Hr. 208

LA-5013-T

A THESIS

State Estimation in Discrete-Time Linear Systems with Quadratic Form Observers

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87544

MASTER

UNITED STATES
ATOMIC ENERGY COMMISSION
CONTRACT W-7405-ENG. 36

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

This report is a printed version of a thesis submitted as partial fulfillment of the requirement for an advanced academic degree and as such represents the independent work of the author. It has not been edited by the Technical Information staff.

Printed in the United States of America. Available from
National Technical Information Service
U. S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22151
Price: Printed Copy ~~$3.00~~; Microfiche $0.95
$5.45

LA-5013-T
A Thesis
UC-32
ISSUED: May 1973

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87544

State Estimation in Discrete-Time Linear Systems with Quadratic Form Observers

by

James Parish Shipley, Jr.

This report is derived from a dissertation submitted to the Department of Electrical Engineering and Computer Science of The University of New Mexico in partial fulfillment of requirements for the Degree of Doctor of Philosophy.

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

TABLE OF CONTENTS

	Page
LIST OF ILLUSTRATIONS	vi
LIST OF TABLES	x
CHAPTER 1. INTRODUCTION	1
1.1 Statement of the Problem	1
1.2 Origin of the Problem	3
1.3 Previous Work on This Problem	4
1.4 Summary of the Dissertation	5
CHAPTER 2. SOME RESULTS ON OBSERVABILITY	7
2.1 Introduction	7
2.2 A Constructive Proof of Complete Observability	7
2.3 Conditions for the Existence of A^{-1}	15
2.4 An Example	16
2.5 A Special Case	19
2.6 Remarks on Necessary Conditions on the Control in the Stochastic Case	21
2.7 Summary	22
CHAPTER 3. MAXIMUM LIKELIHOOD STATE ESTIMATION	24
3.1 Introduction	24
3.2 Development of the Likelihood Function	24
3.3 Calculation of the Initial State Estimate	31
3.4 Sequential Form of the Estimator	34
3.5 Construction of the Filtering Estimator	37
3.6 An Approximate Solution to the Estimator Equation	39
3.7 A Scalar Example	42
3.8 Summary and Conclusions	44

TABLE OF CONTENTS (Continued)

	Page
CHAPTER 4. A BAYESIAN APPROACH TO STATE ESTIMATION	46
4.1 Introduction	46
4.2 Calculation of the Posterior Density Function	47
4.3 Calculation of the Maximum A Posteriori State Estimate	53
4.4 A Geometric Approach to Estimating the State	54
4.5 Calculation of the Approximate Error Covariance Matrix	60
4.6 Some Remarks on Error Propagation	63
4.7 A Scalar Example	64
4.8 Summary and Conclusions	66
CHAPTER 5. ALGORITHM TABULATION AND STRUCTURAL COMPARISONS	67
5.1 Introduction	67
5.2 Problem Restatement	67
5.3 Algorithm Tabulation	68
5.4 Structural Comparisons	71
CHAPTER 6. NUMERICAL RESULTS	75
6.1 Introduction	75
6.2 General Third-Order Model Investigation	75
6.2.1 Description of the First Simulation Model	75
6.2.2 Typical State Trajectories	77
6.2.3 The Maximum A Posteriori Algorithm	81
6.2.4 The Approximate Maximum A Posteriori Algorithm	88
6.2.5 First-Order Conditional Mean Filter	88
6.2.6 Remarks on the Second-Order Conditional Mean Filter	95
6.2.7 The Maximum Likelihood Algorithm	105

TABLE OF CONTENTS (Continued)

	Page
6.2.8 Further Comparison of the Approximate Maximum A Posteriori, First-Order Conditional Mean, and Maximum Likelihood Algorithms	108
6.2.9 The Effect of Small Controls	108
6.3 Investigation of a Third-Order System with Q = I	120
6.4 Summary	120
CHAPTER 7. CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH	128
7.1 Summary of Results	128
7.2 Directions for Further Research	128
APPENDIX A. PROBLEMS IN THE SOLUTION OF EQUATIONS 2.2.7	130
ACKNOWLEDGMENTS	134
BIBLIOGRAPHY	135

LIST OF ILLUSTRATIONS

Figure		Page
4.4.1	2-D Illustration of $x'Qx = z$, $x'Qx = \hat{z}$, and $x = \hat{x}$	56
6.2.2.1	Typical x_1 Trajectories for the Three Control Sequences	78
6.2.2.2	Typical x_2 Trajectories for the Three Control Sequences	79
6.2.2.3	Typical x_3 Trajectories for the Three Control Sequences	80
6.2.3.1	Maximum A Posteriori Filter Average Error, $u(k) = 1$	82
6.2.3.2	Maximum A Posteriori Filter Average Squared Error, $u(k) = 1$	83
6.2.3.3	Maximum A Posteriori Filter Average Error, $u(k) = 1.1 + (-1)^k(.2)$	84
6.2.3.4	Maximum A Posteriori Filter Average Squared Error, $u(k) = 1.1 + (-1)^k(.2)$	85
6.2.3.5	Maximum A Posteriori Filter Average Error, $u(k) = (-1)^{k+1}$	86
6.2.3.6	Maximum A Posteriori Filter Average Squared Error, $u(k) = (-1)^{k+1}$	87
6.2.4.1	Approximate Maximum A Posteriori Filter Average Error, $u(k) = 1$	89
6.2.4.2	Approximate Maximum A Posteriori Filter Average Squared Error, $u(k) = 1$	90
6.2.4.3	Approximate Maximum A Posteriori Filter Average Error, $u(k) = 1.1 + (-1)^k(.2)$	91
6.2.4.4	Approximate Maximum A Posteriori Filter Average Squared Error, $u(1) = 1.1 + (-1)^k(.2)$	92

LIST OF ILLUSTRATIONS (Continued)

Figure		Page
6.2.4.5	Approximate Maximum A Posteriori Filter Average Error, $u(k) = (-1)^{k+1}$	93
6.2.4.6	Approximate Maximum A Posteriori Filter Average Squared Error, $u(k) = (-1)^{k+1}$	94
6.2.5.1	First-Order Conditional Mean Filter Average Error, $u(k) = 1$	96
6.2.5.2	First-Order Conditional Mean Filter Average Squared Error, $u(k) = 1$	97
6.2.5.3	First-Order Conditional Mean Filter Average Error, $u(k) = 1.1 + (-1)^k(.2)$	98
6.2.5.4	First-Order Conditional Mean Filter Average Squared Error, $u(k) = 1.1 + (-1)^k(.2)$	99
6.2.5.5	First-Order Conditional Mean Filter Average Error, $u(k) = (-1)^{k+1}$	100
6.2.5.6	First-Order Conditional Mean Filter Average Squared Error, $u(k) = (-1)^{k+1}$	101
6.2.6.1	Comparison of Measured Average Squared Error and Calculated Variance of the Estimate Error of the First State Variable	102
6.2.6.2	Comparison of Measured Average Squared Error and Calculated Variance of the Estimate Error of the Second State Variable	103
6.2.6.3	Comparison of Measured Average Squared Error and Calculated Variance of the Estimate Error of the Third State Variable	104
6.2.7.1	Maximum Likelihood Filter Average Error, $u(k) = 1$	106
6.2.7.2	Maximum Likelihood Filter Average Squared Error, $u(k) = 1$	107

LIST OF ILLUSTRATIONS (Continued)

Figure		Page
6.2.8.1	Algorithm Comparison for Larger Measurement Noise, Average Error in x_1	109
6.2.8.2	Algorithm Comparison for Larger Measurement Noise, Average Squared Error in x_1	110
6.2.8.3	Algorithm Comparison for Larger Plant Noise, Average Error in x_1	111
6.2.8.4	Algorithm Comparison for Larger Plant Noise, Average Squared Error in x_1	112
6.2.9.1	Comparison of the First Component of the State and State Estimates for u(k) = .1	114
6.2.9.2	Comparison of the Second Component of the State and State Estimates for u(k) = .1	115
6.2.9.3	Comparison of the Third Component of the State and State Estimates for u(k) = .1	116
6.2.9.4	Comparison of the First Component of the State and State Estimates for u(k) = .01	117
6.2.9.5	Comparison of the Second Component of the State and State Estimates for u(k) = .01	118
6.2.9.6	Comparison of the Third Component of the State and State Estimates for u(k) = .01	119
6.3.1	First-Order Conditional Mean Filter Average Error, Q = I	121
6.3.2	First-Order Conditional Mean Filter Average Squared Error, Q = I	122
6.3.3	Approximate Maximum A Posteriori Filter Average Error, Q = I	123
6.3.4	Approximate Maximum A Posteriori Filter Average Squared Error, Q = I	124

LIST OF ILLUSTRATIONS (Continued)

Figure		Page
6.3.5	Maximum Likelihood Filter Average Error, Q = I	125
6.3.6	Maximum Likelihood Filter Average Squared Error, Q = I	126

LIST OF TABLES

Table		Page
5.2.1	Problem Statement	67
5.3.1	The Maximum Likelihood (ML) Filter Algorithm	68
5.3.2	The Maximum A Posteriori (MAP) Filter Algorithm	69
5.3.3	The Approximate Maximum A Posteriori (AMAP) Filter Algorithm	70
5.3.4	The Second-Order Approximate Conditional Mean Filter Algorithm	70
5.4.1	Relative Merits of the Algorithms	74
6.2.6.1	Comparison of x_1 and the Second-Order Conditional Mean Estimate of x_1	95

ABSTRACT

STATE ESTIMATION IN DISCRETE-TIME

LINEAR SYSTEMS WITH QUADRATIC FORM OBSERVERS

James Parish Shipley, Jr.
Department of EECS
The University of New Mexico, 1972

The problem considered in this dissertation is that of estimating the state of a system with a linear plant equation but having a measurement function which is a quadratic form in the state. It is first demonstrated by constructive proof that the problem has a solution if the controls have large enough magnitude, and some requirements on the minimum size of the control are discussed.

Three algorithms for estimating the state are obtained, one using the technique of maximum likelihood, the other two utilizing a more general Bayesian approach. In addition, a first-order approximate conditional mean filter is included for comparison.

It is demonstrated by structural comparisons and numerical results that the algorithm for finding an approximate maximum of the posterior density function is the overall best in terms of speed, storage, practicability, and accuracy. It is also shown that good state estimates can be generated for a large class of problems using any of these algorithms if the control largeness conditions are satisfied.

Finally, some extensions are discussed, the main one being the application of the technique of finding an approximate maximum of the posterior density function to systems with more general nonlinear observers.

Chapter 1

INTRODUCTION

At the present time, the trend in nonlinear filter theory seems to be the development of general algorithms applicable to as large a class of nonlinear systems as possible. This is an admirable goal and one that reflects the urgent desire to know all the answers, to be prepared for every eventuality. To a large extent, this approach has been successful as witness the huge number of books and papers on nonlinear filter theory available today.

However, it is always possible that a specific problem has some characteristic which may be used to advantage and which would be lost in a general treatment. It may also be true that generalizations can be made from these findings that have some bearing on a larger class of problems.

It was with these last two ideas in mind that this topic was chosen. In addition, the problem has some roots in practical application, as discussed in Section 1.2.

1.1 Statement of the Problem

Consider a discrete-time system with linear dynamics described by the difference equation

$$x(k+1) = Fx(k) + Bu(k) + Gw(k) \qquad k = 0,1,\ldots \tag{1.1.1}$$

where k is the time index

$x(k) \varepsilon E^n$ is the system state

$u(k) \varepsilon E^m$ is the known control input

$w(k) \varepsilon E^{\ell}$ is the unknown plant noise

F is the nxn state transition matrix

B is the nxm control distribution matrix

G is the nxℓ plant noise distribution matrix.

The initial state x(0) is assumed to have a gaussian probability density function with mean zero and covariance matrix P(0). Similarly, the plant noise sequence $\{w(k)\}$ is taken to be zero-mean, gaussian, and white with covariance matrix given by

$$E\{w(k)w'(j)\} = S(k)\delta_{kj} \tag{1.1.2}$$

where δ_{kj} is the kronecker delta, and the prime indicates matrix transposition. Clearly, S(k) is $\ell x \ell$.

Observations are made of a function of the system state according to

$$z(k+1) = x'(k+1)Qx(k+1) + v(k+1) \qquad k = 0,1,\ldots \tag{1.1.3}$$

where $z(k+1) \varepsilon E^1$ is the scalar measurement

$v(k+1) \varepsilon E^1$ is the measurement noise

and Q is the nxn measurement matrix.

As with the plant noise, the sequence $\{v(k)\}$ is assumed to be zero-mean, gaussian, and white with covariance matrix

$$E\{v(k)v(j)\} = T(k)\delta_{kj} \tag{1.1.4}$$

In this case, the covariance matrix T(k) is 1x1, i.e., a scalar.

Since Q is the kernel of a quadratic form, it is only necessary, from a modeling viewpoint, to consider those matrices Q that are symmetric. Notice also that there is no measurement at time step $k = 0$.

The desired result is an estimate of the current state based on all the known control inputs and available measurements. Thus the problem is one of filtering for state estimation.

It should be pointed out that the system time-invariance and zero-mean statistics properties are not fundamental to what follows. Removal of these restrictions is straightforward but tedious.

Finally, throughout this work it is assumed that equations 1.1.1-4 describe the system exactly; that is, there is no modeling error.

1.2 Origin of the Problem

This problem was abstracted from a portion of the Van de Graaff beam control problem at Los Alamos Scientific Laboratory, Los Alamos, New Mexico. The original task was to focus and position the ion beam using only a Faraday cup for measurements. Although some focusing information can be obtained, the Faraday cup is primarily a position-measuring device. Under suitable conditions the Faraday cup current is proportional to the distance of the beam centroid from the center of the cup. Thus the measurement is related to the distance of the "state" (i.e., beam position) from the origin of "state" space represented by the center of the Faraday cup. In practice the measurement is much more complicated than this because of focusing - positioning interactions, non-uniformity of the beam across its cross-section, the relative geometries of the beam and cup, edge effects, scattered radiation, etc. However, it is felt that solution of the problem stated in the previous

section is a first step toward more effective control of the Van de Graaff beam.

Of course, the results obtained in this thesis are not limited in application to Van de Graaff generators, but may be useful in any situation where a beam (particle, optical, or otherwise) must be positioned based on only a measurement of the distance of the beam from a reference point. Another application that comes immediately to mind is the infrared tracking of objects in the sky with a telescope. Using the observed intensity of the infrared emission of the object as an indication of the distance of the object from the center of the telescope field of view, the algorithms derived here will provide an estimate of the object position relative to the center of the field of view. This position information can then be used to provide pointing corrections to the telescope drive mechanism.

1.3 Previous Work on This Problem

It appears that quadratic form observers have never been treated in any detail in the literature. Many special cases have been treated, however; the favorite one seems to be a scalar system, either linear or nonlinear, with an observer of the form

$$z(k) = x^2(k) + v(k) \tag{1.3.1}$$

with the same notation as in Section 1.1. Sorenson and Alspach [24] work some examples using such a system and conclude that "When there is no plant nonlinearity, it is impossible from the available measurement data to discriminate between the true value of the state and the negative of that value." This is a true statement for the problem they have posed since there is no control input in the model. It will be shown that if

a control input is available, then the control must not be identically zero if the system is to be observable.

The idea of using a known signal input to aid system identification is not new, of course. The general problem of signal design has received some attention, but the goal is usually optimum signal recovery or system identification again. However, the technique of using non-zero controls in nonlinear state estimation problems is at least not well known, if known at all.

1.4 Summary of the Dissertation

After having stated the problem in Section 1.1, it is legitimate to ask whether it can be solved. The purpose of Chapter 2 is to obtain some sufficient conditions under which it can and to discuss some possible necessary conditions. The most striking result is that the control input cannot be identically zero in contrast to completely linear systems. To the best of the author's knowledge, this idea does not appear in the literature.

Armed with this information, it is not difficult to derive the algorithms of Chapters 3 and 4. Chapter 3 on the maximum likelihood estimator is a straightforward extension of the results of Chapter 2, although a bit of finesse is required to put the algorithm in a form amenable to implementation. Chapter 4 attacks the estimation problem from a Bayesian point of view, first obtaining an approximate form for the posterior density function. Two methods of generating state estimates from this density function are derived, one requiring the solution of a nonlinear matrix algebraic equation, the other not even needing the error covariance matrix.

Chapter 5 gathers all the results into tabular form for easy reference. Also included for comparison are first- and second-order approximate conditional mean filters. Differences and similarities among the structures of the various filters are then pointed out.

The function of Chapter 6, Numerical Results, is to provide the results of tests on the algorithms under simulated actual conditions. Some conclusions are then drawn about the relative merits and any particular difficulties of the algorithms.

The last chapter summarizes the dissertation and indicates some possible areas for future research.

Chapter 2

SOME RESULTS ON OBSERVABILITY

2.1 Introduction

It makes little sense to try to estimate the state of a noisy system based on some subset of control inputs and measurements if the corresponding deterministic system is not observable in terms of the same inputs and measurements. Therefore, the first concern is to determine if some (hopefully) finite sequence of controls and measurements contains enough information to yield an exact state estimate for the noise-free system. In what follows a constructive proof is given that this is indeed the case, for proper choice of control, for the deterministic analog of equations 1.1.1 and 1.1.3, and a method is given for reconstructing the state.

The conditions for state reconstruction that will be derived are sufficient but not necessary. A heuristic discussion is given of necessary conditions (i.e., lower bounds) on the control required. The special case of an incremental model is treated, and some examples illustrating the main points of the chapter are presented.

2.2 A Constructive Proof of Complete Observability

Let the state and measurement be determined by

$$x(k+1) = Fx(k) + Bu(k) \tag{2.2.1}$$

$$z(k+1) = x'(k+1)Qx(k+1) \tag{2.2.2}$$

with x(0) unknown. Equation 2.2.1 can be used to write x(k+1) in terms of x(0) and the control sequence {u(0), u(1), ..., u(k)}.

$$x(k+1) = F^{k+1}x(0) + \sum_{i=0}^{k} F^{i}Bu(k-i) \tag{2.2.3}$$

By the properties of the state space description, if the state at any one time and all the controls are known, then the state at any other time is also known. Therefore, it is sufficient to find x(0).

For notational simplicity, let

$$U(k,0) \triangleq \sum_{i=0}^{k} F^{i}Bu(k-i) \tag{2.2.4}$$

and in general

$$U(k,j) \triangleq \sum_{i=0}^{k-j} F^{i}Bu(k-i) \tag{2.2.5}$$

This last equation will be useful in developing the filtering estimator in Chapter 3. Using 2.2.3 and 2.2.4,

$$x(k+1) = F^{k+1}x(0) + U(k,0) \tag{2.2.6}$$

substituting 2.2.6 into the first n measurement equations,*

$$\begin{aligned}
&[Fx(0) + U(0,0)]'Q[Fx(0) + U(0,0)] = z(1)\\
&[F^2x(0) + U(1,0)]'Q[F^2x(0) + U(1,0)] = z(2)\\
&\qquad\vdots\\
&[F^nx(0) + U(n-1,0)]'Q[F^nx(0) + U(n-1,0)] = z(n)
\end{aligned} \tag{2.2.7}$$

*See Appendix A for a more complete discussion of the problems involved in the solution of 2.2.7 for x(0).

Performing the multiplications,

$$x'(0)F'QFx(0) + 2U'(0,0)QFx(0) + U'(0,0)QU(0,0) = z(1)$$

$$\vdots \qquad (2.2.8)$$

$$x'(0)F'^{n}QF^{n}x(0) + 2U'(n-1,0)QF^{n}x(0) + U'(n-1,0)QU(n-1,0) = z(n)$$

Now make the following definitions:

$$Z^* = [z(1) \quad z(2) \quad \ldots \quad z(n)]' \qquad (2.2.9)$$

$$U^* = [U'(0,0)QU(0,0) \quad \ldots \quad U'(n-1,0)QU(n-1,0)]' \qquad (2.2.10)$$

$$X^*[x(0)] = [x'(0)F'QFx(0) \quad \ldots \quad x'(0)F'^{n}QF^{n}x(0)]' \qquad (2.2.11)$$

$$A' = [F'QU(0,0) \vdots \quad \ldots \quad \vdots F'^{n}QU(n-1,0)] \qquad (2.2.12)$$

Putting these into 2.2.8 and rearranging,

$$Ax(0) = \frac{1}{2}\{Z^* - U^* - X^*[x(0)]\} \qquad (2.2.13)$$

By construction, the matrix A is nxn and if it is nonsingular, 2.2.13 can be written in the form

$$x(0) = \frac{1}{2} A^{-1} \{Z^* - U^* - X^*[x(0)]\} \qquad (2.2.14)$$

Conditions for the existence of A^{-1} will be discussed later. Observe that 2.2.14 has the form $x = P(x)$; in such an equation, a solution, if it exists, can be found by a successive approximation technique if P is a contraction mapping. In order to make these ideas more precise, the following definition and theorem from Kantorovich and Alsidov [12], p. 697, are stated.

<u>Definition</u> Given the equation $x = P(x)$ for $x \varepsilon \Omega$, $\Omega \subseteq E^n$, Ω closed. P is a contraction if

$$d[P(x_1), P(x_2)] \leqslant a d[x_1, x_2] \qquad \forall x_1, x_2 \varepsilon \Omega$$

where $0 \leqslant a < 1$ and d is the metric on E^n.

<u>Theorem</u> If P is a contraction mapping in Ω, a unique solution x^* of the equation $x = P(x)$ exists in Ω and can be obtained as the limit of a sequence $\{x_n\}$, where

$$x_{n+1} = P(x_n) \qquad n = 0, 1, \ldots$$

and x_o is any element of Ω. The rate of convergence of $\{x_n\}$ to x^* is given by

$$d[x_n, x^*] \leqslant \frac{a^n}{1-a} d[x_1, x_o]$$

Applying these ideas to 2.2.14, let

$$P(x) \triangleq \frac{1}{2} A^{-1}\{Z^* - U^* - X^*(x)\} \tag{2.2.15}$$

$$\text{and } d[P(x_1), P(x_2)] \triangleq \| P(x_1) - P(x_2) \|$$

$$= \frac{1}{2} \| A^{-1}\{X^*(x_2) - X^*(x_1)\} \| \tag{2.2.16}$$

As shown in Appendix A, any set of n measurement equations such as 2.2.13 has 2^n solutions in general. The problem in applying the above theorem and definition is to find conditions on the control so that P is a contraction mapping in a region Ω containing x^*, the original solution to 2.2.13, and which also contains a convenient known starting point for the iterations, x_o. Since Ω as defined above is difficult to deal with quantitatively, the approach is to define a new set Ω' in which it can be shown that $P(x)$ is a contraction for proper choice of control sequence and which also contains x^* and x_o. By this definition, $\Omega' \subseteq \Omega$, and x^* can be found by the iterative procedure of the theorem

starting from x_o. With these ideas in mind, let

$$\Omega' \triangleq \{x \varepsilon E^n : \|x-x^*\| \leqslant r, r \geqslant 0\} \tag{2.2.17}$$

That is, Ω' is a hypersphere of radius r centered at x*. It is first necessary to show that P maps Ω' into itself. This will be the case if

$$\|x^* - P(x)\| \leqslant r \quad \forall x \varepsilon \Omega' \tag{2.2.18}$$

Or, from 2.2.15,

$$\|x^* - \frac{1}{2} A^{-1}\{Z^* - U^* - X^*(x)\}\| \leqslant r \quad \forall x \varepsilon \Omega' \tag{2.2.19}$$

But by definition, $x^* = P(x^*)$ so that 2.2.19 becomes

$$\|A^{-1}\{Z^* - U^* - X^*(x^*)\} - A^{-1}\{Z^* - U^* - X^*(x)\}\| \leqslant 2r \quad \forall x \varepsilon \Omega' \tag{2.2.20}$$

Simplifying,

$$\|A^{-1}\{X^*(x) - X^*(x^*)\}\| \leqslant 2r \quad \forall x \varepsilon \Omega' \tag{2.2.21}$$

is the condition for P to map Ω' into itself. By the properties of norms (Bellman [3], p. 162), 2.2.21 is satisfied if

$$\|A^{-1}\| \, \|X^*(x) - X^*(x^*)\| \leqslant 2r \quad \forall x \varepsilon \Omega' \tag{2.2.22}$$

Or, rearranging,

$$\|A^{-1}\| \leqslant \frac{2r}{\|X^*(x) - X^*(x^*)\|} \quad \forall x \varepsilon \Omega' \tag{2.2.23}$$

Reference to equations 2.2.11 and 2.2.12 reveals that A is a linear function of the control sequence, while X* does not depend on the control. Therefore, by choosing the sequence of controls large enough in some sense, 2.2.23 can be made to hold.

Next, it is necessary to show that P is a contraction on Ω', i.e., that

$$\|P(x_1) - P(x_2)\| \leqslant a\|x_1 - x_2\| \quad \forall x \varepsilon \Omega' \tag{2.2.24}$$

with $0 \leqslant a < 1$.

A similar development to that above results in

$$\|A^{-1}\| \leqslant \frac{2a\|x_2 - x_1\|}{\|X^*(x_2) - X^*(x_1)\|} \qquad \forall x_1, x_2 \varepsilon \Omega' \tag{2.2.25}$$

with $0 \leqslant a < 1$. Since $x^* \varepsilon \Omega'$ by hypothesis and $a < 1$, satisfaction of 2.2.25 implies that 2.2.23 holds also, so that the only condition on the control is obtained from 2.2.25.

A more explicit relationship results if a particular norm is selected. A useful norm for convergence proofs is the ℓ_∞ norm:

$$\|x\| \triangleq \max_i |x^i| \tag{2.2.26}$$

where x^i is the ith component of the vector x. The corresponding matrix norm is then determined by (Bellman [3], p. 162),

$$\|A\| \triangleq \max_i \sum_j |A_{ij}| \tag{2.2.27}$$

where A_{ij} is the element in the ith row and j^{th} column of the matrix A. Then,

$$\|X^*(x_2) - X^*(x_1)\| = \max_{i=1,\ldots,n} |x_2' F'^i Q F^i x_2 - x_1' F'^i Q F^i x_1|$$

$$= \max_{i=1,\ldots,n} |(x_2 - x_1)' F'^i Q F^i (x_2 + x_1)| \tag{2.2.28}$$

and $$\|x_2 - x_1\| = \max_i |(x_2 - x_1)^i| \tag{2.2.29}$$

where again the superscript i denotes the ith component of the vector $(x_2 - x_1)$. Let

$$J \triangleq \frac{\|X^*(x_2) - X^*(x_1)\|}{2a\|x_2 - x_1\|} \tag{2.2.30}$$

Or, using 2.2.28 and 2.2.29,

$$J = \frac{\max_i |(x_2 - x_1)\ F'^i Q F^i (x_2 + x_1)|}{2a \max_j |(x_2 - x_1)^j|} \tag{2.2.31}$$

Suppose that the maximum in the denominator occurs for $j = \alpha$, so that

$$J = \frac{1}{2a} \max_i \frac{(x_2-x_1)'}{|(x_2-x_1)^\alpha|} F'^i Q F^i (x_2+x_1) \tag{2.2.32}$$

But, $-1 \leq \frac{(x_2-x_1)^i}{|(x_2-x_1)^\alpha|} \leq 1 \qquad i = 1,2, \ldots,n$

with equality holding for $i = \alpha$.

Therefore,

$$J \leq \frac{1}{2a} \max_i \sum_{k=1}^{n} |[F'^i Q F^i]_k (x_2+x_1)| \tag{2.2.33}$$

where $[F'^i Q F^i]_k$ is the kth row of $F'^i Q F^i$.

Using Schwarz's inequality

$$\begin{aligned} J &\leq \frac{1}{2a} \max_i \sum_{k=1}^{n} \|[F'^i Q F^i]_k\| \|x_2+x_1\| \\ &= \frac{1}{2a} \|x_2+x_1\| \max_i \sum_{k=1}^{n} \max_j |[F'^i Q F^i]_{kj}| \end{aligned} \tag{2.2.34}$$

Again, $[F'^i Q F^i]_{kj}$ is the element in the kth row and j^{th} column of $F'^i Q F^i$.

Remembering that $x_2, x_1 \varepsilon \Omega'$,

$$\begin{aligned} \|x_2+x_1\| &= \|(x_2-x^*) + (x_1-x^*) + 2x^*\| \\ &\leq \|x_2-x^*\| + \|x_1-x^*\| + 2\|x^*\| \\ &\leq 2r + 2\|x^*\| \end{aligned}$$

Inserting all the foregoing in 2.2.25,

$$\|A^{-1}\| \leq \frac{a}{r+\|x^*\|} \{\max_i \sum_{k=1}^{n} \max_j |[F'^i Q F^i]_{kj}|\}^{-1} \tag{2.2.35}$$

with $0 \leq a < 1$ is a sufficient condition for P to be a contraction in Ω'.

There remains the problem of a suitable starting point x_o. Recall from the theorem that x_o must lie within Ω. A convenient x_o is the origin which can be made to lie inside Ω', and thus Ω, by choosing r large enough, namely $r \geq \|x^*\|$. If this is done, the condition on the control

sequence that makes P be a contraction on Ω' and $x_o = 0$ be an element of Ω' is

$$\|A^{-1}\| \leqslant \frac{a}{2r} \{\max_i \sum_{k=1}^{n} \max_j |[F'^i Q F^i]_{kj}|\}^{-1} \tag{2.2.36}$$

with $0 \leqslant a < 1$ and $r \geqslant \|x^*\|$

Notice that if 2.2.36 is not satisfied, the iterative procedure may diverge or converge to another solution of 2.2.13. However, this latter situation has never occurred in a considerable number of examples. On this basis and the form of 2.2.13, it is speculated that if the process converges at all, it converges to x^*, but a proof has not been found.

In any case, if there is some a priori knowledge about $\|x^*\|$, an upper bound say, then 2.2.36 will be easy to satisfy. Even if there is no such information, one way to sequentially check the solution is to substitute it into subsequent measurement equations extrapolated to time zero. If the solution does not satisfy all these equations, it is not x^* and another try can be made with larger controls.

These results have proven the following theorem:

<u>Theorem</u> For a discrete-time system described by 2.2.1 and 2.2.2 with Z^*, U^*, X^* and A defined in 2.2.9, 10, 11, and 12 respectively, the state x(0) is reconstructible if A^{-1} exists and satisfies 2.2.36. Furthermore, x(0) can be found by means of the successive approximation equation

$$x_{i+1}(0) = \frac{1}{2} A^{-1}\{Z^* - U^* - X^*[x_i(0)]\} \qquad i = 0,1, \ldots$$

with $x_o(0) = 0$.

2.3 Conditions for the Existence of A^{-1}

By definition,

$$A = \begin{bmatrix} u'(0)B'QF \\ u'(0)B'F'QF^2 + u'(1)B'QF^2 \\ \vdots \\ \sum_{i=0}^{n-1} u'(n-i-1)B'F'^{i}QF^n \end{bmatrix} \tag{2.3.1}$$

It is difficult to make any statements about the singularity of A for a general control sequence. However, it is possible to investigate the properties of A for a particular choice of control.

In this vein, notice that u(0) cannot be the zero vector if A^{-1} exists, since then the first row of A is all zeroes. If $u(0) \neq 0$, but $u(1) = u(2) = \ldots = u(n-1) = 0$, then the nonsingularity of A is equivalent to saying that the sequence of n-vectors

$$\{u'(0)B'QF, \ldots, u'(0)B'F'^{n-1}QF^n\} \tag{2.3.2}$$

is linearly independent.

Another convenient control sequence is $u(0) = u(1) = \ldots = u(n-1) = c$, a constant. The condition for the existence of A^{-1} is then the sequence of n-vectors

$$\{c'B'QF, c'B'[F+I]'QF^2, \ldots, c'B'[F^{n-1} + \ldots + I]'QF^n\} \tag{2.3.3}$$

is linearly independent. It is clear that for the case of scalar control, if 2.3.3 does not hold for a particular choice of c, then it does not hold for any c.

There may be other equally useful control sequences, particularly if the control is not scalar. Each case must be investigated separately as to the existence of A^{-1}.

2.4 An Example

In order to illustrate some of the preceeding points and to gain some insight into the physical significance of some of the results, consider the following two-dimensional system:

$$x(k+1) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k) \tag{2.4.1}$$

$$z(k+1) = x'(k+1) \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} x(k+1) \tag{2.4.2}$$

and suppose that $x(0) = [1 \ -1]'$ (2.4.3)

Observe that this system is not completely controllable and the control is a scalar. First, see if 2.3.2 is valid.

$$B'QF = [0 \quad 1] \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} = [2 \quad 2]$$

Similarly,

$$B'F'QF^2 = [4 \quad 2]$$

so that

$$A = u(0) \begin{bmatrix} 2 & 2 \\ 4 & 2 \end{bmatrix}$$

and $A^{-1} = -\dfrac{1}{4u(0)} \begin{bmatrix} 2 & -2 \\ -4 & 2 \end{bmatrix}$

Therefore, 2.3.2 holds if $u(0) \neq 0$. The required value of $u(0)$ can be determined from the condition 2.2.36.

$$\|A^{-1}\| = \max_i \sum_{j=1}^{2} |A^{-1}_{ij}| = \frac{1}{4|u(0)|}[4+2] = \frac{3}{2|u(0)|}$$

$$F'QF = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}, \quad F'^2QF^2 = \begin{bmatrix} 16 & 4 \\ 4 & 2 \end{bmatrix}$$

Therefore,

$$\max_i \sum_{k=1}^{2} \max_j |[F'^iQF^i]_{kj}| = 20$$

Let the best (conservative) estimate of $\|x^*\|$ be 2, and suppose $a = .5$. Then, the condition on $u(0)$ is

$$\frac{3}{2|u(0)|} \leqslant \frac{.5}{2(2)} [20]^{-1} = 6.25 \times 10^{-3}$$

Therefore,

$$|u(0)| \geqslant 240$$

in order to satisfy 2.2.29. Applying $u(0) = 240$,

$$x(1) = \begin{bmatrix} 2 \\ -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (240) = \begin{bmatrix} 2 \\ 239 \end{bmatrix}$$

and $z(1) = 115{,}202$

$$x(2) = \begin{bmatrix} 4 \\ 239 \end{bmatrix} \text{ and } z(2) = 116{,}170$$

so that $\quad Z^* = [115{,}202 \quad 116{,}170]'$

$$U(0,0) = Bu(0) = [0 \quad 240]'$$

$$U(1,0) = FBu(0) = [0 \quad 240]'$$

Hence $\quad U'(0,0)QU(0,0) = 115{,}200 = U'(1,0)QU(1,0)$

and $\quad U^* = \begin{bmatrix} 115{,}200 \\ 115{,}200 \end{bmatrix}$

Therefore, using the successive approximation equation,

$$x_1(0) = \frac{1}{2}A^{-1}\{Z^* - U^*\} = \begin{bmatrix} 1.008333 \\ \\ -1.006250 \end{bmatrix}$$

$$x_2(0) = x_1(0) - \frac{1}{2}A^{-1}\ X^*[x_1(0)]$$

$$= \begin{bmatrix} 1.00833 \\ \\ -1.006250 \end{bmatrix} - \begin{bmatrix} .00848155 \\ \\ -.00636334 \end{bmatrix} = \begin{bmatrix} .9998518 \\ \\ -.9998867 \end{bmatrix}$$

This answer is within .02% of the true answer in only two iterations.

Several things are apparent from this example. First, the condition 2.2.36 dictates a relatively large control. Second, the correction term to $x_1(0)$ in the computation of $x_2(0)$ is very small compared to $x_1(0)$. This corresponds to choosing the coefficients in 2.2.14 so that the equation is "almost linear," i.e., the nonlinearity represented by the last term in 2.2.14 is negligible.

It is also felt from this example that 2.2.36 will generally result in a very conservative requirement on the control. In order to verify this statement for this problem, suppose $u(0) = 10$ instead of 240. Then,

$$x(1) = \begin{bmatrix} 2 \\ \\ -1 \end{bmatrix} + \begin{bmatrix} 0 \\ \\ 1 \end{bmatrix} (10) = \begin{bmatrix} 2 \\ \\ 9 \end{bmatrix} \qquad z(1) = 202$$

$$x(2) = \begin{bmatrix} 4 \\ \\ 9 \end{bmatrix} \qquad z(2) = 250$$

so that $\qquad Z^* = [202 \quad 250]'$

$$U(0,0) = Bu(0) = [0 \quad 10]'$$

$$U(1,0) = F' Bu(0) = [0 \quad 10]'$$

Hence, $\quad U'(0,0)QU(0,0) = U'(1,0)QU(1,0) = 200$

and $\quad U^* = \begin{bmatrix} 200 \\ 200 \end{bmatrix}$

Therefore, $\quad x_1(0) = \frac{1}{2} A^{-1}\{Z^* - U^*\} = \begin{bmatrix} 1.200 \\ -1.150 \end{bmatrix}$

$$x_2(0) = x_1(0) - \frac{1}{2} A^{-1} X^*[x_1(0)]$$

$$= \begin{bmatrix} 1.200 \\ -1.150 \end{bmatrix} - \begin{bmatrix} .2940 \\ -.2219 \end{bmatrix} = \begin{bmatrix} .9060 \\ -.9281 \end{bmatrix}$$

It appears that the iterative procedure is also converging to the correct answer for the much more reasonable control $u(0) = 10$.

These results suggest that it may not be practical to use 2.2.36 to determine the magnitude of the control in a real problem. It is probably better to try a particular control sequence to see if it works, especially if there are constraints on the magnitude of the control or state. In any case, the object in developing 2.2.36 was merely to show that $x(0)$ could be reconstructed from the available controls and measurements, not to determine conditions on the controls in a practical problem.

2.5 A Special Case

An interesting special case occurs for $F = I$. This is sometimes called an incremental model and can be used to put the equations for a static system in canonical state variable form.

It is clear that neither 2.3.2 nor 2.3.3 can be satisfied since each of the vectors differs by no more than a scalar factor. However, it is still possible to show that the system is observable under

certain conditions. Making the substitution F = I, 2.2.8 can be written

$$
\begin{aligned}
&x'(0)Qx(0) + 2U'(0,0)QX(0) + U'(0,0)QU(0,0) = z(1)\\
&\qquad\qquad\qquad\qquad\vdots\\
&x'(0)Qx(0) + 2U'(n-1,0)Qx(0) + U'(n-1,0)QU(n-1,0) = z(n)\\
&x'(0)Qx(0) + 2U'(n,0)Qx(0) + U'(n,0)QU(n,0) = z(n+1)
\end{aligned}
\tag{2.5.1}
$$

where one extra measurement equation has been added and

$$U(k,0) = B \cdot \sum_{i=1}^{k} u(k-i).$$

Subtracting the first equation from the second, the second from the third, etc., and using the definition of U(k,0),

$$
\begin{aligned}
&2u'(1)B'Qx(0) + u'(1)B'QB[2u(0) + u(1)] = z(2) - z(1)\\
&\qquad\qquad\qquad\qquad\vdots\\
&2u'(n)B'Qx(0) + u'(n)B'QB[2\sum_{i=0}^{n-1} u(i) + u(n)] = z(n+1) - z(n)
\end{aligned}
\tag{2.5.2}
$$

Rearranging,

$$
2 \cdot \begin{bmatrix} u'(1)B'Q \\ \vdots \\ u'(n)B'Q \end{bmatrix} = \begin{bmatrix} z(2) - z(1) \\ \vdots \\ z(n+1) - z(n) \end{bmatrix} - \begin{bmatrix} u'(1)B'QB[2u(0) + u(1)] \\ \vdots \\ u'(n)B'QB[2\sum_{i=0}^{n-1} u(i) + u(n)] \end{bmatrix}
\tag{2.5.3}
$$

This equation is linear in x(0) and has a unique non-trivial solution if

$$U^{*-1} \triangleq \begin{bmatrix} u'(1)B'Q \\ \vdots \\ u'(n)B'Q \end{bmatrix}^{-1} \tag{2.5.4}$$

exists. This leads to the following theorem:

Theorem U^* is nonsingular and the incremental system is completely observable if,

a) the vectors $u(1), u(2), \ldots, u(n)$ are linearly independent, and

b) $(B'Q)^{-1}$ exists.

