulo n. This last group is usually denoted \( Z_n \).

In the main text, we concern ourselves, for the most part, with the properties of groups and make only passing references to problems involving semigroups and monoids. Thus, in the rest of this section we will concentrate on developing the theory of groups, and thus the binary operations used will be referred to as "group multiplications" or "group operations."

It is appropriate to make a note concerning sloppy notation. When the group operation \( * \) is understood, we often write \( ab \) for \( a * b \) and will refer to "the group \( G \)", as opposed to "the group \( (G, *) \)." Such notation will be used whenever there is no chance of ambiguity.

Before discussing some of the deeper properties of groups, it is appropriate to comment on some relatively immediate consequences of Definition A.5. First, if \( G \) is a group, then the right and left cancellation laws hold -- i.e. \( ba = ca \) implies \( b = c \), and \( ab = ac \) implies \( b = c \). Also the equations \( ax = b \) and \( va = b \) have unique solutions (namely \( x = a^{-1}b \) and \( y = ba^{-1} \)), the identity element \( e \) of \( G \) is unique, as is the inverse of any given element of \( G \), and finally \( (g^{-1})^{-1} = g \) \( \forall g \in G \) and \( (ab)^{-1} = b^{-1}a^{-1} \) \( \forall a, b \in G \). The proofs of these are straightforward applications of the definitions (see [F1], pp. 15-17).
Definition A.6: Let \( G \) be a group and \( a \in G \). Define the powers of
\( a \) as follows: if \( n \) is a positive integer, then \( a^n = a^*a^*...a \) (\( n \) times);
if \( n \) is negative define \( a^n \) as \((a^{-1})^{-n}\); and let \( a^0 = e \).

Definition A.7: Let \( S \) be a subset of a group \( G \). Then \( S \) generates \( G \) if
the set
\[
\langle S \rangle \triangleq \{ s_1^{n_1} s_2^{n_2} ... s_k^{n_k} | k \geq 1, s_i \in S, n_i \in \mathbb{Z} \}
\]
(A.1)
is all of \( G \). The elements of \( S \) are collectively termed a set of
generators of \( G \).

Definition A.8: A group \( G \) is cyclic if there is an \( a \in G \) such that
the singleton set \( \{ a \} \) generates \( G \).

Note that by associativity \( a^n a^m = a^{n+m} = a^m a^n \), and also that if
\( G \) is cyclic, all the elements are of the form \( a^k \). An example of a
cyclic group is \( \mathbb{Z}_n \), where all of the numbers \( 0, ..., n-1 \) can be written
as \( l_k \triangleq \{ (l+1+...+1) \mod n \} \) for various values of \( k \) (we caution the reader
to remember that group multiplication for \( \mathbb{Z} \) is what is usually called
addition, and group multiplication for \( \mathbb{Z}_n \) is addition modulo \( n \); it is
in this sense that \( l_k \) should be read as \( k \mod n \); the reader is referred
to Definition A.9, where we remove some of these problems). In addition,
the set of all integers is cyclic and is also generated by 1. Then why
are these two groups different? The reason is that in the former case
we impose the relation \( l_n^m = 0 \).

The concept of using generators and sets of relations among
the generators to construct groups has been studied in great
detail (see [L1] for more on this subject). We will not go
into detail on this subject but will present one example that
is used in the main text of this manuscript. Consider the group

generated by 2 elements, $x$ and $y$, with the relations

$$x^n = e \quad y^2 = e \quad yxy = x^{-1} = x^{n-1}$$

(A.2)

This group, called the dihedral group $D_n$, can be shown to contain $2n$
elements (namely $e$, $x$, $x^2$, $x^{n-1}$, $y$, $xy$, $x^2y$, $x^{n-1}y$) and can be
physically interpreted as the set of all vertex-preserving rigid
motions of an $n$ sided regular polygon -- i.e. any vertex is sent by
any such rigid motion to a location previously occupied by some other
vertex.

**Definition A.9:** A group $G$ is **abelian** if group multiplication is
commutative.

Note that all cyclic groups are abelian, but that $D_n$ is not (for
$n \geq 3$). For abelian groups, we usually use "+" for the group operation
and write $na$ for $a^n$. To avoid the problems caused, for example, by
referring to integer addition as group multiplication, we will usually
refer to the group operation of an abelian group as addition, in keeping
with the use of the symbol "+".

We now come to the concept of **group isomorphism**. First we note
that all the properties of a given finite group can be summarized in
a group multiplication table. Let $a_1, \ldots, a_n$ be the elements of $G$.
Then the multiplication table of $G$ is an $n \times n$ matrix whose $ij$th element
is $a_ia_j$. Let $G_1 = \{1, -1\}$ with multiplication as the group operation,
and let $G_2 = \mathbb{Z}_2$. Then look at the multiplication tables
It is clear that these groups are "the same" except for the names of the elements. This is precisely the concept of isomorphism.

**Definition A.10:** Let \((G, *)\) and \((H, \circ)\) be groups. A homomorphism \(f : G \to H\) is a function for which

\[
f(a \ast b) = f(a) \circ f(b) \quad \forall a, b \in G
\]

An isomorphism is a homomorphism that is one-to-one and onto (\(f\) is one-to-one or injective if \(f(x) = f(y) \Rightarrow x = y\); \(f\) is onto or surjective if for any \(h \in H\) there exists at least one \(g \in G\) such that \(f(g) = h\); if \(f\) is both surjective and injective, it is called bijective). Two groups \(G\) and \(H\) are isomorphic, denoted \(G \cong H\), if there exists an isomorphism \(f : G \to H\). A homomorphism of a group into itself is called an endomorphism, and an isomorphism of a group into itself is an automorphism.

Note that it follows from the definition, that if \(f\) is a homomorphism, then

\[
f(e) = e \quad f(g^{-1}) = (f(g))^{-1}
\]

In the example given previously the function \(f : G_1 \to G_2\) defined by \(f(1) = 0\) and \(f(-1) = 1\) is an isomorphism. Note that if \(f : G \to H\) is an isomorphism, then \(f^{-1} : H \to G\) is also. Thus \(G \cong H \iff H \cong G\). Also, note that the notion of homomorphism is similar to the notion of linear transformation. In fact, suppose \(V_1\) and \(V_2\) are two vector spaces and
T : $V_1 \rightarrow V_2$ is a linear transformation. It is clear that $V_1$ and $V_2$ are groups under vector addition, and, since $T(a+b) = T(a) + T(b)$, $T$ is a group homomorphism.

**Definition A.11:** Let $A$ and $B$ be two subsets of a group $G$. We define the set $AB$ as follows:

$$AB \overset{\Delta}{=} \{ab | a \in A, b \in B\} \quad (A.5)$$

This set multiplication is obviously associative. If $g \in G$, we write

$$Ag = \{ag | a \in A\} \quad (A.6)$$

**Definition A.12:** A nonempty subset $S$ of a group $G$ is a **subgroup** of $G$ ($S < G$) if $S$ is a group under the binary operation of $G$. The subgroups $G$ and $\{e\}$ of $G$ are called **improper subgroups**. All other subgroups are **proper**. Note that if $S$ is a subgroup of $G$, then

$$\{e\} < S < G \quad (A.7)$$

Note that the concept of a subgroup of a group is analogous to the concept of a subspace of a vector space. The following results are proved in [R1].

**Theorem A.1:** Let $S$ be a subset of a group $G$. Then $S < G$ iff

1. $e \in S$
2. $a \in S \implies a^{-1} \in S$
3. $a, b \in S \implies ab \in S$

or, equivalently, iff $S \neq \emptyset$ and $a, b \in S \implies ab^{-1} \in S$.

**Theorem A.2:** The intersection of any family of subgroups of $G$ is itself a subgroup of $G$. 

•
Theorem A.3: If \( S \subseteq G \), then there is a smallest subgroup \( \langle S \rangle \) that contains \( S \) — i.e., if \( S \subseteq H \subseteq G \), then \( \langle S \rangle \subseteq H \). The subgroup \( \langle S \rangle \), which is given by (A.1), is called the subgroup generated by \( S \).

If \( S = \{g\} \), we write \( \langle g \rangle \) for \( \langle \{g\} \rangle \). Clearly this is a cyclic subgroup (see Def. A.8).

Definition A.13: Let \( S < G \). A right coset of \( S \) in \( G \) is any subset \( St \) (a left coset is \( tS \)), \( t \in G \). We call \( t \) a representative of \( St \) (also of \( tS \)).

Note that the cancellation laws imply that \( S \) and \( St \) have the same number of elements.

The following results are proved in [R1] (analogous results for left cosets can also be proved).

Theorem A.4: Let \( S < G \). Then \( Sa = Sb \iff ab^{-1} \in S \). It then follows that any two right cosets are either disjoint or identical (i.e., \( S \) induces a partition of \( G \)).

Definition A.14: Let \( S < G \). The index of \( S \) in \( G \), denoted \([G:S]\) is the number of distinct right cosets (which equals the number of distinct left cosets — see [R1]) of \( S \) in \( G \).

Definition A.15: The order of \( G \), denoted \(|G|\), is the number of elements in \( G \). \( G \) is called finite if \(|G| < \infty \).

Note that

\[
|G| = [G : \{e\}] \tag{A.8}
\]

where \( \{e\} \) is the trivial subgroup consisting of the identity alone.

The following result is extremely important, since it tells us how to determine all the possible sizes of subgroups from knowledge of the order of \( G \).
Theorem A.5 (Lagrange): If $|G| = n$ and $S \leq G$, then
\[ [G:S] = \frac{|G|}{|S|} \]
(A.9)

(so $|S|$ divides $|G|$, written $|S||G|$).

Proof: The right cosets of $S$ partition $G$ into $[G:S]$ pieces, each of which has precisely $|S|$ elements.

Corollary: If $|G| = p$, where $p$ is a prime, then $G$ is cyclic.

Proof: Choose $a \in G$, $a \neq e$. Then $\langle a \rangle$ is a subgroup of $G$ and $\langle a \rangle \neq \{e\}$. By Theorem A.5, $|\langle a \rangle| = p$, but only 1 and $p$ divide $p$.

Thus $|\langle a \rangle| = p$, or $\langle a \rangle = G$.

Definition A.16: If $a \in G$, the order of $a$ is $|\langle a \rangle|$, which is clearly the smallest positive integer for which $a^k = e$.

Definition A.17: A subgroup $S$ of $G$ is a normal subgroup of $G$, denoted $S \triangleleft G$, if $aS^{-1} = S \forall a \in G$ -- i.e., if $asa^{-1} \in S \forall a \in G, s \in S$.

This rather innocuous-looking definition is one of the most crucial concepts in group theory. First note that for abelian groups
\[ aS^{-1} = aa^{-1}S = S \]
(A.10)
so all subgroups of abelian groups are normal. Also, Definition A.17 is equivalent to the following: every right coset of $S$ in $G$ is also a left coset of $S$ in $G$. Finally the intersection of normal subgroups in $G$ is itself a normal subgroup, so any set $S \subset G$ is contained in a smallest normal subgroup, called the normal subgroup generated by $S$.

The dihedral group $D_4$ contains subgroups which are normal and some which aren't. Recall that
\[ D_4 = \{e, x, x^2, x^3, y, xy, x^2y, x^3y\} \]
(A.11)
with multiplication defined by the rules
\[ x^4 = e \quad y^2 = e \quad yxy = x^3 \] (A.12)

Then
\[ \{e,x^2,y,x^2y\} \triangleleft D_4 \] (A.13)

but
\[ \{e,x^2y\} \not\triangleleft D_4 \] (A.14)

since
\[ x(x^2y)x^{-1} = x^3yx^3 = y \notin \{e,x^2y\} \] (A.15)

However
\[ \{e,x^2y\} \triangleleft \{e,x^2,y,x^2y\} \] (A.16)

Thus, if \( S_1 \) and \( S_2 \) are subgroups of \( G \) with \( S_1 \triangleleft S_2 \) and \( S_2 \triangleleft G \), we cannot conclude that \( S_1 \triangleleft G \).

If \( S \triangleleft G \), we can define coset multiplication. That is, we claim that
\[ (aS)(bS) = abS \] (A.17)

Let \( aS_1 \in aS \), \( bS_2 \in bS \). Since \( S \triangleleft G \),
\[ b^{-1}aS_1b = S_3 \in S \] (A.18)

or
\[ s_1b = bs_3 \] (A.19)

so
\[ aS_1bS_2 = abs_3S_2 \in abS \] (A.20)

Thus
\[ (aS)(bS) \subset abS \] (A.21)

and the inclusion the other way can be proven in a similar manner.
(also \(SaSb = Sab\), since \(aS = Sa\)). Thus, we have

\[
(aH) \cdot (eH) = (eH) \cdot (aH) = aH
\]

(A.22)

\[
(aH) \cdot (a^{-1}H) = (a^{-1}H) \cdot (aH) = eH
\]

(A.23)

and we have proven

Theorem A.6: If \(S \vartriangleleft G\), then the cosets of \(S\) in \(G\) form a group under set multiplication defined by (A.5). This group, denoted \(G/S\) (read "\(G\) modulo \(S\)" or "the quotient group of \(G\) with respect to \(S\)"), has order \([G:S]\).

Note that we do not have to use the words "left" or "right" before the word "coset" if we are dealing with normal subgroups.

Definition A.18: Let \(f : G \rightarrow H\) be a homomorphism. The range or image of \(f\) is the subset of \(H\)

\[
Ra f \equiv \text{Im } f \equiv f(G) \triangleq \{ f(g) \mid g \in G \}
\]

(A.24)

The kernel of \(f\) is the subset of \(G\)

\[
\ker f \triangleq \{ g \in G \mid f(g) = e \}
\]

(A.25)

(here we use \(e\) to denote the identities of both \(G\) and \(H\)).

Theorem A.7 (First Isomorphism Theorem): Let \(f : G \rightarrow H\) be a homomorphism with \(K = \ker f\). Then \(K\) is a normal subgroup of \(G\), \(f(G)\) is a (not necessarily normal) subgroup of \(H\), and

\[
G/K = f(G)
\]

(A.26)

Proof: Since \(f(e) = e\), \(K \neq \phi\) and \(e \in f(G)\). First, if \(a, b \in K\), then \(f(ab^{-1}) = f(a)f(b)^{-1} = e\), so \(ab^{-1} \in K\), and, by Theorem A.1, \(K\) is a subgroup. Also, if \(a \in K\), then for any \(g \in G\), \(f(gag^{-1}) = f(g)ef(g)^{-1} = e\), so \(K \triangleleft G\). To show \(f(G)\) is a subgroup, we need only note that
\[ f(a)f(b)^{-1} = f(ab^{-1}), \text{ so Theorem A.1 applies again.} \]

To prove (A.26), define \( F : G/K \to f(G) \) by

\[ F(Ka) = f(a) \quad (A.27) \]

We must first show that \( F \) is well defined. If \( Ka = Kb \), then \( ab^{-1} \in K \), so \( f(ab^{-1}) = e \Rightarrow f(a) = f(b) \), and \( F \) is well defined. \( F \) is a homomorphism, since

\[ F(KaKb) = F(Kab) = f(ab) = f(a)f(b) = F(Ka)F(Kb) \quad (A.28) \]

Clearly \( F(G/K) = f(G) \), and it remains to show that \( F \) is one-to-one. If \( F(Ka) = F(Kb) \), then \( f(a) = f(b) \); hence \( f(ab^{-1}) = e \), \( ab^{-1} \in K \), so \( Ka = Kb \) by Theorem A.4.

We then have a factorization of the homomorphism

\[
\begin{array}{ccc}
G & \xrightarrow{f} & H \\
\downarrow{\pi} & & \downarrow{F} \\
G/K & & \\
\end{array}
\]

where \( F \) is one-to-one and \( \pi \), defined by \( \pi(x) = Kx \) and called the canonical homomorphism, is onto. Thus, it is clear that \( f \) is one-to-one, if and only if \( \ker f = \{e\} \).

A related result, stated here without proof, is: if \( f : G \to H \) is a homomorphism and \( S < H \), then \( K < f^{-1}(S) < G \).

**Theorem A.8 (Second Isomorphism Theorem):** Let \( S < G, T \triangleleft G \), then \( ST < G, S \cap T \triangleleft S \), and

\[ S/(S \cap T) \cong ST/T \quad (A.29) \]

**Proof:** See [RL] pp. 26-27.
Note that if $T$ is not normal, $ST$ need not be a subgroup. As an example, again consider the group $D_4$, and let

$$S = \{e, x^2y\} \quad T = \{e, xy\} \quad (A.30)$$

Then

$$ST = \{e, x^2y, xy, x\} \not\subseteq G \quad (A.31)$$

**Theorem A.9** (Third Isomorphism Theorem): Let $K < H < G$, where both $H$ and $K$ are normal in $G$. Then

$$H/K \cong G/K \quad (A.32)$$

and

$$(G/K)/(H/K) \cong G/H \quad (A.33)$$

**Proof:** Let $f : G/K \to G/H$ be defined by $f(Ka) = Ha$. It is easy to check that $f$ is a well-defined homomorphism whose kernel is $H/K$ and whose image is $G/H$. We then apply Theorem A.7.

**Definition A.19:** Let $G_1$ and $G_2$ be groups. The direct product group of $G_1$ and $G_2$, denoted $G_1 \times G_2$, consists of the set of all ordered pairs $(g_1, g_2)$; $g_1 \in G_1$, $g_2 \in G_2$, with multiplication defined by

$$(g_1, g_2)(h_1, h_2) = (g_1h_1, g_2h_2) \quad (A.34)$$

Let $G$ be a group, then $\text{Aut } G$ is the group of automorphisms of $G$ with function composition as the group operation (the composition of two homomorphisms is a homomorphism). Given groups $G_1$ and $G_2$ and a homomorphism $\theta : G_2 \to \text{Aut } G_1$, we define the semidirect product $G_1 \times_\theta G_2$ as follows: the set is the Cartesian product $G_1 \times G_2$ and group multiplication is

$$(g_1, g_2)(h_1, h_2) = (g_1\theta(g_2)(h_1), g_2h_2) \quad (A.35)$$
See [ML] for the proof that (A.35) does lead to a group structure on the set $G_1 \times G_2$. Note that $G_1$ and $G_2$ can be regarded as normal subgroups of $G_1 \times G_2$ via the identifications
\[ g_1 \leftrightarrow (g_1, e) \] \[ g_2 \leftrightarrow (e, g_2) \] (A.36)

A.3 Rings and Fields

In this section we present a few of the basic results from the theory of fields and vector spaces. These results are important in understanding the development of the theory of linear sequential circuits and convolutional encoders and also the relationship between the class of linear sequential circuits and the class of finite group homomorphic sequential systems (see Chapter 8).

**Definition A.20:** A ring $(R, +, \cdot)$ is a set $R$ together with two binary operations, $+$ and $\cdot$, called (ring) addition and (ring) multiplication, respectively, such that the following axioms are satisfied:

1. $(R, +)$ is an abelian group
2. Multiplication is associative
3. The left and right distributive laws
   \[ a(b + c) = ab + ac \]
   \[ (a + b)c = ac + bc \] (A.37)

hold for all $a, b, c \in R$.

We remark that we often refer to "the ring $R$" and we also often write $ab$ for $a \cdot b$. Note that multiplication need not be commutative.

For example the set of all $n \times n$ matrices with matrix addition and multi-
plication is a ring with noncommutative multiplication. Another example of a ring is \((\mathbb{Z}_n, +, \cdot)\), where \(+\) and \(\cdot\) are addition and multiplication modulo \(n\). In this case, ring multiplication is commutative.

We will refer to \((R, +)\) as the **additive group** of the ring \(R\). Also, we will write \(-a\) for the additive inverse of \(a \in R\) and \(0\) for the additive identity.

**Theorem A.10:** If \(R\) is a ring, then for any \(a, b \in R\)

1. \(0a = a0 = 0\)
2. \(a(-b) = (-a)b = -(ab)\) \hspace{1cm} \text{(A.38)}
3. \((-a)(-b) = ab\)

**Proof:** See [F1] p. 190.

**Definition A.21:** A function \(f : R_1 \rightarrow R_2\) from one ring into another is a **homomorphism** if

1. \(f(a+b) = f(a) + f(b)\)
2. \(f(ab) = f(a)f(b)\) \hspace{1cm} \text{(A.39)}

We remark that to be precise we should call these various functions **group homomorphisms** or **ring homomorphism** (in general, a **homomorphism** preserves all algebraic properties of \(\cdot\)'s), but we will call all such functions **homomorphisms** except when it is ambiguous to do so.

**Definition A.22:** A ring in which ring multiplication is commutative is a **commutative ring**. A ring \(R\) with a multiplicative identity \(1\) (\(1x = x1 = x\ \forall \ x \in R\)) is a **ring with unity**. An element \(u \in R\), a ring with unity, is a **unit** of \(R\) if it has a multiplicative inverse. If every nonzero element of \(R\) is a unit, \(R\) is a **division ring**. A **field** is
a commutative division ring.

We remark that multiplicative identities and inverses are unique. Also, one can define subrings, subfields, etc., in the same way subgroups were defined.

It is clear that the set $\mathbb{R}^1$ of real numbers is a field, as is the set $\mathbb{Q}$ of rational numbers. In addition, there are many finite fields (fields containing a finite number of elements), as the following result indicates.

**Theorem A.11:** A finite field $\text{GF}(p^n)$ of $p^n$ elements exists for every prime $p$ and every integer $n > 1$.

