

UNCLASSIFIED

AD 282 873

*Reproduced
by the*

**ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA**



Best Available Copy

UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

Best Available Copy

282 873

ASD-TR-61-27
VOLUME I

282873

CATALOGED BY ASTIA
AS 500 140.

FUNDAMENTAL STUDY OF ADAPTIVE CONTROL SYSTEMS

TECHNICAL REPORT No. ASD-TR-61-27, VOLUME I

APRIL 1962

ASTIA
SEP 5 1962
TISIA A

FLIGHT CONTROL LABORATORY
AERONAUTICAL SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

Project No. 8225, Task No. 82181

Best Available Copy

(Prepared under Contract No. AF 33(616)-6952
by RIAS Division, The Martin Company, Baltimore, Maryland.
Authors: R. E. Kalman, T. S. Englar, and R. S. Bucy)

NO OTS

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

ASTIA release to OTS not authorized.

Qualified requesters may obtain copies of this report from the Armed Services Technical Information Agency, (ASTIA), Arlington Hall Station, Arlington 12, Virginia.

Copies of this report should not be returned to the Aeronautical Systems Division unless return is required by security considerations, contractual obligations, or notice on a specific document.

FOREWORD

The original research and development upon which this report is based was accomplished by The Research Institute for Advanced Studies (RIAS, Division of The Martin Company), Baltimore, Maryland, under Air Force Contract AF 33(616)-6952. Fundamental Study of Adaptive Control Systems. This is the first report to be issued under this contract, additional volumes are to follow.

The work was administered under the direction of the Flight Control Laboratory, Aeronautical Systems Division. Lt. L. Schwartz and Lt. P. C. Gregory were task engineers for the Laboratory.

The authors are members of The Center for Differential Equations at RIAS.

Dr. R. E. Kalman served as principal investigator.

Acknowledgement is made of the assistance provided by the computer staff of The Martin Company.

This document is unclassified.

ABSTRACT

This is the first detailed report on Contract AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.

A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.

To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

PUBLICATION REVIEW

The publication of this report does not constitute approval by the Air Force of the findings or conclusions contained herein. It is published only for the exchange and stimulation of ideas.

FOR THE COMMANDER:


C. R. BRYAN
Technical Director
Flight Control Laboratory

TABLE OF CONTENTS

<u>Chapter</u>		<u>Page</u>
	Introduction.	1
I	Conceptual Background	4
II	The Noise-Free Regulator Problem.	15
III	A Third-Order Optimal Regulator Problem . . .	45
IV	Optimal Filtering Theory.	57
V	The General Control Problem	67
VI	The Adaptive Control Program.	71
VII	Guiding Principles of Numerical Computation .	79
VIII	The Exponential Subroutine.	82
IX	The Integral Exponential Subroutine	92
X	The Transient Program	96
XI	The Matrix Riccati Equation	98
	Appendix 1 - [R. E. KALMAN (1961 C)] "New methods and results in linear filtering and prediction theory", Proc. Symp. on Engineering Applications of Probability and Random Functions, Purdue University, November 1960; to be published by Wiley.	109
	Appendix A - The Pseudo-Inverse of a Matrix . .	229
	Appendix B - Gaussian Random Vectors	254
	References	241

INTRODUCTION

This is the first detailed report on Contract AF 33(616)-6952.

This contract is part of a continuing effort in the broad study of complex systems problems. Accordingly, it was thought that an encyclopaedic presentation of all results obtained to date was not particularly desirable; instead we have concentrated on giving a careful account of the conceptual and mathematical foundations of the research problems involved. It is hoped, therefore, that the report will serve as an introduction to papers on modern system theory which are now appearing in the literature with increasing frequency.

The report consists of two main parts.

Chapters 1-6 give the technical and mathematical background of our present approach to system theory, with particular attention to the adaptive problem. This part is primarily concerned with a clear exposition of the fundamental ideas, with numerous illustrative examples. No attempt has been made to state all mathematical facts with absolute precision, and in particular most proofs are omitted. Further details, which are often very involved and technical, may be found in the references.

Chapters 7-11 are concerned with the motivation and description of digital computer techniques used for automatic synthesis of optimal systems. These methods are a fairly radical departure from current engineering practices in the systems field. They are, therefore, explained in considerable detail. It is expected that this material will eventually be incorporated into a "handbook" of instructions for the everyday usage of the automatic optimization program which we are now developing. Besides a complete description of the subroutines, a number of check programs and solutions are given which should facilitate use of the program by others.

It is difficult to give a fair description of the advantages as well as shortcomings of the methods used in this report or of the original contributions which are involved. Two main aspects should be emphasized, however.

(1) A comprehensive analytical theory and method of linear system optimization has been developed (with partial support of this contract) and is currently nearing completion. The most important features of this method are:

- (a) It is new and quite different from conventional methods.
- (b) It is applicable to linear systems of any degree of complexity; in other words, no modifications are needed to treat multi-input or multi-output systems.
- (c) It is applicable in principle without any modification (but possibly at great cost of computation) to linear systems with varying coefficients.
- (d) It provides a unified treatment of control and filtering problems, combinations of the two, etc.
- (e) It provides a canonical block diagram for the optimal system which can serve as the starting point of engineering design.
- (f) It is well suited to high-speed digital computation.

The last property of the new method of analysis (sometimes called the "state-transition" method) gives rise to the second important contribution of the report:

(2) A comprehensive system of numerical computations is being developed to implement the theory. The computer programs will be "automated" to a very large extent and eventually it is hoped that they will be available to engineers without detailed theoretical training. Specifically, the following has been accomplished so far:

- (a) Matrix subroutines have been developed which represent a new approach to the usual computational problems of transient response, stability, etc.
- (b) These subroutines are quite simple from the mathematical point of view and allow good control of numerical errors.
- (c) The method of computations is "eigenvalue-free"; in other words, it does not entail the solution of algebraic equations of high degree which is characteristic of conventional techniques. As a result, our methods can be extended much more easily to large scale systems than the conventional ones.

For ease of cross-referencing with the other volumes of this report - which are to be issued later - each chapter is written in as self-contained a way as possible. Equations and figures are numbered separately within each chapter. References are made by author and year of publication; each chapter contains its own list of references, even when this entails some duplication; duplicate references occurring in different chapters are designated in a consistent fashion.

Since details of the mathematical arguments used in this report are not yet readily available, a recent paper by R. E. Kalman, "New Methods and Results in Linear System Theory" is included as an Appendix. This paper contains a very extensive discussion of the theoretical aspects of the optimal filtering problem.

Chapter 1.

CONCEPTUAL BACKGROUND

1. Introduction.

The fundamental mathematical problem in the design of a control system is the specification of the control law.

We are given a dynamical system to be controlled, called the control object. Examples: (i) an airplane, (ii) a satellite, (iii) a chemical plant. Information about the physical behavior of the control object is conveyed by means of certain physical measurements $z(t)$. These measurements may be (i) altitude, Mach number, pitch angle, etc., of an airplane; (ii) distance of a satellite from the moon; (iii) composition of a chemical formed in a reactor. The behavior of the dynamical system may be affected by changing certain physical parameters $u(t)$, called control variables. Control may be exerted through (i) aileron deflection or engine throttle in an airplane; (ii) jets or flywheels inside a satellite; (iii) heat or catalysts influencing the rate of a chemical reaction.* The control law is a prescription for determining the instantaneous values of the control variables $u(t)$ on the basis of present and past measurements of $z(t)$. The control law may also depend on certain other parameters specifying the desired behavior of the dynamical system under control.

The problem of determining a control law can be easily stated in conceptual terms, but the precise mathematical formulation is not a simple matter. Careful assumptions must be made concerning the mathematical model which is to represent the control object, and one must specify in what sense the control law is to be optimal. Without a clearly defined model and a clearly understood criterion of optimality, sophisticated mathematical techniques are uncalled for, perhaps even detrimental. In simple cases, trial-and-error experimentation will lead to a system design which will be intuitively satisfactory and probably nearly optimal. In complex

* In more conventional terms, $z(t)$ is called the output and $u(t)$ is the input. This usage is rather ambiguous and will be avoided in the sequel.

cases, this procedure becomes inefficient and sometimes impossible. One must rely on mathematical reasoning, and this requires greater precision of problem formulation. In the physical sciences, the dangers caused by sloppy application of mathematics can be checked by physical intuition. In the control systems field - which deals with man-made objects rather than observations of Nature - physical intuition is not always a reliable guide.

Many assumptions must be made to derive rationally a particular control law. Critique of assumptions is especially important when one starts to explore the concept of an adaptive control system. Such a system is characterized by the actual or desired independence of its control law from overly specific assumptions on the nature of the control object. To put it crudely, a control system is adaptive if it can perform well (perhaps after a short start-up period) without detailed prior knowledge of the dynamics of the control object. An adaptive system must be, therefore, capable of some form of learning.

Ideally, an adaptive controller should do just as good a job in controlling a supersonic airplane as in controlling a nylon factory, without being specifically designed for either job. The gap between such desiderata and the present state of technology is very great indeed. This has led us (and other research workers in the control systems field) to re-examine the bases of present knowledge in an attempt to see why the adaptive control problem seems so exquisitely difficult. And, of course, there is also a very real need for a better theory in order to evaluate and, if possible - understand and generalize numerous intuitive proposals now being made for practical adaptive systems.

In the initial phase of research on adaptive systems we have assumed that the equations of motion of the control object are known, and have been concerned primarily with the rigorous mathematical formulation and effective numerical solution of the control problem.

The most interesting problems in adaptive control arise when this assumption is relaxed; extensive preparation is necessary, however, before we can reach that stage. In essence, it is necessary to put conventional control theory in a clearer and more precise form - a process which will be seen to yield important results and suggest new problems even in conventional control theory.

2. Mathematical Models for Dynamical Systems.

Fundamental in the mathematical description of a dynamical system is the concept of state. This is simply a convenient way of expressing what might be loosely called

the Principle of Causality. For clarity, we formalize this idea as follows.

DEFINITION OF STATE. The state of a dynamical system is a minimal set of numbers which, specified at any given time, suffice to determine completely the future evolution of the system, provided the future forces acting on the system are known.

We are accustomed to represent physical dynamical systems by means of a system of n differential equations of the first order. The state of the system is then a (finite-dimensional) vector. The n real numbers constituting the state vector are the n initial conditions needed to uniquely specify the solution of the differential equations. An example of this sort is provided by particle mechanics: a system of N particles free to move in 3-dimensional space has a state vector of $6N$ components, made up of the $3N$ position and $3N$ velocity coordinates. In some cases, even the dimension of the state vector may be infinite, as in partial differential equations. In other cases, the number of states may be finite, as in models for digital computers.

By the equations of motion of a dynamical system we mean a rule which specifies how the state of the system at a given time is transformed into other states in the future. We shall also refer to this process as state transition. Usually the equations of motion are given in the small; that is to say, by differential equations which specify the infinitesimal state transition corresponding to the infinitesimal change $t \rightarrow t + dt$ in the time. By integrating these differential equations we obtain the equations of motion in the large; that is to say, we can specify the state transitions corresponding to arbitrary changes $t_0 \rightarrow t_1$ in the time.

As is common practice, we shall usually assume that the equations of motion are linear differential equations. Without some form of linearity, explicit mathematical treatment of the equations of motion is seldom possible. We emphasize, however, that the conceptual framework presented here remains valid also in the nonlinear case. In fact, with the present formulation of the dynamical problem the transition from the linear to the nonlinear is quite natural - which is not the case with other methods of linear analysis (laplace transform, frequency domain methods, etc.)

A sufficiently general mathematical model for linear dynamical systems is provided by the vector equations:

$$(2.1) \quad dx/dt = F(t)x + G(t)u(t) + J(t)w(t),$$

$$(2.2) \quad y(t) = H(t)x(t),$$

$$(2.3) \quad z(t) = u(t) + v(t),$$

where

x is an n -vector, the state of the system;

y is a p -vector, the output of the system;

z is a p -vector, the observed output of the system;

u is an m -vector, the control of the system;

v is a p -vector, representing the noise in the measurement of y ;

w is a q -vector, the random disturbances acting on the system;

We assume that F , G , H , J , which are arbitrary rectangular matrices, depend continuously on t .

In a purely schematic way, these definitions may be visualized with the aid of Figure 1.

The set of equations (2.1-3) includes most of the situations commonly encountered in engineering practice (see numerous examples of this in the sequel). A similar set of equations may be obtained also in the sampled-data case. Certain complications may arise, however, if continuous and pulsed elements occur in the same system. The setting up of equations then requires rather complicated "bookkeeping", for the details of which the reader may consult [Kalman and Bertram, 1959].

It will always be assumed that $v(t)$ and $w(t)$ are gaussian white-noise processes, i.e., their values occurring at different instants of time are independent gaussian random vectors. This can be done with virtually no loss of generality. We can represent the general gaussian random process as the output of a linear (possibly infinite-dimensional) dynamical system excited by white noise. (This is the content of the Loeve-Karhunen representation theorem [Loeve, 1961].) It is physically reasonable to approximate the resulting dynamical system with a finite-dimensional one. (This means that the power spectra of v and w are assumed to be rational.) The state variables associated in this way with the random processes $v(t)$ and $w(t)$ can be combined with the state variables of the system to be controlled. In other words, all problems in which the assumptions of linearity and gaussianity hold can be reduced -- with a change of variables -- to the standard form (2.1-3).

3. Adaptive Systems; Learning States.

So far the concept of an adaptive system has been discussed in rather vague terms. Certainly, there is no definition at present of an adaptive system which meets with general acceptance. We are, therefore, obliged to introduce our own, somewhat special, definition. This is done as a matter of convenience; we do not wish to claim that ours is the only reasonable point of view with regard to "adaptation".

DEFINITION OF AN ADAPTIVE CONTROL SYSTEM. A control system is adaptive if it is capable of changing its control law as a result of measured changes of the control object and its environment and in such a way as to operate at all times in an optimal or nearly optimal fashion.

A system with a fixed control law may operate quite adequately in a changing environment. Such a system may be more properly called insensitive or invariant, rather than adaptive. The word "adaptive" usually carries the connotation of an organism being able to take advantage of a new situation. Hence we do not regard a system as adaptive unless it is also optimal in some sense.

The operation of any adaptive control system will depend on two groups of data: (i) measured (or estimated) values of the state variables of the control object, which are used to determine the instantaneous values of the control variables; (ii) measured (or estimated) numbers defining the equations of the control object and its environment, which are used to determine the control law. The first group of numbers describes the momentary behavior of the control object; the second group refers to "structural" characteristics. For instance, the position, velocity, and angular momentum of a rigid body belong to the first group of data; the mass, moment of inertia, and internal constitution of the body belong to the second group.

A strict distinction between these two concepts is not always possible, of course. In specific cases this is unlikely to lead to confusion, however, since we are accustomed to identifying the second group - structural characteristics - with those properties of an object which are unchangeable or change slowly in time relative to the first group.

We shall call the first group of data the dynamic state, and introduce a special term for the second group of data.

DEFINITION OF LEARNING STATE. This is the "state of knowledge" -- expressed in mathematical form -- concerning all equations, statistical data, performance indices, etc., which are utilized in arriving at the function specifying the control law.

In other words, the learning state is a set of numbers representing all the quantitative information which an engineer would use in rationally arriving at an optimal control law. For instance, in case of the model (2.1-3), the learning state is the collection of numbers making up the matrices F, G, H, J , statistical information concerning the random processes $v(t)$ and $w(t)$, as well as the mathematical specification of the performance index which is to be minimized or maximized by optimal control.

As time passes, the "state of knowledge" is likely to deteriorate, unless further information becomes available from physical measurements. In an adaptive system, these measurements are utilized to update the "state of knowledge". The way in which the measurements of the structural characteristics are utilized determines the transition law of the learning states.

In short, there are two types of dynamic processes taking place in an adaptive control system: (i) the state variables of the control object are estimated from measurements and corresponding control action is taken in accordance with the control law existing at a given moment; (ii) the structural characteristics of the control object and its environment are monitored by another measurement process, and corresponding adjustments are made in the optimal control law from time to time.

The concept of the learning state introduced here is clearly evident also in [Bellman and Kalaba, 1959]. A schematic picture of an adaptive system is shown in Figure 2. We shall return later to the discussion of this figure.

4. Examples of Adaptation.

The following examples give what we feel is a reasonable interpretation of the notions of "structural characteristics" and "learning states".

Consider the problem of manipulating the control surfaces of an airplane or missile to produce lateral acceleration. The action of the control system will be influenced in the main by the following types of effects:

- (A) Random atmospheric disturbances of various sorts.
- (B) Loss of hydraulic fluid; aging of vacuum tubes; effects of temperature, moisture, radiation, etc. on electronic components.

- (C) Decreasing air density at higher altitudes which (i) decreases the effectiveness of the control surfaces; (ii) decreases aerodynamic drag; (iii) changes the statistical properties of windgusts; etc.

The classical theory of a control system [Truxal, 1955; Newton, et al., 1957] considers the problem of random disturbances as part of optimal linear design. We have emphasized this point by including random effects in the model (2.1-3). Thus only (B) and (C) are to be regarded as "structural changes".

One of the main reasons for the use of feedback is to counteract changes of type (B). Internal feedback is used to render control equipment largely insensitive to changes in the characteristics of electronic and other components. According to our view, emphasized in Section 3, this is not adaptation.

Changes of type (C) are usually the most drastic; they affect the very nature of the control object. In other words, changes of type (C) are not slow changes in the environment but substituting an entirely new environment.

5. Specific Forms of the Learning State.

The process of acquiring information about the structural characteristics of the control object and its environment may take numerous forms depending on the nature of the control problem. We mention briefly some of the problems which have been discussed in the literature [Aseltine, et al., 1958; Levin, 1958].

A large class of adaptive systems is concerned with learning the equations of motion of the control object. In some cases, this learning process may be quite simple in principle.

For instance, we might be able to express the lift and drag coefficients of an airplane in terms of the mass, Mach number, and altitude. These three quantities together would constitute the learning state; if they can be directly measured, the problem of adaptation would reduce to calculating lift and drag by preassigned formulae and utilizing the number so obtained to modify the control law. This type of adaptation can be only moderately effective since it is of the open-loop type; no effort is made to revise the original formulas giving lift and drag in terms of the measured quantities. More effective (closed-loop) adaptation could be obtained by fitting a model of the equations of motion of the airplane to the physical quantities measured during flight. If the data processing can be done fast enough, such a model would obviate dependence on prior aerodynamic data in maintaining a nearly optimal control law under a wide variety of flight conditions.

Another example showing the need for a more sophisticated learning process is the following. The bending modes of a ballistic missile change with the expenditure of fuel. If there is an accurate measurement of the loss of mass, one could, in principle, compute the shift in bending modes. In practice, measuring the loss of mass with sufficient accuracy would be very difficult, and one would rather attempt to measure the instantaneous bending modes. The latter are to be regarded then as constituting the learning state.

Different problems of adaptation arise when the problem is not to learn the equations of motion of the control object, but to estimate the random characteristics of the command signals which the control system is to follow. (This is sometimes called "input adaptation".) The problem might be, for instance, to estimate the correlation functions which are needed for the design of an optimal Wiener filter. The predictability of a random process depends on being able to represent it by a dynamical model, so that some "equation of motion" of the random process would constitute the learning state. Thus this problem is quite similar to the preceding one, though there may be many variations in the details.

Finally, let us mention the so-called performance-criterion sensing or extremum adaptive systems. Some overall performance index is measured experimentally and one attempts to find a control law (by trial and error) which optimizes this performance index. The learning states are here parameters describing the control law, and the learning process consists of the trial-and-error adjustment of these parameters.

Of course, the division of an adaptive controller into the two sets of states is quite arbitrary. It is difficult to conceive of a physical experiment which would always distinguish between the two types of states. The division is made as a matter of convenience in attempting to give a workable definition of adaptation, and is strongly motivated by scientific tradition. We are used to representing physical dynamical systems with linear or quasi-linear models.

A single representation is not likely to fit accurately many situations at the same time. We must have therefore a capability of changing the parameters of the representation, i.e., the equations of motion. A particular learning state corresponds to a particular equation of motion; the transition in the learning states corresponds to changing estimates of the instantaneous equations of motion of the control object. If the learning states can be changed rapidly and accurately, the adaptive controller will be able to handle nonlinear control objects: the learning states will represent the best instantaneous linear approximation to the nonlinear system.

6. References.

- J. A. ASELTINE, A. R. MANCINI, and C. W. SARTURE (1958) "A survey of adaptive control systems", Trans. IRE Prof. Group on Automatic Control, PGAC-6, December 1958, 102-108.
- R. BELLMAN and R. KALABA (1959) "A mathematical theory of adaptive control processes", Proc. Nat. Acad. Sci., USA, 45, 1288-1290.
- R. E. KALMAN and J. E. BERTRAM (1959) "A unified approach to the theory of sampling systems", J. Franklin Inst., 267 (1959) 405-436.
- M. J. LEVIN (1958) "Methods for the realization of adaptive systems", ISA Paper No. FCS 2-58.
- G. C. NEWTON, Jr., L. A. GOULD, and J. F. KAISER (1957) "Analytical design of feedback controls" (book), Wiley.
- J. G. TRUXAL (1955), "Automatic feedback control system synthesis (book)", McGraw-Hill.
- M. LOEVE (1961), "Probability Theory", Second Edition, Van Nostrand.

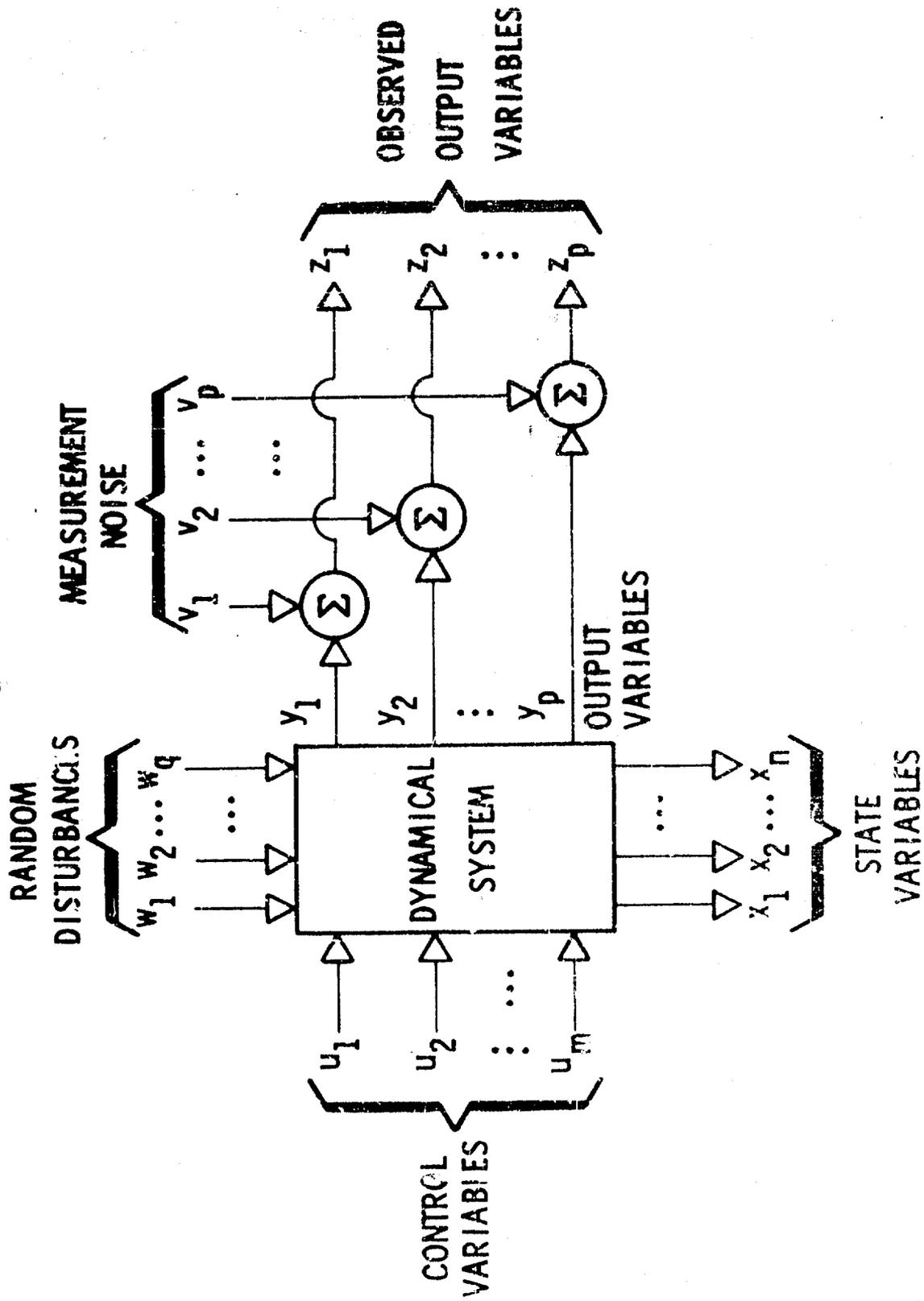


FIG. 1 SCHEMATIC REPRESENTATION OF A DYNAMICAL SYSTEM

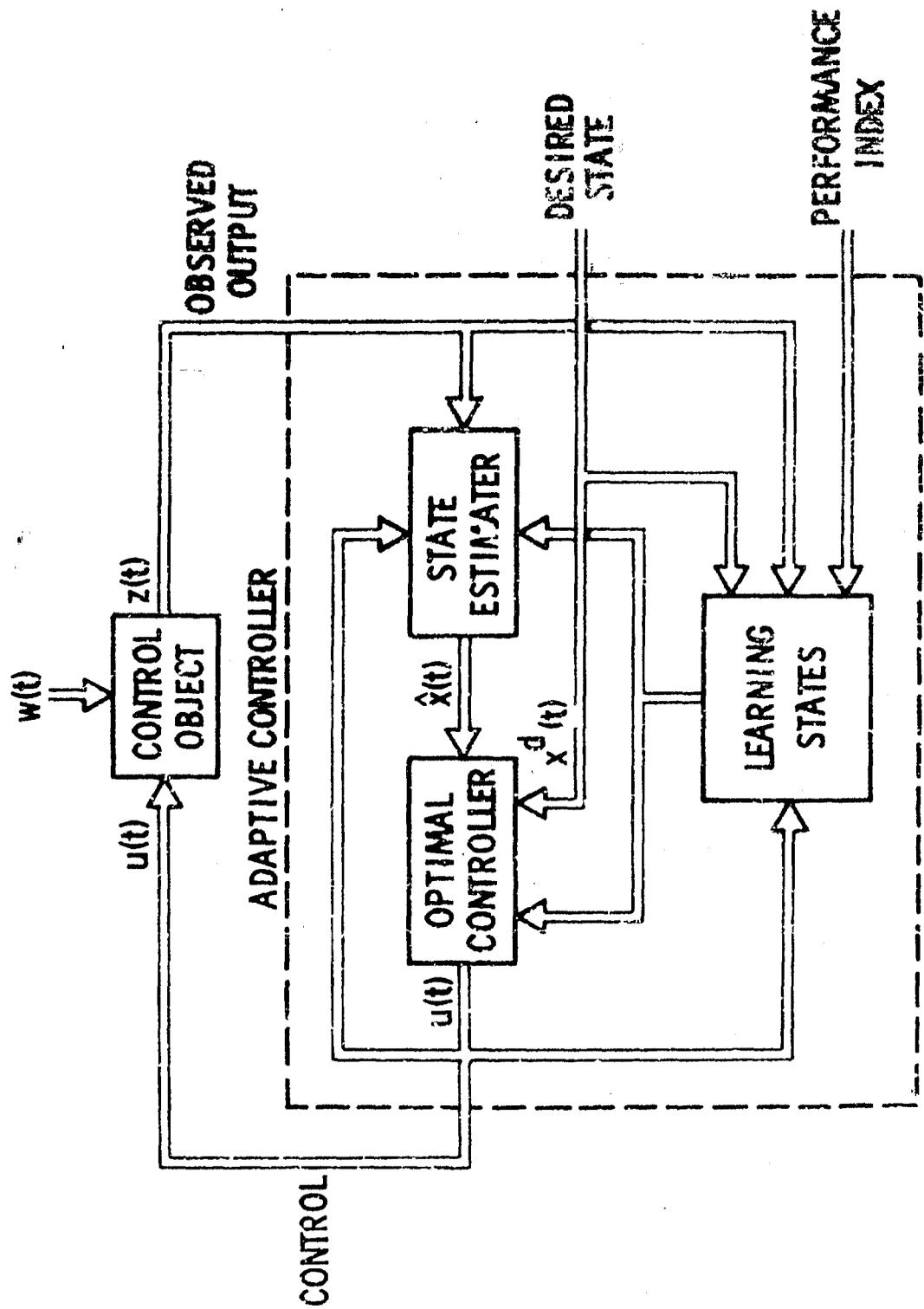


FIG. 2 SCHEMATIC REPRESENTATION OF ADAPTIVE CONTROL SYSTEM

Chapter 2.

THE NOISE-FREE REGULATOR PROBLEM

1. Assumptions and Notation.

For vector-matrix notation, see [Kalman-Bertram, 1960] or the Appendix.

It will be assumed that a sufficiently accurate model of the dynamical system to be controlled is provided by the linear equations discussed in Chapter 1, Section 2. In addition, it will be assumed throughout this chapter that noise effects are absent; in other words, $v(t)$ and $w(t)$ are identically zero. Thus we shall be concerned with the vector equations

$$(1.1) \quad dx/dt = F(t)x + G(t)u(t),$$

$$(1.2) \quad y(t) = H(t)x(t).$$

It is convenient to visualize this system by means of a vector block diagram shown in Figure 1A. This diagram is to be interpreted just as an ordinary block diagram, with two differences: (i) the fat lines used to denote the signal flow serve as a reminder that we are dealing with vector rather than scalar variables; (ii) the boxes denote linear transformations on the signals rather than multiplication by scalars.

In concrete terms the block diagram in Figure 1A is to be interpreted as follows. The box $1/s$ represents a set of n integrators. The output of the j -th integrator is fed back with the coefficient $f_{ij}(t)$ to the input of the i -th integrator. The j -th control variable $u_j(t)$ is fed forward with coefficient $g_{ij}(t)$ to the input of the i -th integrator. Finally, the i -th output $y_i(t)$ is a linear combination of the outputs of all the integrators, where the output of the j -th integrator appears with coefficient $h_{ij}(t)$. See Figure 1B.

We shall assume that $F(t)$, $G(t)$, $H(t)$, and $u(t)$ are piecewise continuous

functions of time. Then given a fixed control $u(t)$, (1.1) will have a unique solution. Aside from the time, this solution will depend also on (i) the initial state x_0 ; (ii) the initial time t_0 ; (iii) the control $u(t)$. It is often convenient to exhibit this dependence explicitly; we shall therefore write a solution of (1.1) in the form

$$\varphi_u(t; x_0, t_0).$$

This notation implies that

$$(1.3) \quad \varphi_u(t_0; x_0, t_0) = x_0$$

and

$$(1.4) \quad \frac{d\varphi_u(t; x_0, t_0)}{dt} = F(t)\varphi_u(t; x_0, t_0) + G(t)u(t).$$

The last equation is the definition of the solution of a differential equation. The equality needs to hold almost everywhere with respect to t ; more precisely, (1.4) may fail at points of discontinuity of $F(t)$, $G(t)$, or $u(t)$.

Instead of the cumbersome term "solution", we shall usually speak of $\varphi_u(t; x_0, t_0)$ as the motion of (1.1) passing through the point x_0 at the time t_0 under the influence of some fixed control $u(t)$.

It is well known in the theory of differential equations [Coddington and Levinson, 1955; Kalman and Bertram, 1960] that the motions of (1.1) can be expressed explicitly by means of the formula

$$(1.5) \quad x(t) = \varphi_u(t; x_0, t_0) = \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, \tau)G(\tau)u(\tau)d\tau,$$

which is valid for any x_0, t, t_0 (and not merely for $t \geq t_0$).

The matrix $\Phi(t, t_0)$ occurring in (1.5) is the transition matrix of (1.1) and is uniquely determined by the following requirements [Kalman and Bertram, 1960].

$$(1.6) \quad \Phi(t, t) = I = \text{unit matrix for all } t,$$

and

$$(1.7) \quad d\phi(t, t_0)/dt = F(t)\phi(t, t_0) \text{ for all } t \text{ and } t_0.$$

From these properties and the uniqueness of solutions of (1.1) one can show at once that

$$(1.8) \quad \phi^{-1}(t, t_0) = \phi(t_0, t) \text{ for all } t, t_0;$$

$$(1.9) \quad \phi(t_3, t_2)\phi(t_2, t_1) = \phi(t_3, t_1) \text{ for all } t_1, t_2, t_3.$$

CONVENTION. During the sequel we shall frequently omit explicit mention of arguments (such as time) if they are obviously implied by the context.

2. Quadratic Performance Indices.

We now define the regulator problem. Given that (1.1) is at some arbitrary state x at time t , we are to select a control $u(t)$ which drives the output of (1.1) to zero.

In general, there will be many control functions which accomplish this. To assign a numerical value to a particular control, we consider the following function of the control function, usually called a performance index:

$$(2.1) \quad 2V(x, t, T; u) = \|\phi_u(T; x, t)\|_S^2 + \int_t^T [\|H(\tau)\phi_u(\tau, x, t)\|_{Q(\tau)}^2 + \|u(\tau)\|_{R(\tau)}^2] d\tau,$$

where we use the special notation

$$(2.2) \quad \|x\|_Q^2 = \sum_{i,j=1}^n x_i q_{ij} x_j$$

for a quadratic form in x whose coefficients constitute the symmetric nonnegative

definite matrix Q . The scalar $\|x\|_Q$ can be regarded as the generalized euclidean distance from the origin. In (2.1) $Q(\tau)$ and $R(\tau)$ are positive definite and continuous in τ . The term $\|\phi_u(T; x, t)\|_S^2$ in (2.1) is the cost of the deviation of the final state of the dynamical system from the origin, measured with the aid of the distance function $\|x\|_S^2$. The terms $\|R(\tau)\phi_u(\tau; x, t)\|_{Q(\tau)}^2$ and $\|u(\tau)\|_{R(\tau)}^2$ in (2.1) represent costs per unit time of the deviation of the output of the dynamical system from the origin and the cost of the control action $u(\tau)$, respectively.

The terminal time T in (2.1) may be finite or infinite. In the latter case special precautions are necessary, as the integral (2.1) must be defined by a limiting process, letting $T \rightarrow \infty$.

This formulation of the regulator problem can be readily generalized to include the servomechanism problem. In that case we are given a certain desired output $y^d(\tau)$ which the system (1.1) is to follow as faithfully as possible. To include this requirement in the definition of V , we simply replace the first term in the integrand of (2.1) by

$$\|y^d(\tau) - H(\tau)\phi_u(\tau; x, t)\|_{Q(\tau)}^2 .$$

Further discussion of this problem will be postponed till a later chapter.

Evidently the performance index V in (2.1) is a quadratic function of the initial state x for any fixed $u(\tau)$. The reason for this assumption is that it leads to a linear control law.

Finally, it should be noted that the problem becomes meaningless if the cost of the control power $\|u(\tau)\|_{R(\tau)}^2$ is not included in the integrand in (2.1), for then V can be made arbitrarily small by using control variables of increasingly large amplitudes.

3. Statement of the Noise-Free Optimal Regulator Problem.

Given that the motion of (1.2) passes through the point x at time t , find a control $u^0(t)$ which minimizes the performance index V . The minimum value of V will depend only on x , t , and T and can be denoted by

$$(3.1) \quad V^0(x, t, T) = \min_u V(x, t, T; u)$$

A rigorous treatment of this problem may be found in [Kalman, 1961 A-B]. We shall now sketch the main features of this theory, omitting most proofs.

The optimal regulator problem is a special case of a general problem in theoretical physics or the calculus of the variations: that of minimizing the action, which is the integral of the lagrangian.* In the present case,

$$V^0(x, t, T) - V^0(x, T, T)$$

is the action and

$$(3.2) \quad L(x, u, t) = \frac{1}{2} [\|H(t)x\|_{Q(t)}^2 + \|u(t)\|_{R(t)}^2]$$

is the lagrangian.

It can be shown further [Kalman, 1961 A-B] that V^0 satisfies the hamilton-jacobi partial differential equation:

$$(3.3) \quad \frac{\partial V^0}{\partial t} + \mathcal{H}(x, V_x^0, t) = 0,$$

where V_x^0 is the gradient of V^0 with respect to x ; setting

$$(3.4) \quad p = V_x^0,$$

the hamiltonian \mathcal{H} is defined by

$$(3.5) \quad \mathcal{H}(x, p, t) = \min_u (L(x, u, t) + p'[F(t)x + G(t)u])^{**}$$

which leads to

$$(3.6) \quad 2\mathcal{H}(x, p, t) = \|H(t)x\|_{Q(t)}^2 + 2p'F(t)x - \|G'(t)p\|_{R^{-1}(t)}^2.$$

* In accordance with modern usage, adjectives and nouns formed from names of mathematicians who died before 1900 are not capitalized.

** The prime denotes the transpose of a matrix or of a (column) vector.

The minimization involved in (3.5) is known as Pontryagin's minimum principle [Pontryagin, 1957; Kalman, 1961 B]. This principle yields at once the optimal control law as a function of p , and t :

$$(3.7) \quad u^0(t) = -R^{-1}(t)G'(t)p .$$

Hence specification of the optimal control law reduces to finding p as a function of x and t , in other words, to the solution of the hamilton-jacobi partial differential equation (3.3).

Formula (3.7) is valid provided $R(t)$ is a nonsingular matrix. More generally, an optimal control law exists if the right-hand side of (3.5) has a minimum with respect to u - for which the nonsingularity of R is a sufficient, though not necessary, condition. These mathematical facts have an instructive physical interpretation, as was pointed out at the end of Section 2.

4. Solution of the Hamilton-Jacobi Equation.

Equation (3.3) may be solved by assuming that

$$(4.1) \quad V^0(x, t, T) = \frac{1}{2} \|x\|_{P(t)}^2 .$$

There is no loss of generality in assuming that P is symmetric. Since

$$(4.2) \quad V^0(x, T, T) = \frac{1}{2} \|x\|_S^2 ,$$

making use of the symmetry it follows that

$$(4.3) \quad P(T) = S .$$

From (3.4) and (4.1) we have,

$$(4.4) \quad p = P(t)x .$$

Substituting (4.1) into (3.3) gives

$$(4.5) \quad \frac{1}{2} x' \frac{dP}{dt} x + \frac{1}{2} x' H' Q H x + x' P F x - \frac{1}{2} x' P G R^{-1} G' P x = 0.$$

This must hold identically for all x ; hence (noting that the symmetrical part of PF is $\frac{1}{2}(F'P + PF)$) (4.5) simplifies to

$$(4.6) \quad - \frac{dP}{dt} = F'P + PF - PGR^{-1}G'P + H'QH,$$

which is the so-called matrix riccati equation. Hence we have arrived at the following result:

(4.7) A solution of the hamilton-jacobi partial differential equation corresponding to the lagrangian (3.2) and the hamiltonian (3.3) is given by the quadratic form (4.1), with time-varying coefficients governed by the matrix riccati equation (4.6). This solution must satisfy the boundary condition (4.2). The riccati equation is to be solved BACKWARDS in time, starting with $P(T) = S$.

It follows from (3.7) and (4.4) that

$$(4.8) \quad u^0(t) = -R^{-1}(t)G'(t)P(t)x(t),$$

which shows that the optimal control law is linear. Hence

(4.9) The optimal controller is a linear feedback system in which all state variables of the system must be known at all times to effect control.

The matrix

$$(4.10) \quad K(t) = R^{-1}(t)G'(t)P(t)$$

will be called the optimal gain. Figure 2 shows the vector matrix block diagram of the optimal control system.

The restriction that all state variables be known at all times will be removed in a later chapter

5. Application of the Theory in a Simple Case.

In order to give the reader a feeling for the details of the theory, we shall give a complete discussion of the first-order case. The model of the control object is taken as

$$(5.1) \quad \frac{dx_1}{dt} = f_{11}x_1 + g_{11}u_1,$$

$$y_1 = h_{11}x_1.$$

This means that the matrices defining (1.1) are given by

$$F = [f_{11}],$$

$$G = [g_{11}],$$

$$H = [h_{11}];$$

all these matrices are assumed to be constant, and $g_{11} \neq 0$, $h_{11} \neq 0$. See Fig. 3A.

The performance index is defined by

$$(5.2) \quad 2V^0(x_1, t) = \min_u \left(s_{11}x_1^2(T) + \int_t^T [q_{11}y_1^2(\tau) + r_{11}u_1^2(\tau)] d\tau \right),$$

and therefore

$$Q = [q_{11}],$$

$$R = [r_{11}],$$

$$S = [s_{11}];$$

all these matrices are also assumed to be constant. Moreover, $q_{11} > 0$, $r_{11} > 0$, $s_{11} \geq 0$.

The riccati equation (4.6) is now

$$(5.3) \quad -\frac{dp_{11}}{dt} = 2f_{11}p_{11} - \frac{s_{11}^2 p_{11}^2}{r_{11}} + h_{11}^2 q_{11}.$$

Since this is a first-order nonlinear differential equation, it is easy to discuss its behavior in qualitative terms. Equation (5.3) has two equilibrium states, \bar{p}_{11} and \tilde{p}_{11} , which are the roots of the quadratic resulting from setting $dp_{11}/dt = 0$:

$$s_{11}^2 \frac{\bar{p}_{11}}{r_{11}} = f_{11} + \sqrt{f_{11}^2 + s_{11}^2 h_{11}^2 \frac{q_{11}}{r_{11}}} > 0,$$

$$s_{11}^2 \frac{\tilde{p}_{11}}{r_{11}} = f_{11} - \sqrt{f_{11}^2 + s_{11}^2 h_{11}^2 \frac{q_{11}}{r_{11}}} < 0.$$

We then find that (see Figure 4)

$$-\frac{dp_{11}}{dt} < 0 \quad \text{if } p_{11} > \bar{p}_{11} \quad \text{or } p_{11} < \tilde{p}_{11},$$

$$-\frac{dp_{11}}{dt} > 0 \quad \text{if } \tilde{p}_{11} < p_{11} < \bar{p}_{11},$$

which shows that \bar{p}_{11} is a stable and \tilde{p}_{11} is an unstable equilibrium point of (5.3) as $t \rightarrow -\infty^*$.

The meaning of letting $t \rightarrow -\infty$ in (5.3) is the following. By the constancy of f_{11} , q_{11} , r_{11} , s_{11} the optimal performance index (5.2) is independent of the origin of time:

* (Remember that $dt < 0$ in this case and hence the inequalities are just opposite of the usual case where $dt > 0$).

$$V^0(x_1, t, T) = V^0(x_1, 0, T - t).$$

Hence $t \rightarrow -\infty$ is equivalent to $T \rightarrow \infty$.

Then the limiting solutions of the variance equation as $t \rightarrow -\infty$ correspond to infinite terminal time T .

Since $P_{11}(0) = s_{11} \geq 0$, it is clear that only the equilibrium state \bar{p}_{11} is of interest. We shall call \bar{p}_{11} the steady-state solution of the Riccati equation (5.3). Note that \bar{p}_{11} is independent of s_{11} as long as $s_{11} \geq 0$. The corresponding steady-state optimal gain is

$$(5.4) \quad \bar{k}_{11} = g_{11}, \quad \frac{\bar{p}_{11}}{r_{11}} = \frac{f_{11}}{g_{11}} + \sqrt{\left(\frac{f_{11}}{g_{11}}\right)^2 + h_{11}^2 \frac{q_{11}}{r_{11}}}$$

the equations of motion of the optimal controller are

$$(5.5) \quad \frac{dx_1}{dt} = (f_{11} - g_{11} \bar{k}_{11})x_1,$$

as shown in Figure 3B.

Recall now that $g_{11}^2 h_{11}^2 q_{11} > 0$. Then $f_{11} - g_{11} \bar{k}_{11} < 0$ and we see that in the steady state the optimal system is stable whether f_{11} is positive or negative, i.e., whether the uncontrolled system was stable or unstable. This is not trivial because the mere fact that a system is optimal does not imply that it is also stable!

Not only is the optimal system stable, but any degree of stability can be accomplished by suitable choice of the ratio q_{11}/r_{11} .

If $q_{11}/r_{11} \gg (f_{11})^2$, using (5.3) we can write, approximately

$$\left(\frac{\bar{p}_{11}}{r_{11}}\right)^2 \approx \left(\frac{q_{11}}{r_{11}}\right) \left(\frac{h_{11}}{g_{11}}\right)^2$$

$$\frac{\bar{p}_{11}}{h_{11}} \approx \frac{h_{11}}{r_{11}} \sqrt{\frac{r_{11}}{q_{11}}}$$

Regarding

$$V^0(x_1) = \frac{1}{2} \bar{p}_{11} x_1^2$$

as a Lyapunov function, we find that its derivative along motions of (5.1) is the integrand in (5.2):

$$\begin{aligned} \dot{V}^0(x_1) &= \bar{p}_{11} x_1 (f_{11} - g_{11} k_{11}) x_1 \\ &= -(q_{11} h_{11}^2 + r_{11} k_{11}^2) x_1^2. \end{aligned}$$

Using again the approximation $q_{11}/r_{11} \gg (f_{11})$, it follows that

$$k_{11} \approx h_{11} \sqrt{\frac{q_{11}}{r_{11}}}.$$

Hence

$$\dot{V}^0(x_1) \approx -2q_{11} h_{11}^2 x_1^2.$$

From the theory of Lyapunov [Kalman-Bertram, 1960, p. 386] it follows that the "time constant" of any dynamic system (linear or nonlinear) is bounded by

$$\tau_0 = 2 \max_{x_1} \left[\frac{V^0(x_1)}{-\dot{V}^0(x_1)} \right] \approx \frac{\bar{p}_{11}}{2q_{11} h_{11}^2}.$$

* That is, $\tau \dot{V} < V$. Regarding V as a measure of the distance of the state from the origin, this leads to the estimate $V(t) \leq e^{-t/\tau} V(0)$ of the transient response.

Letting $\rho = q_{11}/r_{11}$, we can summarize these results as:

$$(5.6) \quad \begin{cases} \tau_{\max} = \frac{1}{2g_{11}h_{11}\sqrt{\rho}} \\ K_{11} = h_{11}\sqrt{\rho} \end{cases}$$

Hence the time constant of the optimal system can be made arbitrarily small by letting ρ be large, but this is always accomplished by increasing the gain K_{11} and hence the amplitude of the control signal $u(t)$.

The question now arises as for what values of $T-t$ (5.3) can be regarded as having practically reached its steady-state value. In other words, on what depends the time constant of (5.3)?

Let

$$\delta p_{11} = p_{11} - \bar{p}_{11}$$

and consider the Lyapunov function

$$V(\delta p_{11}) = \frac{1}{2}(\delta p_{11})^2.$$

Hence

$$\begin{aligned} \dot{V}(\delta p_{11}) &= \frac{\partial V}{\partial \delta p_{11}} \frac{d\delta p_{11}}{dt} = (2f_{11}p_{11} - \frac{g_{11}^2 p_{11}^2}{r_{11}} + h_{11}^2 q_{11}) \delta p_{11} \\ &= (2f_{11} - g_{11}^2 \frac{p_{11} + \bar{p}_{11}}{r_{11}}) \delta p_{11} \\ &= - (g_{11}^2 \frac{p_{11}}{r_{11}} + h_{11}^2 \frac{q_{11}}{\bar{p}_{11}}) \delta p_{11} \end{aligned}$$

Hence the maximum time constant of the riccati equation is

$$(5.7) \quad \tau_1 \approx 2 \left[g_{11}^2 \frac{p_{11}}{r_{11}} + h_{11}^2 \frac{q_{11}}{\bar{p}_{11}} \right]^{-1} = 2 \left[g_{11}^2 \frac{p_{11}}{r_{11}} + \frac{1}{2\tau_0} \right]^{-1}.$$

Evidently the more stable is the optimal control system, the "faster" does the solution of the riccati equation approach its limiting value.

To summarize, we have found that:

- (i) if $p_{11}(T) \geq 0$, all solutions of the riccati equation tend to \bar{p}_{11} as $t \rightarrow -\infty$;
- (ii) \bar{p}_{11} is the solution of the optimal regulator problem when $T = \infty$;
- (iii) if $g_{11}^2 h_{11}^2 q_{11} > 0$, the optimal system is always stable;
- (iv) by making the ratio q_{11}/r_{11} large, any desired degree of stability can be obtained;
- (v) the time constant of the riccati equation is directly related to the time-constant of the optimal filter.

The main aim of the theory of the optimal regulator problem is to extend the results to systems of higher order and to systems with time-varying coefficients. This requires fairly complex matrix analysis, and will be discussed later. For further details, consult [Kalman, 1961 A-C].

6. Existence of Solutions of the Optimal Regulator Problem.

The main result here is expressed by the following theorem, which is proved in [Kalman, 1961 A].

(6.1) The noise-free optimal regulator problem has a solution for every finite $T - t$ if the matrix $R(\tau)$ and the matrix $Q(\tau)$ are positive definite for all τ in the interval (t, T) , while S is nonnegative definite.