Proof If $(B'Q)^{-1}$ exists and the vectors $u(1), \ldots, u(n)$ are linearly independent, then the vectors

$u'(1)B'Q, \ldots, u'(n)B'Q$

are also linearly independent since this property is invariant under a nonsingular linear transformation ($B'Q$ in this case). Hence, U^* has n linearly independent rows and thus is nonsingular.

Notice that requirement a) implies that $u(k) \varepsilon E^n$, while b) insures that B is nxn and nonsingular. A necessary condition for the validity of b) is that the system with $F = I$ be completely controllable, as can be seen from the controllability criterion; i.e., the matrix

$$C = [B \vdots IB \vdots \cdots \vdots I^{n-1}B]$$

must have rank n.

2.6 Remarks on Necessary Conditions on the Control in the Stochastic Case

Return now to the noisy system described by equations 1.1.1 and 1.1.3. Obviously, the same measurement value is obtained whether $x(k)$ or $-x(k)$ is used in the measurement equation. Hence, if the uncertainty in the state is large enough compared to the magnitude of the state, the sign of the state will not be apparent from the controls and measurements. It is difficult to see how any estimation scheme could be very successful under such conditions.

The probability of occurrence of this event can be made small by choosing the control large compared to the state uncertainty; i.e., the state must be kept "far enough" away from the origin by the control. The quantity that measures the uncertainty in the state $x(k)$ at time $k-1$ is the one-step prediction error covariance $P(k|k-1)$, for which expressions are derived in the next two chapters. Therefore, as a practical rule of thumb, it is felt that the matrix product $Bu(k-1)$ should have no component with magnitude smaller than the standard deviation of the corresponding one-step prediction error component. That is,

$$|[Bu(k-1)]_i| \geqslant \{[P(k|k-1)]_{ii}\}^{1/2} \quad i = 1,\ldots,n \quad (2.6.1)$$

where $[Bu(k-1)]_i$ is the ith component of the column vector $Bu(k-1)$ and $[P(k|k-1)]_{ii}$ is the ith diagonal element of the square matrix $P(k|k-1)$.

It is shown in Chapter 6, Numerical Results, that a constant control which does not satisfy 2.6.1 for k small but does for k large exhibits poor transient response. Better results are obtained for somewhat larger controls.

2.7 Summary

This chapter forms the basis for the results found in the following two. The most interesting point is that observability requires a control input that is not identically zero in contrast to filtering in completely linear systems. This idea appears to be unknown in the literature.

Some sufficient conditions on the control are derived and it is shown by example that they may be very conservative. Some possible necessary conditions are discussed and a practicable rule is formulated as equation 2.6.1.

Chapter 3

MAXIMUM LIKELIHOOD STATE ESTIMATION

3.1 Introduction

The method of maximum likelihood is quite old and has been discussed by many authors. A good general treatment is contained in Lee [15], p. 56. The basic idea is to define a likelihood function, normally the probability density of the measurements conditioned on some subset of the past and present states. This likelihood function is then maximized with respect to the states to give the values that are most likely to yield the set of observed measurements.

In this chapter, a sequential estimator is formulated using the above technique for the dynamic system described by equations 3.2.1 and 3.2.2. At each time step an implicit nonlinear relation for the maximum of the likelihood function is obtained. In order to avoid having to solve this nonlinear equation, an approximate solution is developed and some conditions for consistency are discussed. Finally, a scalar example is presented.

3.2 Development of the Likelihood Function

In general, there are two major approaches used in the literature to solve the filtering problem. One is to develop an algorithm that estimates only the current state: a so-called point estimator. The other is to estimate the entire sequence of states from time zero to the present; this has been labeled a trajectory estimator and provides a current state estimate by extracting the last estimate in the sequence. Bryson and Frazier [4], Cox [5], and Rauch, Tung and

Striebel [20] have rudimentary discussions on the relationship between these two approaches but give no reasons for choosing one rather than the other.[1] Indeed, it appears that little work has ever been done on the relative merits of the two viewpoints, except in the case of linear systems.

A detailed development of the two likelihood functions will give some insight into how the two approaches compare. First, the system equations are repeated for convenience:

$$x(i+1) = Fx(i) + Bu(i) + Gw(i) \tag{3.2.1}$$

$$i = 0,1\ldots$$

$$z(i+1) = x'(i+1)Qx(i+1) + v(i+1) \tag{3.2.2}$$

Consider now the point estimator. An appropriate likelihood function is

$$L_1[x(k)] = p[z(1),\ldots,z(k)|x(k)] \tag{3.2.3}$$

where k is the present time step and $p(\cdot|\cdot)$ is a joint conditional probability density function.

By the properties of density functions

$$p[z(1),\ldots,z(k)|x(k)] = \int\cdots\int p[z(1),\ldots,z(k)|x(0),\ldots,x(k)] \cdot p[x(0),\ldots,x(k-1)|x(k)]dx(0)\cdots dx(k-1) \tag{3.2.4}$$

Under the assumption that the sequence $\{v(i)\}$ is independent,

$$p[z(1),\ldots,z(w)|x(0),\ldots,x(k)] = p[z(1)|x(1)]\cdots p[z(k)|x(k)] \tag{3.2.5}$$

[1]Actually, the references cited deal with what Sage [23] calls maximum a posteriori (MAP) state estimation. However, the principles are the same as for maximum likelihood state estimation, and the terminology of Rauch, Tung and Striebel [20] is appropriate.

But $p[z(i)|x(i)]$ is clearly gaussian since $v(i)$ is.

Furthermore,

$$E\{z(i)|x(i)\} = E\{x'(i)Qx(i) + v(i)|x(i)\} = x'(i)Qx(i) \tag{3.2.6}$$

and,

$$\begin{aligned} \text{var}\ \{z(i)|x(i)\} &= E\{[z(i) - x'(i)Qx(i)]^2|x(i)\} \\ &= E\{v^2(i)|x(i)\} = T(i) \end{aligned} \tag{3.2.7}$$

so that

$$p[z(i)|x(i)] = [2\pi T(i)]^{-\frac{1}{2}}\exp\{\frac{[z(i) - x'(i)Qx(i)]^2}{2T(i)}\} \tag{3.2.8}$$

and

$$\begin{aligned} &p[z(1),\ldots,z(k)|x(0),\ldots,x(k)] \\ &= \prod_{i=1}^{k} [2\pi T(i)]^{-\frac{1}{2}}\exp\{-\frac{1}{2}\sum_{i=1}^{k}\frac{[z(i) - x'(i)Qx(i)]^2}{T(i)}\} \end{aligned} \tag{3.2.9}$$

The second term under the integral in 3.2.4 may be rewritten in the form

$$\begin{aligned} p[x(0),\ldots,x(k-1)|x(k)] &= \frac{p[x(0),\ldots,x(k)]}{p[x(k)]} \\ &= \frac{p[x(0)]p[x(1)|x(0)]\cdots p[x(k)|x(k-1),\ldots,x(0)]}{p[x(k)]} \end{aligned} \tag{3.2.10}$$

From the fact that the state is a Markov process, it follows that

$$p[x(i)|x(i-1),\ldots,x(0)] = p[x(i)|x(k-1)] \tag{3.2.11}$$

so that

$$p[x(0),\ldots,x(k-1)|x(k)] = \frac{p[x(0)]p[x(1)|x(0)]\cdots p[x(k)|x(k-1)]}{p[x(k)]} \tag{3.2.12}$$

Furthermore, $p[x(i)|x(i-1)]$ is gaussian since the plant noise sequence is, and

$$\begin{aligned} E\{x(i)|x(i-1)\} &= E\{Fx(i-1) + Bu(i-1) + Gw(i-1)|x(i-1)\} \\ &= Fx(i-1) + Bu(i-1) \end{aligned} \tag{3.2.13}$$

and,

$$\begin{aligned} \text{var}\ \{x(i)|x(i-1)\} &= E\{[Gw(i-1)][Gw(i-1)]'|x(i-1)\} \\ &= GS(i-1)G' \end{aligned} \tag{3.2.14}$$

so that,

$$\begin{aligned} p[x(i)|x(i-1)] = {} & [(2\pi)^n \det GS(i-1)G']^{-\frac{1}{2}} \\ & \cdot \exp\{-\tfrac{1}{2}[x(i) - Fx(i-1) - Bu(i-1)]'[GS(i-1)G']^{-1} \\ & \cdot [x(i) - Fx(i-1) - Bu(i-1)]\} \end{aligned} \tag{3.2.15}$$

It follows that

$$\begin{aligned} p[x(0),\ldots,x(k-1)|x(k)] = {} & \frac{1}{p[x(k)]} \prod_{i=1}^{k} [(2\pi)^n \det GS(i-1)G']^{-\frac{1}{2}} \\ & \cdot \exp\{-\tfrac{1}{2} \sum_{i=1}^{k} [x(i) - Fx(i-1) - Bu(i-1)]'[GS(i-1)]^{-1} \\ & \cdot [x(i) - Fx(i-1) - Bu(i-1)]\} \end{aligned} \tag{3.2.16}$$

Equations 3.2.9 and 3.2.16 may now be substituted into 3.2.4 and integrated (in principle) to yield $p[z(1),\ldots,z(k)|x(k)]$. Note,

however, that the integrand in 3.2.4 is a fourth-order exponential in $x(i)$, $i = 1,\ldots,k$, and that it probably cannot be evaluated in closed form. However, it may be possible to obtain approximate solutions to it by numerical methods or investigation of special cases.

Proceeding to the trajectory estimator, one logical likelihood function is

$$L_2[x(0),\ldots,x(k)] = p[z(1),\ldots,z(k)\,|\,x(0),\ldots,x(k)] \tag{3.2.17}$$

An expression for this has already been obtained and is given by equation 3.2.9. Rather than maximizing 3.2.9 directly, the normal method is to work with the logarithm of it. Therefore, the problem of finding the state trajectory that maximizes 3.2.9 is equivalent to minimizing

$$L_2' = \sum_{i=1}^{k} \frac{[z(i) - x'(i)Qx(i)]^2}{T(i)} \tag{3.2.18}$$

with respect to the sequence of states $\{x(i),\ i = 1,\ldots,k\}$. Obviously, L_2' is minimized if the state trajectory is chosen so that[1]

$$x'(i)Qx(i) = z(i) \qquad i = 1,2,\ldots,k \tag{3.2.19}$$

However, there are additional constraints in the form of equation 3.2.1, and these constraints are stochastic in nature because of the

[1]It is possible that the measurement noise can cause $z(i)$ to lie outside the range of the function $x'(i)Qx(i)$ for some i, so that 3.2.19 has no real solutions (e.g., $z(i) < 0$ for Q positive definite). This problem can always be resolved by setting $z(i) = 0$. In any case, the probability of such an event is low if the control sequence is chosen in accordance with the ideas in Chapter 2.

presence of the plant noise. It is not possible in general to satisfy both 3.2.1 and 3.2.19 nor is it possible to know if they are satisfied since the plant noise is unknown. Clearly, any solution to 3.2.1 and 3.2.19 will necessarily be approximate.

One approach to such problems is to replace the stochastic constraints with deterministic ones obtained by substituting the expected value of the plant noise (in this case, zero) for the value of the plant noise. This is a workable approach, and a good one for completely linear systems, and is related to the Certainty Equivalence Principle (Aoki [1], p. 52) of stochastic control theory. However, it will be shown in subsequent sections that when 3.2.1 is substituted into 3.2.19, quadratic forms in the plant noise are obtained due to the quadratic form observer. In such terms it seems more reasonable to use the second moment of the plant noise.

The result of this discussion is that a trajectory estimator can be found by solving a set of deterministic simultaneous equations quite similar to those developed in Chapter 2 in connection with the question of observability.

It is still not clear whether or not the point estimator is inherently better than the trajectory estimator. As Cox [5] and Rauch, Tung and Striebel [20] point out, they are identical for linear systems, but not necessarily so for nonlinear systems. That is, for nonlinear systems, the most likely state does not necessarily lie on the most likely trajectory. It might seem that the point estimator would be somewhat better since it uses the complete probability density function of the plant noise, whereas the trajectory estimator incorporates only

the first two moments. However, as many authors have pointed out, this can also be construed as a disadvantage for the point estimator since it requires complete knowledge of the plant noise distribution, while the trajectory estimator needs only the mean and covariance which are normally relatively easily determined. This is of no importance in this work since the plant noise is assumed to be gaussian, but it is a factor if extensions of these ideas are made.

Maximum likelihood estimation as developed by Fisher in 1912 (Deutsch [7], p. 135) is an embodiment of the idea that estimates should satisfy the equations which govern the observation process when the actual observed values are substituted. This is the real motivation behind the choice 3.2.18 for the likelihood function. From this viewpoint, it is clear that the discussion preceding 3.2.18 and the corresponding discussion concerning the point estimator are merely devices for treating maximum likelihood estimation in a more general way that includes Fisher's ideas as a particular subclass of maximum likelihood estimators. Thus it may be said that the result of using 3.2.18 is a classical maximum likelihood estimator.

The most compelling reason for choosing the trajectory estimator is its immediately apparent relative computational tractability. It has already been shown in Chapter 2 under what conditions a solution exists for a set of equations similar to those of 3.2.1 and 3.2.19. On the other hand, as pointed out during the development, the point estimator requires integration of a fourth-order exponential in the state which probably cannot be done in closed form.

Considering these factors, it was decided to use the trajectory

estimator approach to obtain a maximum likelihood state estimation algorithm. The remainder of the chapter is devoted to that goal.

One final point needs to be made before the material in the next section. It is clear from the above discussion, and the fact is used in Section 3.5, that equations 3.2.1 and 3.2.19 can be solved approximately for any state x(i), where i = 1,...,k. Therefore, in order to provide continuity with Chapter 2 and to make the results therein more easily applicable, the development of the algorithm begins in Section 3.3 by finding an estimate of the initial state x(0) that approximately satisfies 3.2.1 and 3.2.19. For this development an expression for x(i) in terms of x(0) and the control and plant noise sequences is necessary and is given by

$$x(i) = F^i x(0) + \sum_{j=0}^{i-1} F^j Bu(i-j-1) + \sum_{j=0}^{i-1} F^j Gw(i-j-1) \qquad i = 0,1,\ldots \tag{3.2.20}$$

3.3 Calculation of the Initial State Estimate

In this section, the likelihood function L_2' given by equation 3.2.18 along with the constraints 3.2.1 are used to calculate an estimate of x(0) based on the first n measurements. Later sections show how to update this estimate to the present time step and make the calculation sequential.

For notational ease, it is convenient at this point to make some definitions similar to equation 2.2.4. In connection with 3.2.20, let

$$U(k,0) \triangleq \sum_{i=0}^{k} F^i Bu(k-1) \tag{3.3.1}$$

and

$$W(k,0) \triangleq \sum_{i=0}^{k} F^i Gw(k-i) \tag{3.3.2}$$

Putting 3.2.20, 3.3.1, and 3.3.2 into 3.2.19 with k = n,

$$\begin{aligned} &[Fx(0) = U(0,0) = W(0,0)]' Q[Fx(0) = U(0,0) + W(0,0)] = z(1) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ &[F^n x(0) = U(n-1,0) + W(n-1,0)]' Q[F^n x(0) + U(n-1,0), + W(n-1,0)] = z(n) \end{aligned} \tag{3.3.3}$$

Upon multiplying out the terms in the brackets, the equations will contain terms like

$$2U'(k,0)QU(k,0)$$

$$2W'(k,0)QF^n x(0)$$

and $\quad W'(k,0)QW(k,0)$

since the plant noise is unknown, w(i) is replaced by its expected value, 0, in the first two expressions:

$$\begin{aligned} 2U'(k,0)QW(k,0) &\rightarrow 2U'(k,0)QE\{W(k,0)\} \\ &= 2U'(k,0)QE\{\sum_{i=0}^{k} F^i Gw(k-i)\} \\ &= 2U'(k,0)Q \sum_{i=0}^{k} F^i GE\{w(k-i)\} \\ &= 0 \end{aligned} \tag{3.3.4}$$

where a → b indicates that a is replaced by b.

Similarly, $\quad 2W'(k,0)QF^n x(0) \rightarrow 0 \qquad (3.3.5)$

However, the third expression can be written in the form

$$\begin{aligned} W'(k,0)QW(k,0) &= \text{trace}\ \{QW(k,0)W'(k,0)\} \\ &\rightarrow \text{trace}\ \{QE[W(k,0)W(k,0)]\} \end{aligned}$$

But $$E[W(k,0)W'(k,0)] = E\{\sum_{i=0}^{k}\sum_{j=0}^{k} F^i Gw(k-i)w'(k-j)G'F'^j\}$$

$$= \sum_{i=0}^{k}\sum_{j=0}^{k} F^i GE\{w(k-i)w'(k-j)\}G'F'^j$$

$$= \sum_{i=0}^{k} F^i GS(k-i)G'F'^i$$

$$\triangleq Y(k,0) \tag{3.3.6}$$

by the assumptions on the statistics of the plant noise.

Hence, $$W'(k,0)QW(k,0) \to \text{trace}\ \{QY(k,0)\} \tag{3.3.7}$$

Utilizing these results in 3.3.3 and rearranging,

$$\begin{bmatrix} 2U'(0,0)QF \\ \vdots \\ 2U'(n-1,0)QF^n \end{bmatrix} \hat{x}(0|n) = \begin{bmatrix} z(1) \\ \vdots \\ z(n) \end{bmatrix} - \begin{bmatrix} U'(0,0)QU(0,0) \\ \vdots \\ U'(n-1,0)QU(n-1,0) \end{bmatrix}$$

$$- \begin{bmatrix} \text{trace}\ QY(0,0) \\ \vdots \\ \text{trace}\ QY(n-1,0) \end{bmatrix} - \begin{bmatrix} \hat{x}'(0|n)F'QF\hat{x}(0|n) \\ \vdots \\ \hat{x}'(0|n)F'^n QF^n\hat{x}(0|n) \end{bmatrix} \tag{3.3.8}$$

where $x(0)$ in 3.3.3 has been replaced with $\hat{x}(0|n)$ (following the notation of Meditch [16]) since the solution to 3.3.3 is an approximation to the solution of 3.3.3 based on the first n measurements.

The equation 3.3.8 is identical in structure to equation 2.2.13 with the exception of the constant matrix

$$Y^* \triangleq [\text{trace}\ QY(0,0),\ldots,\ \text{trace}\ QY(n-1,0)]' \tag{3.3.9}$$

on the right hand side which does not depend on $\hat{x}(0|n)$. Therefore, the conditions for convergence of the successive approximation technique to the solution $\hat{x}(0|n)$ of 3.3.8 are exactly the same. That is, the sequence of approximations defined by

$$\hat{x}_{i+1}(0|n) = \frac{1}{2}A^{-1}\{Z^* - U^* - Y^* - X^*[\hat{x}_i(0|n)]\} \quad i = 0,1,\ldots \tag{3.3.10}$$

with $\hat{x}_o(0|n) = 0$, Y^* given by 3.3.9 and Z^*, U^*, X^*, A defined in equations 2.2.9 - 12, respectively, converges to the true solution of 3.3.8 if the conditions on A in Chapter 2 are satisfied.

3.4 Sequential Form of the Estimator

So far, all that has been obtained is an estimate of the initial state based on the first n measurements. The problem now arises of what to do with the next measurement. It could be incorporated into equation 3.3.8 and the pseudo-inverse used to arrive at an analog of 3.3.10. However, this has the difficulty of increasing the dimension of 3.3.10 by one with each new measurement, thus changing the form of the estimation equation and eventually making calculation of the pseudo-inverse unwieldy. Furthermore, the procedure is not sequential. The purpose of this section is to alleviate these problems.

At time step $k = n+1$, an estimate of $x(0)$, $\hat{x}(0|n)$, is available and an estimate of $x(1)$ can be calculated from it:

$$\hat{x}(1|n) = F\hat{x}(0|n) + Bu(0) \tag{3.4.1}$$

where again the notation $\hat{x}(k|j)$ is that of Meditch [16] and denotes an estimate of $x(k)$ based on the first j measurements. This equation can be combined with the measurement at n+1 in the following fashion. Write the (n+1)st measurement in terms of $x(1)$:

$$[F^n x(1) + U(n,1) + W(n,1)]' Q[F^n x(1) + U(n,1) + W(n,1)] = z(n+1) - v(n+1) \tag{3.4.2}$$

where use has been made of the measurement equation

$$z(n+1) = x'(n+1)Qx(n+1) + v(n+1) \tag{3.4.3}$$

Performing the multiplications and rearranging,

$$2[U(n,1) + W(n,1)]' QF^n x(1) = z(n+1) - v(n+1)$$
$$- [U(n,1) + W(n,1)]' Q[U(n,1) + W(n,1)]$$
$$- x'(1)F'^n QF^n x(1) \tag{3.4.4}$$

Now form the partitioned matrix equation

$$\begin{bmatrix} I \\ \hline [U(n,1) + W(n,1)]' QF^n \end{bmatrix} \hat{x}(1) \simeq \begin{bmatrix} F \\ \hline 0 \end{bmatrix} \hat{x}(0|n) + \begin{bmatrix} Bu(0) \\ \hline 0 \end{bmatrix} \tag{3.4.5}$$
$$+ \frac{1}{2} \begin{bmatrix} 0 \\ \hline z(n+1) - [U(n,1) + W(n,1)]' Q[U(n,1) + W(n,1)] - \hat{x}'(1)F'^n QF^n \hat{x}(1) \end{bmatrix}$$

where $\hat{x}(1)$ is an estimate of $x(1)$ based on $\hat{x}(0|n)$ and the (n+1)st measurement. Notice that $v(n+1)$ has been replaced by its mean value, 0, since it is unknown. Now let

$$A_1 \triangleq \begin{bmatrix} I \\ \hline [U(n,1) + W(n,1)]' QF^n \end{bmatrix} \tag{3.4.6}$$

A_1 is obviously of full rank. Then,

$$(A_1' A_1)\hat{x}(1) \simeq F\hat{x}(0|n) + Bu(0) + \frac{1}{2} F'^n Q[U(n,1) + W(n,1)]$$
$$\{z(n+1) - [U(n,1) + W(n,1)]' Q[U(n,1) + W(n,1)]$$
$$- \hat{x}'(1)F'^n QF^n \hat{x}(1)\} \tag{3.4.7}$$

The calculation of the matrix inverse can be avoided by using the matrix inversion lemma (see, for example, Sage and Melsa [23], p. 499).

$$(A_1' A_1) = I + F'^n Q[U(n,1) + W(n,1)]\ [U(n,1) + W(n,1)]' QF^n \tag{3.4.8}$$

by the definition of A_1. Applying the matrix inversion lemma,

$$(A_1'A_1)^{-1} = I - F'^{n}Q[U(n,1) + W(n,1)]$$

$$\cdot \{[U(n,1) + W(n,1)]'QF^{n}F'^{n}Q[U(n,1) + W(n,1)] + 1\}^{-1}$$

$$\cdot [U(n,1) + W(n,1)]'QF^{n} \tag{3.4.9}$$

The required inverse has been reduced to simple division. Substituting 3.4.9 into 3.4.7 and simplifying and rearranging,

$$\hat{x}(1) \simeq \hat{x}(1|n) + \frac{1}{2} K(n+1)\{z(n+1) - 2[U(n,1)+ W(n,1)]'QF^{n}\hat{x}(1|n)$$

$$- [U(n,1) + W(n,1)]'Q[U(n,1) + W(n,1)] - \hat{x}'(1)F'^{n}QF^{n}\hat{x}(1)\} \tag{3.4.10}$$

with $\quad K(n+1) \triangleq F'^{n}Q[U(n,1) + W(n,1)]$

$$\cdot \{[U(n,1) + W(n,1)]'QF^{n}F'^{n}Q[U(n,1) + W(n,1)] + 1\}^{-1} \tag{3.4.11}$$

As before, since W(n,1) is unknown, it is convenient to replace the expressions in which it occurs with easily calculated expected values. The resulting relations are taken as the definition of the maximum likelihood estimate of the state at time 1 based on the first n+1 measurements, i.e.,

$$\hat{x}(1|n+1) \triangleq \hat{x}(1|n) + \frac{1}{2}\hat{K}(n+1)\{z(n+1) - 2U'(n,1)QF^{n}\hat{x}(1|n)$$

$$U'(n,1)QU(n,1) - \hat{x}'(1|n+1)F'^{n}QF^{n}\hat{x}(1|n+1)$$

$$- \text{trace } Q \sum_{i=0}^{n-1} F^{i}GS(n-i)G'F'^{i}\} \tag{3.4.12}$$

with

$$\hat{K}(n+1) = F'^{n}QU(n,1)\{U'(n,1)QF^{n}F'^{n}QU(n,1)$$

$$+ \text{trace } [QF^{n}F'^{n}Q \sum_{i=0}^{n-1} F^{i}GS(n-i)G'F'^{i}] + 1\}^{-1} \tag{3.4.13}$$

Reasoning by induction, the maximum likelihood estimate of x(k) based on the first k+n measurements is given by

$$\hat{x}(k|\,k+n) = \hat{x}(k|\,k+n-1) + \frac{1}{2}\hat{K}(k+n)\{z(k+n)$$

$$- 2U'(k+n-1,k)QF^n\hat{x}(k|\,k+n-1)$$

$$- U'(k+n-1,k)QU(k+n-1,k)$$

$$- \text{trace } Q \sum_{i=0}^{n-1} F^i GS(k+n-i)G'F'^i$$

$$- \hat{x}(k|\,k+n)F'^nQF^n\hat{x}(k|\,k+n)\} \qquad (3.4.14)$$

where

$$\hat{K}(k+n) = F'^nQU(k+n-1,k)\{U'(k+n-1,k)QF^nF'^nQU(k+n-1,k)$$

$$+ \text{trace } [QF^nF'^nQ \sum_{i=0}^{n-1} F^i GS(k+n-i)G'F'^i] + 1\}^{-1} \qquad (3.4.15)$$

Equations 3.4.14 and 3.4.15 form the sequential estimate for $\hat{x}(k|\,k+n)$. However, observe that $\hat{x}(k|\,k+n)$ appears on both sides of 3.4.14; at this point, it is possible to develop an iterative procedure for solving for the estimate. Such a solution has several disadvantages, the main one being that $\hat{x}(k|\,k+n)$ is a fixed lag estimator of x(k) with lag n. The purpose of the next section is to remove this objection.

3.5 Construction of the Filtering Estimator

Equations 3.4.14 and 3.4.15 can be updated to time step k+n using 3.2.20. However, in order to accurately describe the system behavior, it is necessary to go back to equations 3.4.10 and 3.4.11 and include the plant noise. Generalizing from 3.4.1,

$$\bar{x}(k+n|\,k+n-1) \triangleq F^n\hat{x}(k|\,k+n-1) + U(k+n-1,k) + W(k+n-1,k) \tag{3.5.1}$$

and

$$\bar{x}(k+n|\,k+n) \triangleq F^n\hat{x}(k|\,k+n) + U(k+n-1,k) + W(k+n-1,k) \tag{3.5.2}$$

Inserting these into 3.4.10 at time step k,

$$\begin{aligned}
\bar{x}(k+n|\,k+n) = {} & \bar{x}(k+n|\,k+n-1) + \frac{1}{2} F^n K(k+n) \\
& \cdot\{z(k+n) - 2[U(k+n-1) + W(k+n-1,k)]'Q \\
& \cdot [\bar{x}(k+n|\,k+n-1) - U(k+n-1,k) - W(k+n-1,k)] \\
& - [U(k+n-1,k) + W(k+n-1,k)]'Q[U(k+n-1,k) + W(k+n-1,k)] \\
& - [\bar{x}(k+n|\,k+n) - U(k+n-1,k) - W(k+n-1,k)]'Q \\
& \cdot [\bar{x}(k+n|\,k+n) - U(k+n-1,k) - W(k+n-1,k)]\}
\end{aligned} \tag{3.5.3}$$

where K(k+n) is given by 3.4.11 with n replaced by k+n-1. As before, since the plant noise is unknown, the expressions in which $W(\cdot,\cdot)$ appears are replaced by expected values that are relatively easy to calculate. To denote this change, $\bar{x}(\cdot|\,\cdot)$ is replaced by $\hat{x}(\cdot|\,\cdot)$, so that after simplifying,

$$\begin{aligned}
\hat{x}(k+n|\,k+n) = {} & \hat{x}(k+n|\,k+n-1) + \frac{1}{2} F^n\hat{K}(k+n) \\
& \cdot\{z(k+n) - 2U'(k+n-1,k)Q[\hat{x}(k+n|\,k+n-1) - \hat{x}(k+n|\,k+n)] \\
& -\hat{x}'(k+n|\,k+n)Q\hat{x}(k+n|\,k+n)\}
\end{aligned} \tag{3.5.4}$$

$$\begin{aligned}
\hat{K}(k+n) = {} & F'^{\,n}QU(k+n-1,k)\{U'(k+n-1,k)QF^nF'^{\,n}QU(k+n-1,k) \\
& + \text{trace}\,[QF^nF'^{\,n}Q\sum_{i=0}^{n-1} F^iGS(k+n-i)G'F'^{\,i}] + 1\}^{-1}
\end{aligned} \tag{3.5.5}$$

and $\hat{x}(k+n|\,k+n-1) = F\hat{x}(k+n-1|\,k+n-1) + Bu(k+n-1)$ (3.5.6)

Again, equations 3.5.4, 5, and 6 are taken as the defining relations for the maximum likelihood state estimate. These equations are clearly preferable over 3.4.14 and 3.4.15 from a filtering viewpoint since they provide an estimate of the current state rather than a fixed lag estimate.

Note that $\hat{x}(k+n|\,k+n)$ appears on both sides of 3.5.4. Therefore, either a method will have to be given for solving for the state estimate or an approximate algorithm developed. The purpose of the next section is to obtain such an approximate algorithm.

The derivations of equations 3.5.4, 5, and 6 show that the time index k ranges over the positive integers starting at k = 1. The algorithm is started with $\hat{x}(n|\,n)$, which is obtained by updating $\hat{x}(0|\,n)$ using 3.4.1 n times, while $\hat{x}(0|\,n)$ is found using the method given in Section 3.3.

3.6 An Approximate Solution to the Estimator Equation

As already mentioned, 3.5.4 is a nonlinear implicit equation in the current state estimate. It may be difficult and time-consuming to find the solution. Therefore, it is useful to try to find an approximate solution to 3.5.4.

The possibility that immediately suggests itself is to replace $\hat{x}(k+n|\,k+n)$ on the right side of 3.5.4 with the one-step predicted estimate $\hat{x}(k+n|\,k+n-1)$. Then the approximate solution to 3.5.4 is

$$\hat{x}_A(k+n|\,k+n) = \hat{x}_A(k+n|\,k+n-1) + \frac{1}{2} F^n\hat{K}(k+n)$$

$$\cdot\{z(k+n) - \hat{x}_A'(k+n|\,k+n-1)Q\hat{x}_A(k+n|\,k+n-1)\} \qquad (3.6.1)$$

$$\hat{x}_A(k+n|\,k+n-1) = F\hat{x}_A(k+n-1|\,k+n-1) + Bu(k+n-1) \tag{3.6.2}$$

and $\hat{K}(k+n)$ still given by 3.5.5. The subscript A on the state estimates indicates that they are only approximate solutions to 3.5.4. It remains to be shown whether or not the estimator is consistent. Let

$$e(k+n|\,k+n-1) \triangleq x(k+n) - \hat{x}_A(k+n|\,k+n-1) \tag{3.6.3}$$

and

$$e(k+n|\,k+n) \triangleq x(k+n) - \hat{x}_A(k+n|\,k+n) \tag{3.6.4}$$

These are the one-step prediction error and filtering error respectively. Then, using 3.6.1,

$$e(k+n|\,k+n) = e(k+n|\,k+n-1) - \frac{1}{2} F^n\hat{K}(k+n)$$

$$\cdot\{z(k+n) - \hat{x}_A'\,(k+n|\,k+n-1)Q\hat{x}_A(k+n|\,k+n-1)\} \tag{3.6.5}$$

recalling that

$$x'(k+n)Qx(k+n) + v(k+n) = z(k+n)$$

3.6.5 can be written

$$e(k+n|\,k+n) = e(k+n|\,k+n-1) - \frac{1}{2} F^n\hat{K}(k+n)$$

$$\cdot\{v(k+n) + [x(k+n) + \hat{x}_A(k+n|\,k+n-1)]'Qe(k+n|\,k+n-1)\} \tag{3.6.6}$$

where use has been made of the property of quadratic forms that

$$x_1'Qx_1 - x_2'Qx_2 \equiv (x_1 + x_2)'Q(x_1 - x_2) \tag{3.6.7}$$

Rearranging 3.6.6,

$$e(k+n|\,k+n) = \left\{I - \frac{1}{2}F^n\hat{K}(k+n)\,[x(k+n) + \hat{x}_A(k+n|\,k+n-1)]'\,Q\right\}$$

$$\cdot\, e(k+n|\,k+n-1) - \frac{1}{2}F^n\hat{K}(k+n)v(k+n) \tag{3.6.8}$$

To simplify notation, let

$$B_{k+n} \triangleq I - \frac{1}{2}F^n\hat{K}(k+n)\,[x(k+n) + \hat{x}_A(k+n|\,k+n-1)]'\,Q \tag{3.6.9}$$

The one-step prediction error is related to the filtering error at the previous time step by

$$e(k+n|\,k+n-1) = Fe(k+n-1|\,k+n-1) + Gw(k+n-1) \tag{3.6.10}$$

Therefore, the difference equation for the filtering error is

$$e(k+n|\,k+n) = B_{k+n}Fe(k+n-1|\,k+n-1) + B_{k+n}Gw(k+n-1)$$

$$- \frac{1}{2}F^n\hat{K}(k+n)v(k+n) \tag{3.6.11}$$

If the system matrices, the control input, and the noises are all finite, then the last two terms in 3.6.11 look like finite disturbance inputs. In this case, the most desirable error propagation, i.e., maximum decrease of the magnitude of the error on the average at each step, is obtained by insuring that the eigenvalues of $B_{k+n}F$ lie as close as possible to the origin; in particular, they must lie inside the unit circle. This will be the case if

$$\|B_{k+n}F\| < 1 \tag{3.6.12}$$

At this point, it is not clear whether or not a control sequence exists such that 3.6.12 is satisfied in general. Also, since 3.6.12 should be checked at each time step, it is unwieldy to use it to

determine the proper control. Therefore, perhaps a better method of investigating the consistency of 3.6.1 is a simulation. This is one of the purposes of Chapter 6, Numerical Results.

3.7 A Scalar Example

The following example is intended to point out the amount and kind of computation required to estimate the state of a simple system using the previous algorithm.

Let the system be described by

$$x(k+1) = \frac{1}{2}x(k) + u(k) + w(k) \qquad (3.7.1)$$

$$k = 0,1, \ldots$$

$$z(k+1) = \frac{1}{2}x^2(k+1) \qquad (3.7.2)$$

with $x(k)$ a scalar. Suppose that $x(0) = -1$, and arbitrarily choose $u(k) = 1$, $k = 0,1,\ldots$. Let the plant noise covariance be constant at $S = .01$, and further suppose that the particular plant noise sample function for this example is $w(k) = 0$, $k = 0,1,\ldots$. Therefore,

$$x(1) = \frac{1}{2}(-1) + 1 = \frac{1}{2} \qquad z(1) = \frac{1}{8}$$

$$x(2) = \frac{1}{2}(\frac{1}{2}) + 1 = \frac{5}{4} \qquad z(2) = \frac{25}{32}$$

$$x(3) = \frac{1}{2}(\frac{5}{4}) + 1 = \frac{13}{8} \qquad z(2) = \frac{169}{128}$$

$$x(4) = \frac{1}{2}(\frac{13}{8}) + 1 = \frac{29}{16} \qquad z(4) = \frac{841}{512}$$

First, the initial state estimate $x(0)$ must be calculated.

$U(0,0) = u(0) = 1$

$Z^* = z(1) = \frac{1}{8}$

$Y(0,0) = .01$

trace $QY(0,0) = Y^* = .005$

$A = 2U'(0,0)QF = 2(1)(.5)(.5) = .5 \qquad A^{-1} = 2$

Using these values in 3.3.10,

$$\hat{x}_1(0|1) = \frac{1}{2}(2)\{.125 - 1 - .005\} = -.880$$

$$\hat{x}_2(0|1) = \hat{x}_1(0|1) - \frac{1}{2}A^{-1}X^*[\hat{x}_1(2|1)]$$

$$= -.880 - (-.880)(.125)(-.880) = -.880 - .0969 = -.9769$$

$$\hat{x}_3(0|1) = \hat{x}_1(0|1) - \frac{1}{2}A^{-1}X^*[x_2(0|1)]$$

$$= -.880 - (-.9769)(.125)(-.9769) = -.880 - .1192 = -.9992$$

apparently, $\hat{x}(0|1)$ is approaching the value -1 as expected. Therefore, take $\hat{x}(0|1) = -1$. The state estimate at $k = 1$ is found from the update equation.

$$\hat{x}(1|1) = F\hat{x}(0|1) + Bu(0) = \frac{1}{2}(-1) + 1 = \frac{1}{2} = \hat{x}_A(1|1)$$

The next step is to calculate the subsequent state estimates. To do this, the filter gain $\hat{K}(\cdot)$ must be determined. Notice that, since the system is time-invariant and the control is constant, $\hat{K}(\cdot)$ is also time-invariant. From 3.5.5

$$\hat{K} = \frac{1}{2}(\tfrac{1}{2})(1)\{(1)(\tfrac{1}{2})(\tfrac{1}{2})(\tfrac{1}{2})(\tfrac{1}{2})(1) + (\tfrac{1}{2})(\tfrac{1}{2})(\tfrac{1}{2})(\tfrac{1}{2})(.01) + 1\}^{-1}$$

$$= \frac{1}{4}(1.063125)^{-1} \cong .235$$

$$\hat{x}_A(2|1) = F\hat{x}_A(1|1) + Bu(1) = \frac{1}{2}(\tfrac{1}{2}) + 1 = \frac{5}{4}$$

$$\hat{x}_A(2|2) = \frac{5}{4} + \frac{1}{2}(\tfrac{1}{2})(.235)\{\tfrac{25}{32} - \tfrac{25}{32}\} = \frac{5}{4}$$

etc. The last two answers are obtained via the approximate algorithm in Section 3.6. In the noiseless case, the accuracy of the first few estimates is directly dependent on the ability of the algorithm to accurately calculate $\hat{x}(0|n)$. It is also apparent why a constant control is so attractive for a time-invariant system; the filter gain only has to be computed once.

3.8 Summary and Conclusions

In this chapter, the technique of maximum likelihood has been used to derive a nonlinear implicit equation, 3.5.4, for the current state estimate. It is possible to solve this equation directly using methods well known in the theory of nonlinear equations (see e.g., Ostrowski [18], Saaty [21], or Stern [26]). However, any such solution will be expensive in terms of time and computational effort. Therefore, an approximate algorithm (Section 3.6) has been obtained to circumvent this difficulty. The consistency of this estimator is still in doubt, and is one of the topics covered in Chapter 6, Numerical Results.

The forms of the estimators, 3.5.4 and 3.6.1, and the equation for the filter gain, 3.5.5, exhibit a property that is recurrent throughout this thesis. That is, the control cannot be identically zero for these algorithms. Otherwise, $\hat{K}(k) \equiv 0$ and there is no measurement feedback to correct the one-step predicted estimate. Under this condition, there is little reason to expect the state estimates to be very good except, possibly, in the case of small noise. Even then, divergence will occur after a long enough time if the plant noise is not identically zero.