**Proof:** See [F1], p. 367.

We remark that "GF" stands for Galois Field, and for $n = 1$ the Galois field $\text{GF}(p)$ is just $\mathbb{Z}_p$. However this is not the case for $n > 1$. For example, consider the ring $\mathbb{Z}_4$. This cannot be a field, since $2 \cdot 2 = 0$, so 2 has no inverse. Thus $\mathbb{Z}_4 \neq \text{GF}(4)$. In fact the additive group of $\text{GF}(4)$ is $\mathbb{Z}_2 \times \mathbb{Z}_2$ and the multiplication table is

<table>
<thead>
<tr>
<th></th>
<th>(0,0)</th>
<th>(0,1)</th>
<th>(1,0)</th>
<th>(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>(0,0)</td>
<td>(0,0)</td>
<td>(0,0)</td>
<td>(0,0)</td>
</tr>
<tr>
<td>(0,1)</td>
<td>(0,0)</td>
<td>(0,1)</td>
<td>(1,0)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>(1,0)</td>
<td>(0,0)</td>
<td>(1,0)</td>
<td>(1,1)</td>
<td>(0,1)</td>
</tr>
<tr>
<td>(1,1)</td>
<td>(0,0)</td>
<td>(1,1)</td>
<td>(0,1)</td>
<td>(1,0)</td>
</tr>
</tbody>
</table>

**Definition A.23:** A vector space $V$ over the field $F$ is an abelian group together with an operation of scalar multiplication of elements of $V$ by elements of $F$ on the left, satisfying the following for all $\alpha, \beta \in F$ and $v_1, v_2 \in V$.
(1) \( \alpha v_1 \in V \)

(2) \( \alpha(\beta v_1) = (\alpha \beta) v_1 \)

(3) \( (\alpha + \beta)v_1 = \alpha v_1 + \beta v_1 \) \hspace{1cm} (A.40)

(4) \( \alpha(v_1 + v_2) = \alpha v_1 + \alpha v_2 \)

(5) \( lv_1 = v_1 \)

where 1 is the multiplicative identity of \( F \). The elements of \( V \) are vectors and the elements of \( F \) are scalars.

Finally, we make the remark that a vector space homomorphism — i.e. a map \( f : V \rightarrow V' \) that is a group homomorphism of the abelian groups \( V \) and \( V' \) and that also preserves scalar multiplication (\( F(\alpha v) = \alpha F(v) \)) — is obviously the same thing as a linear transformation.

The theory of linear algebra applies to vector spaces over arbitrary fields. For properties of vector spaces, see Halmos [H1] or Hoffman and Kunze [H2].

A.4 Lie Groups and Lie Algebras

The theory of Lie groups and Lie algebras relies heavily on group theory and on the theory of differentiable manifolds. However, for our purposes, we do not need much of this theory and instead will present some specialized definitions for matrix Lie Groups. We will follow the work of Brockett [B1].

Let \( M(n, \mathbb{R}^1) \) be the set of real-valued nxn matrices. This is a real vector space of dimension \( n^2 \).

**Definition A.24:** A (matrix) Lie algebra \( L \) in \( M(n, \mathbb{R}^1) \) is a subspace
of $M(n, \mathbb{R}^1)$ with the additional property that if $A$ and $B$ are in $L$, so is their \textbf{commutator product}, $[A, B] = AB - BA$. A Lie algebra $L$ is \textbf{abelian} if $[A, B] = 0 \ \forall \ A, B \in L$.

Note that the \textbf{commutator product} is \textbf{skew-symmetric} ($[A, B] = -[B, A]$) and that it satisfies the \textbf{Jacobi identity}

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0 \quad (A.41)$$

Also, the intersection of Lie algebras is again a Lie algebra, but the union, direct sum, and commutator of two Lie algebras need not be Lie algebras.

\textbf{Definition A.25}: Let $S$ be any subset of $M(n, \mathbb{R}^1)$. The \textbf{Lie algebra generated by} $S$, denoted $\{S\}_A$, is the smallest Lie algebra containing $S$.

As before, we can define the concept of Lie subalgebra, and we note that all Lie algebras are Lie subalgebras of $M(n, \mathbb{R}^1)$.

\textbf{Definition A.26}: A \textbf{matrix group} is a subset of $M(n, \mathbb{R}^1)$ that is a group under matrix multiplication. If $T$ is a set of nonsingular $n \times n$ matrices, we write $\langle T \rangle_G$ for $\langle T \rangle$, the group generated by $T$. Let $\exp : M(n, \mathbb{R}^1) \to M(n, \mathbb{R}^1)$ be the matrix exponential map

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!} \quad (A.42)$$

A matrix group $G$ is a \textbf{matrix Lie group} if for some Lie algebra $L$

$$G = \{\exp(L)\}_G \quad (A.43)$$

We have used the fact that $\exp(A)$ is nonsingular for all $A \in M(n, \mathbb{R}^1)$ ($\det(\exp A) = e^{\text{tr}(A)}$). Also, it is simple to check that $\{\exp L\}_G$ is abelian $\iff$ $L$ is abelian.
It is appropriate to comment on the topological structure of Lie groups. Let \( L \) be a Lie algebra and take any \( M \in \{ \exp L \}_G \).

Define the map \( \phi_M : L \to \{ \exp L \}_G \)

\[ \phi_M(A) = (\exp A)M \] \hspace{1cm} (A.44)

As noted in [Bl], this map is one-to-one from a neighborhood of 0 in \( L \) onto a neighborhood of \( M \) in \( \{ \exp L \}_G \), and it has a smooth inverse.

In addition, the set of maps \( \phi_M^{-1} \) form a differentiable structure of class \( C^\infty \) on \( \{ \exp L \}_G \) (see [W1], [S2], or [Cl]). These facts imply that \( \{ \exp L \}_G \) is a locally Euclidean space of dimension equal to the dimension of \( L \), and the differentiable structure makes \( \{ \exp L \}_G \) a \( C^\infty \) manifold (here \( C^\infty \) means infinitely differentiable and refers to the fact that, restricted to the neighborhoods on which they are one-to-one, \( \phi_M^{-1} \circ \phi_{M_2} \) is \( C^\infty \) \( \forall M_1, M_2 \in \{ \exp L \}_G \)).

We make one comment on the relationship between this theory and the theory of bilinear control systems. Consider the differential equation

\[ \dot{X}(t) = \left( \sum_{i=1}^{n} u_i(t)A_i \right)X(t) \] \hspace{1cm} (A.45)

where the \( u_i \) are scalar functions of time and the \( A_i \) and \( X(t) \) are \( nxn \) matrices. As proven in [Bl], the set of states reachable from \( X(0) = X \in \mathbb{M}(n, \mathbb{R}^1) \), is just \( \{ \exp \{ \mathbb{R}C_1, C_2, \ldots, C_m \} \}_A \). \( C^\infty \) \( \mathbb{M}(n, \mathbb{R}^1) \).

We will now mention a few examples of matrix Lie groups and their associated Lie algebras. The general linear group \( \mathbb{G}\mathbb{L}(n, \mathbb{R}^1) \) \( \Delta \) \( \{ \exp (\mathbb{M}(n, \mathbb{R}^1)) \}_G \) is the group of all nonsingular \( nxn \) matrices. The special linear group \( \mathbb{S}\mathbb{L}(n, \mathbb{R}^1) \) and its Lie algebra \( \mathbb{S}\mathbb{L}(n, \mathbb{R}^1) \) are defined by
\[ \text{SL}(n, R^1) \triangleq \{ A \in \text{GL}(n, R^1) | \det A = 1 \} \]  
(A.46)

\[ \text{sl}(n, R^1) \triangleq \{ A \in \text{M}(n, R^1) | \text{tr} A = 0 \} \]  
(A.47)

Similarly, the orthogonal group \( \text{O}(n, R^1) \) and its algebra \( \text{o}(n, R^1) \) are

\[ \text{O}(n, R^1) \triangleq \{ A \in \text{GL}(n, R^1) | A' = A^{-1} \} \]  
(A.48)

\[ \text{o}(n, R^1) \triangleq \{ A \in \text{M}(n, R^1) | A' = -A \} \]  
(A.49)

The special orthogonal group \( \text{SO}(n) \) is the intersection of \( \text{O}(n, R^1) \) and \( \text{SL}(n, R^1) \) and the Lie algebra \( \text{so}(n) \) is the corresponding intersection of Lie algebras.

Of particular importance in this manuscript are \( \text{SO}(3) \) and \( \text{SO}(2) \).

The group \( \text{SO}(3) \) is the group of all rigid body rotations in \( R^3 \) (i.e. it is the set of all 3x3 direction cosine matrices \([W2]\)). The associated Lie algebra has dimension 3 and the three "infinitesimal rotations" (see Chapter 2)

\[
R_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \quad R_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad R_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \]  
(A.50)

form a basis for \( \text{so}(3) \). The group \( \text{SO}(2) \) is the group of all rotations about one fixed axis, and its one dimensional Lie algebra has the infinitesimal rotation

\[
R = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \]  
(A.51)

for a basis. Identifying \( R \) and \( R_2 \), we see that \( \text{SO}(2) \) is isomorphic to \( \{ \exp R_2 \}_G < \text{SO}(3) \). Note that \( \text{SO}(2) \) is abelian (as are all one dimensional Lie groups), but \( \text{SO}(3) \) is not.
Finally, as noted earlier, Lie groups need not be matrix groups. For example, there are many other representations of \( SO(2) \). We note that the general element of \( SO(2) \) is of the form

\[
\begin{bmatrix}
\cos \theta & \sin \theta \\
-\sin \theta & \cos \theta
\end{bmatrix}
\]

and thus there is a one-to-one correspondence between points on the unit circle, \( S^1 \), in \( \mathbb{R}^2 \), and elements of \( SO(2) \), and if we defined addition of angles by

\[
\theta + \phi = (\theta + \phi) \mod 2\pi
\]

the group \( S^1 \) is isomorphic to \( SO(2) \), and we will refer to them interchangeably. Furthermore, both groups are isomorphic to the set of complex numbers of unit modulus with complex multiplication as the group operation. See Chapter 2 for more detailed descriptions of the various representations of the circle.
APPENDIX B

SOME TECHNICAL RESULTS RELATED TO
ESTIMATION ON THE CIRCLE

"Tedious it were to tell, and harsh to hear."

- William Shakespeare [S16]

B.1 Introduction

In this appendix, we present some of the technical results used and referred to in Chapter 4. Section B.2 contains some results on the properties of functions on $\mathbb{R}^1$ and their projections onto $S^1$. In Section B.3 we develop rigorously the discrete-time conditional density equations used in Section 4.4. Section B.4 contains a result relating the continuous-time and discrete-time problems.

B.2 Differentiable Functions on $\mathbb{R}^1$ and Their Projections Onto $S^1$

The motivation behind this section is to define conditions under which equation (4.44) holds. That is, given a function $f : \mathbb{R}^1 \to \mathbb{R}^1$ and its $S^1$ projection

$$g(\theta) = \sum_{n=-\infty}^{+\infty} f(\theta + 2n\pi)$$

(B.1)

when can one write

$$\frac{\partial^2 g}{\partial \theta^2} (\theta) = \sum_{n=-\infty}^{+\infty} \frac{\partial^2 f}{\partial \theta^2} (\theta + 2n\pi)$$

(B.2)

We will prove the following result.

**Theorem B.1:** Let $f : \mathbb{R}^1 \to \mathbb{R}^1$ be a twice continuously differentiable
function such that

\[
\max(\lim_{x \to +\infty} |x^s f(x)|, \lim_{x \to -\infty} |x^s f(x)|) < r < \infty
\]

(B.4)

\[
\max(\lim_{x \to +\infty} |x^s f'(x)|, \lim_{x \to -\infty} |x^s f'(x)|) < r < \infty
\]

(B.5)

\[
\max(\lim_{x \to +\infty} |x^s f''(x)|, \lim_{x \to -\infty} |x^s f''(x)|) < r < \infty
\]

(B.6)

for some \( s > 1 \), where \( \cdot = \frac{d}{dx} \).

Then \( g : [-\pi, \pi] \to \mathbb{R} \), defined by (B.1), is twice continuously differentiable, and (B.2) and (B.3) hold.

Proof: We prove this with the aid of the next two lemmas.

**Lemma B.1:** Let \( f : \mathbb{R} \to \mathbb{R} \) be continuous and satisfy (B.4). Then the functions \( g_n : [-\pi, \pi] \to \mathbb{R} \) defined by

\[
g_n(\theta) = \sum_{k=-n}^{+n} f(\theta + 2k\pi)
\]

(B.7)

converge uniformly to \( g \), defined by (B.1), which therefore is (uniformly) continuous.

Proof: Given \( \varepsilon > 0 \), there exists an \( M > 0 \) such that

\[
|f(x)| < \frac{\varepsilon + r}{|x|^s} \quad \forall \ |x| > M
\]

(B.8)

Choose an integer \( N_1 \) so that \( 2N_1 \pi > M \). Then, given \( n > m > N_1 \)

\[
|g_n(\theta) - g_m(\theta)| < \sum_{|k| > m} |f(\theta + 2k\pi)|
\]

\[
\leq \sum_{|k| > m} \frac{\varepsilon + r}{|\theta + 2k\pi|^s}
\]

\[
\leq \sum_{|k| > m} \frac{\varepsilon + r}{|(2k-1)\pi|^s}
\]

(B.9)
The series
\[ \sum_{k=-\infty}^{\infty} \frac{\varepsilon + r}{|k(2k-1)\pi|^s} < \varepsilon \]
converges for \( s > 1 \), and therefore we can choose \( N_2 \gg N_1 \) such that
\[ \sum_{|k|>N_2} \frac{\varepsilon}{|k(2k-1)\pi|^s} < \varepsilon \] (B.10)

Then for \( n > m \gg N_2 \)
\[ |g_n(\theta) - g_m(\theta)| < \varepsilon \quad \forall \theta \in [-\pi, \pi] \] (B.11)

Thus the sequence of continuous functions \( \{g_n\} \) converges uniformly to \( g \), which therefore is uniformly continuous on \([-\pi, \pi]\).

**Lemma B.2:** Let \( g_n \) be a sequence of differentiable functions from \([-\pi, \pi]\) to \( \mathbb{R}^1 \). Suppose that

(1) \( g_n \) converges to \( g \)

(2) the derivatives \( g_n' \) converge uniformly.

Then \( g \) is differentiable and

\[ g' = \lim_{n \to \infty} g_n' \] (B.12)

**Proof:** See Hoffman [H8, p. 272].

The theorem now follows. First, let \( g_n \) in Lemma B.2 be given by (B.7) and \( g \) by (B.1). Then hypothesis (1) of Lemma B.2 is clear and hypothesis (2) follows from Lemma B.1 by replacing \( f \) by \( f' \), \( g_n \) by \( g_n' \) and by using (B.5) instead of (B.4). We then let \( g_n \) in Lemma B.2 be replaced by \( g_n' \), and, using (B.6), we obtain

\[ g'(\theta) = \lim_{n \to \infty} g_n'(\theta) = \lim_{n \to \infty} \sum_{k=-n}^{n} f'(\varepsilon + 2n\pi) \] (B.13)
\[ g''(\theta) = \lim_{n \to \infty} g''_n(\theta) = \lim_{n \to \infty} \sum_{k=-n}^{+n} f''(\theta+2n\pi) \]  \hspace{1cm} (B.14)

Continuity of \( g' \) and \( g'' \) follow, since they are the uniform limits of continuous functions.

We remark that since we have

\[ \lim_{x \to \infty} x^n e^{-x^2} = 0 \hspace{1cm} \forall \ n \]  \hspace{1cm} (B.15)

Theorem B.1 holds for \( f \) a normal density and \( g \) the corresponding folded normal.

**Example B.1:** This example illustrates that if the hypotheses of Lemma B.1 aren't satisfied, the function \( g \) need not be continuous.

Let \( f : \mathbb{R}^1 \to \mathbb{R}^1 \) be defined on the interval \([(2n-1)\pi, (2n+1)\pi]\) by

\[ f(x) = \begin{cases} 
  0 & x \in [(2n-1)\pi, 2n\pi - \frac{1}{2(|n|+1)}] \\
  2x + (\frac{1}{|n|+1} - 4n\pi) & x \in [2n\pi - \frac{1}{2(|n|+1)}, 2n\pi] \\
  -2x + (\frac{1}{|n|+1} + 4n\pi) & x \in [2n\pi, 2n\pi + \frac{1}{2(|n|+1)}] \\
  0 & x \in [2n\pi + \frac{1}{2(|n|+1)}, (2n+1)\pi] 
\end{cases} \]  \hspace{1cm} (B.16)

The graph of \( f \) is illustrated in Figure B.1. On the interval

![Figure B.1: Graph of the Function of Example B.1](image-url)
[(2n-1), (2n+1)n] the graph of f is zero except for a small isosceles triangle of base and height $\frac{1}{|n|+1}$, with the slopes of the sides equal to ±2. Thus f is Lipschitz continuous, and

$$\lim_{x \to \pm \infty} |x^s f(x)| < 1 \quad \forall \ s < 1 \quad \text{(B.17)}$$

However f goes to zero only as $\frac{1}{x}$, and, in fact, we have

$$g(0) = \sum_{n=-\infty}^{\infty} f(2n\pi) = \sum_{n=-\infty}^{\infty} \frac{1}{|n|+1} = \infty \quad \text{(B.18)}$$

B.3 A Single Stage Estimation Problem

In this section we apply the Radon-Nikodym Theorem to the single stage estimation problem that is discussed in Section 4.4. Consider the probability space $(\mathbb{R}_1^+, \mathcal{B}_1, \mathbb{P}_\gamma)$ where $\mathcal{B}_1$ is the Borel $\sigma$-algebra on $\mathbb{R}_1$ and $\mathbb{P}_\gamma$ is any probability measure on $\mathcal{B}_1$. We define two random variables

$$\gamma(\omega) = \omega \quad \text{(B.19)}$$

$$\tilde{\gamma}(\omega) = \omega \mod 2^+ \quad \tilde{\gamma} \in [-\pi, \pi) \quad \text{(B.20)}$$

where $\omega \in \mathbb{R}_1$. Then $\gamma$, the $\sigma$-algebra generated by $\gamma$, is $\mathcal{B}_1$, and

$\gamma_{\tilde{\gamma}}$, the $\sigma$-algebra generated by $\tilde{\gamma}$, is the sub $\sigma$-algebra of $\gamma$ consisting of the following sets

$$\tilde{A} \in \gamma_{\tilde{\gamma}} \iff \tilde{A} \cap [-\pi, \pi) \subseteq A \in \gamma \quad \text{and} \quad \tilde{A} = \bigcup_{n=-\infty}^{\infty} (A+2n\pi) \quad \text{(B.21)}$$

i.e. $\tilde{A}$ is "periodic" in form and is determined by $A$.

We define a sequence of measures on the measurable space $([-\pi, \pi), \mathcal{B}_2)$, where $\mathcal{B}_2$ is the Borel $\sigma$-algebra on $[-\pi, \pi)$:
\[ P_y^n(B) = P_y(B + 2n\pi), \quad n = 0, \pm 1, \pm 2, \ldots \quad \text{(B.22)} \]

\[ \tilde{P}_y(B) = \sum_{n=-\infty}^{\infty} \frac{dP^n_y}{d\tilde{P}_y}(B) \quad \text{(B.23)} \]

where \( B \in \mathcal{B}_2 \). Note that \( \tilde{P}_y \) is the probability distribution for \( \tilde{y} \).

Clearly \( \tilde{P}_y(B) = 0 \Rightarrow P_y^n(B) = 0 \), which means \( P_y^n \) is absolutely continuous with respect to \( \tilde{P}_y \), \( P_y^n \ll \tilde{P}_y \), and thus, by the Radon-Nikodym Theorem [H5], [R3], for each \( n \) there exists a \( \mathcal{B}_2 \)-measurable function \( \frac{dP^n_y}{d\tilde{P}_y}(\omega) \) on \([-\pi, \pi)\), such that

\[ P_y^n(B) = \int_B \frac{dP^n_y}{d\tilde{P}_y}(\omega) d\tilde{P}_y(\omega) \quad \forall B \in \mathcal{B}_2 \quad \text{(B.24)} \]

We wish to compute the conditional probability measure \( P_{y|\tilde{y}} \) for \( y \) given \( \tilde{y} \). The following theorem shows that this conditional measure can be expressed in a form that reflects our uncertainty as to the number of integral multiples of \( 2\pi \) that separate the values of \( y \) and \( \tilde{y} \).

**Theorem B.2:** The conditional probability measure

\[ P_{y|\tilde{y}}(C|\beta) = P_r(y \in C|\tilde{y} = \beta) \quad \beta \in [-\pi, \pi), \, C \in \mathcal{B}_1 \quad \text{(B.25)} \]

can be expressed in the form

\[ P_{y|\tilde{y}}(C|\beta) = \sum_{n=-\infty}^{\infty} \chi_C(\beta + 2n\pi) \frac{dP^n_y}{d\tilde{P}_y}(\beta) \quad \text{(B.26)} \]

where \( \chi_C \) is the indicator function for \( C \) (i.e. \( \chi_C(x) = 1 \) if \( x \in C \) and \( \chi_C(x) = 0 \) if \( x \notin C \)).

**Proof:** From the definition of conditional expectation [D2], [F3], [N1], we have that \( P_{y|\tilde{y}}(C|\beta) \) is the unique (up to \( \tilde{P}_y \)-measurability) \( \tilde{\mathcal{P}} \)-measurable function such that for any \( A \in \mathcal{A}_{\tilde{y}} \)
\[
\int_{A} \chi_{C}(\omega) dP_{y}(\omega) = \int_{A} P_{y} |\tilde{\gamma}(C|\tilde{y})(\omega)) dP_{y}(\omega) \quad (B.27)
\]

The left-hand side of (B.27) equals

\[
\sum_{n=-\infty}^{+\infty} \int_{A+2n\pi} \chi_{C}(\omega) dP_{y}(\omega) = \sum_{n=-\infty}^{+\infty} \int_{A} \chi_{C}(\xi+2n\pi) dP_{y}^{n}(\xi) \quad (B.28)
\]

where \( A = \tilde{A} \cap [-\pi, \pi) \). This last equality follows from the definition of \( P_{y}^{n} \) and the obvious relationship between \( \mathcal{B}_{1} \) - and \( \mathcal{B}_{2} \)-measurability.