In order to understand this result, a simple counter-example will be considered in detail. Define a performance index by

$$(6.2) \quad 2V^0(x_1, t, T) = \max_u (s_{11} x_1^2(T) + \int_t^T [y_1^2(\tau) - u_1^2(\tau)] d\tau)$$

while

$$(6.3) \quad \frac{dx_1(t)}{dt} = \alpha x_1(t) + u_1(t) \quad (\alpha = \text{real}).$$

In other words,

$$F = (\alpha), G = (1), q_{11} = 1, h_{11} = 1, r_{11} = -1.$$

The corresponding equation has a solution

$$2V^0(x_1, t, T) = p_{11}(t)x_1^2,$$

where

$$p_{11}(T) = s_{11}$$

and

$$(6.4) \quad -\frac{dp_{11}(t)}{dt} = 1 + 2\alpha p_{11}(t) + p_{11}^2(t).$$

Integrating (6.4) by separation of variables, we get, setting $s_{11} = 0$,

$$\begin{aligned}
 (6.5) \quad & \left. \begin{aligned}
 p_{11}(t) &= \frac{\frac{1}{\sqrt{1-\alpha^2}} \tan \sqrt{1-\alpha^2} (T-t)}{1 - \frac{\alpha}{\sqrt{1-\alpha^2}} \tan \sqrt{1-\alpha^2} (T-t)} && \text{if } \alpha^2 < 1, \\
 p_{11}(t) &= \frac{\frac{1}{\sqrt{\alpha^2-1}} \tanh \sqrt{\alpha^2-1} (T-t)}{1 - \frac{\alpha}{\sqrt{\alpha^2-1}} \tanh \sqrt{\alpha^2-1} (T-t)} && \text{if } \alpha^2 > 1, \\
 p_{11}(t) &= \frac{T-t}{1+(T-t)} && \text{if } \alpha = -1, \\
 p_{11}(t) &= \frac{T-t}{1-(T-t)} && \text{if } \alpha = +1.
 \end{aligned} \right\}
 \end{aligned}$$

From (6.5) it can be observed that if $\alpha > -1$ the solution $p_{11}(t)$ has a finite escape time and the maximization problem is MEANINGLESS for $T-t > t_e$ where

$$(6.6) \quad t_e = \frac{1}{\sqrt{1-\alpha^2}} \tan^{-1} \sqrt{\frac{1-\alpha^2}{\alpha}}$$

For $\alpha \leq -1$ a solution exists in the steady state, that is, $T \rightarrow +\infty$ if $s_{11} = 0$. However, if $s_{11} \neq 0$ and is a sufficiently large positive number, even when $\alpha = -1$ there is no steady-state solution. This phenomenon is shown clearly by the state-space (1-dimensional) plot of the differential equation (6.4). See Figure 5, where the arrows indicate the direction of motion as $t \rightarrow -\infty$. Indeed if

$a_{11} > -\alpha + \sqrt{\alpha^2 - 1}$, then no steady state solution exists. These cases illustrate the problems which one may encounter by neglecting existence questions.

The classical approach via the Euler equations would not reveal the fact that the optimization problem becomes meaningless for large $T - t$; the Euler equations in the present case always have a unique solution.

7. Existence of the Solution of the Steady-State Optimal Regulator Problem.

The optimal regulator problem makes sense in the steady-state ($T = \infty$) only if the limit

$$(7.1) \quad V(x, t, \infty; u) = \lim_{T \rightarrow \infty} V(x, t, T; u)$$

exists and is finite for some control function $u(\tau)$ defined for $\tau \in (t, \infty)$.

In order to investigate this situation, we introduce a new concept:

DEFINITION OF COMPLETE CONTROLLABILITY. A system (1.1) is said to be completely controllable if at any initial time t any initial state x can be taken to the origin in a finite length of time by the application of a suitable control function.

The abstract definition of complete controllability is equivalent to the following concrete condition (for proof, see [Kalman, Ho, Narendra, 1962]):

(7.2) THEOREM. A system (1.1) is completely controllable if and only if the matrix

$$(7.3) \quad W(t, T) = \int_t^T \Phi(t, \tau) G(\tau) R^{-1}(\tau) G'(\tau) \Phi'(t, \tau) d\tau$$

is positive definite for some $T > t$.

The matrix W has the following interpretation: The minimum energy required to transfer the state x at time t to the origin at time T is

$$\int_t^T \|u(\tau)\|_{R(\tau)}^2 d\tau = \|x\|_{W^{-1}(t, T)}^2.$$

The computation of the matrix W is most easily performed by observing that it, too, is governed by the riccati equation. To see this, we note two facts: (i) If a matrix P is governed by a riccati equation, then its inverse P^{-1} (if it exists) is also governed by a riccati equation. (ii) By the remark in the preceding paragraph, $\|x\|_{W^{-1}(t, T)}^2$ is the performance index for a special optimization problem: take x at time t to the origin at time T , minimizing along the motion the control energy.

Differentiating (7.3) with respect to t , using (1.7-9), we find

$$(7.4) \quad dW/dt = F(t)W + WF'(t) - G(t)R^{-1}(t)G'(t)$$

which is a special case of the riccati equation (4.6). In practical cases, W is usually computed by means of this equation rather than by numerical integration of (7.3).

If the system (1.1) is constant (or stationary) that is to say, if F , and G are constants, then complete controllability can be checked more simply (for proof, see [Kalman, Ho, Narendra, 1961]2):

(7.5) THEOREM. For a constant system (1.1) a necessary and sufficient condition for complete controllability is

$$(7.6) \quad \text{rank } [G, FG, \dots, F^{n-1}G] = n.$$

The condition of complete controllability is not necessary for the existence of the limit (7.1). But if a system is not completely controllable, its state variables can be decomposed into two groups, one of which is completely unaffected by control. See Figure 6. If the part of the system which is not coupled to u is asymptotically stable, then the limit (7.1) exists; but if this part is unstable, the limit will not exist.

It can be shown that if a linear constant system is described by a transfer function, then it is always completely controllable. This is due to the fact that in writing down the transfer function terms which account for lack of complete controllability cancel out of the numerator and denominator.

Because a single-input/single-output control object described by a transfer function is always completely controllable, the importance of controllability was unnoticed for a long time in the literature of control engineering. In simple cases, lack of controllability is easily detected and eliminated by physical considerations. On the other hand, in complicated cases when the equations of motion are written in the normal form (1.1) and there are several inputs and outputs, controllability is not obvious and efficient mathematical means must be devised to test this property of the system. This is the price one has to pay for a more general theory.

If we do not have complete controllability, the limit (7.1) will not exist in general. This is easily seen by the following example

$$dx_1/dt = f_{11}x_1 + u_1(t),$$

$$dx_2/dt = x_3,$$

$$dx_3/dt = -x_2,$$

$$y_1 = x_1,$$

$$Q = [1],$$

$$R = [1],$$

$$S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then V will always contain the term

$$\|x(T)\|_S^2 = [x_1(t) \cos T + x_3(t) \sin T]^2$$

which cannot be affected by $u_1(t)$ and clearly does not have a limit as $T \rightarrow \infty$.

The most important consequence of complete controllability is the following:

(7.7) THEOREM. If a system (1.1) is constant, completely controllable, and $S = 0$, then the limit (7.1) always exists.

By complete controllability, $V^0(x, t, T)$ may be bounded from above, independent of T . The theorem then follows immediately since $V^0(x, t, T)$ is nondecreasing as $T \rightarrow \infty$ and a bounded, monotone sequence always converges. It is an open question at present whether this theorem holds also when $S \neq 0$ (because then $V^0(x, t, T)$ is not necessarily monotone increasing T .)

Slight further arguments prove also

(7.8) THEOREM. Let $P(t; 0, T)$ be a solution of the riccati equation corresponding to $P(T; 0, T) = 0$. Then

$$(7.9) \quad P^* = \lim_{T \rightarrow \infty} P(t; 0, T)$$

always exists, and P^* is an equilibrium state of the riccati equation, i.e., $dP/dt = 0$ when $P = P^*$.

See [Kalman, 1961A].

We note also that, whenever the limit (7.1) exists, the following is true:

$$(7.10) \quad \min_u \{ \lim_{T \rightarrow \infty} V(x, t, T; u) \} = \lim_{T \rightarrow \infty} [\min_u V(x, t, T; u)] \\ = \lim_{T \rightarrow \infty} V^0(x, t, T).$$

The left-hand side is the definition of optimal control for infinite terminal time $T = \infty$. The right-hand side shows that the min and lim operations may be interchanged; in other words, optimal control when $T = \infty$ can be obtained as the limit of optimal controls as $T \rightarrow \infty$.

The proof of (7.10) is almost immediate, appealing to the definition of optimality.

Finally, let us observe that while complete controllability guarantees the existence of the limit (7.9) when $S = 0$, it may happen that for other values of S there will be a different limit.

The equilibrium states of the riccati equation are obtained by setting $dP/dt = 0$. We show that it is possible to have under complete controllability more than one

equilibrium state. Consider the system

$$(7.11) \quad \begin{aligned} \dot{x}_1 &= x_1 + u_1, \\ \dot{x}_2 &= x_2 + u_2 \\ y_1 &= x_1 + x_2. \end{aligned}$$

Here

$$F = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad H = [1 \quad 1].$$

By (7.5), the system is clearly completely controllable.

We let

$$Q = [1], \quad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then the riccati equation is

$$(7.12) \quad \begin{aligned} -dp_{11}/dt &= 2p_{11} - p_{11}^2 - p_{12}^2 + 1, \\ -dp_{12}/dt &= 2p_{12} - p_{12}(p_{11} + p_{22}) + 1, \\ -dp_{22}/dt &= 2p_{22} - p_{12}^2 - p_{22}^2 + 1. \end{aligned}$$

It can be proved [Kalman, 1961C, Example 14.20] that on setting the left-hand sides in (7.12) equal to zero, the resulting set of quadratic equations has precisely two nonnegative definite solutions:

$$P(1) = \frac{1}{2} \begin{bmatrix} 3 + \sqrt{3} & \sqrt{3} - 1 \\ \sqrt{3} - 1 & 3 + \sqrt{3} \end{bmatrix}$$

and

$$\bar{P}(2) = \frac{1}{2} \begin{bmatrix} 1 + \sqrt{3} & 1 + \sqrt{3} \\ 1 + \sqrt{3} & 1 + \sqrt{3} \end{bmatrix}$$

$\bar{P}(1)$ is nonsingular, while $\bar{P}(2)$ is singular.

8. Uniqueness of the Solution of the Steady-State Regulator Problem.

To prove that the steady-state control law is unique, i.e., independent of δ , we need a new concept, which may be regarded as the dual of controllability.

DEFINITION OF COMPLETE OBSERVABILITY. The system (1.12) is said to be completely observable if it is possible to determine the exact value of $x(t_0)$ given the values of $y(t)$ in a finite interval (t_{-1}, t_0) preceding t_0 .

The analog of Theorem (7.2) is proved in [Kalman, 1961C, Lemma (15.7)] and may be stated as follows:

(8.1) **THEOREM.** A system (1.1-2) is completely observable if and only if the matrix

$$(8.2) \quad M(t_0, T) = \int_{t_0}^T \phi'(t, T) H'(t) Q(t) H(t) \phi(t, T) dt$$

is positive definite for some $T > t_0$.

The analog of Theorem (7.5) is:

(8.3) **THEOREM.** For a constant system (1.1-2) a necessary and sufficient condition for complete observability is

$$\text{rank } [H', F'H', \dots, F'^{n-1}H'] = n.$$

According to this criterion the system (7.11) is not completely observable. Thus the example at the end of Section 7 shows that in the absence of complete observability we cannot expect in general to have a unique optimal control law in the steady-state.

The main result of this chapter may be stated as follows:

(8.4) THEOREM. Consider a constant system (1.1-2), i.e., F, G, H, Q, R are constant matrices. Assume that the system is completely controllable and completely observable. Then:

(i) The solution of the riccati equation starting at any nonnegative definite matrix S converges exponentially to a unique, positive definite matrix P as $t \rightarrow -\infty$ (or $T \rightarrow \infty$).

(ii) The optimal control law for $T = \infty$ is constant and the optimal regulator is asymptotically stable.

This theorem can be generalized in a natural and straightforward way also to nonconstant (time-varying) systems. The precise statement of the results is more complicated. For these statements and the proofs the reader is referred to [Kalman, 1961A and 1961C].

The fact that under conditions of complete controllability and complete observability the optimal system is stable is not a triviality since the formulation of the optimization problem in Section 2 did not include this requirement. Nor does stability of the optimal system follow in general. For instance, if we take the matrix $\bar{P}^{(2)}$ of Section 7, we find the corresponding infinitesimal transition matrix of the steady-state optimal system is

$$F^0 = F - G\bar{K} = \frac{1}{2} \begin{bmatrix} 1 - \sqrt{3} & -1 - \sqrt{3} \\ -1 - \sqrt{3} & 1 - \sqrt{3} \end{bmatrix}$$

whose eigenvalues are

$$\lambda_1 = 2\sqrt{3}; \quad \lambda_2 = -2.$$

Thus the optimal system for $T = \infty$ is unstable if we choose

$$S = \bar{P}^{(2)}.$$

But the deeper significance of Theorem (8.4) lies in the result that every solution of the riccati equation starting at a nonnegative definite initial value converges to \bar{P} ; moreover, convergence is exponential. This means that the riccati equation provides a feasible computation procedure for obtaining the optimal system which is not likely to be affected by roundoff errors. Note that, according to the theorem, one could have obtained \bar{P} by setting the left-hand side of (4.6) equal to zero and solving the resulting set of simultaneous quadratic algebraic equations in the elements of P . This procedure can indeed be carried out in simple cases [Kalman, 1961C, Sect. 14] but when the order of the system becomes larger than 2, the approach via the riccati equation is likely to be appreciably more efficient.

9. Some Important Inequalities.

From the point of view of practical numerical computation it is of course by no means enough to know that the solution of the riccati equation converges exponentially, one must have also an estimate of the speed of convergence.

This aspect of the theory is not yet in a definitive form. We shall confine ourselves therefore to the statement of the major results to date. Proofs may be found in [Kalman, 1961A and 1961C].

If A, B are symmetric matrices, let us use the notation $A > B$ [$A \geq B$] to signify the fact that $A - B$ is positive [nonnegative] definite.

We assume (1.1-2) is constant, completely controllable and completely observable.

The desired inequalities are then as follows:

$$(9.1) \quad 0 < P(t) \leq W^{-1}(t_0, t) + M(t_0, t),$$

$$(9.2) \quad P^{-1}(t) \leq M^{-1}(t_0, t) + W(t_0, t), \quad t > t_0$$

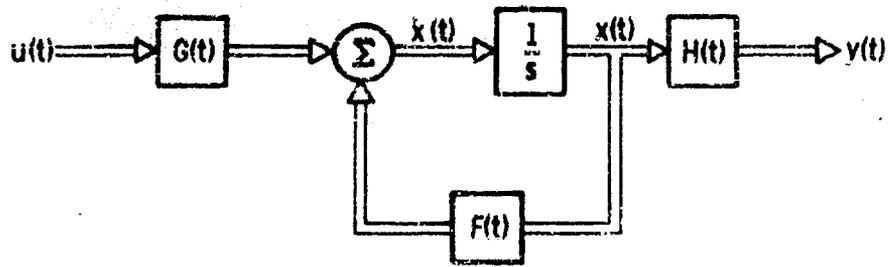
$$(9.3) \quad P(t) - P(t_0) \geq \frac{\lambda_{\min}(M^2(t_0, t_0))}{4 \operatorname{tr} M^2(t_0, t) \operatorname{tr} W(t_0, t)} I$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a symmetric matrix A .

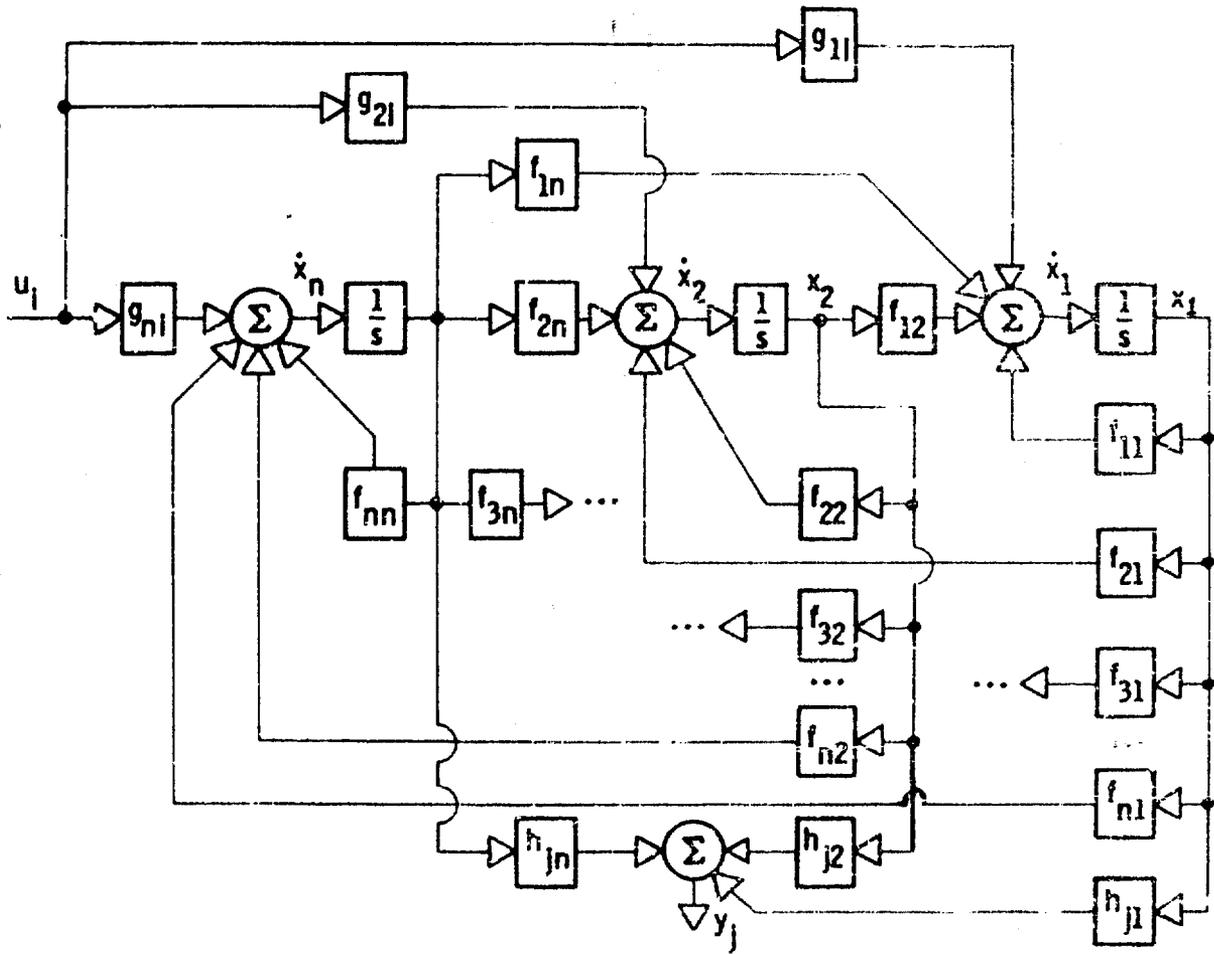
These inequalities are useful in guiding the choice of numerical values of Q and R .

10. References.

- E. A. CODDINGTON and N. LEVINSON (1955) "Theory of ordinary differential equations (book)", McGraw-Hill, 1955.
- R. E. KALMAN (1961A) "Contributions to the theory of optimal control", Proc. Symp. on Ordinary Differential Equations, Mexico City, 1959; to appear in Bol. Soc. Mat. Mexicanas.
- R. E. KALMAN (1961B) "Variational problems in system theory", Proc. RAND Corp. Symposium on Mathematical Optimization, 1960.
- R. E. KALMAN (1961C) "New methods and results in linear filtering and prediction theory", Proc. Symp. on Engineering Applications of Probability and Random Functions, Purdue University, Nov. 1960; to be published by Wiley.
- R. E. KALMAN and J. E. Bertram (1960) "Control system analysis and design via the second method of Lyapunov", J. Basic Engr. (Trans. ASME), 82, 371-400.
- R. E. KALMAN, Y. C. Ho, and K. S. NARENDRA (1962) "Controllability of linear dynamical systems", to appear in Contributions to Differential Equations.
- L. S. PONTRYAGIN (1957) "Optimal control processes (in Russian)", Uspekhi Mat. Nauk, 14, 3-20.



(A) MATRIX BLOCK DIAGRAM OF LINEAR DYNAMICAL SYSTEM



(B) CONVENTIONAL BLOCK DIAGRAM OF LINEAR DYNAMICAL SYSTEM

FIG. 1 LINEAR DYNAMICAL SYSTEM

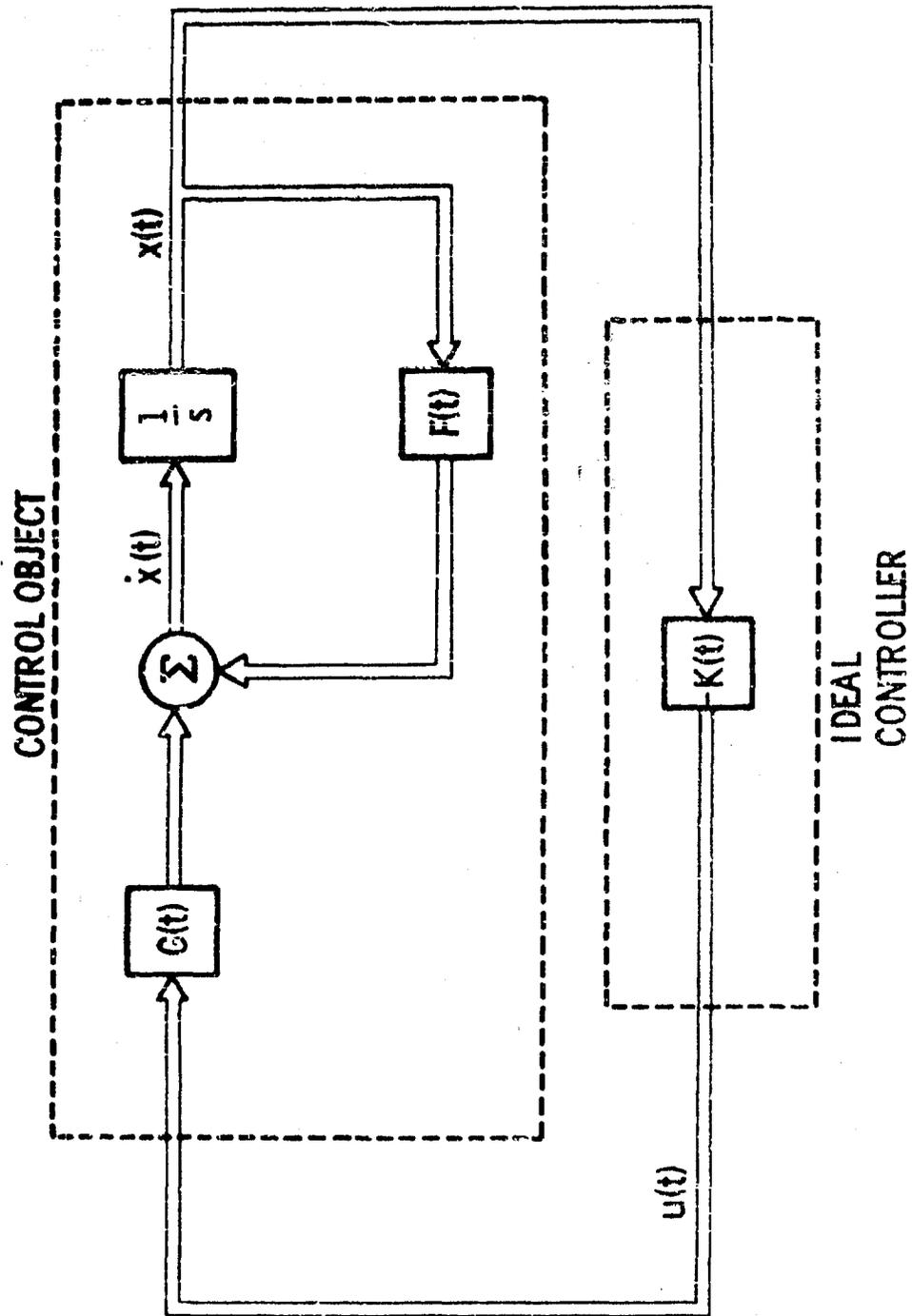


FIG. 2 OPTIMAL NOISE-FREE REGULATOR

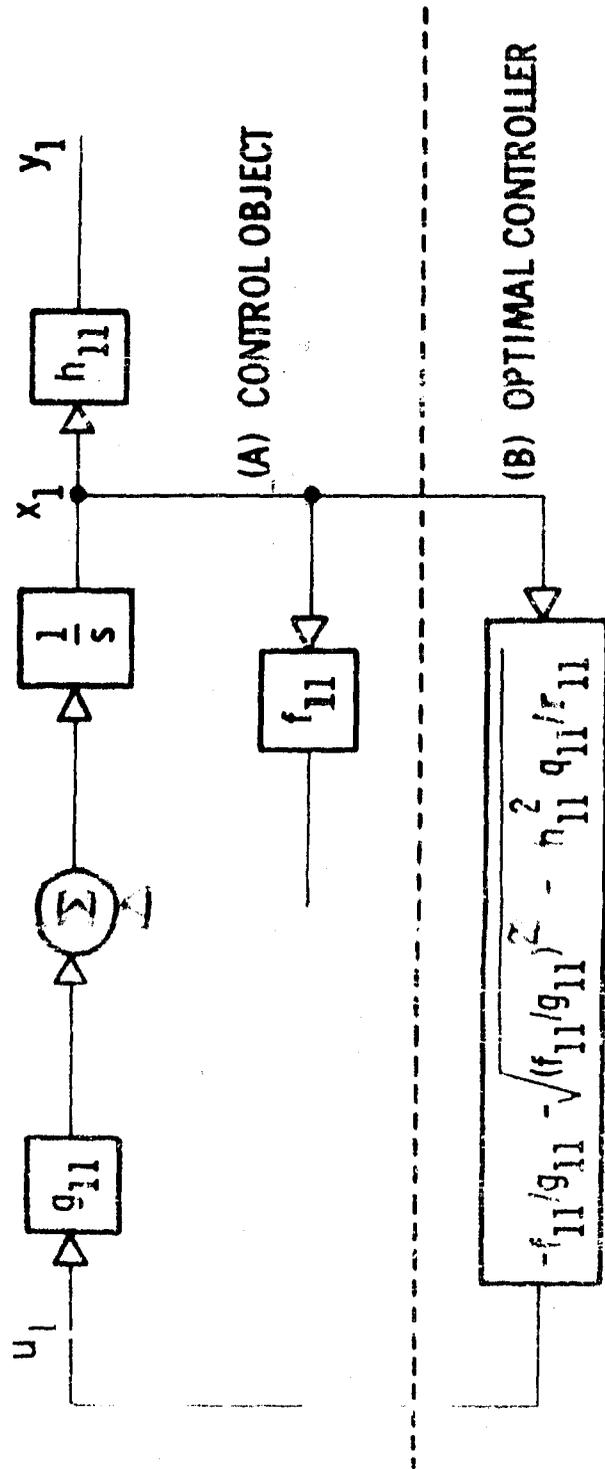


FIG. 3 THE FIRST-ORDER OPTIMAL CONTROL SYSTEM

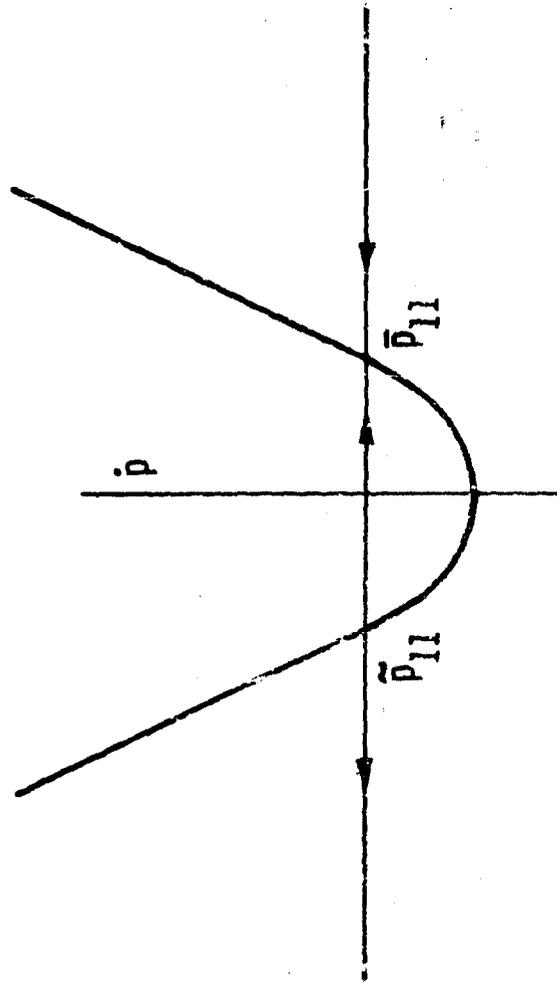


FIG. 4 STABILITY OF THE RICCATI EQUATION

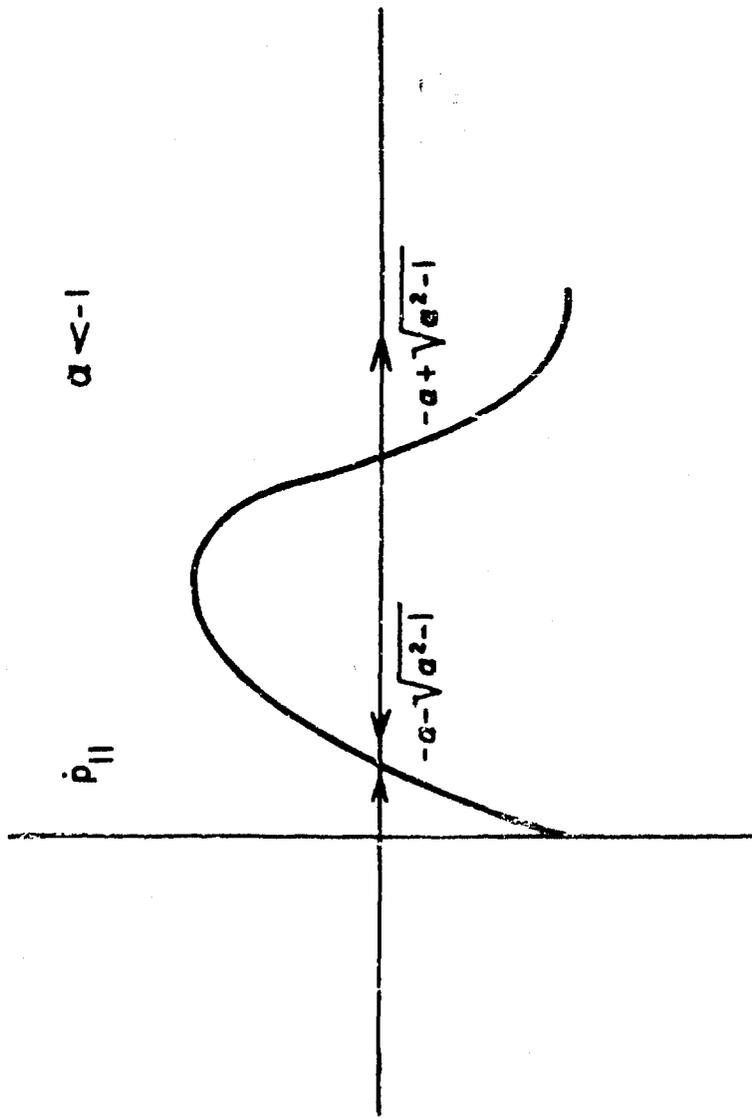


FIG. 5. STATE-SPACE OF A RICCATI EQUATION

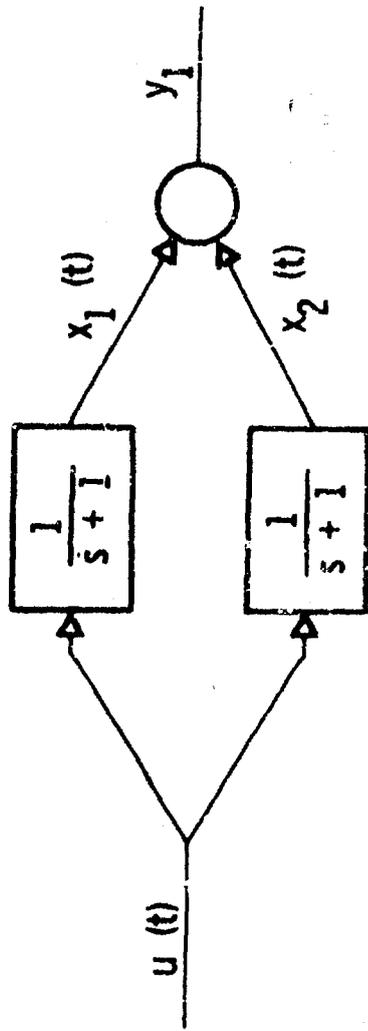


FIG. 6. AN UNCONTROLLABLE SYSTEM

Chapter 3.

A THIRD-ORDER OPTIMAL REGULATOR PROBLEM

1. Introduction.

We shall discuss in this chapter the noise-free optimal regulator problem in a special case when the control object is of the third order ($n = 3$). On the one hand, this example is simple enough to be treatable in part by analytic methods; on the other hand, the example is complicated enough to illustrate various problems encountered in numerical computation.

2. Definition of Control Object; Transition Matrix.

The defining matrices in equation (1.1) of Chapter 2 are taken as

$$(2.1) \quad F = \begin{bmatrix} 0 & 100 & 0 \\ -100 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

The matrix H will remain undefined for the moment.

The (conventional) block diagram of the control object is shown in Figure 1. Note that the element values of F and G may be read off by inspection from the figure.

We now compute the transition matrix corresponding to F given by (2.1). For simplicity we write $\Phi(t) = \Phi(t, 0)$. By (1.7), of Chapter 2 $\Phi(t)$ satisfies the relations

$$(2.2) \quad d\Phi(t)/dt = F\Phi(t), \quad \Phi(0) = I.$$

Taking Laplace transforms on both sides, we have, just as in the scalar case,

(2.3)

$$(sI - F)\phi(s) = I.$$

One can compute $\phi(s)$ by solving (2.3). This is quite complicated, however. A simpler method is this:

We observe that by (2.2) $\phi_{1,j}(s)$ is the transfer function from the input to the j -th integrator to the output of the i -th integrator in Figure 1. Utilizing Mason's loop rule [Mason, 1956], we can easily calculate these transfer functions and obtain

$$\phi(s) = \frac{1}{s(s^2 + 10,001)} \begin{bmatrix} s^2 + 1 & 100s & 100 \\ -100s & s^2 & s \\ 100 & -s & s^2 + 10,000 \end{bmatrix},$$

which checks with (2.3). Taking the inverse Laplace transform of each element of $\phi(s)$, we finally get

$$\phi(t) = \frac{1}{\omega^2} \begin{bmatrix} 1 + (\omega^2 - 1)\cos \omega t & 100 \omega \sin \omega t & 100(1 - \cos \omega t) \\ -100 \omega \sin \omega t & \omega^2 \cos \omega t & \omega \sin \omega t \\ 100(1 - \cos \omega t) & -\omega \sin \omega t & 10^4 + \cos \omega t \end{bmatrix}$$

where $\omega^2 = 10,001$.

3. Controllability.

We can use Theorem (7.5) of Chapter 2 to check whether the system (2.1) is completely controllable. The answer is in the affirmative, for the matrix

$$[G, FG, F^2G] = \begin{bmatrix} 1 & 0 & 10,000 \\ 0 & -100 & 0 \\ 0 & 0 & -100 \end{bmatrix}$$

has rank 3.

This result is merely qualitative. To get a quantitative answer as to how effectively control can be applied, we must compute the controllability matrix W given by (7.3) of Chapter 2, and find its inverse (which always exists by complete controllability) to see how much energy is needed to take the various states to zero.

We could compute W by direct integration, since the integrands in (7.3) of Chapter 2 would involve only simple trigonometric functions. But this task is exceedingly tedious. Calculating crudely, we see that if $T - t$ is several times larger than the period ($2\pi/\omega = 6 \times 10^{-2}$) of $\phi(t)$, then the amount of energy required to take the states

$$x^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{or} \quad x^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

to zero is about 5×10^3 times smaller than the energy required to take

$$x^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

to zero.

In practice, the matrix W is obtained by computing the solution of the differential equation (7.4) of Chapter 2 by means of the methods discussed in Chapter 11. Taking $R = [1]$, we get the matrix

$$W(0, 1) = \begin{bmatrix} 4.9765 & 0.0126 & -0.0503 \\ 0.0126 & 5.0201 & 0.0000 \\ -0.0503 & 0.0000 & 0.0015 \end{bmatrix} \times 10^{-1}$$

whose inverse is

$$W^{-1}(0, 1) = \begin{bmatrix} 0.0003 & 0.0000 & 0.0101 \\ 0.0000 & 0.0002 & 0.0000 \\ 0.0101 & 0.0000 & 1.0001 \end{bmatrix} \times 10^4.$$

The numerical results thus confirm the earlier qualitative conclusions.

4. First Attempt at Design.

We assume that the primary objective of control is to assure that x_1 is small at all times. Therefore we set

$$(4.1) \quad H = [1 \quad 0 \quad 0].$$

It is a good idea to check immediately whether with this choice of H the system is completely observable. In view of Theorem (8.3) of Chapter 2, complete observability is equivalent to

$$\det[H', F'H', F'^2H'] = \det \begin{bmatrix} 1 & 0 & 10,000 \\ 0 & -100 & 0 \\ 0 & 0 & 100 \end{bmatrix} \neq 0.$$

Hence H given by (4.1) assures complete observability.

Guided by the analysis of Section 5, Chapter 2, we now wish to choose the ratio Q/R large in order to assure an adequate degree of stability (in the present case both Q and R are 1×1 matrices, i.e., scalars). Suppose we let $Q/R = 10^4$. Moreover, in view of the method for computing the riccati equation explained in Chapter 11, it is best to take $Q = R^{-1} = 10^2$.

In view of the discussion of Section 4 of Chapter 2, the steady-state value of P can be obtained by setting $dP/dt = 0$ in the riccati equation (4.5). Moreover, this solution is always unique. We observe that if $\bar{P} = I$, then $dP/dt = 0$. Hence $\bar{P} = I$ is the steady-state solution of the riccati equation. Therefore,

$$\bar{K} = R^{-1}G'P = 10^2G' = [100 \quad 0 \quad 0]$$

is the optimal gain. The infinitesimal transition matrix of the closed-loop optimal system is:

$$(4.2) \quad F^0 = F - GK = F - 10^2 GG^T = \begin{bmatrix} -100 & +100 & 0 \\ -100 & 0 & +1 \\ 0 & -1 & 0 \end{bmatrix}$$

The eigenvalues of this matrix are given by

$$(4.3) \quad \begin{cases} \lambda_1 = -0.01000 \\ \lambda_{2,3} = -49.9950 \pm i 86.6054, \end{cases}$$

which shows that the optimal closed loop system is very poorly damped.

The explanation for the poor damping is the following: The criterion of optimality requires only that x_1 be quickly reduced to a small value. This does indeed happen, since the first row of the optimal closed-loop transition matrix is given by

$$\varphi_{11}(t) = 10^{-4}e^{-.01t} + 1.154e^{-49.995t} \sin(86.605t + 2.095),$$

$$\varphi_{12}(t) = -10^{-4}e^{-.01t} + 1.154e^{-49.995t} \sin(86.605t + 10^{-4}),$$

$$\varphi_{13}(t) = 10^{-2}e^{-.01t} - .0115e^{-49.995t} \sin(86.605t + 1.047).$$

This shows the effect of unit initial conditions in x_1, x_2, x_3 on $x_1(t)$.

On the other hand, the criterion of optimality does not require good control over x_3 . Since (see the discussion of controllability) x_3 is very weakly coupled to x_1 , the good transient response of x_1 does not bring about a similarly good transient response in x_3 . Another way of saying the same thing is that the control energy is used more effectively in reducing x_1 than in reducing x_3 , because x_3 does not enter directly in the error criterion $\|Lx\|_Q^2$ and the amount of energy required to quickly take x_3 to zero is enormous.

It should be noted that, in accordance with the general theorem (8.4) in Chapter 2, the closed-loop optimal system is asymptotically stable.

5. Second Design.

Motivated by the first set of results, we now let

$$(5.1) \quad H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and

$$(5.2) \quad Q = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}, \quad R = [10^{-2}].$$

In other words, we weight errors in x_2 and x_1 equally.

The observability matrix corresponding to (5.1) and (5.2) was obtained by computing the solution of the riccati equation (4.6) of Chapter 2 with $R^{-1} = 0$. We found

$$M(0, 1) = \begin{bmatrix} 4.9778 & 0.0126 & 0.0502 \\ 0.0126 & 5.0204 & -0.0001 \\ 0.0502 & -0.0001 & 9.9995 \end{bmatrix} \times 10,$$

and

$$(5.3) \quad M^{-1}(0, 1) = \begin{bmatrix} 2.0090 & -0.0050 & -0.0101 \\ -0.0050 & 1.9919 & 0.0000 \\ -0.0101 & 0.0000 & 1.0001 \end{bmatrix} \times 10^{-2}.$$

The steady-state value of P was obtained from machine calculations as

$$(5.4) \quad \bar{P} = \begin{bmatrix} 0.0101 & -0.0001 & 0.0099 \\ -0.0001 & 0.0101 & -0.0100 \\ 0.0099 & -0.0100 & 1.0098 \end{bmatrix} \times 10^2.$$

This gave an optimal steady-state gain matrix

$$(5.5) \quad \bar{K} = [1.0099 \quad -0.0099 \quad 0.9899] \times 10^2$$

and an infinitesimal transition matrix of the closed-loop optimal system

$$(5.6) \quad F^0 = F - G\bar{K} = \begin{bmatrix} -1.0099 & 1.0099 & -0.9899 \\ -1.00 & 0 & 0.01 \\ 0 & -0.01 & 0 \end{bmatrix} \times 10^{+2}$$

The eigenvalues (5.6) are

$$(5.7) \quad \begin{cases} \lambda_1 = - .9999, \\ \lambda_{2,3} = - 49.9950 \pm i 86.6054, \end{cases}$$

which shows that we have improved the damping by a factor of about 100 by giving x_3 equal weight with x_1 in the error criterion. Defined by H and Q, the new error criterion has forced the system to distribute the control energy better between x_1 and x_3 . This improvement has not been made, however, without a considerable increase in required control energy, even though the first component of the K matrix remains essentially the same. Note also that \bar{P}_{33} is 100 times larger in this case than in the previous section.

Finally, observe that control over x_1 and x_2 is virtually unchanged, and the complex eigenvalues of F^0 have remained the same.

6. Third Design.

Another design was investigated, setting

$$(6.1) \quad H^0QH = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 500 \end{bmatrix}.$$

Error in x_3 are weighted five times more than errors in x_1 . We again have complete observability.

The observability matrix was found to be

$$(6.2) \quad \mathcal{X}(0, 1) = \begin{bmatrix} 0.4984 & 0.0013 & 0.0452 \\ 0.0013 & 0.5023 & -0.0001 \\ 0.0452 & -0.0001 & 4.9991 \end{bmatrix} \times 10^2.$$

Notice that the terms in the third row and third column (i.e., terms associated with x_3) are much larger than in (5.2).

The steady-state value of P was obtained from machine calculations and it was observed that P converged more quickly than it did under the conditions of Section 5:

$$(6.3) \quad \bar{P} = \begin{bmatrix} 0.0102 & -0.0002 & 0.0223 \\ -0.0002 & 0.0102 & -0.0228 \\ 0.0223 & -0.0228 & 2.2862 \end{bmatrix} \times 10^2.$$

We see that all terms associated with x_3 are (approximately) $\sqrt{5}$ times larger than in (5.4).

The optimal steady-state gain matrix was found to be

$$(6.4) \quad \bar{K} = [1.0223 \quad -0.0225 \quad 2.2259] \times 10^2.$$

The infinitesimal transition matrix of the closed-loop optimal system was found to be:

$$(6.5) \quad F^0 = F - G\bar{K} = \begin{bmatrix} -1.0223 & 1.0225 & -2.2259 \\ -1.0000 & 0. & 0.0100 \\ 0. & -0.0100 & 0. \end{bmatrix} \times 10^2.$$

The eigenvalues of (6.5) matrix are

$$(6.6) \quad \left\{ \begin{array}{l} \lambda_1 = -2.2355, \\ \lambda_{2,3} = -49.9949 \pm 1.86.6199. \end{array} \right.$$

The following are some noteworthy aspects of this example:

(A) The elements k_{12} and k_{13} in (6.4) are larger by a factor of $\sqrt{5}$ than corresponding elements of (5.5). This is due to the change in Q/R .

(B) Despite this change, k_{12} is now still only .02 times k_{11} which has remained unchanged. In other words, there is essentially no change in the control over x_1 and x_2 . Hence the complex pair of eigenvalues in (6.6) remains about the same as (5.7).

(C) The ubiquitous factor of $\sqrt{5}$ is to be expected from the scalar analysis in Chapter 2; see equations (5.6) and (5.7).

7. Fourth Design.

Finally, we took

$$(7.1) \quad H'QH = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 200 & 0 \\ 0 & 0 & 500 \end{bmatrix}.$$

Again we have complete observability with

$$(7.2) \quad M(0,1) = \begin{bmatrix} 1.5025 & -0.0013 & 0.0352 \\ -0.0013 & 1.4977 & -0.0000 \\ 0.0352 & -0.0000 & 4.9992 \end{bmatrix} \times 10^2.$$

The riccati equation converged to a steady-state value of P more slowly than in the third design (Sect. 6), indicating that the largest eigenvalue of the closed loop system is somewhat closer to zero.

The steady-state gain matrix

$$(7.3) \quad \bar{K} = [1.5826 \quad -0.7523 \quad 2.2203] \times 10^2.$$

The infinitesimal transition matrix of the closed loop optimal system was

$$(7.4) \quad F^0 = F - GK = \begin{bmatrix} 1.5826 & 1.7523 & -2.2203 \\ -1.00 & 0 & 0.0100 \\ 0 & -0.0100 & 0 \end{bmatrix} \times 10^2.$$

The eigenvalues of this matrix (7.4) are

$$(7.5) \quad \begin{aligned} \lambda_1 &= -1.2910 \\ \lambda_{2,3} &= -78.4830 \pm i 211.2952 \end{aligned}$$

The size of the real eigenvalue accords with the qualitative prediction made from the rate of convergence of the riccati equation.

Qualitatively the shifts in the eigenvalues could be predicted on the basis of more control energy being put into the $x_1 - x_2$ loop at the expense of x_3 control. Quantitatively the picture is considerably more complicated than it was in the one dimensional system analyzed in Chapter 2. For instance, it would appear heuristically inviting to assume that x_1 and x_2 are so tightly coupled that

$$H'QH = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 200 & 0 \\ 0 & 0 & 500 \end{bmatrix}$$

is the same as

$$H'QH = \begin{bmatrix} 300 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 500 \end{bmatrix}$$

and such an assumption is possible to maintain about the observability matrix where a factor of three appears in the $x_1 - x_2$ terms. But this point of view is too naive to account for the changes in \bar{P} and \bar{K} .

A careful analysis of the means of applying the inequalities of Chapter 2 is required in order to obtain information about Q and R in terms of parameters more familiar to the engineer, such as time constants and frequencies.

8. References.

S. J. MASON (1956) "Feedback theory -- further properties of signal flow graphs". Proc. IRE, 44, 92

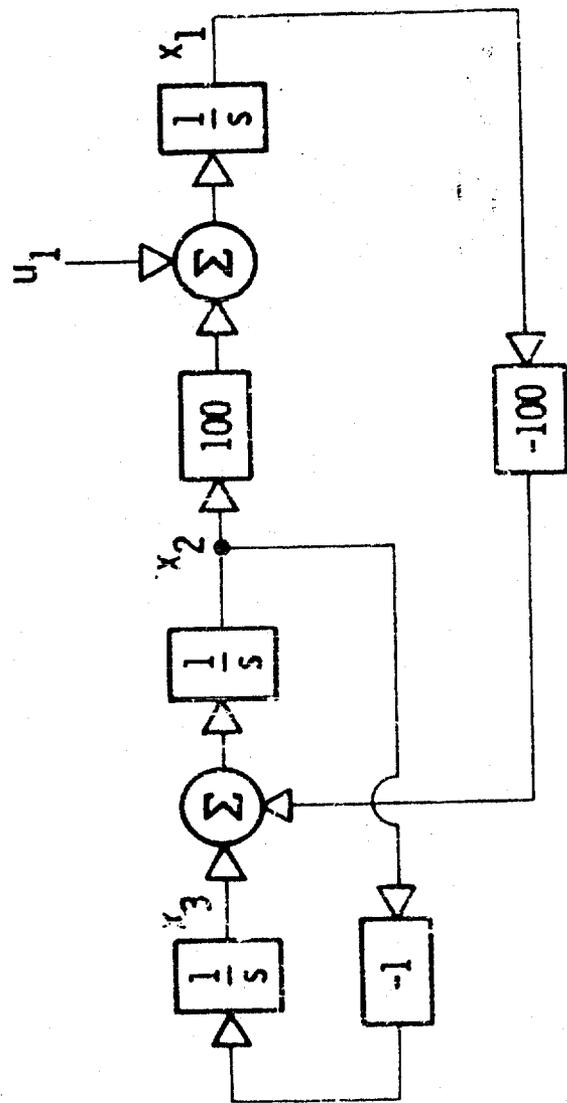


FIG. 1 BLOCK DIAGRAM OF CONTROL OBJECT

Chapter 4.