Practically speaking, the implementation of the results of this chapter is somewhat cumbersome, due mainly to the fact that a different procedure is used to calculate the estimate of $x(0)$ than the other state estimates. Furthermore, no state estimate is available until time step n. The object of the next chapter is to eliminate these difficulties.

Chapter 4

A BAYESIAN APPROACH TO STATE ESTIMATION

4.1 Introduction

In this chapter, a more general approach to the problem of state estimation is taken in an effort to make better use of available data. The basic groundwork has been laid in a paper by Ho and Lee [8].

Assume that some set of measurements $Z^*(k) = [z(1), z(2), \ldots, z(k)]$ is given, and it is desired to determine the best (in some sense) estimate of the current state $x(k)$ based on these measurements. The Bayesian approach is to determine the a posteriori conditional density function $p[x(k)|Z^*(k)]$, since it contains all the pertinent statistical information. Knowing $p[x(k)|Z^*(k)]$, various estimates of $x(k)$ can be determined. Two of these are the conditional mean (CM) estimate given by

$$\hat{x}_{CM}(k|k) \triangleq \int x(k)\ p[x(k)|Z^*(k)]\ dx(k) \tag{4.1.1}$$

and the maximum a posteriori (MAP), or most probable, estimate which is the mode of $p[x(k)|Z^*(k)]$ and is calculated from

$$\left.\frac{\partial p[x(k)|Z^*(k)]}{\partial x(k)}\right|_{\hat{x}_{MAP}(k|k)} = 0 \tag{4.1.2}$$

with

$$\left.\frac{\partial^2 p[x(k)|Z^*(k)]}{\partial x^2(k)}\right|_{\hat{x}_{MAP}(k|k)} < 0 \tag{4.1.3}$$

In problems with nonlinear dynamics and/or observers, it is

usually impractical if not impossible to calculate the conditional mean because the posterior density function is not representable in closed form. A more insidious difficulty is that such density functions are not reproducing; that is, the form of the density function, hence the form of the estimator, changes with time. This makes sequential implementation extremely difficult. In such cases it is common to make the ad hoc assumption that $p[x(k)|Z^*(k)]$ has a reproducing form, usually gaussian. It will become apparent that the problem of state estimation with quadratic form observers possesses all of these difficulties, but some measure of progress can still be made.

4.2 Calculation of the Posterior Density Function

For the convenience of the reader, the system model is restated in abbreviated form.

$$x(k+1) = Fx(k) + Bu(k) + Gw(k) \tag{4.2.1}$$

$$z(k+1) = x'(k+1)Qx(k+1) + v(k+1) \tag{4.2.2}$$

The initial state $x(0)$ is zero-mean and gaussian with covariance $P(0)$. The plant noise sequence $\{w(k)\}$ and measurement noise sequence $\{v(k)\}$ are also zero-mean and gaussian, and in addition are white with respective covariances $S(k)$ and $T(k)$. It is further assumed that the initial state and noises are all independent.

The following result, which depends on Bayes' rule and recognition of the fact that the process described by 4.2.1 is a Markov sequence, is derived by Ho and Lee [8], and in slightly different form by Sage

and Melsa [23], p. 274. The posterior density function can be written in the recursive form

$$p[x(k+1)|Z^*(k+1)] = \frac{p[z(k+1)|x(k+1)]}{p[z(k+1)|Z^*(k)]}$$

$$\cdot\int p[x(k+1)|x(k)]p[x(k)|Z^*(k)]dx(k) \tag{4.2.3}$$

with

$$p[x(0)|Z^*(0)] = p[x(0)] \tag{4.2.4}$$

First, the posterior density function will be computed explicitly for the first two time steps. Then, using these results, a form for the density function at later times will be assumed. By the statement of the problem, $x(0)$ has a zero-mean gaussian distribution with covariance $P(0)$. Therefore,

$$p[x(0)] = [(2\pi)^n \det P(0)]^{-\frac{1}{2}} \exp\{-\frac{1}{2} x'(0)P^{-1}(0)x(0)\} \tag{4.2.5}$$

Since $x(1) = Fx(0) + Bu(0) + Gw(0)$, $p[x(1)|x(0)]$ is gaussian because $w(0)$ is.[1] Moreover

$$E\{x(1)|x(0)\} = Fx(0) + Bu(0) \tag{4.2.6}$$

and $\mathrm{var}\{x(1)|x(0)\} \triangleq E[\{x(1) - E[x(1)|x(0)]\}\{x(1) - E[x(1)|x(0)]\}'|x(0)]$

$$= GS(0)G' \tag{4.2.7}$$

so that

$$p[x(1)|x(0)] = [(2\pi)^n \det GS(0)G']^{-\frac{1}{2}}$$

$$\cdot\exp\{-\frac{1}{2}[x(1) - Fx(0) - Bu(0)]'[GS(0)G']^{-1}[x(1) - Fx(0) - Bu(0)]\} \tag{4.2.8}$$

[1] More precisely, $p[x(1)|x(0),u(0)]$ is gaussian. However, $u(0)$ [and later $u(k)$] may be treated like any other known system parameters and not included in the list of conditioning variables.

A similar relationship holds at all other times also:

$$p[x(k+1)|x(k)] = [(2\pi)^n \det GS(k)G']^{-\frac{1}{2}}$$

$$\cdot\exp\{-\tfrac{1}{2}[x(k+1) - Fx(k) - Bu(k)]'[GS(k)G']^{-1}[x(k+1) - Fx(k) - Bu(k)]\} \tag{4.2.9}$$

The density function $p[z(1)|x(1)]$ is also gaussian, since

$z(1) = x'(1)Qx(1) + v(1)$ and $v(1)$ is gaussian.

$$E\{z(1)|x(1)\} = E\{x'(1)Qx(1) + v(1)|x(1)\} = x'(1)Qx(1) \tag{4.2.10}$$

from the assumption that $\{v(k)\}$ has zero-mean.

$$\mathrm{var}\{z(1)|x(1)\} \triangleq E[\{z(1) - x'(1)Qx(1)\}^2|x(1)] = E\{v^2(1)\} = T(1) \tag{4.2.11}$$

Therefore,

$$p[z(1)|x(1)] = [2\pi T(1)]^{-\frac{1}{2}} \exp\{-\frac{1}{2T(1)}[z(1) - x'(1)Qx(1)]^2\} \tag{4.2.12}$$

Again, a similar equation is true for all time:

$$p[z(k)|x(k)] = [2\pi T(k)]^{-\frac{1}{2}} \exp\{-\frac{1}{2T(k)}[z(k) - x'(k)Qx(k)]^2\} \tag{4.2.13}$$

The density $p[z(k+1)|Z^*(k)]$ can be found (conceptually) from

$$p[z(k+1)|Z^*(k)] = \int p[x(k+1), z(k+1)|Z^*(k)]dx(k+1)$$

$$= \int p[z(k+1)|x(k+1)] \int p[x(k+1)|x(k)]p[x(k)|Z^*(k)]dx(k)dx(k+1) \tag{4.2.14}$$

$$p[z(1)|Z^*(0)] = p[z(1)] = \int p[z(1)|x(1)] \int p[x(1)|x(0)]p[x(0)]\cdot dx(0)dx(1)$$

$$= [2\pi T(1)]^{-\frac{1}{2}} [(2\pi)^n \det GS(0)G']^{-\frac{1}{2}} [(2\pi)^n \det P(0)]^{-\frac{1}{2}} \tag{4.2.15}$$

$$\cdot\int_{-\infty}^{\infty} \exp\{-\frac{1}{2T(1)}[z(1) - x'(1)Qx(1)]^2\}$$

$$\cdot\int_{-\infty}^{\infty} \exp\{-\frac{1}{2}[x(1) - Fx(0) - Bu(0)]'[GS(0)G']^{-1}[x(1) - Fx(0) - Bu(0)]$$

$$-\frac{1}{2}x'(0)P^{-1}(0)x(0)\}dx(0)dx(1) \tag{4.2.16}$$

The second integral can be evaluated by completing the square in the exponent on x(0). When this is done

$$p[z(1)|Z^*(0)] = [2\pi T(1)]^{-\frac{1}{2}}\{(2\pi)^n \det[FP(0)F' + GS(0)G']\}^{-\frac{1}{2}}$$

$$\cdot\int_{-\infty}^{\infty}\exp\{-\frac{1}{2T(1)}[z(1) - x'(1)Qx(1)]^2$$

$$-\frac{1}{2}[x(1) - Bu(0)]'[FP(0)F' + GS(0)G']^{-1}[x(1) - Bu(0)]\}dx(1) \quad (4.2.17)$$

The integral in 4.2.17 has no closed form solution, but it does have the property that it is not a function of x(1). This fact will be useful in what follows. It is now possible to write down the density function $p[x(1)|Z^*(1)]$:

$$p[x(1)|Z^*(1)] = \frac{[2\pi T(1)]^{-\frac{1}{2}}}{p[z(1)|Z^*(0)]}\{(2\pi)^n \det[FP(0)F' + GS(0)G']\}^{-\frac{1}{2}}$$

$$\cdot\exp\{-\frac{1}{2}[x(1) - Bu(0)]'[FP(0)F' + GS(0)G']^{-1}[x(1) - Bu(0)]$$

$$-\frac{1}{2T(1)}[z(1) - x'(1)Qx(1)]^2\} \quad (4.2.18)$$

The next density function required is $p[x(2)|Z^*(1)]$ which can be calculated from the rule

$$p[x(2)|Z^*(1)] = \int p[x(2)|x(1)]\, p[x(1)|Z^*(1)]dx(1)$$

$$= \frac{[2\pi T(1)]^{-\frac{1}{2}}}{p[z(1)|Z^*(0)]}\{(2\pi)^n \det[FP(0)F' + GS(0)G']\}^{-\frac{1}{2}}[(2\pi)^n \det GS(1)G']^{-\frac{1}{2}}$$

$$\cdot\int_{-\infty}^{\infty}\exp\{-\frac{1}{2T(1)}[z(1) - x'(1)Qx(1)]^2$$

$$-\frac{1}{2}[x(1) - Bu(0)]'[FP(0)F' + GS(0)G']^{-1}[x(1) - Bu(0)] \quad (4.2.19)$$

$$-\frac{1}{2}[x(2) - Fx(1) - Bu(1)]'[GS(1)G']^{-1}[x(2) - Fx(1) - Bu(1)]\}dx(1)$$

It is believed that this integral cannot be evaluated in closed

form. In order to continue then, it is necessary to make some assumptions concerning the form of the density functions at later time steps. A common approach, and the one used by Sage and Melsa [23], p. 466, in their derivation of approximate conditional mean filtering for nonlinear systems, is to assume that $p[x(k)|Z^*(k)]$ is gaussian. However, 4.2.18 shows that this assumption may not be very good for this problem, at least at time step $k = 1$. In fact, $p[x(1)|Z^*(1)]$ may not even be unimodal. Therefore, an equally reasonable and perhaps more accurate assumption is that $p[x(2)|Z^*(1)]$ is gaussian since $p[x(1)|Z^*(0)]$ is. Under this assumption, only the conditional mean and covariance of $x(2)$ given $Z^*(1)$ are needed to determine $p[x(2)|Z^*(1)]$. Approximations to these statistics are not difficult to calculate using 4.2.19. The conditional mean is

$$E\{x(2)|Z^*(1)\} = \int_{x(2)} \int_{x(1)} x(2)p[x(2)|Z^*(1)]dx(1)dx(2)$$

$$= \int_{x(2)} \int_{x(1)} x(2)p[x(2)|x(1)]p[x(1)|Z^*(1)]dx(1)dx(2) \qquad (4.2.20)$$

Interchanging the order of integration,

$$E\{x(2)|Z^*(1)\} = \int_{x(1)} p[x(1)|Z^*(1)] \int_{x(2)} p[x(2)|x(1)]dx(2)dx(1) \qquad (4.2.21)$$

The second integral is easily evaluated from equation 4.2.9 to give

$$E\{x(2)|Z^*(1)\} = \int_{x(1)} [Fx(1) + Bu(1)]p[x(1)|Z^*(1)]dx(1) \qquad (4.2.22)$$

Therefore, by the definition of conditional mean,

$$E\{x(2)|Z^*(1)\} = FE\{x(1)|Z^*(1)\} + Bu(1) \qquad (4.2.23)$$

As with the integral in 4.2.19, it is believed that $E\{x(1)|Z^*(1)\}$

cannot be found analytically. Assume for the moment, however, that an estimate of $E\{x(1)|Z^*(1)\}$, call it $\hat{x}(1|1)$, is available by whatever means. One of the functions of the next section is to show how to calculate $\hat{x}(\cdot|\cdot)$. Then,

$$E\{x(2)|Z^*(1)\} \cong \hat{x}(2|1) \triangleq F\hat{x}(1|1) + Bu(1) \tag{4.2.24}$$

In a similar fashion,

$$\text{var}\{x(2)|Z^*(1)\} = \int_{x(1)} p[x(1)|Z^*(1)] \int_{x(2)} [x(2) - E\{x(2)|Z^*(1)\}]$$

$$\cdot[x(2) - E\{x(2)|Z^*(1)\}]' p[x(2)|x(1)]dx(2)dx(1) \tag{4.2.25}$$

After carrying out the indicated operations,

$$\text{var}\{x(2)|Z^*(1)\} = F\text{var}\{x(1)|Z^*(1)\}F' + GS(1)G' \tag{4.2.26}$$

But, $\text{var}\{x(1)|Z^*(1)\} \cong P(1|1)$

$$\triangleq E\{[x(1) - \hat{x}(1|1)][x(1) - \hat{x}(1|1)]'|Z^*(1)\} \tag{4.2.27}$$

so that

$$\text{var}\{x(2)|Z^*(1)\} \cong P(2|1) \triangleq FP(1|1)F' + GS(1)G' \tag{4.2.28}$$

Therefore, $p[x(2)|Z^*(1)]$ is approximately given by

$$p[x(2)|Z^*(1)] \cong K_1 \exp\{-\tfrac{1}{2}[x(2) - \hat{x}(2|1)]'P^{-1}(2|1)[x(2)-\hat{x}(2|1)]\} \tag{4.2.29}$$

where K_1 is an appropriate normalization constant. Using 4.2.29, it is apparent from the development of $p[x(1)|Z^*(1)]$ that $p[x(2)|Z^*(2)]$ has the approximate form

$$p[x(2)|Z^*(2)] \cong K_2 \exp\{-\frac{1}{2T(2)}[z(2) - x'(2)Qx(2)]^2$$

$$-\tfrac{1}{2}[x(2) - \hat{x}(2|1)]'P^{-1}(2|1)[x(2) - \hat{x}(2|1)]\} \tag{4.2.30}$$

where again K_2 is a normalization constant. It is not much of an inductive leap now to make the formal assumption that

$$p[x(k)|Z^*(k-1)] = K_3 \exp\{-\tfrac{1}{2}[x(k) - \hat{x}(k|k-1)]' P^{-1}(k|k-1) \cdot [x(k) - \hat{x}(k|k-1)]\} \tag{4.2.31}$$

with $$P(k|k-1) = FP(k-1|k-1)F' + GS(k-1)G' \tag{4.2.32}$$

and $$P(k|k) \triangleq E\{[x(k) - \hat{x}(k|k)][x(k) - \hat{x}(k|k)]'|Z^*(k)\} \tag{4.2.33}$$

and $$\hat{x}(k|k-1) = F\hat{x}(k-1|k-1) + Bu(k-1) \tag{4.2.34}$$

all of which imply that

$$p[x(k)|Z^*(k)] = K_4 \exp\{-\frac{1}{2T(k)}[z(k) - x'(k)Qx(k)]^2 - \tfrac{1}{2}[x(k) - \hat{x}(k|k-1)]' P^{-1}(k|k-1)[x(k) - \hat{x}(k|k-1)]\} \tag{4.2.35}$$

Equations 4.2.31-35 form the core for the remaining developments in this chapter. Notice that by the arguments already used in connection with 4.2.19 that it will not be possible to calculate $E\{x(k)|Z^*(k)\}$ using 4.2.35. In this regard, these assumptions are not as nice as those used by Sage and Melsa [23], p. 466, in their derivation of the approximate conditional mean filtering algorithm. However, equations 4.2.31-35 may lead to more accurate or easily implementable algorithms if some other estimate is used.

4.3 Calculation of the Maximum A Posteriori State Estimate

One way to generate the quantity $\hat{x}(k|k)$ needed for the density functions derived in Section 4.2 is to use 4.1.2 and 4.1.3 to obtain the maximum a posteriori (MAP) estimate of $x(k)$. Applying 4.1.2,

$$\frac{2}{T(k)}[z(k) - \hat{x}'_{MAP}(k|k)Q\hat{x}_{MAP}(k|k)][\hat{x}'_{MAP}(k|k)Q] - [\hat{x}_{MAP}(k|k) - \hat{x}_{MAP}(k|k-1)]' P^{-1}(k|k-1) = 0 \tag{4.3.1}$$

where

$$\hat{x}_{MAP}(k|k-1) = F\hat{x}_{MAP}(k-1|k-1) + Bu(k-1) \tag{4.3.2}$$

Or, rearranging

$$\hat{x}_{MAP}(k|k) = \hat{x}_{MAP}(k|k-1) + \frac{2}{T(k)} P(k|k-1)Q\hat{x}_{MAP}(k|k)$$

$$\cdot[z(k) - \hat{x}'_{MAP}(k|k)Q\hat{x}_{MAP}(k|k)] \tag{4.3.3}$$

where use has been made of the technique of taking the derivative of the logarithm of $p[x(k)|Z^*(k)]$.

Equation 4.3.3 is a nonlinear algebraic equation in $\hat{x}_{MAP}(k|k)$ that may be difficult to solve, particularly if its solution is not unique. This will be the case if $p[x(k)|Z^*(k)]$ is multimodal which is entirely possible. For this reason 4.3.3 will not be very useful from a practical standpoint. For the purpose of comparison, in the chapter on numerical results a standard function minimization routine will be used to try to find the global minimum of the exponent in 4.2.35 at each time step. It will be seen that such a procedure is very expensive in terms of computer time and may require a large number of iterations. It therefore is important to develop some alternative algorithm that is more feasible computationally.

4.4 A Geometric Approach to Estimating the State

The aim of this section is to find an approximation to the value of $x(k)$ that yields the maximum in 4.2.35 and is easy to calculate. Consider the exponent in 4.2.35 at time step k; call it $f(x)$, dropping the time indices for convenience.

$$f(x) \triangleq \frac{1}{2T}[z - x'Qx]^2 + \frac{1}{2}[x - \hat{x}]'P^{-1}[x - \hat{x}] \tag{4.4.1}$$

As already discussed, minimizing $f(x)$ is the same as maximizing

$p[x(k)|Z^*(k)]$; therefore, an approximation to the minimum of $f(x)$ is desired. $f(x)$ can be written as the sum of two functions of x:

$$f(x) = f_1(x) + f_2(x) \tag{4.4.2}$$

where $f_1(x) = \frac{1}{2T}[z - x'Qx]^2$ (4.4.3)

and $f_2(x) = \frac{1}{2}[x - \hat{x}]'P^{-1}[x - \hat{x}]$ (4.4.4)

Both f_1 and f_2 are non-negative functions of x. One approach to finding an approximate minimum of $f(x)$ is to calculate the minima of $f_1(x)$ and $f_2(x)$ and choose an intermediate value as the estimate of the minimum of $f(x)$.

The minimum value of $f_1(x)$ occurs at all x such that the equation

$$x'Qx = z \tag{4.4.5}$$

is satisfied, if any such x exists. The only possibility of non-existence of solutions to 4.4.5 occurs when measurement noise causes the value of z to lie outside the range of the function $x'Qx$ (e.g., a negative z for positive definite Q). This pathological case can always be resolved by setting $z = 0$. If Q is positive or negative definite, this results in $x = 0$, a unique solution to 4.4.5.

Equation 4.4.5 is that of a hyper-conic section in state space. Different types of surfaces are formed according to the definiteness of Q. If Q is positive definite, the surface is a hyper-ellipsoid; if it is negative definite, a hyper-hyperboloid, etc.

The minimum value of $f_2(x)$ always occurs at $x = \hat{x}$. For the two-dimensional case, the functions $x'Qx = z$, $x = \hat{x}$, and $x'Qx = \hat{z}$ (where $\hat{z}$ is given by $\hat{x}'Q\hat{x} = \hat{z}$) with z and $\hat{x}$ given typical values and Q positive definite are plotted in Figure 4.4.1.

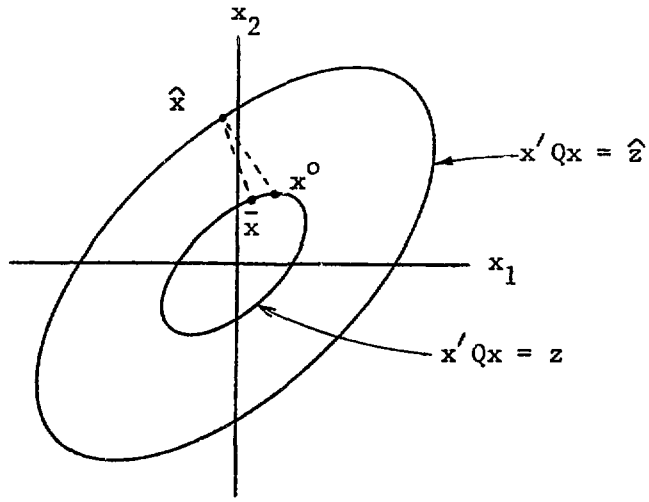

Figure 4.4.1 2-D Illustration of $x' Qx = z$, $x' Qx = \hat{z}$, and $x = \hat{x}$

In the figure the dotted line $\bar{x}\hat{x}$ is normal to the $x' Qx = z$ surface. Therefore, the length of $\bar{x}\hat{x}$ is the shortest distance from $\hat{x}$ to the surface. It is reasonable to expect that x^*, the minimum of $f(x)$, lies somewhere close to this line segment. If x^* is assumed to lie on $\bar{x}\hat{x}$ between $\bar{x}$ and $\hat{x}$, then it can be written

$$x^* = a\hat{x} + (1-a)\bar{x} \tag{4.4.6}$$

where a is a scalar, $a \varepsilon [0,1)$. The normal of unit length perpendicular to $x' Qx = z$ at $\bar{x}$ is given by

$$\bar{h} = \left.\frac{\nabla(x' Qx - z)}{\|\nabla(x' Qx - z)\|}\right|_{x = \bar{x}} = \frac{Q\bar{x}}{\|Q\bar{x}\|} \tag{4.4.7}$$

Observe that $\bar{x}$, which is unknown, appears on the right side of this equation. Using this equation and referring to Figure 4.4.1,

$$\hat{x} = \bar{x} + c'\bar{h} \tag{4.4.8}$$

or $$\hat{x} = \bar{x} + cQ\bar{x} = (I + cQ)\bar{x} \tag{4.4.9}$$

where c is a scalar that can be determined from the fact that $\bar{x}' Q\bar{x} = z$ in the following manner. If $(I + cQ)^{-1}$ exists, then

$$\bar{x} = (I + cQ)^{-1}\hat{x} \tag{4.4.10}$$

and $\hat{x}'(I + cQ)'^{-1}Q(I + cQ)^{-1}\hat{x} = z$ (4.4.11)

In principle, this equation can be solved for c and 4.4.10 used to calculate $\bar{x}$. However, 4.4.11 has no nice closed form solution for c and it requires the existence and calculation of $(I + cQ)^{-1}$. The following development is designed to circumvent these problems.

Instead of calculating $\bar{x}$, find a value x^o, if it exists, that satisfies $x^{o\prime}Qx^o = z$ and lies on the normal to the surface $x'Qx = \hat{z}$ passing through $\hat{x}$. The unit normal to $x'Qx = \hat{z}$ at $\hat{x}$ is given by

$$h = \frac{\nabla(x'Qx - \hat{z})}{\|\nabla(x'Qx - \hat{z})\|}\Bigg|_{x = \hat{x}} = \frac{Q\hat{x}}{\|Q\hat{x}\|} \tag{4.4.12}$$

Usint this relation and Figure 4.4.1, x^o is given by

$$x^o = \hat{x} + bh \tag{4.4.13}$$

where again b is a scalar that can be determined from the fact that $x^{o\prime}Qx^o = z$.

$$x^{o\prime}Qx^o = (\hat{x} + bh)'Q(\hat{x} + bh) = z$$

Or, $bh'Qh + 2bh'Q\hat{x} + [\hat{x}'Q\hat{x} - z] = 0$ (4.4.14)

This is a scalar quadratic equation for b; its solution by the quadratic formula is

$$b = \frac{-h'Q\hat{x} \pm \{[h'Q\hat{x}]^2 - h'Qh[\hat{x}'Q\hat{x} - z]\}^{1/2}}{h'Qh} \tag{4.4.15}$$

As already mentioned, x^o may not exist; that is, the normal to $x'Qx = \hat{z}$ at $\hat{x}$ may not intersect the $x'Qx = z$ surface. In this case, 4.4.15 will result in a complex number for b. One method of resolving this problem is to modify b so that the resulting x^o is real and as close to the calculated complex x^o as possible. This is easily done by simply setting the imaginary part of b to zero. However, this may result in state estimates that are further away from $\hat{x}$ than the distance

between the surfaces $x'Qx = z$ and $x'Qx = \hat{z}$. This can be partially corrected by taking the absolute value of the quantity under the radical in 4.4.15, thus always making b real. Choice of the root of minimum absolute value in 4.4.15 yields a state estimate that is closer to $\hat{x}$.

In any problem for which $\hat{x}$ and x^o are reasonably close together, such as low noise systems, the case of complex b is not very important from a practical point of view. The pathological situation occurs only for a relatively small set of $\hat{x}$ and z. If desirable, the one-step predicted estimate (for a plant controllable over some sequence) can be made to lie near one of the eigenvectors of Q by use of the control input. Since the surfaces of constant z all have coincident normals (namely, the eigenvectors) where they intersect the eigenvectors, x^o always exists for such $\hat{x}$.

In all cases, the root of smallest absolute value should be chosen in 4.4.15. One reason has already been discussed. The other reason is that the remaining root, for b real, corresponds to the intersection of the line $\hat{x}x^o$ with another point on the hyper-surface which is of no interest.

Finally, after calculating b and h, x^* can be found from 4.4.6 and 4.4.13 with $\bar{x}$ replaced by x^o.

$$x^* = a\hat{x} + (1-a)(\hat{x} + bh)$$

$$= \hat{x} + (1-a)bh \qquad (4.4.16)$$

All that remains is a rational choice for the scalar a. From equation 4.4.1, a should be chosen according to the relative uncertainties in the measurement z and the one-step predicted estimate

$\hat{x}$. That is, if $\|T\| >> \|P\|$, then x^* should be much closer to $\hat{x}$ than to $\bar{x}$, and vice-versa. In practice, however, P is not known with any great degree of accuracy, if at all. Furthermore, calculation of P increases the complexity of the algorithm. Therefore, a guess for a, say $a = \frac{1}{2}$, may yield as good results as a value of a based on calculation of P. In the chapter on numerical results, this algorithm is implemented with $a = \frac{1}{2}$ and the results of a simulation compared with those of an approximate conditional mean filter and the maximum likelihood filter developed in the preceeding chapter.

Although the foregoing arguments are based on a two-dimensional picture, the results carry over to the n-dimensional case with only one minor modification. The surfaces $x'Qx = z$ and $x'Qx = \hat{z}$ are no longer conic sections but hyper-conic sections.

Reinserting the time indices and summarizing,

$$\hat{x}_{AMAP}(k|k) = \hat{x}_{AMAP}(k|k-1) + (1-a)b(k)h(k) \tag{4.4.17}$$

$$a \in [0,1]$$

$$\hat{x}_{AMAP}(k|k-1) = F\hat{x}_{AMAP}(k-1|k-1) + Bu(k-1) \tag{4.4.18}$$

$$h(k) = \frac{Q\hat{x}_{AMAP}(k|k-1)}{\|Q\hat{x}_{AMAP}(k|k-1)\|} \tag{4.4.19}$$

$$\begin{aligned} b(k) = \frac{1}{h'(k)Qh(k)} \{ & -h'(k)Q\hat{x}_{AMAP}(k|k-1) \\ & \pm [\{h'(k)Q\hat{x}_{AMAP}(k|k-1)\}^2 - h'(k)Qh(k) \\ & \cdot \{\hat{x}'_{AMAP}(k|k-1)Q\hat{x}_{AMAP}(k|k-1) - z(k)\}]^{\frac{1}{2}}\} \end{aligned} \tag{4.4.20}$$

where the root with minimum magnitude is used.

Finally,

$$\hat{x}_{AMAP}(0|0) = E\{x(0)\} = 0 \tag{4.4.21}$$

4.5 Calculation of the Approximate Error Covariance Matrix

In order that $P(k|k-1)$ may be used to find a, and for purposes of comparison, a method is presented here for its approximate calculation. Since.

$$P(k|k-1) = FP(k-1|k-1)F' + GS(k-1)G' \tag{4.5.1}$$

it is only necessary to compute $P(k-1|k-1)$, the error covariance of $x(k-1)$ conditioned on $Z^*(k-1)$.

By construction, x^* given by 4.4.16 is an approximate solution to 4.3.3 which is repeated for convenience (without the MAP subscript).

$$\hat{x}(k|k) = \hat{x}(k|k-1) + \frac{2}{T(k)}\,[z(k) - \hat{x}'(k|k)Q\hat{x}(k|k)]P(k|k-1)Q\hat{x}(k|k) \tag{4.3.3}$$

If x^* is close to a solution of 4.3.3 then the following results will yield an approximate answer for the covariance of the error $x(k)-x^*(k)$. Let

$$e(k+1|k+1) \triangleq x(k+1) - \hat{x}(k+1|k+1) \tag{4.5.2}$$

so that

$$\begin{aligned} e(k|k) = x(k) - \hat{x}(k|k) &= Fx(k-1) + Bu(k-1) + Gw(k-1) \\ &\quad -F\hat{x}(k-1|k-1) - Bu(k-1) \\ &\quad - \frac{2}{T(k)}\,[z(k) - \hat{x}'(k|k)Q\hat{x}(k|k)]P(k|k-1)Q\hat{x}(k|k) \\ &= Fe(k-1|k-1) + Gw(k-1) \\ &\quad - \frac{2}{T(k)}\,[z(k) - \hat{x}'(k|k)Q\hat{x}(k|k)]P(k|k-1)Q\hat{x}(k|k) \end{aligned} \tag{4.5.3}$$

But

$$\begin{aligned} z(k) - x(k|k)Q\hat{x}(k|k) &= x'(k)Qx(k) + v(k) - \hat{x}'(k|k)Qx(k|k) \\ &= [x(k) + \hat{x}(k|k)]'Qe(k|k) + v(k) \end{aligned} \tag{4.5.4}$$

Inserting 4.5.4 into 4.5.3,

$$I + \frac{2}{T(k)} P(k|k-1)Q\hat{x}(k|k)[x(k) + \hat{x}(k|k)]' Qe(k|k)$$

$$= Fe(k-1|k-1) + Gw(k-1) - \frac{2}{T(k)} v(k)P(k|k-1)Q\hat{x}(k|k) \quad (4.5.5)$$

Since $x(k)$ is unknown, it is replaced with the best available estimate of $x(k)$, $\hat{x}(k|k)$.

$$\bar{e}(k|k) = B_k^{-1}\{F\bar{e}(k-1|k-1) + Gw(k-1) - \frac{2}{T(k)} v(k)P(k|k-1)Q\hat{x}(k|k)\} \quad (4.5.6)$$

where $B_k = I + \frac{4}{T(k)} P(k|k-1)Q\hat{x}(k|k)\hat{x}'(k|k)Q$ (4.5.7)

and the bar over the error indicates that $x(k)$ has been replaced by $\hat{x}(k|k)$.

Note that B_k has no random components at time step k. The error covariance is defined as

$$P(k|k) = E\{[e(k|k) - E\{e(k|k)|Z^*(k)\}][e(k|k) - E\{e(k|k)|Z^*(k)\}]' |Z^*(k)\} \quad (4.5.8)$$

Now $E\{e(k|k)|Z^*(k)\} = E\{x(k) - \hat{x}(k|k)|Z^*(k)\}$

$$= E\{x(k)|Z^*(k)\} - \hat{x}(k|k) \quad (4.5.9)$$

Recall that the conditional expectation in (4.5.9) is not calculable. Therefore, in order to arrive at an approximation for $P(k|k)$, it is assumed that

$$\hat{x}(k|k) \cong E\{x(k)|Z^*(k)\} \quad (4.5.10)$$

so that

$$P(k|k) \cong E\{\bar{e}(k|k)\bar{e}'(k|k)|Z^*(k)\} \quad (4.5.11)$$

Assumption 4.5.10 may be very good if the function 4.4.1 is unimodal and not too unsymmetrical. Otherwise, 4.5.10 can be considerably in error since the solution to 4.3.3 lies on one of the extrema of 4.4.1, while $E\{x(k)|Z^*(k)\}$ and $x^*(k)$ do not necessarily. In any case,

using 4.5.11,

$$P(k|k) \cong B_k^{-1} E\{[F\bar{e}(k-1|k-1) + Gw(k-1) - \frac{2}{T(k)} v(k)P(k|k-1)Q\hat{x}(k|k)]$$

$$\cdot[F\bar{e}(k-1|k-1) + Gw(k-1) - \frac{2}{T(k)} v(k)P(k|k-1)Q\hat{x}(k|k)]' |Z^*(k)\}$$

$$\cdot B_k'^{-1} \tag{4.5.12}$$

Using the assumed independences of the plant and measurement noises

$$P(k|k) \cong B_k^{-1}\{FP(k-1|k-1)F' + GS(k-1)G' + \frac{4}{T(k)} P(k|k-1)Q\hat{x}(k|k)x'(k|k)Q$$

$$\cdot P(k|k-1)\}B'^{-1}_k \tag{4.5.13}$$

Or

$$P(k|k) = B_k^{-1}\{I + \frac{4}{T(k)} P(k|k-1)Q\hat{x}(k|k)\hat{x}'(k|k)Q\}P(k|k-1)B_k'^{-1} \tag{4.5.14}$$

Using equation 4.5.7,

$$P(k|k) = P(k|k-1)B_k'^{-1}$$

$$= \{P^{-1}(k|k-1) + \frac{4}{T(k)} Q\hat{x}(k|k)\hat{x}'(k|k)Q\}'^{-1}$$

$$= \{P^{-1}(k|k-1) + \frac{4}{T(k)} Q\hat{x}(k|k)\hat{x}'(k|k)Q\}^{-1} \tag{4.5.15}$$

Calculation of the matrix inverse can be avoided by use of the matrix inversion lemma

$$\Rightarrow P(k|k) = P(k|k-1) - 4P(k|k-1)Q\hat{x}(k|k)[4\hat{x}'(k|k)QP(k|k-1)Q\hat{x}(k|k) + T(k)]^{-1}$$

$$\cdot\hat{x}'(k|k)QP(k|k-1) \tag{4.5.16}$$

and

$$P(k|k-1) = FP(k-1|k-1)F' + GS(k-1)G' \tag{4.5.17}$$

Equations 4.5.16 and 4.5.17 are approximate recursive relations for the error covariances of $\hat{x}_{MAP}(k|k)$ and $x^*(k)$, the solutions to 4.3.3 and 4.4.16 respectively. If the error covariance of $x^*(k)$ is desired, then $x^*(k)$ should replace $\hat{x}(k|k)$ in 4.5.16.

4.6 Some Remarks on Error Propagation

Equation 4.5.6 is the difference equation that describes approximately the trajectory of the state estimation error. It is important to notice two things: 1) 4.5.6 is a linear difference equation, and 2) the last two terms on the right hand side appear as disturbance inputs. If these disturbances are finite, then the approximate error will decrease on the average if the unforced part of 4.5.6 is uniformly asymptotically stable-in-the-large. Since

$$\bar{e}(k|k) = B_k^{-1}F\bar{e}(k-1|k-1) \tag{4.6.1}$$

is the unforced part of 4.5.6 and is linear, uniform asymptotic stability-in-the-large is equivalent to the stability of the origin (see, e.g., Aoki [1], Kalman and Bertram [11], or Kushner [13] and [14]). Therefore, a necessary and sufficient condition for the stability of 4.6.1 is that the eigenvalues of $B_k^{-1}F$ all lie inside the unit circle. This will occur if

$$\|B_k^{-1}F\| < 1 \tag{4.6.2}$$

Or, using the properties of norms,

$$\|F\| < \|B_k^{-1}\|^{-1} \tag{4.6.3}$$

is a sufficient condition for the stability of 4.6.1. Recalling the definition of B_k, 4.5.7, it is clear that $\|B_k^{-1}\|^{-1}$ is indirectly controllable by $u(k-1)$; i.e., by making $u(k-1)$ large, $\hat{x}(k|k)$ can be moved as far as desired away from the origin. Thus, 4.6.3 can be satisfied for large enough control.

Of course, 4.6.3 is structured such that the required $u(k-1)$ cannot be computed ahead of time. A particular $u(k-1)$ must be used to obtain $x(k)$, $z(k)$, and $\hat{x}(k|k)$. Then these values can be inserted

into the right hand side of 4.6.3 to ascertain whether or not it is true. A better procedure is to try to find a $u(k-1)$ such that the resulting $x(k|k)$ is very likely to satisfy 4.6.3. This can be done by replacing $\hat{x}(k|k)$ with $\hat{x}(k|k-1) = F\hat{x}(k-1|k-1) + Bu(k-1)$. Then a candidate for $u(k-1)$ can be substituted into 4.6.3, to see if it holds before applying it to the system. If it does not, then another $u(k-1)$ can be tried, etc., until 4.6.3 is satisfied.

In actual practice, such a calculation may not be necessary. In the first simulation in the chapter on numerical results, it is found that a constant control, in this case a scalar of $u(k) = 1$ for all k, is sufficient for the algorithm to be statistically consistent. It is entirely possible that some such convenient control will work in the general case.