Now \( P_{y}^{n} \ll \tilde{P}_{y} \), so

\[
\sum_{n=-\infty}^{+\infty} \int_{A} \chi_{C}(\xi+2n\pi) dP_{y}^{n}(\xi) = \sum_{n=-\infty}^{+\infty} \int_{A} \chi_{C}(\xi+2n\pi) \frac{dP_{y}^{n}}{d\tilde{P}_{y}}(\xi) d\tilde{P}_{y}(\xi)
\]

\[
= \int_{A} \left[ \sum_{n=-\infty}^{+\infty} \chi_{C}(\xi+2n\pi) \frac{dP_{y}^{n}}{d\tilde{P}_{y}}(\xi) \right] d\tilde{P}_{y}(\xi) \quad (B.29)
\]

where we have used the Radon-Nikodym Theorem, the fact that \( dP_{y}^{n}/d\tilde{P}_{y} > 0 \) a.e. \( (\tilde{P}_{y}) \), and the monotone convergence theorem, [R3].

Similarly, the right-hand side of (B.27) is equal to

\[
\sum_{n=-\infty}^{+\infty} \int_{A+2n\pi} P_{y} |\tilde{\gamma}(C|\tilde{y})(\omega)) dP_{y}(\omega) = \sum_{n=-\infty}^{+\infty} \int_{A} P_{y} |\tilde{\gamma}(C|\tilde{y})(\xi)) dP_{y}^{n}(\xi) \quad (B.30)
\]

where we have used the periodicity of \( \tilde{\gamma}(\omega) \). Again using the Radon-Nikodym and monotone convergence theorems, we see that the expression on the right-hand side of (B.30) is equal to

\[
\sum_{n=-\infty}^{+\infty} \int_{A} P_{y} |\tilde{\gamma}(C|\tilde{y})(\xi)) \frac{dP_{y}^{n}}{d\tilde{P}_{y}}(\xi) d\tilde{P}_{y}(\xi)
\]

\[
= \int_{A} P_{y} |\tilde{\gamma}(C|\tilde{y})(\xi)) \left[ \sum_{n=-\infty}^{+\infty} \frac{dP_{y}^{n}}{d\tilde{P}_{y}}(\xi) \right] d\tilde{P}_{y}(\xi) \quad (B.31)
\]
Clearly \( \tilde{P}_y \) is a probability measure on \([-\pi, \pi]\), since

\[
\tilde{P}_y([-\pi, \pi]) = \sum_{n=-\infty}^{+\infty} P_y([-\pi, \pi]) \\
= \sum_{n=-\infty}^{+\infty} P_y([-\pi, \pi] + 2n\pi) = P_y((-\infty, \infty)) = 1
\]  

(B.32)

Also, on any \( B \) measurable set \( B, \)

\[
\tilde{P}_y(B) = \sum_{n=-\infty}^{+\infty} P^n_y(B) = \sum_{n=-\infty}^{+\infty} \int_B \frac{dP^n_y}{dP_y}(\xi) d\tilde{P}_y(\xi) \\
= \int_B \left[ \sum_{n=-\infty}^{+\infty} \frac{dP^n_y}{dP_y}(\xi) \right] d\tilde{P}_y(\xi)
\]  

(B.33)

and, since \( \tilde{P}_y \) is a finite measure, this implies that

\[
\sum_{n=-\infty}^{+\infty} \frac{dP^n_y}{dP_y}(\xi) = 1 \quad \text{a.e. \( (P_y) \)}
\]  

(B.34)

Thus, the right-hand side of (B.27) is

\[
\int_A P_y|\tilde{y}(C|\tilde{y}(\xi)) d\tilde{P}_y(\xi)
\]  

(B.35)

Since \( \tilde{P}_y \) is the probability distribution for \( \tilde{y}, \) (B.35) should be viewed as "integrating out" the \( \tilde{y} \)-dependence in the joint distribution for \( y \) and \( \tilde{y} \) to obtain the marginal distribution for \( v. \) In fact, from (B.27) we see that (B.35) equals \( P_v(C \cap \tilde{A}). \)

Comparing (B.29) and (B.35), we see that the conditional measure is given by

\[
P_y|\tilde{y}(C|\tilde{y}(\xi)) = \sum_{n=-\infty}^{+\infty} \chi_{C}(\xi+2n\pi) \frac{dP^n_y}{dP_y}(\xi)
\]  

(B.36)
But this is defined for $\xi \in [-\pi, \pi]$, and in this case $\tilde{y}(\xi) = \xi$. Thus
\[ P_{y|\tilde{y}}(C|\beta) = \sum_{n=-\infty}^{+\infty} \chi_C(\beta+2n\pi) \frac{dP^n}{dP_y}(\beta) \quad (B.37) \]

We note that for fixed $\beta$, $P_{y|\tilde{y}}(C|\beta)$ is a sum of Dirac measures concentrated at the points $\beta+2n\pi$, where
\[ P_{\tau}(y = \beta+2n\pi|\tilde{y} = \beta) = \frac{dP^n}{dP_y}(\beta) \quad (B.38) \]

Thus, in terms of $\delta$-functions, we can write the conditional "density"
\[ P_{v|\tilde{y}}(\xi|\beta) = \sum_{n=-\infty}^{+\infty} \frac{dP^n}{dP_v}(\beta) \delta(\xi-\beta-2n\pi) \quad (B.39) \]

**Corollary:** Suppose $P_y$ is absolutely continuous with respect to Lebesgue measure $\lambda$:
\[ P_y(A) = \int_A P_v(\xi)d\lambda(\xi) \quad A \in \mathcal{B}_1 \quad (B.40) \]

Then the conditional "density" is given
\[ P_{v|\tilde{y}}(\xi|\beta) = \sum_{n=-\infty}^{+\infty} \frac{dP^n(\xi+2n\pi)}{dP_v(\beta+2n\pi)} \delta(\xi-\beta-2n\pi) \quad (B.41) \]
\[ = \frac{dP_v(\xi)}{dP_v(\beta+2k\pi)} \delta(\xi-(\xi \text{mod} 2\pi)) \quad (B.42) \]

**Proof:** It is easy to see that $P^n_y$ is absolutely continuous with respect to Lebesgue measure (also called $\lambda$) on $(-\pi, \pi)$, and
\[ \frac{dP^n_y}{d\lambda}(\eta) = P_y(\eta+2n\pi) \quad (B.43) \]
is a version of the Radon-Nikodym derivative. Clearly \( p_y \geq 0 \)
a.e. (\( \lambda \)), and thus, by monotone convergence and the finiteness of \( \tilde{P}_y \),
\[
\sum_{k=-\infty}^{+\infty} p_y(\eta+2k\pi) < \infty \quad \text{a.e. } (\lambda) \tag{B.44}
\]
and, in fact, it is the Radon-Nikodym derivative \( d\tilde{P}_y/d\lambda \). It is
then clear that
\[
\frac{dP_y^n}{d\tilde{P}_y}(\eta) = \frac{dP_y^n/d\lambda(\eta)}{d\tilde{P}_y/d\lambda(\eta)} = \frac{p_y(\eta+2n\pi)}{\sum_{k=-\infty}^{+\infty} p_y(\eta+2k\pi)} \tag{B.45}
\]
Finally, consider the set where \( p_y|\tilde{y}(\xi|\beta) \) is undefined -- i.e. where
\[
\sum_{k=-\infty}^{+\infty} p_y(\beta+2k\pi) = 0 \tag{B.46}
\]
But this set is a set of \( \tilde{P}_y \)-measure zero. Equation (B.42) follows
immediately from (B.41) and the properties of the \( \delta \)-function.

We make the comment that \( \tilde{P}_y \) is the probability measure for the
random variable \( \tilde{y} \), and thus, a naive application of Baves' rule yields
\[
p_y|\tilde{y}(\xi|\beta) = \frac{p_y|\tilde{y}(\beta|\xi)p_y(\xi)}{p_y(\beta)}
\]
\[
= \frac{\delta(\beta-(\xi \text{mod} 2\pi))p_y(\xi)}{\sum_{k=-\infty}^{+\infty} p_y(\beta+2k\pi)} \tag{B.47}
\]
We now consider a more general version of the single stage
problem of Section 4.4. We assume that we have a probability space
\((\mathbb{R}^2, \mathcal{B}_3, P_{xy})\), where \( \mathcal{B}_3 \) is the Borel \( \sigma \)-algebra of \( \mathbb{R}^2 \) and \( P_{xy} \) is any
probability measure on $\mathcal{B}_3$. We define three random variables

$$x(\omega_1, \omega_2) = \omega_1 \quad (B.48)$$

$$y(\omega_1, \omega_2) = \omega_2 \quad (B.49)$$

$$y(\omega_1, \omega_2) = \omega_2 \mod 2\pi \quad (B.50)$$

and the marginal distributions

$$P_x(A) = P_{xy}(A \times R^1) \quad A \in \mathcal{B}_1 \quad (B.51)$$

$$P_y(B) = P_{xy}(R^1 \times B) \quad B \in \mathcal{B}_1 \quad (B.52)$$

We let $\mathcal{N}_y$ be the minimum sub $\sigma$-algebra of $\mathcal{B}_3$, with respect to which $y$ is measurable, and we define $\mathcal{N}_x$ and $\mathcal{N}_y$ analogously.

We wish to compute the conditional measure $P_{x|\tilde{y}}$. As before, we obtain a form for this measure that reflects our uncertainty as to the number of multiples of $2\pi$ that are "chopped off" of $y$ in the process of observing $\tilde{y}$. To derive the desired result, we will need to consider two other conditional measures, $P_{x|\tilde{y}, \tilde{y}}$ and $P_{x|y}$. Since $\mathcal{N}_y \subseteq \mathcal{N}_\tilde{y}$ (i.e. $\tilde{y}$ is a deterministic function of $y$), we have

$$P_{x|\tilde{y}, \tilde{y}}(A|\alpha, \alpha \mod 2\pi) = P_{x|y}(A|\alpha) \quad (B.53)$$

As before, we define the following measures on $([-\pi, \pi], \mathcal{N}_2)$:

$$P^n_y(S) = P_y(S + 2n\pi) \quad (B.54)$$

$$P_y(S) = \sum_{n=-\infty}^{+\infty} P^n_y(S) \quad S \in \mathcal{N}_2 \quad (B.55)$$
Theorem B.3: The conditional distribution $P_{x|\tilde{y}}(C|\beta)$ is given by

$$P_{x|\tilde{y}}(C|\beta) = \sum_{n=-\infty}^{+\infty} \frac{dP_y}{dP_{\tilde{y}}} (\beta) P_{x|y}(C|\beta + 2n\pi)$$  \hspace{1cm} (B.56)

Proof: It is easy to see from Theorem B.2 that the conditional probability

$$P_{y|\tilde{y}}(y = \xi|\beta) = P_{x,y|\tilde{y}}(z \in R^1, v = \xi|\beta)$$  \hspace{1cm} (B.57)

exists and is given by

$$P_{y|\tilde{y}}(y = \xi|\beta) = \begin{cases} \frac{dP_y}{dP_{\tilde{y}}} (\beta) & \xi = \beta + 2k\pi \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (B.58)

Using the properties of iterated expectations, [N1] and equation (B.53)

$$P_{x|\tilde{y}}(C|\beta) = \mathcal{E}\{ P_{x|y}(C|\xi) | \tilde{y} = \beta \}$$

$$= \int_{R^2} P_{x|y}(C|y(\omega_1, \omega_2) = \xi) \, dP_{x,y|\tilde{y}}(\omega_1, \omega_2 | \tilde{y} = \beta)$$  \hspace{1cm} (B.59)

Since $P_{x|y}(C|y(\omega_1, \omega_2) = \xi)$ is independent of $\omega_1$, we perform the integration with respect to $\omega_1$ first.
\[ P_{X|\bar{y}}(C|\beta) = \int_{-1}^{1} P_{X|y}(C|\omega_2) dP_{Y|\bar{y}}(\omega_2|\beta) \]

\[ = \sum_{n=-\infty}^{+\infty} \frac{dP_{Y}^{n}}{dP_{Y}} (\beta) P_{X|y}(C|y = \beta + 2n\pi) \]

\[ = \sum_{n=-\infty}^{+\infty} P(y = \beta + 2n\pi|\bar{y} = \beta) P_{X|v}(C|v = \beta + 2n\pi) \quad (B.60) \]

The following two corollaries solve the general single stage problem of Section 4.4 and the special linear problem.

**Corollary 1:** Suppose \( x \) and \( v \) are independent real valued random variables and define

\[ y = h(x) + v \quad (B.61) \]

\[ \bar{y} = y \mod 2\pi \quad (B.62) \]

where \( h : R^1 \rightarrow R^1 \) is measurable. Also, suppose \( p_x(\alpha) \) and \( p_v(\nu) \) are the probability densities for \( x \) and \( v \) respectively. Then a version of the probability density \( p_{X|\bar{v}}(\alpha|\beta) \) is given by

\[ p_{X|\bar{v}}(\alpha|\beta) = \sum_{n=-\infty}^{+\infty} \frac{p_{Y}(\beta + 2n\pi)}{p_{\bar{v}}(\beta)} p_{X|y}(\alpha|\beta + 2n\pi) \quad (B.63) \]

\[ = \sum_{n=-\infty}^{+\infty} \frac{p_{Y}(\beta + 2n\pi|\alpha) p_{X}(\alpha)}{p_{\bar{v}}(\beta)} \quad (B.64) \]

where

\[ p_{Y|x}(\beta + 2n\pi|\alpha) = p_{v}(\beta + 2n\pi - h(\alpha)) \quad (B.65) \]

\[ p_{Y}(\beta + 2n\pi) = \int_{-\infty}^{+\infty} p_{Y|x}(\beta + 2n\pi|u) p_{X}(u) du \quad (B.66) \]

\[ p_{\bar{v}}(\beta) = \sum_{n=-\infty}^{+\infty} p_{v}(\beta + 2n\pi) \quad (B.67) \]
Proof: Equation (B.63) follows from Theorem B.3, the corollary to Theorem B.2, and the observation that, if the measure $p_{x|y}(C|\beta)$ has a density with respect to Lebesgue measure, then, from (B.56), so does $p_{x|y}(C|\beta)$, and it is given by (B.63). Equations (B.65), (B.66), and (B.67) are immediate consequences of the definitions and the independence of $x$ and $v$. Equation (B.64) follows from (B.63) and Bayes' rule.

**Corollary 2:** If $x$ and $v$ are normally distributed and $h$ is linear ($h(x) = ax$), then $p_{x|y}(\alpha|\beta)$ is expressible as the linear combination (B.72) of an infinite number of normal distributions, with weighting coefficients that are functions of the measurement and are given by

$$c_n(\beta) \triangleq \frac{p_y(\beta + 2n\pi)}{p_y(\beta)} = \frac{N(\beta + 2n\pi; a\eta; a^2\gamma_1 + \gamma_2)}{\sum_{k=-\infty}^{\infty} N(\beta + 2k\pi; a\eta; a^2\gamma_1 + \gamma_2)} \quad (B.68)$$

where

$$p_x(\alpha) = \frac{1}{\sqrt{2\pi\gamma_1}} e^{-\frac{(\alpha - \eta)^2}{2\gamma_1}} = N(\alpha; \eta, \gamma_1) \quad (B.69)$$

$$p_v(\nu) = \frac{1}{\sqrt{2\pi\gamma_2}} e^{-\frac{\nu^2}{2\gamma_2}} = N(\nu; 0, \gamma_2) \quad (B.70)$$

Proof: We will use the form of $p_{x|y}(\alpha|\beta)$ given in (B.63). The additive properties of independent normally distributed random variables, [M9], yield

$$p_y(\xi) = N(\xi; a\eta, a^2\gamma_1 + \gamma_2) \quad (B.71)$$
and therefore the equation for \(c_n(\beta)\) is correct. Then

\[
P_{x|y}(\alpha|\beta) = \sum_{n=-\infty}^{\infty} c_n(\beta)p_{x|y}(\alpha|\beta + 2n\pi). \tag{B.72}
\]

But \(p_{x|y}\) is the solution of a \underline{linear} filtering problem, and therefore is a normal distribution. In fact

\[
p_{x|y}(\alpha|\beta + 2n\pi) = N(\alpha; \eta_n, \gamma_3), \tag{B.73}
\]

where

\[
\gamma_3 = \frac{\gamma_1\gamma_2}{a^2\gamma_1 + \gamma_2} \tag{B.74}
\]

\[
\eta_n = \frac{\eta\gamma_2 + \gamma_1 a(\beta + 2n\pi)}{a^2\gamma_1 + \gamma_2} \tag{B.75}
\]

B.4 A Limiting Argument

We have seen in Sections 4.4 and B.3 that the ambiguity concerning the number of rotations leads to equations for the conditional density that involve infinite sums. Intuitively, if we observe the process continuously, this ambiguity should disappear — assuming that the random processes involved are continuous. From the rigorous arguments of Sections 4.2 and 4.3 we have seen that this is the case — i.e. in the limit we know \(dy(t)\), not just \(dy(t) \mod 2\pi\). We can also see this by examining the discrete approximation to the continuous problem.

Our discrete equations are
\[ \Delta y_k = (\Delta y_k \mod 2\pi) \]

\[ = [m(x_k, k\Delta t)\Delta t + q^{1/2}(k\Delta t) \Delta w_k] \mod 2\pi \]  \hspace{1cm} (8.76)

where \( x_k = x(k\Delta t) \), and \( x(t) \) is a continuous process, independent of the Brownian motion process \( w(t) \). We also assume \( q(t) \) is continuous, \( m(x, t) \) is measurable in \( x \) for all \( t \) and continuous in \( t \) for all \( x \).

We wish to examine the effect of one additional such measurement at time \( t \), in terms of the size of \( \Delta t \). Thus, we assume we have computed

\[ p_x(t)(\alpha) \overset{\Delta}{=} p_x(t)(\alpha \mid \text{past measurements}) \]  \hspace{1cm} (B.77)

and that we take the measurement

\[ \Delta y_t = (\Delta y_t \mod 2\pi) \]

\[ = (m(x(t), t)\Delta t + q^{1/2}(t) \Delta w_t) \mod 2\pi \]  \hspace{1cm} (B.78)

As indicated in equation (B.77) we will suppress all conditioning on past measurements. Thus, we wish to compute \( p_x(t) | \Delta y_t (\alpha | \beta) \) in terms of \( p_x(t)(\alpha) \) and the new information \( \Delta y_t \) (here \( \beta \) is the observed value of \( \Delta y_t \)).

Using the discrete measurement formulae, we have

\[ p_x(t) | \Delta y_t (\alpha | \beta) = \sum_{n=-\infty}^{+\infty} c_n(\beta)p_x(t) | \Delta y_t (\alpha | \beta + 2n\pi) \]  \hspace{1cm} (B.79)

where we have the explicit formula
\[
\begin{align*}
\frac{c_n(\beta)}{\sum_{r=-\infty}^{\infty} \int_{-\infty}^{\infty} N(\beta+2\pi n-m(u,t)\Delta t; 0, q(t)\Delta t)p_{x(t)}(v)dv} &= \\
&= \frac{\int_{-\infty}^{\infty} \exp - \frac{1}{2q(t)\Delta t} \{\beta+2\pi n-m(u,t)\Delta t\}^2 p_{x(t)}(u)du}{\sum_{r=-\infty}^{\infty} \int_{-\infty}^{\infty} \exp - \frac{1}{2q(t)\Delta t} \{\beta+2\pi n-m(u,t)\Delta t\}^2 p_{x(t)}(u)du} \\
&= (B.80)
\end{align*}
\]

Examining this expression, we see that the numerator contains a term
\[
\exp = \frac{2n^2 \pi^2}{q(t)\Delta t}
\]

which is \(o(\Delta t^2)\) if \(n \neq 0\). Thus, one sees that for small \(\Delta t\), the probability that \(\Delta y_t\) and \(\tilde{\Delta} y_t\) differ by a nonzero multiple of \(2\pi\) appears to go to zero quite fast as \(\Delta t \to 0\). To make a precise statement concerning this, we must make some technical assumptions.

(1) The probability density for \(x(t)\) conditioned on the past measurements, \(p_{x(t)}(\alpha)\), exists.

(2) The conditional density for \(x(t)\), if we were to measure \(\Delta y_t\) (not \(\tilde{\Delta} y_t\)), \(p_{x(t)|\Delta y_t}(\alpha|\beta)\) exists and is bounded uniformly for all \(\alpha\) and \(\beta\).