OPTIMAL FILTERING THEORY

1. Motivation of Assumptions.

Another problem of optimization of importance for the study of adaptive systems is that of statistical estimation theory or generalized Wiener filtering. In this problem it is usually assumed that one observes a signal in the presence of additive noise and one desires to find a "best estimate" of the signal by linear operations on the observations.

In relation to control theory, the assumptions are stated in a somewhat different (but by and large equivalent) form. We assume that the state x of the control object cannot be observed directly. We can, however, measure some linear combinations of the state variables. These measurements are denoted by the vector y . The measurements are not made with perfect accuracy, so that actually we observe a vector z which is the sum of y and a vector v representing measurement errors or noise. In addition, it is also assumed that the control object is subject to certain random disturbances w acting on it.

In accordance with the discussion in Sect. 2, Chapter 1, these assumptions, combined with linearity, yield the equations

$$(1.1) \quad dx/dt = F(t)x + G(t)v(t)$$

$$(1.2) \quad z(t) = y(t) + v(t) = R(t)x(t) + v(t)$$

where $w(t)$ and $v(t)$ are random processes (see below). In the first equation, we omit a term involving $u(t)$ since we are not concerned here with the control problem. This term will reappear again in Chapter 5.

Alternately, one can interpret these equations in the following terms: We are given a random process $x(t)$ and a related random process $z(t)$. Values of $z(t)$ are observed over a certain interval of time. On the basis of these observations, we wish to estimate the value of $x(t_1)$, where t_1 is some arbitrarily chosen instant of time. It can be shown that if $x(t)$ is a markovian and gaussian random

process, it can always be represented by a scheme such as (1.1-2). See [Kalman, 1961C, Sect. 7].

We make the further assumption that $v(t)$ and $w(t)$ are gaussian white-noise processes independent of each other. They are specified mathematically by their covariance matrices:

$$(1.5) \quad \left\{ \begin{array}{l} E v(t) v'(\tau) = [E v_i(t) v_j(\tau)] = \bar{R}(t) \delta(t - \tau), \\ E w(t) w'(\tau) = [E u_i(t) u_j(\tau)] = \bar{Q}(t) \delta(t - \tau), \\ E v(t) v'(\tau) = 0 \quad \text{for all } t, \tau, \\ E w(t) = E v(t) = 0 \quad \text{for all } t, \end{array} \right.$$

where E represents the mathematical expectation, $\delta(t - \tau)$ is the Dirac delta function, \bar{Q} , \bar{R} are positive definite matrixes. The case of nonwhite noise can be reduced to this formulation by a change of variables.

We shall often refer to (1.1-2) as the model of the signal process. The matrix block diagram for the model is seen in Fig. 1.

A much more detailed discussion of the subject of this chapter may be found in [Kalman, 1961C].

2. Statement of the Filtering Problem.

The filtering problem can be stated as follows: Determine a linear operator on the set of observations $\{z(\tau) | \tau \in [t_0, t]\}$ whose value $\hat{x}(t_1 | t)$ at time t_1 has the properties:

$$(i) \quad E \hat{x}(t_1 | t) = E x(t_1),$$

$$(ii) \quad E \|\tilde{x}(t_1 | t)\|_B^2 = \sum_{i,j=1}^n b_{ij} E \tilde{x}_i \tilde{x}_j = \text{minimum} \quad (B \text{ any positive definite matrix}).$$

Above we used the abbreviation

$$(2.1) \quad \tilde{x}(t_1 | t) = x(t_1) - \hat{x}(t_1 | t)$$

for the error in the estimate $\hat{x}(t_1 | t)$. The observations of z start at time t_0 .

(which is taken as fixed), and end at time t (which is taken as a running parameter).

Thus $\hat{x}(t_1|t)$ is to be unbiased, minimum variance estimator of $x(t)$; that is, $\hat{x}(t_1|t)$ minimizes the average value of the squares of the error. One of the interesting properties of the estimator $\hat{x}(t_1|t)$ is that the best estimator of the scalar quantity

$$(2.2) \quad a'x(t_1) = \sum_{i=1}^n a_i x_i(t)$$

turns out to be

$$a'\hat{x}(t_1|t).$$

3. Solution of Filtering Problem.

By a rather involved argument given in detail in [Kalman-Bucy 1961; Kalman 1961C] it can be shown that $\hat{x}(t|t)$ is the output of a dynamical system similar to (1.1) whose input consists of the observations $z(t)$:

$$(3.1) \quad \frac{d\hat{x}(t|t)}{dt} = F(t)\hat{x}(t|t) + K(t)[z(t) - H(t)\hat{x}(t|t)].$$

The dynamical system (3.1) can be physically realized by a feedback system as shown in Fig. 2.

It can be shown that $K(t)$, the gain of the optimal filter, is determined by the covariance matrix of the errors of the optimal filter. In fact, if

$$E\tilde{x}(t|t)\tilde{x}'(t|t) = \Sigma(t)$$

then

$$(3.2) \quad K(t) = \Sigma(t)H'(t)R^{-1}(t)$$

Further, it can be shown [Kalman-Bucy, 1961; Kalman 1961C] that $\Sigma(t)$ is determined as the solution of the following matrix riccati equation:

$$(3.3) \quad \frac{d\Sigma}{dt} = F(t)\Sigma + \Sigma F'(t) - \Sigma H'(t)R^{-1}(t)H(t)\Sigma + Q(t)$$

where

$$\Sigma(t_0) = E x(t_0) x'(t_0).$$

To avoid confusion between (3.3) and the matrix riccati equations of Chapter 2, we shall usually refer to (3.3) as the variance equation.

Notice that this equation is solved forward in time. By solving (3.3) and then using (3.2) the optimal filter (3.1) completely specified

4. Duality Relations.

It should be noted that the solutions to the regulator problem and the filtering problem are quite similar: in each case the problem reduces to the solution of a matrix riccati equation. Actually much more is true. To every filtering problem there corresponds a "dual" control problem so that the same riccati equation provides the answer to both problems. The "duality relations" may be stated explicitly as follows:

<u>Filtering</u>		<u>Control</u>
$\Sigma(t)$	\longleftrightarrow	$P(t)$
F	\longleftrightarrow	F'
G	\longleftrightarrow	H'
H	\longleftrightarrow	G'
t_0	\longleftrightarrow	T
\bar{Q}	\longleftrightarrow	Q
\bar{R}	\longleftrightarrow	R

This shows in particular that the conditions of complete controllability and complete observability are duals of one another. A completely controllable dynamic system of a control problem is the dual of a completely observable dynamical system of a filtering problem. Hence the existence and uniqueness theorems in Chapter 2 are valid also for the filtering problem, with the conditions dualized and the conclusions now pertaining to the optimal filter rather than the optimal control system. Each counter-example of Chapter 2 would serve, after it is dualized, as a counter-example for a filtering theorem.

5. Example of a Filtering Problem.

The following special case of filtering problem will be considered in detail to illustrate the application of the general theory.

The signal process x_1 is given by the differential equation

$$(5.1) \quad \frac{dx_1}{dt} = f_{11}x_1 + w_1,$$

the observed signal is

$$(5.2) \quad z_1 = x_1 + v_1.$$

In accordance with (1.3), it will be assumed that

$$(5.3) \quad \begin{aligned} Ew_1(t)w_1(\tau) &= \bar{q}_{11}\delta(t - \tau) \\ Ev_1(t)v_1(\tau) &= \bar{r}_{11}\delta(t + \tau) \\ Ex_1^2(t_0) &= \sigma_{11}(t_0) = \alpha \end{aligned}$$

or

$$F = [f_{11}], \quad G = [1], \quad H = [1], \quad \bar{Q} = [q_{11}], \quad \bar{R} = [r_{11}].$$

Specializing (3.1) to this case, the equation of motion of the optimal filter is

$$(5.4) \quad \frac{d\hat{x}_1(t|t)}{dt} = f_{11}\hat{x}_1(t|t) + \frac{\sigma_{11}(t)}{r_{11}} [z_1(t) - \hat{x}_1(t|t)]$$

The block diagram of the optimal filter is shown in Fig. 3.

The solution of (3.3) in this case can be found by separation of variables and integrating. The end result is

$$\frac{\sigma_{11}(t)}{r_{11}} = \frac{\frac{\alpha}{r_{11}} \left[\sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} + f_{11} \right] (r_{11} - f_{11}) + \sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} e^{-\sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} (t-t_0)}}{(f_{11} + \sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2}) + \left[\frac{\alpha}{r_{11}} - f_{11} + \sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} \right] e^{-2\sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} (t-t_0)}} - \left[\sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} - f_{11} \right] (r_{11} + \sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2})$$

Equation (5.5)

The solution of the steady-state filtering problem ($t_0 = \infty$), the conventional Wiener problem, exists since the model is completely controllable and completely observable. The solution of the Wiener problem is given by

$$(5.6) \quad \lim_{t \rightarrow +\infty} \sigma_{11}(t) = (f_{11} + \sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2}) \bar{r}_{11} = \frac{q_{11}}{\sqrt{\frac{q_{11}}{r_{11}} + f_{11}^2} - f_{11}} = \bar{\sigma}_{11}$$

This is a well-known result of the conventional theory.

Notice that the optimal filter is stable regardless of whether the signal process was stable or not.

If $f_{11} < \frac{q_{11}}{r_{11}}$ then the time constant of the optimum filter is at most

$\left(\frac{\bar{r}_{11}}{q_{11}}\right)^{\frac{1}{2}}$, which shows that the less noise power in relation to signal power the faster the filtering loop. Hence the time constant of the optimal filter depends directly on the signal to noise ratio. Since the filtering problem is the dual of the control problem, all of the extensive discussion in Chapter 2 is relevant also to the filtering problem.

6. References.

- R. E. KALMAN (1961C) "New methods and results in linear filtering and prediction theory", Proc. Symp. on Engineering Applications of Probability and Random Functions, Purdue University, Nov. 1960; to be published by Wiley.
- R. E. KALMAN and R. S. HUGY (1961) "New Results in Linear Filtering and Prediction Theory", J. Basic Engr. (Trans. ASME), (83) (to appear).

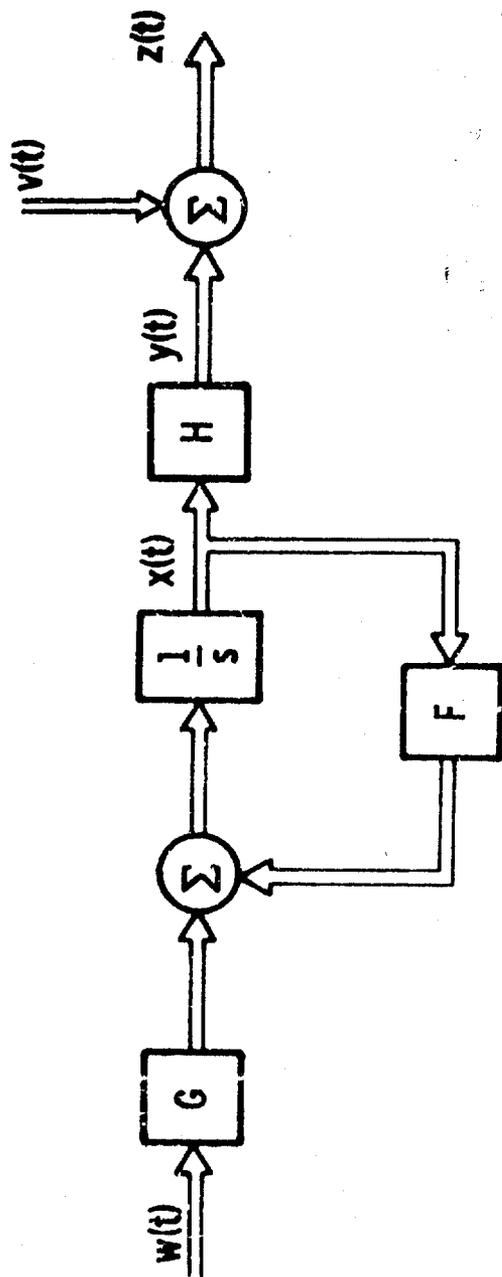


FIG. 1 MODEL OF SIGNAL PROCESS

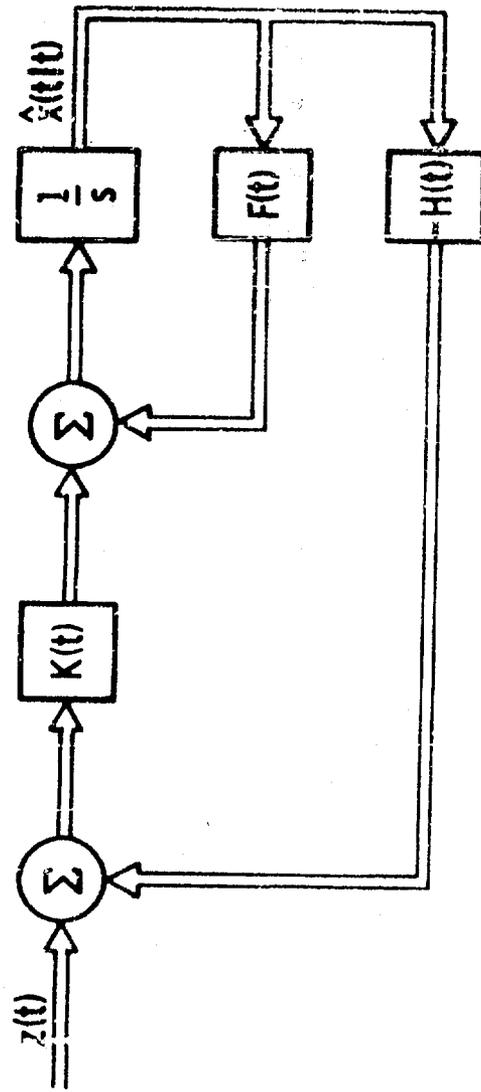


FIG. 2 BLOCK DIAGRAM OF OPTIMAL FILTER

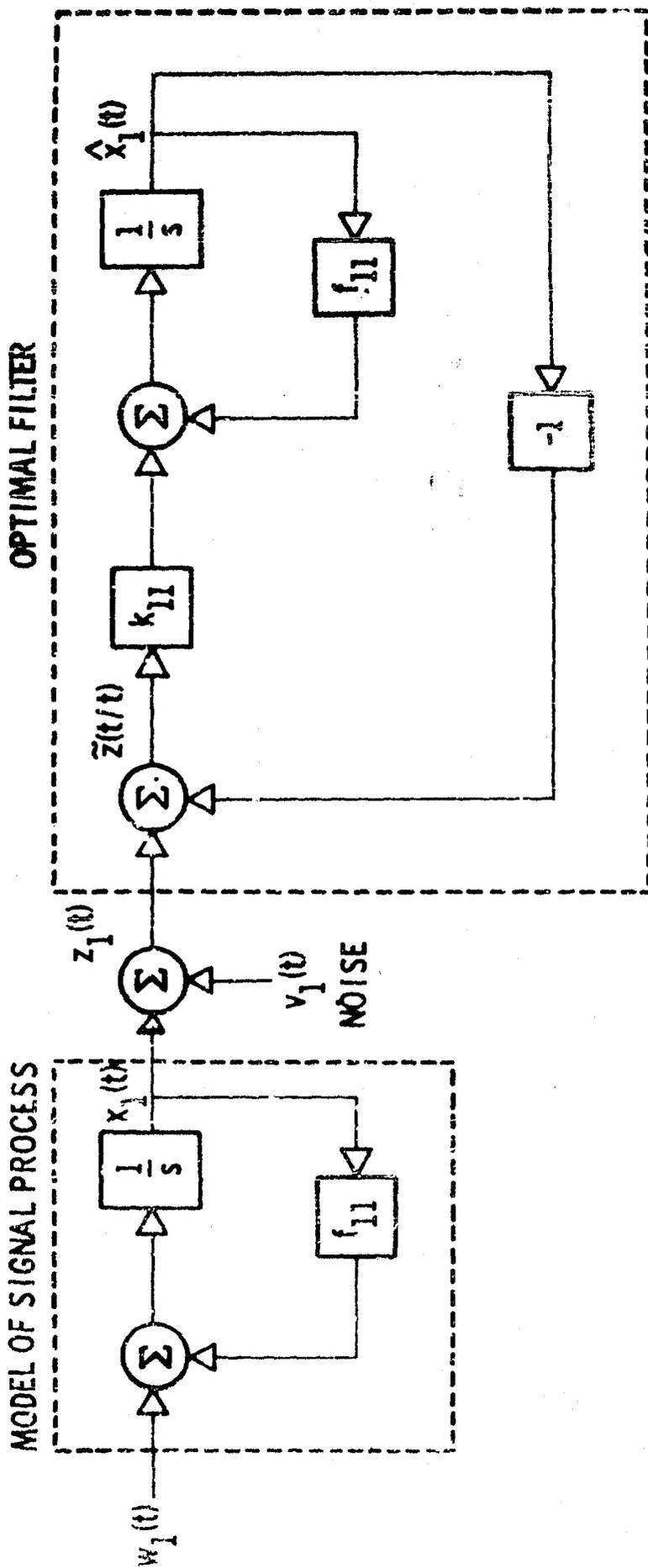


FIG. 3 FIRST-ORDER SIGNAL PROCESS AND OPTIMAL FILTER

Chapter 6.

THE ADAPTIVE CONTROL PROBLEM

1. Orientation.

It is now quite clear how Fig. 2 of Chapter 1 is to be interpreted. We assume that the equations of motion are given in form introduced in Chapter 1. We solve the filtering problem first, yielding the box marked "state estimator". Then, as discussed in the preceding chapter, we form the "optimal controller" by operating on $\hat{x}(t)$. This will be a linear operation, represented by the matrix $K_2(t)$.

Assume now that we have a means of measuring the values of the matrices, $F, G, H, \bar{Q}, \bar{R}$. The estimates of these parameters will form the learning states. The other parameters of the problem, namely S, Q, R specifying the quadratic performance index are usually given exactly.

With these estimates, we compute the solutions of the two relevant riccati equations of the filter and regulator problem, closing the "adaptive" loop.

Of course, no claim can be made at this time that such a procedure is optimal. It is probably not. But the combined problem of instantaneously best control and best estimation of the structural parameters is too difficult at present to be seriously studied.

2. Ideal Adaptation.

Suppose that the learning process were ideal, that is, it is possible to determine the values of $F, G, H, J, \bar{Q}, \bar{R}$ by observing the system output $z(t)$. Then one could design a controller on the basis of the theory of Chapter 5. The combination of the general controller and this ideal learning model will be called an ideal adaptive system. Obviously, it will have the best performance of any adaptive system. Because of the theory of the general control problem presented in Chapter 5, the performance of the ideal adaptive system in a given environment can be determined exactly -- this is just the general time varying control problem.

The concept of an ideal adaptive system has two major practical uses:

The performance index is now defined to be the average value of V , denoted by EV . The average is taken with respect to the probability distribution of w .

The general control problem is then the following: Find a control $u(t)$ such that EV is minimized.

In Chapter 2, the optimal $u(t)$ depended only on the initial state. Here this is no longer true, because the effect of w cannot be predicted at the beginning of the control process. As additional values of the state are measured, more information is obtained and this information must somehow be utilized in computing the optimal $u(t)$.

To define the problem precisely, we must therefore also add the following: Control must be based on the actually observed values of $z(t)$ in the interval (t_0, t) .

3. Solution of the General Control Problem.

In this section we shall give the form of the general solution; details will be omitted, since the theory is not yet complete.

The best estimator $\hat{x}(t|t)$ of $x(t)$ is orthogonal to the error of estimation $\tilde{x}(t|t)$; hence

$$\begin{aligned} (3.1) \quad E\|x(t)\|_{Q(t)}^2 &= E\|\hat{x}(t|t) + \tilde{x}(t|t)\|_{Q(t)}^2 \\ &= E\|\hat{x}(t|t)\|_{Q(t)}^2 + E\|\tilde{x}(t|t)\|_{Q(t)}^2. \end{aligned}$$

It can be shown that $\hat{x}(t|t)$ and $\tilde{x}(t|t)$ satisfy the differential equations

$$(3.2) \quad \frac{d\hat{x}(t|t)}{dt} = F(t)\hat{x}(t|t) + K_1(t)[H(t)\tilde{x}(t|t) + v(t)] + G(t)u(t)$$

and

$$(3.3) \quad \frac{d\tilde{x}(t|t)}{dt} = F(t)\tilde{x}(t|t) - K_1(t)[H(t)\tilde{x}(t|t) + v(t)] + w(t)$$

On the basis of equation (3.1) it follows that

$$(3.4) \quad \begin{aligned} EV = E(\|\hat{\phi}_u(T, \underline{x}, t, w)\|_{\underline{g}}^2) &+ \int_t^T \|\bar{R}(\tau)\hat{\phi}_u(\tau; \underline{x}, t, w)\|_{\underline{Q}(\tau)}^2 + \|u(\tau)\|_{\bar{R}(\tau)}^2 d\tau \\ &+ E(\|\tilde{\phi}_u(T, \underline{x}, t, w)\|_{\underline{g}}^2) + \int_t^T \|\bar{R}(\tau)\tilde{\phi}_u(\tau; \underline{x}, t, w)\|_{\underline{Q}(\tau)}^2 d\tau, \end{aligned}$$

where $\hat{\phi}$ is the motion of (3.2) and $\tilde{\phi}$ is motion of (3.3). From the form of (3.2-4) we see that the problem splits into two parts:

- (A) Estimate $x(t)$ by filtering theory.
- (B) Control the system defined by (3.2) according to the noise-free regulator theory.

This "decoupling" of the problem into the two parts previously discussed is due to the linearity and the fact that the random forcing term in (3.2) is a white-noise process with zero mean. Since such a process is completely unpredictable, it cannot be taken into account in computing the optimal control law. In other words, the solution of the regulator problem when the state can be exactly and instantaneously measured is the same with or without a white-noise type of forcing term.

The canonical form of the optimal control system in the general case is shown in Fig. 1, which is self-explanatory. K_1 is used to denote the optimal feedback gains obtained from the riccati equation of the control problem.

4. Engineering Implications of the Form of the Solution.

As in the filtering problem, the ratio of the disturbance power \bar{Q} to the noise power \bar{R} gives an estimate of the reciprocal time constant of the filter in Fig. 1. If the disturbance power is small in comparison to the noise power one gets a relatively low gain in the filtering loop. However, \bar{Q} may be regarded as a rough measure of how well the dynamic model is known, \bar{Q} being large for the case when the model dynamics are not known well. In practice usually \bar{Q} can be assumed to be fairly large in comparison to \bar{R} .

As in the noise-free regulator problem, the time constant of the control loop can be approximately specified by the choice of RQ^{-1} .

Arguments similar to those in Chapter 2 give equivalent time constants for the riccati equations (16) and (17) and provide verifiable conditions as to when the two gains K_1 and K_2 can be replaced by constants.

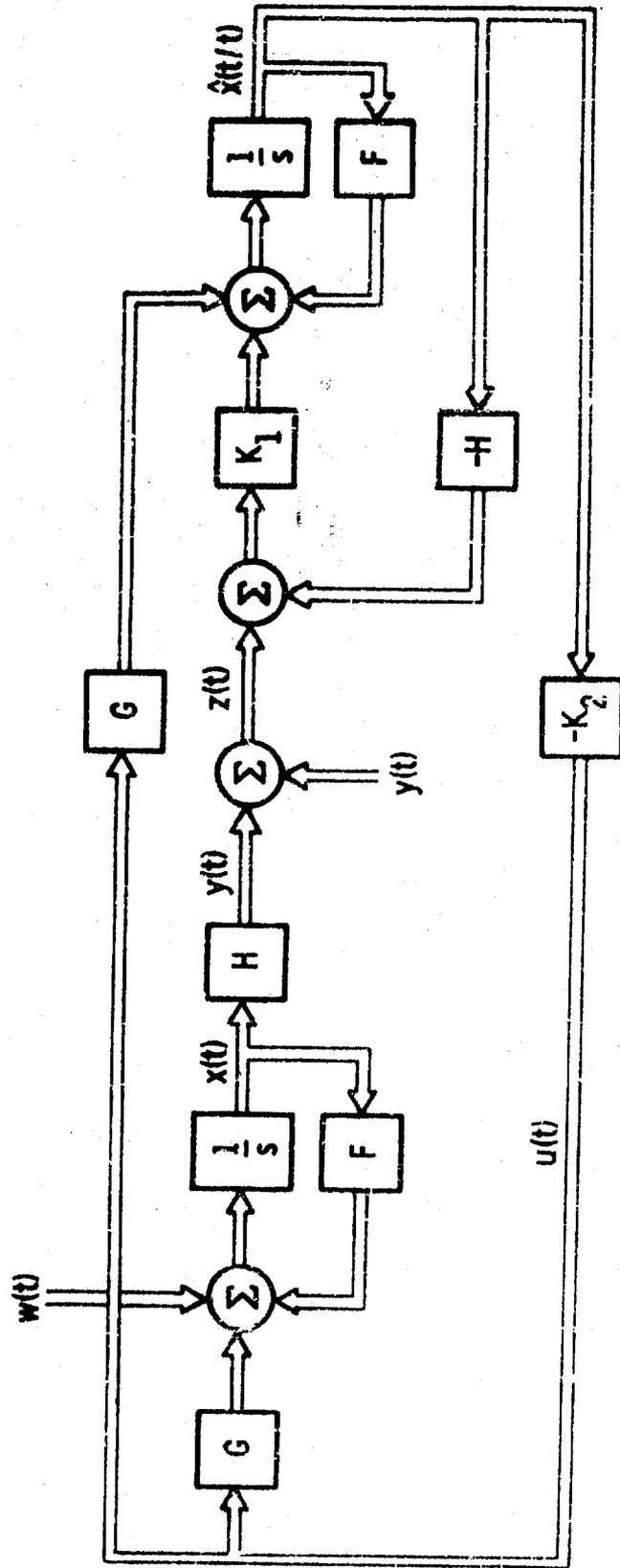


FIG. 1 CANONIC FORM OF OPTIMAL CONTROL SYSTEM I

Chapter 6.

THE ADAPTIVE CONTROL PROBLEM

1. Orientation.

It is now quite clear how Fig. 2 of Chapter 1 is to be interpreted. We assume that the equations of motion are given in form introduced in Chapter 1. We solve the filtering problem first, yielding the box marked "state estimator". Then, as discussed in the preceding chapter, we form the "optimal controller" by operating on $\hat{x}(t)$. This will be a linear operation, represented by the matrix $K_2(t)$.

Assume now that we have a means of measuring the values of the matrices, $F, G, H, \bar{Q}, \bar{R}$. The estimates of these parameters will form the learning states. The other parameters of the problem, namely S, Q, R specifying the quadratic performance index are usually given exactly.

With these estimates, we compute the solutions of the two relevant riccati equations of the filter and regulator problem, closing the "adaptive" loop.

Of course, no claim can be made at this time that such a procedure is optimal. It is probably not. But the combined problem of instantaneously best control and best estimation of the structural parameters is too difficult at present to be seriously studied.

2. Ideal Adaptation.

Suppose that the learning process was ideal, that is, it is possible to determine the values of $F, G, H, J, \bar{Q}, \bar{R}$ by observing the system output $z(t)$. Then one could design a controller on the basis of the theory of Chapter 5. The combination of the general controller and this ideal learning model will be called an ideal adaptive system. Obviously, it will have the best performance of any adaptive system. Because of the theory of the general control problem presented in Chapter 5, the performance of the ideal adaptive system in a given environment can be determined exactly -- this is just the general time varying control problem.

The concept of an ideal adaptive system has two major practical uses:

- (1) The evaluation of various alleged "adaptive" designs, the determination of whether an adaptive controller is really needed, and whether even an ideal adaptive controller can do the job.
- (2) The actual design of adaptive controllers.

These two uses will be explored in the next sections.

3. Evaluation of Adaptive Designs.

To check a given adaptive system design, we prescribe the evolution of the control object in time by specifying $F(t)$, $G(t)$, $H(t)$, $\bar{Q}(t)$, and $\bar{R}(t)$. We then compute the optimal control system based on the knowledge of these parameters. This gives us the performance of the optimal adaptive system. For large learning times, a well-designed adaptive system should approach this ideal.

We can also obtain a lower bound on the performance of an adaptive system. We take some "average" value of $F(t)$, $G(t)$, $H(t)$, $\bar{Q}(t)$ and $\bar{R}(t)$, and design a control system with a constant control law based on these parameters. If a control system is truly adaptive, it must perform better than one whose control law is based on the "average" equations of motion. Of course, there is no guarantee that any design with a constant control law will be stable under the various conditions which may be encountered; if so, this is a sure indication that an adaptive system is called for.

4. Design of Adaptive Systems.

An adaptive filter could be envisioned as follows. Take estimates \hat{F} , \hat{G} , \hat{H} , \hat{Q} , \hat{R} of F , G , H , Q , R supplied by the learning process. Substitute the estimates in the riccati equation, and use the solution of this equation to set the gains of the optimal filter. See Fig. 1, where the adaptive adjustments are indicated by dashed lines. This is, of course, largely an "open loop" process; it could be improved by measuring the performance of the filter and then feeding this information back to the learning process.

If one would try to extend this scheme to the control problem difficulties would arise since the control feedback gains are obtained by solving the riccati equation backwards in time. Therefore in the control problem the learning process must supply predictions of F , G , H at least as far into the future as several time constants

of the riccati equation of the control problem. This is one of the reasons why it is important to understand quantitatively the dynamical behavior of the riccati equation.

This approach separates the art of the design of the learning process from the science of optimal control.

5. Example of an Adaptive Filter.

This example will be explained only in a superficial way since it is quite involved. The interested reader can consult [Bucy, 1959].

The model of the signal process is taken as

$$(5.1) \quad \begin{aligned} dx_1/dt &= x_2, \\ dx_2/dt &= w_1, \end{aligned}$$

while the observed signal is

$$(5.2) \quad z_1 = x_1 + v_1.$$

In other words,

$$(5.3) \quad F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad H = [1 \quad 0]$$

$$Q = \begin{bmatrix} 0 & 0 \\ 0 & \bar{q}_{11} \end{bmatrix},$$

$$(5.4) \quad \bar{R} = \begin{bmatrix} r_{11} \end{bmatrix}$$

Note that this system is completely observable and completely controllable. In this case the optimal filter is described by the equations

$$(5.5) \quad \begin{aligned} dx_1/dt &= x_2 + k_{11}x_1, \\ dx_2/dt &= k_{21}(z_1 - x_1) \end{aligned}$$

where

$$(5.6) \quad k_{11} = \frac{\sigma_{11}}{\bar{r}_{11}}, \quad k_{21} = \frac{\sigma_{12}}{\bar{r}_{11}}, \quad c = \frac{\sigma_{22}}{\bar{r}_{11}}.$$

The scalars k_{11} , k_{21} and c satisfy the equations

$$(5.7) \quad \begin{aligned} dk_{11}/dt &= 2k_{21} + k_{11}^2, \\ dk_{21}/dt &= c - k_{11}k_{21}, \\ dc/dt &= \frac{\bar{q}_{11}}{\bar{r}_{11}} - k_{21}^2, \end{aligned}$$

which follow immediately from the variance equations in Chapter 4.

Now suppose that \bar{q}_{11} and \bar{r}_{11} are the only unknown parameters. Then it follows from (5.7) that if $\bar{q}_{11}/\bar{r}_{11}$ could be estimated then (5.7) could be solved to set the gains in (5.5) and hence achieve adaptive behavior.

The variable $\epsilon = z_1(t) - \hat{x}_1(t|t) = \tilde{x}_1(t|t) + v(t)$ is white-noise, i.e., has flat spectrum when the system (5.5) is optimal. When (5.5) is not optimal ϵ (with λ being the estimate of $\bar{q}_{11}/\bar{r}_{11}$), varies as in Fig. 2. One can compute the area under the spectrum of ϵ from 0 to ω_0 by passing ϵ through a low-pass filter and then rectifying it. Likewise, the area under the spectrum of ϵ from ω_0 to $2\omega_0$ is computed by means of passing ϵ through a band-pass filter and then rectifying it. Therefore a convenient learning process is provided by the nonlinear differential equation

$$(5.8) \quad d\lambda/dt = k_3 [\alpha |\epsilon|_{L.P.} - \beta |\epsilon|_{B.P.}]$$

where

$|\epsilon|_{L.P.}$ = result of passing ϵ through a low-pass filter and rectifying;

$|\epsilon|_{B.P.}$ = result of passing ϵ through a band-pass filter and rectifying;

The final adaptive system shown in Fig. 3 is described by the equations (5.5), (5.7), (5.8).

This system can be made more sophisticated by determining the constants α , β , and k_3 so as to provide the widest stability margins, and passing ϵ through an exact copy of the filter loop.

In [Bucy, 1959] a rather detailed system is described, and results of computer simulation are given which substantiate the theoretical analysis.

6. References.

- R. S. BUCY (1959) "Adaptive Finite-Time Filtering", Johns Hopkins University Applied Physics Laboratory, Internal Memorandum HED -645.

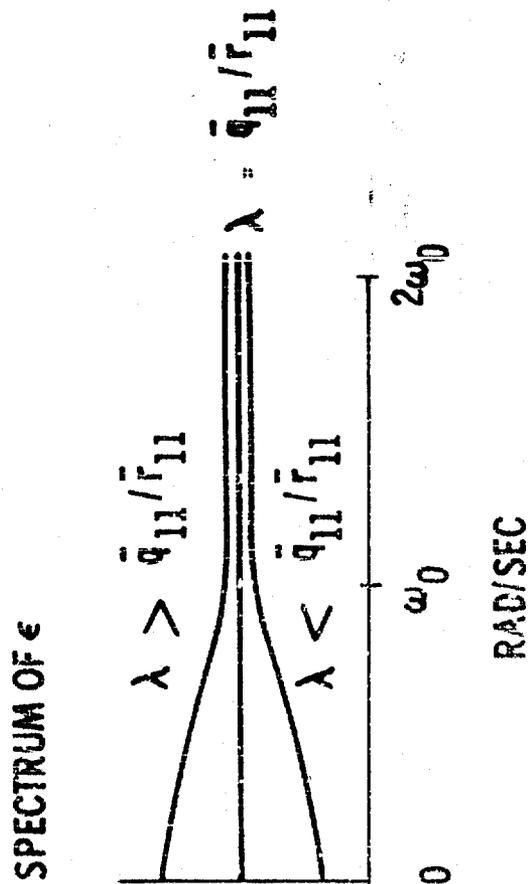


FIG. 2 SPECTRUM OF ϵ AS A FUNCTION OF THE ESTIMATE OF $\bar{q}_{11}/\bar{r}_{11}$

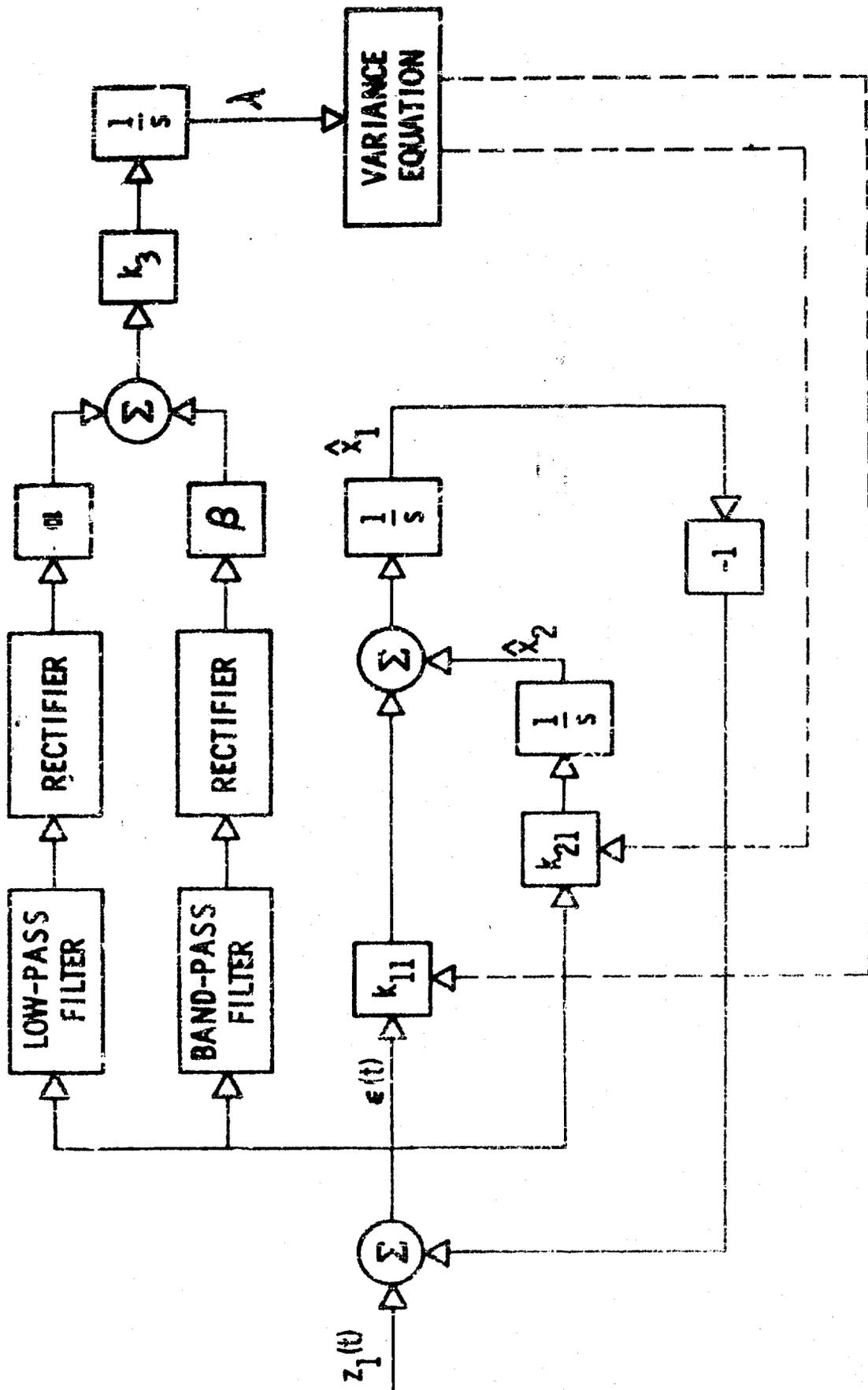


FIG. 3 ADAPTIVE FILTER

Chapter 7.

GUIDING PRINCIPLES OF NUMERICAL COMPUTATION

One of the important objectives of this project is to develop effective methods for dealing with complex control systems. Since emphasis was not on the analysis but on the optimal synthesis of these systems, analog computation was out of the question. From the beginning we were striving to obtain efficient methods of digital computations. As a matter of fact, the methods presented in this report would be rather awkward to apply without the use of a digital computer. This is the price that must be paid to obtain methods which are applicable regardless of complexity of the problem. Only by a combination of imaginative computer utilization and advanced mathematical techniques can engineering problems of complex system design be effectively attacked.

Guided by this philosophy of approach, we developed a general computational framework for problems in systems theory. The following specific objectives have been accomplished.

(A) All computations should take place in the time domain. This was necessitated by the nature of the underlying mathematical theory.

(B) The computations should be "eigenvalueless". That is to say, no inversion of Laplace transforms, solution of high-order algebraic equations, etc. should be required.* The methods we are using work easily for 15-th order systems (which is the maximum size for which they have been designed) and can surely be extended to at least 30-th order systems without the need for basically different numerical methods.

* These approaches run into serious numerical difficulties when the order of the system exceeds perhaps 10; the difficulties become probably fatal when the order exceeds 50.

(C) The specifications of the system should be given in matrix form. This is necessary for ease of programming and desirable in order that the programs have maximum flexibility.

(D) Sampled-data systems should be possible to treat without special techniques. This is really a by-product of the mathematical theory. Much of the elaborate engineering theory on sampled-data systems can be dispensed with. The same programs can be used. The principles of solution of specific problems in the continuous and sampled-data cases differ only in minor details.

(E) The results should be displayed in a neat form. We have adopted standard formats for printing matrices. Every effort has been made to present results of computations in a clean, usable, and complete form. We are nearing the stage where the end-result of a specific system optimization problem is a "book" produced by the computer, which describes the problem, exhibits the answers, provides partial checks in the course of the computation, etc.

The principal program from the numerical point of view is a subroutine for the computation of the exponential of a matrix. This is the central part of transient computations. One can also use this subroutine to solve efficiently some problems which at first sight require only elementary methods.

Consider, for instance, the linear matrix equation

$$(1) \quad F'P + PF = -Q$$

where Q and F are given constant $n \times n$ matrices and one wants to find a symmetric $n \times n$ matrix P satisfying the equation. This problem occurs in the second method of Lyapunov, in constructing a Lyapunov function for a linear system with constant coefficients. Of course (1) is just a set of $n(n+1)/2$ equations in the $n(n+1)/2$ unknown elements $P_{11}, \dots, P_{1n}, P_{22}, \dots, P_{2n}, \dots, P_{nn}$ of the symmetric matrix P . Hence the problem can be solved using a standard matrix inversion subroutine.

There are two difficulties with this method. First, if $n = 10$, then $n(n+1)/2 = 55$, and linear equations of this size cannot be solved by elementary numerical techniques. Second, the data are not arranged in such a way that the matrix defining the $n(n+1)/2$ linear equations in the elements of P can be read off by inspection. As a matter of fact, exceedingly tedious bookkeeping is needed to obtain this matrix, since it has $(55)^2 = 3,025$ elements.

One can compute P by a different method, however. It is well known that if F is a stable matrix, then

$$P = \int_0^{\infty} e^{F't} Q e^{Ft} dt$$

and this expression can be readily evaluated by using the riccati-equation subroutine (which, in turn, is based almost entirely on the exponential subroutine). In fact, consider

$$P(t) = \int_t^{\infty} e^{F'(t-\tau)} Q e^{F(\tau-t)} d\tau$$

Differentiating with respect to t , we see that $P(t)$ satisfies the differential equation

$$(2) \quad -\dot{P} = F'P + PF + Q$$

which is a special case of the riccati equation. If all eigenvalues of F have negative real parts, then as $t \rightarrow -\infty$, $\dot{P} \rightarrow 0$. If $\dot{P} = 0$, then (2) reduces to (1).

The solutions of differential equation (2) can be computed rapidly and accurately by means of the riccati-equation subroutine which is, in effect, a special iteration procedure.

Thus we see that by reducing a trivial algebraic problem (1) to a nontrivial analytic one (2), a great deal can be gained from the point of view of efficiency and ease of numerical computation.

The following four chapters contain a description of the main subroutines which were developed to date. Each subroutine contains some contribution to numerical analysis. To aid the eventual users fairly detailed explanations are given concerning the origin of the subroutine, the methods of computation, and numerical checks.

Chapter 8.

THE EXPONENTIAL SUBROUTINE

1. Theory.

The matrix exponential e^{Ft} may be defined with the aid of the everywhere convergent power series

$$(1.1) \quad e^F = \exp F = \sum_{i=0}^{\infty} \frac{F^i}{i!} \quad *$$

To show the convergence of the power series, note that

$$\|e^F\| = \left\| \sum_{i=0}^{\infty} \frac{F^i}{i!} \right\| \leq \sum_{i=0}^{\infty} \frac{\|F\|^i}{i!} = e^{\|F\|}$$

which shows that the matrix series for e^F converges whenever the scalar series for $e^{\|F\|}$ converges; the latter series is well-known to converge uniformly for all t in any bounded interval $[T, U]$. This function is of interest in this report primarily because it is the transition matrix of the vector differential equation

$$(1.2) \quad dx/dt = Fx \quad (F = \text{constant}).$$

In other words, we can show that

$$(1.3) \quad \phi(t, t_0) = \exp[(t - t_0)F].$$

According to Sect. 1, Chapter 2, we have to verify two properties of $\exp[(t - t_0)F]$ in order to prove (1.3). First, if $t = t_0$,

* It should be noted that $\exp(A + B)t \neq (\exp At)(\exp Bt)$ unless A and B commute.

$$\Phi(t, t) = \exp 0 \cdot F = I,$$

which is trivial; second, we must show that $\exp[(t - t_0)F]$ satisfies the differential equation (1.3). By the definition of the derivative,

$$d(\exp[(t - t_0)F])/dt = \lim_{h \rightarrow 0} \frac{e^{(t+h-t_0)F} - e^{(t-t_0)F}}{h}.$$

Since tF and hF commute,

$$e^{(t-t_0+h)F} = e^{(t-t_0)F} e^{hF}.$$

Hence

$$d e^{(t-t_0)F} / dt = \lim_{h \rightarrow 0} \frac{e^{hF} - I}{h} \cdot e^{(t-t_0)F}.$$

By (1.1),

$$= F \cdot e^{(t-t_0)F}$$

which was to be proved. Note that this proof fails in general if $F \neq$ constant.

This may be demonstrated also by termwise differentiation, (as we have already shown) because the series (1.1) converges uniformly on every interval $[0, t]$.

Some other facts which may be proved about e^{Ft} are (see [Coddington and Levinson, 1955])

$$(1.4) \quad \|e^{At}\| \leq (n-1) + e^{\|A\|t} \quad \text{where } n \text{ is the order of the matrix.}$$

$$(1.5) \quad e^{A+B} = e^A e^B \quad \text{if and only if } A \text{ and } B \text{ commute.}$$

$$(1.6) \quad e^{J^{-1}FJ} = J^{-1} e^{FJ}.$$

$$(1.7) \quad \text{Determinant } e^{Ft} = e^{t \text{ trace } F}.$$

The last theorem shows that e^{tF} is always nonsingular, moreover, the columns of e^{tF} are n linearly independent solutions of (1.2). Thus any solution of (1.2) can be obtained by a linear combination of the column vectors of e^{tF} .

2. Program Algorithm.

For computing e^{tF} , the sum of at most the first thirty-six terms of its defining series (1.1) is used. Thus we compute

$$(2.1) \quad \sum_{i=0}^{35} \frac{t^i F^i}{i!} \approx e^{tF}.$$

The sum (2.1) is actually computed as follows. Let T_i be the i th term of the expansion: $T_0 = I$, $T_1 = Ft$, etc. The sum is accumulated and T_{i+1} is obtained from T_i by multiplication by

$$\frac{tF}{i+1}.$$

The following motivates why thirty-six terms are used in (2.1), and gives a condition under which the result can be expected to be accurate.

In the IBM 709 and 7090 a little more than eight significant digits are carried when operating in the single-precision, floating point mode. This imposes limitations on the accuracy of the program. Consider a simple cosine series. We know that for any value of the argument the absolute value of the function is one or less. Yet if the argument were 20, the term $\frac{20^{20}}{20!}$ is so large that the addition to it of a number of the order of 1 in magnitude does not effect it. That is, if any term of a series exceeds 10^8 , we know that no answer of the order of 1 or less in magnitude will in general be correct. Thus if we want an answer that is correct to four decimal places, no component of the sum may exceed 10^4 . The largest term of the e^t series is T_j where j is the smallest integer such that $\frac{t}{j+1} < 1$; therefore $T_j = \frac{t^j}{j!}$ where $j = [t]$, the greatest integer less than t . For our purposes $\frac{t^j}{j!}$ should always be less than 10^4 , which implies that t should be less than 10; since $\frac{10^{10}}{10!} \approx 10^4$.

Without attempting to discuss the problem more rigorously, it was decided that if $\|F\| \cdot |t| < 10$, then the answers could be depended upon to four decimal places.

Further, since $\frac{10^{35}}{35!} \approx 10^{-35}$, it is evident that no more than thirty-six terms need be carried. On the other hand, if $\frac{t^{35}}{35!}$ is greater than 10^{-5} then $\frac{t^{10}}{10!}$ was greater than 10^4 , and the validity of the answer is open to question anyway. Unfortunately, there is no error return if such a condition occurs. Error return occurs when one of the T_i is greater than 10^{40} (machine overflow); computation of (2.1) stops when a term T_i is less than 10^{-40} (underflow).

3. Checks.

(A) e^{tA} was computed for the 15×15 nilpotent matrix

$$\begin{bmatrix} 0_{14} & I_{14} \\ 0 & 0'_{14} \end{bmatrix} \quad (\text{where } 0_{14} \text{ is the 14-dimensional zero column vector and } I_{14} \text{ is the } 14 \times 14 \text{ identity matrix})$$

for $t = .1, 1., 5.,$ and $10.$ For all these values of $t,$ this checked to six significant figures and approximately thirteen decimal places.

(B) Choosing at "random" a 15×15 matrix F such that $\|F\| = 91$ and

$t = 0.10989016,$ e^{tA} and $((e^{\frac{tA}{8}})^2)^2)^2$ were computed and the results printed to

eight significant figures. It was felt that $((e^{\frac{tA}{8}})^2)^2)^2$ was probably quite close to the exact value of $e^{tA}.$ Comparison showed the two results to be the same to about six decimal places, with the smaller elements losing accuracy, being correct to only four or five significant digits.

For the matrices involved in this check see Figure 1.

(C) The exponential was computed for the 7×7 diagonal matrix $\text{diag}(-10, -4, -1, 0, 1, 4, 10)$ with $t = 1.$ The result is in Fig. 2.

Because this matrix is diagonal, the scalar analysis used in Sect. 2 applied exactly and the answers accord with this very well. In the submatrix (-10) where we were not only at the limit of the acceptable range of application but were taking differences, we barely managed to have accuracy in four decimal places. In fact if the answer had been printed in the four-place rounded format which we use, rounding would have produced the wrong answer. However, where differences were not being taken, as in the submatrix $(10),$ the answer is correct to seven significant figures, which again we expect. The same results are true, with decreasing significance, of the submatrices (4) and (-4) and (1) and $(-1).$

In general however, the accuracy seems adequate for our purposes.