4.7 A Scalar Example

Consider the problem posed in Section 3.7. Remember that

$$x(k+1) = \frac{1}{2}x(k) + u(k) + w(k) \tag{4.7.1}$$

$$z(k+1) = \frac{1}{2}x^2(k+1) \tag{4.7.2}$$

with $x(0) = -1$, $u(k) = 1$ for all k, and $S(k) = .01$ but $w(k) \equiv 0$. The last group of assumptions is made to allow the trial of equation 4.5.3.

a) Calculation of MAP estimate using 4.3.3

$$\hat{x}_{MAP}(1|0) = Bu(0) = 1$$

$$P(1|0) = \frac{1}{2}(1)\frac{1}{2} + .01 = .26$$

$$\Rightarrow \quad \hat{x}_{MAP}(1|1) = 1 + 200(.26)\,\frac{1}{2}\,\hat{x}_{MAP}(1|1)[.125 - \frac{1}{2}\,\hat{x}^2_{MAP}(1|1)]$$

or

$$\hat{x}_{MAP}(1|1) = 1 + 3.25\,\hat{x}_{MAP}(1|1) - 13\,\hat{x}^3_{MAP}(1|1)$$

The solution to this cubic equation is approximately $\hat{x}_{MAP}(1|1) = .558$

and was obtained by trial and error. The next step is to calculate $P(1|1)$ using 4.5.16.

$$P(1|1) = .26 - 4(.26)^2(\tfrac{1}{2})^2(.558)^2[4(.558)^2(\tfrac{1}{2})^2(.26) + .01]^{-1}$$

$$= .26 - .2314 = .0286$$

$$P(2|1) = \tfrac{1}{4}(.0286) + .01 = .01715$$

$$\hat{x}_{MAP}(2|1) = \tfrac{1}{2}(.558) + 1 = 1.279$$

$$\hat{x}_{MAP}(2|2) = 1.279 + 200(.01715)(\tfrac{1}{2})\hat{x}_{MAP}(2|2)[.78125 - \tfrac{1}{2}\hat{x}_{MAP}(2|2)]$$

$$= 1.279 + 1.3398\,\hat{x}_{MAP}(2|2) - .8575\,\hat{x}^3_{MAP}(2|2)$$

The solution to this cubic is approximately $\hat{x}_{MAP}(2|2)=1.2578$. Comparison with the example in Chapter 3 shows that the first two state estimates are close to the two values of the states and appear to be getting closer with time.

b) Calculation of the approximate MAP (AMAP) estimate using 4.4.16.

$$\hat{x}_{AMAP}(1|0) = Bu(0) = 1$$

$h = 1$ so that

$$b(1) = 2[-\tfrac{1}{2} \pm \{\tfrac{1}{4} - \tfrac{1}{2}(\tfrac{1}{2} - \tfrac{1}{8})\}^{\frac{1}{2}}] = -\tfrac{3}{2} \text{ or } -\tfrac{1}{2}$$

$$\Rightarrow \quad b(1) = -\tfrac{1}{2}$$

Then, using $a = \frac{1}{2}$ as will be done in Chapter 6,

$$\hat{x}_{AMAP}(1|1) = x^*(1) = 1 + \tfrac{1}{2}(-\tfrac{1}{2}) = \tfrac{3}{4}$$

$$\hat{x}_{AMAP}(2|1) = \tfrac{1}{2}(\tfrac{3}{4}) + 1 = \tfrac{11}{8}$$

$$b(2) = 2[-\tfrac{11}{16} \pm \{\tfrac{121}{256} - \tfrac{25}{32}\}^{\frac{1}{2}}] = -\tfrac{1}{8} \text{ or } -\tfrac{21}{8}$$

$$\Rightarrow \quad b(2) = -\tfrac{1}{8} \text{ so that}$$

$$\hat{x}_{AMAP}(2|2) = \tfrac{11}{8} - \tfrac{1}{8} = \tfrac{10}{8} = \tfrac{5}{4} \qquad \text{which is exact.}$$

For this example at least, the approximate MAP (AMAP) estimator gives estimates that are about as good as those of the MAP estimator. In addition, the AMAP computational procedure is much simpler, primarily because of the need to solve a cubic equation for the MAP estimate. Also, the MAP estimate requires calculation of the error covariance, while the AMAP does not.

4.8 Summary and Conclusions

In this chapter, an approximate form for the posterior density function of the state conditioned on the measurements is obtained based on logical considerations. After noting that finding the value of the state directly that maximizes this density function is a difficult problem, a simple approximation to this maximum is derived, along with conditions on the control for estimator consistency. Finally, a scalar example is included to demonstrate the relative computational difficulties of the MAP and AMAP estimators.

The AMAP estimator is very appealing because of its simplicity. It should be somewhat faster than a linearized Kalman filter since it requires no error covariance calculation.

Chapter 5

ALGORITHM TABULATION AND STRUCTURAL COMPARISONS

5.1 Introduction

The purpose of this chapter is twofold; first, the basic problem is restated and each of the algorithms discussed are tabulated for handy future reference. In addition, first- and second-order approximate conditional mean filters (Sage and Melsa [23], p. 466) are presented for comparison. Second, the differences and similarities among the algorithms are brought out with a special view toward computational difficulties.

5.2 Problem Restatement

The original problem is restated in tabular form in Table 5.2.1.

Table 5.2.1 Problem Statement

Plant Dynamics	$x(k+1) = Fx(k) + Bu(k) + Gw(k) \qquad k = 0,1, \ldots$
Observation	$z(k+1) = x'(k+1)Qx(k+1) + v(k+1), \; Q = Q'$
Initial State Statistics	$E\{x(0)\} = 0 \qquad var\{x(0)\} = P(0)$ gaussian
Plant Noise Statistics	$E\{w(k)\} = 0$ for all k $\quad var\{w(k)\} = S(k)$ gaussian, white
Measurement Noise Statistics	$E\{v(k)\} = 0$ for all k $\quad var\{v(k)\} = T(k)$ gaussian, white
Desired Result	An estimate of the current state based on all the known control inputs and available measurements.

5.3 Algorithm Tabulation

The next four tables contain the three algorithms obtained in this thesis plus the first- and second-order approximate conditional mean filters of Sage and Melsa [23], p. 466.

Table 5.3.1 The Maximum Likelihood (ML) Filter Algorithm

Main Filter Equation	$\hat{x}_{ML}(k+n\mid k+n) = \hat{x}_{ML}(k+n\mid k+n-1) + \frac{1}{2} F^n \hat{K}(k+n)\{z(k+n) - \hat{x}'_{ML}(k+n\mid k+n-1)Q\hat{x}_{ML}(k+n\mid k+n-1)\} \quad k = 1,2, \ldots$
Subsidiary Equation	$\hat{K}(k+n) = F'^{n}QU(k+n-1,k) \cdot\{U'(k+n-1,k)QF^nF'^{n}QU(k+n-1,k) + \text{trace } QF^nF'^{n}Q \sum_{i=0}^{n-1} F^iGS(k+n-i)G'F'^{i} + 1\}^{-1}$ $U(k,0) = \sum_{i=0}^{k} F^iBu(k-i)$ $\hat{x}_{ML}(k+n\mid k+n-1) = F\hat{x}_{ML}(k+n-1\mid k+n-1) + Bu(k+n-1)$
Initialization	$\hat{x}(n-1\mid n) = F^{n+1}x(0\mid n) + \sum_{i=0}^{n} F^iBu(n-i)$ $\hat{x}_{i+1}(0\mid n) = \frac{1}{2} A^{-1}\{Z^* - U^* - Y^* - X^*[x_i(0\mid n)]\}$ $i = 0,1, \ldots \quad \hat{x}_o(0\mid n) = 0$ $A' = [F'QU(0,0) \vdots \text{---} \vdots F'^{n}QU(n-1,0)]$ $Z^* = [z(1), \ldots, z(n)]'$ $U^* = [U'(0,0)QU(0,0), \ldots, U'(n-1,0)QU(n-1,0)]'$ $Y^* = [\text{trace } QY(0,0), \ldots, \text{trace } QY(n-1,0)]'$ $X^*[x] = [x'F'QFx, \ldots, x'F'^{n}QF^nx]'$ $Y(k,0) = \sum_{i=0}^{k} F^iGS(k-i)G'F'^{i}$

Table 5.3.1, continued

Control Constraints	For initialization, A^{-1} exist and satisfy 2.2.36. (This generally results in a very conservative control policy.) After initialization, the control should be somewhat larger than the plant noise (Section 2.6).
Text Reference	Chapter 3 and parts of Chapter 2

Table 5.3.2	The Maximum A Posteriori (MAP) Filter Algorithm
Main Filter Equation	$\hat{x}_{MAP}(k\|k) = \hat{x}_{MAP}(k\|k-1) + \frac{2}{T(k)} P(k\|k-1)Q\hat{x}_{MAP}(k\|k) \cdot \{z(k) - \hat{x}'_{MAP}(k\|k)Q\hat{x}_{MAP}(k\|k)\}$
Subsidiary Equations	$P(k\|k-1) = FP(k-1\|k-1)F' + GS(k-1)G'$ $P(k\|k) = P(k\|k-1) - 4P(k\|k-1)Q\hat{x}_{MAP}(k\|k) \cdot \{\hat{x}'_{MAP}(k\|k)QP(k\|k-1)Q\hat{x}_{MAP}(k\|k) + T(k)\}^{-1} \cdot \hat{x}'_{MAP}(k\|k)QP(k\|k-1)$ $\hat{x}_{MAP}(k\|k-1) = F\hat{x}_{MAP}(k-1\|k-1) + Bu(k-1)$
Initialization	$\hat{x}_{MAP}(0\|0) = E\{x(0)\} = 0$ $P(0\|0) = P(0)$
Control Constraints	$\|F\| < \|B_k^{-1}\|^{-1}$ where $B_k = I + \frac{4}{T(k)} P(k\|k-1)Q\hat{x}(k\|k)\hat{x}'(k\|k)Q$ i.e., the control must be chosen large enough so that the above holds in order for the error to decrease on the average at each time step.
Text Reference	Chapter 4

Table 5.3.3 The Approximate Maximum A Posteriori (AMAP) Filter Algorithm

Main Filter Equation	$\hat{x}_{AMAP}(k\|k) = \hat{x}_{AMAP}(k\|k-1) + (1-a)b(k)h(k)$ $a\epsilon[0,1]$
Subsidiary Equations	$\hat{x}_{AMAP}(k\|k-1) = F\hat{x}_{AMAP}(k-1\|k-1) + Bu(k-1)$ $h(k) = \dfrac{Q\hat{x}_{AMAP}(k\|k-1)}{\|Q\hat{x}_{AMAP}(k\|k-1)\|}$ $b(k) = \dfrac{1}{h'(k)Qh(k)}\{-h'(k)Q\hat{x}_{AMAP}(k\|k-1)$ $\pm[\{h'(k)Q\hat{x}_{AMAP}(k\|k-1)\}^2 - h'(k)Qh(k)$ $\cdot\{\hat{x}_{AMAP}(k\|k-1)Q\hat{x}_{AMAP}(k\|k-1) - z(k)\}]^{\frac{1}{2}}\}$ where the root with minimum magnitude is used.
Initialization	$\hat{x}_{AMAP}(0\|0) = E\{x(0)\}=0$
Control Constraints	Same as the MAP Filter
Text Reference	Chapter 4

Table 5.3.4 The Second-Order Approximate Conditional Mean Filter Algorithm

Main Filter Equation	$\hat{x}_{CM}(k\|k) = \hat{x}_{CM}(k\|k-1) + K(k)\{z(k)$ $- \hat{x}'_{CM}(k\|k-1)Q\hat{x}_{CM}(k\|k-1) - \text{trace } QP(k\|k-1)\}$

Table 5.3.4, continued

Subsidiary Equations	$\hat{x}_{CM}(k\|k-1) = F\hat{x}_{CM}(k-1\|k-1) + Bu(k-1)$ $K(k) = \frac{2}{T(k)} P(k\|k)Q\hat{x}_{CM}(k\|k-1)$ $P(k\|k-1) = FP(k-1\|k-1)F' + GS(k-1)G'$ $P(k\|k) = P(k\|k-1) - 4P(k\|k-1)Q\hat{x}_{CM}(k\|k-1)$ $\cdot\{4\hat{x}_{CM}(k\|k-1)QP(k\|k-1)Q\hat{x}_{CM}(k\|k-1)$ $+ T(k) + \theta(k\|k-1)\}^{-1}\hat{x}'_{CM}(k\|k-1)QP(k\|k-1)$ $\theta(k\|k-1) = \sum_{i,j,k,\ell=1}^{n} \{[P(k\|k-1)]_{ik}[P(k\|k-1)]_{j\ell}$ $+ [P(k\|k-1)]_{i\ell}[P(k\|k-1)]_{jk}\}Q_{ij}Q_{k\ell}$
Initialization	$\hat{x}_{CM}(0\|0) = E\{x(0)\} = 0$ $P(0\|0) = P(0)$
Control Constraints	None. See the discussion in Section 5.4.
Text Reference	See Sage and Melsa [23], p. 466.

The first-order algorithm is obtained by setting to zero the terms trace $QP(k|k-1)$ in the main filter equation and $\theta(k|k-1)$ in the error covariance.

5.4 Structural Comparisons

First of all, notice that the form of the estimate is the same in all cases in that it consists of a correction to the one-step predicted estimate based on the latest measurement. The only real differences lie in the calculation of the correction and in the initialization for the maximum likelihood estimator.

The MAP algorithm is the only one that requires the current state estimate in the calculation of the correction terms; that is, the MAP estimate is a solution to a matrix nonlinear algebraic equation. This unpleasant fact of life is caused by assuming that $p[x(k)|Z^*(k-1)]$, rather than $p[x(k)|Z^*(k)]$, is gaussian.

In all except the AMAP filter, the correction term can be broken down into the product of a filter gain matrix and a measurement residual; i.e., the correction term has the form

$$K[z - \hat{x}'Qx] \tag{5.4.1}$$

where $\hat{x}$ is the appropriate state estimate. For the maximum likelihood estimator, K is a function of the past n controls but is independent of the state estimate. This is very convenient from a computational point of view since all the filter gains can be found as soon as a control sequence is chosen. This is not the case with the MAP and first-order conditional mean filters where K depends on either the one-step predicted estimate or the current estimate of the state.

The correction term structure for the AMAP filter is different in that it looks like a vector with magnitude and direction calculated separately. However, the magnitude and direction both depend on the one-step predicted estimate as before. The computational procedure is very simple requiring only two different matrix-vector products, four dot products, and the solution of a scalar quadratic equation. Most of the simplicity comes about because there is no need to calculate the filtering error covariance matrix if a value is arbitrarily assumed for the scalar a. In the next chapter it will be

shown that the AMAP filter works quite well in all the cases considered if $a=\frac{1}{2}$ is used.

The maximum likelihood filter also does not require calculation of the error covariance, but it has the disadvantages that (1) no state estimate is available until time step k=n, (2) a relatively large amount of storage is required for the various matrix products and past controls, and (3) the estimator structure changes at step n because of the initialization procedure.

It is interesting to note that the error covariance matrix for the MAP and AMAP filters is nearly identical to that for the first-order conditional mean filter, the only difference being that the filtered state estimate is replaced by the one-step predicted state estimate in the latter. Again, this is due to the differing assumptions on the forms of the posterior density functions. This means that the orders of calculation are slightly different. For the MAP and AMAP filters, the current state estimate is computed first, followed by the error covariance (if desired for the AMAP filter). The order is reversed for the first-order conditional mean filter.

Finally, it should be remarked that, as already proved for the maximum likelihood algorithm, identically zero control sequences are not allowed for any of these algorithms. Examination of the tables shows that such a control sequence yields zero estimates for all time for the MAP and first-order conditional mean filters, and undefined estimates for the AMAP filter. No mention of this possibility is made by Sage and Melsa [23] in their formulation of the first-order conditional mean filter.

Table 5.4.1 shows the relative merits of the various algorithms with respect to two characteristics that are important in any practical implementation: memory required and computational speed. These results are based on experience gained in obtaining the data in the next chapter. The algorithms are ranked on a scale of one to four with one indicating best, four worst.

Table 5.4.1 Relative Merits of the Algorithms

	Memory Required	Speed
Maximum Likelihood	4	1
MAP	3	4
AMAP	1	2
1st Order CM	2	3

Chapter 6

NUMERICAL RESULTS

6.1 Introduction

In this chapter the algorithms developed in Chapters 3 and 4 are tested, both to ascertain the validity of the statements made herein, and to be able to make comparisons between algorithms. The next section contains the main results of the chapter; a randomly-generated model is used to test the relative merits of the algorithms and some different control sequences are investigated. In addition, some cases with marginal (small) controls are examined and the effects of larger noise are observed. Section 6.3 considers in less detail a model similar to that in Section 6.2 except that the kernel of the quadratic form is the unit matrix.

It is not the purpose of this chapter to exhaustively test the algorithms; that is an enormous task. The idea is to give some insight into the relative merits of the filters and whether they are useful for several different problems. This chapter should not be viewed as a substitute for performing a simulation on the problem at hand. It is incumbent on the user of these ideas to protect himself.

6.2 General Third-Order Model Investigation

In this section a randomly-generated third-order problem is examined with regard to each of the algorithms of Chapter 5, various control sequences, and various noise magnitudes.

6.2.1 Description of the First Simulation Model

A third-order system model was chosen with the following values:

$$F = \begin{bmatrix} .502 & .406 & .543 \\ -.365 & .735 & -.290 \\ .344 & -.064 & .549 \end{bmatrix}$$

$$B = [1.093 \quad 2.720 \quad -1.113]'$$

$G = I_3$, the third-order unit matrix

$P(0)$ was not used to generate $x(0)$.

$S(k) = S = .01I_3$

$T(k) = T = .01$

$$Q = \begin{bmatrix} 1.317 & 1.063 & 1.025 \\ 1.063 & .461 & 4.07 \\ 1.025 & .407 & 1.258 \end{bmatrix}$$

It is easily, but tediously, verified that Q is symmetric but not positive definite. The numbers in the F, B, and Q matrices were obtained with a random number generator. The values for F came from a gaussian, zero-mean distribution with a variance of one, those for B from a similar distribution with a variance of four. The Q-matrix values came from a uniform distribution on (0,1.5).

Each of the averages plotted below is calculated from the results of 100 runs with different noise sequences but fixed $x(0)$. Eight different values for $x(0)$ are chosen, one in each octant of the three-dimensional system. This is done in order to observe the effect of different starting points on the algorithms.

Three different types of control sequences are also investigated as to their effect on filtering error. The first type of control sequence is $u(k)$ equal to a constant, in this case $u(k)=1$. The second type is a dither about a constant, namely $u(k)=1.1+(-1)^k(.2)$. This is

a control commonly used in systems with a dead zone, or hysteresis, for stabilization. The third type of control is oscillatory, $u(k)=(-1)^{k+1}$.

These three control sequences can be shown to result in state trajectories (see Section 6.2.2) such that equation 4.6.1 is stable, that is, the matrix $B_k^{-1}F$ has all its eigenvalues inside the unit circle. In fact, for this problem, it is not possible to choose a control such that 4.6.1 is unstable. The reason for this may be seen from the following arguments. First, F has no eigenvalues outside the unit circle, as is apparent from the state trajectories in Section 6.2.2. Remembering that B_k is given by

$$B_k = I + \frac{4}{T(k)} P(k|k-1)Q\hat{x}(k|k)\hat{x}'(k|k)Q \tag{4.5.7}$$

it is clear that since the second term on the right side of 4.5.7 is positive semidefinite, B_k has no eigenvalues inside the unit circle (there may be some on the boundary) and therefore, B_k^{-1} has none outside. Hence, $B_k^{-1}F$ has all its eigenvalues inside the unit circle and 4.6.1 is stable, no matter what control sequence is chosen.

Therefore, the lower bound on the magnitude of the control is not given by the arguments in Section 4.6, but by those in Section 2.6 that the control should be somewhat larger than the state uncertainty. This is the case for all three control sequences as can be seen by referring to Section 6.2.9. The effects of smaller controls will be considered later.

6.2.2 Typical State Trajectories

Figures 6.2.2.1, 2, and 3 show typical state trajectories for

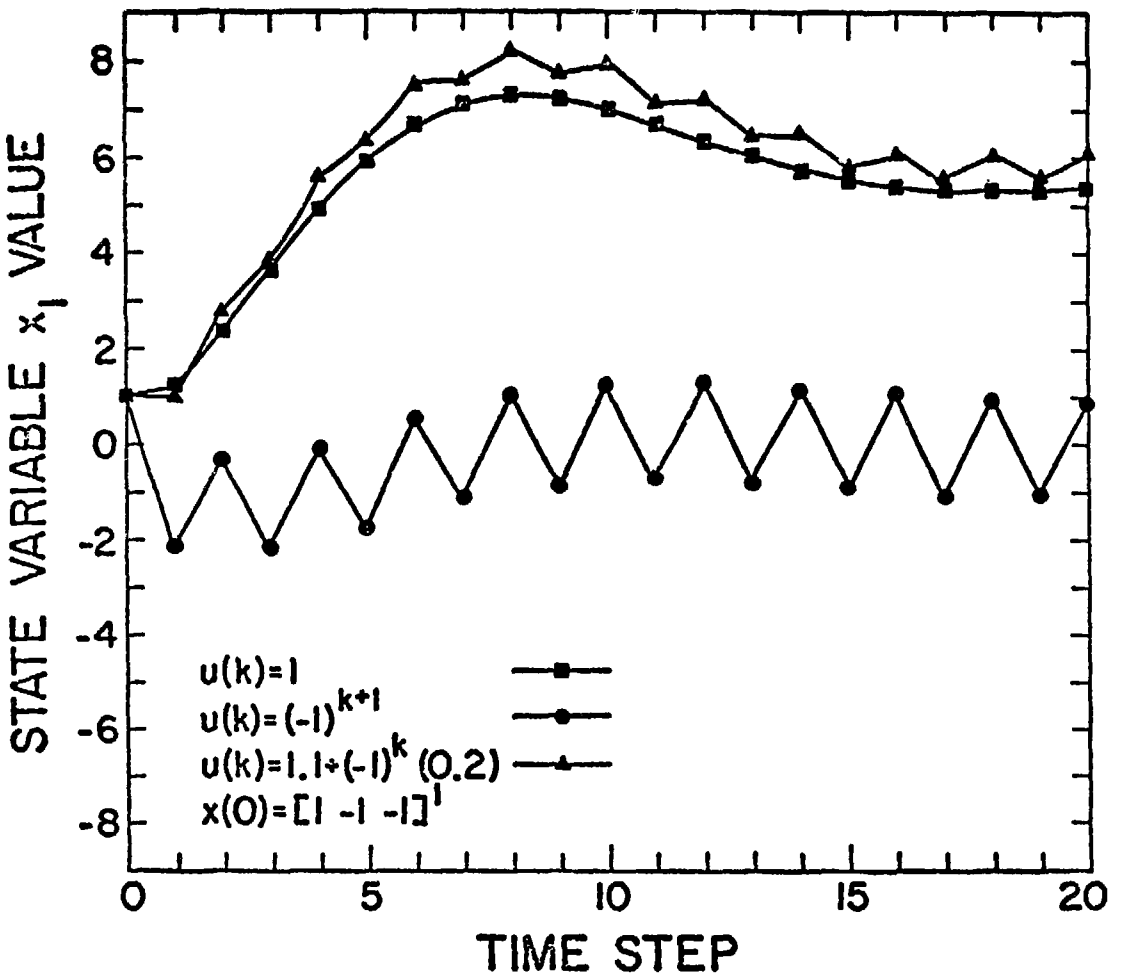

Figure 6.2.2.1 Typical x_1 Trajectories for the Three Control Sequences

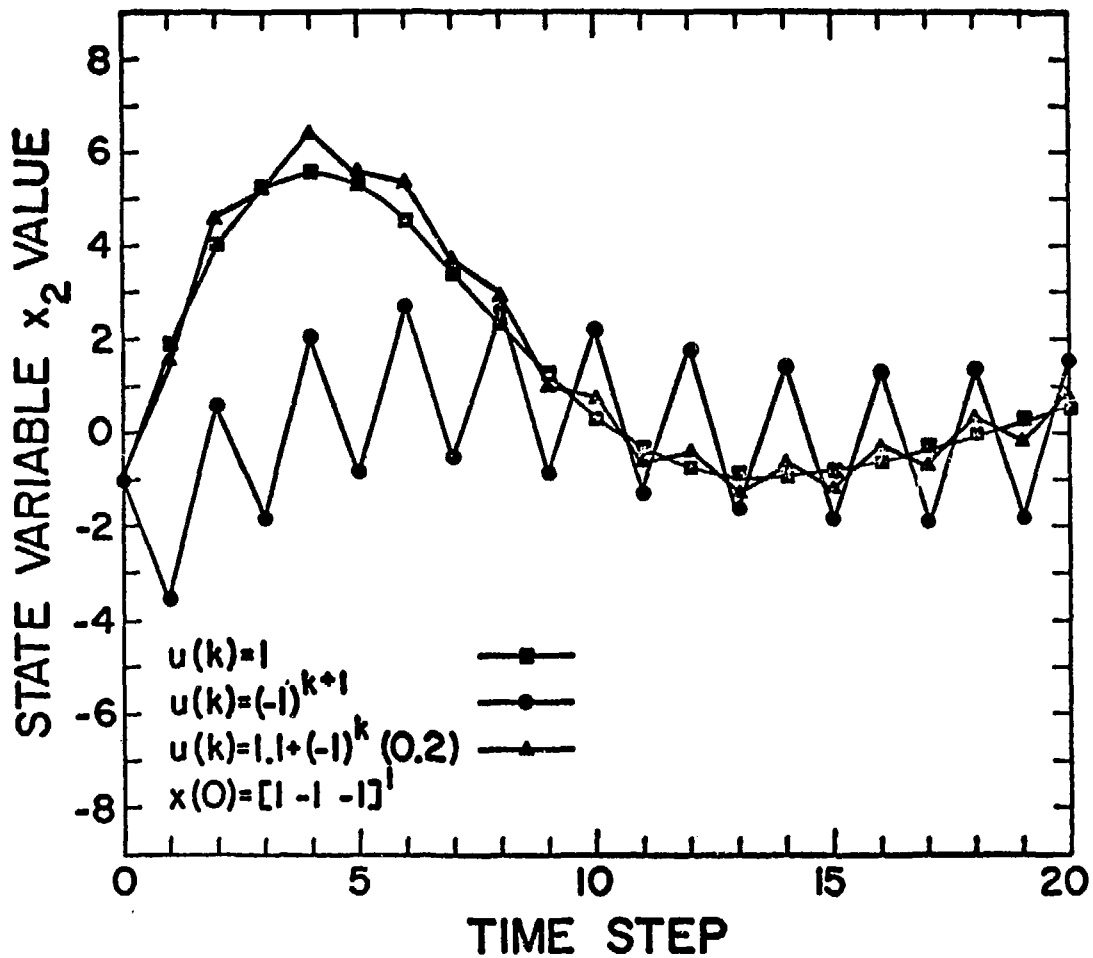

Figure 6.2.2.2 Typical x_2 Trajectories for the Three Control Sequences

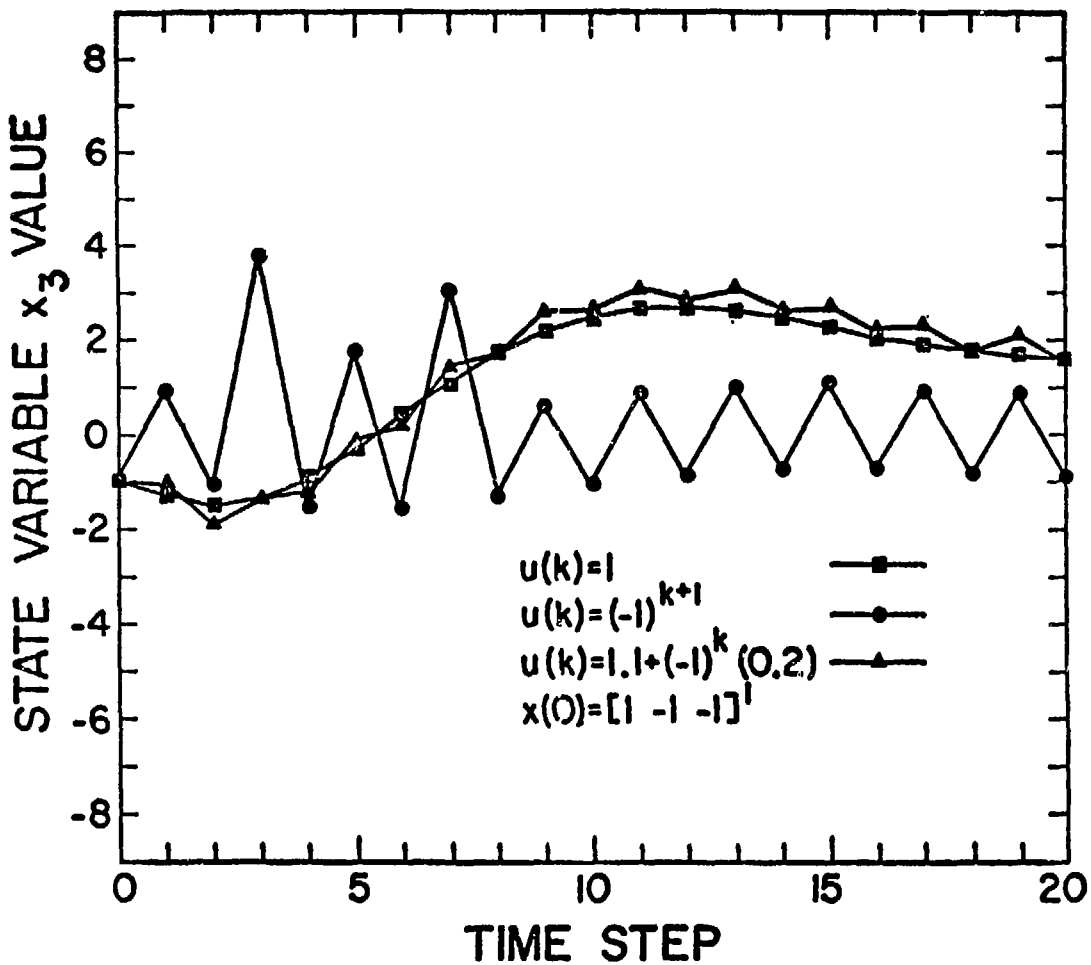

Figure 6.2.2.3 Typical x_3 Trajectories for the Three Control Sequences

each of the three control sequences starting from $x(0)=[1\ -1\ -1]'$. Such plots are also available for other starting points. Since there is no essential difference between any of them, only the ones for this $x(0)$ are given. Note that, for this system, the oscillatory control generally results in the state being closer to the origin. This may be important in cases where there are state constraints.

6.2.3 The Maximum A Posteriori Algorithm

This algorithm is obtained by attempting to solve the estimator equation 4.3.3. This was done using a standard Fortran subroutine to find the maximum of the density function in 4.2.35.

Graphs of the results are shown in Figures 6.2.3.1-6. Again, only the values for one starting point, $x(0)=[1\ -1\ -1]'$ are shown, since there is no fundamental difference between the values for any of the starting points. This will be the case throughout this numerical study.

It is evident that there is not much difference in filtering error between using $u(k)=1$ and $u(k)=1.1\pm.2$. Figures 6.2.3.1 and 3 indicate that there is some residual error even after twenty time steps and that this error is constant in sign; i.e., the estimate is biased.

A different problem occurs for the oscillatory control. At time step 9, the subroutine for the solution of the estimation equation is not able to find the solution within the number of iterations allowed. Consequently, there is a discontinuity in the estimation error at this point. Note that this has the effect of nearly erasing the data available from steps previous to 9. That is, the filter essentially has to start

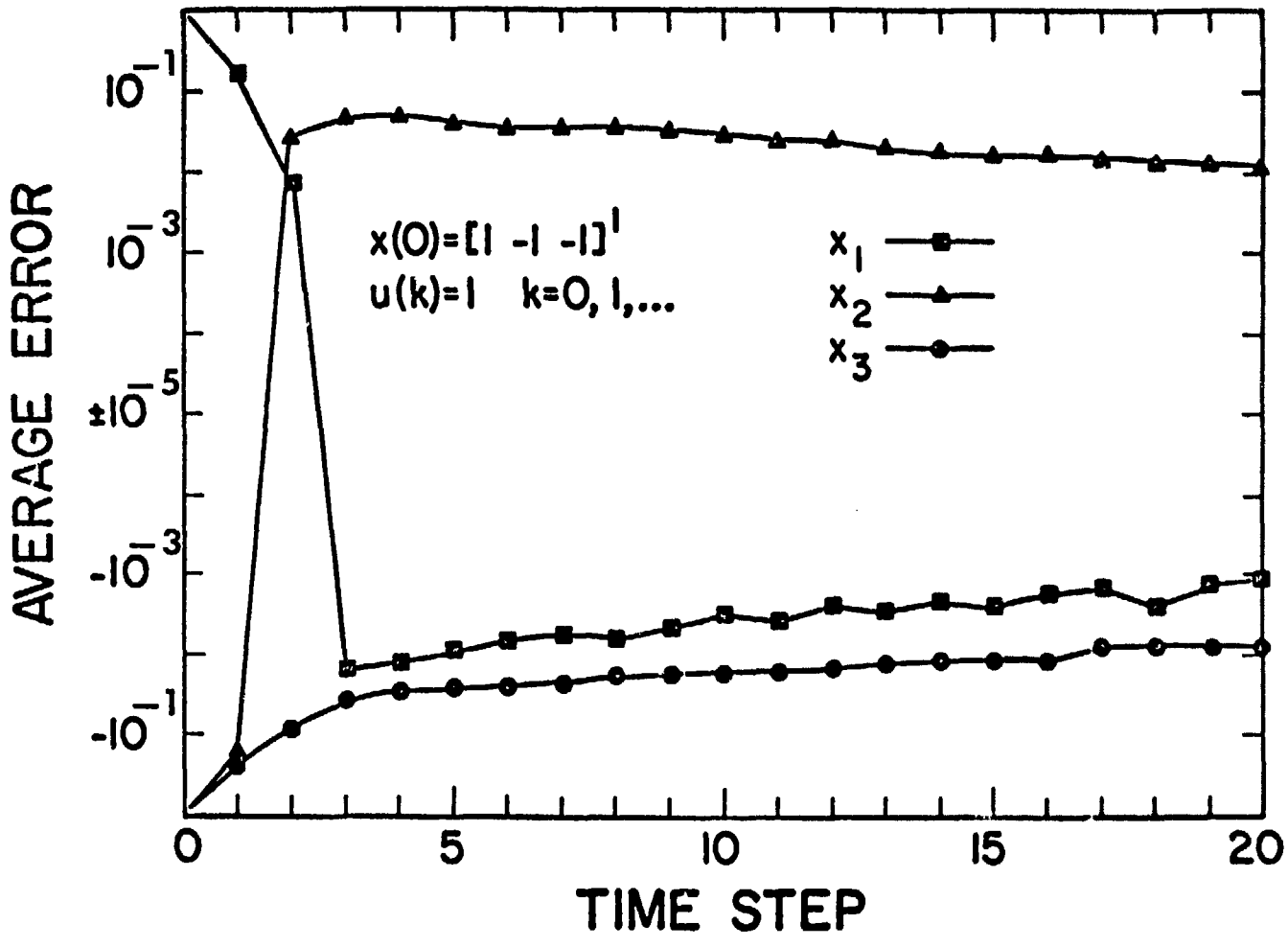

Figure 6.2.3.1 Maximum A Posteriori Filter Average Error, u(k) = 1

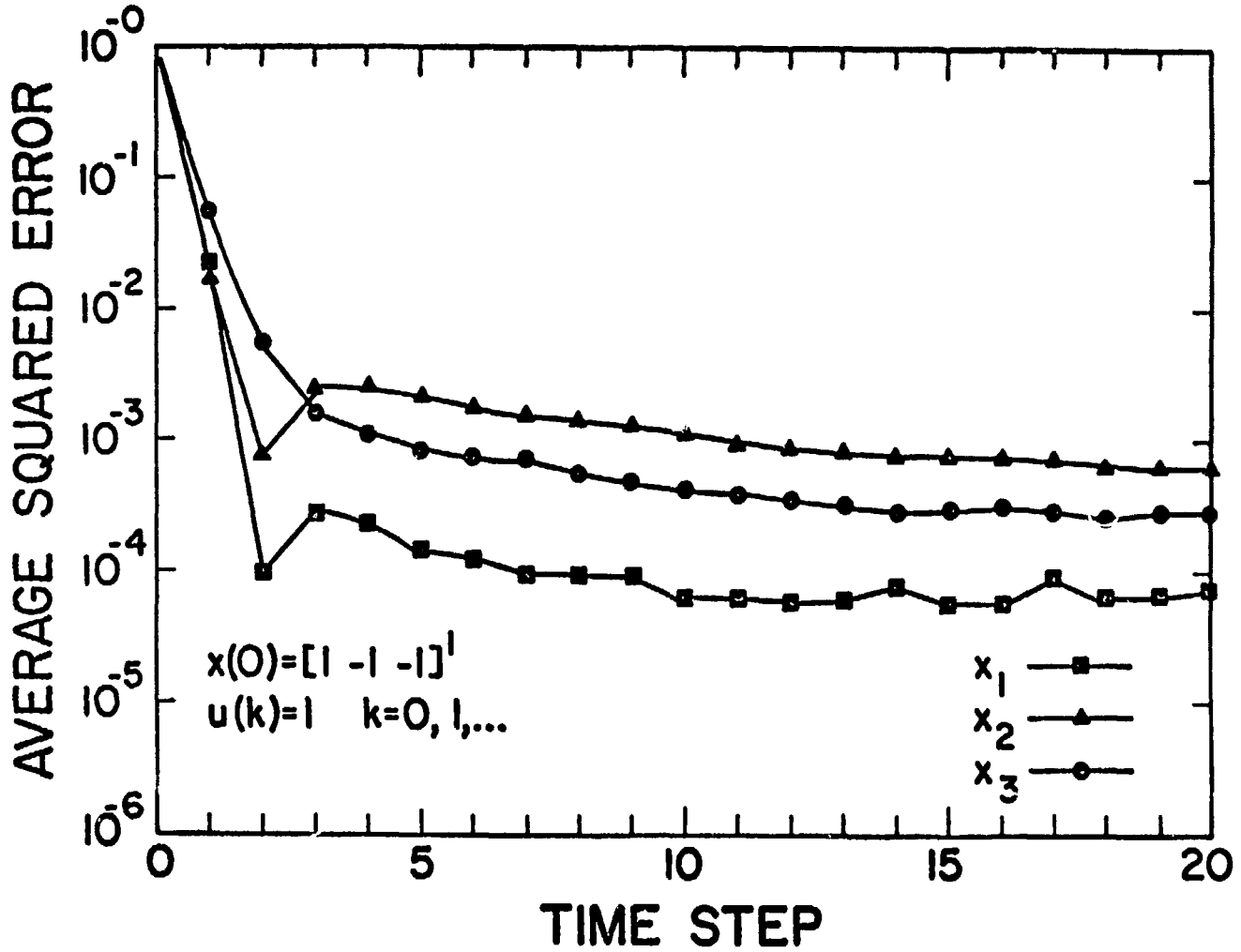

Figure 6.2.3.2 Maximum A Posteriori Filter Average Squared Error, u(k) = 1

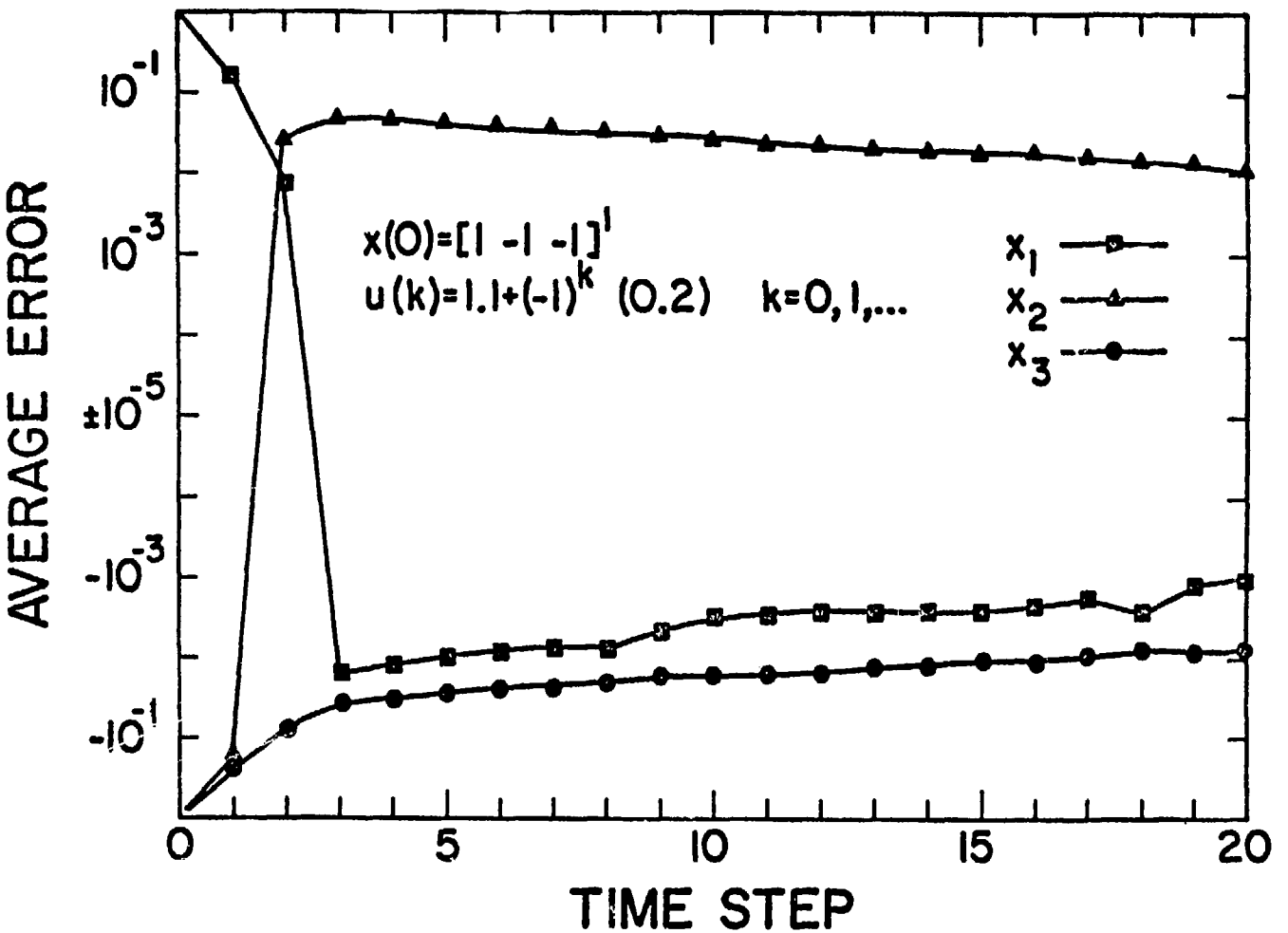

Figure 6.2.3.3 Maximum A Posteriori Filter Average Error, $u(k) = 1.1 + (-1)^k(.2)$