(3) We have the following bound:
\[
\int_{-\infty}^{\infty} e^{-\alpha^2 m^2(u,t) + \xi m(u,t)} p_{x(t)}(u)du \leq K(\alpha^2) e^{k(\alpha^2)\xi^2} \quad (B.81)
\]

where \(K(\alpha^2), k(\alpha^2)\) are bounded for \(\alpha^2 \in [0, \gamma]\), for some \(\gamma > 0\).
We can now prove the following

**Theorem B.4**: If the assumptions above hold, we have the following relationships:

\[ \sum_{n \neq 0} \frac{c_n(\beta)}{c_0(\beta)} = o(\Delta t^\ell) \quad \forall \ell \geq 0 \quad (B.82) \]

\[ \left| \left| p_x(t) | \Delta y (\alpha | \beta) - p_x(t) | \bar{\Delta} y (\alpha | \beta) \right|_L \right|^\infty = o(\Delta t^\ell) \quad \forall \ell \geq 0 \quad (B.83) \]

We will need the following technical lemma:

**Lemma B.3**: Let \( y(t) \) be a continuous, real-valued function of time, and define

\[ \bar{\Delta} y (s) = (y(s) - y(t)) \bmod 2\pi \quad (s > t) \quad (B.84) \]

Let \( q \) be a positive constant, \( \rho(x) \) a measurable, real-valued function on \( \mathbb{R}^1 \), \( h > 0 \), an element of \( L^1(\mathbb{R}) \) such that \( ||h||_{L^1} > 0 \), and \( \{s_n\}_{n=1}^\infty \) a sequence of real numbers decreasing to \( t \). Then there exists an integer \( N_0 \), such that

\[ \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2q} \left\{ o^2(x)(s_n - t) - 2\bar{\Delta} y (s_n) \rho(x) \right\} \right) h(x) dx \]

\[ \geq \frac{1}{2} ||h||_{L^1} \quad \forall \; n \geq N_0 \quad (B.85) \]

**Proof**: Let \( F_n(x) \) be the integrand of the left-hand side of (B.85).

Then, for fixed \( x \)

\[ \lim_{n \to \infty} F_n(x) = h(x) \quad (B.86) \]

By Fatou's Lemma ([H5], [R3])
\[
\lim \inf_n \int f_n \, dx \geq \lim \inf_n \int f_n \, dx = \| h \|_{L^1} \tag{B.87}
\]

Choose \( N_0 \), such that for \( m \geq N_0 \)
\[
\int f_m \, dx \geq \lim \inf_n \int f_n \, dx - \frac{1}{2} \| h \|_{L^1} \tag{B.88}
\]

Then
\[
\int f_m \, dx \geq \frac{1}{2} \| h \|_{L^1} \quad \forall \ m \geq N_0 \tag{B.89}
\]

Proof of Theorem B.4: The proof of this result requires some straightforward but rather lengthy computations. Thus, we shall only sketch the proof, leaving the details to the interested reader.

Consider the infinite sum
\[
g(\beta) = \sum_{n \neq 0} \int_{-\infty}^{\infty} \exp \left\{ \frac{1}{2q(t)\Delta t} (2\beta(2n\pi - m(u,t))\Delta t) + (2n\pi - m(u,t))^2 \right\} p_{x(t)}(u) \, du \tag{B.90}
\]

Using (B.81), we have
\[
g(\beta) \leq \sum_{n \neq 0} \exp \left\{ \frac{n\pi(2n\pi - 2\beta)}{q(t)\Delta t} \right\} K(\Delta t)\exp \left\{ k(\Delta t) \right\} \left( \frac{2n\pi + 2\beta}{q(t)} \right)^2 \tag{B.91}
\]

We now note that \( w(t) \) is a continuous random process, and, therefore, we assume that we are given a continuous sample path, \( w^0(t) \). We then choose a \( \delta > 0 \), \( K_0 > 0 \), \( k_0 > 0 \), such that
\[
K \left( \frac{\Delta t}{2q} \right) \leq K_0 \tag{B.92}
\]
\[
k_0 \left( \frac{\Delta t}{2q} \right) \leq k_0 \tag{B.93}
\]
\[
| \Delta y^0(t) | \leq \pi/2 \tag{B.94}
\]
for all $\Delta t < \delta$ (here $\tilde{\Delta}y^0_t$ is the value for the particular sample path $w^0(t)$ selected). Then, for $\Delta t < \delta$.

\[
g(\tilde{\Delta}y^0_t) \leq K_0 \sum_{n \neq 0} \exp \left\{ -\frac{(2n^2 - |n|)}{q(t)\Delta t} \pi^2 \right\} \exp k_0 \left( \frac{2|n| + 1}{q(t)} \pi \right)^2 \quad (B.95)
\]

and for $\Delta t < \min(\delta, q(t)/2k_0)$, the right hand side of (B.95) is finite.

Examining equation (B.80), we can write

\[
\frac{1}{\Delta t} \sum_{n \neq 0} \frac{c_n(\tilde{\Delta}y^0_t)}{c_0(\tilde{\Delta}y^0_t)} = \frac{g(\tilde{\Delta}y^0_t)}{\int_{-\infty}^{\infty} \exp -\frac{1}{2q(t)} \left\{ m^2(u, t)\Delta t - 2\tilde{\Delta}y^0_t m(u, t) \right\} p_x(t)(u)du} \quad (B.96)
\]

Taking a sequence $\{\Delta t_r\}_{r=1}^\infty$ decreasing to zero and using Lemma B.3, we see that there is an $R_0$ such that

\[
\int_{-\infty}^{\infty} \exp -\frac{1}{2q(t)} \left\{ m^2(u, t)\Delta t_r - 2\tilde{\Delta}y^0_r m(u, t) \right\} p_x(t)(u)du \geq \frac{1}{2}
\]

\[
\forall \quad r \geq R_0 \quad (B.97)
\]

where

\[
\tilde{\Delta}y^0_r(t) = \{m(x(t), t)\Delta t_r + q(t)\Delta r w^0_t\mod 2\pi \quad (B.98)
\]

and

\[
\Delta r w^0_t = w^0(t + \Delta r_r) - w^0(t) \quad (B.99)
\]

Then, for $r \geq R_0$

\[
\frac{1}{\Delta t_r} \sum_{n \neq 0} \frac{c_n(\tilde{\Delta}y^0_r(t))}{c_0(\tilde{\Delta}y^0_r(t))} \leq \frac{2g(\tilde{\Delta}y^0_r(t))}{\Delta t_r} \quad (B.100)
\]
Using (B.95), it can easily be shown that for any \( \varepsilon > 0 \), there exists a positive integer \( r(\varepsilon) \) such that

\[
\frac{1}{\Delta t} \sum_{n \neq 0} \frac{c_n(\Delta y_t^0(\tau))}{c_0(\Delta y_t^0(\tau))} \leq \varepsilon \quad \forall \ r \geq r(\varepsilon)
\]  \hspace{1cm} (B.101)

Thus

\[
\lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{n \neq 0} \frac{c_n(\Delta y_t^0)}{c_0(\Delta y_t^0)} = 0
\]  \hspace{1cm} (B.102)

for any continuous sample function \( w^0 \).

To prove (B.83), we use the assumption that \( p_{x(t)|\Delta y_t}(\alpha|\beta) \) is bounded for all \( \beta \) and \( \alpha \). Let \( M \) be an upper bound. Then rewriting equation (B.79) for the particular sample path chosen, we have

\[
p_{x(t)|\Delta y_t}(\alpha|\Delta y_t^0) = p_{x(t)|\Delta y_t}(\alpha|\Delta y_t^0)
\]

\[
- \frac{\sum_{\tau \neq 0} c_{\tau}/c_0}{1 + \sum_{\tau \neq 0} c_{\tau}/c_0} p_{x(t)|\Delta y_t}(\alpha|\Delta y_t^0)
\]

\[
+ \sum_{n \neq 0} \frac{c_n/c_0}{1 + \sum_{\tau \neq 0} c_{\tau}/c_0} p_{x(t)|\Delta y_t}(\alpha|\Delta y_t^0 + 2n\pi)
\]  \hspace{1cm} (B.103)

Thus

\[
|p_{x(t)|\Delta y_t}(\alpha|\Delta y_t^0) - p_{x(t)|\Delta y_t}(\alpha|\Delta y_t^0)| \leq \varepsilon
\]

\[
2M \left[ \frac{\sum_{n \neq 0} c_n/c_0}{1 + \sum_{n \neq 0} c_n/c_0} \right] = o(\Delta t^\varepsilon) \quad \forall \ \varepsilon \geq 0
\]  \hspace{1cm} (B.104)
We note that Theorem B.4 may still be true even if (B.81) is not satisfied. An examination of the proof shows that all we require is the following: let

$$
\rho(\alpha^2, \xi) = \int_{-\infty}^{+\infty} e^{-\alpha^2 m^2(u, t) + \xi m(u, t)} p_x(t)(u) \, du \tag{B.105}
$$

Then we must have

$$
\sum_{n \neq 0} \exp \left\{ \frac{n\pi(2n\pi - 2\tilde{A}y_0)}{q(t)\Delta t} \right\} \rho \left[ \frac{\Delta t}{2q(t)}, \frac{(2n\pi + 2\tilde{A}y_0)}{q(t)} \right] = o(\Delta t^2)
$$

$$
\psi \ell \geq 0 \tag{B.106}
$$

For instance

$$
\rho(\alpha^2, \xi) \sim K_1 e^{-K_2(\xi + c)^2} + K_3 \xi^r \tag{B.107}
$$

for given c, r, K_1, K_2, K_3, which depend only on \( \alpha \) and are bounded as \( \alpha \to 0 \), will satisfy. Thus, for example, Theorem B.4 holds if \( m \) is linear and \( p_x(t) \) is normal.
APPENDIX C

EQUIVALENCE CLASS OBSERVATIONS AND TRUNCATION
OF SERIES OF NORMAL DENSITIES

"That is best which lieth nearest;
Shape from that thy work of art.


C.1 Introduction

In this appendix we consider in more detail some of the questions raised in Section 4.4, where we studied a discrete-time estimation problem on $S^1$. In that section we saw how the mod $2\pi$ ambiguity of the measurements led to infinite series expressions for the conditional density. In Section C.2 we extend this measurement-ambiguity concept to other problems, and in Section C.3 we study methods for truncating the series that arise in these problems.

C.2 Measurement Ambiguity Problems

In Section 4.4 the type of observation process used was

\[ y = h(x) + \nu \]  \hspace{1cm} (C.1)

\[ \tilde{y} = y \mod 2\pi \]  \hspace{1cm} (C.2)

where $x$ and $\nu$ are independent real-valued random variables. An equivalent way of expressing (C.2) is that our measurement $\tilde{y} = m(\nu)$ where $m : \mathbb{R} \to \mathbb{R}^1/2\pi\mathbb{Z}$ is the map

\[ m(x) = \{x+2\pi n | n \in \mathbb{Z} \} \]  \hspace{1cm} (C.3)
That is, our measurement process is, in fact, an equivalence class measurement, and thus a "measurement ambiguity" is introduced into the problem.

One is then led to consider the more general problem. Let "=" be an equivalence relation on $\mathbb{R}^1$ (see [K3]) -- i.e. a relation between elements of $\mathbb{R}^1$ that satisfies

(1) $x = x$  \hspace{1cm} (C.4)
(2) $x = y \iff y = x$  \hspace{1cm} (C.5)
(3) $x = y, y = z \implies x = z$  \hspace{1cm} (C.6)

for all $x, y, z \in \mathbb{R}^1$. This relation splits $\mathbb{R}^1$ into equivalence classes (an equivalence class, written $[x]$, is the set of all $y \in \mathbb{R}^1$ such that $y \equiv x$) and the set of equivalence classes is denoted $\mathbb{R}^1/\equiv$. Clearly $x \equiv y \iff x \equiv y \mod 2\pi$ is an equivalence relation and in this case $\mathbb{R}^1/\equiv$ is just $\mathbb{R}^1/2\pi\mathbb{Z}$.

We now wish to consider the following: suppose we are given two independent real-valued random variables $x$ and $v$, with densities $p_x(x)$ and $p_v(v)$, respectively, an equivalence relation "=" on $\mathbb{R}^1$, and the maps $h : \mathbb{R}^1 \rightarrow \mathbb{R}^1$, which is measurable, and $m : \mathbb{R}^1 \rightarrow \mathbb{R}^1/\equiv$ defined by

$$m(z) = [z]$$  \hspace{1cm} (C.7)

The problem is to compute $p_x(y|x|\beta)$, where

$$y = h(x) + v$$  \hspace{1cm} (C.8)
$$\tilde{y} = m(y)$$  \hspace{1cm} (C.9)

To define conditional probabilities, we need to define a sub-$\sigma$-algebra of $\mathcal{B}_1$, the $\mathbb{R}^1$ Borel sets, that $\tilde{y}$ "generates." We do this in
the natural way -- i.e. we take the largest \( \sigma \)-algebra \( \mathcal{B}_2 \) on
\( R^1/\sim \) with respect to which \( m : (R^1, \mathcal{B}_1) \to (R^1/\sim, \mathcal{B}_2) \) is measurable,
and then our sub-\( \sigma \)-algebra \( \mathcal{A}_y \) of \( \mathcal{B}_1 \) is
\[
\mathcal{A}_y = m^{-1}(\mathcal{B}_2) = \{m^{-1}(\tilde{A}) | \tilde{A} \in \mathcal{B}_2 \}
\]
(C.10)

Also it is clear that
\[
\mathcal{B}_2 = \{\tilde{A} \subseteq R^1/\sim | m^{-1}(\tilde{A}) \in \mathcal{B}_1 \}
\]
(C.11)

(note that it is easy to show that \( \mathcal{B}_2 \) is the Borel \( \sigma \)-algebra associated
with the quotient topology [K3] on \( R^1/\sim \)). Note that these definitions
are consistent with the mod 2\( \pi \) case of Appendix B, and that now we wish
to find
\[
P_{x|y}(C|\tilde{y}(\omega)) \overset{\Delta}{=} P_{x|\widetilde{y}}(C|\mathcal{A}_y)(\omega)
\]
(C.12)
and its density if \( P_{x|y} << \lambda \) (Lebesque measure).

For simplicity, in the remainder of this Appendix we will assume
that \([x] \in \mathcal{A}_y \) \( \forall x \in R^1 \) -- i.e. that each equivalence class is Borel
measurable. This is not such an unreasonable assumption. For example,
if our actual measurement is
\[
\tilde{v} = m_1(y)
\]
(C.13)
where \( m_1 : R^1 \to R^1 \) is measurable, then an equivalent measurement is
\[
\tilde{v} = m(y)
\]
(C.14)
where \( m \) is the projection map for the equivalence relation
\( y_1 \sim y_2 \iff m_1(y_1) = m_1(y_2) \). In this case \([x] = m^{-1}_1(m_1(x)) \) is measurable
\( \forall x \in R^1 \). Clearly the mod 2\( \pi \) case is of this type.
As with the mod $2\pi$ equivalence relation, the key to the computation of $P_{x|\tilde{y}}$ and $p_{x|\tilde{y}}$ is finding a way to characterize the conditional probability distribution $P_{x|\tilde{y}}(\cdot|\beta)$ for $C$ an $\mathbb{R}^1$ Borel set (see Theorems B.2 and B.3). In the general case we cannot write down equations as explicit as those of Section B.3; however we can say some things. It is clear that $P_{x|\tilde{y}}(C|\beta)$ is concentrated on the equivalence class $\beta$. If the unconditional probability $P_{x}(\beta) = 0$, then $P_{x|\tilde{y}}$ will be singular ([H5], [R3], [S9]) with respect to $P_{y}$ (as in the mod $2\pi$ case) -- i.e. there exists a set, namely $\beta$ such that $P_{y}(\beta) = 0$, but $P_{y|\tilde{y}}(\beta|\beta) = 1$. If $P_{y}(\beta) > 0$, we can use the usual conditional probability formula [P2]

$$P_{y|\tilde{y}}(C|\beta) = \frac{P_{y}(C \cap \beta)}{P_{y}(\beta)} \quad \text{(C.15)}$$

Referring to Theorem B.3, we see that equation (B.60) holds for a general equivalence relation -- i.e. given random variables $x$ and $v$ and $\tilde{y} = m(y)$, where $m$ is given by (C.7), we have

$$P_{x|\tilde{y}}(C|\beta) = \int_{\mathbb{R}^1} P_{x|v}(C|\omega)dP_{y|\tilde{y}}(\omega|\beta) \quad \text{(C.16)}$$

and if $P_{x|y} \ll \lambda$ (Lebesgue measure) and the density $p_{x|y}(\alpha|\omega)$ is jointly measurable in $\alpha$ and $\omega$, then $p_{x|\tilde{y}}(\alpha|\beta)$ exists and is given by

$$p_{x|\tilde{y}}(\alpha|\beta) = \int_{\mathbb{R}^1} p_{x|y}(\alpha|\omega)dP_{y|\tilde{y}}(\omega|\beta) \quad \text{(C.17)}$$

where we have used the Fubini Theorem [H5], [R3]. In particular, if $x$ and $v$ are independent random variables with densities $p_{x}(\alpha)$ and $p_{v}(\nu)$ respectively and $y$ is given by (C.8), it is easy to see that (C.17) holds and $p_{x|y}$ is given by
\[ p_{x|y}(\alpha|\omega) = \frac{p_{y|x}(\omega|\alpha)p_{x}(\alpha)}{p_{y}(\omega)} \]  \hspace{1cm} (C.18)

where

\[ p_{y|x}(\omega|\alpha) = p_{y}(\omega-h(\alpha)) \]  \hspace{1cm} (C.19)

\[ p_{y}(\omega) = \int_{-\infty}^{\infty} p_{y|x}(\omega|u)p_{x}(u)du \]  \hspace{1cm} (C.20)

If \( h(x) = hx \), \( p_{x}(\alpha) = N(\alpha;\eta_{1},\gamma_{1}) \), \( p_{y}(\nu) = N(\nu;0,\gamma_{2}) \), then we have the linear-Gaussian-ambiguity equation

\[ p_{x|\tilde{y}}(\alpha|\beta) = \int_{-\infty}^{\infty} N(\alpha;\eta(\omega),\gamma_{3})dP_{y}(\omega) \]  \hspace{1cm} (C.21)

where

\[ \gamma_{3} = \frac{\gamma_{1}\gamma_{2}}{h^{2}\gamma_{1}+\gamma_{2}} \]  \hspace{1cm} (C.22)

\[ \eta(\omega) = \frac{\gamma_{2}\gamma_{1}h\omega}{h^{2}\gamma_{1}+\gamma_{2}} \]  \hspace{1cm} (C.23)

This result holds for an arbitrary equivalence relation (such that \([x] \in B_{1} \forall x \in \mathbb{R}^{1}\), and the results of Section B.3 can be derived from these equations.

Also, if \( P_{y}(\beta) > 0 \), we can say even more by applying (C.15).

For instance (C.21) becomes

\[ p_{x|\tilde{y}}(\alpha|\beta) = \frac{1}{P_{y}(\beta)} \int_{B} N(\alpha;\eta(\omega),\gamma_{3})dP_{y}(\omega) = \frac{1}{P_{y}(\beta)} \int_{-\infty}^{\infty} \chi_{B}(\omega)N(\alpha;\eta(\omega),\gamma_{3})dP_{y}(\omega) \]  \hspace{1cm} (C.24)

An example of an equivalence relation, for which \( P_{y}(\beta) > 0 \) for at least some \( \beta \), and which is of some practical importance, is
\[ x = y \iff [x] = [y] \quad \text{i.e. we quantize our information (all numbers in the interval \((n, n+1]\)} \text{ are given the quantized value } n. \text{ See [C5]}\]

for further discussion of this example.

Note that for mod\(2\pi\) equivalence \(P_y(\beta) = 0\), and we were forced to play a measure-theoretic game to characterize \(P_y|_{\tilde{y}}\). Our success was due to the "nice" structure of the equivalence classes. Each equivalence class had a unique element in the interval \([-\pi, \pi]\) (i.e. a "distinguished" element) and

\[
[x] = \{x + 2\pi n \mid n \in \mathbb{Z}\} \quad (C.25)
\]

which led directly to the infinite sum form for \(P_y|_{\tilde{y}}\). Thus, if \(P_y(\beta) = 0\), we should examine the structure of the equivalence relation to see if we can use it to advantage.

Another example of an equivalence relation whose structure can be utilized effectively is

\[ x = y \iff |x| = |y| \quad (C.26) \]

and in this case

\[ m(x) = \{\pm x\} \quad (C.27) \]

Consider independent real random variables \(x\) and \(v\) with \(p_x(x) = N(\alpha; \gamma_1, \gamma_2)\) and \(p_v(v) = N(0; 0, \gamma_2)\), and define the observation \(\tilde{y}\) by

\[ y = hx + v \quad (C.28) \]

\[ \tilde{y} = m(y) = \{\pm y\} \quad (C.29) \]

As in the mod\(2\pi\) case, we choose a distinguished element of each equivalence class, namely the nonnegative element. Thus our measurement
process is

\[ \tilde{y} = |hx + v| \]  

(C.30)

Using the analysis of this section it is easy to show that

\[ p_x|\tilde{y}(a|B) = c_1(B)N(\alpha;\eta(B),\gamma_3) + (1-c_1(B))N(\sigma;\eta(-B),\gamma_3) \]  

(C.31)

where \( \gamma_3 \) is given by (C.22),

\[ c_1(B) = \Pr(y=B|\tilde{y}=B) = \frac{N(\beta;\eta\eta,\gamma_1+\gamma_2)}{N(\beta;\eta\eta,\gamma_1+\gamma_2)+N(-\beta;\eta\eta,\gamma_1+\gamma_2)} \]  

(C.32)

\[ 1-c_1(B) = \Pr(y=-B|\tilde{y}=B) = \frac{N(-\beta;\eta\eta,\gamma_1+\gamma_2)}{N(\beta;\eta\eta,\gamma_1+\gamma_2)+N(-\beta;\eta\eta,\gamma_1+\gamma_2)} \]  

(C.33)

and

\[ \eta(\omega) = \frac{\gamma_2\eta + \gamma_1\hbar\omega}{\hbar^2\gamma_1+\gamma_2} \]  

(C.34)

Thus, the conditional density is computed by two linear filters, one taking \( B \) as the measurement, the other \(-B\). The analogy between this result and the mod 2\( \pi \) result is clear.