4. References.

E. A. CODDINGTON and N. LEVINSON (1955) "Theory of ordinary differential equations, (book)", McGraw-Hill, 1955.

ADAPTIVE SYSTEMS

INPUT MATRIX

NUMBER OF ROWS IS

NUMBER OF COLUMNS IS EXPONENT = 0

-8.0000	1.0000	6.0000	7.0000	-7.0000	-1.0000	-4.0000	1.0000	-2.0000	1.0000	-6.0000	2.0000	0.0000	-8.0000	3.0000
4.0000	9.0000	5.0000	1.0000	-4.0000	7.0000	0.0000	-1.0000	-3.0000	2.0000	3.0000	1.0000	4.0000	-7.0000	0.0000
-3.0000	-7.0000	-2.0000	-5.0000	-6.0000	-8.0000	-9.0000	4.0000	-2.0000	5.0000	9.0000	-4.0000	7.0000	-9.0000	2.0000
-6.0000	7.0000	-9.0000	0.0000	-1.0000	7.0000	9.0000	9.0000	5.0000	6.0000	-1.0000	0.0000	-8.0000	-8.0000	-4.0000
0.0000	-2.0000	-1.0000	-7.0000	4.0000	-8.0000	9.0000	9.0000	-2.0000	-4.0000	8.0000	-5.0000	-3.0000	-1.0000	-1.0000
-6.0000	1.0000	-2.0000	1.0000	-2.0000	4.0000	7.0000	7.0000	-3.0000	7.0000	-9.0000	3.0000	-3.0000	-2.0000	-2.0000
0.0000	0.0000	8.0000	6.0000	3.0000	-8.0000	-1.0000	8.0000	-5.0000	-6.0000	-9.0000	8.0000	4.0000	5.0000	8.0000
0.0000	-3.0000	6.0000	5.0000	0.0000	7.0000	0.0000	-7.0000	-1.0000	8.0000	-8.0000	-8.0000	7.0000	-9.0000	-2.0000
-1.0000	-5.0000	7.0000	-4.0000	-8.0000	2.0000	-1.0000	-8.0000	-2.0000	-3.0000	-5.0000	7.0000	3.0000	6.0000	-8.0000
2.0000	8.0000	0.0000	3.0000	4.0000	-2.0000	-4.0000	3.0000	1.0000	0.0000	-4.0000	6.0000	-9.0000	1.0000	9.0000
1.0000	9.0000	-7.0000	1.0000	-2.0000	-6.0000	-7.0000	-4.0000	-6.0000	0.0000	-2.0000	-4.0000	-5.0000	0.0000	-8.0000
4.0000	1.0000	-7.0000	-6.0000	-2.0000	6.0000	2.0000	5.0000	-2.0000	1.0000	8.0000	-3.0000	-3.0000	7.0000	9.0000
-6.0000	-4.0000	-6.0000	-5.0000	-7.0000	9.0000	7.0000	2.0000	-5.0000	6.0000	2.0000	-2.0000	3.0000	-4.0000	6.0000
5.0000	3.0000	-2.0000	4.0000	0.0000	6.0000	2.0000	9.0000	-1.0000	9.0000	6.0000	7.0000	9.0000	6.0000	-8.0000
-9.0000	-8.0000	-5.0000	0.0000	-9.0000	-7.0000	-9.0000	-9.0000	7.0000	-8.0000	5.0000	-7.0000	7.0000	-8.0000	9.0000

THIS IS A CHECK RUN FOR ENTRY
 ROW 7 = 91. WE WILL COMPUTE EXP(11) WITH
 0.10989011 IN TWO WAYS, BY DIRECT CALCULATION
 AND BY CALCULATING WITH T = 0.01372628 AND
 SQUARING THE RESULT THREE TIMES.

FIGURE 1-A. INPUT MATRIX

ADAPTIVE SYSTEMS

INPLT MATRIX DIR

NUMBER OF ROWS IS NUMBER OF COLUMNS IS

-0.22218787E-00	-0.50328908E-00	0.1671935E-00	0.2777078E-00	-0.10211703E-01	-0.23310187E-00
-0.8925972E-00	-0.75222511E-00	0.85416301E-00	0.26775208E-00	-0.47217357E-00	-0.4315063E-00
-0.2378480E-00	0.12235492E-01	0.6828078E-01	-0.7711732E-00	0.19807809E-01	-0.52705599E-01
-0.53714028E-01	-0.2813260E-01	0.31426548E-00	-0.10201598E-01	-0.25660017E-01	0.59380366E-00
-0.26078590E-00	0.56793280E-00	-0.8121133E-00	0.12635527E-01	-0.26727775E-01	0.26209917E-01
-0.63749611E-00	-0.10550769E-01	0.5976491E-00	-0.32662622E-00	-0.37251966E-00	-0.10805725E-01
-0.13991890E-01	0.70786273E-02	-0.17882422E-01	0.49373668E-00	0.33026599E-00	-0.19966970E-01
0.11258169E-01	-0.13641365E-01	-0.1336091E-01	-0.1264963E-01	0.13572391E-01	0.16525400E-01
0.82604189E-00	0.15718372E-00	-0.10999392E-00	0.43320397E-00	-0.50110895E-00	0.13266493E-01
0.71302804E-00	-0.66169764E-00	0.11764736E-01	-0.1888550E-01	-0.49181569E-00	0.17279337E-01
0.76578731E-00	-0.46292203E-00	0.69310536E-01	-0.63339405E-01	0.19867356E-01	-0.28163870E-01
-0.43581645E-00	0.29317627E-00	-0.83007394E-00	-0.79089652E-00	0.14591961E-01	-0.16854889E-01
-0.51300172E-00	-0.20994175E-00	-0.13250230E-01	-0.12891741E-00	-0.2838749E-01	-0.16383570E-01
-0.19758335E-00	0.33023073E-00	0.78367675E-00	0.58880378E-00	-0.54703908E-00	0.29904800E-00
-0.12828694E-00	-0.49599066E-00	0.1085587E-01	-0.10166109E-01	0.53715835E-00	0.15185814E-01
-0.41315345E-00	-0.20334435E-01	0.90864801E-01	-0.3194121E-00	-0.25478170E-01	-0.12076566E-01
0.4435385E-00	0.81242519E-00	-0.74739700E-01	-0.12221251E-01	-0.83228344E-01	-0.10242563E-01
0.13272933E-01	-0.21352696E-00	0.14524336E-00	-0.75843833E-00	-0.16416346E-01	0.37360784E-01
-0.3070297E-00	0.9535813E-00	-0.4346872E-00	0.6786897E-00	0.53307853E-00	0.4721024E-00
0.11693773E-00	-0.13372750E-01	-0.8004948E-00	0.31673691E-00	-0.88060749E-00	0.4804329E-00
0.2393706E-00	-0.40063692E-00	0.14138429E-00	-0.2074882E-00	-0.47535753E-00	0.15580810E-01
0.6513068E-00	0.57643313E-00	-0.19727424E-00	0.21532109E-00	-0.24523984E-00	0.14061531E-01
0.74020771E-00	0.19134433E-01	-0.1284562E-01	0.4567658E-00	0.14028591E-01	-0.14547885E-00
0.2707327E-00	0.12007973E-01	-0.7751971E-00	-0.20091133E-01	-0.19389360E-01	0.10564943E-01
0.55237102E-00	0.11400916E-01	0.4979344E-00	-0.9458749E-00	-0.44087836E-00	0.15980787E-01
0.46985768E-00	0.32468721E-01	0.17885982E-00	0.96040737E-00	-0.10531916E-01	0.50844105E-00
-0.11678861E-01	-0.97055934E-00	-0.80497538E-01	0.1378928E-01	0.44602749E-00	0.29572239E-00
-0.12086219E-01	0.12394713E-01	0.56586181E-00	0.5224095E-00	0.16456747E-01	-0.79165759E-00
0.30552207E-00	0.15975265E-01	0.1227037E-01	-0.64813795E-00	-0.13673109E-01	-0.12886042E-01
-0.20009608E-00	0.96847241E-00	0.16283748E-01	-0.38880190E-00	0.74516223E-00	0.13180646E-01
-0.12759680E-01	0.19324139E-01	-0.8362387E-00	-0.6919589E-00	0.3593550E-00	0.15074947E-01
0.24585696E-01	0.10637777E-01	-0.10090399E-01	-0.12223542E-01	-0.14493091E-01	-0.53236385E-00
0.25752135E-01	0.16985682E-00	-0.49826385E-01	0.1790834E-00	0.1592336E-01	-0.57323647E-00
0.92961971E-00	0.27378657E-00	0.3202764E-01	0.22455384E-01	0.12035149E-01	-0.77209575E-00
-0.10437836E-01	-0.11380778E-01	0.17860748E-01	-0.14216337E-00	0.12409910E-01	0.72633759E-00
-0.84038093E-00	-0.21766149E-01	-0.90510622E-00	-0.33265311E-00	-0.14670154E-01	0.44325139E-01
-0.44025085E-00	-0.68222337E-00	0.62742636E-00	-0.17698197E-01	0.2316563E-00	-0.69375186E-00
0.19543276E-01	0.72767679E-00	-0.13683879E-01			

THIS IS THE MATRIX AS OBTAINED DIRECTLY.

FIGURE 1-B. EXPONENTIAL OBTAINED DIRECTLY

ADAPTIVE SYSTEMS

INPUT MATRIX CSC

NUMBER OF ROWS IS NUMBER OF COLUMNS IS

-0.22518678E-00	-0.50328709E-00	0.18917927E-00	0.27770473E-00	-0.10216775E-01	-0.23580771E-00
-0.80259911E-00	-0.75622462E-00	0.85416240E-00	0.26775923E-00	-0.47217313E-00	-0.43415036E-00
0.27985568E-00	-0.12235452E-01	0.84828897E-00	-0.77171409E-00	0.19467863E-01	-0.51785708E-01
-0.53714649E-01	-0.28132580E-01	0.31426515E-00	-0.18201524E-01	-0.25660798E-01	0.59358340E-00
-0.26075888E-00	0.58792555E-00	-0.21211285E-00	0.13235578E-01	-0.287277150E-01	0.28209895E-01
-0.63745567E-00	-0.10550760E-01	0.59764947E-00	-0.32462591E-00	-0.37257919E-00	-0.10805171E-01
-0.13991880E-01	0.70726830E-02	-0.17882317E-01	0.49373635E-00	0.33828566E-00	-0.19986986E-01
0.11258162E-01	-0.15641555E-01	-0.13360934E-01	-0.12648863E-01	0.11572478E-01	0.16525399E-00
0.82760059E-00	0.15718347E-00	-0.10998055E-00	0.43320487E-00	-0.50111166E-01	0.13286852E-01
0.76376685E-00	-0.66169697E-00	0.11786728E-01	-0.18885534E-01	-0.49180505E-00	0.17279325E-01
-0.43581615E-00	-0.82927150E-00	0.69310510E-00	-0.63339330E-00	0.19407357E-01	-0.28165846E-01
-0.51300158E-00	0.99317554E-00	-0.83007329E-00	-0.79089520E-00	0.24591949E-01	-0.16354478E-01
-0.19758318E-00	-0.20995170E-00	-0.13250220E-01	-0.12891719E-00	-0.23387760E-01	-0.16365357E-00
0.28285674E-00	0.35073034E-00	0.78369530E-00	0.58880319E-00	-0.54703875E-00	0.29904849E-00
-0.41315329E-00	-0.20334421E-01	0.90864707E-01	-0.10168103E-01	0.53715748E-00	0.15185803E-01
0.24445520E-00	0.81262467E-00	-0.14738855E-01	-0.12221244E-01	-0.83224633E-01	-0.12076557E-01
0.13272926E-01	-0.21352690E-00	0.14324327E-00	-0.75243580E-00	-0.25478836E-01	-0.10362558E-01
-0.10109285E-00	0.93935747E-00	-0.83608706E-00	0.69868854E-00	-0.16416334E-01	0.57306789E-01
0.11693760E-00	-0.15372742E-01	-0.80049512E-00	0.31673086E-00	0.53507435E-00	0.89232592E-00
0.23993593E-00	-0.40063633E-00	0.14739477E-00	-0.20748818E-00	-0.87535713E-00	0.15540807E-01
0.64130649E-00	0.57643276E-00	-0.19727417E-00	0.21532100E-00	-0.24523962E-00	0.14061523E-01
0.74080715E-00	0.19134619E-01	-0.12459531E-01	0.45676556E-00	0.14028519E-01	-0.14547883E-00
0.27073256E-00	-0.12007965E-01	-0.77519122E-00	-0.20091141E-01	-0.34087800E-00	0.10568937E-01
0.52370650E-00	0.17400970E-01	0.89793272E-00	-0.98587495E-00	-0.40878934E-01	0.15980756E-01
0.44985734E-00	0.35468695E-01	0.17885967E-00	0.90040679E-00	-0.10537907E-01	0.50844075E-00
-0.12086212E-01	-0.97055870E-00	-0.80497453E-01	0.15789856E-01	0.84682191E-00	0.29572228E-00
0.30952177E-00	-0.15975255E-01	0.12270828E-01	-0.65815763E-00	-0.13675702E-01	-0.12788007E-01
-0.20009596E-00	0.96447124E-00	0.16283741E-01	-0.38880181E-00	0.24516179E-00	0.13180638E-01
-0.12754451E-01	-0.19324724E-01	-0.83627744E-00	-0.69196513E-00	0.35939339E-00	0.15074933E-01
0.24585681E-01	0.10637772E-01	-0.10090394E-01	-0.12223536E-01	-0.14493082E-01	-0.53236373E-00
0.25752120E-01	0.16985684E-00	-0.99826287E-01	0.17900027E-00	0.15792724E-01	-0.57323629E-00
0.92961988E-00	0.27378440E-00	0.32032734E-01	0.22845564E-01	0.12035140E-01	-0.77209528E-00
0.10437830E-01	-0.11380769E-01	0.19880735E-01	-0.19216533E-00	0.12909899E-01	0.72653734E-00
-0.84038035E-00	-0.21766132E-01	-0.90510551E-00	-0.33245288E-00	-0.14670143E-01	0.44325259E-01
-0.44025047E-00	-0.648222295E-00	0.62767554E-00	-0.17698207E-01	0.23136938E-00	-0.695757153E-00
0.19545264E-01	0.78767829E-00	-0.13681872E-01			

THIS IS THE MATRIX AS OBTAINED BY SQUARING.

FIGURE 1-G. EXPONENTIAL OBTAINED BY SQUARING

ADAPTIVE SYSTEMS

INPUT MATRIX USE

NUMBER OF ROWS IS NUMBER OF COLUMNS IS

-0.12852252E-06	-0.82312206E-06	0.1776335E-06	0.3837045E-06	-0.83440503E-06	-0.15273690E-06
-0.40349701E-06	-0.47173832E-06	0.61094761E-06	0.24959845E-06	-0.44330955E-06	-0.627194619E-06
0.10803342E-06	-0.70281801E-06	0.50663948E-06	-0.52899122E-06	0.24733928E-05	0.11501838E-06
0.24214387E-07	-0.35033951E-05	0.53644180E-06	-0.14007092E-05	-0.10774463E-05	0.26077032E-06
-0.20976632E-07	0.65233135E-06	-0.51460008E-06	0.90897633E-06	-0.23543033E-06	0.21755695E-05
-0.43958426E-06	-0.87916851E-06	0.33527613E-06	-0.30919909E-06	-0.46938658E-06	-0.78974154E-06
-0.78281801E-05	-0.57666524E-07	0.9778870E-08	0.3129239E-06	0.32410626E-06	-0.11622008E-05
0.48545341E-06	-0.92367199E-06	-0.74505806E-06	-0.1389989E-06	0.13560057E-05	0.11362135E-06
0.11544400E-05	0.2331974E-06	0.79651513E-07	0.65565169E-06	-0.23847858E-06	0.49837780E-06
0.68310167E-06	-0.57055225E-06	0.73015690E-06	-0.16391277E-05	-0.64074993E-06	0.12964010E-06
0.67055275E-06	-0.57526593E-06	0.4544852E-06	-0.7525086E-06	0.20563702E-05	-0.24437904E-05
-0.3074851E-06	0.72270632E-06	-0.64820051E-06	-0.71525574E-06	0.12218952E-05	-0.11026859E-05
-0.4023135E-06	-0.50269419E-07	0.723E7199E-06	-0.21975273E-06	0.10943690E-07	-0.10430813E-06
-0.1713635E-06	0.37370490E-06	0.10430813E-06	0.58859587E-06	-0.33527413E-06	0.15273690E-06
0.1955774E-06	-0.6072232E-06	0.67055225E-06	-0.65565169E-06	0.36507845E-06	0.98347664E-06
-0.20116588E-06	-0.61341045E-05	0.94994903E-07	-0.23847858E-06	0.60303137E-07	-0.90897083E-06
0.25331974E-06	0.3746000E-08	-0.7543779E-07	-0.7003548E-06	-0.10244548E-07	-0.46193600E-06
0.66545341E-06	-0.62917871E-07	0.65406987E-07	-0.52899122E-06	-0.11026859E-05	-0.51222742E-08
-0.12685987E-06	0.71525574E-06	-0.14901051E-06	0.43213387E-06	0.20115588E-05	0.53155084E-06
0.12293458E-06	-0.84936619E-06	-0.36507845E-06	0.44703484E-07	-0.67055225E-06	0.30919909E-06
0.13224781E-06	-0.58859587E-06	0.1173468E-06	-0.24028122E-06	-0.37623432E-05	0.99837780E-06
0.15767787E-06	0.34507845E-06	0.68917871E-07	0.87406967E-07	0.22351742E-05	0.7003548E-06
0.58193516E-06	0.11709038E-05	-0.923E7199E-06	0.3129239E-06	0.11622008E-05	-0.24214387E-07
0.33096200E-06	0.7003548E-06	-0.58859587E-06	-0.11622008E-05	-0.11324863E-05	0.52154064E-06
0.36507845E-06	0.25254877E-06	0.1736335E-06	-0.52899122E-06	-0.34272677E-06	0.11324863E-06
0.3129239E-06	0.25928020E-06	0.15459958E-06	0.78231656E-06	-0.81956387E-06	0.30547139E-06
-0.71525574E-06	-0.64074993E-06	-0.71711838E-07	0.7876153E-06	0.55133298E-06	0.11920929E-06
-0.44074993E-06	0.92327199E-06	0.40978193E-06	0.3129239E-06	0.17473894E-05	-0.37997961E-06
0.16018748E-06	-0.10221801E-05	0.7897615E-06	-0.37037451E-06	-0.64074993E-06	-0.35739504E-07
-0.11920929E-06	0.45193600E-06	0.7003548E-06	-0.8568167E-07	0.44703484E-06	0.77486038E-06
-0.92367199E-06	-0.14305115E-05	-0.6329935E-06	-0.5587935E-06	0.10803342E-06	0.14097992E-05
0.14901161E-05	0.44703484E-06	-0.55133298E-06	-0.67055225E-06	-0.95367432E-06	-0.12685987E-06
0.84406967E-06	0.16391277E-06	-0.9778870E-07	0.12479722E-06	0.12768010E-05	-0.3797461E-06
0.6565109E-06	-0.84936619E-06	0.25928020E-06	0.1519913E-05	0.80936619E-06	-0.46938658E-06
-0.58114329E-06	-0.17285347E-06	0.12945010E-05	-0.3725293E-07	0.1175871E-05	0.24546816E-05
-0.4723231E-06	-0.4723231E-06	-0.69290400E-05	-0.22724271E-06	-0.11920929E-05	-0.11920929E-06
0.11920929E-05	0.50663948E-06	-0.67055225E-06	-0.11473894E-05	0.25704503E-06	-0.3354713E-06

THIS IS THE DIFFERENCE MATRIX.

END OF FILE CONDITIONS ENCOUNTERED ATTEMPTING TO READ DATA. PROGRAM RETURNS TO MONITOR.

FIG. 1-D. DIFFERENCE MATRIX

5.3792528E-05	0	0	0	0	0	0	0
0	1.8315654E-02	0	0	0	0	0	0
0	0	-3.6787941E-01	0	0	0	0	0
0	0	0	0.9999999E 00	0	0	0	0
0	0	0	0	2.7132817E 00	0	0	0
0	0	0	0	0	5.4598145E 01	0	0
0	0	0	0	0	0	2.2028461E 04	0

FIG. 2-A

Computed Value of Matrix Exponential

4.599930E-05	0	0	0	0	0	0	0	0	0
0	1.831569E-02	0	0	0	0	0	0	0	0
0	0	3.678944E-01	0	0	0	0	0	0	0
0	0	0	0	1.0	0	0	0	0	0
0	0	0	0	0	2.718281E	0	0	0	0
0	0	0	0	0	0	5.4998150E 01	0	0	0
0	0	0	0	0	0	0	0	2.2026465E 04	0

Fig. 2-E

Correct Value of Matrix Exponential

4.5599950E-05	0	0	0	0	0	0	0	0	0
0	1.8315659E-02	0	0	0	0	0	0	0	0
0	0	3.678944E-01	0	0	0	0	0	0	0
0	0	0	1.0	0	0	0	0	0	0
0	0	0	0	2.7182818	0	0	0	0	0
0	0	0	0	0	5.4598150E 01	0	0	0	0
0	0	0	0	0	0	0	0	2.2026466E 04	0

Fig. 2-B

Correct Value of Matrix Exponential

Chapter 9

THE INTEGRAL EXPONENTIAL SUBROUTINE

1. Theory.

The concrete definition of the integral exponential

$$(1.1) \quad \int_0^t e^{\tau F} d\tau$$

can be obtained by integrating term-by-term the series defining the matrix exponential:

$$\int_0^t e^{\tau F} d\tau = \sum_{i=0}^{\infty} \frac{F^i t^{i+1}}{(i+1)!}$$

If F^{-1} exists, this may be written as

$$F^{-1}(e^{\tau F} - I).$$

Notice, however, that the matrix (1.1) does not depend for its existence upon the existence of F^{-1} .

In Chapter 8 it was stated that the solution of a free system

$$\frac{dx}{dt} = Fx$$

may be expressed as

$$x(t) = e^{(t - t_0)F} x(t_0).$$

Often, however, we are interested in a controlled system, which in the linear, constant-coefficient case may be written as:

$$(1.2) \quad \dot{x} = Fx + G u(t) \quad (F, G \text{ constant}).$$

The solution to (1.2) for $x(0) = 0$ may be written as

$$x(t) = \int_0^t e^{(t-\tau)F} G u(\tau) d\tau.$$

To obtain a complete system of solutions when \underline{u} is constant, replace the vector \underline{u} by the identity matrix. Then

$$\Gamma(t) = \int_0^t e^{(t-\tau)F} G d\tau.$$

Making the substitution $\tau' = t - \tau$, the integral assumes the simpler form:

$$\Gamma(t) = \int_0^t e^{\tau'F} d\tau' G.$$

This is the matrix which the subroutine "Integral Exponential" computes.

2. Program Algorithm.

The terms T_i of the defining matrix are obtained as they are for the matrix exponential

$$T_{i+1} = T_i \frac{tF}{i+1};$$

however, $T_0 = tI$, not 1.

The arguments concerning permissible range of t given in Chapter 8, Sect. 2 are applicable here also.

3. Checks.

A) The integral exponential was computed for the 15×15 nilpotent matrix

$$\begin{bmatrix} 0_{14} & I_{14} \\ 0 & 0'_{14} \end{bmatrix}$$

(where 0_{14} is the 14-dimensional zero column vector and I_{14} is the 14×14 identity matrix) for $t = .1, 1., 5.,$ and $10.$ For all values of t , this checked to six significant figures and approximately thirteen decimal places.

B) The integral exponential was computed for the 7×7 matrix

$\text{diag}(-10, -4, -1, 0, 1, 4, 10)$

with $t = 1$. The result is seen in Fig. 1.

9.9992417E-01	0	0	0	0	0	0	0	0
0	2.4542104E-01	0	0	0	0	0	0	0
0	0	6.3212050E-01	0	0	0	0	0	0
0	0	0	0.9999999E-00	0	0	0	0	0
0	0	0	0	1.7182817E-00	0	0	0	0
0	0	0	0	0	0	1.3399536E 01	0	0
0	0	0	0	0	0	0	2.2025461E 04	0

Fig. 1

Computed Value of Integral Exponential

Chapter 10

THE TRANSIENT PROGRAM

1. Theory.

The Transient Program is designed to give a time history of the state vector y of the system:

$$(1.1) \quad \begin{aligned} \dot{x} &= Fx + Gu \\ u &= Jr(t) - Kx \\ y &= Hx. \end{aligned}$$

This problem, as explained in the first Sections of Chapters 8 and 9 has the solution:

$$(1.2) \quad x(t_0 + t) = e^{t(F - GK)}x(t_0) + \int_{t_0}^t e^{\tau(F - GK)}GJ r(\tau)d\tau.$$

Under the assumption that r is a constant vector, the solution (1.2) may be written stepwise as follows:

$$x(t_0 + T) = e^{T(F - GK)}x(t_0) + \int_0^T e^{\tau(F - GK)}d\tau GJr$$

where T is the sampling period.

This enables us, as in the Riccati Program, to obtain a step-wise sampling of the analytic solution of the problem, after computing the exponential and integral exponential of $T(F - GK)$ only once.

2. Program Algorithm.

The equations which have been mechanized are the following:

$$\begin{aligned} x(t + T) &= \Phi x(t) + \Gamma u \\ u &= Ar - Ex \\ y &= Hx. \end{aligned}$$

Where the matrices Φ , Γ , J , Λ , Σ are inputs, as well as T , $x(t_0)$, N , the maximum number of steps, and another sampling period $\tau = Tk$, k an integer, the use of which will be explained below.

Input might be, for instance to solve the system (1.1)

$$\dot{\phi} = e^{T(F - GK)}$$

$$\Gamma = \int_0^T e^{\tau(F - GK)} d\tau G$$

$$\Lambda = J, \Sigma = 0.$$

It is apparent that if r is a constant and the inputs are as they appear above, then $u = Jr$, a constant vector, and the equation for u is superfluous in the stepwise procedure.

To accommodate the case when r is a piecewise constant (sampled) function of time, the second, τ , sampling period has been provided. Every k steps of length T in the computation of x , the value of r is used to compute a new value of u . If r is a constant this offers a saving in machine time since choice of a k greater than N will prevent reference to the second equation except at the beginning of the run. In this way, it is easy to find the response of a sampled-data system between sampling points. As in the Riccati Program, observe that this is not a stepwise integration procedure; the only errors are round-off in the computer and whatever is involved in assuming r to be sufficiently well approximated by a step function.

Another feature of the program is that not the state variables but linear combinations thereof, are printed. In several problems which have been considered, e.g., one concerning satellite orientation, the state variables could not be observed, but only linear combinations of the state variables. To be able to compare machine results with experimental data, the Transient Program has an additional input matrix H and the vector of observables $y = Hx$ is printed.

At every interval T the complete output is printed. This consists of the time, two components of the r vector, seven components of y , and three components of u . The problem is terminated by exceeding the maximum number of steps N . Because of programmed space limitations N must be less than 200.

Chapter 11

THE MATRIX RICCATI EQUATION

1. Theory.

A system of $2n$ linear differential equations is said to be hamiltonian (or canonical) if it can be written in the form:

$$(1.1) \quad \dot{x}_i = dx_i/dt = \frac{\partial \mathcal{H}}{\partial p_i}, \quad \dot{p}_i = dp_i/dt = -\frac{\partial \mathcal{H}}{\partial x_i}, \quad i = 1, \dots, n.$$

where \mathcal{H} , the hamiltonian, is a homogeneous quadratic polynomial in the x_i and p_i with coefficients which are functions of time.

The most general such function may be written as

$$(1.2) \quad 2\mathcal{H} = [x, y]H \begin{bmatrix} x \\ y \end{bmatrix},$$

where H is a symmetric matrix. H can be partitioned as

$$(1.3) \quad H = \begin{bmatrix} A & C \\ C' & -B \end{bmatrix}$$

where A and B are symmetric, C is arbitrary; the negative sign before B is purely for our later convenience. All these matrices may be functions of time.

Substituting (1.2-) into (1.1), we get

$$(1.4) \quad \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = Z \begin{bmatrix} x \\ y \end{bmatrix}.$$

This equation possesses an important kind of symmetry which we can state as follows:

(1.5) THEOREM. A $2n$ order system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = Z \begin{bmatrix} x \\ y \end{bmatrix}$$

is hamiltonian if and only if

$$Z = JZ'J$$

where

$$J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}^*$$

Proof. Only if: This follows from the fact that the $2n$ order system (1.4) above -- which was derived from the most general possible hamiltonian function -- satisfies the condition.

If: Consider

$$Z = \begin{bmatrix} C & -B \\ -A & D \end{bmatrix}$$

where A, B, C, D are arbitrary. Then

$$JA'J = \begin{bmatrix} -E' & C' \\ D' & -B' \end{bmatrix} = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

$$JZ'J = \begin{bmatrix} -D' & -B' \\ -A' & -C' \end{bmatrix} = \begin{bmatrix} C & -B \\ -A & D \end{bmatrix},$$

* $I = n \times n$ identity matrix and $0 = n \times n$ zero matrix.

and $A = A'$, $B = B'$, and $D = -C'$. But these are just the conditions satisfied by the matrix Z in (1.4) above, and it is apparent that the form of Z determines that the system is hamiltonian.

(1.6) COROLLARY. If λ_1 is an eigenvalue of Z so is $-\lambda_1$.

Proof. $Zx = \lambda_1 x$ implies $JZ'Jx = \lambda_1 x$ and, since $J^{-1} = -J$, we have $Z'Jx = -\lambda_1 Jx$.

(1.7) THEOREM. Let

$$\theta(t, t_0) = \begin{bmatrix} \theta_{11}(t, t_0) & \theta_{12}(t, t_0) \\ \theta_{21}(t, t_0) & \theta_{22}(t, t_0) \end{bmatrix}$$

$$(1.8) \quad \begin{bmatrix} \theta_{11}(t_0, t) & \theta_{12}(t_0, t) \\ \theta_{21}(t_0, t) & \theta_{22}(t_0, t) \end{bmatrix} = \begin{bmatrix} \theta_{22}(t, t_0) & -\theta_{12}(t, t_0) \\ -\theta_{21}(t, t_0) & \theta_{11}(t, t_0) \end{bmatrix}$$

Proof. If $\theta(t, t_0)$ is the transition matrix of $\dot{x} = Zx$, then, as is well-known, the transition matrix of $\dot{y} = Z'y$ is $\theta'(t_0, t)$. Note that $J' = J^{-1} = -J$. By Theorem (1.5)

$$Z' = -J'ZJ' = -J^{-1}ZJ' = J^{-1}ZJ.$$

Let $Jy = x$. Then $\dot{x} = J\dot{y} = JZ'y = -JZ'J^{-1}x = -JZ'J'x = JZ'Jx = Zx$, the last step following by Theorem (1.5). Hence

$$J'\theta'(t_0, t)J = \theta(t, t_0)$$

or

$$\theta(t_0, t) = J\theta'(t, t_0)J'$$

is identical with (1.8). Q. E. D.

We shall now explain the connection between the riccati equation and the hamiltonian system (1.4). Consider the n by n matrix $P(t)$ defined by:

$$(1.9) \quad P(t) = [\theta_{21}(t, t_0) + \theta_{22}(t, t_0)P(t_0)] [\theta_{11}(t, t_0) + \theta_{12}(t, t_0)P(t_0)]^{-1}$$

over an interval of time where

$$[\theta_{11}(t, t_0) + \theta_{12}(t, t_0)P(t_0)] \text{ is nonsingular.}$$

We now determine the differential equation satisfied by $P(t)$. Writing out in detail the relation $\dot{\theta} = Z\theta$ we get

$$\begin{aligned} \dot{\theta}_{11} &= C\theta_{11} + B\theta_{21}, \\ \dot{\theta}_{12} &= C\theta_{12} - B\theta_{22}, \\ \dot{\theta}_{21} &= A\theta_{11} - C'\theta_{21}, \\ \dot{\theta}_{22} &= -A\theta_{12} - C'\theta_{22}. \end{aligned}$$

Let us write $P_0 = P(t_0)$, and temporarily drop all arguments. Differentiating both sides of

$$(1.10) \quad P(\theta_{11} + \theta_{12}P_0) = \theta_{21} + \theta_{22}P_0$$

with respect to t and using the preceding relations, we get

$$\begin{aligned} \dot{P}(\theta_{11} + \theta_{12}P_0) + P[C\theta_{11} - B\theta_{21} + (C\theta_{12} - B\theta_{22})P_0] \\ = -A\theta_{11} - C'\theta_{21} - (A\theta_{12} + C'\theta_{22})P_0. \end{aligned}$$

Rearranging and using (1.9) and (1.10), we get the matrix riccati equation

$$(1.11) \quad -\dot{P} = -PB + C'P + P(-A - C'B) - A.$$

Choosing $P(t_0)$ to be symmetric guarantees that \dot{P} is symmetric and then $P(t)$ is always symmetric. Although formula (1.9) does not seem to be symmetric, it is, as can be verified using relations (1.8).

In the special case where the riccati equation is linear ($B = 0$):

$$\dot{\theta}_{12} = C\theta_{12}$$

But $\theta_{12}(t_0, t_0) = 0$ because $\theta(t_0, t_0) = I$, therefore $\theta_{12} = 0$. Then

$$P(t) = [\theta_{21}(t, t_0) + \theta_{22}(t, t_0)P(t_0)]\theta_{11}^{-1}(t, t_0).$$

But from equation (1.8) we see that $\theta_{11}^{-1} = \theta_{22}$. and

$$(1.12) \quad P(t) = [\theta_{21}(t, t_0) + \theta_{22}(t, t_0)P(t_0)]\theta_{22}(t, t_0) \quad \text{if } B = 0.$$

Thus we have proved:

(1.13) **THEOREM.** The solution of the riccati equation (1.11) may be expressed in terms of the transition matrix of the hamiltonian system (1.4) if $B = 0$, the solution always exists and is given by (1.12). If $B \neq 0$, the solution is given by (1.9), in an interval of time where $[\theta_{11}(t, t_0) + \theta_{12}(t, t_0)P_0]$ is nonsingular.

The last condition rules out the so-called "conjugate points" of the calculus of variations. An example where a conjugate point does occur was given in Sect. 7 of Chapter 2.

In case of the riccati equation (4.6) of Chapter 2, the matrix Z has the form

$$(1.14) \quad Z = \begin{bmatrix} F & -GR^{-1}G' \\ -H'QH & F' \end{bmatrix}$$

In the case of the riccati equation (3.3) of Chapter 4, the matrix Z has the form

$$(1.15) \quad Z = \begin{bmatrix} -F' & -H'R^{-1}H \\ -GG' & F \end{bmatrix}.$$

2. Program Algorithm.

The transition matrix of (1.4) can be computed easily only when Z is a constant matrix. In that case,

$$(2.1) \quad \Theta(t, t_0) = \exp[(t - t_0)Z].$$

The program first computes (2.1) and then $P(t)$, by substituting the four submatrices of Θ into (1.9) or (1.2). In this way one obtains a step-wise solution of the riccati equation without any truncation errors, subject only to roundoff errors in computing Θ by means of the exponential subroutine of Chapter 8.

At each step $P(t)$ is symmetrized before proceeding to the next step by replacing it by

$$\frac{P(t) + P'(t)}{2}$$

Symmetrization is absolutely essential because otherwise uncontrollable roundoff errors may accumulate in the antisymmetric part of $P(t)$.

The input to the program consists of the matrices A, B, C (C arbitrary, A, B symmetric), a symmetric matrix $P(t_0)$, a sampling period T at which intervals the matrix P will be computed, a convergence criterion number ϵ , a maximum number of intervals N , and various printing codes.

The "sampling period" $T = t - t_0$ for the riccati equation may be arbitrary, subject only to the restriction that

$$\|T\| \cdot \|Z\| < 10$$

which is necessary for the convergence of the exponential subroutine (see Chapter 8). Since only the ratio of Q to R matters in (1.14) or (1.15), one should state these quantities so that $GR^{-1}G'$ and $H'QH$ are approximately of the same order of magnitude -- this will keep $\|A\|$ small.

This problem is terminated in one of two ways. Either the maximum number of steps is exceeded or the convergence criterion is satisfied.

The convergence criterion is that

$$\sum_{i=1}^n |p_{ii}(t+T) - p_{ii}(t)| / \sum_{i=1}^n p_{ii}(t+T)$$

be less than ϵ , an input number.

When $P(t)$ is computed, $K(t) = R^{-1}G'P(t)$ or $K'(t) = P(t)GR^{-1}$ can be also computed and printed, if desired.

Print controls enable the customer to print $K(t)$ and/or $P(t)$ at every step, every fifth step, only at the final step, or never.

As stated above, the Riccati Program must be provided with the matrices A , B , and C . If the customer does not care to pre-compute these, the Entrance to Riccati Routine, may be used with inputs G , H , Q , and R^{-1} . This program will compute $A = H'QH$, $B = GR^{-1}G'$, and $R^{-1}G'$.

3. Checks.

A) The Program was run with

$$F = \begin{bmatrix} 0_6 & I_6 \\ 0 & 0_6' \end{bmatrix}$$

where $0_6 =$ six dimensional zero column vector and $I_6 = 6 \times 6$ identity matrix.

$$G = \begin{bmatrix} 1 \\ 0_6 \end{bmatrix}$$

$Q = H = 7 \times 7$ zero matrix

$$R = [0.6075]$$

$$P(0) = 0.6075 I,$$

$$T = -0.2.$$

This was iterated for ten steps and the result compared with a hand-computed result expressed exactly in four-place decimals. The Riccati printout appears in Fig. 1. The hand computed result is:

$$P_{10} = \begin{bmatrix} 0.2025 & 0.4050 & 0.4050 & 0.2700 & 0.1350 & 0.0540 & 0.0180 \\ 0.4050 & 1.4175 & 2.0250 & 1.7550 & 1.0800 & 0.5130 & 0.1980 \\ 0.4050 & 2.0250 & 3.8475 & 4.1850 & 3.1050 & 1.7280 & 0.7650 \\ 0.2700 & 1.7550 & 4.1850 & 5.8275 & 5.4450 & 3.7170 & 1.9680 \\ 0.1350 & 0.0800 & 3.1050 & 5.4450 & 6.6375 & 5.8410 & 3.8730 \\ 0.0540 & 0.5130 & 1.7280 & 2.7170 & 5.8410 & 6.8319 & 5.9178 \\ 0.0180 & 0.1980 & 0.7050 & 1.9680 & 3.8730 & 5.9178 & 6.8623 \end{bmatrix}$$

2) The Program was run with

$$F = \begin{bmatrix} 0_6^1 & 0 \\ -I_6 & 0_6 \end{bmatrix} \quad H = [1 \quad 0_6^1]$$

$$G = R^{-1} = 7 \times 7 \text{ zero matrix}$$

$$Q = 2.0250$$

$$P(0) = 2.0250 I_7$$

$$\alpha = -0.2$$

This was iterated for ten steps and the result compared with a hand-computed result which was expressed exactly in four place decimals. The Riccati printout appears in Fig. 2. The hand-computed result is:

$P_{10} = 10.$

2.6935	-1.9758	1.2990	-0.6720	0.2790	-0.0900	0.0180
-1.9758	2.2869	-1.9710	1.2870	-0.6480	0.2430	-0.0540
1.2990	-1.9710	2.2725	1.9350	1.2150	-0.5400	0.1350
-0.6720	1.2870	-1.9350	2.1825	-1.7550	0.9450	-0.2700
0.2790	-0.6480	1.2150	-1.7550	1.8225	-1.2150	0.4050
-0.0900	0.2430	-0.5400	0.9450	-1.2150	1.0125	-0.4050
0.0180	-0.0540	0.1300	-0.2700	0.4050	-0.4050	0.2025

ADAPTIVE SYSTEMS

INPUT MATRIX A

NUMBER OF ROWS 7

NUMBER OF COLUMNS 7 EXPONENT = 0

0.20249997E-00	0.40499997E-00	0.40499997E-00	0.40499997E-00	0.26999997E-00	0.13499998E-00	0.53999998E-01
0.17299997E-01	0.40499997E-00	0.14174999E-01	0.20249997E-01	0.20249997E-01	0.17549998E-01	0.10799998E-01
0.51299991E-00	0.19789996E-00	0.40499997E-00	0.20249997E-01	0.20249997E-01	0.38474999E-01	0.41849994E-01
0.31049995E-01	0.17279997E-01	0.76499987E-00	0.26999997E-00	0.26999997E-00	0.17549998E-01	0.41849994E-01
0.58274991E-01	0.54499991E-01	0.37169994E-01	0.19679996E-01	0.19679996E-01	0.13499998E-00	0.10799998E-01
0.31049995E-01	0.54499991E-01	0.66374990E-01	0.58409990E-01	0.58409990E-01	0.38729993E-01	0.53999998E-01
0.51299991E-00	0.17279997E-01	0.37169994E-01	0.58409990E-01	0.58409990E-01	0.68318989E-01	0.59177990E-01
0.17299997E-01	0.12799996E-00	0.76499987E-00	0.19679996E-01	0.19679996E-01	0.38729993E-01	0.59177990E-01
0.69622997E-01						

FIG 1 FIRST CHECK PROBLEM

ADAPTIVE SYSTEMS

INPUT MATRIX A

NUMBER OF ROWS 7

NUMBER OF COLUMNS 7 EXPONENT = 0

0.263498E 02	-0.197579E 02	0.129899E 02	-0.671997E 01	0.278999E 01	-0.899997E 00
0.179999E 00	-0.197579E 02	0.228699E 02	-0.197099E 02	0.128699E 02	-0.647997E 01
0.242999E 01	-0.535998E 00	0.129899E 02	-0.197099E 02	0.227249E 02	-0.193499E 02
0.121499E 02	-0.535998E 01	0.134999E 01	-0.671997E 01	0.128699E 02	-0.193499E 02
0.218249E 02	-0.175499E 02	0.944998E 01	-0.269999E 01	0.278999E 01	-0.647997E 01
0.121499E 02	-0.175499E 02	0.182249E 02	-0.121499E 02	0.404999E 01	-0.899997E 00
0.242999E 01	-0.535998E 01	0.944998E 01	-0.121499E 02	0.101249E 02	-0.404999E 01
0.179999E 00	-0.535998E 00	0.134999E 01	-0.269999E 01	0.404999E 01	-0.404999E 01
0.202499E 01					

END OF FILE CONDITION ENCOUNTERED ATTEMPTING TO READ DATA. PROGRAM RETURNED TO MONITOR.

FINISH TIME 04.23.01E 02.18.61

FIG. 2 SECOND CHECK PROBLEM

FUNDAMENTAL STUDY OF ADAPTIVE CONTROL SYSTEMS

VOLUME I - APPENDICES

by

R. E. Kalman

NEW METHODS AND RESULTS IN LINEAR PREDICTION AND FILTERING THEORY*

by

R. E. Kalman

Research Institute for Advanced Studies

Baltimore 12, Md.

Table of Contents.

1. Introduction	111
2. Preview of Contents	113
3. Historical remarks and acknowledgements	119
4. Notation and other preliminaries	121
5. The gauss-markov sequence	125
6. The gauss-markov process	128
7. Axiomatic definition of the gauss-markov sequence and process.	133
8. A simple prediction problem	138
9. Statement and examples of the filtering problem	141
10. Other formulations of the filtering problem	145
11. Solution of the filtering problem for random sequences	148
12. Examples of discrete filtering	153
13. Solution of the filtering problem for random processes	162
14. Examples of continuous filtering	168
15. Minimal-variance unbiased estimation	194
16. Properties of the variance equation	208
Appendix A The pseudo-inverse of a matrix	229
Appendix B Gaussian random vectors	234
References	241

* Presented at the Symposium on Engineering Applications of Random Function Theory and Probability at Purdue University in November 1960. To appear in the proceedings of the symposium, to be published by John Wiley.

1. Introduction. There is no doubt that Wiener's theory of statistical prediction and filtering is one of the great contributions to engineering science. Yet the theory has found few practical applications so far. This is probably due to the difficulty of measuring the statistical characteristics of random processes, which is the starting point of the theory.

But even if one is willing to accept physically motivated assumptions in place of experimental statistical data, there remains a major problem: computation of the optimal filter. Current textbooks [1-3] contain several methods for doing this. These methods yield analytical answers only for a few trivial academic examples, and they are rather poorly suited for numerical computations. Most of these procedures terminate with the impulse response of the optimal filter. This is not a complete solution of the problem, however, since there is in general no simple method for synthesizing a filter with a prescribed impulse response. Another shortcoming of the conventional approach is that the treatment of time-variable problems is very awkward.

This paper is concerned with overcoming difficulties of the second type mentioned above. The required statistical data are assumed to be given as part of the problem statement. Moreover, these data are given in such a form that computation of the optimal filter is highly simplified, with a single equation covering all cases.

The Wiener problem is reduced to the classical hamiltonian formalism of the calculus of variations; many long-standing difficulties of the theory are resolved or greatly clarified. The solution consists in the specification of the differential equation of the optimal filter.

The reader should be warned right away that the Wiener problem is not really one in statistics. It belongs to the realm of pure probability theory; it is similar in some ways to the law of large numbers, the central limit theorem, etc. Wiener's approach, as ours, requires that the probabilistic structure of the random processes be known exactly. Therefore confidence limits, statistical decision rules, etc. do not enter the picture.

Wiener assumes stationarity and describes the random process by its power spectral density or covariance function. In this paper, we assume slightly more:

the random process is to be markovian; in other words, it can be thought of as being generated by a linear dynamical system (of finitely many degrees of freedom) excited by white noise. This is very nearly the only case where explicit solutions of the Wiener problem have been found in the past. Very roughly, Wiener's point of view is to admit the possibility of denumerably infinite degrees of freedom -- this is important in some cases. But, in engineering problems, our assumption is frequently more natural.

In any case, the difference between the classical point of view and ours becomes important only if one wants to form an estimate of the power spectral density of a random process on the basis of actual measurements. We shall not be concerned with this question here and will study the Wiener problem solely from the standpoint of probability theory.

The exposition given here summarizes the contributions of two earlier papers [4-5], although many details will be different. There are also a number of theorems and examples which appear now for the first time.

The intent of the paper is primarily expository, and we shall not hesitate to omit certain mathematical technicalities connected with the rigorous definition of continuous random processes. Everything else will be stated in precise mathematical terms; the reader wishing to fill in the missing details should have no difficulty in making contact with the pure mathematical literature.

2. Preview of Contents. We begin with a survey of the main topics of the paper, leaving the definitions intentionally somewhat vague for the moment. We give here also an account of the general philosophy of approach taken in this paper. The reader might read through this section lightly at first, returning to it after a detailed study of the succeeding material. Starred sections and examples may be omitted without interrupting the logical sequence of exposition.

A random process is a family of functions $x(\tau, \omega)$ depending on two arguments: (i) the time τ , which is a real number; and (ii) a random event, which is denoted abstractly by the symbol ω . If $\omega = \omega_0$ is a fixed random event, then $x(\tau, \omega_0)$ is some function of time, usually called the sample-function. If $\tau = \tau_0$ is a fixed instant of time, then $x(\tau_0, \omega)$ is a random variable, which is frequently written simply as $x(\tau_0)$. Instead of letting τ be a real number, we can take τ to be an integer; in this case we call the family of functions $x(\tau, \omega)$ a random sequence. A random process or sequence can also be regarded as a 1-parameter family of random variables.

A random process or sequence is described mathematically by specifying (i) a collection ("ensemble") of sample functions $\{x(\cdot, \omega)\}$ and (ii) the probabilities of the random events ω . In general, there are nondenumerably many random events and therefore the probability of a single event must be set equal to 0. One gets around this difficulty by defining the probability of sets of events. A rigorous definition of random process is a complicated problem; see Doob [6, Chapter 1] and Loeve [7, Chapter 9].

Suppose that we have observed values of $x(\tau, \omega)$ corresponding to some interval of time, say $t_0 \leq \tau \leq t$, where t_0 denotes the starting point of observations and t refers to the present instant of time. Let F_t be the set of all sample functions which agree with the observations made during the interval $t_0 \leq \tau \leq t$. Let Ω_t be the set of all ω 's such that sampling functions corresponding to them belong to the set F_t . By dividing the probability of any subset of Ω_t itself, we obtain the conditional probability of the occurrence of sample functions for values $\tau \geq t$. We can now state the:

PREDICTION PROBLEM. Given the actually observed values of a random process over some interval of time, find the conditional probabilities of all future values of the random process.

Thus the prediction problem consists in calculating conditional probabilities -- often of a very complicated type.

The filtering problem is very similar: instead of observing $x(\tau, \omega)$, we observe a random process $z(t, \omega')$ related to $x(t, \omega)$, i.e., the probabilities of sets of events ω and ω' are dependent on one another. For example, x may be a signal and z the signal plus noise. Stated formally:

FILTERING PROBLEM. Given the actually observed values of a random process over some interval of time, find the conditional probabilities of all values of another, related, random process.

Once the conditional probabilities are known, one can of course answer in principle any problem concerning the probable future evolution of the process: the conditional probabilities incorporate all the information inherent in the observed values of the random process. The adjective "optimal" as used in prediction and filtering theory refers to the fact that all information contained in the observed data is taken into account.