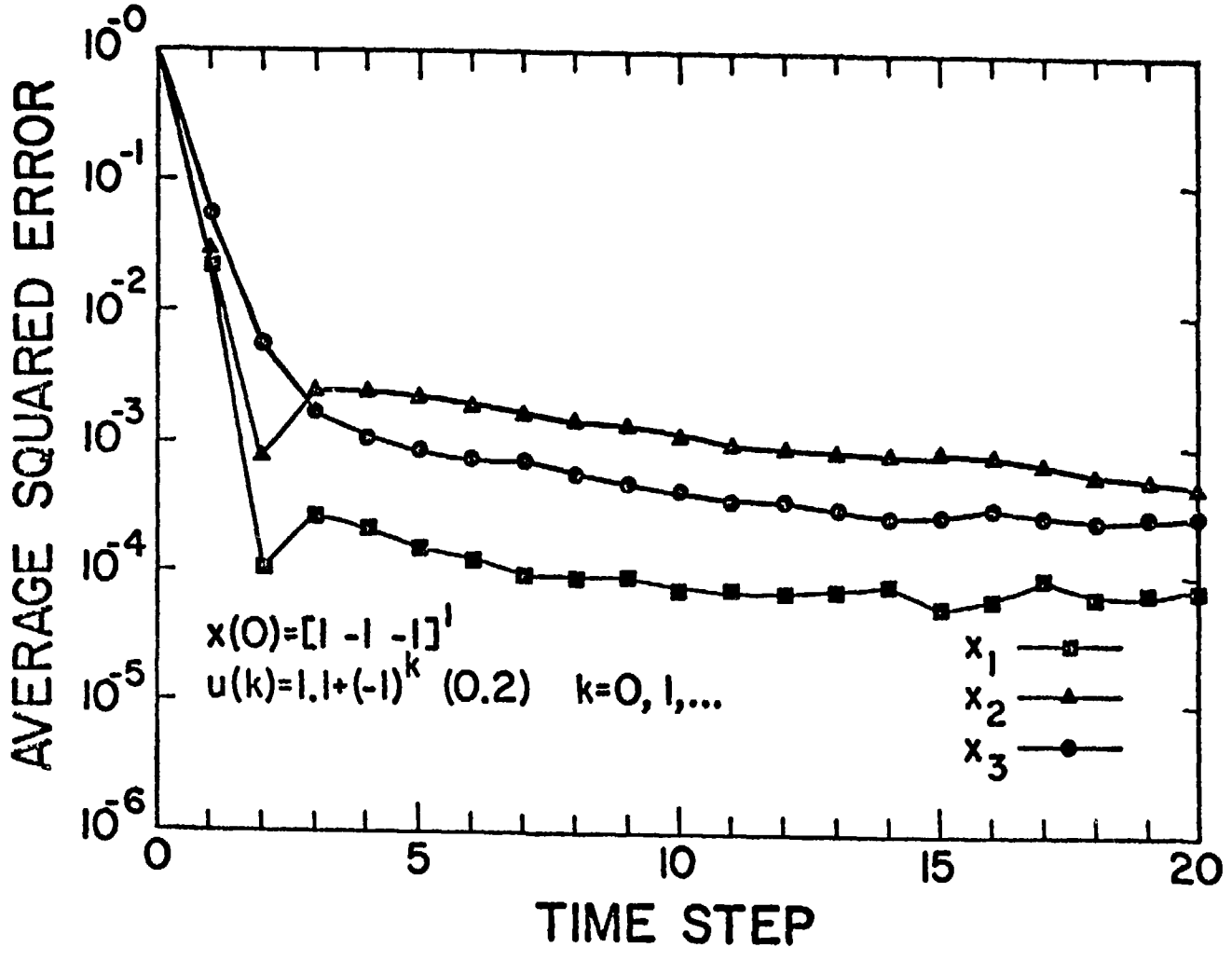

Figure 6.2.3.4 Maximum A Posteriori Filter Average Squared Error, $u(k) = 1.1 + (-1)^k(.2)$

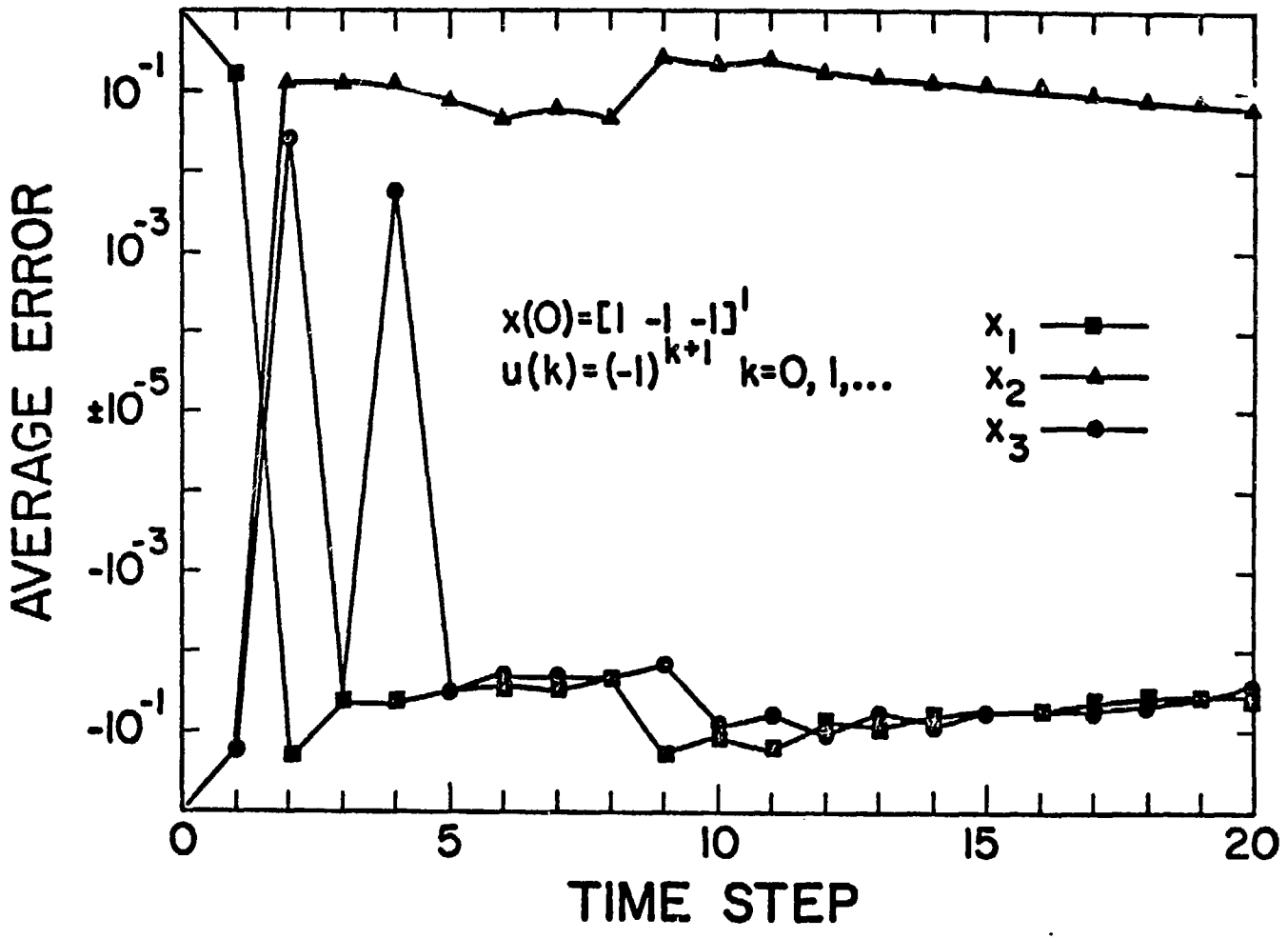

Figure 6.2.3.5 Maximum A Posteriori Filter Average Error, $u(k) = (-1)^{k+1}$

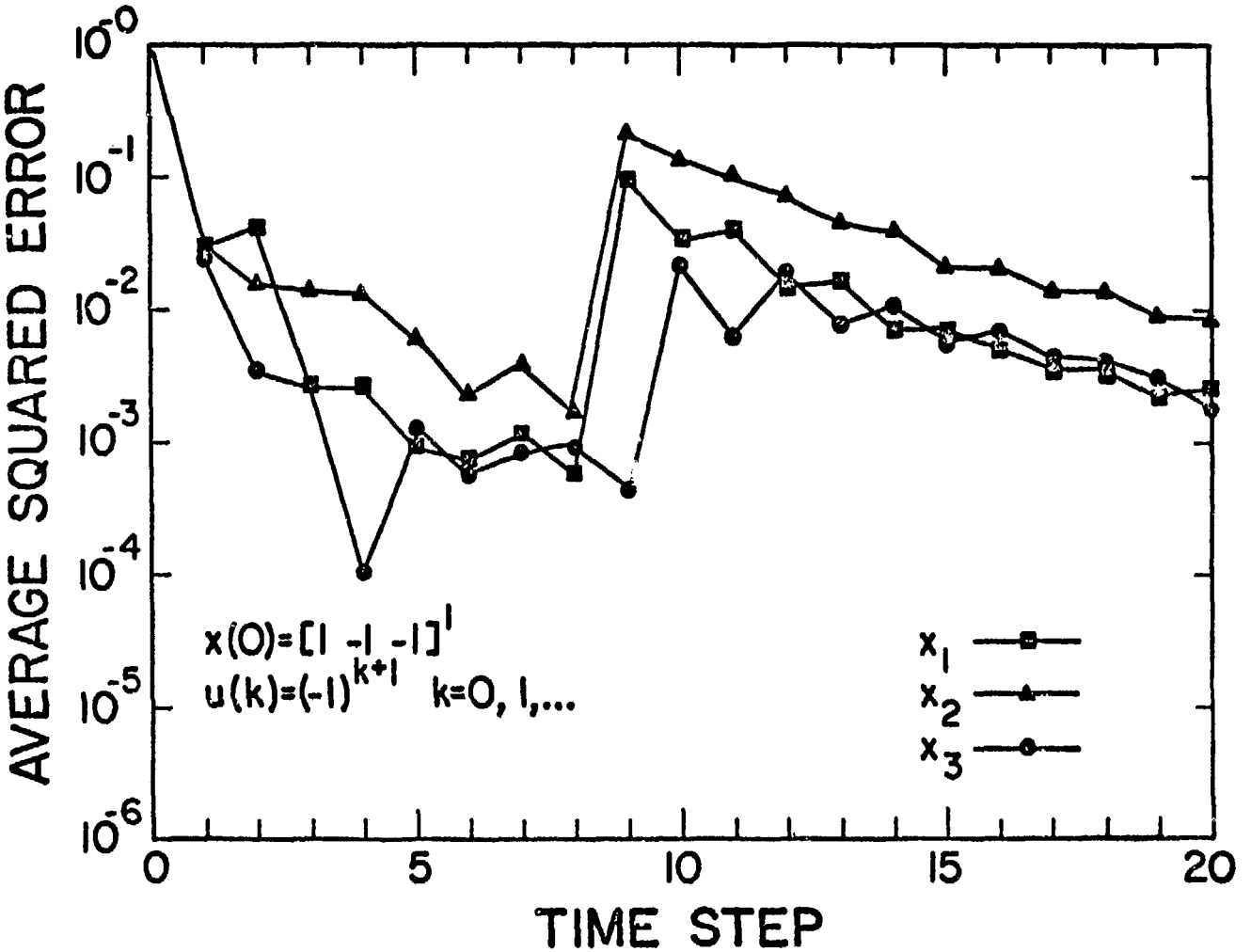

Figure 6.2.3.6 Maximum A Posteriori Filter Average Squared Error, $u(k) = (-1)^{k+1}$

over. It should also be remarked that this algorithm is extremely expensive in terms of computer time.

6.2.4. The Approximate Maximum A Posteriori Algorithm

In this section the results for the most useful of all the algorithms are discussed. Figures 6.2.4.1-6 are plots of the average error and average squared error for the three control sequences.

Again, there appears to be little difference between using $u(k)=1$ and $u(k)=1.1\pm.2$. However, this algorithm seems to yield a much more nearly unbiased estimate as evidenced by the bipolar behavior in Figures 6.2.4.1 and 3. The oscillatory control sequence still gives a biased estimate (Figure 6.2.4.5).

This algorithm has slightly smaller ultimate (at time step 20) error, but the MAP algorithm has a little better transient response as measured by the average squared error in Figures 6.2.4.2, 4, and 6.[1]

6.2.5. First-Order Conditional Mean Filter

In order to make comparisons, a first-order approximate conditional mean filter was simulated using the same starting points and control sequences. Whereas the first two control sequences resulted in more-or-less unbiased estimates for the AMAP filter, this is not true for this algorithm. The more nearly unbiased behavior occurs for the oscillatory control. Examination of the best average squared error performances for the two algorithms shows that they are nearly

---

[1]Note: The scalar $\hat{a}$ in equation 4.4.12 was arbitrarily chosen to be .5 for this simulation.

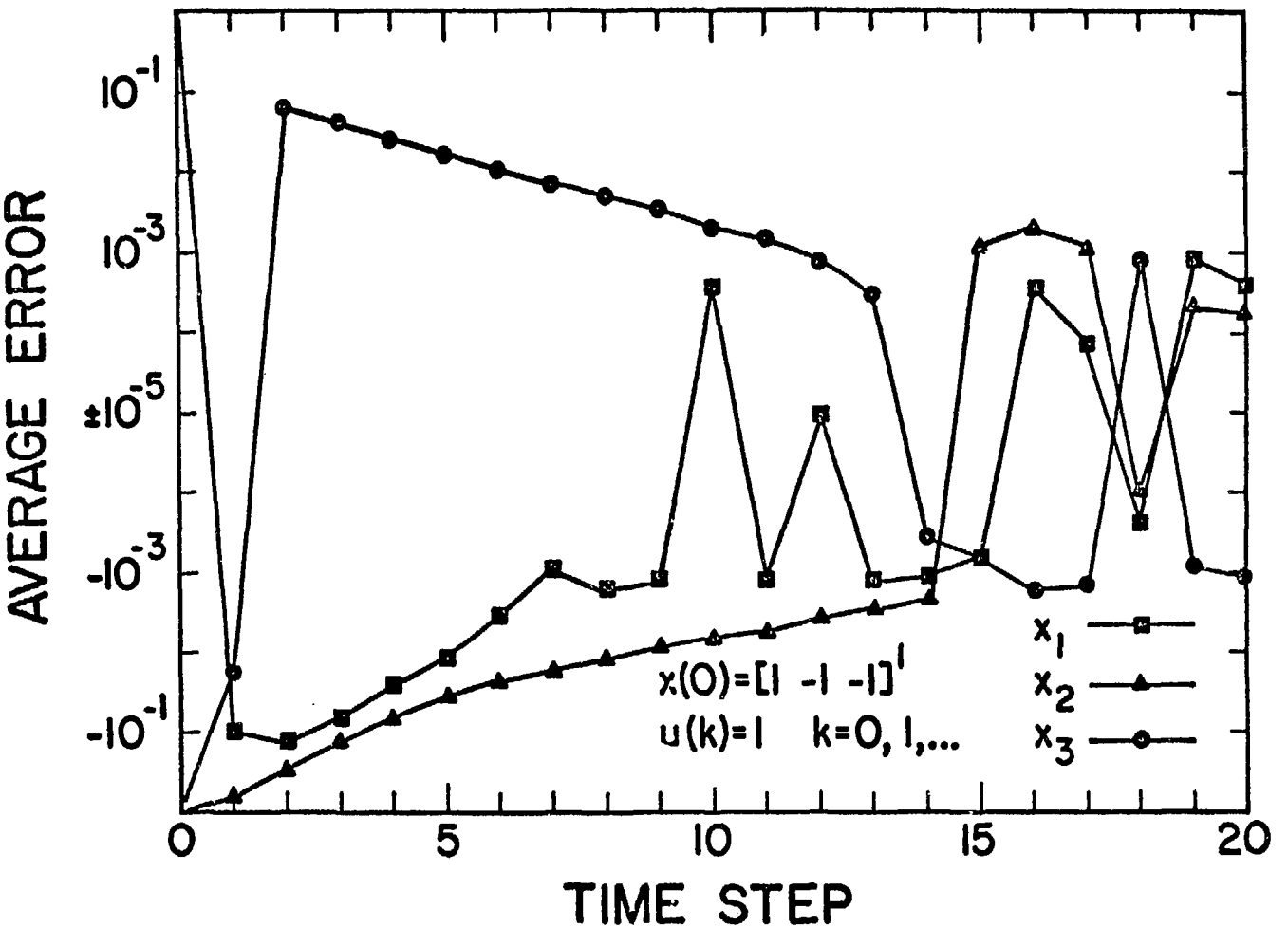

Figure 6.2.4.1 Approximate Maximum A Posteriori Filter Average Error, u(k) = 1

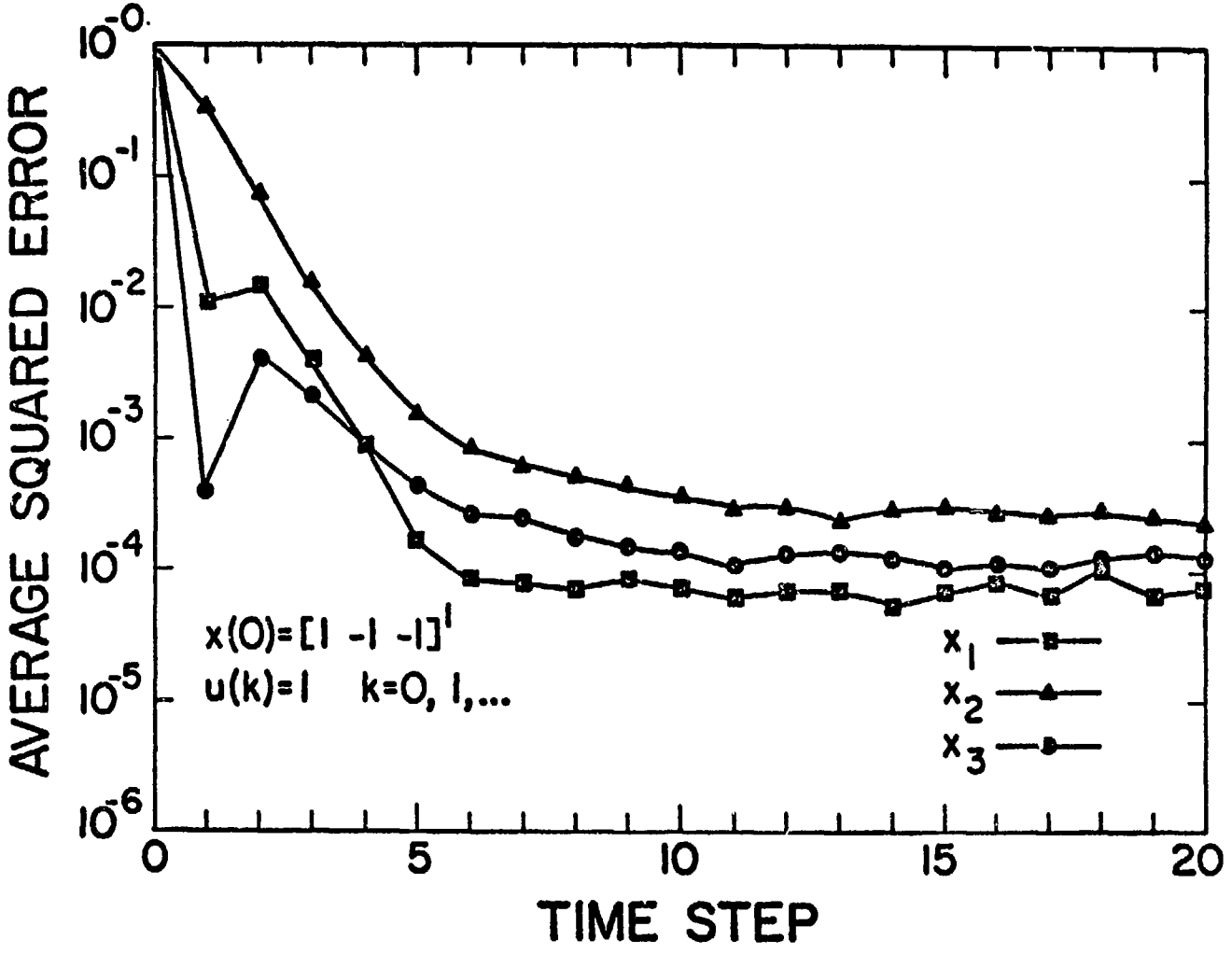

Figure 6.2.4.2 Approximate Maximum A Posteriori Filter Average Squared Error, u(k) = 1

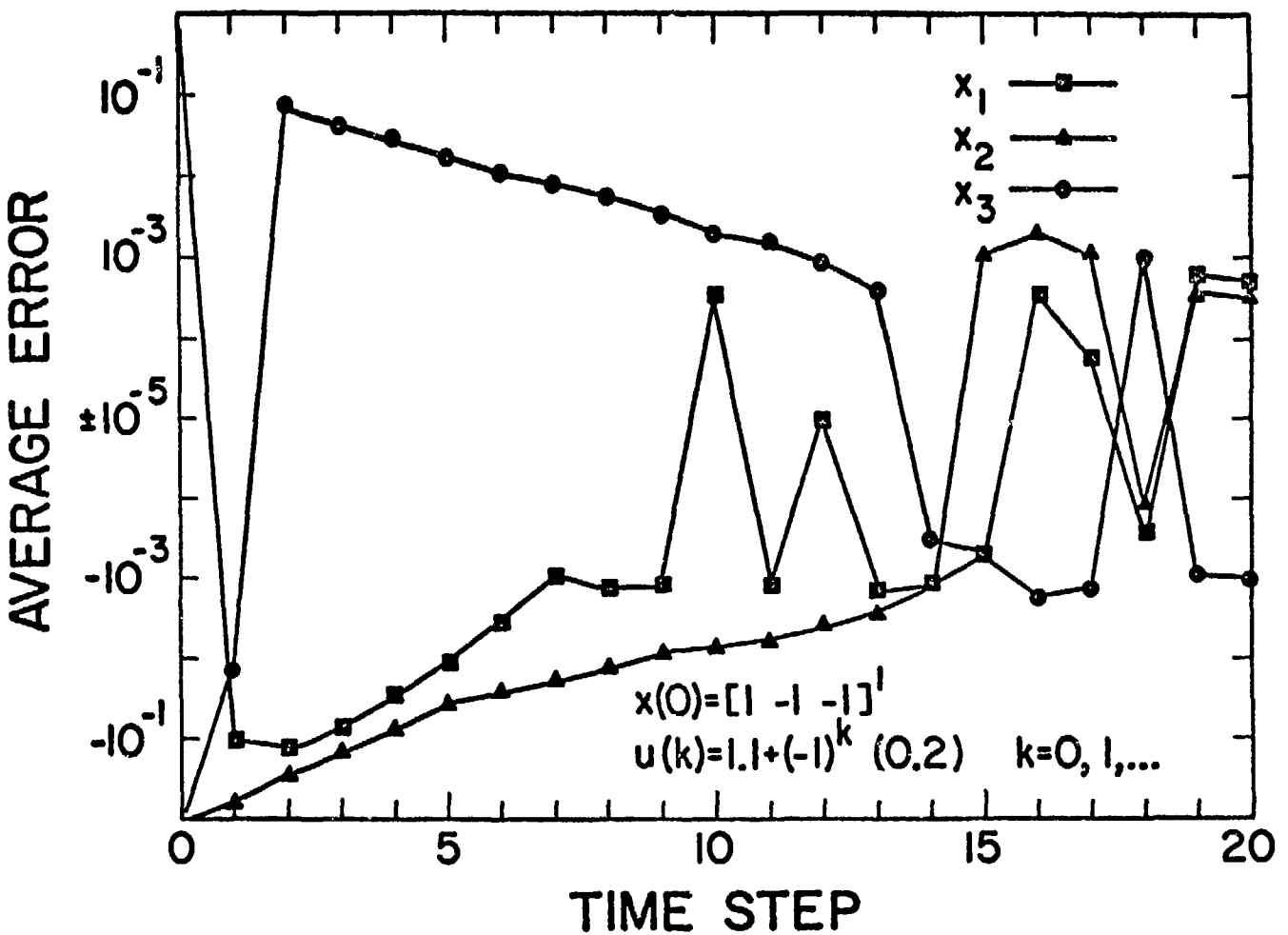

Figure 6.2.4.3 Approximate Maximum A Posteriori Filter Average Error, $u(k)=1.1+(-1)^k(.2)$

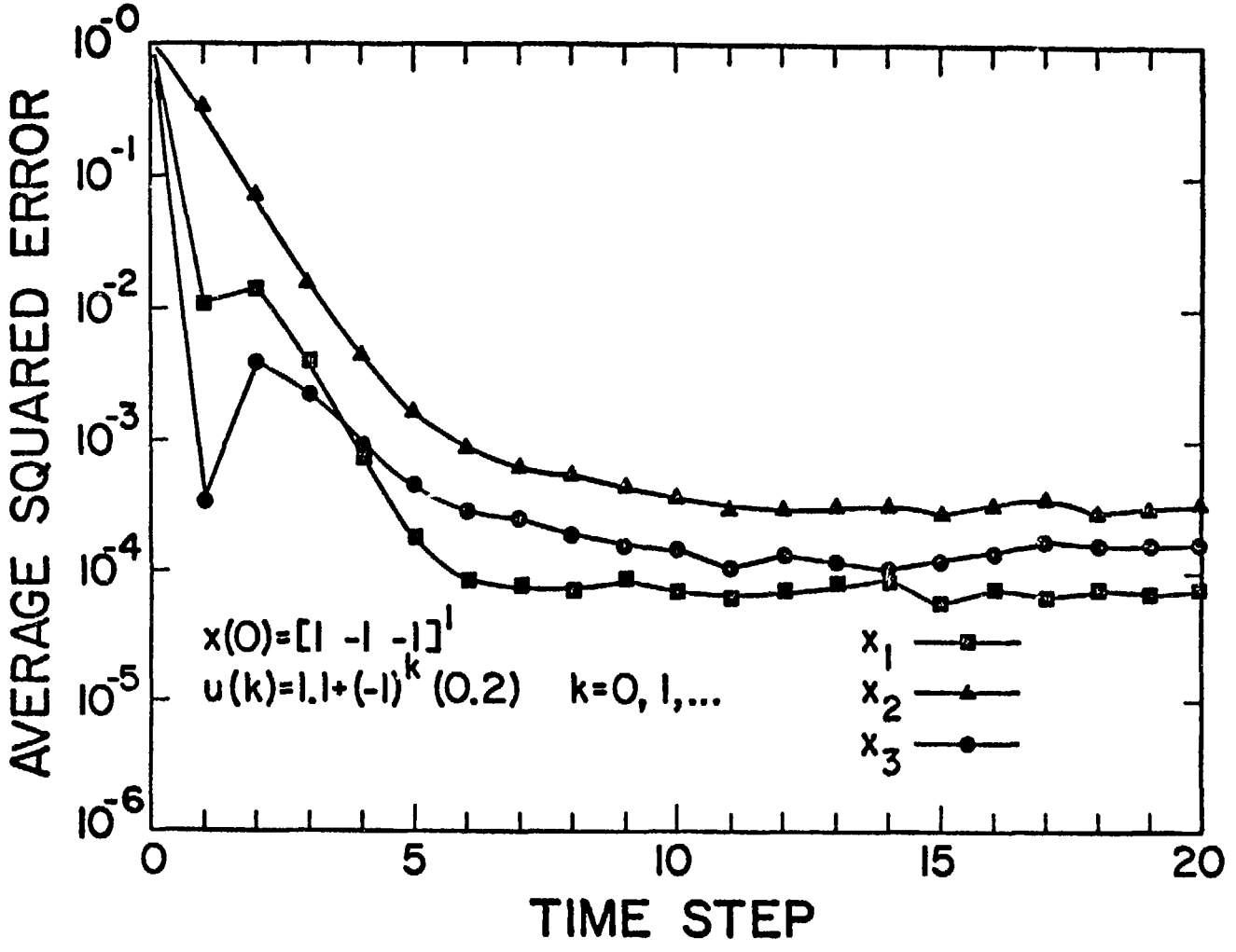

Figure 6.2.4.4 Approximate Maximum A Posteriori Filter Average Squared Error, $u(k)=1.1+(-1)^k(.2)$

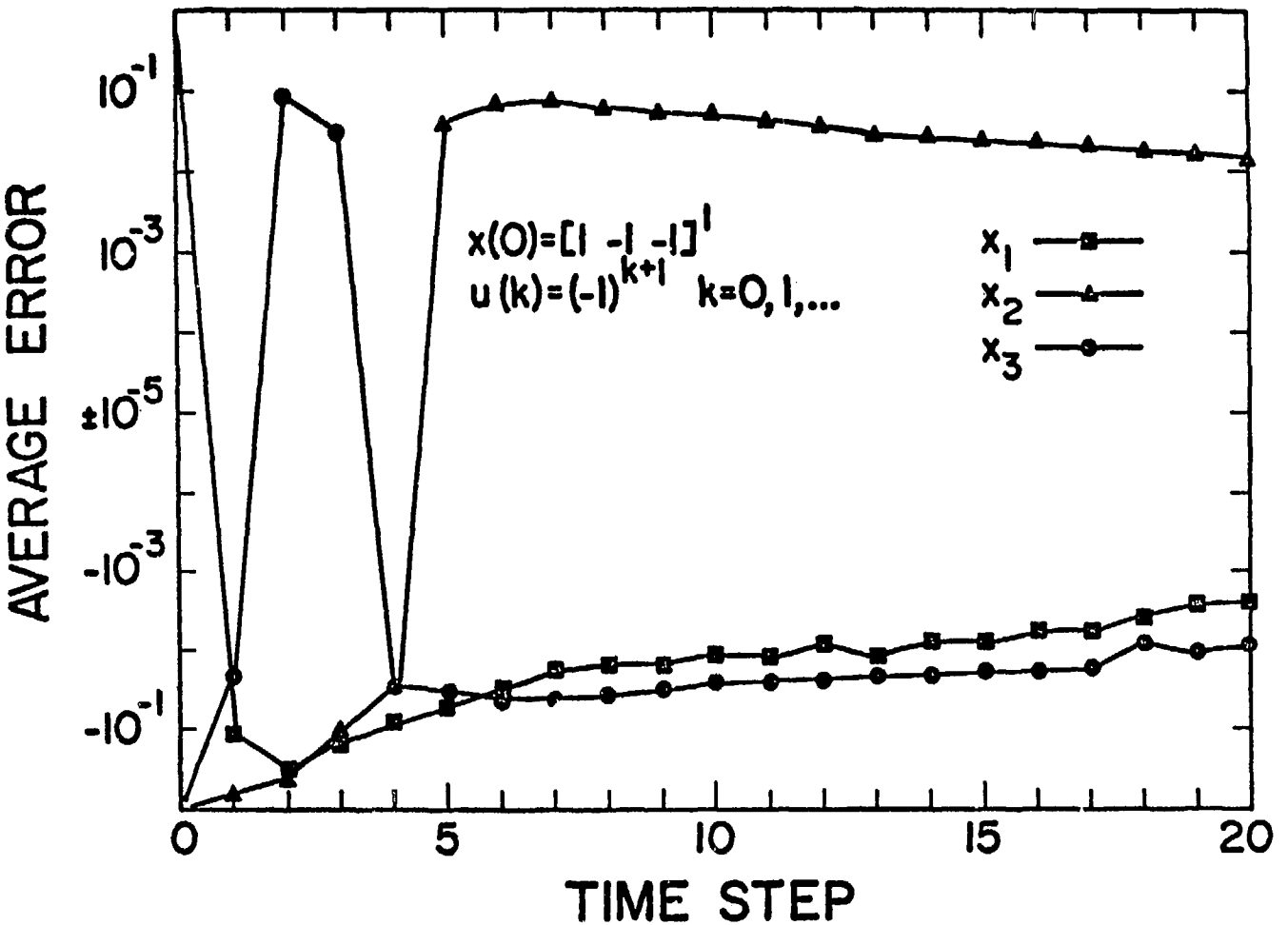

Figure 6.2.4.5 Approximate Maximum A Posteriori Filter Average Error, $u(k)=(-1)^{k+1}$

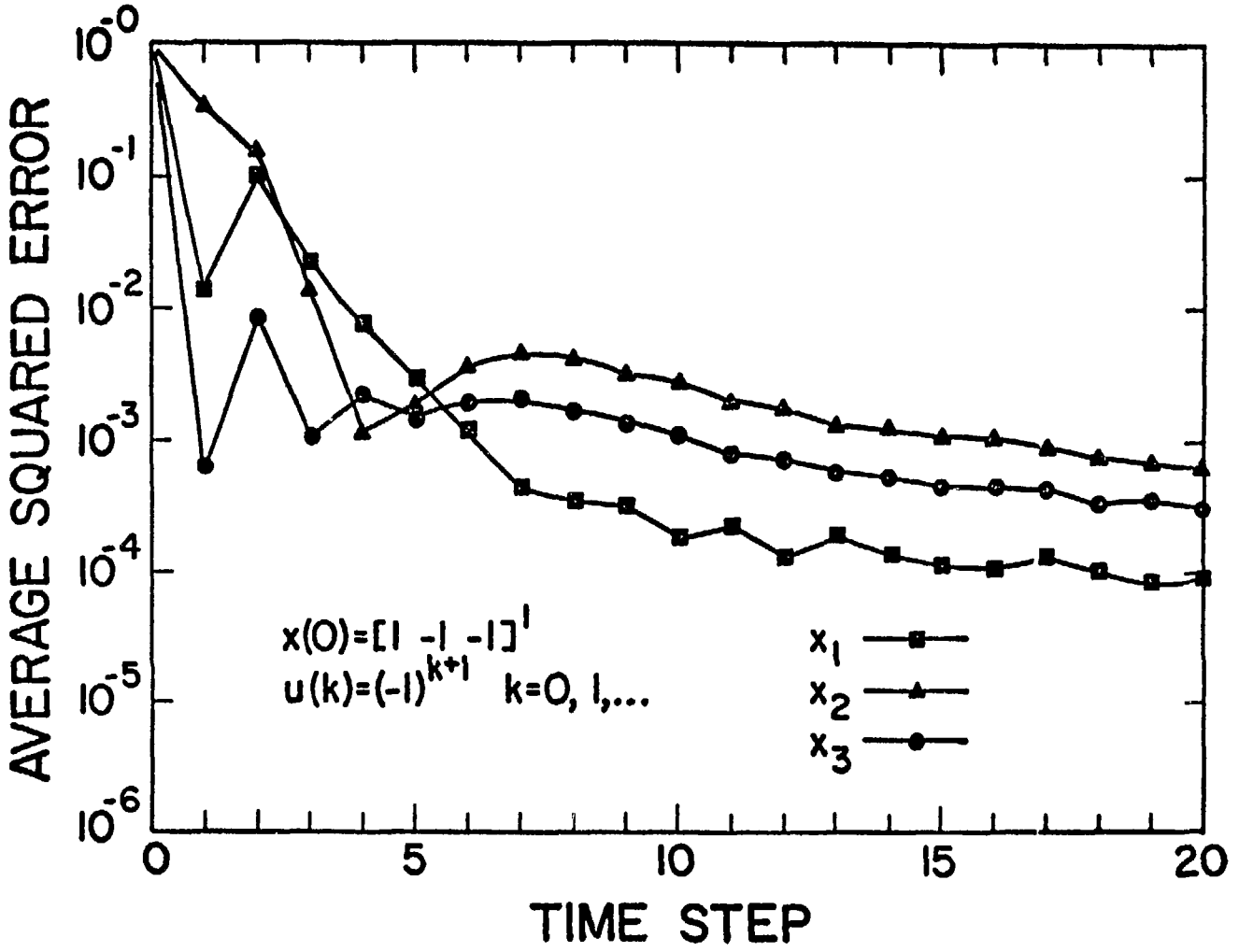

Figure 6.2.4.6 Approximate Maximum A Posteriori Filter Average Squared Error, $u(k)=(-1)^{k+1}$

identical. Figures 6.2.5.1-6 contain the results in graphical form. Further comparisons with larger noise will be made in another section.

6.2.6. Remarks on the Second-Order Conditional Mean Filter

The second-order filter gives grossly inaccurate state estimates as can be seen by comparing the numbers for the true value of x_1 and the estimated value of x_1, the first component of the state, in Table 6.2.6.1.