Note that if we take \( N \) measurements with independent normally distributed noises

\[ \tilde{y}_k = |hx + v_k| \quad k=1,\ldots,N \]  

(C.35)

the conditional density \( p_x|y_1,\ldots,y_N \) is the linear combination of \( 2^N \) normal densities, each of which is computed by a linear estimator that takes either \( \tilde{y}_k \), \( k=1,\ldots,N \) as its measurement history. The corresponding coefficient of this term is the probability that \( y_k \), \( k=1,\ldots,N \) equals
the chosen sequence (either $+\tilde{y}_k$ or $-\tilde{y}_k$, $k=1,...,N$). We can also let $x$ be a discrete process, described by a linear difference equation

$$x_{k+1} = a_k x_k + b_k^{1/2} v_k \quad x_0 = 0$$  \hspace{1cm} (C.36)

where the $w_k$ are normally distributed and independent of each other and of the $v_j$. In this case, if we observe

$$\tilde{y}_k = |hx_k + v_k| \quad k=1,...,N$$  \hspace{1cm} (C.37)

the density $p_{x_N|y_1,...,y_N}$ is also the sum of $2^N$ normal densities (see Section 4.4 for the mod $2\pi$ analogy).

Note that there are less than $2^N$ terms if any of the $\tilde{y}_k = 0$, since in this case $y_k = \tilde{y}_k$ (note, however, that (C.31) - (C.34) still hold). The special nature of the value 0 is due to the fact that

$$\text{card}([x]) = 2 \quad \text{if} \quad x \neq 0$$  \hspace{1cm} (C.38)

while

$$\text{card}([0]) = 1$$  \hspace{1cm} (C.39)

In the next section we will discuss methods for truncating series of normal densities that arise in this manner. We will find it necessary to use the equations for updating the coefficients of the terms in the series. In the mod $2\pi$ case, see equations (4.169) - (4.173). We now derive the analogous results for the absolute value problem.

Let $p_{\nu}(\nu) = N(\nu; 0, \gamma_2)$ and

$$p_x(\alpha) = \sum_{k=1}^{M} d_k N(\alpha; \eta_k, \gamma_1)$$  \hspace{1cm} (C.40)
\[ \sum_{k=1}^{M} d_k = 1 \quad d_k > 0 \]  \hspace{1cm} (C.41)

Consider the measurement given by (C.30). In this case, a straightforward computation yields

\[ p_x | \gamma (\alpha | \beta) = \sum_{k=1}^{M} e_k^+(\beta)N(\alpha; \eta_k^+(\beta), \gamma_3) + \sum_{k=1}^{M} e_k^-(\beta)N(\alpha; \eta_k^-(\beta), \gamma_3) \]  \hspace{1cm} (C.42)

where \( \gamma_3 \) is given by (C.22) and

\[ \eta_k^\pm(\beta) = \frac{\gamma_2 \eta_k + \gamma_1 h\beta}{h^2 \gamma_1 \pm \gamma_2} \]  \hspace{1cm} (C.43)

\[ e_k^+(\beta) = \frac{d_k N(\beta; h\eta_k, h^2 \gamma_1 + \gamma_2)}{z(\beta)} \]  \hspace{1cm} (C.44)

\[ e_k^-(\beta) = \frac{d_k N(-\beta; h\eta_k, h^2 \gamma_1 + \gamma_2)}{z(\beta)} \]  \hspace{1cm} (C.45)

where

\[ z(\beta) = \sum_{k=1}^{M} d_k [N(\beta; h\eta_k, h^2 \gamma_1 + \gamma_2) + N(-\beta; h\eta_k, h^2 \gamma_1 + \gamma_2)] \]  \hspace{1cm} (C.46)

C.3 Truncation of Series of Normal Densities

In this section we will discuss a number of methods for truncating the series of normal densities that arise in the discrete-time estimation problems discussed in Chapter 4, Appendix B, and the preceding section. The approaches that will be described are motivated by the work of Buxbaum and Haddad [B10]. Their reason for considering this problem is somewhat different, since the sums of normal densities they considered
arose from nongaussian input signals. The reader is referred to [Bl0] to see the similarity between the truncation problem of Buxbaum and Haddad and ours. We only note that in our case the coefficients of the various terms in the series have a clear interpretation as probabilities of various events (e.g. \( \Pr(\gamma = \gamma + 2k\pi | \gamma) \)).

The precise problem statement is as follows: we are given a positive integer \( N \), a discrete-time signal process \( \{x(k)\} \), and a sequence of measurements of the signal. We assume that the exact conditional probability density is of the form

\[
p(\alpha, k) = \sum_n c_n(k)N(\alpha; \eta_n(k), \gamma(k))
\]

\[
\sum_n c_n(k) = 1 \quad c_n(k) > 0
\]

and that because of computational considerations we cannot compute the density. Instead, we wish to devise a technique for recursively choosing \( 2N \) real numbers \( \{d_n(k)\}_{n=1}^N \), \( \{\mu_n(k)\}_{n=1}^N \) with

\[
\sum_{n=1}^N d_n(k) = 1 \quad d_n(k) > 0
\]

such that the density

\[
f(\alpha, k) = \sum_{n=1}^N d_n(k)N(\alpha; \mu_n(k), \gamma(k))
\]

"approximates" \( p(\alpha, k) \).

The motivation for making this type of approximation deserves some comment. First of all, we assume that the variance \( \gamma(k) \) of all the densities is the same. This is consistent, for example, with the conditional densities derived for the \( \text{mod} \ 2\pi \) and absolute value problems.
Also, for these problems we have the physically appealing interpretation of the form of $p(\alpha,k)$ as being computed by a bank of discrete-time linear filters that operate with the same variance $\gamma(k)$ but different observations ($\tilde{y}(k)$ in the absolute value case and $\tilde{v}(k)+2n\pi$ in the mod $2\pi$ case). The difficulty in computing $p(\alpha,k)$ arises from the fact that either the bank of filters is infinite or the number of filters grows exponentially. Thus, for practical application of these results, we need to truncate this bank of filters -- i.e., we wish to keep only $N$ filters. Thus we have the approximation form given in (C.50). We allow the possibility that $\{\eta_n\}_{n=1}^N$ need not be a subset of $\{\eta_n\}_{n=1}^\infty$, although cases in which $\{\eta_n\}_{n=1}^N \subseteq \{\eta_n\}_{n=1}^\infty$ will be considered. In addition, the restrictions on the $d_n$'s given in (C.49) are motivated by the fact that the $c_n$ have interpretations as probabilities (see Sections 4.4 and B.3) and the sum of the $c_n$'s is the certain event (i.e., $\sum c_n = 1$; also this is necessary so that $\int_{-\infty}^{\infty} p(\alpha,k)d\alpha = 1$). We shall devise schemes using this probabilistic interpretation, and, therefore, it is natural to make the assumption (C.49).

Therefore, our problem is not a simple curve fitting problem, since we cannot compute $p(\alpha,k)$. However, we wish to approximate $p(\alpha,k)$ by using its structure and physical interpretation. Thus, we are looking for sub-optimal filter banks, and, for estimation applications, a reasonable figure of merit to be used to compare the performance of these schemes is some measure of the estimation error. For this discussion, we will consider the problem of devising schemes for approximating the minimum variance estimate of $x(k)$ given that we have the storage constraints of (C.49) and (C.50). The unrealizable minimum variance estimate is
\[ \hat{x}(k) = \sum_{n} c_n(k) y_n(k) \]  

while the limited storage minimum variance estimate will be taken to be

\[ \tilde{x}(k) = \sum_{n=1}^{N} d_n(k) w_n(k) \]  

Thus a scheme will be considered "good" if it yields a small mean square error with \( \hat{x}(k) \) taken as the estimate. At the end of this section we will briefly discuss the problem of analyzing the performance of these schemes. We will not discuss the companion problem of folded normal densities and estimates that minimize criteria like \( \mathcal{E}[1-\cos(\theta-\tilde{\theta})] \), but we note that the approaches described here are applicable to that problem. Also, we will not present numerical results here but refer to those in [B10], which indicate that these approaches are reasonable. Clearly, some additional numerical work is necessary for the particular problems of interest in Sections 4.4 and C.2.

To make the following discussion more specific, we will consider the absolute value problem. Specifically, we wish to estimate a random variable \( x \), with a priori density \( N(\alpha; \eta_0(0), \gamma(0)) \), given the sequence of measurements \( \tilde{y}(k) \), \( k=1,2,... \)

\[ y(k) = x + v(k) \]  

\[ \tilde{y}(k) = |y(k)| \]  

where the \( v(k) \) are independent of each other and \( x \) and have distributions \( N(\nu; 0, r(k)) \). In this case \( p(\alpha, k) = p_x(\alpha, |\tilde{y}(i)|, i=1,...,k) \) can be computed recursively using (C.40) - (C.46). Thus \( p(\alpha, k) \) has \( 2^k \) terms,
and we wish to design a suboptimal filter that produces a density with only $N$ terms.

Before starting, we make several comments. As mentioned earlier, $p(\alpha, k)$ is still the sum of $2^k$ densities if $x$ is a random process arising from a linear discrete-time system driven by a white Gaussian sequence (see (C.36) and Section 4.4). In fact, as discussed in Section 4.4, the coefficients $c_n$ are unchanged by such a diffusion update (see (4.168)). Thus the techniques we will devise are also applicable if $x$ is defined by (C.36).

In addition, the techniques we will talk about are applicable to the mod $2\pi$ problem with some modifications. The basic difference between the absolute value and mod $2\pi$ problems is that $p(\alpha, k)$ has infinitely many terms in the mod $2\pi$ case. We will briefly discuss how we might adapt these techniques to the mod $2\pi$ problem after we have discussed the various methods.

The most obvious method of truncation is to choose the $N$ largest $c_n(k)$ at each stage. Call them $c_{n_1}(k), \ldots, c_{n_N}(k)$, and let

$$d_i(k) = \frac{c_{n_i}(k)}{\sum_{j=1}^{N} c_{n_j}(k)} \quad i = 1, \ldots, N$$  \hspace{1cm} (C.55)$$

$$\mu_i(k) = n_{n_i}(k)$$  \hspace{1cm} (C.56)$$

Then for the next stage, we assume that the density (C.50) is the a priori density and use (C.40) - (C.46) to compute an updated density, choose the $N$ largest $c_i(k+1)$ and repeat the process. We note that Buxbaum and Haddad refer to this approach as a "maximum likelihood"
approximation. In their case, where the $c_i(k)$ are not, in general, related to various hypotheses, Buxbaum and Naddad needed the quotes around the words \textit{maximum likelihood}; however, in our case, we are, in fact choosing the \textit{most likely} "paths" (sets of values \{y(j)\} given \{|y(j)|\}), excluding the paths discarded at preceding steps.

As mentioned in [B10], one has no guarantee that a term dropped at one stage would not have led to terms that would later have significant effects on the density shape. For example, referring to (C.44), the fact that $c_j(k) > c_i(k)$ does not mean that

$$c_j(k)N(\tilde{y}(k+1); \eta_j(k), \gamma(k)) > c_i(k)N(\tilde{y}(k+1); \eta_i(k), \gamma(k))$$

(C.57)

Thus, discarding $c_i$ in favor of $c_j$ may not be the best thing to do. An example of how this might happen is illustrated in Figure C.1.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure_c1.png}
\caption{Illustrating a Density Function Consisting of Three Terms}
\end{figure}

Suppose we have a three term distribution with the three peaks at $-\varepsilon$, 0, and $\varepsilon$ with the $\pm \varepsilon$ peaks having large $c_n$'s and the 0 peak a small one.
Then it is possible that the overall peak of the distribution will be at 0, and thus the sizes of the coefficients coming from the 0 peak term are likely to grow relative to the other terms. We see that the relative sizes of the various $c_n$ are highly sample path dependent. Therefore it does not appear possible to choose a priori which of the $c_n$ to keep (although any effective method of this type would be advantageous since it would eliminate the need for decision logic).

Returning to the maximum likelihood algorithm, we note that despite their reservations regarding the logic behind the scheme, Buxbaum and Haddad found that this method worked quite well even when the number of terms $N$ was fairly small. However, we wish to consider variations on the max likelihood approximation in order to avoid the problem of throwing away terms that are likely to be important later.

Referring to (C.57), we see that the relative size of the terms is related to the sizes of $N(\tilde{y}(k) + \eta_j(k), \gamma(k))$, and such a term is largest for the smallest value of $|\tilde{y}(k+1) - \eta_j(k)|$. Then, in addition to keeping some of the terms with the largest values of $c_j(k+1)$, we could also consider keeping some terms that correspond to the smallest values of $|\tilde{y}(k+1) - \eta_j(k)|$.

Another method employs the philosophy of using the present probability density to predict the sizes of future terms. Suppose we have processed the measurements $\tilde{y}(1), \ldots, \tilde{y}(k-1)$ to obtain $f(\alpha, k-1)$ given in (C.50), have observed $\tilde{y}(k)$ and computed $\{e^{\pm}_L(\tilde{y}(k))\}_{L=1}^N$ from (C.44) and (C.45). We wish to keep $N$ of these terms. Using the untruncated (at time $k$) density
(C.42) the measurement equations (C.53) and (C.54), and the density for \( v(k+1) \), we can compute the density \( f(\tilde{y}(k+1)|\tilde{y}(1),...,\tilde{y}(k)) \) (we use "\( f \)" instead of "\( p \)" to indicate that truncations have been performed at earlier steps).

\[
f(\tilde{y}(k+1)=z|\tilde{y}(1),...,\tilde{y}(k)) = \sum_{\ell=1}^{N} e^\ell_\ell(\tilde{y}(k))[N(z;\eta^\ell_\ell(k),\gamma(k)+r(k+1))
+ N(-z;\eta^-\ell_\ell(k),\gamma(k)+r(k+1))] + e^-\ell_\ell(\tilde{y}(k))[N(z;\eta^-\ell_\ell(k),\gamma(k)+r(k+1))
+ N(-z;\eta^-\ell_\ell(k),\gamma(k)+r(k+1))]
\tag{C.58}
\]

where here \( \eta^\ell_\ell(k) \) and \( e^\ell_\ell(k) \) are computed as in (C.43) - (C.45) with the \( d_\ell \) and \( \eta_\ell \) on the right hand sides replaced by the \( d_\ell(k-1) \) and \( \eta_\ell(k-1) \) of (C.50). Also, \( \beta \) is replaced by \( \tilde{y}(k) \).

Now given any \( c^j_j(k) \in \{e^\ell_\ell(y(k))\}_{\ell=1}^{N} \), the associated \( \eta^j_j(k) \), and an hypothesized value \( \xi \) of \( \tilde{y}(k+1) \), we can compute the updated coefficients \( c^\pm_j(k+1) \)

\[
c^\pm_j(k+1;\xi) = \frac{c^j_j(k)N(\pm\xi;\eta^j_j(k),\gamma(k)+r(k+1))}{z(\xi)} \tag{C.59}
\]

where \( z(\xi) \) is computed via (C.46) and is simply a normalizing constant. We wish to compute

\[
E(c^\pm_j(k+1)|\tilde{y}(1),...,\tilde{y}(k)) = \int_0^\infty c^\pm_j(k+1;\xi)f(\tilde{y}(k+1)=\xi|\tilde{y}(1),...,\tilde{y}(k))d\xi
\tag{C.60}
\]

Noting (C.59) and (C.46), we see that this is extremely complicated because of the denominator \( z(\xi) \). Thus, we might consider an ad hoc approach suggested by the fact that all of the \( c^j_j(k+1;\xi) \) have the same denominator. That is, suppose we drop the denominators and compute
\[ q^+(c_j(k), k) = c_j(k) \int_0^\infty N(\xi; \eta_j(k), \gamma(k) + r(k+1)) \mathcal{E}(\eta_j(k), \gamma(k) + r(k+1)) \, d\xi \]
\[ (C.61) \]

We note that \( q^+(c_j(k), k) > q^+(c_j(k), k) \) does not guarantee that \( \mathcal{E}(c_j(k), k) > \mathcal{E}(c_j(k), k) \), but \( \mathcal{E}(c_j(k), k) \) is computationally simpler than \( \mathcal{E}(c_j(k), k) \).

Looking at \( (C.58) \) and \( (C.61) \), we see that the typical term in \( (C.61) \) excluding constants is of the form
\[ \int_0^\infty N((-1)^{l_1} x; m_1, \gamma) N((-1)^{l_2} x; m_2, \gamma) \, d\xi = s(m_1, m_2, \gamma, r, s) \]
\[ (C.62) \]

This can be rearranged:
\[ s(m_1, m_2, \gamma, r, s) = N(m_1; (-1)^r, m_2, 2\gamma) t(0; \frac{(-1)^r m_1 + (-1)^r m_2}{2}, \gamma/2) \]
\[ (C.63) \]

where
\[ t(\alpha; \mu, \delta) = \int_0^\infty N(\xi; \mu, \delta) \, d\xi \]
\[ (C.64) \]

is related to the error function and can be tabulated. We now have
\[ q^+(c_j(k), k) = c_j(k) \sum_{l=1}^N e^+_l(\tilde{\gamma}(k))[s(\eta_j(k), l, \gamma(k) + r(k+1), 0, 0) + s(\eta_j(k), l, \gamma(k) + r(k+1), 0, 1)] + e_l(\tilde{\gamma}(k))[s(\eta_j(k), l, \gamma(k) + r(k+1), 0, 0) + s(\eta_j(k), l, \gamma(k) + r(k+1), 0, 1)] \]
\[ (C.65) \]

and \( q^- \) is found by changing from 0 to 1 the fourth arguments of each of the \( s(m_1, m_2, \gamma, r, s) \) terms that appears in \( (C.65) \).

We have the following algorithm:
(1) Given \( f(\alpha,k-1) \) (see (C.50)) and \( \tilde{y}(k) \), compute 
\[ \{ e_\pm^\pm(k), \eta_\pm^\pm(k) \}_{\pm=1}^N \] and \( \gamma(k) \) from (C.22) and (C.43) - (C.46).

(2) For each \( c_j^+(k) \in \{ e_\pm^\pm(k) \}_{\pm=1}^N \) compute \( q^+(c_j^+(k),k) \) and choose the \( N \) \( c_j(k) \) that yield the largest values of \( q^+(c_j(k),k) \).

(3) Using (C.55) and (C.56) we compute \( f(\alpha,k) \) and repeat the process.

This algorithm uses more of the information in \( f(\alpha,k) \) than the straight max likelihood scheme, since it attempts to predict what will happen one stage ahead. Of course, we can develop analogous algorithms that predict \( M \) stages ahead. However, the computational demands even in the one stage case are large, and, although it is likely that this method will perform better than the max likelihood algorithm (since it indicates which peaks truly have low probability), it is not clear that we should go to this trouble. Some numerical work is necessary here.

We now mention two other schemes discussed in [B10]. Suppose we process the first \( M \) measurements without truncation. Then half of the terms correspond to \( y(1) = \tilde{y}(1) \) and the other half to \( y(1) = -\tilde{y}(1) \).

If we add up the two sets of coefficients, the first sum equals 
\[ \Pr(y(1)=\tilde{y}(1)|\tilde{y}(1),...,\tilde{y}(M)) \] and the second is \( \Pr(y(1)=-\tilde{y}(1)|\tilde{y}(1),...,\tilde{y}(M)) \).

We then choose the set corresponding to the larger of the two probabilities and discard the other set. The process is repeated by processing \( \tilde{y}(M+1) \) and dividing the terms into the two groups \( y(2) = \tilde{y}(2) \) and \( y(2) = -\tilde{y}(2) \).

This scheme is called the "hypothesis test" approximation in [B10]. Simulations indicating that this is indeed a reasonable approach are reported in [B13].
The other method, called the "fusion" approximation, uses the stability of the discrete-time Kalman Bucy filter [J2], [D3]. It is known [J2], [B10] that the outputs of two such filters differing only in initial conditions and the first few inputs (measurements) converge. Thus, for our problem, consider processing the first $M$ measurements without truncation and grouping the terms that correspond to paths that are the same over the last $K < M$ stages (i.e. that make the same choices $y(l) = \pm \tilde{y}(l)$, $l = M-K+1, \ldots, M$). "Fuse" each group into a single normal density with mean

$$\mu = \sum_{n \in \text{particular group}} c_n(M) \eta_n(M) \quad (C.66)$$

and coefficient

$$d = \sum_{n \in \text{particular group}} c_n(M) \quad (C.67)$$

At each subsequent stage we repeat this fusion of paths that are the same over the last $K$ steps. Again, successful simulations are reported in [B13].

We note that these approaches can also be applied to the mod $2\pi$ equivalence class problem. However, in that case, since the sums are all infinite, we must a priori chop off the tails of the series. That is, we must assume that only a finite number of terms are nonzero since we can only compute a finite number of them. Precisely, we must assume that for any value of $\tilde{y}$, there are only a finite number of values of $n$ such that $\Pr(y = \tilde{y} + 2\pi n | \tilde{y}) > 0$ (note that in the absolute value problem there are only 2 possible values of $y$ that lead to any given value of $|y|$).
Having made this a priori assumption, we can apply the techniques just discussed to the mod 2π problem. Note that this "chopping off" of the tails of the series is not too unreasonable, since one can show that for large N \( \Pr(y = \bar{y} + 2N\pi | \bar{y}) \) goes as \( e^{-N^2} \) (see Section B.4).

Finally, we will discuss the analysis of the performance of such schemes. As mentioned earlier, we will take the error variance as our figure of merit for comparing schemes. The analytical computation of the error variance involves \( p(\alpha, k) \), and, since we are assuming that we cannot compute this, it appears that the only feasible scheme for computing the error variance is via simulation. For completeness, we will derive the exact equation for the conditional mean squared error for an arbitrary approximation scheme. Let \( \hat{x}(k) \) be the exact conditional mean given in (C.51), and let \( \bar{x}(k) \) be the approximate mean given by (C.52). Then the exact conditional mean squared error for the estimate \( \tilde{x}(k) \) is

\[
P(k) = E[(x(k) - \tilde{x}(k))^2] = \int_{-\infty}^{\infty} (\alpha - \bar{x}(k))^2 p(\alpha, k) d\alpha \tag{C.68}
\]

We can simplify (C.68)

\[
P(k) = [\bar{x}(k)]^2 - 2\bar{x}(k)\bar{x}(k) + E(x^2(k)) = \bar{x}(k)[\bar{x}(k) - 2\bar{x}(k)] + \gamma(k) + \sum_{n} c_n(k) n^2(k) \tag{C.69}
\]

The last two terms are independent of the choice of approximation, so a figure of merit for evaluating approximation schemes is

\[
R(k) = \bar{x}(k)[\bar{x}(k) - 2\bar{x}(k)] \tag{C.70}
\]

and the minimum value of \( R(k) \) occurs when \( \bar{x}(k) = \tilde{x}(k) \).
\[ R(k) \geq R_{\text{min}}(k) = -\bar{x}^2(k) \] (C.71)

Note that just prior to the truncation at each stage, we have more than \( N \) terms (2N terms in the absolute value problem), and it makes sense to use all of these terms to compute \( \bar{x}(k) \).