At present, there is but one class of problems in which the prediction or filtering problems can be effectively solved: both x and z must be gauss-markov processes. In this case the solution is quite simple in principle. The conditional probability distribution of a gaussian process is completely described by its mean values and covariances. If in addition the process is also markovian, then it suffices to know the means and covariances at one instant of time.

The solution is as follows. We must compute the conditional means and the conditional covariances. The conditional means will depend on the observed values of the z process. They are computed by putting the observed values through the so-called optimal filter. The conditional covariances are independent of the observed values. Therefore they can be computed separately, even before any observations have been made. Knowledge of the conditional variances at time $\tau = t$ is necessary to compute the conditional means at time $\tau = t$. The equation for the conditional mean is linear, the equation for the covariances is nonlinear.

We digress momentarily to emphasize some consequences of the markovian and gaussian assumptions.

Any random process may be regarded as markovian with a suitable definition of the state of the process. For instance, the state may be taken as the observed past history of the process. The important thing is to find the "smallest" state space for which the markovian property holds. We shall

assume here that the state of the process can be described by a vector with finitely many components; in other words, the state space is finite-dimensional. This assumption is highly desirable because it leads to differential equations of finite order which can be treated by standard methods. We could, for the sake of greater generality, operate in a "larger" (i.e., infinite-dimensional) state space, but the mathematical subtleties which arise do not seem to have physical significance. And, after all, physical systems can be (and often must be) approximated by differential equations of finite order.

The reasons for the gaussian assumption call for a longer explanation. Since the strict prediction or filtering problems cannot be solved in general, it is natural to take a look at weaker versions of the same problem. For instance, let us consider the:

LINEAR FILTERING PROBLEM. Find an estimate $\hat{x}(t_1)$ of $x(t_1)$ which is (i) a linear function of the observed values of $z(\tau)$ and (ii) minimizes the mean-square $E\{x(t_1) - \hat{x}(t_1)\}^2$.

It turns out that the solution of this problem is identical with the solution of the (strict) gaussian filtering problem: the optimal gaussian filter is a linear filter which minimizes the mean-square error.

The proof of this dual interpretation of the optimal filter is the following. If we seek the best linear estimate, then only the first and second order moments (i.e., the means and covariances) of the z process need to be known. Given any random process with prescribed means and covariances, one can find a unique gaussian process with the same means and covariances. (This is trivial since a gaussian random process is uniquely determined by its means and covariances.) Hence the solution of the linear filtering problem in the gaussian case must be the same as in the general case. But in the gaussian case the solution of the linear filtering problems is simultaneously also the solution of the strict filtering problem*. This is because in the gaussian case the mean-square error is minimized by the conditional mean which (as we have noted before) is computed by means of a linear operation on the z process.

* This and similar observations have led Doob [6, pp. 71-78] to introduce the notion of "strict sense" and "wide sense" properties. These concepts are motivated as follows: Suppose a random process has a certain property P which can be expressed in terms of means and covariances. Suppose also that the unique gaussian process with the same means and covariances has a corresponding but stronger property P' . Then P' is a strict-sense property and P is a wide-sense property. In particular, the filter which is optimal in the strict sense in the gaussian case has the wide-sense property that it is the optimal linear filter, without any assumption of gaussianity.

It is a matter of taste which of the two questions we pose; the answer is always the same. If we demand a strict answer, we must also accept the highly restrictive gaussian assumption. If we look only for the best linear filter, then knowledge of the first and second moments of the random process suffices and nothing more has to be assumed about the nature of the probability distribution. To put it differently, if we know only the first and second moments of the random process, then we have only the first (linear) approximation to the dynamical model for the process. (Knowledge of only the first moments would give the zero-th order approximation, i.e., a model which is neither dynamic nor stochastic.) Virtually nothing is known at present even about the second-order approximation.

We now turn to a description of the mathematical results section-by-section.

The concept of a gauss-markov sequence is introduced in Sect. 5. We use as the basic definition a linear dynamical system excited by a gaussian white-noise sequence. This is physically appealing and avoids the unnecessary restriction to stationarity.

The gauss-markov process is defined similarly in Sect. 6. This involves an unpleasant technical difficulty because it is not possible to give a rigorous definition of the gaussian white-noise process by elementary means. We give here only an intuitive definition in terms of an (improper) limit process. A rigorous definition via generalized functions is nowadays fairly straightforward, as may be seen from the literature cited.

Sect. 7 is concerned with showing that the representations of the gauss-markov sequence and gauss-markov process given in Sect. 5 and 6 can be deduced by postulating merely the gaussian and markovian properties. Hence representing such processes as the motion of a linear dynamical system acted on by white noise is not a loss of generality.

Sect. 8 introduces the main topics of the paper by a rigorous but elementary discussion of a standard prediction problem. This provides an interesting comparison of old and new methods.

A precise statement of the filtering problem with which the paper is concerned appears in Sect. 9, together with a discussion of some traditional problems, all of which reduce to our filtering problem by a suitable choice of notation or by minor supplementary assumptions. Equations (I_d) and (I_c) denote the model of the random sequence and process.

Since the present problem formulation is unconventional, its relations with more standard formulations are explored in Sect. 10. Unfortunately, the

standard mean-square approach to filtering rather tends to obscure the theoretical issues involved. The main point is that mean-square optimal linear filtering is optimal also with respect to many criteria other than mean square.

Sect. 11 is the solution of the optimal filtering problem when the time is discrete. This is an improvement over the presentation in [4], achieved by the use of the so-called pseudo-inverse of a matrix. We emphasize strongly that a finite number of parameters (the conditional mean and the conditional covariance matrix) can be regarded as the "state" of the filtering problem; the resulting simplicity of the solution is due solely to this fact. The principal equations of the theory in the discrete case are: (II_d) , the optimal filter and (III_d) , the variance equation. In Sect. 12, we give two examples which illustrate in detail the mathematical and also the physical significance of these equations.

In Sect. 13, we obtain the continuous analogs (II_c) and (III_c) of the optimal filter and of the variance equation. Because of the difficulty in giving a rigorous definition of random differential equations when the excitation is a white-noise process we do not give a rigorous derivation of these results but apply the improper limit argument -- already used to define the white-noise process in Sect. 6 -- to deduce (II_c) and (III_c) from (II_d) and (III_d) . Again, a rigorous derivation requires the use of generalized functions or other advanced tools. The same equations (II_c) and (III_c) were obtained before in a different way [5]. Since the variance equation (III_c) is nonlinear, even the proof of the existence of its solutions is nontrivial -- but easy. Once this is established, we see that the variance equation can be related to a hamiltonian system (IV_c) of $2n$ first-order linear differential equations familiar from the calculus of variations or from theoretical physics. The latter can be solved more-or-less explicitly; this is important because one thereby avoids having to solve the variance equations by numerical quadrature, which would be quite unpleasant because of the many variables involved. This surprising and yet natural connection between the Wiener filtering and the calculus of variations was first pointed out by Kalman and Bucy in [5]. This opens up many promising possibilities of research!

Sect. 14 applies (I_c-IV_c) to a wide variety of problems. The explicit steps are elementary but at times, particularly in Example (14.20), very intricate. Some of these problems are among the most complex ever solved in

Wiener filtering theory. Additional examples are given in [5].

The very great difficulties encountered in the direct and explicit solution of filtering problems literally enforce a change in point of view. Resigned to the fact that explicit answers can only be obtained by numerical computation, one wants to have at least a good qualitative understanding of the filtering problem, particularly as far as the dynamical behavior of the variance equation is concerned. This is indeed the chief task of filtering theory -- in the opinion of the writer.

In Sect. 15 the much simpler problem of minimum-variance unbiased parameter estimation is considered from this point of view. We obtain an important criterion -- complete observability -- for the existence of such an estimator.

This is then used in Sect. 16 to prove the most important theorem of the paper, concerning the existence and uniqueness of limiting solutions of the variance equation. Finally, we exhibit a canonical form for the Hamiltonian equations (IV_c) which can be regarded as a generalization of Wiener's well-known method of spectral factorization. It follows in particular that if the steady-state solution of the Wiener problem is known, a complete solution of the same problem with the observation interval being finite can be constructed using only elementary algebraic steps.

Throughout Sects. 13-16 we avoid appealing to the duality relations which exist between the optimal filtering and the optimal control problem [4-5]; the proofs are given by direct arguments whenever possible.

Appendix A presents some relevant facts concerning the pseudo-inverse and generalized inverse of a matrix. Appendix B is a convenient summary of the properties of gaussian random vectors. A noteworthy feature is a new formula for conditional expectation which is valid even if the gaussian random vectors involved have a singular covariance matrix.

3. Historical remarks and acknowledgements. A characteristic feature of this paper is consistent adherence to the "time-domain" point of view. In the 1940's and early 1950's most of the engineering literature in prediction and filtering theory was written from the "frequency-domain" point of view. This was in harmony with the fashions of system analysis then prevailing, and can be explained by the fact that most stochastic problems in engineering at that time arose in the field of communications where the "frequency-domain" description of systems is quite natural. However, the frequency-domain method as it now stands is not well suited for the study of nonlinear systems or even linear systems with time-varying parameters. Progress in the latter fields has re-awakened interest in "time-domain" methods.

One of the first effective solutions of a time-variable filtering problem was given about 1956 or 1957 by Shinbrot [8]. Although his results can now be obtained more easily by other methods (see Sect. 14), Shinbrot's work contributed substantially to a better understanding of the time-variable filtering problem.

Concurrently or perhaps slightly earlier Pugachev began a systematic study of time-domain methods, culminating in his excellent textbook [9], now in second edition, which is still little known outside the Soviet Union.

More recently, in a series of important papers, Parzen [10-12] has laid the foundations of a general theory of statistical estimation by coordinate-free methods; that is, independently of the particular Hilbert-space representation of the random process.

Very crudely speaking, our approach is the most effective but also the least general of the two. Parzen's work occupies the other extreme; Pugachev is in the middle. Parzen's and Pugachev's starting assumptions lie closer to experimental data, but the calculations which they must perform to get explicit answers are more involved. Ultimately, some synthesis of the three methods is likely to evolve.

The other characteristic feature of this paper is the insistence that prediction and filtering is primarily the determination of conditional distributions and only secondarily the computation of certain functionals of the conditional distributions. This point of view is new; it has been forcefully brought forth in the work of Furstenberg [13]. Surely, this is the clearest and most convenient way of studying prediction and filtering of Gaussian

process; surely, this will be the starting point for future studies in non-linear prediction and filtering theory.

This exposition was prepared with the partial support of the U. S. Air Force under Contracts AF 49(638)-382 and AF 33(616)-6952. The writer is indebted to his colleagues, particularly R. S. Bucy, for numerous stimulating conversations.

4. Notation and other preliminaries. We shall employ in the main the notations and terminology of [5, 14]. Small boldface letters \underline{u} , \underline{v} , ..., \underline{z} denote vectors with coordinates u_1, v_1, \dots, z_1 . Boldface Roman and Greek capitals $\underline{C}, \underline{F}, \underline{G}, \dots, \underline{H}, \underline{\Gamma}$ denote matrices whose elements are written as $c_{1j}, \dots, \gamma_{1j}$. The unit matrix is \underline{I} . Small Greek letters usually denote constants. The time is denoted by t, t_0, t_1 , or τ ; these may be arbitrary real numbers (continuous-time) or arbitrary integers (discrete-time). The letters i, j, \dots, q are reserved for integers.

The transpose of a matrix is denoted by the prime. The inner product of \underline{x} and \underline{y} is denoted by $\underline{x}'\underline{y}$ and the tensor product by \underline{xy}' which is just a matrix with elements $x_i y_j$. The norm $\|\underline{x}\|$ is $(\underline{x}'\underline{x})^{1/2}$. If \underline{A} is a symmetric, nonnegative definite matrix we use the abbreviation $\|\underline{x}\|_{\underline{A}}^2$ for the quadratic form $\underline{x}'\underline{Ax}$. Numerical quantities will be always real, never complex.

The symbol $E\{ \}$ denotes the expectation operator (or ensemble average). We shall retain the curly brackets for greater clarity even if several symbols E are used in the same formula. Sometimes we write covariance matrices as

$$\text{cov}[\underline{x}] = E\{(\underline{x} - E(\underline{x}))(\underline{x} - E(\underline{x}))'\}$$

and

$$\text{cov}[\underline{x}, \underline{y}] = E\{(\underline{x} - E(\underline{x}))(\underline{y} - E(\underline{y}))'\}$$

A continuous-time linear dynamical system in this paper will mean the system of equations

$$dx/dt = \underline{F}(t)\underline{x} + \underline{G}(t)\underline{u}(t), \quad (4.1)$$

$$\underline{y}(t) = \underline{H}(t)\underline{x}(t), \quad (4.2)$$

and a discrete-time linear dynamical system will be the system of difference equations

$$\underline{x}(t+1) = \underline{Q}(t+1, t)\underline{x}(t) + \underline{\Gamma}(t+1, t)\underline{u}(t), \quad (4.3)$$

$$\underline{y}(t) = \underline{H}(t)\underline{x}(t). \quad (4.4)$$

In both cases, we call the n -vector \underline{x} the state of the system, the m -vector $\underline{u}(t)$ is the input or control function, and the p -vector \underline{y} is the output. When the input is an uncontrollable, say random, quantity, we replace $\underline{u}(t)$ by $\underline{w}(t)$. \underline{F} , \underline{G} , \underline{H} resp. \underline{Q} , $\underline{\Gamma}$, \underline{H} are $n \times n$, $n \times m$, $p \times n$ matrices. If all these matrices are constant, then the system is said to be constant (or stationary); if $\underline{u}(t) = \underline{0}$, then the system is free.

The general solution of (4.1) is well known to be [15]:

$$\underline{x}(t) = \underline{Q}(t, t_0)\underline{x} + \int_{t_0}^t \underline{Q}(t, \tau)\underline{G}(\tau)\underline{u}(\tau)d\tau \quad (4.5)$$

with arbitrary \underline{x} , t , t_0 . This formula is valid if, for instance, $\underline{u}(t)$ is a continuous function, in which case the function $\underline{x}(t)$ defined by (4.5) has the following properties:

- (i) it satisfies the initial condition: $\underline{x}(t_0) = \underline{x}$;
- (ii) it is differentiable and satisfies everywhere the differential equation (4.1);
- (iii) it is uniquely determined by the choice of \underline{x} , t_0 .

The matrix $\underline{Q}(t, \tau)$ occurring in (4.5) is called the transition matrix of (4.1); it is characterized by the properties:

$$\underline{Q}(t_0, t_0) = \underline{I} \quad \text{for all } t_0, \quad (4.6)$$

(this follows from (i));

$$\underline{Q}(t_2, t_1)\underline{Q}(t_1, t_0) = \underline{Q}(t_2, t_0) \quad \text{for all } t_0, t_1, t_2 \quad (4.7)$$

(this follows from (4.5) and (iii));

in addition, \underline{Q} satisfies its own differential equation

$$d\Phi(t, t_0)/dt = F(t)\Phi(t, t_0) \quad (4.8)$$

(this follows by setting $\underline{u}(t) \equiv \underline{0}$ and then differentiating (4.5)). From (4.7) it is clear that Φ is never singular.

It can be shown that properties (4.6-4.8) uniquely determine the transition matrix of the differential equation (4.1).

When F is constant, the transition matrix depends only on the difference $t - t_0$ and can be explicitly defined as the exponential of the matrix F :

$$\Phi(t, t_0) = \exp \sum_{i=0}^{\infty} [F(t - t_0)]^i / i! \quad (4.9)$$

When F is not constant, there is no simple way to compute Φ explicitly.

Turning now to the definition of a discrete-time linear dynamical system, it is not necessary to assume that $\Phi(t+1, t)$ is nonsingular. But it is virtually no restriction at all to add this assumption; then one can define by induction $\Phi(t, t_0)$ so that relations (4.6-7) are satisfied for all integers t, t_0 .

It is easy to reduce (4.1) to (4.3). As far as the transition matrix is concerned, we simply consider it for integer values of time only. We must assume, however, that $\underline{u}(t)$ is piecewise constant, that is

$$\underline{u}(t) = \underline{u}(k) \quad \text{when } k \leq t < k+1, \quad \text{where } k = \text{integer.}$$

Then the integral in (4.5) can be computed explicitly and we find

$$\Phi(t+1, t) = \int_t^{t+1} \Phi(t+1, \tau) G(\tau) d\tau. \quad (4.10)$$

The converse is not true, as may be seen at once by considering

$$x_1(t+1) = -x_1(t), \quad (4.11)$$

since the number -1 does not have a real logarithm. In other words, it is impossible to "imbed" (4.11) in a continuous-time dynamical system with a real one-dimensional state space.

A linear dynamical system is said to be stable if

$$\|\underline{q}(t, t_0)\| \leq \alpha \quad \text{for all } t \geq t_0. \quad (4.12)$$

It is asymptotically stable if, in addition,

$$\lim_{t \rightarrow \infty} \|\underline{q}(t, t_0)\| = 0 \quad \text{for all } t_0. \quad (4.13)$$

Finally, the system is uniformly asymptotically stable if

$$\|\underline{q}(t, t_0)\| \leq \alpha e^{-\beta(t - t_0)} \quad \text{for all } t \geq t_0, \text{ where } \alpha, \beta > 0. \quad (4.14)$$

These definitions follow by specializing the more general definitions for arbitrary (possibly nonlinear) dynamical systems [14]. For a constant system, (4.14) is equivalent to: all eigenvalues of F have negative real parts.

5. The gauss-markov sequence. This is a sequence of random vectors $\underline{x}(t)$, $\underline{x}(t+1)$, ... generated by the recursion relation

$$\underline{x}(t+1) = \underline{\Phi}(t+1, t)\underline{x}(t) + \underline{\Gamma}(t+1, t)\underline{w}(t), \quad (5.1)$$

where $\underline{w}(t_0)$, $\underline{w}(t_0+1)$, ... is a sequence of gaussian random vectors, any two of which taken at different times are independent. By gaussianity, the last property is equivalent to the vanishing of the cross-variance matrix:

$$\text{cov}[\underline{w}(t_1), \underline{w}(t_2)] = 0 \quad \text{if } t_1 \neq t_2. \quad (5.2)$$

Though not logically necessary, for the purposes of this paper it will be assumed that the sequence \underline{w} has zero mean:

$$\underline{E}[\underline{w}(t)] = 0 \quad \text{for all } t. \quad (5.3)$$

Then, by gaussianity, it follows also that the sequence \underline{w} is uniquely determined by its auto-covariance matrix:

$$\text{cov}[\underline{w}(t)] = \underline{q}(t). \quad (5.4)$$

It should be noticed that this definition is not complete until the initial state $\underline{x}(t_0)$ of the dynamical system (5.1) is specified. It is natural to assume that $\underline{x}(t_0)$ is a random variable, in fact, a gaussian random variable, with zero mean and arbitrary variance, independent of \underline{w} . Since linear combinations of gaussian random variables are gaussian, it follows that $\underline{x}(t_0)$, $\underline{x}(t_0+1)$, ... is a sequence of gaussian random variables with zero mean.

By repeated application of (5.1), we can write:

$$\underline{x}(t_1) = \underline{\Phi}(t_1, t_0)\underline{x}(t_0) + \sum_{t=t_0}^{t_1-1} \underline{\Phi}(t_1, t)\underline{\Gamma}(t, t-1)\underline{w}(t-1). \quad (5.5)$$

Since the $\underline{w}(t)$ occurring at different times are independent, it follows that, for $t_1 > t_0$,

$$\begin{aligned} \Pr(x_1(t_1) \leq \xi_1, \dots, x_n(t_n) \leq \xi_n | x(t_0), x(t_0 - 1), \dots) \\ = \Pr(x_1(t_1) \leq \xi_1, \dots, x_n(t_n) \leq \xi_n | x(t_0)). \end{aligned} \quad (5.6)$$

In other words, the conditional probability distribution of $x(t_1)$ given $x(t_0)$ and preceding observed values of the state variable is identical with the probability distribution of $x(t_1)$ given the last observation $x(t_0)$. Relation (5.5) is usually called the (strict) markov property.

We have now justified the use of the adjectives "gauss" and "markov" with the sequence generated by (5.1).

By analogy with the common usage in random processes, we may call v a (gaussian) white-noise sequence. Thus a gauss-markov random sequence is a discrete-time linear dynamical system excited by gaussian white noise.

The sequence (5.1) serves as an idealized linear model for random processes observed in nature. In general, the state $x(t)$ of (5.1) is an abstract entity, not amenable to direct physical measurement. To make the model more realistic, we add the assumption: all observables $y(t)$ are as linear functions of $x(t)$. Thus we adjoin to (5.1) the equation

$$z(t) = H(t)x(t) + v(t) = y(t) + v(t) \quad (5.1')$$

where $v(t)$ is a white-noise sequence, specified by

$$\text{cov}[v(t_1), v(t_2)] = 0 \quad \text{if } t_1 \neq t_2,$$

$$E[v(t)] = 0 \quad \text{for all } t,$$

$$\text{cov}[v(t)] = R(t).$$

Adding $v(t)$ to $y(t) = H(t)x(t)$ is intended to reflect the fact that physical measurements of observables can never be made with infinite precision. We shall reserve a detailed motivation and critique of this assumption until later.

Evidently, z given by (5.1') is a random sequence which is related to (more precisely, correlated with) the sequence x . Not only does $z(t)$ depend on $x(t)$, but there may be a correlation between $w(t)$ and $y(t)$:

$$\text{cov}[\underline{w}(t), \underline{v}(t)] = \underline{c}(t) \quad (5.7)$$

We shall adopt (5.1-1') as the standard form of the gauss-markov sequence. Any gaussian white-noise sequence can be put into this form, as we shall prove in Sect. 7.

The system (5.1-1') is shown schematically in Fig. 1. This is a conventional block diagram, except for the fact that the rectangular blocks denote matrices (not scalars); the signals are vectors. To differentiate Fig. 1 from scalar block diagrams, the signal flow is depicted by fat arrows.

6. The gauss-markov process. Intuitively, this concept is most readily understood as the limiting case of a gauss-markov sequence, when the distance between successive values of time tends to zero. We have already noted in Sect. 4 that -- if the forcing function is piecewise constant -- any linear differential equation may be converted into a linear difference equation in such a way that at integer values of time the solutions of the differential equation agree with the solutions of the difference equation. This procedure will now be used in the reverse order.

Let us replace (5.1-1') formally by the system

$$\underline{dx}/dt = \underline{F}(t)\underline{x} + \underline{G}(t)\underline{w}(t), \quad (6.1)$$

$$\underline{z}(t) = \underline{H}(t)\underline{x}(t) + \underline{v}(t). \quad (6.1')$$

The block diagram of this system is shown in Fig. 2; the box $1/s$ symbolizes integration with respect to time.

The terms \underline{v} and \underline{w} in (6.1-1') should be limiting cases of the gaussian white-noise sequences denoted by the same symbols \wedge in Sect. 5. The problem -- to make this notion precise.

First we define the random processes \underline{v} and \underline{w} in such a way that at integer values of time the random processes \underline{x} and \underline{z} generated by (6.1-1') agree with the random sequences \underline{x} and \underline{z} generated by (5.1-1').

To accomplish this, the sample functions are to be piecewise constant over intervals of length 1. We set

$$\underline{v}(t) = \underline{v}(k) \quad \text{and} \quad \underline{w}(t) = \underline{w}(k) \quad (6.2)$$

where

$$k = \text{integer} \quad \text{and} \quad k \leq t < k + 1;$$

the right-hand sides of (6.2) are to be the gaussian white-noise sequences denoted by the same letters in Sect. 5. See Fig. 3.

The solutions of the differential equation (6.1) corresponding to these sample functions constitute the sample functions of a random process \underline{x} . The probabilities of these sample functions can be readily calculated since the driving terms in (6.1) are gaussian. In this way the random process \underline{x} is rigorously defined, and so is the random process \underline{z} given by (6.1'). If the difference equation (5.1) is derived from the differential equation (6.1), these random processes will agree at integer values of time with the random sequences generated by (5.1-1').

The mathematical structure of the random processes just defined is no more complicated than the structure of the corresponding random sequence; we have merely introduced a continuous time parameter.

Now we come to a delicate matter, the definition of the gaussian white noise process. We shall not attempt to give a rigorous definition (which would require advanced analytical tools) but hope that the following discussion will lend some intuitive meaning to this important concept.

Let $\underline{v}^{(q)}$ and $\underline{v}^{(q)}$ be the gaussian random processes defined above, but now we assume that the intervals over which the sample functions are constant are of length q^{-1} (where q is a positive integer). We let $q \rightarrow \infty$. While this happens, we must multiply the covariance matrices by q , in order to preserve the physical characteristics of these processes. This can be seen easily as follows. Let

$$\underline{x} = \int_0^1 \underline{v}^{(q)}(t) dt; \quad (6.3)$$

this is a well-defined random variable for all q . The mean of \underline{x} is zero because the mean of $\underline{v}^{(q)}$ is zero; we compute

$$\begin{aligned} \text{var } \|\underline{x}\| &= E\left\{ \left[\int_0^1 \underline{v}^{(q)}(t) dt \right]^T \cdot \left[\int_0^1 \underline{v}^{(q)}(\tau) d\tau \right] \right\} \\ &= \text{trace} \int_0^1 \int_0^1 E\{ \underline{v}^{(q)}(t) \underline{v}^{(q)'}(\tau) \} dt d\tau \end{aligned}$$

The last expression can be explicitly evaluated from the definition of the \underline{v} process; the result is

$$\text{var } \|x\| = \sum_{i=0}^{q-1} \bar{R}(1/q) \cdot q^{-2}$$

If \bar{R} is constant, this expression tends to zero as q^{-1} . In other words, if \bar{R} is kept constant as $q \rightarrow \infty$, then the effect of white noise in the differential equation

$$dx/dt = \underline{v}^{(q)}(t)$$

would eventually reduce to zero -- this is physically absurd. Hence to keep $\text{var } \|x\|^2 = \text{const.}$ as $q \rightarrow \infty$, we must multiply the covariance matrices of $\underline{v}^{(q)}$ and $\underline{w}^{(q)}$ by q . This means that the amplitudes of the random steps in the sample functions of $\underline{v}^{(q)}$ and $\underline{w}^{(q)}$ increase as \sqrt{q} ; on the other hand, the areas of the random steps tend to 0 since they change as $\sqrt{q} \cdot q^{-1}$,

Guided by these observations, we define the Gaussian white-noise* process \underline{v} and \underline{w} as the formal limit of $\underline{v}^{(q)}$ and $\underline{w}^{(q)}$ as $q \rightarrow \infty$. Since they are Gaussian, \underline{v} and \underline{w} are specified by

$$E\{\underline{v}(t)\} = \underline{0}, \quad E\{\underline{w}(t)\} = \underline{0} \quad \text{for all } t; \quad (6.4)$$

$$E\{\underline{v}(t)\underline{v}'(\tau)\} = \delta(t - \tau)\bar{R}(t), \quad E\{\underline{w}(t)\underline{w}'(\tau)\} = \delta(t - \tau)\underline{Q}(t) \quad \text{for all } t, \tau; \quad (6.5)$$

$$E\{\dot{\underline{v}}(t)\underline{v}'(\tau)\} = \delta(t - \tau)\underline{C}(t) \quad \text{for all } t, \tau, \quad (6.6)$$

where δ is the Dirac delta function.

The preceding discussion shows also that the values of the sample functions of \underline{v} and \underline{w} are to be regarded as delta functions of vanishingly small areas.

Mathematically speaking, this definition is of course sheer nonsense since $\delta(t)$ is not a well-defined function; it is even more absurd to speak of sample

* The term "white" is due to the fact that these processes, like ordinary light, may be thought of as containing waves of every frequency with equal probability. When values of $\underline{v}(t)$ occurring at different times are not independent, this is no longer true and then one sometimes talks of "colored" noise.

functions whose "values" are delta functions of zero area. Still, the idea of white noise is a very useful one. How is this to be reconciled with one's mathematical conscience?

Two points should be emphasized here. First, in the usual applications one never deals with the covariance matrices (6.3-4) directly, but only in conjunction with the computation of integrals. For instance, consider the gaussian random process \underline{x} generated by (6.1A), with $\underline{x}(t_0) = \underline{0}$.

Since $E\{\underline{x}(t)\} = \underline{0}$, its covariance matrix is

$$\text{cov} [\underline{x}(t)] = E\{\underline{x}(t)\underline{x}'(t)\}.$$

By (4.5),

$$\begin{aligned} &= E\left\{ \int_{t_0}^t \underline{\phi}(t, \tau) \underline{G}(\tau) \underline{w}(\tau) d\tau \right. \\ &\quad \times \left. \int_{t_0}^t \underline{w}'(\tau') \underline{G}'(\tau') \underline{\phi}'(t, \tau') d\tau' \right\} \end{aligned}$$

We proceed formally, interchanging the expected-value operation and integration with respect to τ , and using (6.4):

$$= \int_{t_0}^t d\tau \int_{t_0}^t d\tau' \underline{\phi}(t, \tau) \underline{G}(\tau) \underline{G}'(\tau') \underline{\phi}'(t, \tau') \delta(\tau - \tau').$$

Utilizing properties of the δ function, we finally obtain

$$\text{cov} [\underline{x}(t)] = \int_{t_0}^t \underline{\phi}(t, \tau) \underline{G}(\tau) \underline{G}'(\tau) \underline{\phi}'(t, \tau) d\tau \quad (6.7)$$

The derivation of this result is purely formal, because the random process \underline{x} has not even been defined. Since the sample functions of \underline{w} and \underline{w}' are mathematically meaningless, so is also the differential equation (6.1A).

But we could have obtained (6.7) rigorously by the following rigorous procedure. The integral (6.7) certainly exists in relation to the process $\underline{y}(q)$; its limit $q \rightarrow \infty$ also exists, and we can regard this limit as the

covariance matrix of some random process x , not yet completely defined. Carrying this method farther, we can define a random process by specifying the values of its integrals (linear functionals) and not assigning any meaning to its sample functions. This idea was developed by Wiener in the 1920's and still constitutes one of the main tools of the rigorous theory of random processes.

The second point is this. The subterfuge of dealing only with integrals of a random process is not really satisfactory because no meaning is attached to the differential equation (6.1) itself. In recent years, a new approach has evolved which is relatively free of difficulties of this sort. The white-noise process is regarded as a generalized random process which is the random counterpart of the concept of a generalized function (or distribution) invented by Sobolev and L. Schwartz. This technique is used successfully by the Russian school led by I. M. Gel'fand [16, 17].

As mentioned in the Introduction, as far as the present paper is concerned, we regard the difficulties surrounding the rigorous definition of random processes as purely technical; we shall not hesitate therefore to take limits formally, interchange the expected value operation and integration with respect to time, etc. (The reader will notice that the "inadmissible steps" are used only to derive integrals of the type (6.7) -- these results could be rigorously justified by Wiener's technique.) We shall devote a future paper to such problems, using the Gel'fand theory.

The definition of a random process by means of a linear dynamical system excited by white noise was emphasized in the engineering literature particularly by Bode and Shannon [18] and Zadeh and Ragazzini [19]. Not only is this assumption physically pleasing* but it leads to a clear and convenient mathematical framework.

* We imagine that noise originates on a microscopic level. Macroscopic noise is clearly gaussian, because of the superposition of many small random effects; and it is white, because the dynamics of microscopic phenomena are very fast on the time-scale of the microscopic observer. Appreciable dynamical effects come about only on a macroscopic scale and are represented by (6.1).

7. Axiomatic definition of the gauss-markov sequence and process.

The definitions given in the previous sections may seem to be highly arbitrary: guided by physical intuition, we postulate a "mechanism" is simply a natural representation of the process; we can derive this representation by taking the gaussian and markovian properties as fundamental axioms. In other words, from a logical point of view there is no loss in generality in starting with (5.1-1') or (6.1-1') as the basic definition.

AXIOMATIC DEFINITION. We say that a 1-parameter family of random vectors $\underline{x}(t)$ ($t = \text{integer}$) is a gauss-markov sequence if it has the following properties:

(I) the sequence is gaussian; that is to say, for any fixed integers t, τ the random vectors $\underline{x}(t), \underline{x}(\tau)$ have a joint gaussian distribution with mean $\underline{\mu}(t), \underline{\mu}(\tau)$ and cross-covariance matrix $\underline{\Sigma}(t, \tau)$.

(II) the sequence is markovian, in other words, for any integer $t_1 > t_0$, the strict markovian property (5.6) is satisfied.

Similarly, we say that the 1-parameter family of random vectors $\underline{x}(t)$ ($t = \text{real number}$) is a gauss-markov process, if the preceding properties hold with t, τ being real numbers, and if $\underline{\Sigma}(t, t)$ is nonsingular while $\underline{\Sigma}(t, \tau)$ is a continuously differentiable function of t, τ . (End of definition.)

First we shall study random sequences. We assume for the moment that $\underline{\mu}(t) \equiv \underline{0}$. We let $t > \tau$ be integers and write

$$\underline{\Phi}(t, \tau) = \underline{\Sigma}(t, \tau) \underline{\Sigma}^{\#}(t, \tau) \quad (7.1)$$

where $(\)^{\#}$ denotes the generalized inverse of Penrose. (See Appendix A.) By (B.11), $\underline{v}(t) = \underline{x}(t+1) - E\{\underline{x}(t+1) | \underline{x}(t)\}$, is a gaussian random vector with zero mean which is independent of $\underline{x}(t)$. Hence we can write

$$\underline{x}(t+1) = \underline{\Phi}(t+1, t) \underline{x}(t) + \underline{v}(t) \quad (7.2)$$

which is identical with (5.1), except that $\underline{\Gamma} = \underline{I}$.

Moreover, \underline{w} in (7.2) is a white-noise sequence. For

$$E(\underline{x}(t+1) | \underline{x}(t), \underline{x}(t-1)) = E(\underline{x}(t+1), \underline{w}(t-1));$$

by definition, $\underline{x}(t-1)$ and $\underline{w}(t-1)$ are independent of each other, so that

$$\begin{aligned} &= E(\underline{x}(t+1) | \underline{x}(t-1)) + E(\underline{x}(t+1) | \underline{w}(t-1)) \\ &= \underline{\phi}(t+1, t-1)\underline{x}(t-1) + E(\underline{\phi}(t+1, t)\underline{x}(t) | \underline{x}(t-1)) \\ &\quad + E(\underline{w}(t) | \underline{x}(t-1)); \end{aligned}$$

by (7.2) the middle term in the preceding equation is just $\underline{\phi}(t+1, t)\underline{w}(t-1)$.
Hence

$$E(\underline{x}(t+1) | \underline{x}(t), \underline{x}(t-1)) = \underline{\phi}(t+1, t)\underline{x}(t) + E(\underline{w}(t) | \underline{w}(t-1))$$

By the markovian property, this is also equal to

$$= E(\underline{x}(t+1) | \underline{x}(t))$$

which implies that

$$E(\underline{x}(t) | \underline{x}(t-1)) = \underline{0}$$

proving that $\underline{w}(t)$ and $\underline{w}(t-1)$ (being gaussian) are mutually independent.

Further,

$$\begin{aligned} E(\underline{w}(t) | \underline{x}(t-1)) &= E(\underline{x}(t+1) - \underline{\phi}(t+1, t)\underline{x}(t) | \underline{x}(t-1)) \\ &= [\underline{\phi}(t+1, t-1) - \underline{\phi}(t+1, t)\underline{\psi}(t, t-1)]\underline{x}(t-1) \end{aligned}$$

so that the independence of $\underline{w}(t)$ and $\underline{x}(t-1)$ will follow if we prove the identity

$$\underline{\Phi}(t_3, t_2)\underline{\Phi}(t_2, t_1) = \underline{\Phi}(t_3, t_1) \text{ whenever } t_1 < t_2 < t_3 \quad (7.3)$$

In fact, let $\underline{x}(t_1)$ be arbitrary and consider

$$\underline{\Phi}(t_3, t_1)\underline{x}(t_1) = E(\underline{x}(t_3) | \underline{x}(t_1))$$

by an elementary property of conditional expectations [12; p. 35],

$$= E(E(\underline{x}(t_3) | \underline{x}(t_2), \underline{x}(t_1)) | \underline{x}(t_1)),$$

using the markov property,

$$= E(E(\underline{x}(t_3) | \underline{x}(t_2)) | \underline{x}(t_1));$$

using gaussianness, we calculate the conditional expectation by (B.6) and write by (7.1),

$$= E(\underline{\Phi}(t_3, t_2)\underline{x}(t_2) | \underline{x}(t_1)),$$

$$= \underline{\Phi}(t_3, t_2)\underline{\Phi}(t_2, t_1)\underline{x}(t_1),$$

which proves (7.2) since $\underline{x}(t_1)$ was arbitrary.

Extending these arguments by induction, it follows easily that

$$\begin{aligned} \text{cov}[\underline{v}(t), \underline{x}(\tau)] &= \underline{0} & \text{if } t \geq \tau, \\ \text{cov}[\underline{v}(t), \underline{v}(\tau)] &= \underline{0} & \text{if } t \neq \tau. \end{aligned} \quad (7.4)$$

Conversely, (7.3-4) show that the recursion relation (7.2) defines a gauss-markov sequence, which has zero mean and the prescribed covariance matrix $\underline{\Sigma}(t, \tau)$ (provided we let $\underline{x}(t_0)$ be a gaussian random vector with $\text{cov } \underline{x}(t_0) = \underline{\Sigma}(t_0, t_0)$).

The covariance matrix of \underline{w} is nonnegative definite but may be singular. For convenience, we factor it into the form $\underline{\Gamma} \underline{\Gamma}'$ and then (7.2) becomes

$$\underline{x}(t+1) = \underline{\Sigma}(t+1, t)\underline{x}(t) + \underline{\Gamma}(t+1, t)\underline{w}(t), \quad (7.5)$$

where \underline{w} is now defined as an m -vector white-noise sequence with zero mean and unit variance.

Now we can remove the zero mean assumption by considering $\underline{x}(t) - \underline{\mu}(t)$ instead of $\underline{x}(t)$; then (7.5) becomes

$$\underline{x}(t+1) - \underline{\mu}(t+1) = \underline{\Phi}(t+1, t)(\underline{x}(t) - \underline{\mu}(t)) + \underline{\Gamma}(t+1, t)\underline{w}(t) + \underline{d}(t),$$

where the deterministic component $\underline{d}(t)$ is defined as

$$\underline{d}(t) = \underline{\mu}(t+1) - \underline{\Phi}(t+1, t)\underline{\mu}(t).$$

Now let \underline{z} be a gaussian random sequence which is causally related to $\underline{x}(t)$, that is to say, the conditional distribution of $\underline{z}(t)$ given $\underline{x}(t)$ is identical with the conditional distribution of $\underline{z}(t)$ given $\underline{x}(t), \underline{z}(t-1), \underline{z}(t-2), \dots$. Define

$$E(\underline{z}(t) | \underline{x}(t)) = \underline{H}(t)\underline{x}(t),$$

$$\underline{v}(t) = \underline{z}(t) - \underline{H}(t)\underline{x}(t).$$

Proceeding as before, it can be proved that \underline{v} is a gaussian white-noise sequence (possibly correlated with $\underline{w}(t)$).

We have now proved that (5.1-1') is a representation of the abstractly defined gaussian sequences \underline{x} and \underline{z} . This representation is clearly unique, aside from the unimportant arbitrariness in defining $\underline{\Gamma}$ and hence also \underline{v} .

The derivation of the representation (6.1) proceeds similarly. There are some minor points which require comment, however.

Since $\underline{\Sigma}(t, t)$ is assumed to be nonsingular, it follows from (7.1) that

$$\underline{\Phi}(t, t) = \underline{I} \quad \text{for all } t.$$

If $\underline{\Sigma}(t, \tau)$ is continuously differentiable in t for all τ , then

$$\underline{F}(t) = \left. \frac{\partial \underline{\Phi}(t, \tau)}{\partial t} \right|_{\tau = t} = \lim_{\substack{h \rightarrow 0 \\ \tau \rightarrow t}} \left[\frac{\underline{\Phi}(t+h, t) - \underline{I}}{h} \right] \underline{\Phi}(t, \tau)$$

is defined. Therefore by (7.3) $\underline{\Phi}(t, \tau)$ satisfies the differential equation

$$d\underline{\Phi}/dt = \underline{F}(t)\underline{\Phi} \quad \text{for all } t \geq \tau.$$

By (4.7), $\underline{\Phi}$ will then satisfy (7.3) for all real numbers t_1, t_2, t_3 .

On the other hand, any scalar of the form

$$\varphi(t, \tau) = \exp[\alpha(t) - \alpha(\tau)]$$

satisfies (7.3); but if φ is not differentiable, we cannot regard it as the solution of a differential equation, so that the representation (6.1) cannot exist.

If the gauss-markov process is stationary, i.e., $\underline{\Sigma}(t, \tau)$ depends only on $t - \tau$, then it suffices to assume that $\underline{\Sigma}$ is continuous in t at $t = \tau = 0$ [20, p. 46] to assure that $\underline{\Phi}(t, \tau)$ is the transition matrix of a differential equation, and thus to prove the representation (6.1).

Similar results were obtained a long time ago by Doob [21] and Wang and Uhlenbeck [22], but the present derivation is simpler.*

*

Incidentally, neither Doob nor Wang and Uhlenbeck make any assumptions about continuity of $\underline{\Sigma}$. Without some such assumption their results are incorrect. For instance, assuming stationarity $\sigma(t, t) = 1$, and considering the scalar case, (7.3) reduces to $\sigma(t_3 - t_1) = \sigma(t_3 - t_2)\sigma(t_2 - t_1)$. Moreover, since σ is a correlation coefficient, $|\sigma| \leq 1$ and σ is an even function of its argument. Doob [22] asserts that the only nonzero function $\sigma(t)$ satisfying these conditions is $\sigma(t) = \exp[-\alpha|t|]$, where α is a nonnegative constant. But this is false, for one can construct -- using the axiom of choice -- functions which satisfy these requirements but are everywhere discontinuous so that they cannot be represented by an exponential.

2. A simple prediction problem. Before embarking on the detailed and unavoidably complex study of prediction and filtering in the general case, it will be helpful to pause for a minute and solve a simple problem. Consider a gauss-markov process generated by

$$\left. \begin{aligned} dx_1/dt &= -\alpha x_1 + x_2 \\ dx_2/dt &= -\alpha x_2 + w_1(t) \\ z_1(t) &= y_1(t) = x_1(t) \end{aligned} \right\} \quad (\alpha > 0) \quad (8.1)$$

The ordinary block diagram of the system is shown in Fig. 4A. Note that the output of the system can be observed without any corrupting noise.

The matrices \underline{F} , \underline{G} , \underline{H} can be read off by inspection from Fig. 4A. They are:

$$\underline{F} = \begin{bmatrix} -\alpha & 1 \\ 0 & -\alpha \end{bmatrix}, \quad \underline{G} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \underline{H} = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad (8.2)$$

We wish to estimate the value of $x_1(t + \theta)$, where $\theta > 0$, having observed all past outputs $z_1(\tau)$ of the system up to time t . This problem is identical with one stated (in different language) in [3, p. 406]. We shall solve this problem here after very few preliminaries and in a very much simpler way than in [3], where the solution, using older methods, appears after 400 pages of preparation.

By linearity, the quantity $x_1(t + \theta)$ depends on two things: (i) the state $\underline{x}(t)$ and (ii) the excitation $w_1(\tau)$ in the interval $[t, t + \theta]$. Since w_1 is a white-noise process, its future values cannot be estimated in any way from past observations; or, more precisely, the best estimate is simply the mean, which in this case is zero. Expressing this in writing, we have:

$$\begin{aligned}
E(x_1(t + \theta) | z_1(\tau), \tau \leq t) &= H \underline{\phi}(t + \theta, t) E(x(t) | z_1(\tau), \tau \leq t) \\
&\quad + E(H \int_t^{t+\theta} \underline{\phi}(t + \theta, \tau) G w(\tau) d\tau | z_1(\tau), \tau \leq t) \\
&= H \underline{\phi}(t + \theta, t) E(x(t) | z_1(\tau), \tau \leq t)
\end{aligned} \tag{8.3}$$

Now we calculate the transition matrix of (8.1). This is easily done by noting that $\phi_{1j}(t, \tau)$ = response observed at the 1-th integrator in Fig. 4 at time t if a unit impulse is applied to the input of the j -th integrator at time $\tau < t$. The result is:

$$\underline{\phi}(t, \tau) = \begin{bmatrix} e^{-\alpha(t - \tau)} & (t - \tau)e^{-\alpha(t - \tau)} \\ 0 & e^{-\alpha(t - \tau)} \end{bmatrix} \tag{8.4}$$

What is the conditional expectation of $x(t)$, given all the observations $z_1(\tau)$ up to time t ? Clearly, $x_1(t)$ is known exactly because the observations are not corrupted by noise. On the other hand, by (8.1),

$$x_2(t) = dx_1(t)/dt + \alpha x_1(t) = dx_1(t)/dt + \alpha x_1(t) \tag{8.5}$$

But the white noise process w_1 passes through two "smoothing" operations as shown in Fig. 4. Thus x_2 is "smoother" than w_1 (x_2 is the so-called Ornstein-Uhlenbeck process [21]), and x_1 is smoother than x_2 ; in particular, the x_1 process has a derivative, $-\alpha x_1 + x_2$, which is a well-defined random process. Hence we may evaluate the right-hand side of (8.5). Thus we get by (8.5)

$$\bar{x}(x(t) | z(\tau), \tau \leq t) = \hat{x}(t | t) = \begin{bmatrix} z_1(t) \\ \alpha x_1(t) + dx_1(t)/dt \end{bmatrix},$$

and, using (8.4),

$$E(x_1(t + \theta) | z(\tau), \tau \leq t) =$$

$$= \hat{x}_1(t + \theta | t) = e^{-\alpha\theta} [(1 + \alpha\theta)z_1(t) + \theta dz_1(t)/dt]$$

This agrees with [3, p. 408, eq. 73]. The optimal predictor is shown in Fig. 4B. The symbol $\dot{}$ denotes differentiation with respect to time.

A most interesting feature of this result is that it is independent of the variance of w_1 . Using the concept of white noise one can almost completely dispense with the machinery of probability theory to get the answer.

Another important point is the fact that the optimal prediction involves the operation of differentiation. This operation is not realizable in practice: mathematically, because differentiation is an unbounded operator; and physically, because the ideal differentiator has infinite bandwidth. We shall see later that this unpleasant feature of optimal prediction is a consequence of the assumption that the output of the system (8.1) can be observed exactly. If we introduce white noise in the observations, with no matter how little energy per unit time, the difficulty disappears. We have therefore two choices in formulating a prediction or filtering problem in continuous time:

(i) either we assume that the observations on the random process can be made with infinite accuracy -- then we must approximate the ideal predictor which is not physically realizable;

(ii) or we assume that the observations are contaminated with white noise -- then the optimal predictor is always realizable.

We shall always choose the second assumption which is far more natural from the physical point of view.

It is clear that this difficulty does not arise in discrete-time systems and therefore the question of whether or not the observations are exact is immaterial.

9. Statement and examples of the filtering problem. We have now arrived at the main part of the material. For ease of reference we restate the essence of the discussion in Sects. 2, 5-7 as follows:

FILTERING PROBLEM. Consider the gauss-markov sequence

$$\underline{x}(t+1) = \underline{a}(t+1, t)\underline{x}(t) + \underline{\Gamma}(t+1, t)\underline{w}(t), \quad (I_a)$$

$$\underline{z}(t) = \underline{H}(t)\underline{x}(t) + \underline{v}(t)$$

where v, w are gaussian white-noise sequences.

Or consider the gauss-markov process

$$dx/dt = \underline{F}(t)\underline{x} + \underline{G}(t)\underline{w}(t), \quad (I_c)$$

$$\underline{z}(t) = \underline{H}(t)\underline{x} + \underline{v}(t),$$

where v, w are gaussian white-noise processes.

In either case, v and w are explicitly defined by the relations

$$E(\underline{v}(t)) = \underline{0}, \quad E(\underline{w}(t)) = \underline{0}, \quad \text{for all } t;$$

$$E(\underline{v}(t)\underline{v}(\tau)) = \delta(t - \tau)\underline{R}(t), \quad E(\underline{w}(t)\underline{w}'(\tau)) = \delta(t - \tau)\underline{Q}(t), \quad \text{for all } t, \tau;$$

$$E(\underline{v}(t)\underline{w}'(\tau)) = \delta(t - \tau)\underline{C}(t) \quad \text{for all } t.$$

(In these expressions t, \tau are integers resp. real numbers; \delta(t - \tau) is the kronecker delta resp. the Dirac delta function.)

Now suppose the actual values of the random variable z(\tau) have been observed in the interval t_0 \le \tau \le t.

What is the conditional probability distribution of x(t_1)?

We shall refer to (I) as the model of the message process. This terminology is motivated by communication theory: one may regard intuitively

$$\underline{y}(t) = \underline{H}(t)\underline{x}(t)$$

as the message, \underline{y} is the noise, \underline{z} is the signal (message plus noise); \underline{y} is the reason why \underline{z} is a random variable.

It will be convenient to use from now on certain special notations.