Table 6.2.6.1 Comparison of x_1 and the Second-Order Conditional Mean Estimate of x_1

Time Step	True Value of x_1	Estimated Value of x_1
1	3.14502	1584.11
2	4.29209	-9147.05
3	5.16390	13230.9
4	5.85287	1835.18
5	6.36879	9785.34

The reason for this misbehavior can be seen if the graphs of Figures 6.2.6.1-3 are examined. These are plots of the calculated (using equation 4.5.16) error variances for the three components of the state, and of the measured average squared error. If the filter equations truly describe the situation, then these two graphs should be nearly coincident in every case. Note, however, that the calculated value is at worst some two orders of magnitude greater than the measured value for each component. It is apparent from Table 5.3.4 that this can have a large effect on the state estimate since the second order

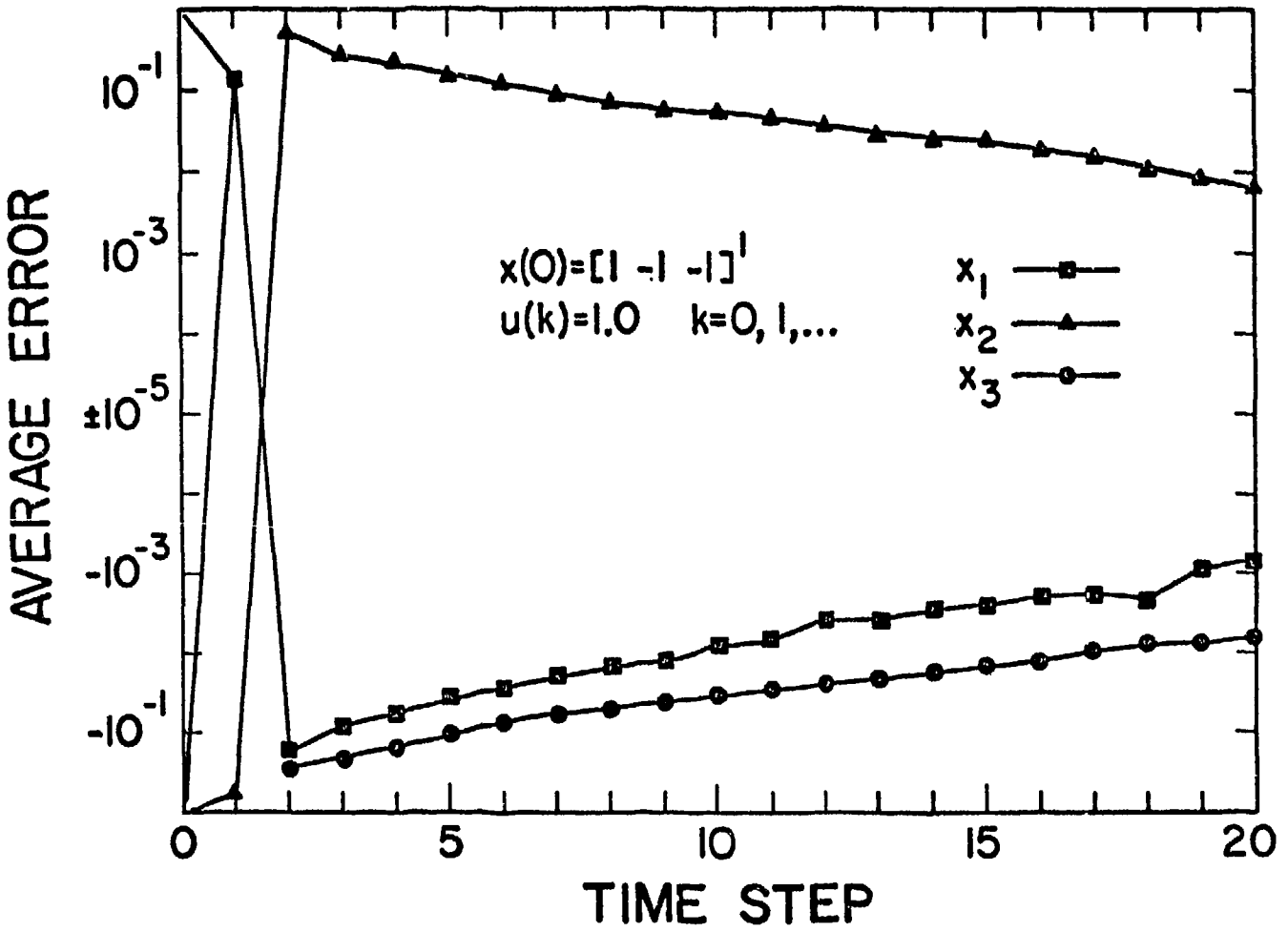

Figure 6.2.5.1 First-Order Conditional Mean Filter Average Error, u(k) = 1

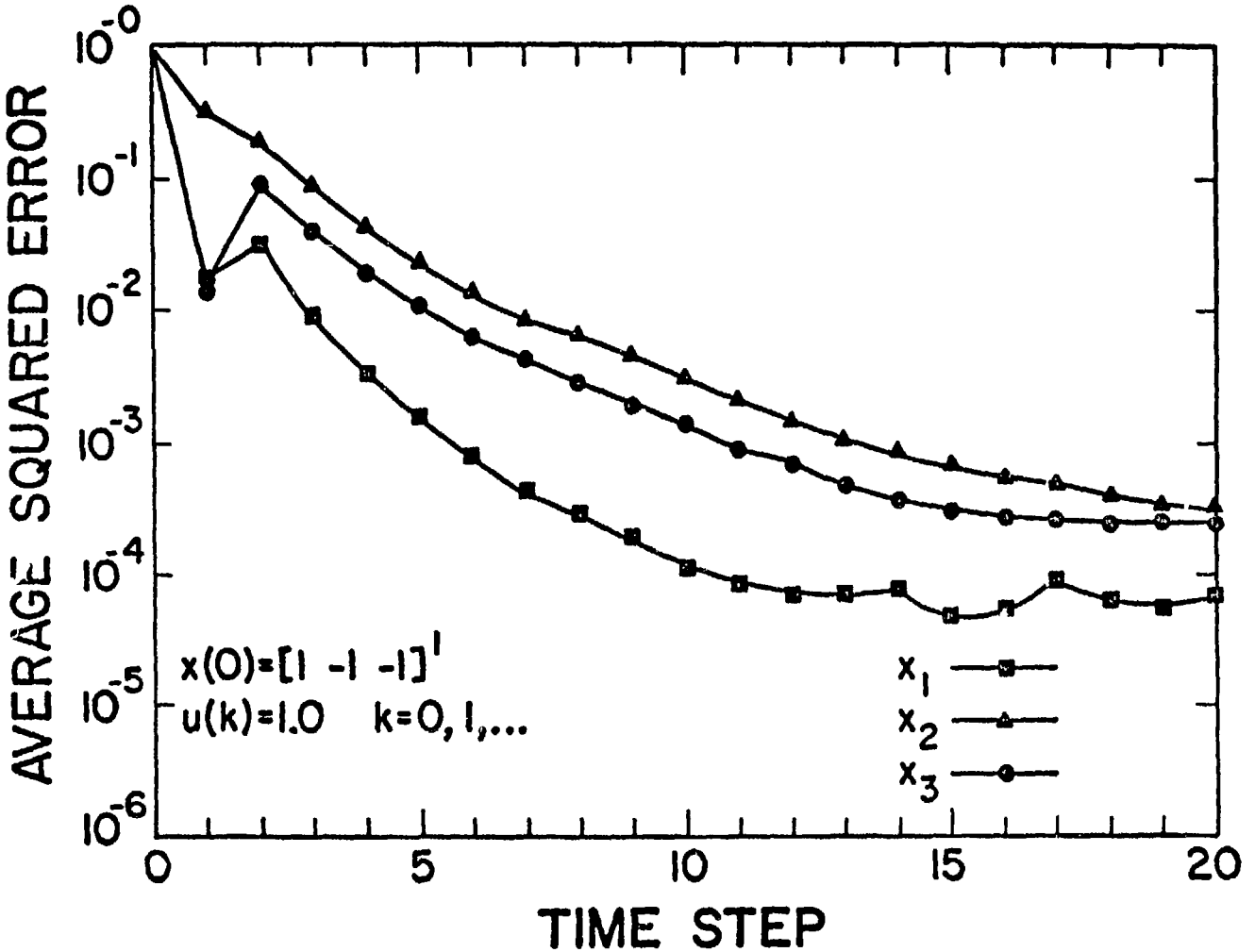

Figure 6.2.5.2 First-Order Conditional Mean Filter Average Squared Error, u(k) = 1

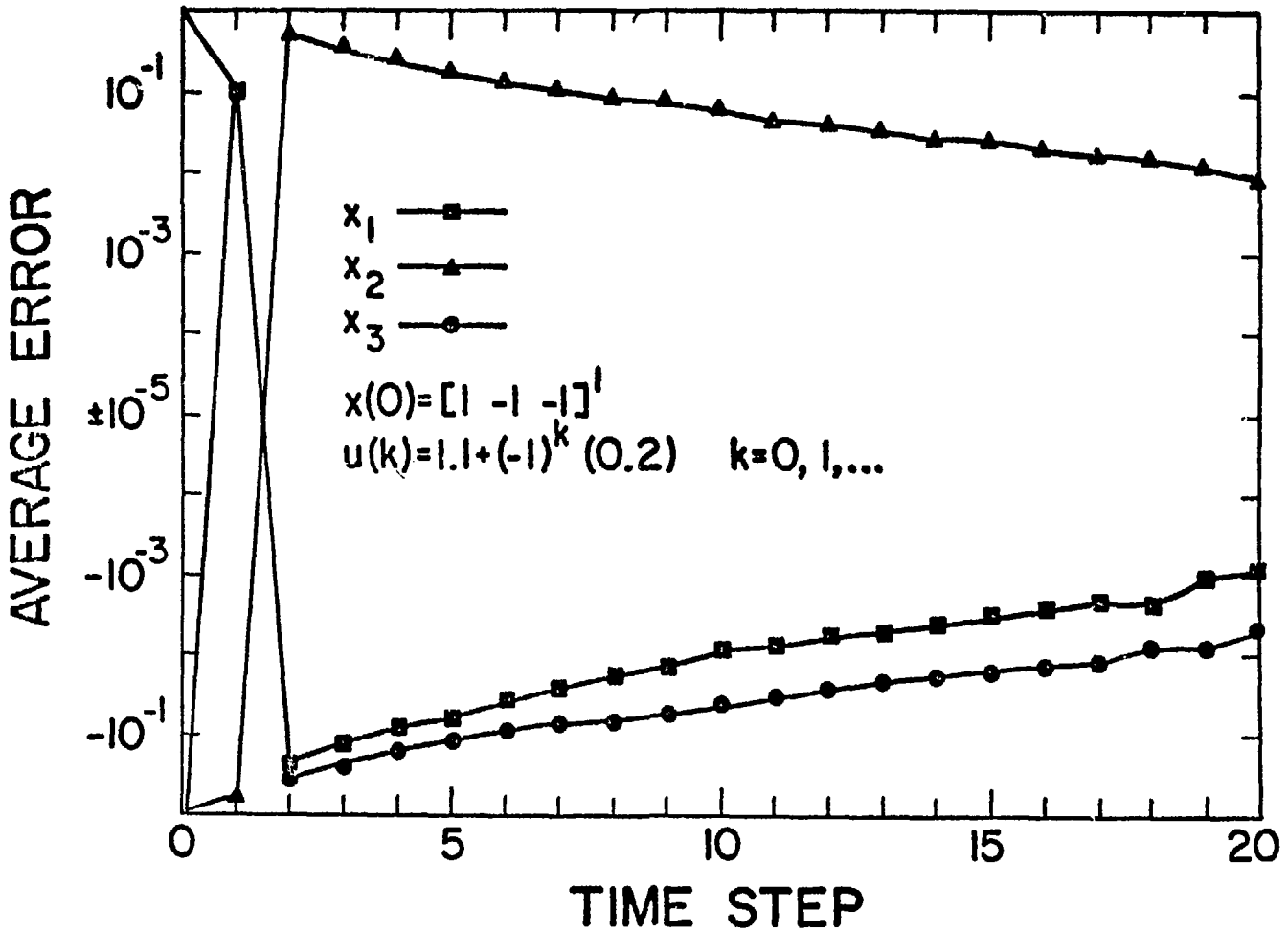

Figure 6.2.5.3 First-Order Conditional Mean Filter Average Error, $u(k) = 1.1 + (-1)^k(.2)$

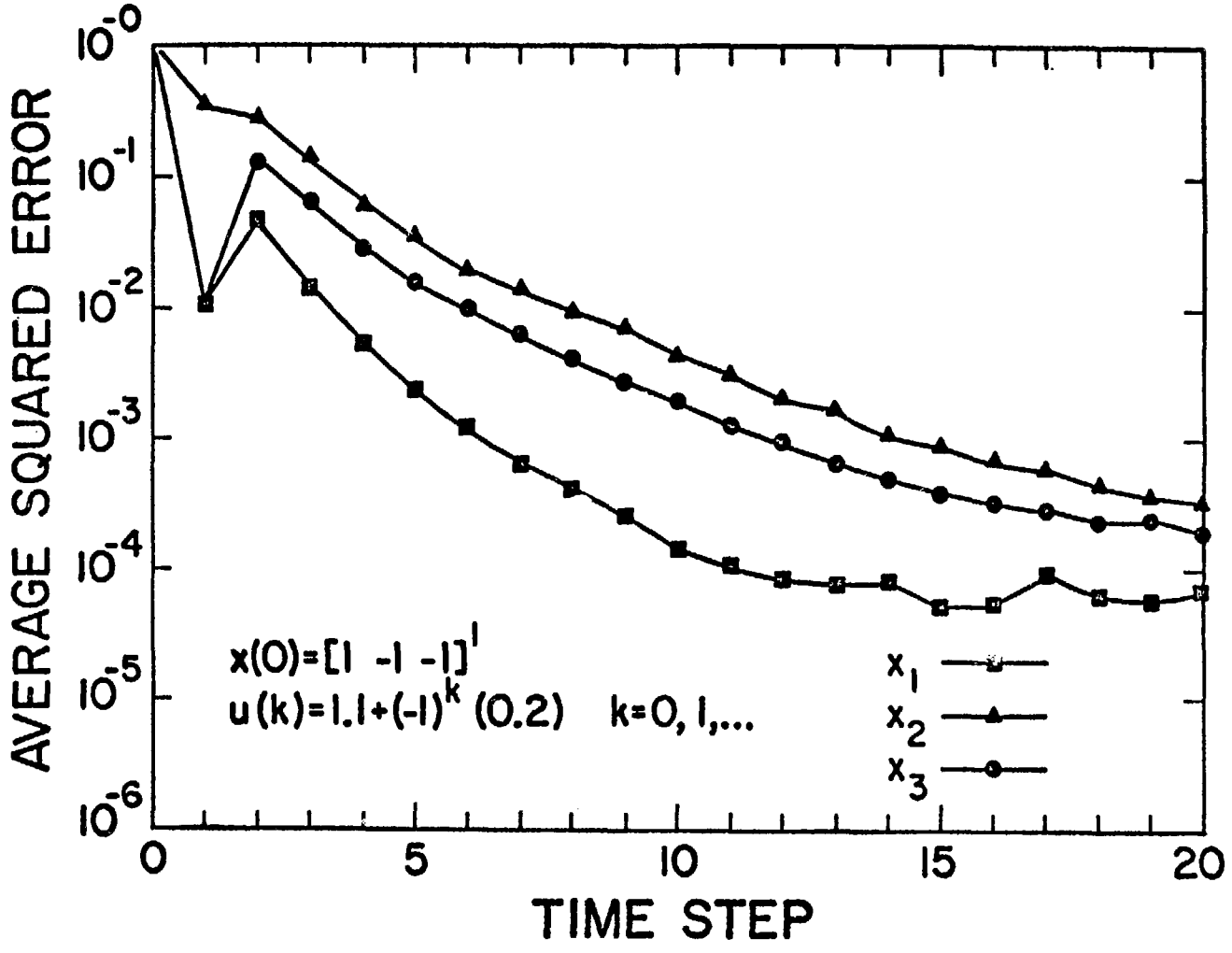

Figure 6.2.5.4 First-Order Conditional Mean Filter Average Squared Error, $u(k) = 1.1 + (-1)^k(.2)$

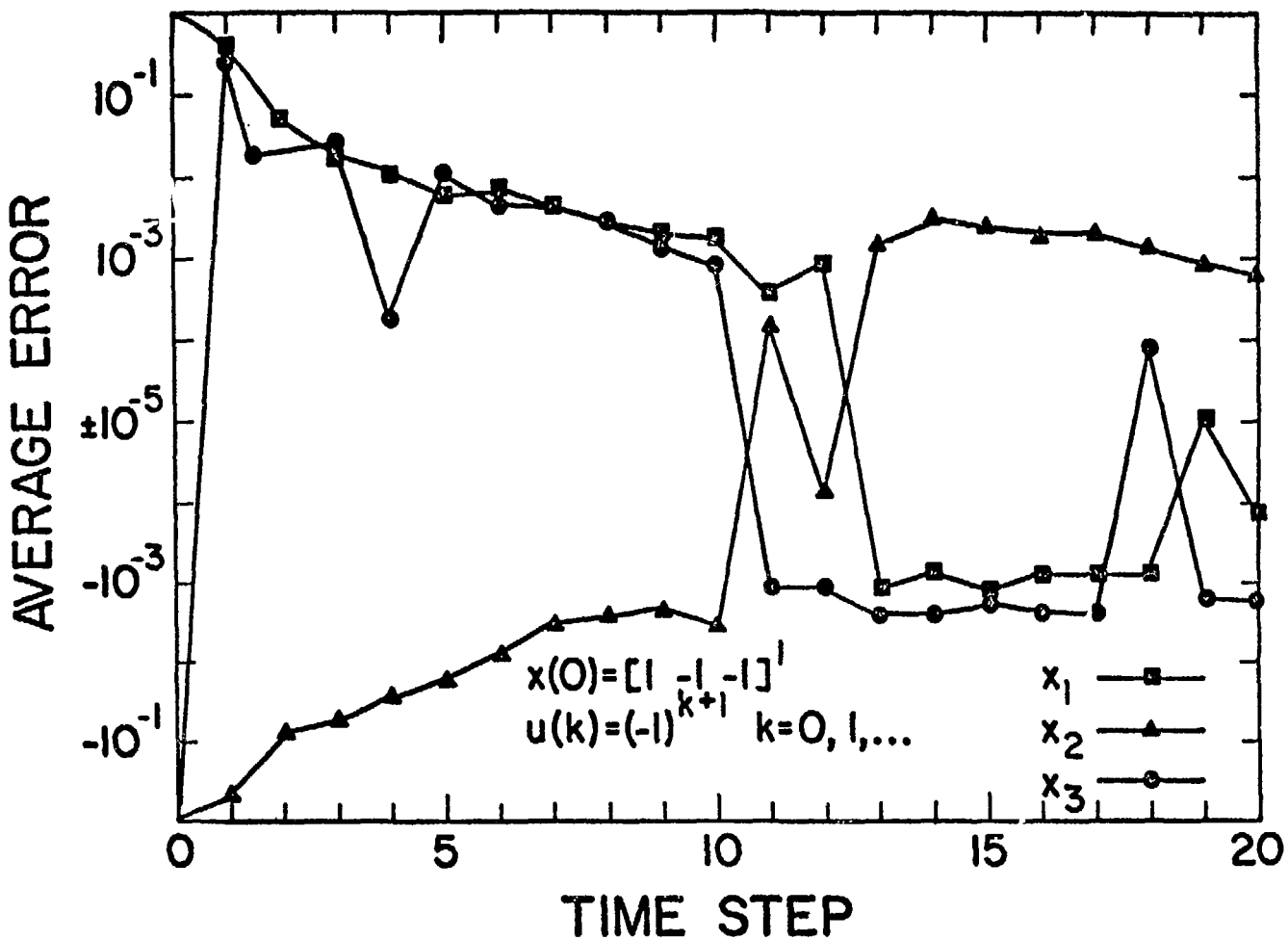

Figure 6.2.5.5 First-Order Conditional Mean Filter Average Error, $u(k) = (-1)^{k+1}$

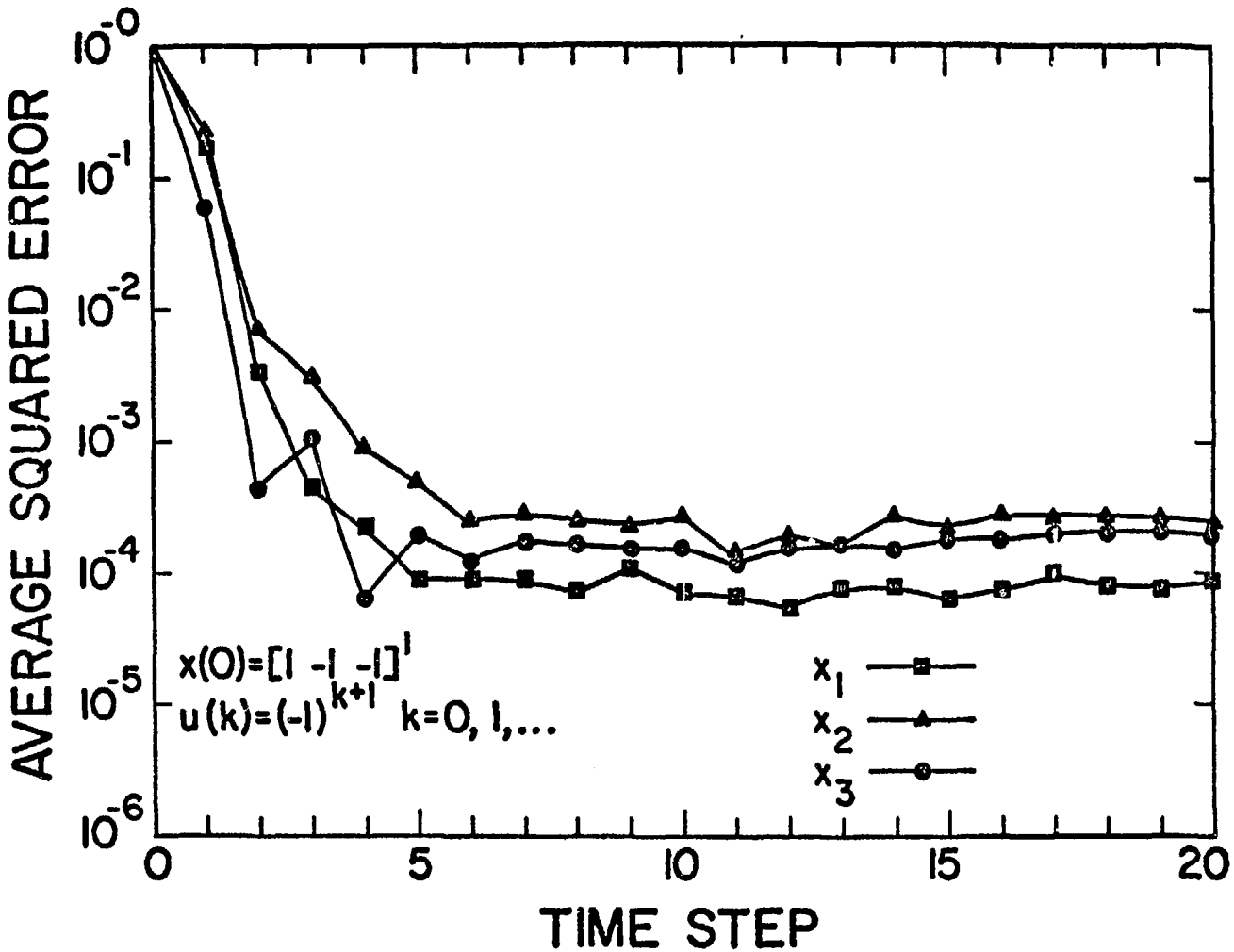

Figure 6.2.5.6 First-Order Conditional Mean Filter Average Squared Error, $u(k) = (-1)^{k+1}$

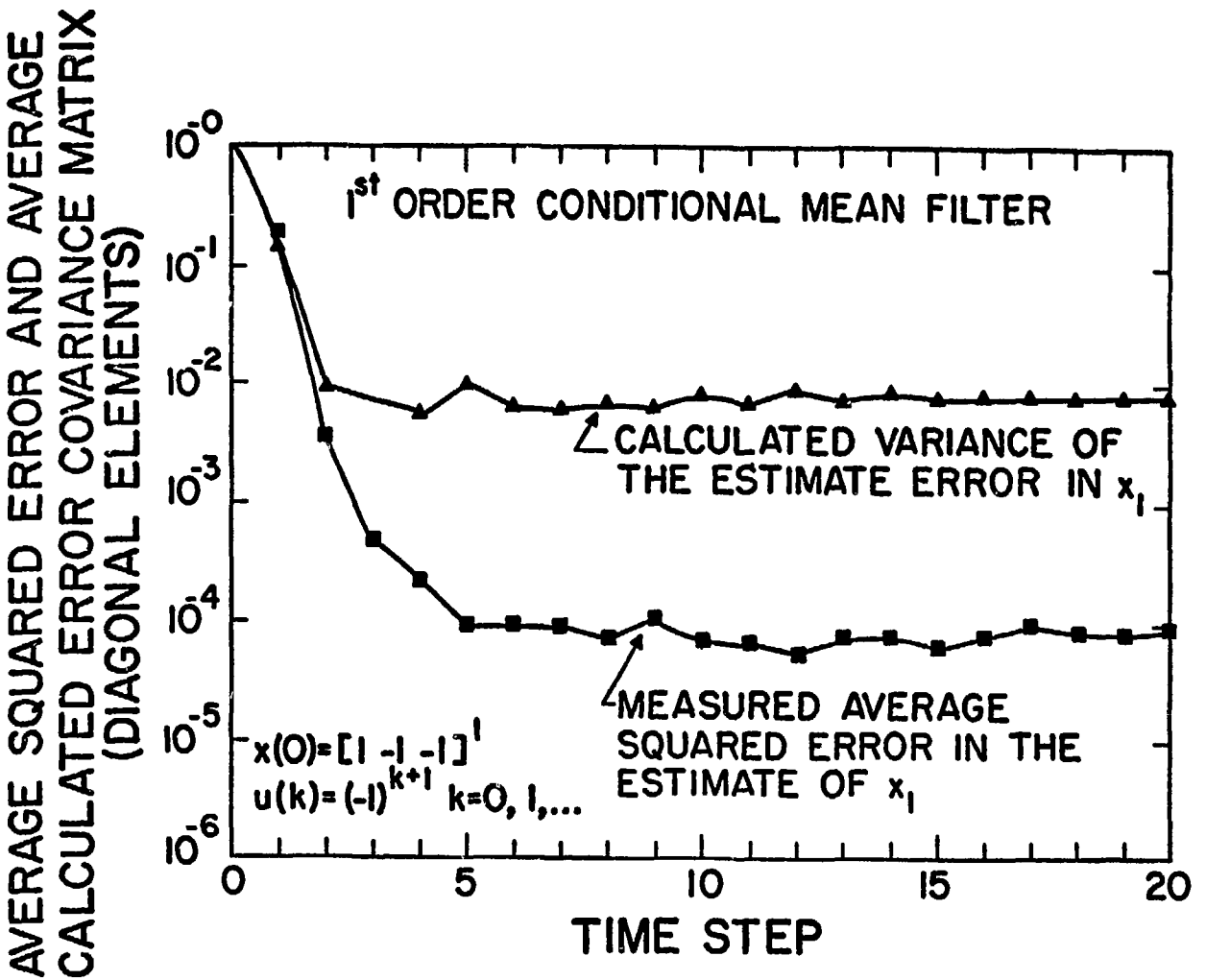

Figure 6.2.6.1 Comparison of Measured Average Squared Error and Calculated Variance of the Estimate Error of the First State Variable

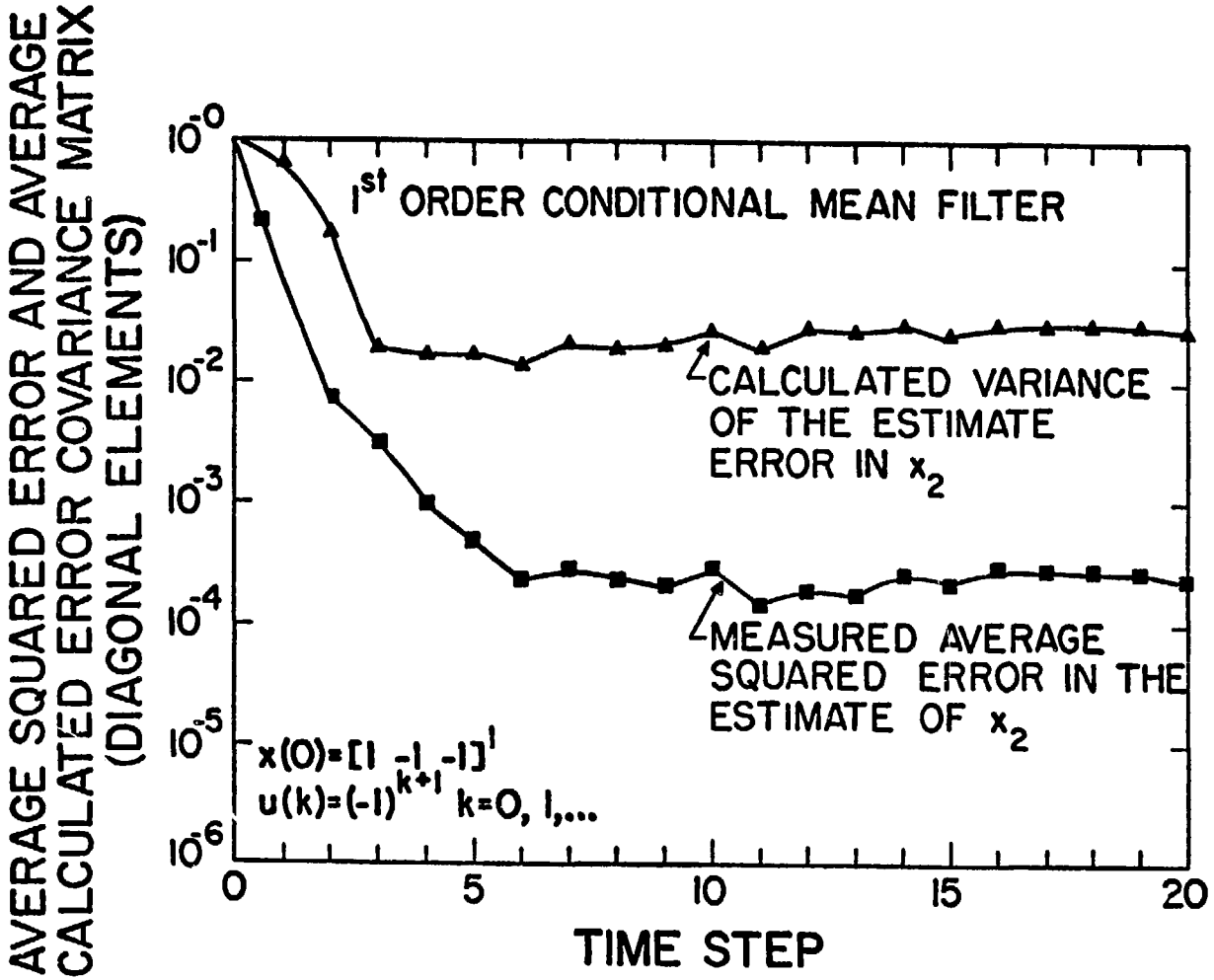

Figure 6.2.6.2 Comparison of Measured Average Squared Error and Calculated Variance of the Estimate Error of the Second State Variable

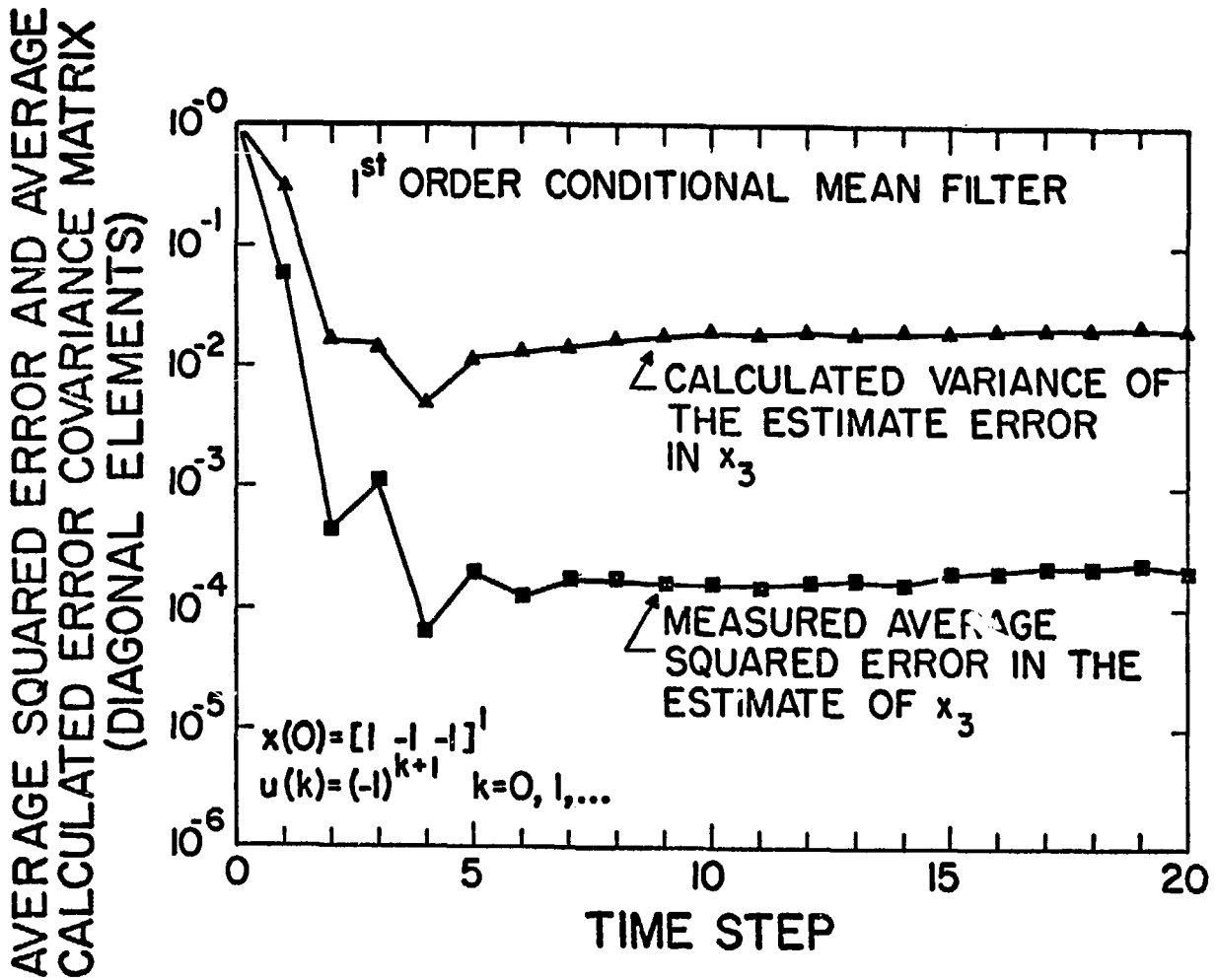

Figure 6.2.6.3 Comparison of Measured Average Squared Error and Calculated Variance of the Estimate Error of the Third State Variable

correction term has the form

$$- \text{trace } QP(k+1|k+1)$$

This will also cause the quantity $\theta(k+1|k)$ given in Table 5.3.4 to be too large, thus making the calculated error covariance at the following step too large also.

For these reasons, the second order filter is not useful for this particular problem. It may be possible to find a more accurate method of calculating the error covariance. Denham and Pines discuss this problem and its possible solution with an iteration filter in reference 6.

6.2.7 The Maximum Likelihood Algorithm

It was found empirically that the constant control $u(k)=1$ $k=0,1, \ldots$ was sufficient for the iterative solution for the initial state to converge. Therefore, this control was used in the simulation. The results are shown in Figures 6.2.7.1 and 2 for the same starting point as the other algorithms. Note that no information is available until time step 3.

It is apparent from Figure 6.2.7.1 that there is a residual bias present in the estimate similar to that in the other filters. However, the magnitude of the bias is generally less. This is consistent with the notion that maximum likelihood estimators frequently result in unbiased estimates.

The average squared error performance, given in Figure 6.2.7.2, is slightly inferior for long times, but generally better, in some cases by two orders of magnitude, for time steps less than 5.

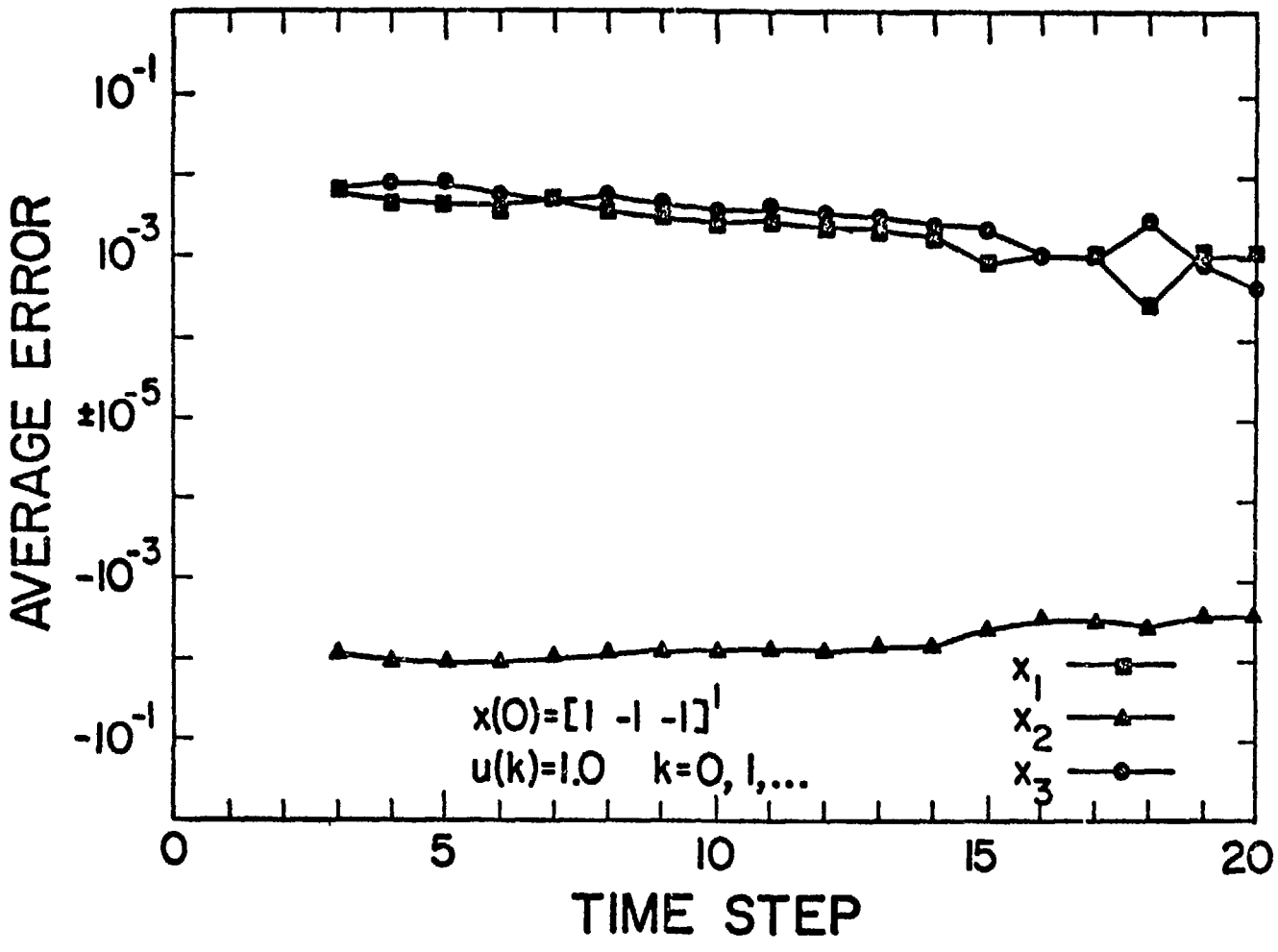

Figure 6.2.7.1 Maximum Likelihood Filter Average Error, u(k) = 1

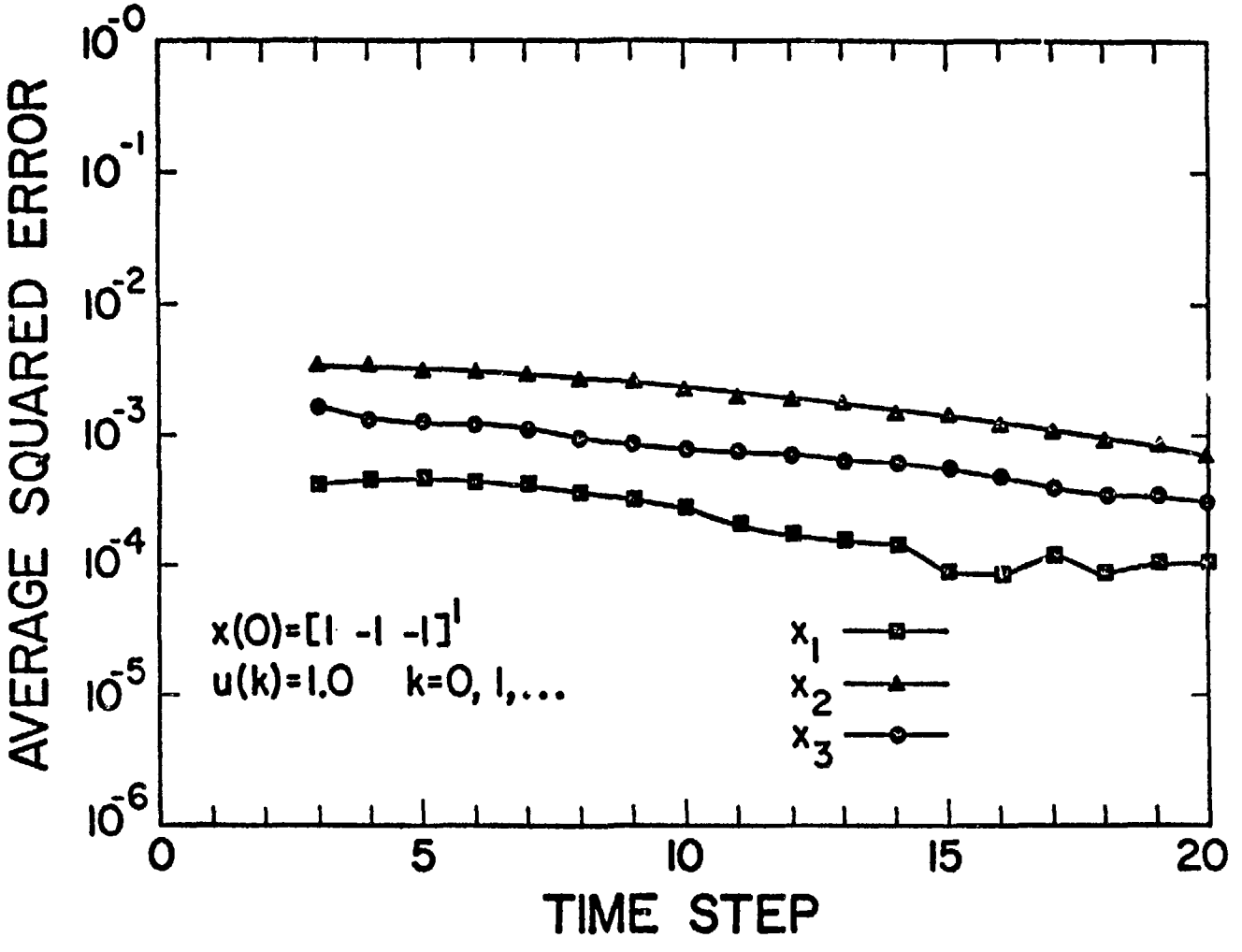

Figure 6.2.7.2 Maximum Likelihood Filter Average Squared Error, u(k) = 1

6.2.8 Further Comparison of the Approximate Maximum A Posteriori, First-Order Conditional Mean, and Maximum Likelihood Algorithms

In this section comparisons are made among the three most interesting algorithms for larger noises. For this reason, a new control sequence of $u(k)=10$ $k=0,1, \ldots$ is chosen. Furthermore, in each of the following graphs, only the first component of the state is plotted. The remaining components behave in a similar manner. Again, the same initial state is chosen as before.