An interesting idea for off-line analysis of various approximation techniques is to use an approximation scheme that retains a large number of terms and is too large for real-time use to approximate \( \hat{x}(k) \) and \( R(k) \) for the various simpler schemes that are being considered for on-line use. Also, it would be interesting to study the convergence properties of these schemes as the number of terms increases. Such a study would provide a comparison of the various methods in terms of accuracy and the number of terms required to achieve a desired level of performance.
APPENDIX D

MOMENT TRUNCATION METHODS WITH APPLICATIONS TO PHASE TRACKING PROBLEMS

"She cut off their tails with a carving knife"

- Mother Goose [M17]

In Chapter 5 we devised some techniques for solving phase tracking problems using Fourier analysis. The optimal solutions derived in that chapter are physically unrealizable, since the filters are infinite dimensional. In this appendix we will briefly discuss a few methods for truncating the infinite sets of moment equations encountered in Chapter 5. We will specifically treat the continuous time problems of Section 5.3; however, these techniques are also applicable to the discrete time problems of Section 5.2. In addition, as discussed in [L11], some of these techniques are quite general and can be applied to large classes of nonlinear estimation problems. In this appendix we only give a few examples of the types of approximations that can be used. The reader is referred to [L11], [J2], and [B9] for more thorough treatments of numerical methods in filtering theory.

We first consider the pure phase tracking problem treated in Example 5.6. Specifically we wish to truncate the infinite sequence of coupled differential equations for the Fourier coefficients $a_n$ and $b_n$ (see equations (5.50) and (5.51) and Figure 5.5). In this case we wish to approximate the $S^1$ density

$$p(\theta, t) = \frac{1}{2\pi} + \sum_{n=1}^{\infty} a_n(t) \sin n\theta + b_n(t) \cos n\theta$$

(D.1)

by a finite set of parameters. As mentioned in Example 5.6, a natural
The approximation is
\[ \tilde{p}(\theta, t) = \frac{1}{2\pi} + \sum_{n=1}^{N} a_n(t) \sin n\theta + b_n(t) \cos n\theta \]  \hspace{1cm} (D.2)

- i.e. just assume \( b_m(t) = a_m(t) = 0 \) \( \forall \ m > N \). As discussed in [Kl0], the Fourier coefficients usually fall off at least as \( 1/n^2 \), and thus, for large enough \( N \), this straightforward truncation method may work quite well. Of course one needs some numerical results to determine how many terms are enough. The reader is referred to Chapter 6 for a discussion that indicates that this straightforward method is not as good as it first appears. See also Appendix C for some analysis on the order of the errors introduced by this approximation method.

Another truncation scheme mentioned in Example 5.6 involves the use of a low-pass filter. The phase to be tracked is
\[ \theta(t) = \omega_c t + \int_0^t q^{1/2}(s) dv(s) + \theta_0 \]  \hspace{1cm} (D.3)

and thus the Fourier coefficients look like
\[ a_n(t) = \frac{1}{\pi} \sin(n\omega_c t + \alpha_n(t)) \quad b_n(t) = \frac{1}{\pi} \cos(n\omega_c t + \beta_n(t)) \]  \hspace{1cm} (D.4)

Thus, if we use a low pass filter that allows terms with frequencies \( \leq N\omega_c \) to pass unamplified and not phase shifted, but chops off all frequencies \( \geq (N+1)\omega_c \), we can effectively truncate the bank of filters depicted in Figure 5.5. Consider the term in the differential equation for \( a_N \) that causes the difficulty -- i.e. that contains the coupling to the higher modes (we can consider the \( b_N \) equation in an analogous manner). Specifically consider the term
\[ c_N(t) = \frac{1}{2} \left( b_{N-1}(t) - b_{N+1}(t) \right) - \pi a_N(t) a_1(t) \]  \hspace{1cm} (D.5)

Using (D.4) we compute

\[ a_N(t) a_1(t) = \frac{1}{\pi^2} \sin(N \omega_c t + \alpha_N(t)) \sin(\omega_c t + \alpha_1(t)) \]

\[ = \frac{1}{2\pi^2} \left[ \cos((N-1)\omega_c t + \alpha_N(t) - \alpha_1(t)) - \cos((N+1)\omega_c t + \alpha_N(t) + \alpha_1(t)) \right] \]  \hspace{1cm} (D.6)

Thus, if we actually had the signal \( c_N(t) \) and put it through an ideal low-pass filter of the desired type, the output would be

\[ \tilde{c}_N(t) = \frac{1}{2} \left[ b_{N-1}(t) - \frac{1}{\pi} \cos((N-1)\omega_c t + \alpha_N(t) - \alpha_1(t)) \right] \]  \hspace{1cm} (D.7)

Note that the higher order coefficient \( b_{N+1}(t) \) does not appear in \( \tilde{c}_N \) and that the effect of the low pass filter on \( c_N \) is precisely the same as if we had assumed that

\[ \beta_{N+1}(t) = [\alpha_N(t) + \alpha_1(t)] \mod 2\pi \]  \hspace{1cm} (D.8)

which is, in fact, true if we are tracking perfectly. Thus, if we are tracking well, (D.8) may be a reasonable assumption. It is this argument that provides the basic motivation for this approach.

We note that since we are attempting to devise an approximation that does not involve the explicit computation of \( b_{N+1}(t) \) (i.e. we want to use only the first \( N \) Fourier coefficient filters), we cannot compute \( c_N(t) \).

However, by low-pass filtering \( \pi a_N a_1 \) and subtracting the output from \( \frac{1}{2} b_{N-1} \), we can compute \( \tilde{c}_N \) without knowing \( c_N \). We also remark that we can avoid the low-pass filtering of \( \pi a_N a_1 \) in the following manner: given \( a_N \) and \( a_1 \), we compute \( \alpha_N \) and \( \alpha_1 \) by comparing the phases of \( a_N \) and \( a_1 \) with the phases of \( \sin N \omega_c t \) and \( \sin \omega_c t \) (see Chapter 6, in which we discuss the
computation of total phase from the observation of a sinusoidal signal with the aid of cycle counters and inverse trigonometric functions). Having $\alpha_{N}$ and $\alpha_{1}$ and connecting up the (N-1) filter, we can easily compute $\tilde{c}_{N}$. Finally, we note that only experimental results will determine the usefulness of this approximation.

We now discuss several examples of what has been called the "assumed density" type of approximation (see [I1.1] and [K19]). The basic idea of the approximation is the following: we assume that the conditional density has some known form that is specified by a finite set of parameters; then, having $\{a_{n}, b_{n}\}_{n=1}^{N}$, we compute the assumed density parameters and the associated values of $a_{N+1}$ and $b_{N+1}$. For example, for the assumed density form (D.2), we have $a_{N+1} = b_{N+1} = 0$. A slightly more complicated example involves the assumption that $p(\theta, t)$ is a folded normal density

$$
\tilde{p}(\theta, t) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} e^{-n^{2} \gamma(t)/2} \left[ \sin n\eta(t) \sin n\theta + \cos n\eta(t) \cos n\theta \right]
$$

(D.9)

In this case, if we compute $a_{1}(t)$ and $b_{1}(t)$, assumption (D.9) implies

$$
e^{-\gamma(t)} = \pi^{2} \left[ a_{1}^{2}(t) + b_{1}^{2}(t) \right]
$$

(D.10)

$$
\sin \eta(t) = \frac{a_{1}(t)}{[a_{1}^{2}(t) + b_{1}^{2}(t)]^{1/2}} \quad \cos \eta(t) = \frac{b_{1}(t)}{[a_{1}^{2}(t) + b_{1}^{2}(t)]^{1/2}}
$$

(D.11)

Having made these computations, we can then compute

$$
a_{2}(t) = \frac{1}{\pi} e^{-2\gamma(t)} \sin 2\eta(t) = \frac{2}{\pi} \left[ e^{-\gamma(t)} \right]^{2} \sin \eta(t) \cos \eta(t)
$$

(D.12)

$$
b_{2}(t) = \frac{1}{\pi} e^{-2\gamma(t)} \cos 2\eta(t) = \frac{1}{\pi} \left[ e^{-\gamma(t)} \right]^{2} \left[ \cos^{2} \eta(t) - \sin^{2} \eta(t) \right]
$$

(D.13)

and truncate the filter equations.
As mentioned in Section 3.3, the set of densities of the form

\[ p(\theta) = \sum_{n=1}^{M} c_n F(\theta; \gamma_n, \eta_n) \]  \hspace{1cm} (D.14)

is dense in \( L^1(-\pi, \pi) \), and thus assuming that \( p(\theta, t) \) is of the form
given in (D.14) would seem to be a reasonable idea. We then must compute
\( \{ \gamma_n, \eta_n \}_{n=1}^{M} \) in terms of the Fourier coefficients \( \{ a_n, b_n \}_{n=1}^{N} \). This is
not an easy problem (except when \( M = 1 \)) since these parameters are
related by the equations

\[ a_k = \frac{1}{\pi} \sum_{n=1}^{M} c_n e^{-k^2 \gamma_n/2} \sin kn \eta_n \]  \hspace{1cm} (D.15)

\[ b_k = \frac{1}{\pi} \sum_{n=1}^{M} c_n e^{-k^2 \gamma_n/2} \cos kn \eta_n \]  \hspace{1cm} (D.16)

Even if we take \( \gamma_n = \gamma \psi n \), it is not clear how to solve these
equations in closed form for \( M > 1 \).

As an approximation, we can consider taking at least one
Newton-Raphson step [W8] to approximate the solution to (D.15) and (D.16).
That is, let \( y \) be the vector of parameters \( \{ \gamma_n, \eta_n \}_{n=1}^{M} \) (or \( e^{-\gamma_n} \) and
\( \cos \eta_n, \sin \eta_n \)), and let \( a \) and \( b \) be vectors of the Fourier coefficients.
Equations (D.15) and (D.16) can then be written as

\[ f(y; a, b) = 0 \]  \hspace{1cm} (D.17)

Suppose we have an initial guess \( y_0 \). Then the Newton-Raphson step
computes \( y_1 = y_0 + \Delta y \) by solving (D.17) "to first order" -- i.e., we
write

\[ f(y_0 + \Delta y; a, b) = f(y_0; a, b) + \frac{\partial f}{\partial y}(y_0; a, b) \Delta y \]  \hspace{1cm} (D.18)

and solve
\[ \Delta y = - \left[ \frac{\partial^2}{\partial y} (v_0; a, b) \right]^{-1} f(y_0; a, b) \]  

(D.19)

An example where we might use this occurs in the discrete measurement problem. Suppose \( \theta(t) \) is an \( S^1 \) Brownian motion process and we take measurements \( y_k \) at the time instants \( kA \). Let \( \gamma_n(k|k) \), \( \eta_n(k|k) \), \( a_n(k|k) \), and \( b_n(k|k) \) be the parameters of \( p(\theta, kA|k) \) where we have processed the first \( k \) measurements. The propagation of these coefficients just prior to the observation of \( y_{k+1} \) is performed as in (5.25) - (5.29).

In this case, \( a_n(k+1|k) \) and \( b_n(k+1|k) \) are computed as in (5.29), and we also have

\[ \gamma_n(k+1|k) = \gamma_n(k|k) + \Delta \quad \eta_n(k+1|k) = \eta_n(k|k) \]  

(D.20)

Then, having computed \( a_n(k+1|k+1) \) and \( b_n(k+1|k+1) \), we must compute \( \gamma_n(k+1|k+1) \) and \( \eta_n(k+1|k+1) \). A reasonable assumption would be to use (D.19) using \( \gamma_n(k+1|k) \) and \( \eta_n(k+1|k) \) as initial guesses.

There are a number of other approximate techniques involving folded normal densities. We will mention only one of these. Suppose we compute \( \{a_n, b_n\}_{n=1}^N \), and fix \( M \) points \( \eta_1, \ldots, \eta_M \) around the circle. We can then consider approximating \( p(\theta,t) \) by

\[ \hat{p}(\theta,t) = \sum_{n=1}^M c_n(t) F(\theta; \eta_n, \gamma(t)) \]  

(D.21)

- e.g. choose \( c_n(t) \) and \( q(t) \) to minimize the least squares criterion

\[ \sum_{i=1}^N \left\{ \sum_{j=1}^M c_j(t) \sin \eta_j \right\}^2 + \sum_{j=1}^M c_j(t) \cos \eta_j \right\}^2 \]  

(D.22)

where \( q(t) = e^{-\gamma(t)/2} \) and \( d_i, e_i > 0 \). This problem is still somewhat
complex, but is more tractable than solving (D.15) and (D.16).

Also, if we choose \( \gamma(t) \) a priori, (D.22) is a linear least squares problem in the \( c_j(t) \). Thus, some numerical work may yield a reasonable method for choosing the time function \( \gamma(t) \), and then the truncation procedure (i.e. the solution for the \( c_j \)) is easily implemented.

Finally, one could consider approximating \( p(\theta,t) \) by a uniform density

\[
p(\theta) = \begin{cases} 
\frac{1}{\theta_b - \theta_a} & -\pi \leq \theta_a < \theta < \theta_b \leq \pi \\
0 & \text{otherwise}
\end{cases}
\]  

We will not discuss this here but will discuss an \( R^1 \) analog of this approximation technique later in this appendix.

We now consider some \( R^1 \) approximation techniques. As discussed in Examples 5.7 and 5.8, in demodulating sinusoidal signals we have a somewhat more complex situation than in the phase tracking problem, since we must estimate functions like \( x^n \sin mx \). Again to be specific, suppose we consider the phase demodulation problem of Example 5.7. In this case, we wish to truncate the coupled set of equations for \( a_{nm}(t) \) and \( b_{nm}(t) \) (see (5.58) - (5.62) and Figure 5.7). Suppose we compute only \( a_{nm} \) and \( b_{nm} \) for \( n = 1, \ldots, N \) and \( m=1, \ldots, M \). From (5.58) and (5.59), we see that to truncate the filter equations effectively, we must compute approximations for \( \{b_{N+1,m}, a_{N+1,m} \}_{m=1}^M \) and \( \{b_{n,M+1}, a_{n,M+1} \}_{n=1}^N \). In some sense this truncation problem is more difficult than the \( S^1 \) problem discussed previously, since, for example, the moments

\[
b_{no}(t) = \frac{1}{\pi} \mathbb{E}[x^N(t)|z(s), 0 \leq s \leq t]
\]  

(D.24)
do not necessarily go to zero for large \( n \) (if \( |x| \leq 1-\varepsilon \) probability 1 for some \( 0 < \varepsilon < 1 \), then \( b_{nm} \) does go to zero as \( n \to \infty \), and we can truncate the moments by neglecting all \( b_{nm} \) for \( n > N \)).

We first note that we can come up with truncation techniques analogous to those for the \( S^1 \) problem by replacing folded normal densities with normal densities (note that finite linear combinations of normal densities are dense in \( L'(-\infty, \infty) \), [L7], [A2]). We will not discuss this type of approximation here except to note that for such techniques we need to know the form of \( \mathcal{E}(x^n \cos mx) \) if \( x \) has the density \( N(x; \eta, \gamma) \).

For such a random variable \( x \), let

\[
\begin{align*}
p_{nm} &= \mathcal{E}(x^n \cos mx) \quad (D.25) \\
q_{nm} &= \mathcal{E}(x^n \sin mx) \quad (D.26)
\end{align*}
\]

We can then show that for all \( m \geq 0 \), we have the following equations for \( p_{nm} \) and \( q_{nm} \)

\[
\begin{align*}
p_{0m} &= e^{-m^2 \gamma/2} \cos m\eta \\
q_{0m} &= e^{-m^2 \gamma/2} \sin m\eta \quad (D.27) \\
p_{lm} &= \eta p_{0m} - \gamma \eta q_{0m} \\
q_{lm} &= \eta q_{0m} + \gamma \eta p_{0m} \quad (D.28)
\end{align*}
\]

\[
\begin{align*}
p_{nm} &= \eta p_{n-1,m} + \gamma [(n-1)p_{n-2,m} - mq_{n-1,m}] \quad (D.29) \\
q_{nm} &= \eta q_{n-1,m} + \gamma [(n-1)q_{n-2,m} + mp_{n-1,m}] \quad (D.30)
\end{align*}
\]

where (D.29) and (D.30) hold for \( n \geq 2 \). Note that \( p_{00} = \mathcal{E}(x^n) \).

We will briefly discuss two other truncation methods. The reader is referred to [J2], [B9], [S13], and [L11] for a variety of other techniques, including those using Hermite polynomials and quasi-moments.
We now consider a simple assumed-form density technique and a variation of it. In this method we use only $b_{10}(t)$ and $b_{20}(t)$ to compute approximations for $\{b_{N+1,m}(t), a_{N+1,m}(t)\}_{m=1}^M$ and $\{b_{n,N+1}(t), a_{n,N+1}(t)\}_{n=1}^N$. As discussed in [L11], we assume that $\tilde{p}(x,t)$, our approximation to the conditional density $p(x,t)$, is the uniform density

$$p(x,t) = U(x; \eta(t), \sigma(t)) = \begin{cases} \frac{1}{2\sigma(t)} & \eta(t) - \sigma(t) \leq x \leq \eta(t) + \sigma(t) \\ 0 & \text{otherwise} \end{cases} \tag{D.31}$$

where

$$\eta(t) = \pi b_{10}(t) \tag{D.32}$$

$$\sigma(t) = [\pi b_{20}(t) - \eta^2(t)]^{1/2} \tag{D.33}$$

As with the normal density, we can compute $d_{nm} = \mathcal{E}(x^n \cos mx)$ and $e_{nm} = \mathcal{E}(x^n \sin mx)$ if $x$ has the density $U(x; \eta, \sigma)$.

$$d_{n0} = \frac{(\eta+\sigma)^{n+1} - (\eta-\sigma)^{n+1}}{2\sigma(n+1)} \quad e_{n0} = 0 \tag{D.34}$$

and for $n \geq 0$ and $m \geq 1$

$$d_{nm} = \frac{1}{2m\sigma} [(\eta+\sigma)^n \sin m(\eta+\sigma) - (\eta-\sigma)^n \sin m(\eta-\sigma)] - \frac{n}{m} e_{n-1,m} \tag{D.35}$$

$$e_{nm} = \frac{1}{2m\sigma} [(\eta-\sigma)^n \cos m(\eta-\sigma) - (\eta+\sigma)^n \cos m(\eta+\sigma)] + \frac{n}{m} d_{n-1,m} \tag{D.36}$$

We can use these equations to truncate the phase demodulation equations (5.58) and (5.59), where we note that from (5.56) and (5.57)
\[ a_{nm}(t) = \frac{1}{\pi} \cos m\omega_c t \mathcal{E}[x^n(t) \sin mx(t) | z(s), 0 \leq s \leq t] \]
\[ + \frac{1}{\pi} \sin m\omega_c t \mathcal{E}[x^n(t) \cos mx(t) | z(s), 0 \leq s \leq t] \quad (D.37) \]

\[ b_{nm}(t) = \frac{1}{\pi} \cos m\omega_c t \mathcal{E}[x^n(t) \cos mx(t) | z(s), 0 \leq s \leq t] \]
\[ + \frac{1}{\pi} \sin m\omega_c t \mathcal{E}[x^n(t) \sin mx(t) | z(s), 0 \leq s \leq t] \quad (D.38) \]

An obvious variation of this involves the approximation of the conditional density \( p(x,t) \) by

\[ \tilde{p}(x,t) = N(x; \pi b_{10}(t), \pi b_{20}(t) - \pi^2 b_{10}(t)) \quad (D.39) \]

In this case equations (D.27) - (D.30) can be used to truncate the phase demodulation equations.

Another technique involves the notion of cumulants of a probability density [S13]. Let \( x \) be a real-valued random variable with density \( p(x) \), and let \( \Theta_x(u) \) be the characteristic function, [S13], of \( x \)

\[ \Theta_x(u) = \mathcal{E}(e^{iux}) = \int_{-\infty}^{+\infty} e^{iux} p(x) dx \quad (D.40) \]

If we write

\[ \Theta_x(u) = \exp\left\{ \sum_{n=1}^{\infty} \frac{(iu)^n}{n!} k_n \right\} \quad (D.41) \]

the \( k_n \) are called the cumulants of \( x \). The cumulants are related to the moments

\[ m_n = \mathcal{E}(x^n) \quad (D.42) \]

by the formulae

\[ k_1 = m_1 \]
\[ k_2 = m_2 - m_1^2 \]
\[ k_3 = m_3 - 3m_1 m_2 + 2m_1^3 \]
\[ k_4 = m_4 - 3m_1 m_2^2 - 4m_1 m_3 + 12m_1^2 m_2 - 6m_1^4 \]
\[ \vdots \quad (D.43) \]
As mentioned earlier, we cannot assume the $\mathcal{C}(x^n)$ tends to 0 as $n$ approaches $\infty$; however, as discussed in [SL3], assuming that the cumulants tend to zero is a reasonable assumption. Thus suppose we compute $\{b_n(t)\}_{n=1}^N$ and assume $k_n(t) = 0 \quad \forall \ n > N$. We can then use (D.43) to compute $k_n(t)$, $n \in N$ (where we note that $b_{n0}(t) = \frac{1}{\pi} m_n(t)$).