Let

$$\hat{\underline{x}}(t_1|t) = E(\underline{x}(t_1) | \underline{z}(t_0), \dots, \underline{z}(t))$$

be the conditional mean of $\underline{x}(t_1)$ given observed values of $\underline{z}(\tau)$ for $t_0 \leq \tau < t$. Similarly, let

$$\tilde{\underline{x}}(t_1|t) = \underline{x}(t_1) - \hat{\underline{x}}(t_1|t)$$

be the "error" between the actual value of $\underline{x}(t_1)$ and its conditional expectation. We note, by (B.11), that $\hat{\underline{x}}(t_1|t)$ and $\tilde{\underline{x}}(t_1|t)$ are independent random variables. Finally, let the conditional covariance matrix of $\tilde{\underline{x}}(t_1|t)$ be

$$\underline{\Sigma}(t_1|t) = E(\tilde{\underline{x}}(t_1|t)\tilde{\underline{x}}'(t_1|t))$$

The quantities $\hat{\underline{y}}(t_1|t)$, $\tilde{\underline{y}}(t_1|t)$, $\hat{\underline{v}}(t_1|t)$, etc. are defined similarly.

By gaussianity, the solution of the filtering problem is equivalent to computing $\hat{\underline{x}}(t_1|t)$ and $\underline{\Sigma}(t_1|t)$.

Very many different problems are included in the matrix equations (I_c) or (I_d).

(9.1) EXAMPLE: Dynamical systems subject to random disturbances and measurement noise. Consider a physical dynamical system (an airplane, space vehicle, or chemical plant). Assume the system is linear. The state of the system cannot be observed directly but only through the output $\underline{y}(t)$, which can be measured only in the presence of additive gaussian noise \underline{v} . In addition, the system is subject to random disturbances (atmospheric turbulence, meteorites, chemical impurities) in the form of the gaussian white-noise process \underline{w} . The equations of motion of the system are evidently (5.1) and (6.1), provided we add a deterministic forcing term $\underline{u}(t)$ to (5.1) and (6.1) to account for control variables (rudder, control jets, catalysts). In order to control the system, it is necessary to know the state variables. They can be "reconstructed" by means of an optimal filter. The variance of \underline{v} and \underline{w} can often be specified (within an order of magnitude) by physical considerations.

(9.2) EXAMPLE. It is not necessary to assume in the preceding example that the measurement noise or the random disturbances are white. One can always represent correlation by adding more state variables. For instance, the instruments which measure $y(t)$ may have dynamics of their own and the noise may enter at the output as well as the input of the instruments. See Example (14.52) for a problem of this sort. After the additional dynamical effects have been taken into account, the describing equations can always be reduced again to the standard form, perhaps after some redefinition of variables.

(9.3) EXAMPLE: Estimation of parameters. This is a very common problem in statistics [23, Chapters 32-34]. Suppose we are given a family of functions $\eta_{1j}(\tau)$, $i = 1, \dots, m$ and $j = 1, \dots, n$. We can measure m random variables

$$\zeta_i(\tau) = \sum_{j=1}^n \theta_j \eta_{1j}(\tau) + v_i(\tau) \quad (i = 1, \dots, m) \quad (9.4)$$

in the presence of gaussian white noise $v_i(\tau)$. The problem is to form the "best possible estimate" $\hat{\theta}$ of θ based on observations of $\eta_i(\tau)$ in some interval $[t_0, t]$.

We can easily reduce the problem to the context of (I_a) or (I_c) as follows. Let $y = \theta$. We can then regard θ as the unknown state $x(t)$ of (I_a) or (I_c) , provided that we can represent the functions η_{1j} as

$$\eta_{1j}(\tau) = \sum_{k=1}^n h_{1j}(\tau) \varphi_{jk}(\tau, t) \quad (9.5)$$

where $\varphi_{jk}(\tau, t)$ are elements of a transition matrix. Of course, (9.4) imposes a restriction on the admissible functions η_{1j} : but since any continuous function can be approximated by solutions of ordinary differential equations, this is not a serious limitation in practice.

Intuitively, this problem can be visualized as generalized curve fitting. We have m experimental curves represented by values of the random variable $\zeta_i(\tau)$. These experimental curves are to be fitted simultaneously by linear combinations of smooth curves from the family η_{1j} .

This problem can be solved easily even without the assumption that y is a white-noise process. See Sect. 15.

(9.6) EXAMPLE: Communication system. A very elementary model of a communication system might be the following. A message is a sample-function

of the random process y , defined over some interval, say, $[t_0, t_1]$. The transmitted message y is contaminated by noise v before it reaches the receiver. The mathematical problem in receiver design is the following. Given the observed values of $z(\tau)$ on the interval $[t_0, t]$, what is the best estimate of that sample function of the y process which actually occurred? In other words, find

$$\hat{y}(\tau|t) \quad \text{for all } \tau \text{ in } [t_0, t_1].$$

This problem is quite difficult because it might involve simultaneously both prediction and smoothing; no adequate solution exists as yet in the framework of this paper.

It should be noted that this is also an estimation problem. Unlike Example (9.3) where the unknown parameter was a (finite-dimensional) vector, here the unknown parameter is a more complex mathematical object -- a real-valued function.

The formulation of the filtering problem given here is different from the conventional formulation in the engineering literature [1-5]. The two points of view can be easily reconciled as was mentioned briefly in Sect. 2 and as is discussed in more detail in the next section.

10. Other formulations of the filtering problem. Connections between our version of the filtering problem and other points of view appeared already in Sect. 2. To aid the reader in certain applications of the theory, we summarize here some well-known facts.

Often the filtering problem is formulated as follows. Find an estimate $\underline{x}^*(t_1)$ of $\underline{x}(t_1)$, based on observations of $\underline{z}(\tau)$, $t_0 \leq \tau \leq t$, which minimizes the expected loss

$$E\{E(L(\underline{x}(t_1) - \underline{x}^*(t_1)) | \underline{z}(\tau), t_0 \leq \tau \leq t)\}. \quad (10.1)$$

The loss function L is defined as follows. Let $\rho(\underline{x})$ be a real-valued nonnegative, convex function of \underline{x} :

$$\rho(\lambda \underline{x} + (1 - \lambda)\underline{y}) \leq \lambda \rho(\underline{x}) + (1 - \lambda)\rho(\underline{y}), \quad \text{where } 0 \leq \lambda \leq 1.$$

Then L is a real-valued function of \underline{x} such that

$$L(0) = 0,$$

$$L(\underline{x}_2) \geq L(\underline{x}_1) \geq 0 \quad \text{when} \quad \rho(\underline{x}_2) \geq \rho(\underline{x}_1) \geq 0. \quad (10.2)$$

Evidently $\rho(\underline{x})$ measures the distance of \underline{x} from the origin, and the loss is nondecreasing with this distance. Observe that L need not be convex.

The solution of the preceding problem is contained in the theorem:

(10.3) Let \underline{x} be an n -dimensional random vector with mean $\underline{\mu}$ and distribution function $F(\underline{x})$. If

- (A) F is symmetrical about $\underline{\mu}$ and
 - (B) F is unimodal (i.e., convex for $\hat{x}_i \leq \mu_i$, $i = 1, \dots, n$),
- Then $E(L(\underline{x} - \hat{\underline{x}}))$ is minimized by setting $\hat{\underline{x}} = \underline{\mu}$.

For the proof, see Sherman [24].

The preceding conditions are obviously satisfied by a gaussian distribution. Hence (10.1) is minimized by taking for $\underline{x}^*(t_1)$ the conditional expectation

$$\underline{x}^*(t_1) = \underline{x}(t_1 | t) = E(\underline{x}(t_1) | \underline{z}(\tau), t_0 \leq \tau \leq t),$$

whose calculation is a part of the filtering problem, as stated in Sect. 2.

A special loss function is $L(\underline{x}) = \|\underline{x}\|_{\underline{P}}^2$, where \underline{P} is nonnegative definite in this case, (10.3) is true without any assumption on the distribution function:

$$\begin{aligned} E(\|\underline{x} - \underline{x}^*\|_{\underline{P}}^2) &= E(\|\underline{x}\|_{\underline{P}}^2) + 2E(\underline{x}'\underline{P}\underline{x}^*) + E(\|\underline{x}^*\|_{\underline{P}}^2), \\ &= \text{const.} + 2\underline{\mu}'\underline{P}\underline{x}^* + \|\underline{x}^*\|_{\underline{P}}^2, \\ &= \|\underline{x}^* - \underline{\mu}\|_{\underline{P}}^2 + \text{const.}, \end{aligned}$$

which is obviously minimized again by $\underline{x} = \underline{\mu}$; it should be noted that this result does not depend in any way on \underline{P} . In particular, suppose $\underline{P} = \underline{a}\underline{a}'$. Then the best estimate of $\underline{a}'\underline{x}$ is $\underline{a}'\underline{\mu}$.

In the literature one often reads snap judgments to the effect that only squared loss functions can be treated. This is incorrect or at least misleading. The preceding discussion shows that the conditional mean supplies the minimum expected loss for many loss functions. Thus the loss function plays a secondary role. Of course if the conditional distribution is known, the best estimate \underline{x}^* can be computed for any loss function.

Finally, we may wish to find the best estimate \underline{x}^* which is a linear function of the data $\underline{z}(\tau)$, $t_0 \leq \tau \leq t$. We have seen that for a "reasonable" loss function the best estimate $\underline{x}^*(t_1)$ in the gaussian case is always the conditional expectation $\hat{\underline{x}}(t_1 | t)$, which is, again by gaussianity, linear in $\underline{z}(\tau)$. The calculation of this estimate involves only the means and covariance matrices of the gaussian process. Thus (as we have pointed out already in Sect. 2) $\underline{x}^*(t_1) = \hat{\underline{x}}(t_1 | t)$ is clearly the best linear estimate for the class of all random processes with the same means and covariance matrices as the gaussian process for which $\hat{\underline{x}}(t_1 | t)$ was computed.

We have now proved:

(10.4) If a linear estimate $x^*(t_1)$ is optimal for one loss function of type (10.2), it is optimal for all such loss functions.

Hence the linear minimum mean square estimate is optimal for all loss functions (10.2).

11. Solution of the filtering problem for random sequences. According to the problem statement in Sect. 9, we are to compute the conditional distribution of $\underline{x}(t_1)$, given observations $\underline{z}(\tau)$ in the interval $t_0 \leq \tau \leq t$. By gaussianity, this is of course equivalent to computing conditional means and covariance matrices.

It will be convenient to work in terms of $\hat{\underline{x}}(t+1|t)$ and $\underline{\Sigma}(t+1|t)$.

First we show how to reduce the problem to the computation of these quantities.

Let $t_1 \geq t+2$. By repeated use of the defining equation (I_d) of a random sequence, we obtain the expression

$$\underline{x}(t_1) = \Phi(t_1, t+1)\underline{x}(t+1) + \sum_{\tau=t+1}^{t_1-1} \Phi(t_1, \tau+2)\Gamma(\tau+1, \tau)\underline{v}(\tau), \quad (11.1)$$

which is valid for all $t_1 \geq t+2$. Taking conditional expectations of both sides with respect to $\underline{z}(t_0), \dots, \underline{z}(t)$, we obtain the relation

$$\underline{x}(t_1|t) = \Phi(t_1, t+1)\underline{x}(t+1|t) \quad \text{when} \quad t_1 \geq t+1, \quad (IV_d)$$

using the fact that $\underline{v}(t+1), \underline{v}(t+2), \dots$ have zero mean and are independent of $\underline{z}(t_0), \dots, \underline{z}(t)$. We see that $\underline{x}(t_1|t)$ is obtained by extrapolating $\underline{x}(t+1|t)$ by means of the transition matrix of the random sequence (I_d).

Let $t_1 = t$. Taking conditional expectations on both sides of (I_d), we get

$$\underline{x}(t+1|t) = \Phi(t+1, t)\underline{x}(t|t) + \Gamma(t+1, t)\underline{v}(t|t). \quad (11.2)$$

It is easy to see that $\underline{v}(t)$ is independent of $\underline{x}(t), \underline{x}(t-1), \dots$; since $E[\underline{v}(t)] = 0$, it follows by (B.6) that $\underline{v}(t|t) = \underline{A}(t)\underline{x}(t)$, where $\underline{A}(t) = \underline{Q}(t)\text{cov}[\underline{x}(t)]^{-1}$. Since we have assumed (see Sect. 4) that $\Phi(t+1, t)$ is nonsingular, (11.2) can be solved for $\underline{x}(t|t)$. In other words, we can express $\underline{x}(t|t)$ as a linear function of $\underline{x}(t+1|t)$ and $\underline{x}(t)$.

If $\Phi(t+1, t)$ is singular, this procedure fails and we must work with $\underline{x}(t|t)$ instead of $\underline{x}(t+1|t)$ as the basic quantity — the required modifications are easy but the resultant formulas are less simple.

If $t_1 < t$, we cannot express $\hat{\underline{x}}(t_1|t)$ solely in terms of $\hat{\underline{x}}(t+1|t)$ and $\underline{z}(t)$. As a matter of fact, $\hat{\underline{x}}(t_1|t)$ will be in general a linear combination of $\hat{\underline{x}}(t+1|t)$, $\underline{z}(t)$, ..., $\hat{\underline{x}}(t_1+1|t)$, $\underline{z}(t_1)$. Fortunately, this is seldom required in practice. The details are messy, and we omit them.

The computation of $\underline{\Sigma}(t_1|t)$ is similar. Since the explicit expressions for $\underline{\Sigma}(t_1|t)$ will not be needed in the sequel, the details are again omitted.

The remainder of this section is concerned primarily with computing $\hat{\underline{x}}(t+1|t)$ and $\underline{\Sigma}(t+1|t)$ in an explicit form.

We shall compute $\hat{\underline{x}}(t+1|t)$ by induction, supposing that $\hat{\underline{x}}(t|t-1)$ is known. The conditional expectation of (\underline{I}_t) with respect to $\underline{z}(t_0)$, ..., $\underline{z}(t)$ may be decomposed into two parts:

- (1) the conditional expectation given $\underline{z}(t_0)$, ..., $\underline{z}(t-1)$, and
- (2) the conditional expectation given

$$\tilde{\underline{z}}(t|t-1) = \underline{z}(t) - \underline{H}(t)\hat{\underline{x}}(t|t-1) = \underline{H}(t)\tilde{\underline{x}}(t|t-1) + \underline{v}(t). \quad (11.3)$$

gaussian

These two sets of random variables are independent; hence the conditional expectations may be computed separately (see (B.1')). Taking conditional expectations on both sides of (\underline{I}_t) with respect to $\underline{z}(t)$, $\underline{z}(t-1)$, ... yields:

$$\begin{aligned} \hat{\underline{x}}(t+1|t) &= \underline{g}(t+1, t)\hat{\underline{x}}(t|t-1) + \underline{\Gamma}(t+1, t)\hat{\underline{w}}(t|t-1) \\ &\quad + \underline{E}\{\underline{x}(t+1)|\tilde{\underline{z}}(t|t-1)\}. \end{aligned} \quad (11.4)$$

We have seen already that $\hat{\underline{w}}(t|t-1) = \underline{0}$.

We compute the conditional expectation in (11.4) with the aid of Theorem (B.6). For this purpose, we need two covariance matrices. The first of these is

$$\text{cov}\{\tilde{\underline{z}}(t|t-1)\} = \text{cov}\{\underline{H}(t)\tilde{\underline{x}}(t|t-1) + \underline{v}(t)\};$$

since $\underline{v}(t)$ and $\underline{x}(t)$ are independent,

$$= \underline{H}(t)\underline{\Sigma}(t|t-1)\underline{H}'(t) + \underline{\Sigma}(t). \quad (11.5)$$

The other matrix is, by (I_d) ,

$$\begin{aligned} \text{cov}[\underline{x}(t+1), \tilde{\underline{z}}(t|t-1)] &= \text{cov}[\underline{a}(t+1, t)\underline{x}(t) + \underline{\Gamma}(t+1, t)\underline{w}(t), \tilde{\underline{z}}(t|t-1)], \\ &= \text{cov}[\underline{a}(t+1, t)\tilde{\underline{x}}(t|t-1), \tilde{\underline{z}}(t|t-1)], \\ &+ \text{cov}[\underline{\Gamma}(t+1, t)\underline{w}(t), \underline{z}(t)], \\ &= \underline{a}(t+1, t)\underline{\Sigma}(t|t-1)\underline{H}'(t) + \underline{\Gamma}(t+1, t)\underline{C}(t). \end{aligned} \quad (11.6)$$

Hence

$$\begin{aligned} E[\underline{x}(t+1) | \tilde{\underline{z}}(t|t-1)] &= E[\underline{x}(t+1|t-1), \tilde{\underline{z}}(t|t-1)] \\ &= [\underline{a}(t+1, t)\underline{\Sigma}(t|t-1)\underline{H}'(t) + \underline{\Gamma}(t+1, t)\underline{C}(t)] [\underline{H}(t)\underline{\Sigma}(t|t-1)\underline{H}'(t) + \underline{R}(t)]^{-1} \tilde{\underline{z}}(t|t-1). \end{aligned} \quad (11.7)$$

Combining (11.5-7), we obtain the equations of the optimal filter:

$$\hat{\underline{x}}(t+1|t) = \underline{v}(t+1, t)\hat{\underline{x}}(t|t-1) + \underline{K}(t)\underline{z}(t),$$

where

$$\underline{K}(t) = [\underline{a}(t+1, t)\underline{\Sigma}(t|t-1)\underline{H}'(t) + \underline{\Gamma}(t+1, t)\underline{C}(t)] [\underline{H}(t)\underline{\Sigma}(t|t-1)\underline{H}'(t) + \underline{R}(t)]^{-1} \quad (11.8)$$

$$\underline{v}(t+1, t) = \underline{a}(t+1, t) - \underline{K}(t)\underline{H}(t).$$

Of course, the initial state $\hat{\underline{x}}(t_0|t_0-1)$ of (11.8) must be specified also. This is to be taken as zero, since initially there are no observations and the mean of $\underline{x}(t_0)$ is zero.

The general block diagram of the filter is shown in Fig. 5. It is a feedback system built around the model of the random sequence (I_d) . The error signal $\tilde{\underline{z}}(t|t-1)$ is fed forward into the model with gain $\underline{K}(t)$. The gain is such that the input to the model is the conditional expectation of $\underline{x}(t+1)$ given the observed difference $\underline{z}(t) - \hat{\underline{z}}(t|t-1)$. One part of this conditional expectation is due to estimating $\tilde{\underline{x}}(t+1|t-1)$, and the other part is due to estimating $\underline{w}(t)$.

The value of $\hat{x}(t+1|t)$ is known immediately after time t , but it is not needed for computing the next estimate until time $t+1$. This time delay makes it possible to perform the computations indicated by (II_d).

The magnitude of $K(t)$ is indicative of the amount of information contained in the signal $\tilde{z}(t|t-1)$ about the state $x(t+1)$. This property of $K(t)$ can be made precise because the quantity of information in the sense of Shannon can be explicitly calculated for gaussian random processes [25]. One can then show [26] that $K(t)$ is to be determined in such a way as to maximize the information conveyed by $\tilde{z}(t|t-1)$ about $x(t+1)$.

We complete the solution of the filtering problem by deriving a recursion relation for the conditional covariance matrix $\Sigma(t|t-1)$, which is the only remaining unknown in (II_d). This can be obtained by inspection from Theorem (B.13), remembering that

$$\begin{aligned} \text{cov}[\tilde{z}(t+1|t-1)] &= \text{cov}[\Phi(t+1, t)\tilde{x}(t|t-1) + \Gamma(t+1, t)v(t)], \\ &= \Phi(t+1, t)\Sigma(t|t-1)\Phi'(t+1, t) \\ &\quad + \Gamma(t+1, t)Q(t)\Gamma'(t+1, t). \end{aligned}$$

Thus, by (B.13)

$$\begin{aligned} \Sigma(t+1|t) &= \Phi(t+1, t) \{ \Sigma(t|t-1) - [\Sigma(t|t-1)H'(t) + \Gamma(t+1, t)C(t)] [H(t)\Sigma(t|t-1)H'(t) + R(t)]^{-1} \\ &\quad \times [H(t)\Sigma(t|t-1) + C'(t)\Gamma'(t+1, t)] M^{-1}(t-1, t) \\ &\quad + \Gamma(t+1, t)Q(t)\Gamma'(t+1, t) \}. \end{aligned} \quad (\text{III}_d)$$

We shall call (III_d) the variance equation.

Several features of this equation are noteworthy.

First, the equation does not involve the observations $z(t)$. This is a special property of the multivariate gaussian distributions: the conditional covariance matrix does not depend on the values of the conditioning variables. Since the gains of the optimal filter are governed by the variance equation, this means that the structure of the optimal filter (i.e., its element values) can be determined independently of the random data $z(t)$.

Second, equations (II_d - III_d - IV_d) together completely determine the conditional distribution of the random sequence for all $t \geq t_0$, given

$\underline{x}(t_0), \dots, \underline{x}(t)$. In other words, the quantities $\hat{\underline{x}}(t|t-1)$ and $\underline{\Sigma}(t|t-1)$ appearing in (II_q - III_q) may be regarded as the state of the filtering problem: the conditional distributions can be specified by a finite number of parameters. This happy state of affairs is due to the gaussian and markovian assumptions. There are no other cases known at present where the conditional distributions can be specified with comparable simplicity; this is precisely where the basic difficulties of the nonlinear prediction and filtering problems lie.

Third, the variance equation is just another form of the celebrated Wiener-Hopf equation [1-3]. (See [5] for a detailed discussion of the vector form of the Wiener-Hopf equation in the continuous case.) This equation states that $\hat{\underline{x}}(t|t-1)$ and $\tilde{\underline{x}}(t|t-1)$ ^{are} uncorrelated (orthogonal) random variables; in other words, the variances of $\hat{\underline{x}}$ and $\tilde{\underline{x}}$ add. The variance equation is just one of many ways of expressing the same thing. The variance equation for random processes can be derived directly from the Wiener-Hopf equation as was done in [5]. The variance equation is also closely related to the calculus of variations, as will be discussed further in Sect. 15.

Fourth, the solution of the variance equation is not determined until the initial state $\underline{\Sigma}(t_0|t_0-1)$ is given. This should be regarded as part of the problem statement, since obviously $\underline{\Sigma}(t_0|t_0-1) = \underline{\Sigma}(t_0) = \text{cov}[\underline{x}(t_0)]$. To avoid any possible misunderstanding, let us mention how $\underline{\Sigma}(t_0|t_0-1)$ is determined in the conventional Wiener theory. There it is assumed that the random sequence is stationary, in other words, $\underline{\omega}(t+1, t)$, $\underline{\Gamma}(t+1, t)$, $\underline{H}(t)$, $\underline{Q}(t)$, $\underline{R}(t)$, $\underline{C}(t)$ are constants; moreover $\underline{\omega}$ is a stable matrix. Then

$$\underline{x}(t) = \sum_{\tau=-\infty}^{t-1} \underline{\omega}(t, \tau+1) \underline{\Gamma}(\tau+1, \tau) \underline{w}(\tau) \quad (11.8)$$

is a well-defined random vector with zero mean whose covariance matrix \underline{S} is independent of t and can be readily calculated. Thus $\underline{\Sigma}(t_0|t_0-1) = \underline{S}$; while not explicitly given, the value of $\underline{\Sigma}(t_0|t_0-1)$ is implied by the assumptions of the problem. Finally, if $\underline{\Sigma}(t_0)$ is nonnegative definite, then $\underline{\Sigma}(t+1|t)$ is also nonnegative definite for all $t \geq t_0$. This is obvious since $\underline{\Sigma}(t+1|t)$ is a covariance matrix.

(11.9) SOLUTION OF THE FILTERING PROBLEM FOR RANDOM SEQUENCES.

Under the assumptions of Sect. 9, the solution consists of calculating the conditional expectations and conditional covariances, by means of equations (II_d - III_d - IV_d).

The conditional means are computed by the 'optimal filter' (II_d) which is a feedback system with its input being the observations $z(t)$. The initial state of the filter is $\hat{x}(t_0 | t_0 - 1) = 0$.

The conditional variances are solutions of the variance equation (III_d) and are calculated independently of the observations $z(t)$. The conditional variances determine the gain $K(t)$ of the optimal filter. The initial state $\Sigma(t_0)$ of the variance equation is given as part of the problem statement.

The solution of the filtering problem is given in a convenient form only if $t_1 \geq t$; then $\hat{x}(t+1|t)$ and $\Sigma(t+1|t)$ contain all necessary information for computing the conditional probability distributions of the future of the random sequence $x(t)$.

This result was first obtained by Kalman [4] in 1954 except for a slightly less general problem statement and the unnecessary assumption that the inverse of the covariance matrix of $\tilde{z}(t|t-1)$ exists. The latter difficulty is now eliminated by the use of the pseudo-inverse.

In the conventional Wiener problem we assume that Φ, Γ, H, Q, R, C are constants; in addition, t_0 is taken as $-\infty$. In this case the variance equation (III_d) should have a constant nonnegative definite solution (equilibrium state) $\bar{\Sigma}$ to which will correspond a constant gain \bar{K} and therefore a constant optimal filter. In Sect. 16 we shall discuss the conditions under which $\bar{\Sigma}$ exists, is unique, and is the limit of every solution of the variance equation (III_d) which starts at a nonnegative definite initial state. We hasten to point out already here that this is always the case if we add the last remaining assumption of the Wiener theory: the model is asymptotically stable. Hence under the conventional assumptions, the solution of the Wiener problem reduces to the determination of the unique equilibrium state $\bar{\Sigma}$ of (III_d) which is purely an algebraic problem involving the solution of simultaneous quadratic equations. This can be carried out explicitly only in simple cases and will be discussed extensively in Sect. 12.

The chief remaining task in filtering theory is the study of the variance equation. This is difficult because the equation is nonlinear. The problem can be best appreciated from the study of detailed examples. Two of these are given in Sect. 12. A summary of what is known about the qualitative behavior of the variance equation appears in Sects. 15-16.

12. Examples of discrete filtering. In the two examples discussed here we assume for simplicity that $Q(t)$ is identically zero. This will not entail a great loss of generality. We shall write $\underline{\Sigma}(t)$ and $\sigma_{ij}(t)$ instead of $\underline{\Sigma}(t|t-1)$ and $\sigma_{ij}(t|t-1)$ to save space.

The simplest possible case is the following:

(12.1) EXAMPLE. Consider a constant, first-order model. Setting the constants γ_{11} and h_{11} equal to 1, we have:

$$\left. \begin{aligned} x_1(t+1) &= \phi_{11}^2 x_1(t) + v_1(t), \\ z_1 &= x_1(t) + v_1(t). \end{aligned} \right\} \quad (12.2)$$

The variance equation follows by inspection from (III_d):

$$\sigma_{11}(t+1) = \phi_{11}^2 \left[\sigma_{11}(t) - \frac{\sigma_{11}^2(t)}{\sigma_{11}(t) + r_{11}} \right] + q_{11}. \quad (12.3)$$

The equation of the optimal filter is:

$$\hat{x}_1(t+1|t) = \phi_{11}^2 (\hat{x}_1(t|t-1)) + \frac{\sigma_{11}(t)}{\sigma_{11}(t) + r_{11}} [z_1(t) - \hat{x}_1(t|t-1)] \quad (12.4)$$

There are several cases of interest, depending on the values of the parameters ϕ_{11} , q_{11} , and r_{11} .

Case (1): $r_{11} = 0$. Equation (12.3) immediately reduces to $\sigma_{11}(t) = q_{11} = \text{const. for } t > t_0$. Therefore

$$\hat{x}_1(t+1|t) = z_1(t).$$

In other words, the filter has no memory and the best estimate is the last piece of data.

In all other cases, the transient behavior of $\sigma_{11}(t)$ will be more complicated. To analyze it, let $\bar{\sigma}_{11}$ stand for an equilibrium point of system (12.3), defined by

$$\bar{\sigma}_{11} = \left(\frac{\varphi_{11}^2 r_{11}}{\bar{\sigma}_{11} + r_{11}} \right) \bar{\sigma}_{11} + q_{11}. \quad (12.5)$$

Of course the requirement $\bar{\sigma}_{11} \geq 0$ must be satisfied also.

We define deviations from equilibrium by

$$\delta\sigma_{11}(t) = \sigma_{11}(t) - \bar{\sigma}_{11}.$$

Substituting this into (12.3) and using (12.5) gives

$$\delta\sigma_{11}(t+1) = \left(\frac{\varphi_{11}^2 r_{11}}{\bar{\sigma}_{11} + r_{11}} \right) \left(\frac{r_{11}}{\sigma_{11}(t) + r_{11}} \right) \delta\sigma_{11}(t). \quad (12.6)$$

We are now ready to discuss the remaining cases.

Case (ii): $r_{11} > 0$, $q_{11} = 0$, $|\varphi_{11}| \leq 1$. Equation (12.5) has only one solution, which is $\bar{\sigma}_{11} = 0$. If $\sigma_{11}(t) = \delta\sigma_{11}(t) > 0$, then the factor on the right-hand side of (12.6) is always positive and less than 1. Hence $\delta\sigma_{11}(t)$ decreases monotonically, and all solutions of (12.3) converge to 0 if they start at $\sigma_{11}(t_0) \geq 0$. Negative values of $\sigma_{11}(t_0)$ are of course ruled out.

Case (iii): $r_{11} > 0$, $q_{11} = 0$, $|\varphi_{11}| > 1$. Now (12.5) has two solutions: $\bar{\sigma}_{11} = 0$ and $\bar{\sigma}_{11} = (\varphi_{11}^2 - 1)r_{11}$. Substitute the second value of $\bar{\sigma}_{11}$ into (12.6); then the factor on the right-hand side of (12.6) is less than one. Thus $|\sigma_{11}(t) + (\varphi_{11}^2 - 1)r_{11}|$ decreases monotonically and every solution of (12.3) with $\sigma_{11}(t_0) > 0$ converges to the second equilibrium point. The only exception is the solution $\sigma_{11}(t) \equiv \sigma_{11}(t_0) = 0$, which is an unstable equilibrium point.

Case (iv): $r_{11} > 0$, $q_{11} > 0$. Now equation (12.5) has a single solution $\bar{\sigma}_{11}$. The first factor of (12.6) is less than one as a consequence of (12.5). The second factor is less than or equal to 1. $|\delta\sigma_{11}(t)|$ decreases monotonically and all solutions converge to the unique equilibrium point $\bar{\sigma}_{11}$.

What can be said about the stability of the optimal filter? In Case (i), this question is vacuous. Otherwise the 1×1 transition matrix of the optimal filter is:

$$\psi_{11}(t+1, t) = \phi_{11} r_{11} / [\sigma_{11}(t) + r_{11}].$$

In Case (ii), ψ_{11} tends to ϕ_{11} as $t \rightarrow \infty$; in Case (iii), ψ_{11} tends to $1/\phi_{11}$. In both cases, the optimal filter is asymptotically stable unless $|\phi_{11}| = 1$. In Case (iv), the optimal filter is always asymptotically stable, since by (12.4)

$$\lim_{t \rightarrow \infty} |\psi_{11}(t+1, t)| = \left| \frac{\phi_{11} r_{11}}{\sigma_{11} + r_{11}} \right| < 1.$$

(12.7) EXAMPLE. When the model (12.2) of the random sequence is non-constant, the discussion is similar but much less elementary. The main point is this: we must assume that the parameters $|\phi_{11}(t+1, t)|$, $q_{11}(t)$ and $r_{11}(t)$ describing the model are roughly of the same order of magnitude at all instants of time; in other words, they cannot become arbitrarily large or arbitrarily small. A convenient condition assuring this is:

$$\begin{aligned} 0 < \alpha_1 \leq |\phi_{11}(t+1, t)| \leq \beta_1 < \infty, \\ 0 < \alpha_2 \leq q_{11}(t) \leq \beta_2 < \infty, \\ 0 < \alpha_3 \leq r_{11}(t) \leq \beta_3 < \infty. \end{aligned} \quad (12.8)$$

We shall assume (12.8) for the sequel. What happens when these conditions are not met remains an open problem.

An immediate consequence of (12.8) is that, even though $\sigma_{11}(t_0) > 0$ is arbitrary,

$$\alpha_2 \leq \sigma_{11}(t) \leq \beta_1^2 \beta_3 + \beta_2 \quad \text{for all } t \geq t_0 + 1 \quad (12.9)$$

In other words, the solutions of the variance equation are uniformly bounded. Using the variance equation, one gets immediately the inequality

$$\frac{\varphi_{11}^2(t+1, t)r_{11}(t)}{\sigma_{11}(t) + r_{11}(t)} \leq \left[1 - \frac{\alpha_{11}(t)}{\sigma_{11}(t+1)} \right] \frac{\sigma_{11}(t+1)}{\sigma_{11}(t)}.$$

By (12.9), the bracketed term is bounded by

$$\lambda^2 = 1 - \frac{\alpha_2}{\beta_1^2 \beta_3 + \beta_2} < 1,$$

and therefore

$$|\psi_{11}(t+1, t)| = \left| \frac{\varphi_{11}(t+1, t)r_{11}(t)}{\sigma_{11}(t) + r_{11}(t)} \right| \leq \lambda \sqrt{\frac{\sigma_{11}(t+1)}{\sigma_{11}(t)}}.$$

Iterating this relation and again using (12.9), we obtain

$$|\psi_{11}(t, t_0)| \leq \lambda^{t-t_0} \sqrt{\frac{\sigma_{11}(t)}{\sigma_{11}(t_0)}} \leq \sqrt{\frac{\beta_1^2 \beta_3 + \beta_2}{\alpha_2}} \lambda^{t-t_0},$$

which proves that the optimal filter is uniformly asymptotically stable.

Now let $\sigma_{11}^{(a)}(t)$ and $\sigma_{11}^{(b)}(t)$ be any two solutions of the variance equation. Let

$$\delta\sigma_{11}(t) = \sigma_{11}^{(a)}(t) - \sigma_{11}^{(b)}(t)$$

be the difference between these two solutions. Then

$$\begin{aligned} r_{11}(t+1) &= \left(\frac{\varphi_{11}(t+1, t)r_{11}(t)}{\sigma_{11}^{(a)}(t) + r_{11}(t)} \right) - \left(\frac{\varphi_{11}(t+1, t)r_{11}(t)}{\sigma_{11}^{(b)}(t) + r_{11}(t)} \right) \delta\sigma_{11}(t), \\ &= \psi_{11}^{(a)}(t+1, t) \psi_{11}^{(b)}(t+1, t) \delta\sigma_{11}(t), \end{aligned} \quad (12.10)$$

and the preceding results shows that the difference between any two solutions of the variance equation will tend to zero uniformly with t. This means that

every solution of the variance equation will tend toward some particular solution $\bar{\sigma}_{11}(t)$ contained in the region (12.9). This solution is conveniently defined by taking its starting point $\sigma_{11}(t_0) = 0$ and then letting $t_0 \rightarrow \infty$. The function $\bar{\sigma}_{11}(t)$ ^{so obtained*} may be regarded as the "moving" equilibrium state of the variance equation.

We carried out the discussion in so much detail in order to indicate the method of proof in the general case. Even though the variance equation is nonlinear, its transient behavior can be studied conveniently by means of formula (12.10) and its generalizations. See Sect. 16.

The next example concerns a second-order model; this slight increase of complexity makes the explicit discussion quite involved, even for the steady-state behavior.

(12.11) **EXAMPLE.** Consider the random sequence $x(t)$ generated in the following fashion:

$$x(t) = k(t) + m(t),$$

where

$$k(t+1) = k(t) + w_1(t),$$

$$m(t+1) = m(t) + n(t),$$

$$n(t+1) = n(t) + w_2(t);$$

$w_1(t)$, $w_2(t)$ are gaussian white-noise sequences with zero mean. In other words, $x(t)$ is the sum of two random sequences: one with independent gaussian random increments (first differences), and one with independent gaussian random second differences. Moreover, values of $x(t)$ are measured with an error $v_1(t)$ which is also a gaussian white-noise sequence with zero mean. Thus

$$z_1(t) = x(t) + v_1(t)$$

It is easy to see that $x_1 = x$ and $x_2 = n$ is a suitable definition of the state variables in this case. The matrices in (I_d) are:

* The limit $\lim_{t \rightarrow -\infty} \sigma_{11}(t; 0, t_0)$ always exists: see Sect. 16.

$$\underline{F} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \underline{r} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \underline{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

The variance equations are:

$$\sigma_{11}(t+1) = \sigma_{11}(t) + 2\sigma_{12}(t) + \sigma_{22}(t) - \frac{[\sigma_{11}(t) + \sigma_{12}(t)]^2}{\sigma_{11}(t) + r_{11}} + q_{11}$$

$$\sigma_{12}(t+1) = \sigma_{12}(t) + \sigma_{22}(t) - \frac{\sigma_{12}(t)[\sigma_{11}(t) + \sigma_{12}(t)]}{\sigma_{11}(t) + r_{11}}$$

$$\sigma_{22}(t+1) = \sigma_{22}(t) - \frac{[\sigma_{12}(t)]^2}{\sigma_{11}(t) + r_{11}} + q_{22}$$

The optimal filter is given by

$$\hat{x}_1(t+1|t) = \hat{x}_1(t|t-1) + \hat{x}_2(t|t-1) = \frac{\sigma_{11}(t) + \sigma_{12}(t)}{\sigma_{11}(t) + r_{11}} [z_1(t) - \hat{x}_1(t|t-1)]$$

$$\hat{x}_2(t+1|t) = \hat{x}_2(t|t-1) + \frac{\sigma_{12}(t)}{\sigma_{11}(t) + r_{11}} [z_1(t) - \hat{x}_1(t|t-1)]$$

The detailed analysis of this example is so tedious that we shall consider only the steady-state behavior. In other words, we shall analyze the equilibrium states of the variance equations, given by

$$(2\bar{\sigma}_{12} + \bar{\sigma}_{22} + q_{11})(\bar{\sigma}_{11} + r_{11}) = (\bar{\sigma}_{11} + \bar{\sigma}_{12})^2, \quad (12.12)$$

$$\bar{\sigma}_{22}(\bar{\sigma}_{11} + r_{11}) = \bar{\sigma}_{12}(\bar{\sigma}_{11} + \bar{\sigma}_{12}), \quad (12.13)$$

$$q_{22}(\bar{\sigma}_{11} + r_{11}) = \bar{\sigma}_{12}^2, \quad (12.14)$$

and subject to the condition that the steady-state variance matrix $[\bar{\sigma}_{ij}]$ be nonnegative definite, which will be true if and only if

$$\sigma_{11} \geq 0, \quad (12.15)$$

$$\bar{\sigma}_{11}\bar{\sigma}_{22} - \bar{\sigma}_{12}^2 \geq 0. \quad (12.16)$$

To avoid discussing cumbersome special cases, we assume that q_{11}, q_{22}, r_{11} are all positive. Introducing the abbreviations

$$\alpha = q_{11}/r_{11} > 0, \quad \beta = q_{22}/r_{11} > 0;$$

$$\xi = \bar{\sigma}_{11}/r_{11}, \quad \eta = \bar{\sigma}_{12}/\sqrt{q_{22}r_{11}}, \quad \zeta = \bar{\sigma}_{22}/r_{11},$$

eliminating ξ and ζ , relations (12.14, 12.13, 12.12, 12.15, 12.16) become respectively

$$\xi = \eta^2 - 1, \quad (12.17)$$

$$\eta\zeta = \sqrt{\beta}(\eta^2 - 1 + \sqrt{\beta}\eta) \quad (12.18)$$

$$\eta^4 = \sqrt{\beta}\eta^3 - (2 + \alpha)\eta^2 - \sqrt{\beta}\eta + 1 = 0, \quad (12.19)$$

$$\xi \geq 0, \quad (12.20)$$

$$\xi\zeta \geq \beta\eta^2. \quad (12.21)$$

By (12.17) and (12.20) $\eta^2 \geq 1$. This, (12.21), and $\beta > 0$ imply that $\xi > 0$ and $\zeta > 0$. Combining (12.17) and (12.18), (12.21) becomes

$$\eta\left(\eta - \frac{1}{\eta}\right)^2 \geq \sqrt{\beta}. \quad (12.22)$$

With (12.17), this yields

$$\eta > 1. \quad (12.23)$$

Turning now to (12.19), we notice the symmetry of the coefficients. This means that if η is a root then $1/\eta$ is also a root. Zero is never a

root. Denoting by $\eta_1, 1/\eta_1, \eta_2, 1/\eta_2$ the four roots of (12.19), we obtain the following conditions:

$$(\eta_1 + 1/\eta_1) + (\eta_2 + 1/\eta_2) = \sqrt{\beta}, \quad (12.24)$$

$$2 + (\eta_1 + 1/\eta_1) \cdot (\eta_2 + 1/\eta_2) = -(2 + \alpha).$$

This is equivalent to a quadratic equation in the unknown $\eta_1 + 1/\eta_1$, which has the solution:

$$\eta_1 + 1/\eta_1 = \frac{1}{2}(\sqrt{\beta} \pm \sqrt{16 + 4\alpha + \beta}) \quad (12.25)$$

In view of (12.23) we must choose the + sign; the minus sign will then correspond to $\eta_2 + 1/\eta_2$ in (12.24). Solving (12.25) for η_1 we get

$$\eta_1 = \frac{1}{4}(\sqrt{\beta} + \sqrt{16 + 4\alpha + \beta} \pm \sqrt{2\beta + 4\alpha + 2\sqrt{16 + 4\alpha + \beta}\beta})$$

The root corresponding to the - sign is the reciprocal of the root corresponding to the + sign. In view of (12.23) we must choose the larger root, so that

$$\eta = \frac{1}{4}(\sqrt{\beta} + \sqrt{16 + 4\alpha + \beta} + \sqrt{2\beta + 4\alpha + 2\sqrt{(16 + 4\alpha + \beta)\beta}}). \quad (12.26)$$

is the only root of (12.19) which could lead to a positive definite matrix.

It remains now only to check whether (12.22) holds. By (12.26), and (12.25) we have

$$\left(\eta - \frac{1}{\eta}\right)^2 = \frac{1}{2}(2\alpha + \beta + \sqrt{(16 + 4\alpha + \beta)\beta}) \geq 2\sqrt{\beta} > \sqrt{\beta}/\eta.$$

Hence we have proved that there exists one and only one solution of (12.19) which is nonnegative definite; this solution is actually positive definite.

and is given by (12.17, 12.18, 12.26). It can be shown (see Sect. 16) that all solutions of the variance equation converge to the equilibrium state $(\bar{q}_{11}, \bar{q}_{12}, \bar{q}_{22})$.

Although this problem appears to be quite elementary, the author is not aware of any detailed study of it in the literature. As a matter of fact, in a recent note to the Soviet Academy, A. L. Brundno and A. L. Lants [27] erroneously assert (without proof) that the solution of this problem is not unique ^{when} $q_{12} = 4q_{22}$.

13. Solution of the filtering problem for random processes. The main object of this section is to establish relations analogous to (II₂) and (III₂). A rigorous proof of this must be preceded by a rigorous definition of the white-noise processes in (I₂). We shall not do this here but will appeal to the semi-rigorous limiting arguments already used in Sect. 6. A different derivation (rigorous except for the use of delta functions in the definition of the covariance matrices of white-noise processes) may be found in [5].

As in Sect. 6, let q be a positive integer and let the time t be discrete so that its successive values differ by q^{-1} . Then, assuming that \underline{F} is the transition matrix of a continuous-time linear dynamical system and $\underline{\Gamma}$ is given by (4.10), we have

$$\underline{q}(t + q^{-1}, t) = \underline{I} + q^{-1}\underline{F}(t) + o(q^{-1}),^*$$

$$\underline{\Gamma}(t + q^{-1}, t) = q^{-1}\underline{G}(t) + o(q^{-1}).^*$$

In view of the discussion of Sect. 6, the covariance matrices $\underline{C}(t)$, $\underline{Q}(t)$, and $\underline{R}(t)$ in (III₂) are to be replaced by

$$q^{-1}\underline{C}(t), \quad q^{-1}\underline{Q}(t), \quad \text{and} \quad q^{-1}\underline{R}(t)$$

as $q \rightarrow \infty$.

* The symbol $o(q^{-1})$ denotes a matrix which is zero in the limit $q = \infty$.

Substituting these expressions in (III_q), we obtain:

$$\frac{\underline{\Sigma}(t + q^{-1}|t) - \underline{\Sigma}(t|t - q^{-1})}{q^{-1}} = \underline{F}(t)\underline{\Sigma}(t|t - q^{-1}) + \underline{\Sigma}(t|t - q^{-1})\underline{F}'(t) \\ + [\underline{\Sigma}(t|t - q^{-1})\underline{H}'(t) + \underline{Q}(t)\underline{C}(t)][q^{-1}\underline{H}(t)\underline{\Sigma}(t|t - q^{-1})\underline{H}'(t) + \underline{H}(t)]^\dagger [\underline{H}(t)\underline{\Sigma}(t|t - q^{-1}) + \underline{C}'(t)\underline{G}'(t)] \\ + \underline{Q}(t)\underline{Q}(t)\underline{G}'(t) + \underline{Q}(q^{-1}).$$

Since $(\alpha\underline{A})^\dagger = \alpha^{-1}\underline{A}^\dagger$ if $\alpha \neq 0$ but $\underline{0}^\dagger = \underline{0}$, we must be careful not to introduce a discontinuity in the term $[\dots]^\dagger$ while taking the limit $q \rightarrow \infty$. The trouble is most easily avoided by assuming, once and for all, that

$$\underline{R}(t) \text{ is positive definite for all } t. \quad (13.1)$$

Passing to the limit $q \rightarrow \infty$, we obtain the variance equation:

$$d\underline{\Sigma}/dt = \underline{F}(t)\underline{\Sigma} + \underline{\Sigma}\underline{F}'(t) - [\underline{H}'(t) + \underline{Q}(t)\underline{C}(t)]\underline{R}^{-1}(t)[\underline{H}(t)\underline{\Sigma} + \underline{C}'(t)\underline{G}'(t)] \\ + \underline{G}(t)\underline{Q}(t)\underline{G}'(t), \quad (III_c)$$

whose solution is the covariance matrix $\underline{\Sigma}(t|t)$.

The same limiting process applied to (II_q) yields the equations of the optimal filter in continuous time:

$$\left. \begin{aligned} d\hat{\underline{x}}/dt &= \underline{F}(t)\hat{\underline{x}} + \underline{K}(t)[z(t) - \hat{\underline{z}}], \\ \text{where} \quad \underline{K}(t) &= [\underline{\Sigma}(t|t)\underline{H}'(t) + \underline{Q}(t)\underline{C}(t)]\underline{R}^{-1}(t). \end{aligned} \right\} \quad (II_c)$$

The solution of the above differential equation is the conditional expectation $\hat{\underline{x}}(t|t)$.

We have already noted in Sect. 8 that if $\underline{R}(t)$ is singular (i.e., some linear combination of components of $\underline{y}(t)$ can be observed exactly) then the

optimal filter may be an unbounded operator -- such as differentiation -- which cannot be realized by means of a linear dynamical system. Hence condition (13.1) cannot be readily relaxed, as is clear from the expression for the optimal gain.

The matrix block diagram of the optimal filter is shown in Fig. 6.

Remarks concerning the initial conditions of $(II_d - III_d)$ apply without modification to $(II_c - III_c)$.

Equation (IV_d) generalizes trivially to:

$$\hat{x}(t_1|t) = \Phi(t_1, t)\hat{x}(t|t) \quad \text{for all } t_1 \geq t. \quad (IV_c)$$

Hence we have:

(13.2) SOLUTION OF THE FILTERING PROBLEM FOR RANDOM PROCESSES. Under the assumptions of Sect. 9 and (13.1), the solution consists of calculating the conditional expectations and conditional covariances, by means of equations $(II_c - III_c - IV_c)$.

The conditional variances $\Sigma(t|t)$ are solutions of the variance equation (III_c) and are calculated independently of the observations $z(t)$. The conditional variances determine the gain $K(t)$ of the optimal filter. The initial state $\Sigma(t_0|t_0)$ of the variance equation is given as part of the problem statement.

The solution of the filtering problem is given in convenient form only if $t_1 \geq t$; then $\hat{x}(t|t)$ and $\Sigma(t|t)$ contain all necessary information for computing the conditional probability distributions of the future of the random process $x(t)$.

This result was first published in [5].

Although Theorem (13.2) appears to be completely analogous to Theorem (11.9), there is one major difference: the 'solution' of the problem in Theorem (13.2) is tied to obtaining a solution of the variance equation (III_c) .

Since (III_c) satisfies a Lipschitz condition, it follows [14-15] that solutions of (III_c) will exist for arbitrary $\underline{\Sigma}(t_0|t_0)$ in some small interval of time containing t_0 . But it is not clear without further investigation that solutions exist for all $t \geq t_0$. (As a matter of fact, this may not even be true for arbitrary $\underline{\Sigma}(t_0|t_0)$.) However, we can readily show:

(13.3) If $\underline{\Sigma}(t_0|t_0)$ is nonnegative definite, then (III_c) has a unique solution which exists for all $t \geq t_0$.

This may be proved as follows. Let $\underline{\Sigma}(t)$ be the covariance matrix of $\underline{x}(t)$ defined by (I_c) , if the covariance matrix of the initial state $\underline{x}(t_0)$ is $\underline{\Sigma}(t_0) = \underline{\Sigma}(t_0|t_0)$. Utilizing the formulas of Sect. 4 and recalling that $E\{\underline{x}(t)\} = \underline{0}$, we have

$$\begin{aligned} \underline{\Sigma}(t) &= \text{cov}[\underline{x}(t)], \\ &= \underline{\Phi}(t, t_0) \underline{\Sigma}(t_0) \underline{\Phi}'(t, t_0) \\ &\quad + E\left\{ \int_{t_0}^t d\tau \int_{t_0}^t d\tau' \underline{\Phi}(t, \tau) \underline{G}(\tau) \underline{W}(\tau) \underline{W}'(\tau') \underline{G}'(\tau') \underline{\Phi}'(t, \tau') \right\}, \\ &= \underline{\Phi}(t, t_0) \underline{\Sigma}(t_0) \underline{\Phi}'(t, t_0) + \int_{t_0}^t \underline{\Phi}(t, \tau) \underline{G}(\tau) \underline{Q}(\tau) \underline{G}'(\tau) \underline{\Phi}'(t, \tau) d\tau. \end{aligned} \quad (13.4)$$

This shows that $\underline{\Sigma}(t)$ is bounded whenever $t \geq t_0$. But the definition of conditional covariance matrix (see B.13) shows that

$$\underline{\Sigma}(t|t) \leq \underline{\Sigma}(t) \quad \text{for all } t \geq t_0; \quad (13.5)$$

in other words, it is known a priori that solutions of the variance equation must be bounded for any $t \geq t_0$. Substituting this fact into the standard existence proof of solutions of nonlinear differential equations [15] proves (13.3). It should be noted that the argument leading to the inequality (13.5) does not hold if $\underline{\Sigma}$ is not a covariance matrix, in other words, if $\underline{\Sigma}(t_0)$ is indefinite.