Figures 6.2.8.1 and 2 compare the three algorithms for the same plant noise as before, but with $T(k)$, the measurement noise covariance, one hundred times greater, $T(k)=1$. Evidently, the measurement noise increase has little effect on the ultimate quality of the state estimate. The transient response is changed slightly.

Figures 6.2.8.3 and 6.2.8.4 are the same graphs as 6.2.8.1 and 6.2.8.2, respectively, except that the measurement noise covariance is $T(k)=.01$ and the plant noise covariance matrix is $S(k)=I_3$, the third-order unit matrix. It is obvious that the increase in plant noise has a much more detrimental effect on the quality of the state estimate than a corresponding increase in measurement noise.

6.2.9 The Effect of Small Controls

This section is a confirmation of some of the ideas in Section 2.6. First, a control sequence $u(k)=.1$ for all k that does not satisfy 2.6.1 for k small but does for k large is tried. The results of one run using the first-order conditional mean and approximate maximum a posteriori algorithms are plotted in Figures 6.2.9.1-3. The matrix $P(1|0)$ calculated from the equation $P(1|0) = FP(0)F' + GS(0)G'$

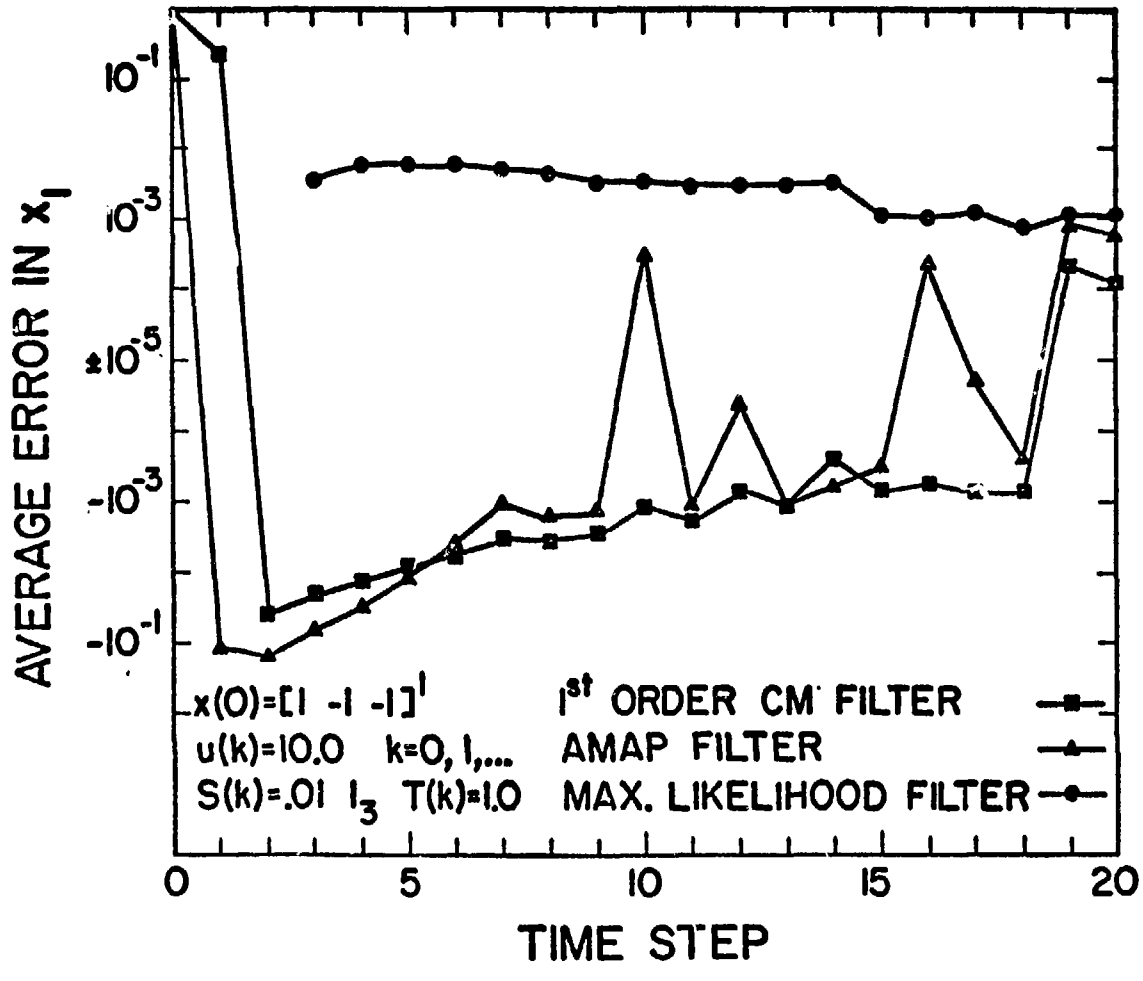

Figure 6.2.8.1 Algorithm Comparison for Larger Measurement Noise, Average Error in x_1

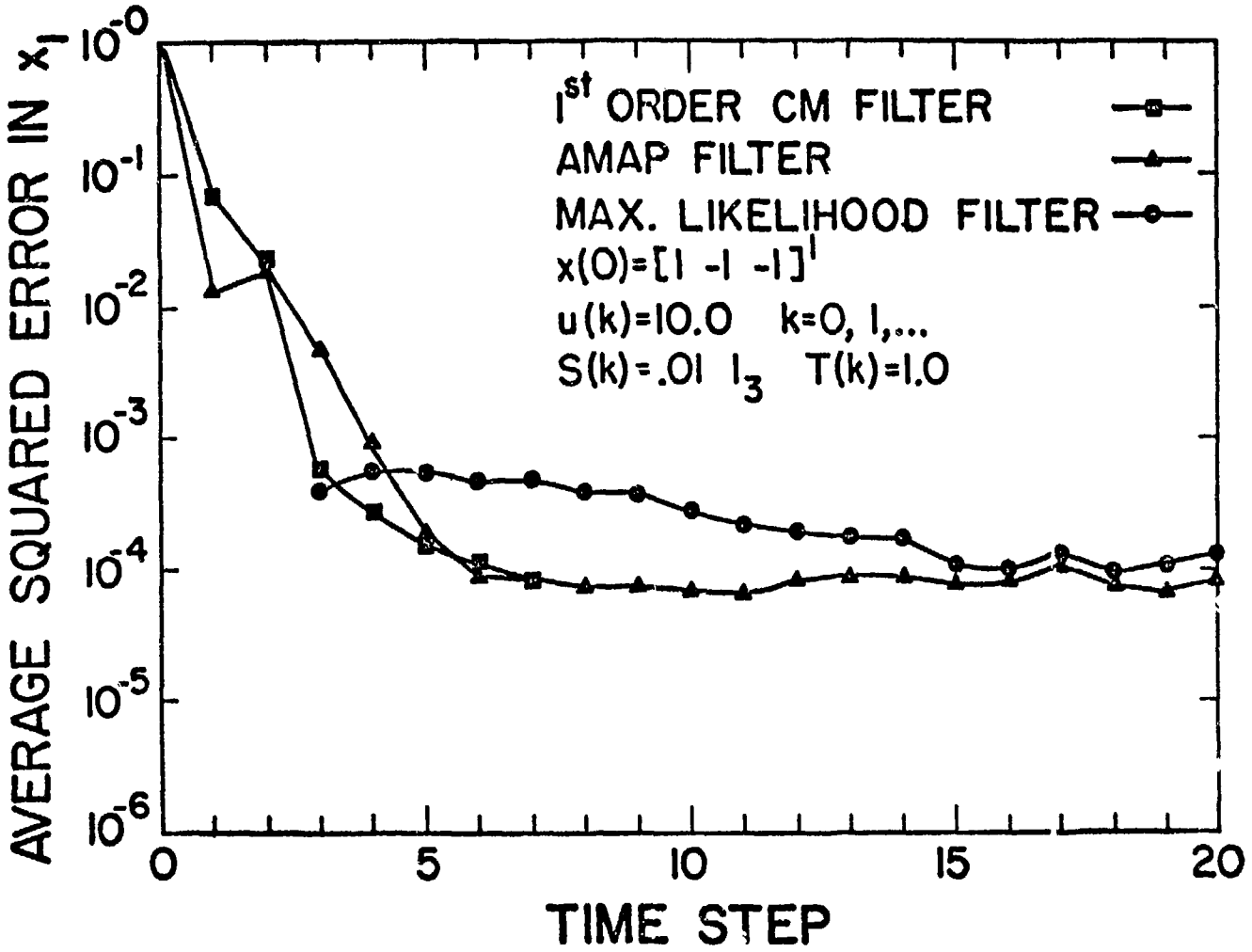

Figure 6.2.8.2 Algorithm Comparison for Larger Measurement Noise, Average Squared Error in x_1

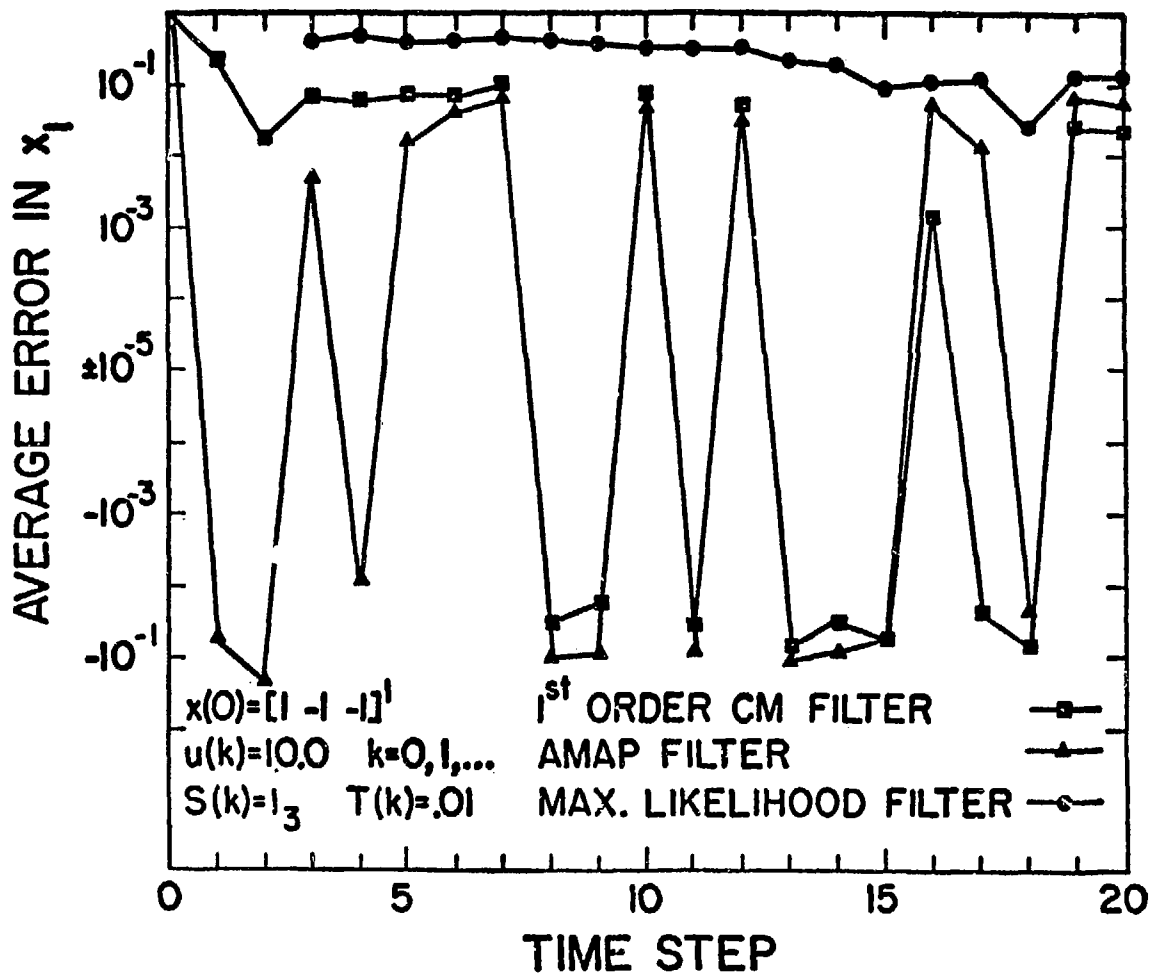

Figure 6.2.8.3 Algorithm Comparison for Larger Plant Noise, Average Error in x_1

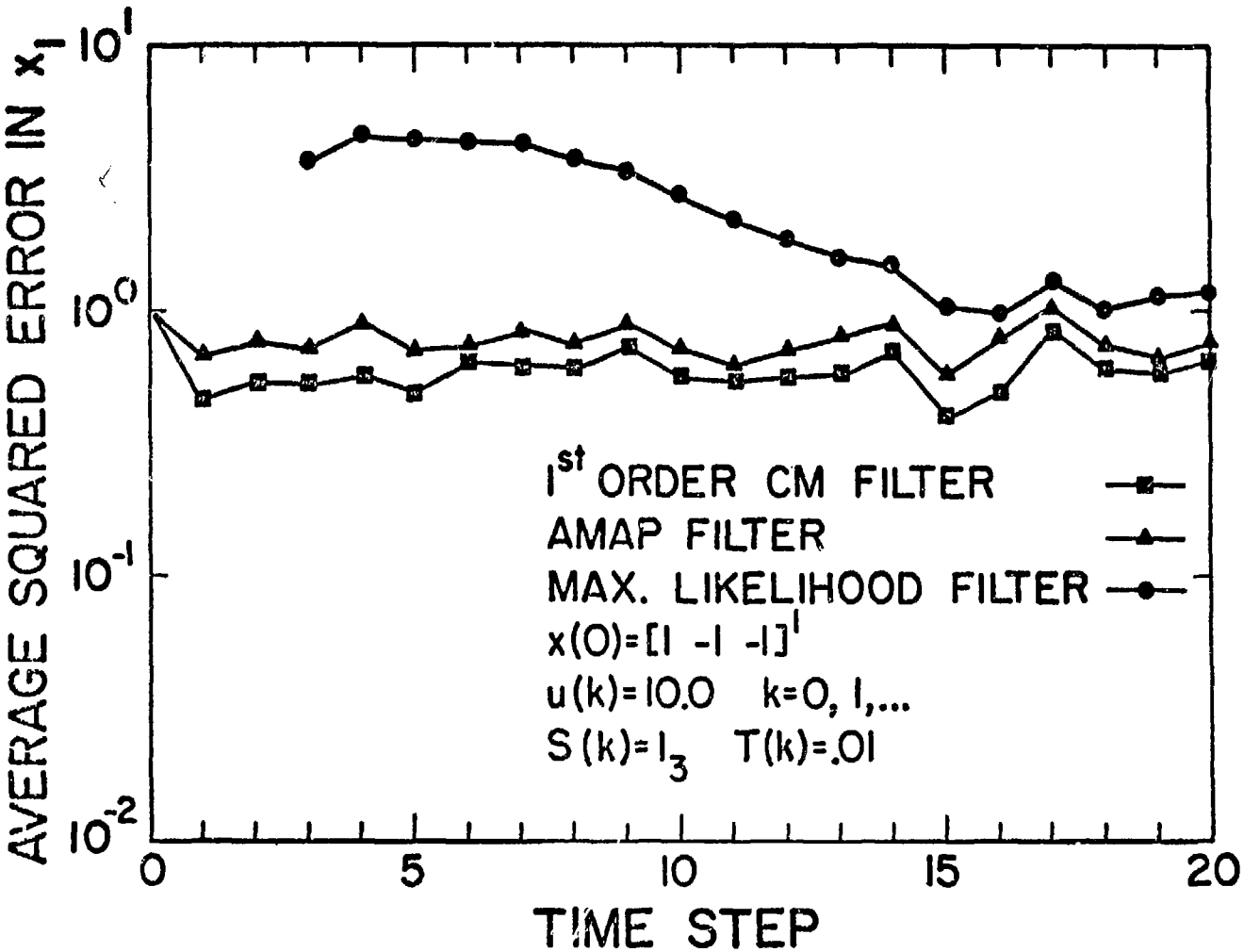

Figure 6.2.8.4 Algorithm Comparison for Larger Plant Noise, Average Squared Error in x_1

has the value

$$P(1|0) = \begin{bmatrix} .722 & -.042 & .449 \\ -.042 & .768 & -.332 \\ .448 & -.332 & .424 \end{bmatrix}$$

and $Bu(0) = [.1698 \quad .2720 \quad -.1113]'$

Equation 2.6.1 is clearly not satisfied at time step k=0. The effect of this as shown in Figures 6.2.9.1-3 is striking. The filter transient responses are very poor, particularly for the first-order conditional mean estimator. This points out another advantage of the AMAP algorithm.

Next, three similar graphs are plotted for u(k)=.01 in Figures 6.2.9.4-6. These graphs show vividly the kinds of problems encountered when the controls are too small. Observe that the first-order conditional mean filter has very large error for k smaller than about 10, and the signs of the estimates of both algorithms are incorrect for most of the observation interval. However, the AMAP estimate looks much more like a mirror image of the true state than does the first-order conditional mean estimate.

The results in Sections 6.2.4 and 6.2.5 were obtained with u(k)=1 for all k so that

$$Bu(k) = [1.698 \quad 2.720 \quad -1.113]'$$

Therefore, assuming that the state uncertainty decreases with time, equation 2.6.1 is satisfied for all time. It would appear that the conditions of Section 2.6 are justified, at least for this problem.

The maximum likelihood estimator does not work for these small controls because of the breakdown of the initialization procedure. Therefore, those results are not presented.

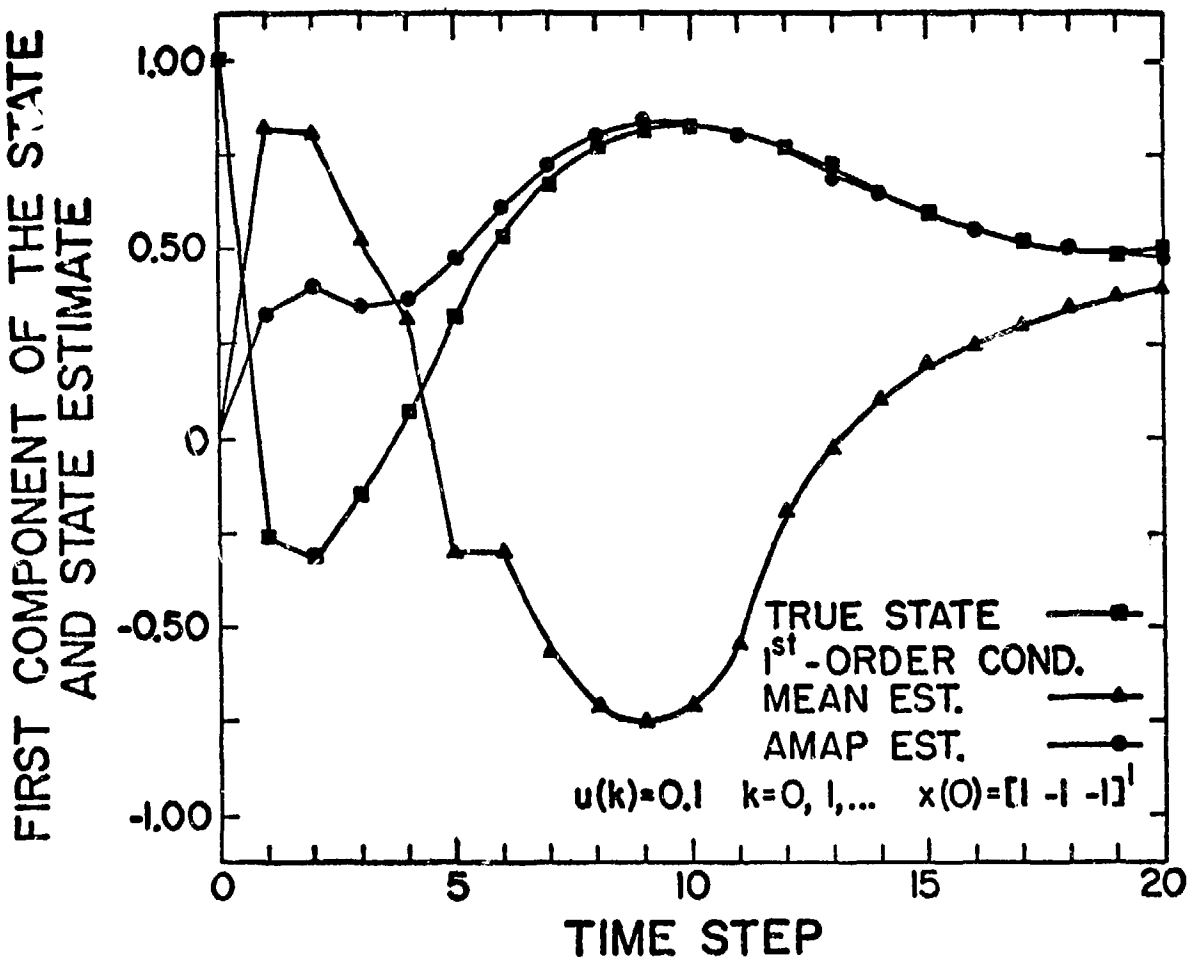

Figure 6.2.9.1 Comparison of the First Component of the State and State Estimates for u(k) = .1

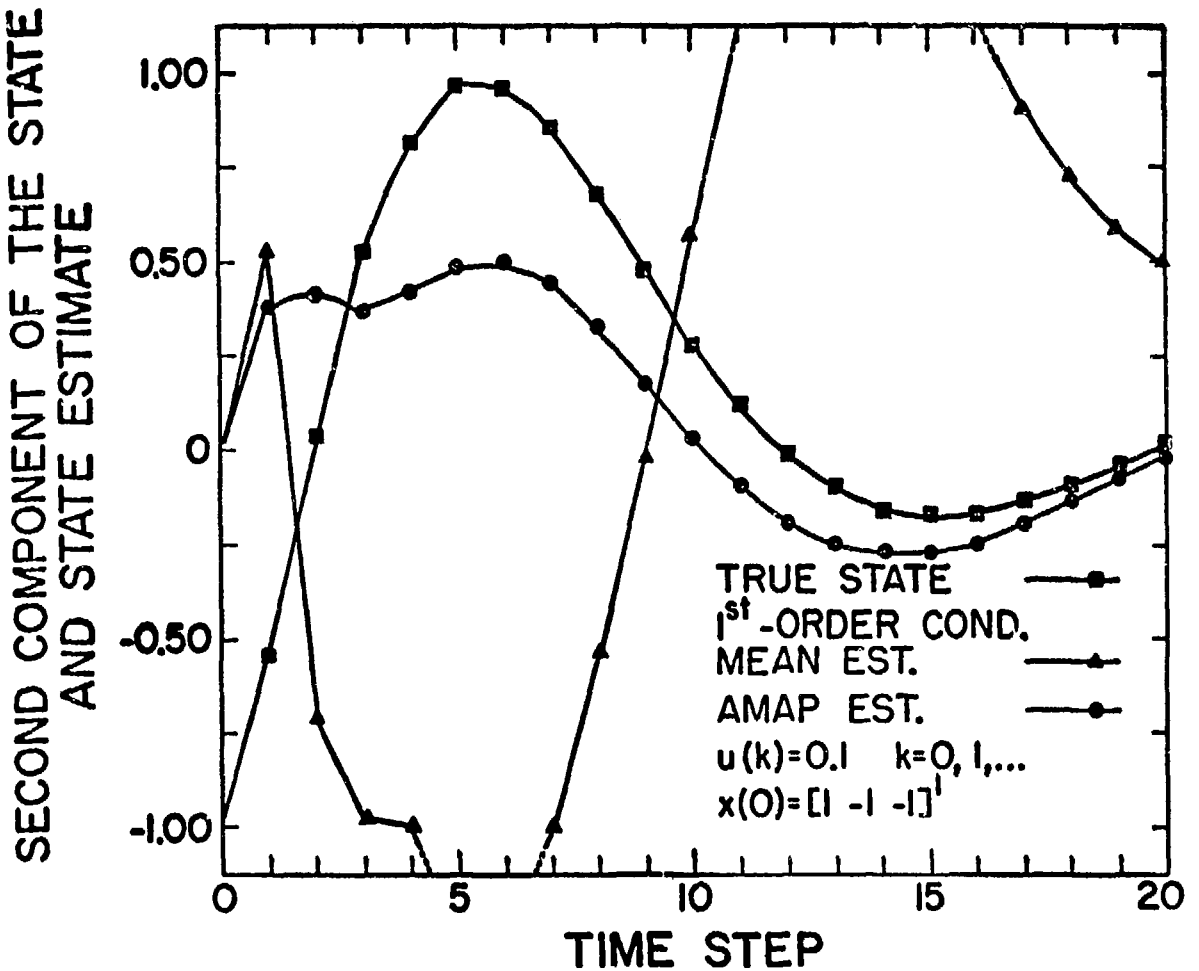

Figure 6.2.9.2 Comparison of the Second Component of the State and State Estimates for u(k) = .1

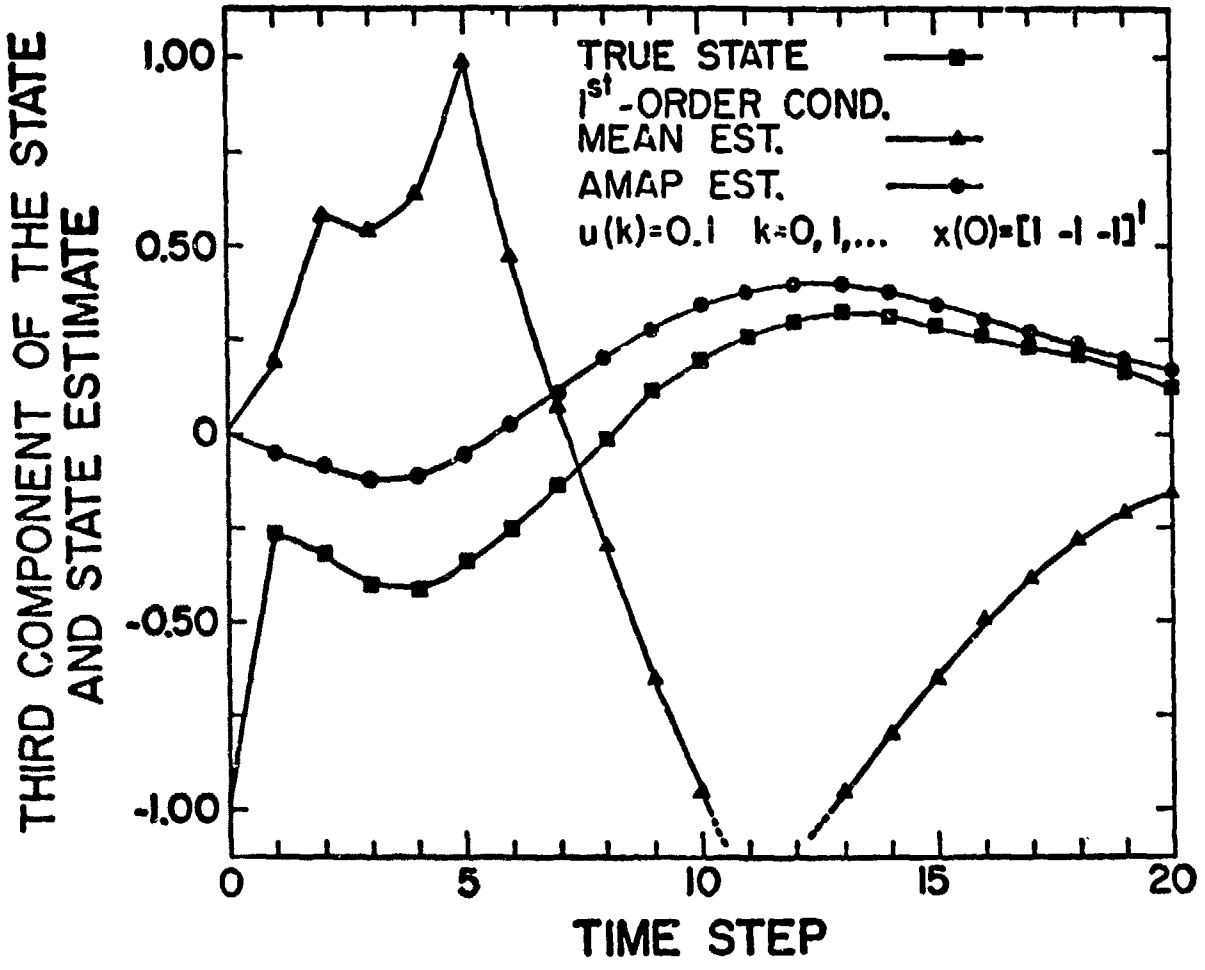

Figure 6.2.9.3 Comparison of the Third Component of the State and State Estimates for u(k) = .1

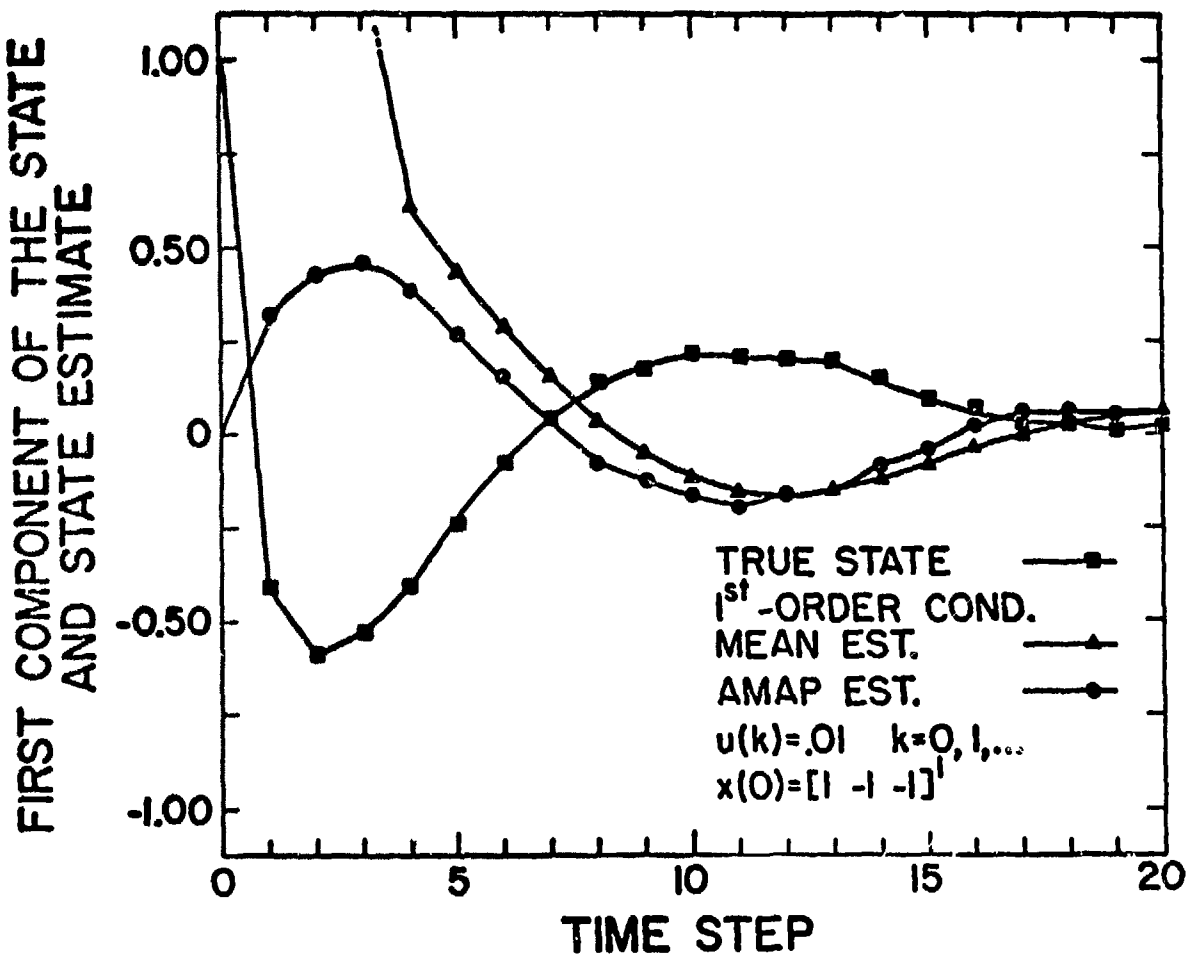

Figure 6.2.9.4 Comparison of the First Component of the State and State Estimates for u(k) = .01

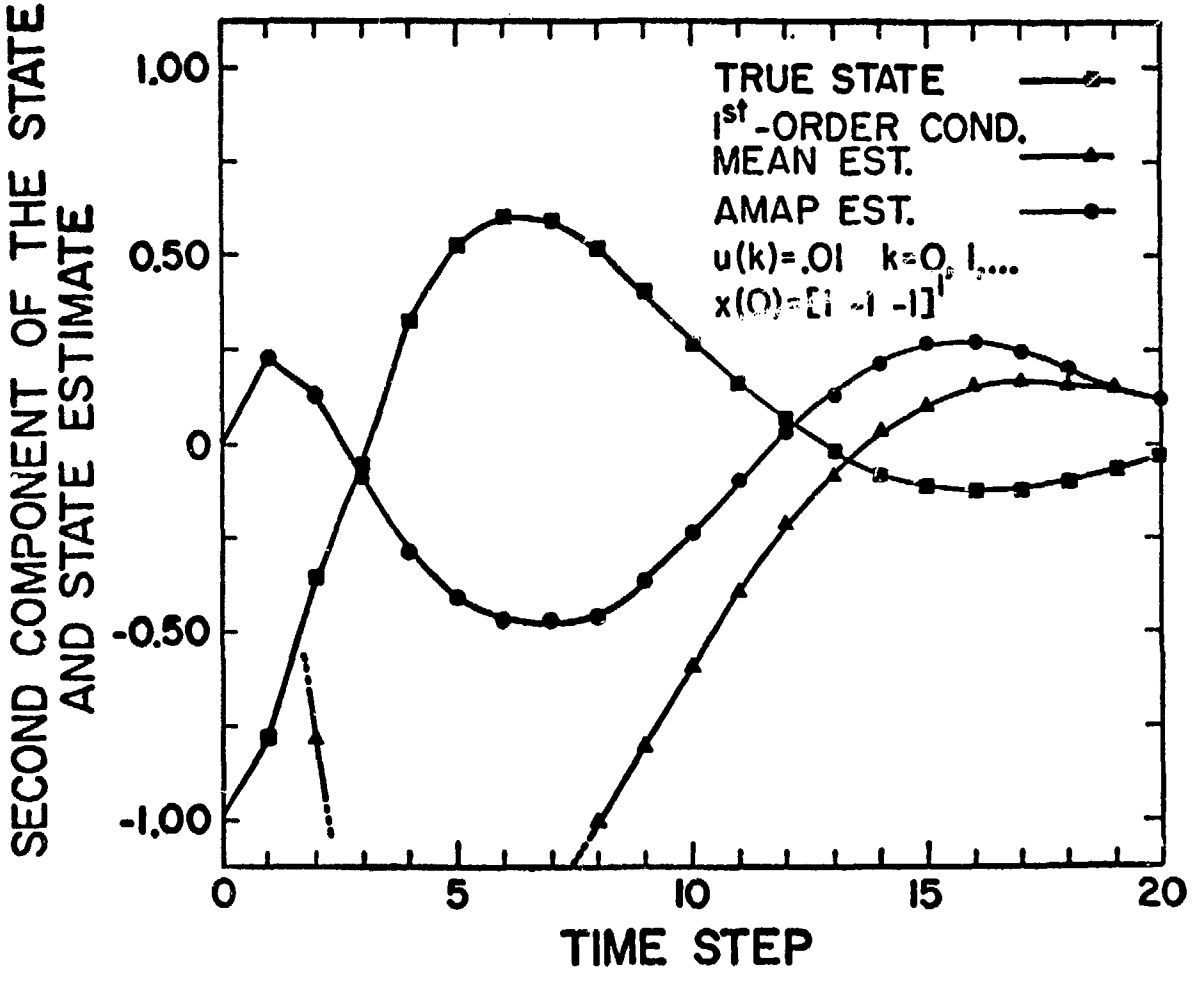

Figure 6.2.9.5 Comparison of the Second Component of the State and State Estimates for u(k) = .01

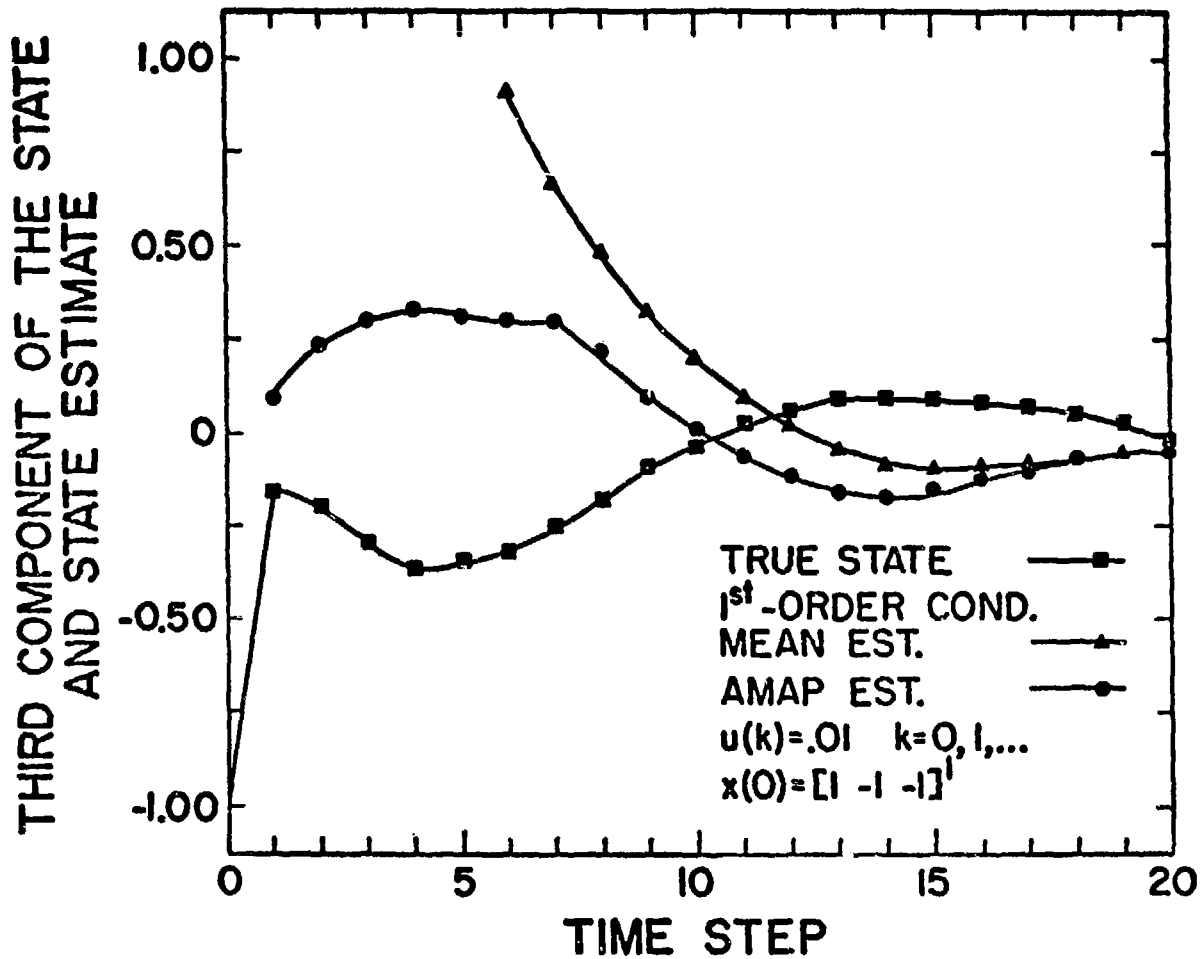

Figure 6.2.9.6 Comparison of the Third Component of the State and State Estimates for u(k) = .01

6.3 Investigation of a Third-Order System with Q = I

This section considers in less detail the same problem as Section 6.2 except that $Q = I_3$, the third-order unit matrix. The only control sequence used is $u(k)=1 \quad k=0,1, \ldots,$ for which the conditions of Section 2.6 are clearly still satisfied. The simulations were done for the same eight initial states as before, and again only the results for $x(0)=[1 \quad -1 \quad -1]'$ are plotted.