Using this approximation

$$\tilde{\rho}_x(u, t) = \exp \left\{ \sum_{n=1}^{N} \frac{(iu)^n}{n!} k_n(t) \right\} \quad \text{(D.44)}$$

we can compute

$$\tilde{p}(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iux} \tilde{\rho}_x(u, t) du \quad \text{(D.45)}$$

and the appropriate expectations.
APPENDIX E

SOME RESULTS FROM ALGEBRAIC SYSTEM THEORY

"Faith in machinery is our besetting danger; often in machinery most absurdly disproportional to the end which this machinery, if it is to do any good at all, is to serve; but always in machinery, as if it had a value in and for itself."

- Matthew Arnold [Al6]

In this appendix we will review some of the central concepts from automata theory and abstract realization theory. The purpose is to provide some background material for the algebraic systems problems considered in Part II of this manuscript. We use Kalman, Falb, and Arbib [K11] as a primary reference. The reader is also referred to Arbib [A3], [A4], Hartmanis and Stearns [H11], Nerode [N2], Myhill [M12], and Zeiger [Z1].

Definition E.1: An automaton or machine is a quintuple $M = (U, \mathcal{Y}, X, \lambda, \delta)$ where

$U = \text{a finite input set}$

$\mathcal{Y} = \text{a finite output set}$

$X = \text{the set of states (not necessarily finite)}$

$\lambda : X \times U \to X = \text{the next state function}$

$\delta : X \to \mathcal{Y} = \text{the next output function}$

The automaton $M$ is called a finite state machine if $X$ is a finite set.

This definition is to be interpreted as defining a sequential machine

$$x_{k+1} = \lambda(x_k, u_k)$$  \hspace{1cm} (E.1)

$$y_k = \delta(x_k)$$  \hspace{1cm} (E.2)
We note that a more general definition of an automaton [K11] allows \( \delta \) to be an explicit function of \( u_k \). Also, by considering sequences of input strings, we can introduce some other very important concepts. We first need to define the input string monoid.

**Definition E.2**: Let \( U \) be a set. Then the string monoid \( U^* \) of \( U \) consists of the set of all strings of elements of \( U \) plus the empty string, \( \Lambda \)

\[
U^* = \{u_1, u_2, \ldots, u_k | u_i \in U, k \geq 1\} \cup \{\Lambda\} \tag{E.3}
\]

with the binary operation "\( \circ \)" of concatenation -- i.e. if \( \omega = u_1, \ldots, u_k \) and \( \omega = v_1, \ldots, v_\ell \) are two elements of \( U^* \), then

\[
\omega_1 \circ \omega_2 = u_1, \ldots, u_k, v_1, \ldots, v_\ell \in U^* \tag{E.4}
\]

The **length** \( |\omega| \) of a string \( \omega \in U^* \) is the number of elements in the string. Also, the string \( \omega = u_1, \ldots, u_\ell \) is to be interpreted as "first" we apply \( u_1 \), then \( u_2 \), etc.

Clearly \( \Lambda \) is the identity of \( U^* \)

\[
\omega \circ \Lambda = \Lambda \circ \omega = \omega \quad \forall \omega \in U^* \tag{E.5}
\]

Also, using the definition of \( U^* \) and system equations (E.1) and (E.2), we can write down equations for the manner in which the machine \( M \) processes any sequence \( \omega \in U^* \). We can define extended versions of \( \lambda \) and \( \delta \), \( \lambda : X \times U^* \to X \) and \( \delta : X \times U^* \to Y \) by

\[
\lambda(x, \omega) = \text{the state after the application of the string } \omega, \text{ starting in state } x \tag{E.6}
\]

\[
\delta(x, \omega) = \text{the output after the application of the string } \omega, \text{ starting in state } x \tag{E.7}
\]
For \( \omega \in U^* \) of length one, we define \( \lambda(x, \omega) \) as before. Also, given a string \( u_1, \ldots, u_k \) and initial state \( x \), the state after the application of this string is the same as the state after the application of \( u_{m+1}, \ldots, u_k \) starting in the state \( \lambda(x, u_1, \ldots, u_m) \). That is, we have the equality

\[
\lambda(x, \omega_1 \circ \omega_2) = \lambda(\lambda(x, \omega_1), \omega_2)
\] (E.8)

Thus, we can compute \( \lambda(x, u_1, \ldots, u_k) \) recursively using the definition of \( \lambda \) on strings of length one. Also, by convention we set \( \lambda(x, \lambda) = x \), and we note that (E.7) implies

\[
\delta(x, \omega) = \delta(\lambda(x, \omega))
\] (E.9)

where \( \delta \) on the right-hand side of (E.9) is the same as that of equation (E.2).

Suppose that we have a distinguished initial state \( x_0 \in X \). Then we can define the input-output response function \( f_{x_0} : U^* \to Y \)

\[
f_{x_0}(\omega) = \delta(x_0, \omega) = \delta(\lambda(x_0, \omega))
\] (E.10)

Note that specifying \( f \) is much like specifying the weighting pattern \([B2]\) of a linear system (in that case the distinguished initial state is 0).

We wish now to invert this procedure. Specifically, we wish to consider the synthesis problem, the design of a system that has a desired input-output behavior. Given sets \( U \) and \( Y \) and a function \( f : U^* \to Y \), find a (recursive) realization of \( f \) -- i.e. find a state set \( X \) with some distinguished element \( x_0 \) and maps \( \lambda : X \times U \to X \) and \( \delta : X \to Y \) such that when we extend the definition of \( \lambda \) and \( \delta \) to strings in \( U^* \) (see (E.6) and (E.7)) we have

\[
f(\omega) = f_{x_0}(\omega) = \delta(x_0, \omega) \quad \forall \omega \in U^*
\] (E.11)
This automaton with the particular initial state is said to realize the input-output map \( f \).

Of course we are not usually interested in finding just any realization but rather a realization that is in some sense "canonical." A logical definition of a "canonical realization" is one that has no "extraneous" states -- i.e. no state that can't be reached from the initial state and also no two states which if taken as initial states yield the same subsequent input-output behavior. These concepts are clearly related to the abstract notions of controllability and observability (see Section 8.3). Another possible definition of canonical realization is one that may include a few extraneous states but that has some desirable structural properties. We will discuss realizations of both types here and refer the reader to Chapter 8, in which realization theory is discussed in a much more specific setting.

We will first describe a canonical realization of the "non-extraneous state"-type. The concept we will talk about is called Nerode equivalence [N2], [Kl]. Given a function \( f : U^* \rightarrow Y \), we define an equivalence relation "\( \sim_f \)" on \( U^* \):

\[ \omega \sim_f \omega' \iff f(\omega \circ \mu) = f(\omega' \circ \mu) \quad \forall \mu \in U^* \quad (E.12) \]

That is, two input sequences are equivalent if the future output of the system for any subsequent input string \( \mu \) is the same if we start with initial string \( \omega \) or \( \omega' \). This definition is very closely connected with the concept of "state." Intuitively, the state of a system is a set of information about the past history of the system's behavior that is sufficient to determine the future behavior when we are given the future
inputs. In this case, let our system be described by the input-output function
f : U* → Y. Suppose we are told that the input string ω has been applied to the system. Then given any future input string u, we know that the future output is f(ω ⊙ u). Thus, knowledge of ω could be taken as the system state.

However, this definition of state contains extraneous information. That is, from (E.12) if ω ϑf ω', then the future behavior of the system is the same from either "state." Thus, all we really need know is the Nerode equivalence class [ω]N of the past input history, since any two strings in this equivalence class satisfy (E.12). Also, one can show that if we know less information about the past inputs than their Nerode class, we cannot determine the future behavior of the system. Therefore, in some sense, the set of Nerode equivalence classes, Xf, is a minimal state set (i.e. provides enough information to predict the future, given future inputs, but no more information).

Using the set Xf as the state set, we can write the recursive equations for a realization of f:

\[ \lambda_f([ω]_N, u) = [ω ⊙ u]_N \quad u \in U \]  \hspace{1cm} (E.13)

\[ \delta_f([ω]_N) = f(ω) \]  \hspace{1cm} (E.14)

Note that these are well defined -- i.e. if ω ϑf ω', one can see from (E.12) that ω ⊙ u ϑf ω' ⊙ u. Also, if ω, ω' ∈ [ω]_N, then

f(ω) = f(ω ⊙ Λ) = f(ω' ⊙ Λ) = f(ω'). The initial state for the realization is the equivalence class [Λ]_N. Then, we can extend (E.13) and (E.14) to strings
\[ \lambda_f([\omega]_N, \omega') = [\omega \circ \omega']_N \]  
(E.15)

\[ \delta_f([\omega]_N, \omega') = \delta_f(\lambda_f([\omega]_N, \omega')) \quad \omega' \in U^* \]  
(E.16)

and

\[ \delta_f([\Lambda]_N, \omega) = \delta_f(\lambda_f([\Lambda]_N, \omega)) = \delta_f([\Lambda \circ \omega]_N) = \delta_f([\omega]_N) = f(\omega) \]  
(E.17)

Thus this automaton does realize \( f \). Also, given any equivalence class \([\omega]_N\),

\[ \lambda([\Lambda]_N, \omega) = [\omega]_N \]  
(E.18)

so all states are reachable from the initial state. In addition, if \([\omega]_N \neq [\omega']_N\), then there exists a \( \mu \in U^* \) such that

\[ f(\omega \circ \mu) = \delta_f([\omega]_N, \mu) \neq \delta_f([\omega']_N, \mu) = f(\omega' \circ \mu) \]  
(E.19)

and we see that distinct states are distinguishable (see Definition 8.1).

Therefore, this realization is canonical in the sense of having no extraneous states.

Of course this realization is quite abstract, but the above arguments indicate that any other realization must have at least as many states as the Nerode realization (see \([K11]\)). Thus a map \( f : U^* \rightarrow Y \) is realizable by a finite state machine if and only if the number of distinct Nerode equivalence classes is finite. Using the Nerode realization as a guide, we can determine how "efficient" various realizations are. See Section 8.3 for a discussion of concrete realizations that have the same state set size as the Nerode realization. In that section we discuss abstract realizations in a slightly different setting (we look at strings of outputs.
as opposed to just the present output value), but the end result is the same.

As mentioned earlier, a second class of canonical realizations concerns itself more with the structural properties of the system. The motivation is to make a sacrifice in storage requirements (increase the number of states) in order to gain some desired system properties. For example, in Section 8.3 we see that the Nerode-minimal realization of a FGHSS need not be a FGHSS. Thus, if we are interested in working strictly within the class of FGHSS's, we might define a canonical realization as the FGHSS that realizes the given input-output map with the smallest number of states. This idea has also been considered by Arbib [A5], [A6], and Arbib and Manes [A7], [A8].

Another useful canonical realization is that provided by considering the state transition functions for a system. Suppose we have a system described by the recursive equations (E.1) and (E.2). Given an input string $\omega$ we can define the state transition function $\phi_\omega$ that tells us what the state after the application of $\omega$ will be for any initial state:

$$\phi_\omega(x) \triangleq \lambda(x, \omega) \quad (E.20)$$

By the concatenation property (E.8)

$$\phi_{\omega_1 \circ \omega_2}(x) = \lambda(\lambda(x, \omega_1), \omega_2) = \lambda(\phi_{\omega_1}(x), \omega_2) = \phi_{\omega_2}(\phi_{\omega_1}(x)) \quad (E.21)$$

That is

$$\phi_{\omega_1 \circ \omega_2} = \phi_{\omega_2} \circ \phi_{\omega_1} \quad (E.22)$$

where the binary operation on the right is ordinary function composition.
Let \( F(X, X) \) be the set of all functions from \( X \) into \( X \). This is a semigroup with function composition as the semigroup operation. It is, in fact, a monoid since the identity function \( i(x) = x \) serves as the identity -- i.e.

\[
f \circ i = i \circ f = f \quad \forall f \in F(X, X)
\]

(E.23)

Thus we have defined a map \( b : U^* \to F(X, X) \)

\[
b(\omega) = \phi^\omega
\]

(E.24)

and we have that

\[
b(\omega_1 \circ \omega_2) = b(\omega_2) \circ b(\omega_1)
\]

(E.25)

which means that \( b \) is an anti-homomorphism (the order of \( \omega_1 \) and \( \omega_2 \) get reversed). We could have avoided this -- i.e. we could have made \( b \) a homomorphism -- if we ordered strings from right to left instead of from left to right or if we had written function operations on the right (\( xf = f(x) \); in this case \( xfg = g(f(x)) = g \circ f(x) \)), as many algebraists do (see [Fl]).

We now consider the following automaton. We take the set of state transition functions \( b(U^*) = \{ \phi^\omega | \omega \in U^* \} \) as our state set. Define \( U \) \( u \in U \)

\[
\alpha(\phi^\omega, u) = b(u) \circ \phi^\omega = \phi^u \circ \phi^\omega = \phi^{\omega \circ u}
\]

(E.26)

Suppose that the original system (E.1), (E.2) has a distinguished initial state \( x_0 \). Define

\[
\beta_{x_0}^\omega (\phi^\omega) = \delta(\phi^\omega(x_0)) = \delta(\lambda(x_0, \omega)) = \delta(x_0, \omega)
\]

(E.27)
As with linear systems, we take as our initial state $\phi_A = I$, the identity function. Then the input-output map for this system is
\[ f_{\phi_A}(\omega) = \beta_{x_0} (\phi_A, \omega) = \beta_{x_0} (\alpha(\phi_A, \omega)) = \beta_{x_0} (\phi) = \delta(x_0, \omega) \] (E.28)
which is precisely the input-output function $f_{x_0}$ of (E.1), (E.2) started in the state $x_0$. Thus the automaton $(U, Y, b(U^*), \alpha, \beta_{x_0})$ with initial state $\phi_A$ is another realization of the input-output function $f_{x_0}$.

We make several remarks about this type of realization. First, since $\phi_2 \circ \phi_1 = \phi_1 \circ \phi_2$ we see that the set $b(U^*)$ is a subsemigroup of $F(X, X)$. Note that the recursive next state function $\alpha$, defined in (E.26), is of the form $s_1 \cdot s_2$ where $s_1, s_2 \in$ a semigroup (here $b(U^*)$) and $s_1$, the input, is restricted to some subset of the semigroup (here $b(U) \subseteq b(U^*)$). It is the simplicity of the form of this next state function that makes this particular type of realization appealing. See Chapter 9 in which we make use of this simple structure to perform certain computational tasks efficiently.

We also remark that the function $b : U^* \rightarrow F(X, X)$ defines a second equivalence relation on $U^*$. We say
\[ \omega_1 = \omega_2 \iff \phi_{\omega_1} = \phi_{\omega_2} \] (E.29)
These concepts can be used to define Myhill equivalence [M12], [K11]. Given an input-output map $f : U^* \rightarrow Y$, we construct the Nerode equivalence classes and Nerode realization (see eqns. (E.13), (E.14)). We then construct the state transition semigroup realization (E.26), (E.27), where $x_0 = [\Lambda]_N$. This realization is called the Myhill realization, and the
semigroup of state transition functions is called the **Myhill semigroup.**

The equivalence relation "\( \equiv_f \)" defined by

\[
\omega_1 \equiv_f \omega_2 \iff \phi_{\omega_1} = \phi_{\omega_2}
\]

(E.30)

for the Myhill realization, is called **Myhill equivalence.**

Equation (E.30) deserves further comment. Recall that the Nerode states are equivalence classes of inputs. Then using (E.20) and (E.30) we have

\[
\omega_1 \equiv_f \omega_2 \iff \lambda_f ([\omega]_N, \omega_1) = \lambda_f ([\omega]_N, \omega_2)
\]

(E.31)

for all Nerode classes \([\omega]_N\). Then from (E.13), we must have

\[
[\omega \circ \omega_1]_N = [\omega \circ \omega_2]_N \quad \forall \omega \in U^*
\]

(E.32)

and referring to (E.12) we see that

\[
\omega_1 \equiv_f \omega_2 \iff f(\omega \circ \omega_1 \circ \mu) = f(\omega \circ \omega_2 \circ \mu) \quad \forall \omega, \mu \in U^*
\]

(E.33)

This is the definition given in [K11]. Note that we have

\[
\omega_1 \equiv_f \omega_2 \Rightarrow \omega_1 \sim_f \omega_2
\]

(E.34)

(consider (E.32) with \( \omega = \Lambda \)), but the reverse implication need not be true. Thus given a Nerode equivalence class \([\omega]_N\), the Myhill class \([\omega]_M\) is contained in it, and we see that the number of states (Myhill classes) in the Myhill realization is at least as large as and perhaps larger than the number of Nerode classes (in [K11] it is shown that if there are \( n \) Nerode classes and \( n' \) Myhill classes, then \( n \leq n' \leq n^2 \)).

Also note that an alternative form for the Myhill realization is
\[ \tilde{a}(\omega)_M = [\omega_M \circ u] \]  \hspace{1cm} (E.35)

\[ \tilde{b}(\omega)_M = f(\omega) \]  \hspace{1cm} (E.36)

We now present an example from [K11] that illustrates the Nerode and Myhill realizations.

**Example E.1.** Let \( U = Y = \{0,1\} \) and define \( f : U^* \to Y \)

\[ f(\omega) = \begin{cases} 1 & \text{if } \omega \in R = \{0,01,011,0111,\ldots\} \\ 0 & \text{for all other } \omega \end{cases} \]  \hspace{1cm} (E.37)

For those familiar with the concept, the set \( R \) is a **regular language** with alphabet \( U = \{0,1\} \), and the function \( f \) is an **acceptor** for this language (see [A4], [B15]). We claim that there are three Nerode equivalence classes --- \( q_0 = \{\lambda\} \), \( q_1 = R \), and \( q_2 = U^* - (R \cup \{\lambda\}) \). We see this as follows: for any \( \omega \in q_2 \), \( f(\omega \circ u) = 0 \) \( \forall u \in U^* \), so all elements of \( q_2 \) are Nerode equivalent; for any \( \omega \in q_1 \) \( f(\omega \circ u) = 1 \) if \( u \) is a sequence of \( n \) 1's for \( n > 0 \) and \( f(\omega \circ u) = 0 \) for any other \( u \); thus all elements of \( q_1 \) are equivalent to each other and inequivalent to any element of \( q_2 \); finally \( f(\lambda) = 0 \), which means \( \lambda \not\in f \omega \forall \omega \in q_1 \) and \( f(\lambda \circ 0) = 1 \), so that \( \lambda \not\in f \omega \forall \omega \in q_2 \).

Using (E.13) and (E.14), we can define \( \lambda_f \) and \( \delta_f \):

\[ \lambda_f(q_0,0) = q_1 \]  \hspace{1cm} (E.38)

\[ \lambda_f(q_0,1) = q_2 \]  \hspace{1cm} (E.39)

\[ \lambda_f(q_1,0) = q_2 \]  \hspace{1cm} (E.40)

\[ \lambda_f(q_1,1) = q_1 \]  \hspace{1cm} (E.41)

\[ \lambda_f(q_2,0) = q_2 \]  \hspace{1cm} (E.42)

\[ \lambda_f(q_2,1) = q_2 \]  \hspace{1cm} (E.43)

\[ \delta(q_0) = \delta(q_2) = 0 \]  \hspace{1cm} (E.44)

\[ \delta(q_1) = 1 \]  \hspace{1cm} (E.45)

A diagram of this system is given in Figure E.1. The circles are the
various states, the double circle indicating that the output from that state is 1, while it is zero in the others. The arrows indicate the state transitions for the particular input values associated with the arrows (the unmarked arrow leading to $q_o$ indicates that $q_o$ is the initial state).

Similarly we can compute the Myhill equivalence classes. We leave it to the interested reader to check that the Myhill classes are $s_o = q_o$, $s_1 = q_1$, $s_2 = \{1,11,111,\ldots\}$, $s_3 = \text{all the others}$. The state transition functions $\phi_\omega$ corresponding to $s_o$, $s_1$, $s_2$, $s_3$ are given by the table:

\[
\begin{array}{cccc}
  & s_o & s_1 & s_2 & s_3 \\
 q_0 & q_0 & q_1 & q_2 & q_2 \\
 q_1 & q_1 & q_2 & q_1 & q_2 \\
 q_2 & q_2 & q_2 & q_2 & q_2 \\
\end{array}
\]

and thus by computing the various function compositions we have the semigroup multiplication table

\[
\begin{array}{cccc}
  & s_o & s_1 & s_2 & s_3 \\
 s_o & s_o & s_1 & s_2 & s_3 \\
 s_1 & s_1 & s_3 & s_1 & s_3 \\
 s_2 & s_2 & s_3 & s_2 & s_3 \\
 s_3 & s_3 & s_3 & s_3 & s_3 \\
\end{array}
\]

where here we use the multiplication

\[s_i \cdot s_j = \phi_{s_j} \circ \phi_{s_i}\]
(see equation (E.22)), where by \( \Phi_{s_i} \) we mean \( \Phi_\omega \) for any \( \omega \in s_i \).

Referring to (E.26) and (E.27) we can write the Myhill realization equations (the initial state is \( s_0 \)):

\[
\alpha(s_1, u) = s_1 \cdot b(u) \tag{E.43}
\]

\[
b(0) = s_1 \quad b(1) = s_2 \tag{E.44}
\]

\[
\beta(s_0) = \beta(s_2) = \beta(s_3) = 0 \quad \beta(s_1) = 1 \tag{E.45}
\]

Figure E.2 illustrates this realization.