At first sight, it may appear that the solution of the nonlinear differential equation (III₀) would in general require numerical quadrature -- a disagreeable prospect because of the $n(n+1)/2$ variables involved (which are the distinct elements of the $n \times n$ symmetric matrix $\underline{\Sigma}$). But (III₀) is not a general nonlinear differential equation: it is a very special one, the matrix riccati equation, which is well-known from the calculus of variations. We shall utilize this fact to derive an exact formula for the solutions of (III₀).

Consider the hamiltonian function \mathcal{H} defined by

$$2\mathcal{H}(\underline{x}, \underline{p}, t) = -\|\underline{Q}'(t)\underline{x}\|_{\underline{Q}(t)}^2 - 2\underline{p}'\underline{F}'(t)\underline{x} + \|\underline{H}(t)\underline{C}'(t)\underline{G}'(t)\underline{x}\|_{\underline{R}^{-1}(t)}^2 \quad (13.6)$$

and the associated canonical differential equations of Hamilton:

$$d\underline{x}/dt = \partial\mathcal{H}/\partial\underline{p}^* = -\underline{F}'(t)\underline{x} + \underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)\underline{p} + \underline{H}'(t)\underline{R}^{-1}(t)\underline{C}'(t)\underline{G}'(t)\underline{x}$$

$$d\underline{p}/dt = -\partial\mathcal{H}/\partial\underline{x}^* = \underline{G}(t)\underline{Q}(t)\underline{C}'(t)\underline{x} + \underline{F}(t)\underline{p} - \underline{G}(t)\underline{C}(t)\underline{R}^{-1}(t)\underline{H}(t)\underline{p} \\ - \underline{G}(t)\underline{C}(t)\underline{R}^{-1}(t)\underline{C}'(t)\underline{G}(t)\underline{x}.$$

Let $\underline{X}(t)$, $\underline{P}(t)$ denote the unique pair of matrix solutions of this equation corresponding to the initial conditions

$$\underline{X}(t_0) = \underline{I} \quad \text{and} \quad \underline{P}(t_0) = \underline{\Sigma}(t_0|t_0). \quad (13.7)$$

Then we have the identity

$$\underline{p}(t) = \underline{\Sigma}(t|t)\underline{X}(t) \quad \text{for all } t \geq t_0. \quad (13.8)$$

which can be easily verified by substituting (13.9) and (III₀) into (13.7).

We see then that $\underline{X}(t)$ satisfies the differential equation

$$d\underline{X}(t)/dt = [-\underline{F}'(t) + \underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)\underline{\Sigma}(t|t) + \underline{H}'(t)\underline{R}^{-1}(t)\underline{C}'(t)\underline{Q}'(t)]\underline{X}(t), \quad (13.9)$$

which is defined for all $t \geq t_0$ because of (13.3). (In fact, (13.10) is the adjoint of the differential equation of the optimal filter.) In view of

* The components of the vector $\partial\mathcal{H}/\partial\underline{x}$ are $\partial\mathcal{H}/\partial x_i$.

(13.9), it follows that $\underline{X}(t)$ is the transition matrix of (V_c) and thus $\underline{X}(t)$ is never singular for $t \geq t_0$. Hence (13.9) becomes

$$\underline{\Sigma}(t|t) = \underline{P}(t)\underline{X}^{-1}(t). \quad (13.10)$$

Let us partition the transition matrix $\underline{\Theta}$ of the $2n \times 2n$ linear system (13.7) into $n \times n$ blocks:

$$\underline{\Theta}(t, t_0) = \begin{bmatrix} \underline{\Theta}_{11}(t, t_0) & \underline{\Theta}_{12}(t, t_0) \\ \underline{\Theta}_{21}(t, t_0) & \underline{\Theta}_{22}(t, t_0) \end{bmatrix}. \quad (13.11)$$

Then (13.11) can be written explicitly as

$$\underline{\Sigma}(t|t) = [\underline{\Theta}_{21}(t, t_0) + \underline{\Theta}_{22}(t, t_0)\underline{\Sigma}(t_0|t_0)] [\underline{\Theta}_{11}(t, t_0) + \underline{\Theta}_{12}(t, t_0)\underline{\Sigma}(t_0|t_0)]^{-1} \quad (13.12)$$

Thus the solutions of the variance equation for $t \geq t_0$ can be expressed exactly in terms of the transition matrix of the hamiltonian system (13.7).

The connection between the matrix riccati equation and the canonical equations of the calculus of variations has been known for a long time [28], but it was relatively unnoticed until recently [29]. To the best of the writer's knowledge, [5] was the first instance in which the relation of the Wiener problem to the classical calculus of variations was explicitly noted.

14. Examples of continuous filtering. The number of cases where it is possible to obtain closed-form solutions of the filtering problem is surprisingly small. We present below some typical samples of these cases; other examples of this sort are discussed in [5]. Being too simple, the examples to be discussed here are of very limited practical interest. But they are useful in conveying insight into the behavior of the variance equation and they serve as a useful guide in obtaining general results, such as those presented in Sects. 15 and 16.

We write $\hat{x}(t) \equiv \hat{x}(t|t)$ and $\Sigma(t) \equiv \Sigma(t|t)$ for the sake of simplicity, and assume again that $C(t) \equiv C$.

(14.1) EXAMPLE. What is obviously the simplest filtering problem appears in Fig. 7A. This is a constant, first-order system and was treated in detail already by Wiener [30]. As may be expected after 15 years of progress in the field, the present treatment is a good deal simpler and more general. The discussion is very similar to that of Example (12.1).

The describing matrices can be read off by inspection from Fig. 7A:

$$\underline{F} = [f_{11}], \quad \underline{G} = [1], \quad \underline{H} = [1], \quad \underline{Q} = [q_{11}], \quad \text{and} \quad \underline{R} = [r_{11}].$$

We assume of course that $r_{11} > 0$. Then the variance equation is:

$$d\sigma_{11}/dt = 2f_{11}\sigma_{11} - \sigma_{11}^2/r_{11} + q_{11}. \quad (14.2)$$

The optimal filter is shown in Fig. 7B, where

$$k_{11}(t) = \sigma_{11}(t)/r_{11}.$$

Setting $\dot{\sigma}_{11} = 0$, we see that the equilibrium states of the variance equation are given by the roots of a quadratic:

$$\bar{\sigma}_{11} = (r_{11} \pm \sqrt{r_{11}^2 + q_{11}/r_{11}})r_{11}. \quad (14.3)$$

Since $\bar{\sigma}_{11}$ is a variance, it must be nonnegative; thus we conclude:

(14.4) The variance equation (14.2) has a unique equilibrium state $\bar{\sigma}_{11}$ if $q_{11} > 0$ or if $f_{11} \leq 0$. Otherwise there are two equilibrium states, 0 and $2f_{11}r_{11}$.

In the classical formulation of the Wiener problem, the message process must be stationary. This requires $f_{11} < 0$. Moreover, one assumes of course also that $q_{11} > 0$, since otherwise the variance of the message process would be zero in the steady state. Under these assumptions, the steady-state gain of the optimal filter is given by

$$\bar{K}_{11} = \bar{\sigma}_{11}/r_{11} = f_{11} + \sqrt{r_{11}^2 + q_{11}/r_{11}}, \quad (14.5)$$

and the steady-state optimal filter is described by the equation

$$d\hat{x}_1/dt = \bar{f}_{11}\hat{x}_1 + \bar{K}_{11}z_1(t), \quad (14.6)$$

where

$$\bar{f}_{11} = f_{11} - \bar{K}_{11} = -\sqrt{r_{11}^2 + q_{11}/r_{11}}. \quad (14.7)$$

In particular, (14.6) shows that the optimal filter is always asymptotically stable. These results are well known [1-3].

In accordance with Remark (14.4), these formulas continue to hold if either $f_{11} \leq 0$ or $q_{11} > 0$. If on the other hand $f_{11} > 0$ and $q_{11} = 0$, then there are two possible equilibrium states and it is not obvious at first which of these corresponds to the solution of the filtering problem with $t_0 = -\infty$. Inspection of the first-order nonlinear differential equation (14.2) shows that of the two possible equilibrium states $\bar{\sigma}_{11} = 0$ is always unstable and $\bar{\sigma}_{11} = 2f_{11}r_{11}$ is always stable at $t \rightarrow \infty$. All solutions starting at $\bar{\sigma}_{11}(t_0)$ converge to the second equilibrium state as $t \rightarrow \infty$. The optimal gain \bar{K}_{11} corresponding to the second equilibrium state is positive, and therefore the optimal filter is asymptotically stable. Hence:

(14.8) The optimal filter is asymptotically stable, except perhaps in the trivial case $\sigma_{11}(t_0) = q_{11} = 0$.

Note that stability does not depend on the model itself being stable. The optimal filter always provides feedback around the model so as to make the closed-loop system stable.

In this example, it is easy to obtain an explicit solution of the variance equation. We consider the associated Hamiltonian system (V_c):

$$\begin{aligned} dx_1/dt &= -f_{11}x_1 + (1/r_{11})p_1, \\ dp_1/dt &= q_{11}x_1 + f_{11}p_1. \end{aligned} \quad (14.9)$$

We assume that either $f_{11} \neq 0$ or $q_{11} \neq 0$. The other case is trivial. Then $\bar{f}_{11} < 0$ and the transition matrix of (14.9) is [5]:

$$\Phi(t + \tau, t) = \begin{bmatrix} \cosh \bar{f}_{11}\tau - \frac{f_{11}}{\bar{f}_{11}} \sinh \bar{f}_{11}\tau & \frac{1}{r_{11}\bar{f}_{11}} \sinh \bar{f}_{11}\tau \\ \frac{q_{11}}{\bar{f}_{11}} \sinh \bar{f}_{11}\tau & \cosh \bar{f}_{11}\tau + \frac{f_{11}}{\bar{f}_{11}} \sinh \bar{f}_{11}\tau \end{bmatrix}. \quad (14.10)$$

Applying formula (13.12), we find, for $t \geq 0$,

$$\sigma_{11}(t) = \frac{(q_{11}/\bar{f}_{11})\sinh \bar{f}_{11}(t-t_0) + [\cosh \bar{f}_{11}(t-t_0) + (f_{11}/\bar{f}_{11})\sinh \bar{f}_{11}(t-t_0)]\sigma_{11}(t_0)}{\cosh \bar{f}_{11}(t-t_0) - (f_{11}/\bar{f}_{11})\sinh \bar{f}_{11}(t-t_0) + [(1/r_{11}\bar{f}_{11})\sinh \bar{f}_{11}(t-t_0)]\sigma_{11}(t_0)}.$$

If $\sigma_{11}(t_0) = q_{11} = 0$, then $\sigma_{11}(t)$ vanishes identically. Otherwise, we have, since $\bar{f}_{11} < 0$,

$$\lim_{t \rightarrow \infty} \sigma_{11}(t) = \frac{-q_{11} + (f_{11} - \bar{f}_{11})\sigma_{11}(t_0)}{f_{11} + \bar{f}_{11} - (1/r_{11})\sigma_{11}(t_0)}.$$

By (14.7) and (14.5)

$$\lim_{t \rightarrow \infty} \sigma_{11}(t) = r_{11}(\bar{f}_{11} - f_{11}) = \bar{\sigma}_{11},$$

which checks with the previous conclusions.

Although the preceding developments provide a complete and explicit picture, the resulting formulas are quite complicated and difficult to understand intuitively. More information can be gained by transcribing the qualitative analysis of Example (12.7).

(14.11) **EXAMPLE.** We consider again the system shown in Fig. 7A-B, but now f_{11} , q_{11} , r_{11} are assumed to be functions of time. Analogously to (12.8), we impose the "uniformity" conditions:

$$\begin{aligned} |f_{11}(t)| &\leq \beta_1 < \infty, \\ 0 < \alpha_2 \leq q_{11}(t) &\leq \beta_2 < \infty, \\ 0 < \alpha_3 \leq r_{11}(t) &\leq \beta_3 < \infty. \end{aligned} \tag{14.12}$$

Applying these conditions to the variance equation (14.2), we conclude that

$$\begin{aligned} \dot{\sigma}_{11} > 0 & \text{ if } 0 \leq \sigma < \alpha_k = -\alpha_2\beta_1 + \sqrt{\alpha_2^2\beta_1^2 + \alpha_2\alpha_3}, \\ \dot{\sigma}_{11} < 0 & \text{ if } \sigma > \beta_k = \beta_1\beta_3 + \sqrt{\beta_1^2\beta_3^2 + \beta_2\beta_3}. \end{aligned}$$

Hence every solution of the variance equation starting at $\sigma_1(t_0) \geq 0$ must eventually enter the region

$$0 < \alpha_k - \epsilon \leq \sigma \leq \beta_k + \epsilon < \infty, \tag{14.13}$$

provided $\epsilon > 0$ is suitably small. Hence it will suffice to restrict attention to solutions of the variance equation lying entirely in the region (14.13).

It is now easy to show that

(14.14) The optimal filter

$$d\hat{x}_1/dt = [f_{11}(t) - \sigma_{11}(t)/r_{11}(t)]\hat{x}_1$$

is uniformly asymptotically stable.

The solution $\sigma_{11}(t)$ of the variance equation will enter the region (14.13) at some time $t_1 \geq t_0$. It suffices to consider the behavior of the optimal filter after time t_1 . We introduce the Lyapunov function [14]

$$V(\hat{x}_1, t) = \frac{1}{\sigma_{11}(t)} \hat{x}_1^2, \quad (14.15)$$

and verify that

$$\frac{1}{\beta_4 + \epsilon} \hat{x}_1^2 \leq V(\hat{x}_1, t) \leq \frac{1}{\alpha_4 - \epsilon} \hat{x}_1^2 \quad \text{when } t \geq t_1;$$

in other words, V is uniformly bounded from above and below. The derivative \dot{V} of V along motions of the optimal filter is given by

$$\begin{aligned} \dot{V} &\equiv \frac{\partial V}{\partial t} + \frac{\partial V}{\partial \hat{x}_1} \cdot \frac{d\hat{x}_1}{dt}, \\ &= - \left[\frac{d\sigma_{11}/dt}{\sigma_{11}(t)} - 2 \frac{d\hat{x}_1}{dt} \frac{1}{\hat{x}_1} \right] V, \\ &= - \left[\frac{q_{11}(t)}{\sigma_{11}(t)} + \frac{\sigma_{11}(t)}{r_{11}(t)} \right] V, \\ &\leq - \left[\frac{\alpha_2}{\beta_4 + \epsilon} + \frac{\alpha_4 + \epsilon}{\beta_3} \right] V, \quad \text{when } t \geq t_1, \end{aligned} \quad (14.16)$$

which shows that \dot{V} is strictly negative when $t \geq t_1$ unless $\hat{x}_1^2 = 0$. This proves (14.14), in view of the well-known theorem of Lyapunov [14; Theorem 1].

We can now complete the qualitative analysis of the variance equation. Let $\sigma_{11}^{(a)}(t)$ and $\sigma_{11}^{(b)}(t)$ be two such solutions of the variance equation.

It is easily verified that if

$$\delta\sigma_{11}(t) = \sigma_{11}^{(a)}(t) + \sigma_{11}^{(b)}(t),$$

then

$$d\delta\sigma_{11}(t)/dt = [r_{11}^{(a)}(t) + r_{11}^{(b)}(t)]\delta\sigma_{11}(t), \quad (14.17)$$

where

$$r_{11}^{(a)}(t) = f_{11}(t) - \sigma_{11}^{(a)}(t)/r_{11}(t),$$

and $r_{11}^{(b)}(t)$ is defined similarly.

If $\psi_{11}^{(a)}(t, \tau)$ and $\psi_{11}^{(b)}(t, \tau)$ are the 1×1 transition matrices of the optimal filter corresponding to $\sigma_{11}^{(a)}(t)$ and $\sigma_{11}^{(b)}(t)$, then

$$\delta\sigma_{11}(t) = \psi_{11}^{(a)}(t, t_0)\psi_{11}^{(b)}(t, t_0)\delta\sigma_{11}(t_0), \quad (14.18)$$

as is immediately verified by differentiating and using (14.17). Hence the distance between any two solutions of the variance equation which start at $\sigma_{11}(t_0) \geq 0$ tends uniformly to zero with $t \rightarrow \infty$.

Equation (14.18) is the obvious analog of equation (12.10). As before, we can define

$$\lim_{t_0 \rightarrow -\infty} \sigma_{11}(t; 0, t_0) = \bar{\sigma}_{11}(t)$$

as the moving equilibrium state of the variance equation.

A particularly noteworthy feature of the Lyapunov function (14.15) is that it provides a quantitative measure of the factors influencing the stability of the optimal filter. This may be seen from the bracketed term in (14.16) which contains the two ratios

$$\sigma_{11}(t)/r_{11}(t) \quad \text{and} \quad q_{11}(t)/\sigma_{11}(t). \quad (14.19)$$

The first of these is just the message-to-noise ratio

$$\text{var}[\tilde{y}_1(t|t)]/\text{var}\{v_1(t)\}$$

of the error signal $\tilde{y}_1(t|t)$ of the optimal filter. The less noisy is the error signal, the more stable is the optimal filter. The second ratio in (14.19) is a measure of how effective the optimal filter is in counteracting the randomness introduced by the white-noise process v_1 . Both ratios can be related precisely to information-theoretical concepts, as is discussed elsewhere [26].

If the complexity of the problem is increased just a little more, the discussion of even the steady-state properties of the variance equation becomes quite involved. A good illustration of this state of affairs is the following

(14.20) **EXAMPLE[†]** The model of the random process is as shown in Fig. 8A. It consists of two constant first-order systems whose outputs are observed separately in the presence of independent white noise; the complications which arise are due to the fact that the random inputs w_1, w_2 may be correlated -- this introduces an "interaction" between two first-order problems.

The matrices corresponding to Fig. 8A are

$$\underline{F} = \begin{bmatrix} f_{11} & 0 \\ 0 & f_{22} \end{bmatrix}, \quad \underline{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \underline{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \underline{Q} = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}, \quad \underline{R} = \begin{bmatrix} r_{11} & 0 \\ 0 & r_{22} \end{bmatrix}.$$

Since the measurement noises are independent, we must set $r_{12} = 0$. The optimal filter is shown in Fig. 8B.

We assume of course that

$$r_{11} > 0, \quad r_{22} > 0. \quad (14.21)$$

[†] This problem was suggested by R. S. Bucy.

$$\begin{aligned}
 \delta_{11} &= 2r_{11}\sigma_{11} - \frac{\sigma_{11}^2}{r_{11}} - \frac{\sigma_{12}^2}{r_{22}} + q_{11}, \\
 \delta_{12} &= (r_{11} + r_{22})\sigma_{12} - \frac{\sigma_{11}\sigma_{12}}{r_{11}} - \frac{\sigma_{12}\sigma_{22}}{r_{22}} + q_{12}, \\
 \delta_{22} &+ 2r_{22}\sigma_{22} - \frac{\sigma_{12}^2}{r_{11}} - \frac{\sigma_{22}^2}{r_{22}} + q_{22}.
 \end{aligned}
 \tag{14.22}$$

We shall investigate only the problem: How many real nonnegative-definite equilibrium states has (14.22)?

To simplify the notation, we set

$$\begin{aligned}
 \rho_{11} &= \bar{\sigma}_{11}/r_{11}, \quad \rho_{12} = \bar{\sigma}_{12}/\sqrt{r_{11}r_{22}}, \quad \rho_{22} = \bar{\sigma}_{22}/r_{22}; \\
 \mu_{11} &= q_{11}/r_{11}, \quad \mu_{12} = q_{12}/\sqrt{r_{11}r_{22}}, \quad \mu_{22} = q_{22}/r_{22}; \\
 \alpha_1 &= \mu_{11} + r_{11}^2, \quad \alpha_2 = \mu_{22} + r_{22}^2.
 \end{aligned}
 \tag{14.23}$$

Letting the δ_{ij} be equal to zero and using these abbreviations, (14.22) reduces to

$$(\rho_{11} - r_{11})^2 + \rho_{12}^2 = \alpha_1, \tag{14.24}$$

$$(\rho_{11} + \rho_{22} - r_{11} - r_{22})\rho_{12} = \mu_{12}, \tag{14.25}$$

$$(\rho_{22} - r_{22})^2 + \rho_{12}^2 = \alpha_2. \tag{14.26}$$

In addition, ρ must be a nonnegative-definite matrix so that

$$\rho_{11} \geq 0, \tag{14.27}$$

$$\rho_{22} \geq 0, \tag{14.28}$$

$$\det[\rho] = \rho_{11}\rho_{22} - \rho_{12}^2 \geq 0. \tag{14.29}$$

It will be convenient to utilize an auxiliary relation which is obtained as follows. Add (14.24) multiplied by ρ_{22} to (14.26) multiplied by ρ_{11} , and then subtract (14.25) multiplied by $2\rho_{12}$. This gives

$$[\rho_{11} + \rho_{22} - 2(f_{11} + f_{22})](\det \rho) = \mu_{11}\rho_{22} - 2\mu_{12}\rho_{12} + \mu_{22}\rho_{11}, \quad (14.30)$$

which can also be obtained by setting $d(\det \rho)/dt$ equal to 0.

The discussion now proceeds by considering numerous special cases.

Case I. $\rho_{12} = 0$. By (14.25), this can happen only if $\mu_{12} = 0$. Then equations (14.24) and (14.26) are decoupled and the problem is reduced to two separate first-order problems which were discussed in detail in Example (14.1).

Case II. $\rho_{12} \neq 0$. Then by (14.27-29) $\rho_{11} > 0$ and $\rho_{22} > 0$.

Now we must consider several subcases:

Case II-A. $\rho_{12} \neq 0$, $\det \rho = 0$. By (14.30), this implies

$$\mu_{11}\rho_{22} - 2\mu_{12}\rho_{12} + \mu_{22}\rho_{11} = 0; \quad (14.31)$$

therefore

$$(\mu_{11}\rho_{22} + \mu_{22}\rho_{11})^2 = 4\mu_{12}^2\rho_{12}^2 = 4\mu_{12}^2\rho_{11}\rho_{22}$$

and

$$(\mu_{11}\rho_{22} - \mu_{22}\rho_{11})^2 = -4(\det \mu)\rho_{11}\rho_{22} \leq 0,$$

which is possible if and only if

$$\det \mu = 0, \quad (14.32)$$

and (using also the fact that $\rho_{12} \neq 0$),

$$\frac{\mu_{11}}{\rho_{11}} = \frac{\mu_{22}}{\rho_{22}} = \frac{\mu_{12}}{\rho_{12}}. \quad (14.33)$$

Substituting (14.31) into (14.24-26), we obtain

$$\left. \begin{aligned} \rho_{11} + \rho_{22} - 2r_{11} &= \mu_{11}/\rho_{11}, \\ \rho_{11} + \rho_{22} - r_{11} - r_{22} &= \mu_{12}/\rho_{12}, \\ \rho_{11} + \rho_{22} - 2r_{22} &= \mu_{22}/\rho_{22}, \end{aligned} \right\} (14.34)$$

this shows that $\det \rho = 0$ only if

$$r_{11} = r_{22}. \quad (14.35)$$

Thus all three equations (14.24-26) reduce to one:

$$\rho_{11} + \rho_{22} - 2r_{11} = \mu_{12}/\rho_{12}. \quad (14.36)$$

There are now again two subcases:

Case II-A-1. $\rho_{12} \neq 0$, $\det \rho = 0$, $\mu_{12} = 0$. This can happen only if $\mu_{11} = \mu_{22} = 0$. Then (14.24-29) is now equivalent to

$$\left. \begin{aligned} \rho_{11} + \rho_{22} - 2r_{11} &> 0, \\ \rho_{11} > 0, \quad \rho_{22} > 0, \\ \rho_{11}\rho_{22} &= r_{12}^2. \end{aligned} \right\}$$

Written out explicitly, these relations are equivalent

$$\left. \begin{aligned} \rho_{11} &= f_{11} + \sqrt{f_{11}^2 - \rho_{12}^2}, \\ \rho_{22} &= f_{11} - \sqrt{f_{11}^2 - \rho_{12}^2}, \end{aligned} \right\} \quad (14.37)$$

and

$$0 \leq |\rho_{12}| < f_{11}.$$

There are of course many matrices which satisfy (14.37). For instance, if $f_{11} = f_{22} = 1$, $q_{11} = q_{12} = q_{22} = 0$, and $f_{11} = f_{22} = 10$, the matrices

$$\begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}, \quad \begin{bmatrix} 18 & 6 \\ 6 & 2 \end{bmatrix}, \quad \begin{bmatrix} 18 & -6 \\ -6 & 2 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 10 & 10 \\ 10 & 10 \end{bmatrix}$$

all have zero determinant and all are equilibrium states of the variance equation (14.22).

Case 14-A-11. $\rho_{12} \neq 0$, $\det \rho = 0$, $\mu_{12} \neq 0$. Then also $\mu_{11} > 0$ and $\mu_{22} < 0$. We can now eliminate ρ_{12} and ρ_{22} from (14.36) with the aid of (14.37) and solve the resulting quadratic equation. Remembering that ρ_{11} must be positive, we get

Since $\rho_{11} > 0$ and $\rho_{22} > 0$, this implies

$$\rho_{11} + \rho_{22} - f_{11} - f_{22} > 0; \quad (14.39)$$

in view of $\rho_{12} \neq 0$, this and (14.25) implies

$$\mu_{12} \neq 0,$$

and

$$\text{sign } \rho_{12} = \text{sign } \mu_{12}. \quad (14.40)$$

Equations (14.24) and (14.26) may be solved by radicals:

$$\left. \begin{aligned} \rho_{11} &= f_{11} \pm \sqrt{\alpha_1 - \rho_{12}^2} \\ \rho_{22} &= f_{22} \pm \sqrt{\alpha_2 - \rho_{12}^2} \end{aligned} \right\} \quad (14.41)$$

Substituting (14.41) in (14.25), we obtain a quadratic equation in ρ_{12}^2 . This equation has four roots; utilizing (14.39), the number of roots reduces to two:

$$\rho_{12} = \mu_{12} / \sqrt{\epsilon_0}, \quad (14.42)$$

where $\epsilon_0 = \pm 1$, and

$$\sqrt{\epsilon_0} = \sqrt{\alpha_1 + \epsilon_2 + 2\alpha_0 \beta}, \quad \beta = \sqrt{\alpha_1 \alpha_2 - \mu_{12}^2}.$$

We have to verify of course that β and ϵ_0 defined by these formulas are real numbers. In fact

$$\beta^2 = \alpha_1 \alpha_2 - \mu_{12}^2 = (\mu_{11} + f_{11})(\mu_{22} + f_{22}) - \mu_{12}^2 \geq |f_{11} f_{22}|^2 \geq 0. \quad (14.43)$$

Similarly,

$$r^2(\epsilon_0) = \alpha_1 + \alpha_2 + 2\epsilon_0\beta > 0;$$

the equality sign is ruled out here due to the fact that $\mu_{12} \neq 0$ by (14.40).

Substituting (14.42) into (14.41) and letting $\epsilon_1 = \pm 1$, we get

$$\left. \begin{aligned} \rho_{11} - f_{11} &= \pm \frac{|\alpha_1 + \epsilon_0\beta|}{r(\epsilon_0)} = \epsilon_1 \frac{\alpha_1 + \epsilon_0\beta}{r(\epsilon_0)}, \\ \rho_{22} - f_{22} &= \pm \frac{|\alpha_2 + \epsilon_0\beta|}{r(\epsilon_0)} = \epsilon_2 \frac{\alpha_2 + \epsilon_0\beta}{r(\epsilon_0)}. \end{aligned} \right\} (14.44)$$

We now have to check whether ρ_{11} , ρ_{22} , and ρ_{12} defined by (14.44) and (14.42) actually satisfy (14.25). Since (14.40) fixes the sign of ρ_{12} , we have to consider 2^3 cases corresponding to various signs of ϵ_0 , ϵ_1 , and ϵ_2 . We can immediately rule out some of these cases by noting that (14.25) is equivalent to

$$(\epsilon_1 - 1)\alpha_1 + (\epsilon_2 - 1)\alpha_2 + \epsilon_0(\epsilon_1 + \epsilon_2 - 2)\beta = 0$$

and remembering that $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta \geq 0$.

If $\epsilon_0 = 1$, then the only possibility is $\epsilon_1 = \epsilon_2 = 1$. Moreover, it is easy to verify that ρ_{11} and ρ_{22} given by (14.44) are always positive in this case.

If $\epsilon_0 = -1$, then ϵ_1 and ϵ_2 may have the following values, without violating any obvious condition: (1, 1); (-1, 1), provided that $\alpha_1 = \beta$ and $\alpha_2 > \beta$; and (1, -1), provided that $\alpha_2 = \beta$ and $\alpha_1 > \beta$.

Now we turn to the condition $\det \rho > 0$. We want to prove that $\epsilon_0 = \epsilon_1 = \epsilon_2 = 1$ is the only case where this is true. In other words, we want to prove the inequalities

$$[r(1)f_{11} + \alpha_1 + \beta][r(1)f_{22} + \alpha_2 + \beta] > \mu_{12}^2, \quad \text{when } \epsilon_0 = 1;$$

$$[r(-1)f_{11} + \alpha_1 - \beta][r(-1)f_{22} + \alpha_2 - \beta] \leq \mu_{12}^2, \quad \text{when } \epsilon_0 = 1, \epsilon_1 = 1, \epsilon_2 = 1;$$

$$r(-1)f_{11}[r(-1)f_{22} - \alpha_2 + \beta] \leq \mu_{12}^2, \quad \text{when } \epsilon_0 = 1, \epsilon_1 = -1, \epsilon_2 = 1;$$

$$[r(-1)f_{11} - \alpha_1 + \beta]r(-1)f_{11} \leq \mu_{12}^2, \quad \text{when } \epsilon_0 = -1, \epsilon_1 = 1, \epsilon_2 = -1.$$

The terms in the square brackets must be always positive. All these inequalities are implied by

$$\epsilon_0(\alpha_1 + \epsilon_0[\beta - r(\epsilon_0)|r_{11}|])(\alpha_2 + \epsilon_0[\beta - r(\epsilon_0)|r_{22}|]) > \epsilon_0 \mu_{12}^2. \quad (14.45)$$

Expanding, we obtain the equivalent inequality:

$$r(\epsilon_0)[\beta + \epsilon_0|r_{11}r_{22}|] > |r_{11}|(\alpha_2 + \epsilon_0\beta) + |r_{22}|(\alpha_1 + \epsilon_0\beta).$$

In view of (14.42), the left-hand side is nonnegative. The preceding discussion shows that the right-hand side is also nonnegative. Squaring both sides we obtain after some calculation the following equivalent relation:

$$\delta = \gamma^2(\det \underline{\mu}) + \mu_{12}^2(|r_{11}| - |r_{22}|) > 0 \quad (14.46)$$

Since $\gamma > 0$ and $\mu_{12}^2 > 0$, the inequality will be satisfied if either

$$\det \underline{\mu} \neq 0 \quad \text{or} \quad |r_{11}| \neq |r_{22}|.$$

Hence we distinguish between two subcases:

Case II-B-1. $\mu_{12} \neq 0$, $\det \underline{\mu} > 0$, $|r_{11}| \neq |r_{22}|$ or $\det \underline{\mu} > 0$.

Then the only possibility is $\epsilon_0 = \epsilon_1 = \epsilon_2 = 1$.

Case II-B-11. $\rho_{12} \neq 0$, $\det \rho > 0$, $\det \mu = 0$, and $|f_{11}| = |f_{22}|$.

If $\epsilon_0 = -1$, then considering all three cases we see that $\delta = 0$ in (14.45) always implies $\det \rho \leq 0$. Hence this case cannot arise.

If $\epsilon_0 = 1$ and

$$f_{11} = f_{22} > 0, \quad (14.47)$$

and then $\det \rho > 0$, so that this case is possible.

We now collect all results, and state them in terms of the matrix $[\mu_{ij}]$.

If $\det \mu > 0$, then we must have Case I or Case II-B-1 because of (14.32).

If $\det \mu = 0$ but $\mu_{12} \neq 0$, then $\rho_{12} \neq 0$ by (14.25) and we must have Case II-B-1, Case II-A-11, or Case II-B-11. If $f_{11} = f_{22} > 0$, then both of the last two cases could arise. For example, take $f_{11} = f_{22} = 2$, $q_{11} = 1$, $q_{22} = 4$, and $q_{12} = 2$. Then $\alpha_1 = 5$, $\alpha_2 = 8$, $\beta = 6$, $r(1) = 5$, $r(-1) = 1$. Substituting into (14.38) respectively (14.44) and (14.42), we find the following two nonnegative definite equilibrium states of (14.22):

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \quad \text{when } \epsilon_0 = -1, \quad \text{and} \quad \frac{1}{5} \begin{bmatrix} 21 & 2 \\ 2 & 24 \end{bmatrix} \quad \text{when } \epsilon_0 = 1.$$

If $\det \mu = 0$ and $\mu_{12} = 0$, we have either Case I or Case II-A-1, because Case II-B is ruled out by (14.40).

Collecting our findings and recalling the results of Example (14.1), the following picture emerges concerning the equilibrium states of the variance equation (14.22).

(14.48) THEOREM. (A) If $\det \rho > 0$, (14.22) has a unique, positive definite equilibrium state \bar{x} :

$$\bar{\sigma}_{12} = \frac{\sqrt{r_{11}r_{22}}}{\sqrt{\frac{q_{11}}{r_{11}} + \frac{q_{22}}{r_{11}} + r_{11}^2 + r_{22}^2 + 2\sqrt{\left(\frac{q_{11}}{r_{11}} + r_{11}^2\right)\left(\frac{q_{22}}{r_{22}} + r_{22}^2\right)} - \frac{q_{12}^2}{r_{11}r_{22}}}}$$

$$\bar{\sigma}_{11} = r_{11} \left[f_{11} + \sqrt{\frac{q_{11}}{r_{11}} + r_{11}^2 - \frac{\sigma_{12}^2}{r_{11}r_{22}}} \right], \quad (14.49)$$

$$\bar{\sigma}_{22} = r_{22} \left[f_{22} + \sqrt{\frac{q_{22}}{r_{22}} + r_{22}^2 - \frac{\sigma_{12}^2}{r_{11}r_{22}}} \right].$$

The formulas for $\bar{\sigma}_{11}$ and $\bar{\sigma}_{12}$ reduce to (14.5) if $q_{12} = 0$.

(B) If $\det q = 0$ and $q_{12} \neq 0$, there are several possibilities:

(1) If $f_{11} \neq f_{22}$, there is a unique equilibrium state given by (14.49), which is nonsingular.

(2) If $f_{11} = f_{22} \neq 0$, there is a unique equilibrium state given by (14.38), which is singular.

(3) If $0 < f_{11} = f_{22}$, then there are precisely two equilibrium states, one singular and given by (14.38), one nonsingular and given by (14.49).

(C) If $q_{11} = q_{22} = q_{12} = 0$, then the following possibilities arise:

(1) If the condition $f_{11} = f_{22} > 0$ does not hold, then necessarily $\bar{\sigma}_{12} = 0$ and the problem reduces to two uncoupled first-order problems: there may be (i) one, (ii) two, or (iii) four equilibrium states depending on whether
 (i) $f_{11} \leq 0$ and $f_{22} \leq 0$, (ii) $f_{11} > 0$ or $f_{22} > 0$ but not both, (iii)
 $f_{11} > 0$, $f_{22} > 0$.

(2) If $0 < f_{11} = f_{22}$, then there are infinitely many equilibrium states, given by (14.37).

We have seen that a complete discussion of even the steady-state behavior of the variance equation is exceedingly tedious. What has been gained? First, the results provide a check on the theorems of Sect. 16. Second, the various special cases which arise serve as a warning that strong, general results can be obtained only under restrictive conditions; we see that the hypotheses of the theorems of Sect. 16 cannot be easily relaxed.

With the information now available, one could actually write down explicit formulas for the solutions of the variance equation. The steps are elementary but very tedious and little insight would be gained. A numerical illustration of the dynamical behavior of the variance equations is provided by the next example; a case where the solutions of the variance equations can be expressed in closed form occurs in Example (14.52).

The very complexity of the present-- relatively elementary -- example shows that a detailed, analytic discussion of the variance equation is out of the question for higher-order systems. We must therefore try to clarify the qualitative properties of the variance equation by abstract methods. See Sect. 16. Once the qualitative behavior is well understood, it is easy to obtain numerical answers by machine computation.

(14.50) EXAMPLE. Consider a dynamical system in which the acceleration is white noise. This situation occurs frequently in guidance problems (see also the next Example). We assume that both the position and the velocity of the system can be observed; these observations are contaminated with independent, additive, white noise. See Fig. 9A.

From the figure, the defining matrices are

$$F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad H = \begin{bmatrix} h_{11} & 0 \\ 0 & h_{22} \end{bmatrix}$$

The optimal filter is shown in Fig. 9B. The variance equations are:

$$d\sigma_{11}/dt = 2\sigma_{12} - h_{11}^2 \sigma_{11}^2 / r_{11} - h_{22}^2 \sigma_{12}^2 / r_{22}$$

$$d\sigma_{12}/dt = \sigma_{22} - h_{11}^2 \sigma_{11} \sigma_{12} / r_{11} - h_{22}^2 \sigma_{12} \sigma_{22} / r_{22}$$

$$d\sigma_{22}/dt = h_{11}^2 \sigma_{12}^2 / r_{11} - h_{11}^2 \sigma_{12}^2 / r_{11} - h_{22}^2 \sigma_{22}^2 / r_{22} + q_{11}$$

The optimal gains are:

$$k_{11}(t) = h_{11} \sigma_{11}(t) / r_{11},$$

$$k_{21}(t) = h_{11} \sigma_{12}(t) / r_{11},$$

$$k_{12}(t) = h_{22} \sigma_{12}(t) / r_{22},$$

$$k_{22}(t) = h_{22} \sigma_{12}(t) / r_{22},$$

$$k_{22}(t) = h_{22} \sigma_{22}(t) / r_{22}.$$

We have assumed of course that $r_{11} > 0$, $r_{22} > 0$.

If $h_{11} \neq 0$, then the variance equations will have at least one equilibrium state; and if furthermore $q_{11} > 0$, the equilibrium state will be unique. These facts follow from Theorem(16.18).

Introducing the abbreviations

$$\alpha = |h_{11}| \sqrt{q_{11}/r_{11}}, \quad \beta = |h_{22}| \sqrt{q_{11}/r_{22}},$$

it is easy to verify that the equilibrium state $\bar{\Sigma}$ given by

$$\frac{h_{11}^2}{r_{11}} \sigma_{11} = \alpha \frac{\sqrt{2\alpha + \beta^2}}{\alpha + \beta^2},$$

$$\frac{h_{11}^2}{r_{11}} \sigma_{12} = \frac{\alpha^2}{\alpha + \beta^2} = \frac{h_{22}^2}{r_{22}} \sigma_{12}, \quad (14.51)$$

$$\frac{h_{22}^2}{r_{22}} \sigma_{22} = \frac{2\sqrt{2\alpha + \beta^2}}{\alpha + \beta^2},$$

is positive definite. If $h_{11} = 0$, the variance equation has no equilibrium state, unless $q_{11} = 0$ also.

The solutions of the variance equation were computed numerically for two sets of values:

Case A

$$\begin{aligned} h_{11} &= 1, & h_{22} &= 0; \\ q_{11} &= 1, & r_{11} &= 16, & r_{22} &= 1; \\ \sigma_{11}(0) &= 1, & \sigma_{12}(0) &= \sigma_{22}(0) = 0. \end{aligned}$$

Case B

$$\begin{aligned} h_{11} &= 1, & h_{22} &= 2; \\ q_{11} &= 1, & r_{11} &= 16, & r_{22} &= 1; \\ \sigma_{11}(0) &= 1, & \sigma_{12}(0) &= \sigma_{22}(0) = 0. \end{aligned}$$

The solutions of the variance equation are shown in Figs. 10 and 11 and the corresponding optimal gains in Figs. 12 and 13. Step responses of the optimal filters appear in Figs. 14-15.

It is clear that the availability of a relatively accurate velocity signal greatly reduces the error and speeds up σ_{22} . On the other hand, σ_{11} is actually somewhat slowed down by the addition of the velocity signal. This phenomenon can be explained easily by looking in detail at the variance equations.

(14.52) EXAMPLE. We shall consider a data-smoothing problem encountered in determining the position and velocity of space vehicles.* This will provide a convenient illustration of the hamiltonian technique for obtaining solutions of the variance equation. Moreover, the example shows how to obtain the model of the random process directly from physical considerations.

The physical picture is as follows. The position of a satellite is measured by means of a radio signal. It is assumed that the measurement contains additive noise which may be taken to be approximately gaussian and white relative to the bandwidth of the satellite motion. A second measurement of the satellite motion is available from an accelerometer. This reading is also subject to noise; but here the noise is due to drift and other very slowly varying effects, and may be considered to be a constant random variable during the interval of interest. The motion of the satellite is linearized and assumed to be one-dimensional, and subject to a constant, gaussian random acceleration.

The problem is to design an optimal filter which provides the best running estimates of the position and velocity of the satellite based on the two types of measurement noise and the variance of the acceleration.

The preceding assumptions are formalized by setting up a model for the message process. Let z_1 denote the radio signal and a_1 the reading of the accelerometer. Both signals are supposed to be known exactly. The equation of motion (linearized, one-dimensional, with unit mass) is

$$\ddot{x}_1 = a(t) = \text{acceleration} = \text{constant} = a$$

where a is a gaussian random variable with zero mean. The accelerometer measures a plus a constant gaussian random variable with zero mean (the bias error of the accelerometer) b :

* This problem was suggested by a paper of E. L. Peterson [31]. See also the writer's discussion of this paper [32].

$$a_1 = a + b.$$

Let $Ea^2 = r_a$ and $Eb^2 = r_b$ and define

$$\rho = \frac{r_a r_b}{r_a + r_b}.$$

We introduce two new random variables which are orthogonal to each other (and thus, by gaussianness, independent). The first of these is exactly known and the second is to be estimated:

$$u_1 = \frac{r_b}{r_a + r_b} a_1, \quad x_3 = \frac{r_a}{r_a + r_b} a_1 - \frac{r_b}{r_a + r_b} b. \quad *$$

Then the equations of motion are:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = a = u_1 + x_3.$$

The model is now fully described and is shown in Fig. 20.A. By inspection of Fig. 20.A we have:

$$\underline{F} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \underline{g} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad \underline{H} = [1 \ 0 \ 0]. \quad (14.53)$$

The variance equations are

$$\begin{aligned} d\sigma_{11}/dt &= 2\sigma_{12} - \sigma_{11}^2/r_{11}, \\ d\sigma_{12}/dt &= \sigma_{13} + \sigma_{22} - \sigma_{11}\sigma_{12}/r_{11}, \\ d\sigma_{13}/dt &= \sigma_{23} - \sigma_{11}\sigma_{13}/r_{11}, \\ d\sigma_{22}/dt &= 2\sigma_{23} - \sigma_{12}^2/r_{11}, \\ d\sigma_{23}/dt &= \sigma_{33} - \sigma_{12}\sigma_{13}/r_{11}, \\ d\sigma_{33}/dt &= -\sigma_{13}^2/r_{11}. \end{aligned} \quad (14.54)$$

* In [32], u_1 and x_3 are not independent, and for this reason results stated there must be corrected.

The Hamiltonian equations (V_c) are

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{p}_1 \\ \dot{p}_2 \\ \dot{p}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1/r_{11} & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix}. \quad (14.55)$$

The transition matrix corresponding to these equations is easily found using (4.9). (The sixth power of the matrix on the right-hand side of (14.55) is zero so that (4.9) is a finite sum.) The result is:

$$\Theta(t,0) = \begin{bmatrix} 1 & 0 & 0 & t/r_{11} & t^2/2r_{11} & t^3/6r_{11} \\ -t & 1 & 0 & -t^2/2r_{11} & -t^3/6r_{11} & -t^4/24r_{11} \\ t^2/2 & -t & 1 & t^3/6r_{11} & t^4/24r_{11} & t^5/120r_{11} \\ 0 & 0 & 0 & 1 & t & t^2/2 \\ 0 & 0 & 0 & 0 & 1 & t \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (14.56)$$

We assume that the initial value of $\Sigma(0)$ is: $\sigma_{11}(0) = \sigma_{22}(0) = 0$, while $\sigma_{33}(0) = \rho$ is the effect due to the bias in the reading of the accelerometer. (Of course, all off-diagonal terms of $\Sigma(0)$ are zero.) Substituting (14.55) into (13.12), we find that the solution of the variance equation corresponding to these initial conditions is:

$$\Sigma(t) = \frac{r_{11}}{t^5/20 + r_{11}/\rho} \begin{bmatrix} t^4/4 & t^3/2 & t^2/2 \\ t^3/2 & t^2 & t \\ t^2/2 & t & 1 \end{bmatrix}. \quad (14.57)$$

It is easily verified by direct substitution that this is indeed a solution of the variance equation which satisfies the initial conditions stated above.

The optimal time-varying gains can now be obtained at once from the relation $K(t) = \Sigma(t)A'R^{-1}$; they are:

$$k_{11}(t) = t^4/4\alpha(t), \quad k_{21}(t) = t^3/2\alpha(t), \quad k_{31}(t) = t^2/2\alpha(t); \quad (14.58)$$

where

$$\alpha(t) = t^5/20 + r_{11}/\rho.$$

The detailed block diagram of the optimal filter is shown in Fig. 20.B. It should be noted that the signal u_1 enters the message process and the model of the message process inside the filter at exactly the same point. This follows from the fact that u_1 is a known constant, independent of the other random variables.

The differential equations of the optimal filter can be read off by inspection of the figure. They are:

$$\begin{bmatrix} \hat{dx}_1/dt \\ \hat{dx}_2/dt \\ \hat{dx}_3/dt \end{bmatrix} = \begin{bmatrix} -t^4/4\alpha & 1 & 0 \\ -t^3/2\alpha & 0 & 1 \\ -t^2/2\alpha & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} + \begin{bmatrix} t^4/4\alpha \\ t^3/2\alpha \\ t^2/2\alpha \end{bmatrix} z_1 + \begin{bmatrix} 0 \\ u_1 \\ 0 \end{bmatrix}. \quad (14.59)$$

This differential equation is difficult to solve. Considerable simplification is obtained by introducing a new set of state variables:

$$\begin{aligned} w_1 &= \hat{x}_3, & 2\hat{x}_1 &= t^2 w_1 + w_2 + t w_3, \\ w_2 &= 2\hat{x}_1 - t\hat{x}_2, & \hat{x}_2 &= t w_1 + w_3, \\ w_3 &= \hat{x}_2 - t\hat{x}_3, & \hat{x}_3 &= w_1. \end{aligned}$$

Then by (14.59)

$$\begin{bmatrix} dw_1/dt \\ dw_2/dt \\ dw_3/dt \end{bmatrix} = \begin{bmatrix} -t^4/4\alpha & -t^2/4\alpha & -t^3/4\alpha \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} + \begin{bmatrix} t^2/2\alpha \\ 0 \\ 0 \end{bmatrix} z_1 + \begin{bmatrix} 0 \\ -t \\ 1 \end{bmatrix} u_1$$

The transition matrix corresponding to this equation is

$$\underline{\Psi}^{(w)}(t, \tau) = \begin{bmatrix} \beta/\alpha & -\delta/2\alpha & -(\gamma - \delta\tau/2)/\alpha \\ 0 & 1 & t - \tau \\ 0 & 0 & 1 \end{bmatrix} \quad (14.60)$$

where

$$\beta(\tau) = \tau^5/20 + r_{11}/\rho, \quad \gamma(t, \tau) = (t^4 - \tau^4)/8, \quad \delta(t, \tau) = (t^3 - \tau^3)/6.$$

Thus the transition matrix corresponding to (14.59) is found to be

$$\underline{\Psi}^{(\hat{x})}(t, \tau) = \frac{1}{\alpha} \begin{bmatrix} \alpha - \delta t^2/2 & (t - \tau)\alpha - (\gamma - \delta\tau)t^2/2 & (\tau^2/2 - t\tau)\alpha + (\beta + \gamma - \delta\tau^2/2)t^2/2 \\ -\delta t & \alpha - (\gamma - \delta\tau)t & -\tau\alpha + (\beta + \gamma - \delta\tau^2/2)t \\ -\delta & -(\gamma - \delta\tau) & \beta + \gamma - \delta\tau^2/2 \end{bmatrix}.$$

The impulse response of the optimal filter relating \hat{x}_2 to x_1 is given by:

$$\begin{aligned} g_r(t, \tau) &= \psi_{21}^{(\hat{x})}(t, \tau)k_{11}(\tau) + \psi_{22}^{(\hat{x})}(t, \tau)k_{21}(\tau) + \psi_{23}^{(\hat{x})}(t, \tau)k_{31}(\tau) = t \cdot \psi_{11}^{(w)}(t, \tau) \cdot \frac{r_{11}}{2\alpha} \\ &= \frac{t\tau^2}{2(t^5/20 + r_{11}/\rho)}, \quad t \geq \tau. \end{aligned} \quad (14.61)$$

The impulse response relating \hat{x}_2 to a_1 is

$$\begin{aligned}
g_1(t, \tau) &= \frac{\rho}{K_b} \hat{v}_{22}^{(x)}(t, \tau) \\
&= \frac{\rho}{K_b} (-t[\hat{v}_{22}^{(w)}(t, \tau)\tau - \hat{v}_{13}^{(w)}(t, \tau)] + \hat{v}_{33}^{(w)}(t, \tau)) \\
&= \frac{\rho}{K_b} \left[1 - \frac{3t^3 - 4t^2\tau + t\tau^2}{24(t^3/20 + r_{11}/\rho)} \right], \quad t \geq \tau. \quad (14.62)
\end{aligned}$$

The expressions for $P_{22}(t)$ and $g_1(t, \tau)$ agree with those given by Peterson [31]. However, he obtains

$$g_1(t, \tau) = \frac{\rho}{K_b} \left[1 - \frac{t^3}{t^3 + 20r_{11}/\rho} \right], \quad t \geq \tau. \quad (14.62a)$$

Even though the two impulse responses g_1 are not the same, either answer is correct! This is due to the fact that we need to consider only constant signals a_1 . The convolution integral formed with either (14.62) or (14.62a) gives the same answer when applied to a constant signal.