The average error and average squared error performances are shown in Figures 6.3.1 and 2, respectively, for the first-order conditional mean filter, those for the approximate maximum a posteriori filter in Figures 6.3.3 and 4, and those for the maximum likelihood estimator in Figures 6.3.5 and 6. Note that in general the transient response is poorer than that for the problem of Section 6.2. It is speculated that this is due to the lack of state variable cross-coupling in the measurement equation with Q=I. The maximum likelihood estimator again seems to be the most unbiased on the average, although the average squared error is ultimately larger.

6.4 Summary

In this chapter, the algorithms previously developed were partially tested and compared and some different control sequences investigated.

The first-order conditional mean and approximate maximum a posteriori algorithms are about equal in error performance if the best (of the ones investigated) control sequence is used for each. However, the AMAP filter does not require calculation of the error covariance and in practice is about three times faster. The correction to the one-step predicted estimate must be calculated on-line in both cases.

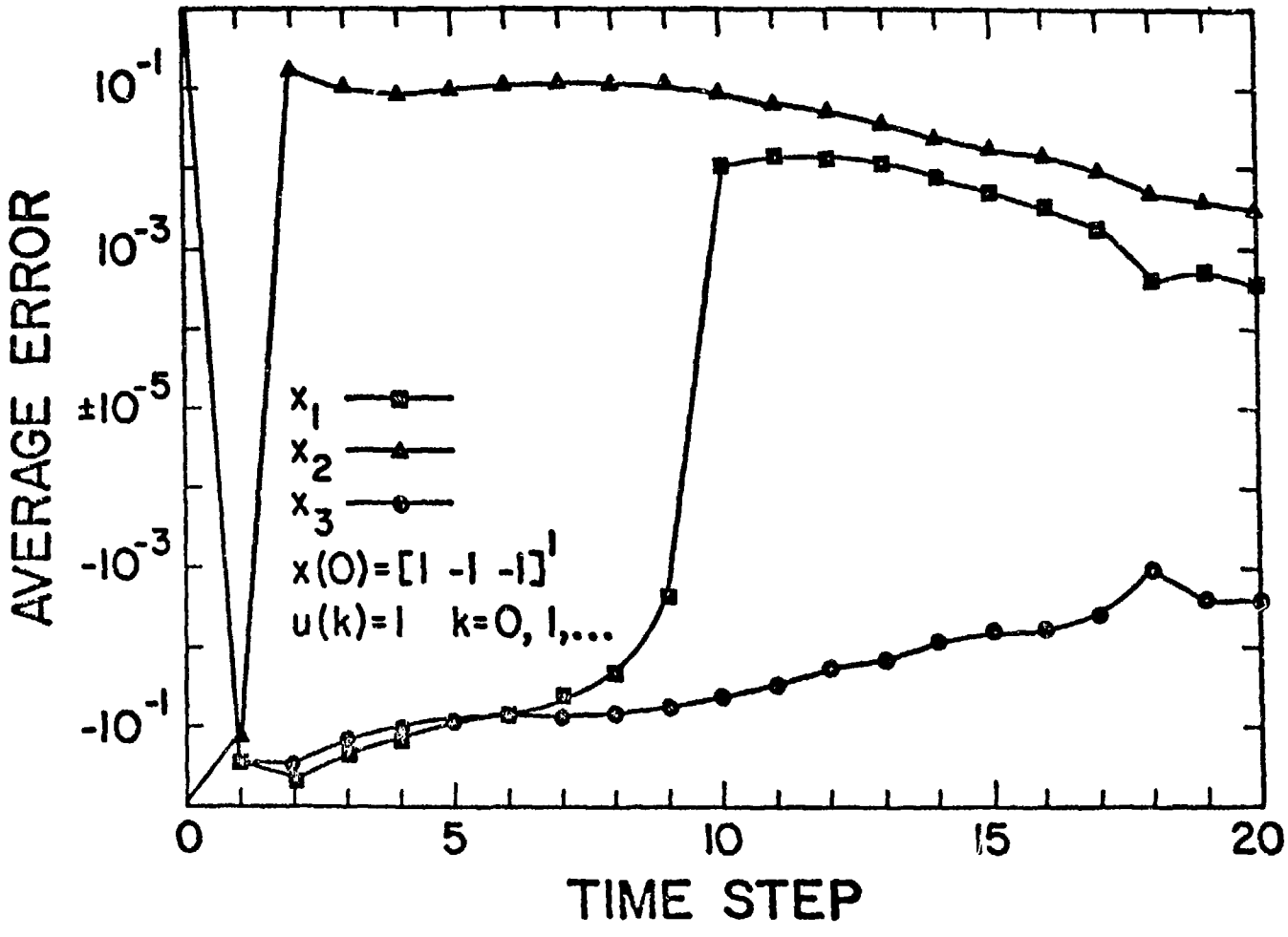

Figure 6.3.1 First-Order Conditional Mean Filter Average Error, Q = I

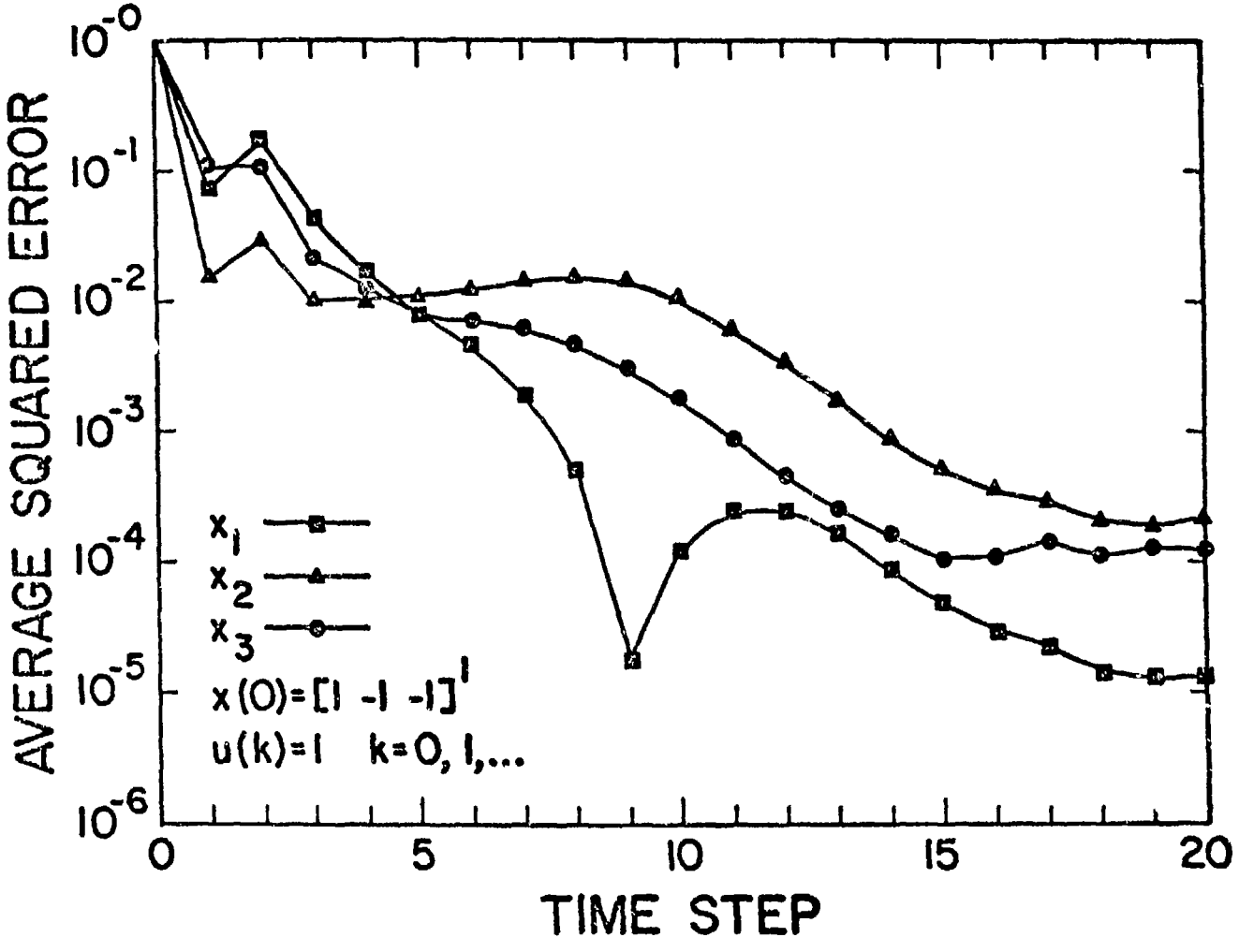

Figure 6.3.2 First-Order Conditional Mean Filter Average Squared Error, Q=I

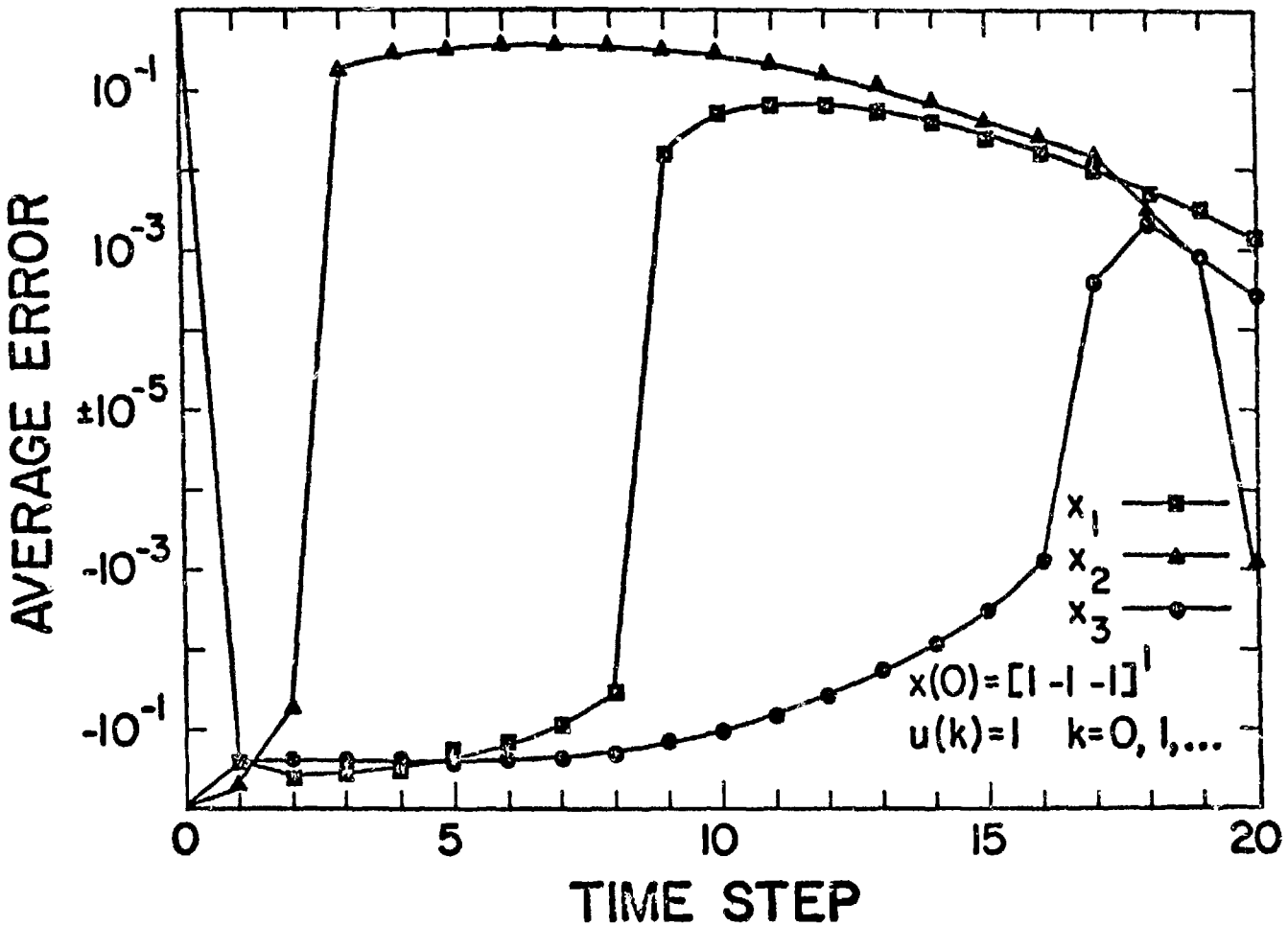

Figure 6.3.3 Approximate Maximum A Posteriori Filter Average Error, Q = I

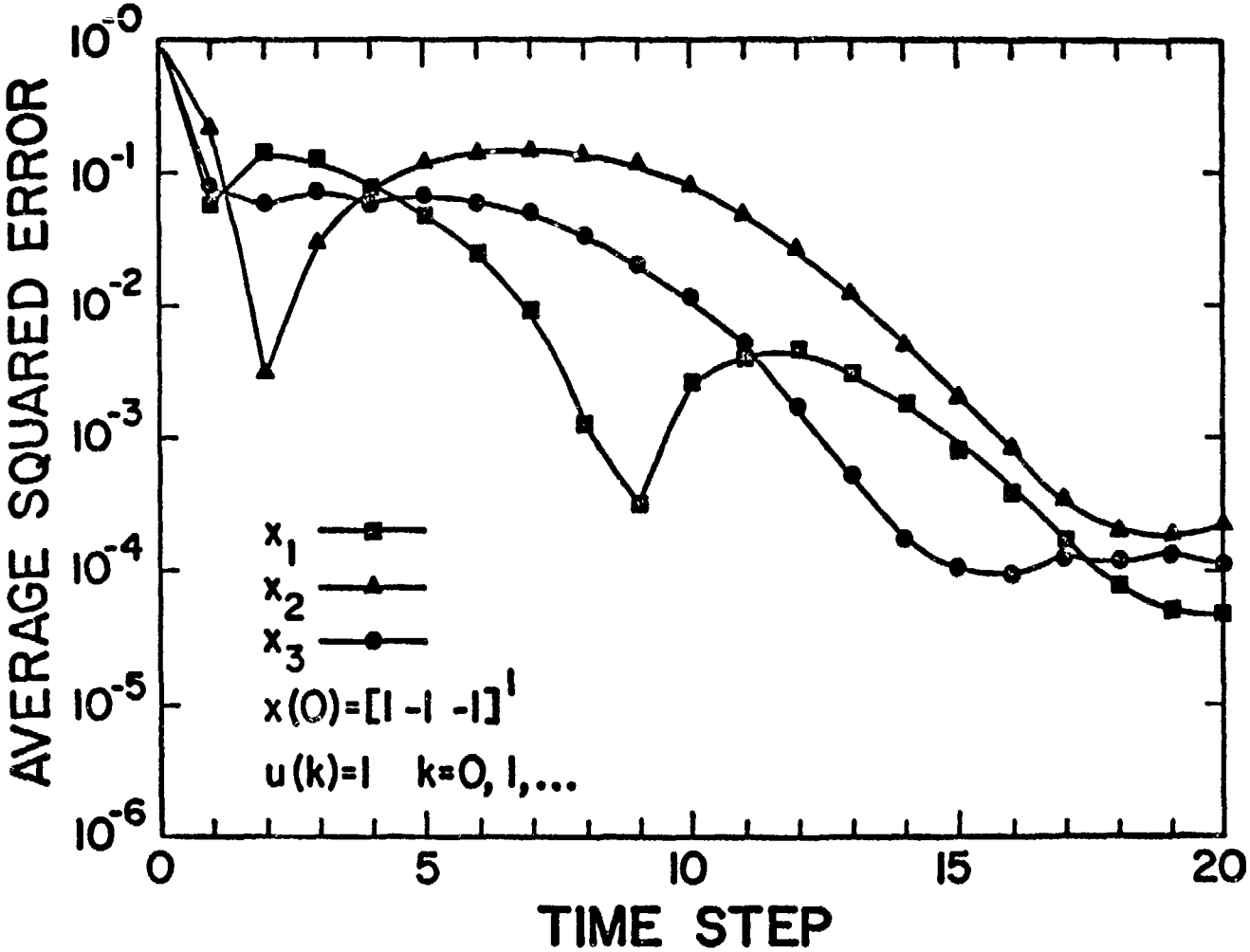

Figure 6.3.4 Approximate Maximum A Posteriori Filter Average Squared Error, Q = I

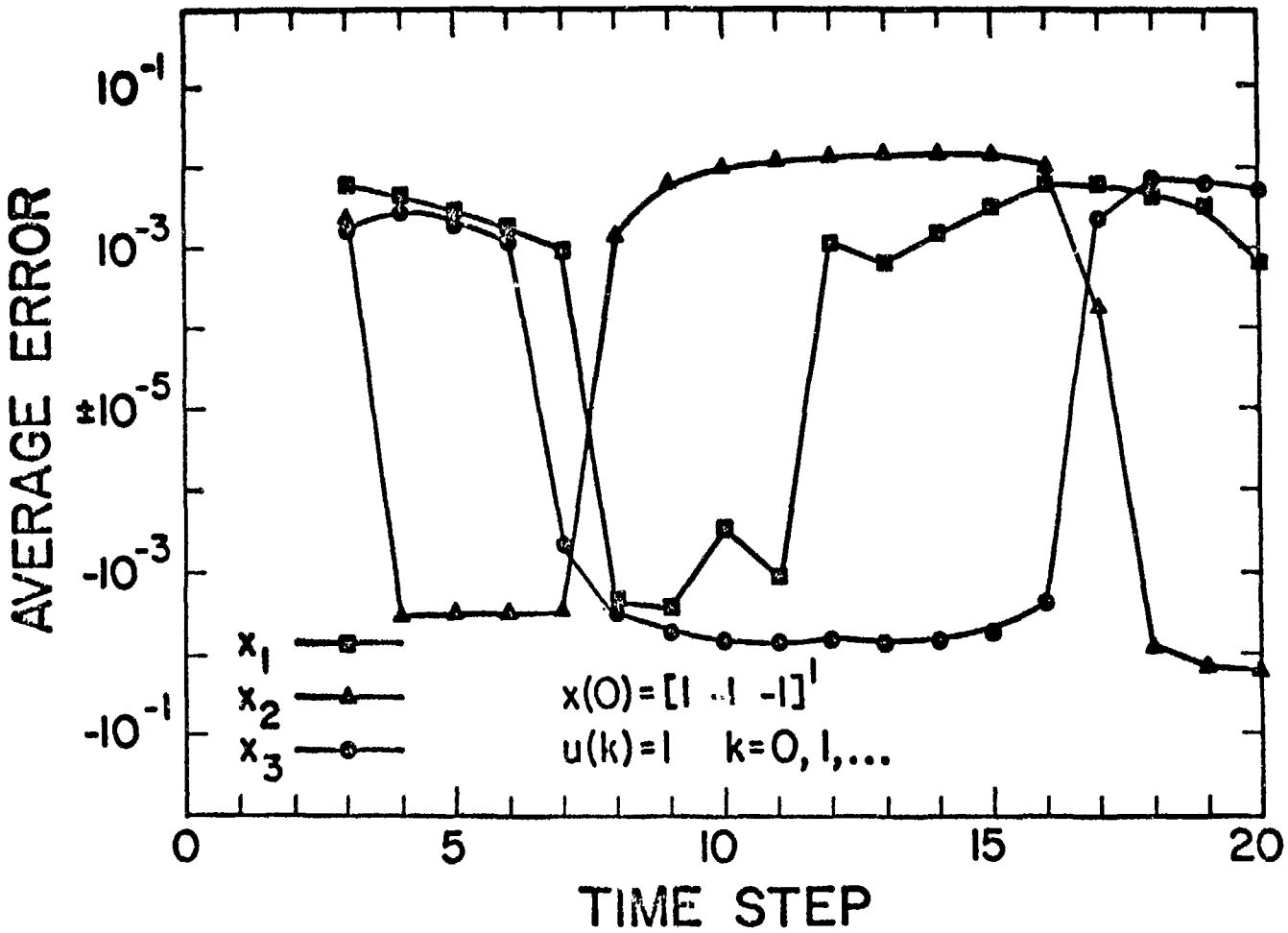

Figure 6.3.5 Maximum Likelihood Filter Average Error, Q = I

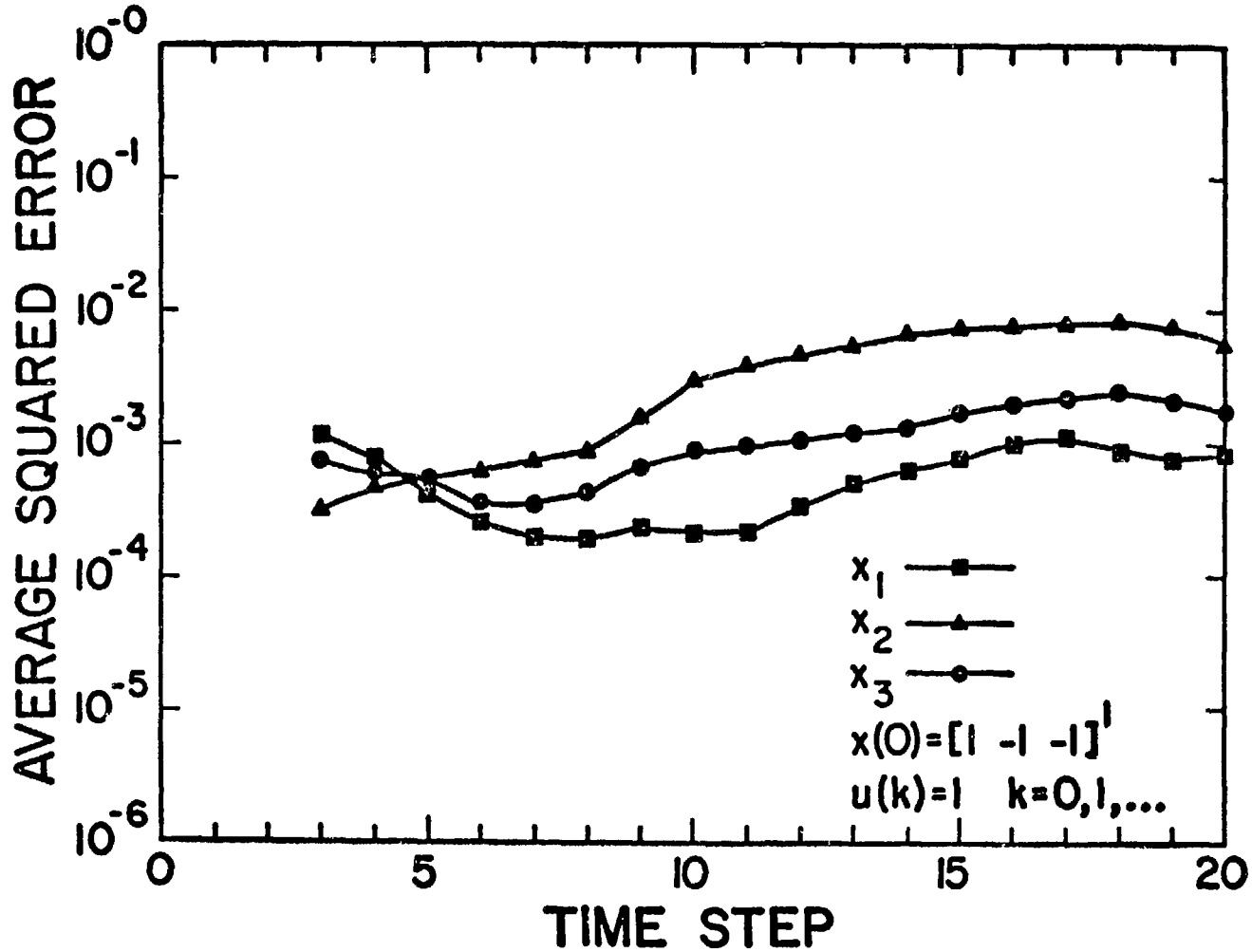

Figure 6.3.6 Maximum Likelihood Filter Average Squared Error, Q = I

The maximum likelihood filter is the cheapest to use in terms of computational time required. However, it has the disadvantages that it does not produce a state estimate until the nth time step, and it generally has larger ultimate estimate error, although the bias is less. A significant advantage is that the filter gain is not dependent on the state, and once a control sequence has been chosen, it can be calculated off-line as in the case of a linear system.

It is safe to conclude on the basis of the results in this chapter that the algorithms developed in this thesis will work on a large class of systems and that they possess some significant advantages over the more commonly used nonlinear filters such as those described by Sage and Melsa [23].

Chapter 7

CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

7.1 Summary of Results

It has been shown in Chapter 2 that the problem posed in Chapter 1 has a solution; i.e., the past controls and measurements contain enough information to estimate the state unambiguously if the controls are large enough. Comforted with this knowledge, three algorithms were obtained for estimating the state in a sequential fashion. The first algorithm was derived by means of the technique of maximum likelihood, the last two by utilizing the more general Bayesian approach espoused by Ho and Lee [8]. In addition, a first-order conditional mean filter was investigated for purposes of comparison. On the basis of the contents of Chapters 5 and 6, the AMAP filter derived in Chapter 4 appears to be the overall best with regard to memory required, speed, accuracy, and utility. It is also clear that quite good state estimates can be obtained for a large class of systems using any of the algorithms if the largeness conditions on the controls are satisfied.

7.2 Directions for Further Research

One obvious extension is to the case of "colored" noise. It is felt that this extension can be easily made using either the techniques of augmenting the state with the noise or measurement differencing (see e.g., Meditch [16].)

Another area of concern is the minimum control required for good state estimation. There surely exist practical problems with control

magnitude constraints that preclude satisfaction of the conditions discussed in Chapter 2. It would be interesting to see if usable algorithms could be developed for this case. Indeed, a useful result would be the elimination of any control. This last goal may be unattainable in the framework of the present problem.

The maximum likelihood estimator could be made much more attractive by (1) streamlining the initialization procedure, (2) providing for state estimates before step n, and (3) obtaining more meaningful lower control bounds in order for the initialization to work. Again, some of these extensions may not be feasible.

The class of problems to which these algorithms are applicable has not been clearly delineated. There may be certain pathological systems other than those discussed for which there are simple tests similar to the controllability and observability criteria for linear systems. Such criteria would obviously be helpful to the design engineer.

Finally, a most promising idea is the application of the technique used for deriving the AMAP filter to more general nonlinear observers. If this technique could be shown to work as well as the traditional linearized Kalman filter (or any of its many variants) on a general nonlinear filter, this would be a very significant result in practical terms.

APPENDIX A

PROBLEMS IN THE SOLUTION OF EQUATIONS 2.2.7

Consider a system of equations similar to 2.2.7, except not restricted to n equations and written in the compact form

$$X^*[x(0)] + 2A\,x(0) + U^* = Z^* \tag{A.1}$$

with all quantities defined similarly to those in equations 2.2.9-12. It is useful and instructive to first consider the case of zero controls. Then A.1 becomes

$$X^*[x(0)] = Z^* \tag{A.2}$$

Remembering that $X^*[x(0)]$ is a column vector of quadratic forms in $x(0)$, it is clear that A.2 has no unique solution. That is, if $x(0)$ is a solution to A.2, then so is $-x(0)$; conversely, there may be no solution.

In order to make things more clear, consider the following two-dimensional example with positive definite Q:

$$x(k+1) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k) \qquad \text{with } u(k) \equiv 0$$

and $$z(k+1) = x'(k+1) \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} x(k+1)$$

and suppose $x(0) = [1 \quad 1]'$

Then, $x(1) = [1 \quad 2]' \qquad z(1) = 10$

$x(2) = [1 \quad 4]' \qquad z(2) = 26$

$x(3) = [1 \quad 8]' \qquad z(3) = 82$

etc.

Using these measurement values, the normal procedure is to substitute them into A.2 and attempt to solve for $x(0)$. When the substitution is made, A.2 becomes

$$2x_1^2(0) + 4x_1(0)x_2(0) + 4x_2^2(0) = 10$$

$$2x_1^2(0) + 8x_1(0)x_2(0) + 16x_2^2(0) = 26$$

$$2x_1^2(0) + 16x_1(0)\ x_2(0) + 64x_2^2(0) = 82$$

etc.

where the subscripts indicate the components of x(0). Since Q is positive definite, these are the equations of ellipses in state space, all centered at the origin. Simultaneous solution of the first pair of equations (corresponding to using only the first two measurements) yields four answers which are the points of interesection of the two ellipses. They are

$$\hat{x}(0) = \pm\ [1 \quad 1]', \pm\ [3.13 \quad -1.79]'$$

When these four values are substituted into the third equation, it is found that only the first pair of solutions satisfy it. It is not difficult to verify that any set of three or more measurement equations have the pair of solutions $\hat{x}(0) = \pm\ [1 \quad 1]'$, and only this pair of solutions, in common. Therefore, the measurements on the system of this example do not contain enough information to give a unique answer.

This problem comes about because all the ellipses have the same center; i.e., x(0) does not appear linearly in A.2. One way to circumvent this difficulty is to apply a non-zero control. Suppose that in the example just given u(2) = 1 so that

$$x(3) = \begin{bmatrix}1\\8\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}1\\9\end{bmatrix} \qquad z(3) = 101$$

and the first three equations of A.2 become

$$2x_1^2(0) + 4x_1(0)x_2(0) + 4x_2^2(0) = 10$$

$$2x_1^2(0) + 8x_1(0)x_2(0) + 16x_2^2(0) = 26$$

$$2x_1^2(0) + 16x_1(0)x_2(0) + 64x_2^2(0) + 2x_1(0) + 16x_2(0) = 100$$

The first two equations result in the same four solutions as before. When these are substituted into the third equation, it is found that the only one which satisfies it is

$$\hat{x}(0) = [1 \quad 1]'$$

which is the true solution. Therefore, for this example there is enough information in these measurements to uniquely determine the state if a non-zero control is applied at the second step.

In general for n^{th} order systems with zero control, no set of measurements can uniquely determine the state because of the symmetry of the surfaces of constant measurement about the origin. Under suitable restrictions on the control distribution matrix and control sequence this symmetry can be destroyed sufficiently so that some number of measurement surfaces (usually greater than n) intersect in only one common point yielding a unique state estimate.

A practical disadvantage to generalizing to n^{th} order systems the method used in the example to find x(0) is that all the solutions to the first n equations must be found, and in general there are 2^n of these. A less important problem is that the estimate of x(0) is not available until some number of measurements have been taken.

The development of Section 2.2 gives a method of solution that alleviates some of these difficulties. The basic idea is to try to find a control sequence such that an iterative procedure for finding

a solution to the first n equations converges to x^*, the original solution to 2.2.13, when the starting point is the origin. It is shown that a control sequence satisfying equations 2.2.36 and such that A^{-1} exists results in this situation.

ACKNOWLEDGMENTS

I wish to express my appreciation to all those persons who offered me their valuable time, experience, and advice in the preparation of this work.

I am particularly indebted to Professor Harold K. Knudsen, Chairman of the Doctoral Committee, for his patience in every discussion and assistance in helping me to become more familiar with the general theory of control systems while developing this topic.

I gratefully acknowledge the interest and assistance of Professors Daniel Petersen, Arnold Koschmann, and Theodore Guinn, members of the Doctoral Committee. In particular, I thank Professor Petersen for enkindling my initial interest in filter theory and for many helpful suggestions.

My sincerest thanks to Richard Hiebert, John Lamb, and William Briscoe of Los Alamos Scientific Laboratory for their aid and encouragement and to the Laboratory itself which provided me with the time and support without which this work would have been impossible.

The financial assistance provided by a National Science Foundation Graduate Traineeship has been invaluable and I extend my gratitude for this grant.

BIBLIOGRAPHY

1. Aoki, Masanao. Optimization of Stochastic Systems. New York: Academic Press, 1967.

2. Aoki, Masanao. "On Some Convergence Questions in Bayesian Optimization Problems." IEEE Transactions on Automatic Control, AC-10 (April, 1965), 180-182.

3. Bellman, Richard. Introduction to Matrix Analysis. New York: McGraw-Hill, 1960.

4. Bryson, A. E., and Frazier, M. "Smoothing for Linear and Nonlinear Dynamic Systems." Aeronautical Systems Division, Wright-Patterson AFB, Ohio, Tech. Rept. No. ASD-TDR-63-119. February, 1963).

5. Cox, Henry. "On the Estimation of State Variables and Parameters for Noisy Dynamic Systems." IEEE Transactions on Automatic Control, AC-9 (January, 1964), 5-12.

6. Denham, Walter F., and Pines, Samuel. "Sequential Estimation When Measurement Function Nonlinearity is Comparable to Measurement Error." AIAA Journal, Vol. 4, No. 6 (June, 1966), 1071-1076.

7. Deutsch, Ralph. Estimation Theory. New Jersey: Prentice-Hall, 1965.

8. Ho, Y. C., and Lee, R. C. K. "A Bayesian Approach to Problems in Stochastic Estimation and Control." IEEE Transactions on Automatic Control, AC-9 (October, 1964), 333-339.

9. Johnson, George W. "A Deterministic Theory of Estimation and Control." IEEE Transactions on Automatic Control, AC-14 (August, 1969), 380-384.

10. Kalman, R. E. "On the General Theory of Control Systems." Proceedings of the IFAC Moscow Congress, Vol. 1 (1960), 481-492.

11. Kalman, R. E., and Bertram, J. E. "Control System Analysis and Design Via the 'Second Method' of Lyapunov, II, Discrete-Time Systems." Transactions of the ASME, Journal of Basic Engineering, Vol. 82D (1960), 371-393.

12. Kantorovich, L. V., and Alsidov, G. P. Functional Analysis in Normed Spaces. New York: Pergamon, 1964.

13. Kushner, Harold. Introduction to Stochastic Control. New York: Holt, Rinehart, and Winston, 1967.

14. Kushner, Harold. Stochastic Stability and Control. New York: Academic Press, 1967.

15. Lee, R. C. K. Optimal Estimation, Identification, and Control, Research Monograph No. 28. Cambridge, Mass.: MIT Press, 1964.

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

17. Ohap, R. J., and Stubberud, A. R. "A Technique for Estimating the State of a Nonlinear System." IEEE Transactions on Automatic Control, AC-10 (April, 1965), 150-155.

18. Ostrowski, A. M. Solution of Equations and Systems of Equations. New York: Academic Press, 1960.

19. Papoulis, Athanasios. Probability, Random Variables, and Stochastic Processes. New York: McGraw-Hill, 1965.

20. Rauch, H. E., Tung, F., and Striebel, C. T. "Maximum Likelihood Estimates of Linear Dynamic Systems." AIAA Journal, Vol. 3, No. 8 (August 1965), 1445-1450.

21. Saaty, Thomas L. Modern Nonlinear Equations. New York: McGraw-Hill, 1967.

22. Saaty, Thomas L., and Bram, Joseph. Nonlinear Mathematics. New York: McGraw-Hill, 1964.

23. Sage, Andrew P., and Melsa, James L. Estimation Theory With Applications to Communication and Control. New York: McGraw-Hill, 1971.

24. Sorenson, H. W., and Alspach, D. L. "Gaussian Sum Approximations for Nonlinear Filtering." IEEE Symposium on Adaptive Processes (9th) (1970), XIX3.1-XIX3.9.

25. Spragins, John. "A Note on the Iterative Application of Bayes' Rule." IEEE Transactions on Information Theory, IT-11 (October, 1965), 544-549.

26. Stern, T. E. Theory of Nonlinear Networks and Systems. Reading, Mass.: Addison-Wesley, 1965.

27. Wishner, R. P., Tabaczynski, J. A., and Athans, M. "A Comparison of Three Non-Linear Filters." Automatica, Vol. 5 (1969), 487-496.

☆ U. S. GOVERNMENT PRINTING OFFICE: 1973 — 784-278/16

ALT: 349(110)