![Diagram](image)

**Figure E.1**: Illustrating the Nerode Realization of Example E.1

![Diagram](image)

**Figure E.2**: Illustrating the Myhill Realization of Example E.1
APPENDIX F

LIE GROUP HOMOMORPHIC SEQUENTIAL SYSTEMS

"Our little systems have their day"

- Alfred, Lord Tennyson [T3]

In this appendix we introduce a class of discrete-time systems evolving on Lie groups. This class of systems should be viewed as being the Lie group analog of the class of FGMSS's described in Chapter 8, and, as we shall see, there is an interesting relationship between this class of Lie group systems and linear systems. To discuss these systems, we need to consider the abstract definitions of a Lie group and its Lie algebra. We will not present these here, and the reader is referred to [WL] and [C1]. We note only that a Lie algebra \( L \) is a finite dimensional vector space over \( \mathbb{R}^1 \) together with an operation \([ , ] : L \times L \to L\) such that for all \( x, y, z \in L \) and all \( a, b \in \mathbb{R}^1\)

\[
[x, y] = -[y, x] \quad \text{(F.1)}
\]

\[
[[x, y], z] + [[y, z], x] + [[z, x], y] = 0 \quad \text{(F.2)}
\]

\[
[ax + by, z] = a[x, z] + b[y, z] \quad \text{(F.3)}
\]

Note that if \( L \) is a matrix Lie algebra (see Definition A.24), the commutator product satisfies (F.1)-(F.3). Also, as a manifold [WL], the dimension of a Lie group \( G \) is the same as the dimension of its Lie algebra \( L_G \), regarded as a vector space.

Suppose \( G \) and \( H \) are Lie groups. A map \( f : G \to H \) is a Lie group homomorphism if it is a group homomorphism and if it is smooth (see
The differential of this map at the identity \( e \in G \), \((df)_e \), is a linear map from the Lie algebra \( L_G \) of \( G \) into the Lie algebra \( L_H \) of \( H \). In fact, \((df)_e : L_G \rightarrow L_H \) is a Lie algebra homomorphism — i.e., it is a linear map such that

\[
(df)_e([x,y]) = [(df)_e(x), (df)_e(y)] \quad \forall \; x,y \in L_G \quad (F.4)
\]

The reader is referred to [Wl], [Cl], and [Sl4] for details of these concepts. An example of a Lie group homomorphism is \( f : SO(2) \rightarrow SO(2) \) defined by

\[
f(X) = X^2 \quad (F.5)
\]

The Lie algebra of \( SO(2) \) is

\[
L = \{ aR \mid a \in \mathbb{R} \} \quad (F.6)
\]

where

\[
R = \begin{bmatrix}
0 & 1 \\
-1 & 0
\end{bmatrix} \quad (F.7)
\]

and \((df)_I : L \rightarrow L\) is given by

\[
(df)_I(A) = 2A \quad \forall \; A \in L \quad (F.8)
\]

We can now define the class of Lie Group Homomorphically Sequential Systems (LGHSS's). Let \( X \), \( Y \) and \( Z \) be Lie groups with associated Lie algebras \( L_x \), \( L_y \), and \( L_z \), respectively. Let \( a : X \rightarrow X \), \( b : U \rightarrow X \), and \( c : X \rightarrow Y \) be Lie group homomorphisms, and let \((da)_e \hat{=} A : L_x \rightarrow L_x\), \((db)_e \hat{=} B : L_u \rightarrow L_x \), and \((dc)_e \hat{=} C : L_x \rightarrow L_y \) be the associated Lie algebra homomorphisms (here we use \( e \) to denote the identities of all three
groups). A LCHSS is a discrete time system of the form

\[ x(k+1) = b[u(k)]a[x(k)] \]  \( \text{(F.9)} \)

\[ y(k) = c[x(k)] \]  \( \text{(F.10)} \)

We now define the concepts of local controllability and local observability.

**Definition F.1:** Consider the LCHSS (F.9), (F.10). The system is **locally controllable at** \( x \in X \) if there exists a neighborhood \( U \) of \( x \) such that every element of \( U \) can be reached if \( x \) is the initial state. The system is **locally observable at** \( x \in X \) if there exists a neighborhood \( V \) of \( x \) such that any two elements of \( V \) are distinguishable (see Definition 8.1).

We note that, as with a PCHSS, **two states in a LCHSS are distinguishable if and only if the identity input sequence distinguishes between them** (see Theorem 8.9).

The following theorem contains sufficient conditions for local controllability and observability at \( e \in X \).

**Theorem F.1:** Consider the LCHSS (F.9), (F.10). The system is locally controllable at \( e \in X \) if the linear system

\[ z(k+1) = A[z(k)] + B[v(k)] \]  \( \text{(F.11)} \)

\[ w(k) = C[z(k)] \]  \( \text{(F.12)} \)

is controllable. Also, the LCHSS is locally observable at \( e \) if the linear system (F.11), (F.12) is observable.

**Proof:** Let \( n = \dim X = \dim L_x \). Consider the input-state map \( f : U^n \to X \) associated with the LCHSS started in the state \( e \).

\[ f(u_0, \ldots, u_{n-1}) = b(u_{n-1})a(u_{n-2}) \ldots a^{n-1}b(u_0) \]  \( \text{(F.13)} \)

We note that \( U^n \) can be regarded as a Lie group with Lie algebra \((L_U)^n\).
Let $F = (df)_{e,\ldots,e}$. We have that $F : \mathcal{L}_u^n \to L_x$. Using the chain rule and product rule for differentials \[ F(v_0,\ldots,v_{n-1}) = B(v_{n-1}) + AB(v_{n-2}) + \ldots + A^{n-1}B(v_o) \] The map $F$ is onto $L_x$ if and only if the linear system (F.11) is controllable -- i.e. if and only if
\[ \text{rank} \ [B,AB,\ldots,A^{n-1}B] = n \] Thus, if (F.15) holds, $F$ is onto $L_x$, and, by the inverse function theorem [S14], [M15], [L12], there is a neighborhood $U$ of $e \in X$, such that given any $x \in U$, there exists $(u_0,\ldots,u_{n-1}) \in U^n$ such that $f(u_0,\ldots,u_{n-1}) = x$ -- i.e. the LGHSS is locally controllable at $e$.

Now consider the map $g : X \to Y^n$ defined by
\[ g(x) = (c(x),ca(x),\ldots,ca^{n-1}(x)) \] We regard $Y^n$ as a Lie group with Lie algebra $(L_y)^n$ and define $G = (dg)_e : \mathcal{L}_x \to (L_y)^n$. Using the chain rule, we have
\[ G(z) = (C(z),CA(z),\ldots,CA^{n-1}(z))' \] Thus $G$ is one-to-one if and only if the linear system (F.11), (F.12) is observable -- i.e. if and only if
\[ \text{rank} \ [C',A'C',\ldots,(A^{n-1})'C'] = n \] Therefore, if (F.11), (F.12) is observable, the inverse function theorem implies that there exists a neighborhood $V$ of $e \in X$ such that if $x, v \in V$, we have
\[ g(x) = g(v) \Rightarrow x = v \] That is, the LGHSS is locally observable at $e$. 
APPENDIX C

AN EXAMPLE OF THE ERRORS INTRODUCED BY FOURIER SERIES TRUNCATION

"Those oft are stratagems which errors seem"

- Alexander Pope [P7]

In this appendix we present an example that describes the type of errors that are caused by truncating the Fourier series expressions for probability densities on $S^1$ and the associated measurement update equations. This discussion is based on the results contained in Section 5.2. For simplicity, we will consider only the simple truncation of Fourier series (i.e., we will not consider any of the more sophisticated techniques discussed in Appendix D).

We assume that we wish to estimate a random variable $\theta$ on $S^1$ with a prior density

$$p(\theta) = p(\theta \mid \theta) = \frac{1}{2\pi} + \sum_{n=1}^{N} a_n(0) \sin n\theta + b_n(0) \cos n\theta \quad (C.1)$$

Since diffusion update equations of the type discussed in Examples 5.4 and 5.5 do not require truncation if $p$ is given by (C.1) (see Section 5.2), we will not consider them here -- i.e., we will only consider computing the conditional density $p(\theta \mid n)$ for $\theta$ given the (possibly nonlinear) measurements $\nu_1, \ldots, \nu_n$. We suppose that for each $k$ we have

$$p(\nu_k \mid \theta) = d_0(\nu_k) + \sum_{n=1}^{N} c_n(\nu_k) \sin n\theta + d_n(\nu_k) \cos n\theta \quad (C.2)$$

(the dependence of the $c_n$'s and $d_n$'s on $k$ allows for the possibility that there are different types of measurements). Note that the series in (C.1)
and (C.2) are taken to be finite. In general this need not be the case, but such a priori truncations are necessary in order to construct useful filtering algorithms.

Using equations (5.5) - (5.9), we have

$$ p(\theta | 1) = \frac{1}{2\pi} + \sum_{n=1}^{2N} a_n(1) \sin n\theta + b_n(1) \cos n\theta \quad (C.3) $$

where

$$ a_n(1) = \frac{\alpha_n(1)}{2\pi c(1)} \quad b_n(1) = \frac{\beta_n(1)}{2\pi c(1)} \quad (C.4) $$

$$ c(1) = \frac{d_o(1)}{2\pi} + \frac{1}{2} \sum_{n=1}^{N} \left( a_n(0)c_n(1) + b_n(0)d_n(1) \right) \quad (C.5) $$

$$ \alpha_k(1) = \begin{cases} 
\frac{c_k(1)}{2\pi} + \frac{1}{2} \sum_{n=1}^{k-1} (a_{n+k}(0)d_n(1) + b_{n+k}(0)c_n(1)) \\
+ \frac{1}{2} \sum_{n=1}^{N-k} [(a_{n+k}(0)d_n(1) + b_n(0)c_{n+k}(1)) - (a_n(0)d_{n+k}(1) + b_{n+k}(0)c_n(1))]
\end{cases}, \quad 1 \leq k \leq N $$

$$ \beta_k(1) = \begin{cases} 
\frac{1}{2} \sum_{n=K-N}^{N} (a_n(0)d_{n+k}(1) + b_n(0)c_{n+k}(1)) \quad \forall 1 \leq K \leq 2N 
\end{cases} \quad (C.6) $$

and we have a similar expression for $\beta_k(1)$. Note that in order to keep these equations from becoming even more complicated, we have not shown the explicit dependence of the various coefficients on the measurement value $y_1 = v$.

A straightforward truncation of $p(\theta | 1)$ (assuming we wish to keep the first $N$ modes) yields

$$ \tilde{p}(\theta | 1) = \frac{1}{2\pi} + \sum_{n=1}^{N} \tilde{a}_n(1) \sin n\theta + \tilde{b}_n(1) \cos n\theta \quad (C.7) $$
where
\[ \tilde{a}_n(1) = a_n(1) \quad \tilde{b}_n(1) = b_n(1) \quad (G.8) \]

Now suppose we wish to compute the optimal estimate using the cost criterion \( \mathcal{E}[(1 - \cos(\theta - \hat{\theta}))^2] \). As shown in Example 3.2, \( \hat{\theta} \) depends on \( a_1, b_1, a_2, \) and \( b_2 \). Thus we must have \( N \geq 2 \). In this case (G.8) shows that the truncation (G.7) introduces no error into the value of \( \hat{\theta}(1) \). However, errors will arise in the second stage of this process -- i.e. when we process \( v_2 \).

We first compute the actual conditional density
\[
p(\theta|2) = \frac{1}{2\pi} + \frac{3N}{\sum_{n=1}^{3N} a_n(2)\sin n\theta + b_n(2)\cos n\theta} \quad (G.9)
\]
\[
a_k(2) = \frac{a_k(2)}{2\pi c(2)} \quad b_k(2) = \frac{b_k(2)}{2\pi c(2)} \quad (G.10)
\]
\[
c(2) = \frac{d_0(2)}{2\pi} + \frac{1}{2} \sum_{n=1}^{N} [a_n(1)c_n(2) + b_n(1)d_n(2)] \quad (G.11)
\]
\[
\begin{align*}
p(\theta|2) &= d_0(2)a_k(1) + \frac{c_k(2)}{2\pi} + \frac{1}{2} \sum_{n=1}^{K-1} (a_n(1)d_{k-n}(2) + b_n(1)c_{k-n}(2)) \\
&\quad + \frac{1}{2} \sum_{n=1}^{N} (a_{n+k}(1)d_n(2) - b_{n+k}(1)c_n(2)) \\
&\quad + \frac{1}{2} \sum_{n=1}^{N-k} (b_n(1)c_{n+k}(2) - a_n(1)d_{n+k}(2)) \quad 1 \leq k \leq N
\end{align*}
\[
\begin{align*}
p(\theta|2) &= d_0(2)a_k(2) + \frac{1}{2} \sum_{n=K-N}^{K-1} (a_n(1)d_{k-n}(2) + b_n(1)c_{k-n}(2)) \\
&\quad + \frac{1}{2} \sum_{n=1}^{2N-K} (a_{n+k}(1)d_n(2) - b_{n+k}(1)c_n(2)) \quad N+1 \leq k \leq 2N \\
&\quad + \frac{1}{2} \sum_{n=K-N}^{2N-K} (a_n(1)d_{k-n}(2) + b_n(1)c_{k-n}(2)) \quad 2N+1 \leq k \leq 3N
\end{align*}
\quad (G.12)
and a similar expression for \( \tilde{\theta}_k(2) \).

We now compute the approximate density \( \tilde{p}(\theta|2) \), where we take 
\( \tilde{p}(\theta|1) \) as the a priori density.

\[
\tilde{p}(\theta|2) = \frac{1}{2\pi} + \frac{1}{N} \sum_{n=1}^{N} \tilde{a}_n(2) \sin n\theta + \tilde{b}_n(2) \cos n\theta
\]

\[
\tilde{a}_k(2) = \frac{\tilde{d}_o(2)}{2\pi \tilde{c}(2)} + \frac{1}{2} \sum_{n=1}^{N} \tilde{a}_n(1) c_{n+k}(2) \]

\[
\tilde{b}_k(2) = \frac{\tilde{d}_o(2)}{2\pi \tilde{c}(2)} + \frac{1}{2} \sum_{n=1}^{N} \tilde{b}_n(1) d_{n+k}(2)
\]

\[
\tilde{c}(2) = \frac{d_o(2)}{2\pi} + \frac{1}{2} \sum_{n=1}^{N} \tilde{a}_n(1) c_{n+k}(2) + \tilde{b}_n(1) d_{n+k}(2)
\]

\[
\tilde{a}_k(2) = d_o(2) \tilde{a}_k(1) + \frac{c_k(2)}{2\pi} + \frac{1}{2} \sum_{n=1}^{N-k} (\tilde{a}_n(1) d_{n+k}(2) + \tilde{b}_n(1) c_{n+k}(2))
\]

\[
+ \frac{1}{2} \sum_{n=1}^{N-k} [(\tilde{a}_{n+k}(1) d_{n+k}(2) + \tilde{b}_n(1) c_{n+k}(2)) - (\tilde{a}_n(1) d_{n+k}(2) + \tilde{b}_n(1) c_{n+k}(2))]
\]

and a similar equation for \( \tilde{\theta}_k(2) \). We can now compare the actual and approximate coefficients. Note that \( C.8 \), \( C.11 \), and \( C.15 \) yield \( \tilde{c}(2) = c(2) \), and thus

\[
a_k(2) - \tilde{a}_k(2) = \frac{1}{2\pi \tilde{c}(2)} \left( \frac{1}{N-k+1} \sum_{n=N-k+1}^{N} \left[ a_n d_n(2) - b_n(1) c_n(2) \right] \right)
\]

Assuming that the coefficients \( a_n, b_n, c_n, d_n = \frac{1}{n} \) \([K10]\), the dominating term in \( a_k(2) - \tilde{a}_k(2) \) goes as \( \frac{1}{(N+1)^2} \frac{1}{(N-k+1)^2} \), and thus

\[
a_k(2) - \tilde{a}_k(2) = \frac{k}{(N+1)^2 (N-k+1)^2}
\]

We can continue this procedure to compute the \( \{a_n(k), b_n(k)\} \)
and the \( \{\tilde{a}_n(k), \tilde{b}_n(k)\} \) from the \( \{a_n(k-1), b_n(k-1)\} \) and the
\( \{\tilde{a}_n(k-1), \tilde{b}_n(k-1)\} \), respectively. It is clear that the error due to
truncation grows at each stage, while (C.18) shows that the error decreases markedly as we increase \( N \). Some analysis is needed to see how to choose \( N \) in order to keep the error within reason over some given measurement interval.

We close this appendix by deriving an equation for the first order effect that errors in the Fourier coefficients have on the estimate of \( \theta \). Let \( \hat{\theta}_u \) be the optimal estimate if we have the untruncated density, and let \( \hat{\theta}_t \) be the estimate if we use the truncated density. Write

\[
\Delta \theta = \hat{\theta}_t - \hat{\theta}_u \tag{C.19}
\]

\[
\Delta a_1 = \tilde{a}_1 - a_1 \quad \Delta b_1 = b_1 - \bar{b}_1 \tag{C.20}
\]

\[
\Delta a_2 = \tilde{a}_2 - a_2 \quad \Delta b_2 = b_2 - \bar{b}_2 \tag{C.21}
\]

and assume we compute \( \hat{\theta}_u \) and \( \hat{\theta}_t \) using the necessary condition (3.25), where we use \( a_1, a_2, b_1, \) and \( b_2 \) for \( \hat{a}_u, \tilde{a}_2, \bar{a}_1, \tilde{b}_1, \) and \( \bar{b}_2 \) for \( \hat{a}_t, \tilde{b}_2, \).

Thus we have

\[
(a_2 \cos 2\tilde{\theta}_u - b_2 \sin 2\tilde{\theta}_u) - 2(a_1 \cos \tilde{\theta}_u - b_1 \sin \tilde{\theta}_u) = 0 \tag{C.22}
\]

Writing \( \hat{\theta}_u = \hat{\theta}_t + \Delta \theta \), making similar substitutions for the \( a \)'s and \( b \)'s, and assuming small perturbations so that we can write \( \sin \Delta \theta \approx \Delta \theta \)

\[
\Delta \theta^2 = \frac{2(\Delta a_1 \cos \tilde{\theta}_t - \Delta b_1 \sin \tilde{\theta}_t) - (\Delta a_2 \cos 2\tilde{\theta}_t - \Delta b_2 \sin 2\tilde{\theta}_t)}{2[(\tilde{a}_1 \sin \tilde{\theta}_t + \tilde{b}_1 \cos \tilde{\theta}_t) - (\tilde{a}_2 \sin 2\tilde{\theta}_t + \tilde{b}_2 \cos 2\tilde{\theta}_t)]} \tag{C.23}
\]

where we have used

\[
(\tilde{a}_2 \cos 2\tilde{\theta}_t - \tilde{b}_2 \sin 2\tilde{\theta}_t) - 2(\tilde{a}_1 \cos \tilde{\theta}_t - \tilde{b}_1 \sin \tilde{\theta}_t) = 0 \tag{C.24}
\]
Also, referring to (3.26), we see that the denominator of the right-hand side of (G.23) must be \( \geq 0 \), and thus must be strictly positive if the first order analysis that led to (G.23) is justified.
BIBLIOGRAPHY

A1. Aitchison, J. and Brown, J.A.C., The Lognormal Distribution,

A2. Achieser, N.I., Theory of Approximation, C.J. Hyman, transl.,

A3. Arbib, M.A., ed, Algebraic Theory of Machines, Languages, and


    Syst. Sci., to be published.


A8. Arbib, M.A. and Manes, E.G., Machines in a Category: An Expository
    Introduction, Computer and Information Science Tech. Rent. 71C-6,
    Univ. of Mass., 1972.

A9. Arbib, M.A., "Decomposition Theory for Automata and Biological
    Systems," IEEE Control Systems Society, Catalog 71C61CSS, A.S. Morse, 

A10. Astrom, K.J., "Optimal Control of Markov Processes with Incomplete

    Functions with Formulas, Graphs, and Mathematical Tables, National


L11. Licht, B.W., Approximations in Optimal Nonlinear Filtering, Report No. SRC 70-1, Systems Research Center, Case Western Reserve University, Cleveland, Ohio, April 1970.


M17. Mother Goose, The Original Mother Goose's Melody, Damrell and Upham, Boston, 1892.


R5. Rader, C.M., "Discrete Fourier Transforms When the Number of Data Samples is Prime," Proc. of the IEEE, Vol. 56, No. 6, June 1968, pp. 1107-1108.


BIOGRAPHICAL NOTE

Alan Steven Willsky was born in Newark, New Jersey on March 16, 1948. He attended public schools in Morris Township and Morristown, New Jersey and graduated as valedictorian from Morristown High School in June, 1965. Mr. Willsky entered M.I.T. in September, 1965 and graduated from M.I.T. in June, 1969, receiving the degree of Bachelor of Science from the Department of Aeronautics and Astronautics. He was awarded the Salisbury Award as the outstanding senior in the Department and the DeFlorez Award for outstanding undergraduate research (development of a linear displacement detection scheme based on laser-interferometric techniques). Mr. Willsky also was elected to Tau Beta Pi and Sigma Gamma Tau. As an undergraduate, he was active in a number of student activities, including the M.I.T.-Wellesley cross-registration planning committee, Student Committee on Environment (chairman), and the Dept. of Aeronautics and Astronautics Creative Engineering Symposium (student organizer).

Mr. Willsky continued his studies as a graduate student in the Department of Aeronautics and Astronautics at M.I.T. starting in September, 1969. From September, 1969 through May, 1973 his graduate study was supported by a Fannie and John Hertz Foundation Fellowship.

During summers Mr. Willsky has been employed by Nicolet Industries, Inc. (1967) developing comparative testing procedures for gasketing materials, and by Bell Telephone Laboratories, (1968) developing a guidance strategy for an underground tunneling device, and (1969) working on control and numerical analysis problems.

Mr. Willsky is married to the former Gail Ruth Fleischman (M.I.T., '70) of Elizabeth, New Jersey, who is presently working on her Ph.D. in Molecular Biology at Tufts University.