The conventional treatment of this problem (see Peterson [31] who follows Shinbrot [8]) stops when the impulse responses relating the measurements (z_1 and a_1) to the desired estimates (say, \hat{x}_2) are obtained. This is not completely satisfactory because the impulse responses may be difficult to realize physically. For instance, it may be easier to build the filter in Fig. 20.B (where only the gains $k_{11}(t)$ are time-varying) than to implement the impulse responses (14.61-62).

Note also that if only $g_1(t, \tau)$ is desired, it can be realized by the 1-dimensional dynamical system

$$\begin{aligned}
dw_1/dt &= -(t^4/4\alpha)w_1 + (t^2/2\alpha)x_1 \\
\hat{x}_2 &= tw_1.
\end{aligned}$$

(14.63) EXAMPLE. Consider the third-order system defined by the matrices:

$$\underline{F} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & f_{33} \end{bmatrix}, \quad \underline{g} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \underline{h} = [1 \quad 0 \quad 0].$$

The model consists of two cascaded integrators, preceded by a single-lag element with transfer function $1/(s - f_{11})$.

$$\begin{aligned} d\sigma_{11}/dt &= 2\sigma_{12} - \sigma_{11}^2/r_{11}, \\ d\sigma_{12}/dt &= \sigma_{13} + \sigma_{22} - \sigma_{11}\sigma_{12}/r_{11}, \\ d\sigma_{13}/dt &= f_{33}\sigma_{13} + \sigma_{23} - \sigma_{11}\sigma_{13}/r_{11}, \\ d\sigma_{22}/dt &= 2\sigma_{23} + \sigma_{12}^2/r_{11}, \\ d\sigma_{23}/dt &= f_{33}\sigma_{23} + \sigma_{33} - \sigma_{12}\sigma_{13}/r_{11}, \\ d\sigma_{33}/dt &= 2f_{33}\sigma_{33} + \sigma_{13}^2/r_{11} + q_{11}. \end{aligned} \tag{14.64}$$

Assuming $r_{11} > 0$, it follows that the steady-state value of σ_{11} is given by the quartic equation

$$\frac{[\bar{\sigma}_{11}/r_{11}]^2 [\bar{\sigma}_{11}/r_{11} - 2f_{33}]^2}{\bar{\sigma}_{11}/r_{11} - f_{33}} = 3[q_{11}/r_{11}]^{1/2}. \tag{14.65}$$

The remaining σ_{ij} can be now easily determined using (14.64). If $f_{33} \leq 0$, then equation (14.65) has a unique solution; if $f_{33} > 0$, then there are two solutions, one of which is ruled out, however, by the nonnegativeness requirement on $\underline{\sigma}$.

A detailed analytical discussion of this example would be pointless. Instead, we must rely on the theorems to be stated in Sect. 16 for a qualitative understanding of the behavior of (14.64).

On the basis of the examples discussed in this section, it is conjectured that the behavior of the variance equation can be ascertained by elementary (though not simple) algebraic means whenever the model of the message process is at least of the second order. By Example (14.63), this is no longer possible when the model is of order three, unless the matrix of the hamiltonian equations is nilpotent as in Example (14.50). From a practical point of view, such algebraic methods are useless; one should resort to the digital computer for numerical answers and to more advanced analysis for understanding the qualitative behavior of the optimal systems.

15. Minimal-variance unbiased estimation. This section is preparation for the detailed theory of the variance equation; at the same time, we establish interesting connections between our methods and the classical theory of parameter estimation [23, Chapters 32-34].

We shall study only the case of continuous time. This will appreciably simplify the formulas. The case of discrete time differs only in trivial details.

Consider a special case of the model (\tilde{I}_0):

$$\underline{z}(t) = \underline{H}(t)\underline{Q}(t, T)\underline{x}(T) + \underline{v}(t), \quad (15.1)$$

where $\underline{z}(T)$ is some fixed but unknown state, $\underline{Q}(t, \tau)$ is the transition matrix corresponding to $\underline{F}(t)$, and $\underline{v}(t)$ is a gaussian white-noise process with $\text{cov}[\underline{v}(t)] = \underline{R}(t)$. We assume that $\underline{R}(t)$ is nonsingular.

Given that $\underline{z}(t)$ has been observed in the interval $t_0 \leq t \leq T$, we wish to obtain the "best" estimate of the "parameters" $\underline{x}(T)$. This problem was discussed already in Example (9.3). We may think of (15.1) as representing noisy observations on a free dynamical system. At time t_0 it is decided that the "best" estimate of the state is desired at time T .

Let us consider first the problem of estimating the scalar quantity

$$\pi = \underline{p}'\underline{x}(T) \quad (15.2)$$

where \underline{p} is an arbitrary but known vector. Let $\hat{\pi}$ denote the estimator of π .

It is clear that $\hat{\pi}$ will be in general a random variable, since it is to be a function of the observations $\underline{z}(t)$. To define the "best" estimator, it is not sufficient to minimize the variance of π , for the trivial estimator which is constant (nonrandom) has zero variance. One could require that

$$\mathbb{E}\{(\hat{\pi} - \pi)^2 | \underline{z}(T)\} \quad (15.3)$$

be a minimum. It is more customary, however, to require first that $\hat{\pi}$ be unbiased

$$E(\hat{\pi} | \underline{x}(T)) = \underline{p}' \underline{x}(T) = \pi, \quad (15.4)$$

and then minimize (15.3) subject to this constraint. Then $\hat{\pi}$ is a minimal-variance unbiased estimator.

Such an estimator does not necessarily exist for the model (15.1), as is immediately obvious upon setting $\underline{h}(t)$ identically equal to zero. Evidently, the existence of an unbiased estimator is a characteristic property of the system (15.1). This motivates the following important concept.

(15.5) DEFINITION. A system (15.1) (or (I_c)) is said to be completely observable if for every t_0 there exists a $T(t_0)$ such that for any parameter π given by (15.2) one can construct an unbiased estimator $\hat{\pi}$ which is a linear function of the observations $\underline{z}(t)$ in the interval $t_0 \leq t \leq T(t_0)$.

By a linear estimator $\hat{\pi}$ we mean

$$\hat{\pi} = \int_{t_0}^t \underline{g}'(t) \underline{z}(t) dt, \quad (15.6)$$

where $\underline{g}(t)$ is an arbitrary (at least piecewise continuous) vector function of time.

It may be that this property holds for some but not all vectors \underline{p} in (15.2). A vector \underline{p} which by (15.2) defines a parameter π having an unbiased estimator $\hat{\pi}$ is called an observable constant. It turns out that observability is the dual of the concept of controllability which is briefly mentioned in Sect. 15 and discussed in much further detail in [26, 55].

For the present purposes, two characterizations of complete observability will be sufficient:

(15.7) OBSERVABILITY TESTS. A system (15.1) ^{(or (I_c))} is completely observable (1) if and only if the matrix

$$\underline{M}(t_0, T) = \int_{t_0}^T \underline{\phi}'(t, T) \underline{H}'(t) \underline{R}^{-1}(t) \underline{H}(t) \underline{\phi}(t, T) dt$$

is positive definite for some $T > t_0$.

Proof [33]. (1) If $\underline{M}(t_0, T)$ is positive definite, then

$$\underline{z}^0(t) = \underline{R}^{-1}(t) \underline{H}(t) \underline{\phi}(t, T) \underline{M}^{-1}(t_0, T) \underline{p} \quad (t_0 \leq t \leq T) \quad (15.8)$$

will define an unbiased estimator (15.6).

(ii) Suppose that system (15.1) is completely observable but that $\underline{M}(t_0, T)$ is singular. Then there is a vector $\underline{p} \neq \underline{0}$ such that $\|\underline{p}\|_{\underline{M}(t_0, T)}^2 = 0$. Then

$$\underline{z}^1(t) = \underline{H}(t) \underline{\phi}(t, T) \underline{p}$$

is identically zero in the interval $[t_0, T]$ since it is a continuous function of time and since

$$\int_{t_0}^T \|\underline{z}^1(t)\|_{\underline{R}^{-1}(t)}^2 dt = \|\underline{p}\|_{\underline{M}(t_0, T)}^2 = 0.$$

Now let $\underline{z}^2(t)$ define an unbiased estimator of $\underline{p}' \underline{x}(T)$. Then

$$0 = \int_{t_0}^T \underline{z}^2'(t) \underline{z}^1(t) dt = \int_{t_0}^T \underline{z}^2'(t) \underline{H}(t) \underline{\phi}(t, T) \underline{p} dt = \|\underline{p}\|^2,$$

which contradicts the hypothesis that $\underline{p} \neq \underline{0}$. Q. E. D.

Several points should be noted here.

Even if the matrix \underline{H} is singular, it supplies valuable information. A modification of the preceding proof shows namely that \underline{p} is an observable state relative to the interval $[t_0, T]$ if and only if

$$[\underline{I} - \underline{M}(t_0, T)] \underline{M}^\dagger(t_0, T) \underline{y} = \underline{0}, \quad (15.9)$$

where \underline{M}^\dagger is any pseudo-inverse of \underline{M} (see Appendix A).

While (15.7) is of central theoretical importance, it is not convenient to apply in concrete cases, because \underline{M} is difficult to calculate. If system (15.1) has constant coefficients, the following purely algebraic criterion is equivalent to (15.7):

$$\text{rank}[\underline{H}', \underline{F}'\underline{H}', \dots, (\underline{F}')^{n-1}\underline{H}'] = n. \quad (15.10)$$

This is proved in [35], using (15.7).

Differentiating the integral defining \underline{M} with respect to T leads to the differential equation

$$d\underline{M}/dT = -\underline{F}'(T)\underline{M} - \underline{M}\underline{F}'(T) + \underline{H}'(T)\underline{R}^{-1}(T)\underline{H}(T). \quad (15.11)$$

If \underline{M} has an inverse, then this equation becomes

$$d\underline{M}^{-1}/dT = \underline{F}(T)\underline{M}^{-1} + \underline{M}^{-1}\underline{F}'(T) - \underline{M}^{-1}\underline{H}'(T)\underline{R}^{-1}(T)\underline{H}(T)\underline{M}^{-1}, \quad (15.12)$$

which is a special case of the variance equation (III₀). Thus \underline{M}^{-1} is analogous to $\underline{\Sigma}$. This is easily seen also by computing the covariance matrix of the unbiased estimator $\hat{\underline{x}}(T)$ defined by:

$$\hat{\underline{x}}(T) = \underline{M}^{-1}(t_0, T) \int_{t_0}^T \underline{\Phi}'(t, T) \underline{H}'(t) \underline{R}^{-1}(t) \underline{y}(t) dt \quad (15.13)$$

(see (15.8)). We find

$$E(\underline{x}(T)) = \underline{M}^{-1}(t_0, T),$$

$$E([\hat{\underline{x}}(T) - \underline{x}(T)][\hat{\underline{x}}(T) - \underline{x}(T)]' | \underline{x}(T)) = \underline{M}^{-1}(t_0, T). \quad (15.14)$$

Except in the constant case, however, \underline{M} will not be invertible in general and therefore we must usually deal with (15.11) rather than the variance equation.

The matrix $\underline{M}(t_0, T)$ is well known in classical statistics. If $\underline{v}(t)$ is gaussian, then \underline{M} is the Fisher information matrix [24]. The definition of the latter is as follows. Let $f_{\underline{z}}(\underline{z}|\underline{x}(T))$ be the conditional probability density functional of the observations $\underline{z}(t)$ in the interval $[t_0, T]$, given $\underline{x}(T)$. (In the case of continuous time, this is a probability density function of curves $\underline{z}(t)$, the rigorous definition of which is somewhat delicate.) The Fisher information matrix is defined as

$$\underline{M} = [m_{ij}] = E \left(\frac{\partial^2 f_{\underline{z}}(\underline{z}|\underline{x}(T))}{\partial x_i(T) \partial x_j(T)} \middle| \underline{x}(T) \right). \quad (15.15)$$

In the case of gaussian noise $\underline{v}(t)$ we have, purely formally,

$$f_{\underline{z}}(\underline{z}|\underline{x}(T)) = \text{const.} \exp \left[-\frac{1}{2} \int_{t_0}^T \|\underline{z}(t) - \underline{H}(t)\underline{\phi}(t, T)\underline{x}(T)\|_{\underline{R}^{-1}(t)}^2 dt \right], \quad (15.16)$$

and one can check easily that the two definitions of \underline{M} coincide. Notice that in the gaussian case the information matrix is independent of the parameter $\underline{x}(T)$.

The most important application of the Fisher information matrix is the famous Cramer-Rao or information inequality [23, Sect. 32.5; 10, Sect. 7]. If $\hat{\underline{x}}(T)$ is any unbiased estimator of $\underline{x}(T)$, then the information inequality is*

$$E \left([\hat{\underline{x}}(T) - \underline{x}(T)] [\hat{\underline{x}}(T) - \underline{x}(T)]' \middle| \underline{x}(T) \right) \geq \underline{M}^{-1}, \quad (15.17)$$

which is valid of course only if \underline{M} is positive definite.

We have just seen that for the estimator defined by (15.15) the equality sign is actually attained. Assuming $\underline{x}(T)$ is not constant, it can be shown [35, §38] that the equality sign in (15.17) can arise if and only if

* If \underline{A} , \underline{B} are two symmetric matrices, we write $\underline{A} > \underline{B}$ [$\underline{A} \geq \underline{B}$] to express the fact that $\underline{A} - \underline{B}$ is positive definite [nonnegative definite].

the probability density functional $f_{\underline{x}}(\underline{\zeta}|\underline{x}(T))$ can be factored as

$$f_{\underline{x}}(\underline{\zeta}|\underline{x}(T)) = g(\underline{\zeta}) e^{\hat{\underline{x}}^r(T) \underline{a}(\underline{x}(T))} e^{\underline{b}(\underline{x}(T))}, \quad (15.18)$$

where g , a , b are arbitrary functions.

By expanding the integrand in (15.16), we see immediately that this condition is true in the gaussian case; in fact, then

$$\underline{a}(\underline{x}(T)) = \underline{x}(T), \quad 2\underline{b}(\underline{x}(T)) = \|\underline{x}(T)\|_{\underline{M}(t_0, T)}^2$$

Whenever the probability density functional $f_{\underline{x}}$ can be factored as

$$f_{\underline{x}}(\underline{\zeta}|\underline{x}(T)) = g(\hat{\underline{x}}(\underline{\zeta}), \underline{x}(T)) h(\underline{\zeta}), \quad (15.19)$$

one says that $\hat{\underline{x}}$ is a sufficient statistic. As is obvious from (15.19), in this case $\hat{\underline{x}}$ contains all information which the data $\underline{\zeta}(t)$ (= the observed values of $\underline{x}(t)$) convey about the parameter $\underline{x}(T)$. This explains intuitively why the equality sign would hold in the information inequality.

Since the (strict-sense) minimal-variance unbiased estimator turns out in the gaussian case, to be linear, this estimator constitutes at the same time the solution of the (wide-sense) problem of finding the minimal-variance linear estimator. (See Sects. 2 and 10.) We now prove this fact by methods independent of the preceding discussion.

(15.20) GAUSS-MARKOV THEOREM. Assume the process (15.1) is completely observable. Let $\underline{v}(t)$ be a white-noise process (not necessarily gaussian) with a nonsingular covariance matrix $\underline{R}(t)$. Then the minimal-variance linear unbiased estimator of any parameter $\underline{w} = \underline{p}'\underline{x}(T)$ of the process (15.1) is $\underline{w} = \underline{p}'\hat{\underline{x}}(T)$, where $\hat{\underline{x}}(T)$ is defined by (15.15).

Proof. Let \underline{p} be an unbiased linear estimator of \underline{w} , defined by

$$\rho = \int_{t_0}^T \underline{x}'(t) \underline{z}(t) dt .$$

Then

$$\begin{aligned} \text{var } \rho &= \mathbb{E} \left[\int_{t_0}^T \underline{x}'(t) \underline{y}(t) dt \right]^2, \\ &= \int_{t_0}^T \|\underline{x}(t)\|_{\underline{R}(t)}^2 dt, \\ &= \int_{t_0}^T \|\underline{z}^0(t) - [\underline{x}(t) - \underline{z}^0(t)]\|_{\underline{R}(t)}^2 dt . \end{aligned}$$

Since ρ is unbiased,

$$\int_{t_0}^T \underline{x}'(t) \underline{R}(t) \underline{z}^0(t) dt = \int_{t_0}^T \underline{x}'(t) \underline{R}(t) \underline{z}(t, T) \underline{M}^{-1}(t_0, T) \underline{z} dt = \|\underline{z}\|_{\underline{M}^{-1}(t_0, T)}^2 .$$

Hence

$$\text{var } \rho = \int_{t_0}^T [\|\underline{z}^0(t)\|_{\underline{R}(t)}^2 + \|\underline{z}^0(t) - \underline{x}(t)\|_{\underline{R}(t)}^2] dt \geq \text{var } \hat{\underline{w}} .$$

Since $\underline{R}(t)$ is positive definite for all $t_0 \leq t \leq T$, the equality sign can occur only if $\underline{x}(t) = \underline{z}(t)$ everywhere in this interval, that is, only if $\rho = \hat{\underline{w}}$. Q. E. D.

It should be noted that the gauss-markov theorem actually does not require the assumption that \underline{y} be a white-noise process. In fact, if \underline{y} has the nonsingular covariance matrix

$$\text{cov}[\underline{y}(t), \underline{y}(\tau)] = \underline{R}(t, \tau);$$

then letting

$$\underline{M}(t_0, T) = \int_{t_0}^T \int_{t_0}^T \underline{Q}'(t', T) \underline{H}'(t') \underline{R}^{-1}(t', t) \underline{M}(t) \underline{x}(t, T) dt' dt,$$

the minimal variance linear unbiased estimator of $\underline{x}(T)$ is given by

$$\hat{\underline{x}}(T) = \underline{M}^{-1}(t_0, T) \int_{t_0}^T \left[\int_{t_0}^T \underline{Q}'(t', T) \underline{H}'(t') \underline{R}^{-1}(t', t) dt' \right] \underline{y}(t) dt.$$

We now consider the question of physically realizing the minimal variance unbiased estimator $\hat{\underline{x}}(T)$ by means of a dynamical system. In this case, the assumption that \underline{y} is a white-noise process is a very appreciable simplification.

(15.21) The minimal variance unbiased estimator $\hat{\underline{x}}(T)$ given by (15.13) is the terminal state of the dynamical system

$$\frac{d\underline{x}}{dt} = \underline{F}(t)\underline{x} + \underline{K}(t)[\underline{z}(t) - \underline{H}(t)\underline{x}], \quad (15.22)$$

where

$$\underline{K}(t) = \underline{M}^\dagger(t_0, t) \underline{H}'(t) \underline{R}^{-1}(t).$$

The matrix block diagram of this equation is identical with Fig. 5, which refers to the optimal filter (Π_c). The only difference lies in the definition of $\underline{K}(t)$, but this is only superficial since we have already noted that $\underline{M}^\dagger(t_0, t)$ is formally the same as $\underline{\Sigma}(t|t)$.

Proof. Let $\underline{Y}(t, \tau)$ be the transition matrix of the filter (15.22). First we show that

$$\underline{Y}(T, t) = \underline{M}^{-1}(t_0, T) \underline{Q}'(t, T) \underline{M}(t_0, t) \quad \text{for } t_0 < t \leq T. \quad (15.23)$$

This formula is clearly true if $t = T$ because then the right-hand side of (15.23) reduces to the unit matrix. $\underline{Y}(T, t)$, regarded as a function of t with T fixed, satisfies the differential equation

$$-\dot{\underline{Y}}(T, t) = \underline{Y}(T, t) [\underline{Q}(t) - \underline{H}'(t_0, t) \underline{H}^{-1}(t) \underline{H}(t)], \quad (15.24)$$

as is easily seen by differentiating

$$\underline{Y}(T, t) \underline{X}(t, T) = \underline{I}$$

and using (4.8). Differentiating the right-hand side of (15.23) with respect to t , using (15.11) and the pseudo-inverse lemma (A.4), we verify easily that (15.24) holds. Thus \underline{Y} defined by (15.23) is indeed the transition matrix of (15.22).

(We see immediately from (15.23) that

$$\underline{X}(T, t) = \underline{0} \quad (t_0 \leq t \leq t_1), \quad (15.25)$$

where t_1 is the largest value of time such that

$$\underline{M}(t_0, t_1) = \underline{0}.$$

Since a transition matrix is never singular, this consequence of (15.23) is of course absurd. In fact, $\underline{Y}(T, t)$ is given by (15.23) only for $t_1 < t \leq T$; further

$$\lim_{h \rightarrow 0} \underline{Y}(T, t_1 + h) = \underline{0}$$

and we simply define $\underline{Y}(T, t)$ to be zero for $t_0 \leq t \leq t_1$.)

In view of (15.20), we have to prove only two things. First, that

$$\hat{\underline{x}}(T) = \int_{t_0}^T \underline{Y}(T, t) \underline{X}(t) \underline{z}(t) dt \quad (15.26)$$

is an unbiased estimator of $\underline{x}(T)$; second, that the covariance matrix of \underline{x} is $\underline{M}^{-1}(t_0, T)$.

It is easy to see that $\underline{x}(T)$ given by (15.26) is unbiased if and only if

$$\int_{t_0}^T \underline{\Psi}(T, t) \underline{K}(t) \underline{H}(t) \underline{\phi}(t, T) dt$$

is the unit matrix. Using (15.11), the preceding integral becomes

$$\int_{t_0}^T [\dot{\underline{\Psi}}(T, t) \underline{\phi}(t, T) + \underline{\Psi}(T, t) \dot{\underline{\phi}}(t, T)] dt,$$

$$= \underline{I} - \underline{\Psi}(T, t_0) \underline{\phi}(t_0, T),$$

$$= \underline{I}$$

because of (15.25).

Further,

$$\text{cov}[\underline{x}(T)] = \int_{t_0}^T \underline{\Psi}(T, t) \underline{K}(t) \underline{R}(t) \underline{K}'(t) \underline{\Psi}'(T, t) dt.$$

By (15.22) and (15.23), the integral is

$$\underline{M}^{-1}(t_0, T), \int_{t_0}^T \underline{\phi}'(t, T) \underline{M}(t_0, t) \underline{M}'(t_0, t) \underline{H}'(t) \underline{R}^{-1}(t) \underline{H}(t)$$

The pseudo-inverse lemma (A.4) shows that the bracketed term is equal to $\underline{M}(t_0, T)$. Hence

$$\text{cov}[\underline{x}(T)] = \underline{M}^{-1}(t_0, T).$$

The proof of (15.21) is complete. (Remark added in proof: The preceding argument shows that actually $\hat{\underline{x}}(t)$ defined by (15.22), is an unbiased estimator of $\underline{x}(t)$ for all $t \geq t_1 > t_0$ where t_1 is the first value of t for which $\underline{M}(t_0, t_1)$ is nonsingular.)

(15.27) REMARK. The theorem just proved shows that the filtering and unbiased estimation problems are governed by essentially the same theory. This is a familiar state of affairs in the calculus of variations. One basic equation, the hamilton-jacobi partial differential equation, covers a wide variety of problems, the differences between the various types of problems being represented by the boundary conditions. The hamilton-jacobi partial differential equation is equivalent to our variance equation. The boundary condition in the filtering case is that $\underline{\Sigma}(t_0 | t_0)$ is some nonnegative definite matrix. In the unbiased estimation problem, the initial condition is $\underline{\Sigma}(t_0 | t_0) = \infty$, which is the same as $\underline{\Sigma}(t_0, z_0) \equiv \underline{\Sigma}^{-1}(t_0 | t_0) = 0$.

Clearly, the solutions of the filtering and estimation problems will in general be different. We can see with reference to Example (14.50) and Figs. 18 and 19 that the optimal filter is usually not unbiased. In other words, if the signal $x_2(t)$ has a mean component, this mean component will be reproduced with an error because the unit step in $x_2(t)$ does not result in a unit step in $\hat{x}_2(t)$.

(15.28) REMARK. Let us indicate briefly how the minimal variance unbiased estimator can be computed in real time. There are essentially four possibilities:

(A) In the most obvious case shown in Fig. 21A, one simply performs the multiplication indicated in the integrand of (15.15) and then integrates with respect to time. The multiplying signals can be generated by a linear dynamical system whose initial state is taken as $\underline{\phi}(t_0, T)\underline{M}^{-1}(t_0, T)\underline{R}$. This method is rather inconvenient if analog computing equipment is used, because of the difficulty of accurate multiplication.

(B) By inspection of (15.15), we note that $\underline{\phi}'(t, T)\underline{H}'(t)\underline{R}^{-1}(t)$ may be interpreted as the (generalized) impulse response of the differential equation

$$d\underline{x}/dt = -\underline{F}'(t)\underline{x} + \underline{H}'(t)\underline{R}^{-1}(t)\underline{u}(t)$$

(the free part of which is the adjoint of (4.1).) Hence the optimal estimator has the obvious physical realization shown in Fig. 21.B. It may be convenient to change variables in such a way that the matrix \underline{H} becomes the identity. Such a system can be easily built using standard analog components but it has a serious disadvantage. If $\underline{F}(t)$ defines an asymptotically stable differential equation, then the corresponding adjoint system defined by $-\underline{F}'(t)$ is usually asymptotically unstable. A special case of this method was noted by Mishkin [56].

(2) Noting the difficulty just mentioned, several authors (most prominently Huggins [57]) have suggested the following alternative. Suppose that the record of the observed function $\underline{z}(t)$ ($t_0 \leq t \leq T$) is inverted in time, that is to say, we introduce a new time variable t' defined by

$$t' - T = T - t,$$

and consider $\underline{z}(t') = \underline{z}(2T - t)$ instead of $\underline{z}(t)$.

Since $\underline{q}'(t, T) = [\underline{q}'(T, t)]^{-1}$ is the transition matrix of the adjoint differential equation

$$d\underline{x}/dt = -\underline{F}'(t)\underline{x}$$

of (4.1), it is clear (by changing variables) that $[\underline{q}'(2T - t', T)]^{-1}$ is the transition matrix $\underline{y}(t, T)$ of the dual system of (1_c)

$$d\underline{x}/dt' = \underline{F}'(t')\underline{x} + \underline{H}'(t')\underline{y}(t'), \quad (15.29)$$

where

$$\underline{F}'(t') = \underline{F}'(2T - t).$$

Hence

$$\underline{q}'(2T - t', T) = [\underline{y}(t', T)]' = \underline{y}(T, t).$$

Therefore after the change of variables $t \rightarrow t'$ we obtain the physical realization shown in Fig. 21B.

The required time-inversion may be performed, for instance, by recording $\underline{z}(t)$ on the tape recorder and then running it backwards.

(D) Finally, we have the realization provided by Theorem (15.21). It should be noted that in this case the minimal variance unbiased estimator is asymptotically stable (see next section) no time-inversion is required, and it is not necessary to change coordinates so as to make \underline{M} unity. On the other hand, one requires time-varying gains $\underline{K}(t)$. See Fig. 21D.

(15.30) **EXAMPLE.** The simplest estimation problem concerns the detection of the sine wave in white noise:

$$z_1(t) = x_1(t)\cos(t - T) + x_2(t)\sin(t - T) + v_1(t).$$

This problem is of importance in dynamic testing [38].

The corresponding matrices are

$$\underline{F} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \underline{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \text{and} \quad \underline{R} = \begin{bmatrix} r_{11} \end{bmatrix}.$$

The matrix \underline{M} given by (15.7) is found to be

$$\underline{M}(t_0, T) = \frac{1}{4} \begin{bmatrix} 2(T - t_0) + \sin 2(T - t_0) & -1 + \cos 2(T - t_0) \\ -1 + \cos 2(T - t_0) & 2(T - t_0) - \sin 2(T - t_0) \end{bmatrix}. \quad (15.31)$$

It is easily checked that this matrix is positive definite for all $T > t_0$.

Appreciable simplification results if we take advantage of the orthogonality properties of sine and cosine. Thus one is led to assume that

$$T - t_0 = q\pi \quad (q = \text{positive integer}) \quad (15.32)$$

in which case

$$\underline{M}(t_0, T) = \frac{1}{2}(T - t_0)\underline{I}.$$

Let us give also an explicit expression for the time-varying gains of the minimal variance unbiased estimator (15.13). Using (15.22) and (15.31) and assuming (15.32), we find that

$$\begin{aligned} r_{11}k_{11} &= \frac{2(t - t_0) + \sin 2(t - t_0)}{4[(t - t_0)^2 - \sin^2(t - t_0)]}, \\ r_{11}k_{21} &= \frac{\sin^2(t - t_0)}{2[(t - t_0)^2 - \sin^2(t - t_0)]}. \end{aligned} \tag{15.33}$$

At t_0 , the values of k_{11} and k_{21} are both infinity. They decrease monotonically to $1/2(T - t_0)$ and 0.

The block diagram of the filter is shown in Fig. 24.

16. Properties of the variance equation. The main purpose of this section is to generalize the results of Examples (12.7) and (14.11). We do this by trying to imitate the methods used to study these examples. The desired generalization can indeed be carried out, provided we assume that the system (I) has two important properties: it is completely observable and completely controllable. Most of the discussion is concerned with the case of continuous time; unless explicitly pointed out, the treatment of the case of discrete time is very similar. Of necessity, this section is rather technical and may be omitted at first reading.

In the interests of simplifying the notation, we assume that the original model (I_0) of the random process is one in which the cross-covariance matrix $\underline{C}(t)$ of \underline{w} and \underline{v} is zero. This does not entail any loss of generality. For let us replace $\underline{F}(t)$ by

$$\underline{F}(t) - \underline{Q}(t)\underline{C}(t)\underline{R}^{-1}(t)\underline{H}(t), \quad (16.1)$$

and $\underline{Q}(t)$ by

$$\underline{Q}(t) + \underline{C}(t)\underline{R}^{-1}(t)\underline{C}'(t). \quad (16.2)$$

Writing

$$\hat{\underline{w}}(t) = E(\underline{w}(t) | \underline{y}(t)) = \underline{C}(t)\underline{R}^{-1}(t)\underline{y}(t)$$

and

$$\tilde{\underline{w}}(t) = \underline{w}(t) - \hat{\underline{w}}(t),$$

we obtain the matrix block diagram shown in Fig. 23, in which the random excitations \underline{v} and \underline{w} are independent. The effect of dependence between \underline{v} and \underline{w} in the original model is now represented by the feedforward term $\underline{Q}(t)\underline{C}(t)\underline{R}^{-1}(t)$ in the block diagram of the optimal filter.

Hence from now on $\underline{Q}(t)$ is taken to be identically zero.

Suppose that $\underline{\Sigma}(t|t)$ is nonsingular. Then (III_c) may be written equivalently as

$$\begin{aligned} d\underline{\Sigma}^{-1}(t|t)/dt = & -\underline{F}'(t)\underline{\Sigma}^{-1}(t|t) - \underline{\Sigma}^{-1}(t|t)\underline{F}(t) + \underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t) \\ & - \underline{\Sigma}^{-1}(t|t)\underline{Q}(t)\underline{Q}'(t)\underline{\Sigma}^{-1}(t|t) \end{aligned} \quad (III_c^*)$$

This equation has precisely the same general form as (III_c) but the symbols designating the four terms on the right are somewhat different. We regard (III_c^{*}) the adjoint of (III_c). We can formalize this notion as follows:

(16.3) DEFINITION. The adjoint of the modal (I_c) is given by

$$\begin{aligned} d\underline{x}/dt = & -\underline{F}'(t)\underline{x} + \underline{H}'(t)\underline{w}(t), \\ \underline{z}(t) = & \underline{Q}'(t)\underline{x}(t) + \underline{v}(t), \end{aligned}$$

where

$$\begin{aligned} \text{cov}[\underline{v}(t), \underline{v}(\tau)] = & \underline{Q}^{-1}(t)\delta(t - \tau), \\ \text{cov}[\underline{w}(t), \underline{w}(\tau)] = & \underline{R}^{-1}(t)\delta(t - \tau). \end{aligned} \quad (II_c^*)$$

This definition is in agreement with the usual terminology in differential equations. (In [4 - 5] a somewhat different concept ("duality") was used, but in the present case the concept of the adjoint is more convenient.)

(16.4) REMARK. There is no loss of generality in assuming that the matrix $\underline{Q}(t)$ in (I_c) is invertible. In fact, we can even assume that the covariance matrix of \underline{w} is \underline{I} , as was discussed in Sect. 7. Thus the adjoint system always exists.

It is natural to introduce the

(16.5) DEFINITION. A system (I_c) is said to be completely controllable if its adjoint is completely observable.

In view of (15.7), complete controllability is equivalent to the following:

(16.6) CONTROLLABILITY LEMMA. The system (I_c) is completely controllable if and only if the matrix

$$W(t_0, T) = \int_{t_0}^T \Phi(T, t) Q(t) Q'(t) \Phi'(T, t) dt$$

is positive definite.

One can of course also define complete controllability directly [33]: The system (I_c) is completely controllable if there exists some forcing function $\underline{u}(t)$ which takes the system initially at rest ($\underline{x}(t_0) = \underline{0}$) to any arbitrary state \underline{x} in a finite length of time ($\underline{x}(T) = \underline{x}$). Furthermore, it follows [33] that the minimum amount of "control energy" necessary to accomplish this is given by

$$\int_{t_0}^T \|\underline{u}(t)\|_{Q^{-1}(t)}^2 dt = \|\underline{x}\|_{W^{-1}(t_0, T)}^2$$

The matrix \underline{W} , which is the adjoint of the information matrix \underline{M} , is thus seen to represent "reciprocal energy".

Just as in the case of Examples (12.7) and (14.11), it is desirable to impose conditions which guarantee that observability and controllability are essentially unaffected by the choice of t_0 .

(16.7) DEFINITION. A system (I_c) is uniformly completely observable if there exist fixed positive constants σ, α, β such that

$$0 < \alpha I \leq \underline{M}(t-\sigma, t) \leq \beta I^*$$

for all t . The system (I_c) is uniformly completely controllable if

* See footnote on page 198

$$0 < \alpha I \leq W(t - \sigma, t) \leq \beta I^*$$

for all t . (For simplicity, it is assumed that the same constants σ , α , β occur in both inequalities.)

(16.8) REMARK. The transformation introduced at the beginning of this section does not affect observability or controllability. Indeed, if $\underline{Y}(t, \tau)$ is the transition matrix corresponding to (16.1), then

$$\begin{aligned} \underline{x}(t) &= \underline{H}(t)\underline{Y}(t, T)\underline{x}(T) \\ &= \underline{H}(t)\left[\underline{\phi}(t, T)\underline{x}(T) + \int_t^T \underline{\phi}(t, \tau)\underline{G}(\tau)\underline{C}(\tau)\underline{R}^{-1}(\tau)\underline{H}(\tau)\underline{x}(\tau)d\tau\right] \end{aligned}$$

by (4.5). Hence

$$\underline{x}(t) + \int_t^T \underline{H}(t)\underline{\phi}(t, \tau)\underline{G}(\tau)\underline{C}(\tau)\underline{R}^{-1}(\tau)\underline{Y}(\tau)d\tau = \underline{H}(t)\underline{\phi}(t, T)\underline{x}(T),$$

which shows that there exists an unbiased estimator for (16.1) if and only if there exists one for \underline{F} , and it is clear that the variances of the unbiased estimators are the same in the two cases. Hence the matrix \underline{M} is the same in both cases. Passing to the adjoint system, we reach the same conclusion regarding the matrix \underline{N} .

Enroute to the main theorem (16.12), we establish the following facts, which are of interest in themselves.

(16.9) LEMMA. If (I_0) is uniformly completely observable and uniformly completely controllable and if $\underline{M}(t_0|t_0)$ is nonnegative definite, then $\underline{M}(t|t)$ is uniformly bounded for all $t \geq t_0 + \sigma$, and we have

$$\underline{M}(t|t) \leq \underline{M}^{-1}(t - \sigma, t) + \underline{M}(t - \sigma, t) \text{ when } t \geq t_0 + \sigma^*$$

To prove this, we make use of Theorem (15.21) which provides a filter for unbiased estimation. This filter is of course not optimal in the sense

* See footnote on page 198.

of Sect. 9, and therefore provides an upper bound for the variance of the optimal filter. Now the most important feature of unbiased estimation is the fact that errors due to the initial variance of \underline{x} are reduced to zero in a finite length of time. To put it differently, it suffices to operate on data over the time interval $[t-\sigma, t]$.

The covariance matrix of the filter (15.22) at any time $t \geq t_0 + \sigma$ is therefore given by

$$\underline{M}^{-1}(t-\sigma, t) + \text{cov} \left[\int_{t-\sigma}^t d\tau \underline{Y}(t, \tau) \underline{K}(\tau) \underline{H}(\tau) \int_{t-\sigma}^{\tau} dv \underline{Q}(\tau, v) \underline{Q}(v) \underline{V}(v) \right],$$

where $\underline{Y}(t, \tau)$ is the transition matrix of (15.22). Adding

$$\text{cov} \left[\int_{t-\sigma}^{\tau} d\tau \underline{Y}(t, \tau) \underline{K}(\tau) \underline{H}(\tau) \int_{\tau}^t dv \underline{Q}(\tau, v) \underline{Q}(v) \underline{V}(v) \right]$$

to the preceding expression and making use of (15.23) establishes the desired inequality.

(16.10) LEMMA. If (I_c) is uniformly completely controllable and uniformly completely observable and if $\underline{\Sigma}(t_0 | t_0)$ is positive definite, then

$$[\underline{K}^{-1}(t-\sigma, t) + \underline{M}(t-\sigma, t)]^{-1} \leq \underline{\Sigma}(t | t) \quad \text{when } t \geq t_0 + \sigma.$$

Let us compute first an expression for the rate of change of the determinant of the covariance matrix. Elementary but extensive manipulations yield

$$d[\det \underline{\Sigma}] / dt = [2 \text{tr}(\underline{K} - \underline{K}\underline{K}) + \text{tr}(\underline{\Sigma}^{\frac{1}{2}} \underline{H}' \underline{R}^{-1} \underline{H} \underline{\Sigma}^{\frac{1}{2}}) + \text{tr}(\underline{\Sigma}^{\frac{1}{2}} \underline{Q} \underline{Q}' \underline{\Sigma}^{\frac{1}{2}})] \det \underline{\Sigma}, \quad (16.11)$$

which is valid of course only if $\det \underline{\Sigma} > 0$. We have already encountered a special case of this formula in the form (14.30).

The second and third terms on the right-hand side of (16.11) are non-negative because they are the traces of nonnegative definite matrices. Hence

$$\det \underline{\Sigma}(t|t) \geq \left(\exp \int_{t_0}^t \text{tr}[\underline{F}(\tau) - \underline{K}(\tau)\underline{H}(\tau)]d\tau \right) \det \underline{\Sigma}(t_0|t_0).$$

Now \underline{F} , \underline{H} are assumed to be a continuous function of t (see Sect. 4), and so is $\underline{K}(t)$ by (3.15). Hence if $\det \underline{\Sigma}(t_0|t_0) > 0$ then $\det \underline{\Sigma}(t|t) > 0$ for all $t > t_0$.

From this and (15.3) it follows that if $\underline{\Sigma}(t_0|t_0)$ is positive definite, $\underline{\Sigma}^{-1}(t|t)$ exists for all $t > t_0$ and is the unique solution of (III*) having the initial value $\underline{\Sigma}^{-1}(t_0|t_0)$. Applying (16.9) to the adjoint case, the desired inequality follows at once.

(continued on next page)

(16.12) LEMMA. Suppose the system (I_c) is uniformly completely controllable and that $\underline{\Sigma}(t_0|t_0)$ is nonnegative definite. Then $\underline{\Sigma}(t|t)$ is nonnegative definite for all $t \geq t_1 = t_0 + \sigma$.

We have already seen in the course of the proof of the preceding lemma that if $\underline{\Sigma}(t|t)$ is once nonsingular, it will remain nonsingular thereafter. Hence it is sufficient to prove that $\underline{\Sigma}(t_1|t_1)$ is nonsingular.

Let us assume the contrary. Then there is a fixed nonzero vector \underline{p} such that

$$\|\underline{p}\|_{\underline{\Sigma}(t_1|t_1)}^2 = 0. \quad (16.13)$$

We shall show that actually

$$\|\underline{Y}(t_1, t)\underline{p}\|_{\underline{\Sigma}(t|t)}^2 = 0 \quad \text{for} \quad t_0 \leq t \leq t_1, \quad (16.14)$$

where \underline{Y} denotes the transition matrix of the optimal filter.

Let $\underline{g}(t) = \underline{Y}(t_1, t)\underline{\Sigma}(t|t)\underline{Y}'(t_1, t)$. By (II_c) and (III_c) it follows that $\underline{g}(t)$ satisfies the integral equation

$$\begin{aligned} \underline{g}(t) - \underline{g}(t_0) &= \int_{t_0}^t [\underline{g}(t)\underline{Y}(t, t_1)\underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)\underline{Y}'(t, t_1)\underline{g}(t) \\ &\quad + \underline{Y}(t_1, t)\underline{g}(t)\underline{g}(t)\underline{g}'(t)\underline{Y}'(t_1, t)] dt. \end{aligned} \quad (16.15)$$

The integrand is nonnegative definite. So is $\underline{g}(t_0)$. On the other hand, $\|\underline{p}\|_{\underline{g}(t_1)}^2 = 0$ by hypothesis. Hence $\|\underline{p}\|_{\underline{g}(t)}^2$ must vanish identically on $[t_0, t_1]$, which proves (16.14).

Instead of (16.13), we can write also

$$\underline{\Sigma}(t|t)\underline{Y}'(t_1, t)\underline{p} = \underline{0} \quad \text{when} \quad t_0 \leq t \leq t_1. \quad (16.16)$$

Let

$$q(t) = \underline{X}'(t_1, t)p.$$

Differentiating with respect to t and using (Π_c) and (16.14) we get

$$\begin{aligned} dq(t)/dt &= [d\underline{X}'(t_1, t)/dt]p, \\ &= [- \underline{F}'(t) + \underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)\underline{X}(t|t)]p, \\ &= - \underline{F}'(t)q(t). \end{aligned}$$

Thus $q(t)$ satisfies a differential equation which is the adjoint of (Π_c) ; this equation has the unique solution

$$q(t) = \underline{q}(t_1, t)p = \underline{X}'(t_1, t)p, \quad (16.17)$$

which satisfies the initial condition $q(t_1) = p$. Substituting (16.14) and (16.17) into (16.15), we get

$$\|\underline{p}\|_{\underline{q}(t)}^2 + \underline{q}(t_0) = \|\underline{p}\|_{\underline{H}(t_0, t_1)}^2.$$

By hypothesis (16.13), we have then

$$- \|\underline{p}\|_{\underline{q}(t_0)}^2 = \|\underline{p}\|_{\underline{H}(t_0, t_1)}^2.$$

Since $\underline{p} \neq 0$, the right-hand is positive by the assumed uniform complete controllability (see (16.6)). This contradiction proves that $\underline{X}(t_1|t_1)$ is positive definite.

We are now in a position to prove the chief result of the paper.

(16.18) MAIN THEOREM. Suppose the system (I_c) is uniformly completely observable and uniformly completely controllable. Then the optimal filter is uniformly asymptotically stable.

(16.19) REMARK. It should be mentioned right away that "optimality" by no means implies "stability". But in physical applications, the uniform asymptotic stability of the optimal filter is an indispensable requirement. If a system is not uniformly asymptotically stable, then a bounded input may result in an unbounded output [14]. Hence small bias errors can ruin the performance of the filter. Perturbations in the values of \underline{x} would be disastrous unless the filter is at least stable. The search for conditions under which "optimality" implies various forms of "stability" is the central problem of filtering theory. In the classical Wiener approach, this problem is completely ignored, but it turns out (see below) that the classical assumptions guarantee stability anyway.

(16.20) EXAMPLES. The conditions of the theorem are clearly satisfied in case of Examples (12.1), Case (iv); (12.7); (14.1), with $q_{11} > 0$; (14.11). In these examples we were able to show uniform asymptotic stability of the optimal filter by direct methods.

(16.21) COUNTEREXAMPLES. What happens if we do not have complete controllability? In Example (12.1), Case (ii), with $|g_{11}| = 1$, the optimal filter is stable but not asymptotically stable. On the other hand, in Case (iii) of the same example the optimal filter is asymptotically stable. Similar comments apply to Example (14.1). Another illustration of the theorem is provided by Case II-A-2 of Example (14.20). Every steady-state optimal filter corresponding to the matrices on page 178 is unstable; the eigenvalues of $\underline{F} - \underline{KH}$ are ± 10 in each case.

Proof of the main theorem. Let $\underline{Y}(t, \tau)$ be the transition matrix of the optimal filter (Π_c) . Then

$$\Delta(t, \tau) = \underline{Y}'(\tau, t)$$

is the transition matrix of the adjoint of (Π_c) :

$$d\underline{x}/dt = - [\underline{F}'(t) - \underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)\underline{\Sigma}(t|t)]\underline{x}. \quad (16.22)$$

We shall prove that, for all t , there are positive constants c_1, c_2 such that

$$\|\Delta(t, t + \sigma)\| \leq c_1 e^{-c_2 \sigma}. \quad (16.23)$$

This implies

$$\|\underline{Y}'(t + \sigma, t)\| = \|\underline{Y}(t + \sigma, t)\| \leq c_1 e^{-c_2 \sigma},$$

and thus the uniform asymptotic stability of (Π_c) .

To establish (16.23), we introduce the Lyapunov function [14]

$$V(\underline{x}, t) = \|\underline{x}\|_{\underline{\Sigma}(t|t)}^2.$$

By lemmas (16.9), (16.10), (16.12) we know that $V(\underline{x}, t)/\|\underline{x}\|^2$ is uniformly bounded from above and below, at least for $t \geq t_1 = t_0 + 2\sigma$:

$$(\alpha + \beta^{-1})\|\underline{x}\|^2 \leq V(\underline{x}, t) \leq (\alpha^{-1} + \beta)\|\underline{x}\|^2. \quad (16.24)$$

The derivative of V along motions of (16.22) is given by:

$$\begin{aligned} \dot{V} &= \frac{\partial V}{\partial t} + \underline{x}' \frac{\partial V}{\partial \underline{x}} \\ &= \|\underline{H}(t)\underline{\Sigma}(t|t)\underline{x}\|_{\underline{R}^{-1}(t)}^2 + \|\underline{Q}(t)\underline{x}\|_{\underline{Q}(t)}^2. \end{aligned} \quad (16.25)$$

Thus V is nondecreasing along any motion as $t \rightarrow \infty$. We shall show that

$$V(\underline{x}(t), t) \leq \gamma V(\underline{x}(t + \sigma), t + \sigma) \text{ when } 0 < \gamma < 1. \quad (16.26)$$

which will prove (16.25), in view of the well-known theorem of Lyapunov [14].

The problem is to find a lower bound for the integral of the right-hand side of (16.25).

Let

$$u(\tau) = R^{-1}(\tau)H(\tau)\Sigma(\tau)\Delta(\tau, t + \sigma)\underline{x}(t + \sigma).$$

Then, by (16.22) and (4.5)

$$\begin{aligned} \Delta(\tau, t + \sigma)\underline{x}(t + \sigma) &= \underline{\Phi}'(t + \sigma, \tau)\underline{x}(t + \sigma) \\ &\quad + \int_{\tau}^{t+\sigma} \underline{\Phi}'(v, \tau)H'(v)u(v)dv. \end{aligned}$$

Let

$$\int_t^{t+\sigma} \|u(v)\|_{R(v)}^2 dv = \epsilon^2 \|\underline{x}(t + \sigma)\|^2;$$

and

$$\int_t^{t+\sigma} \|\underline{\Phi}'(v)\Delta(v, t + \sigma)\underline{x}(t + \sigma)\|_{Q(v)}^2 dv = \eta^2 \|\underline{x}(t + \sigma)\|^2$$

where ϵ and η depend on $t + \sigma$ and $\underline{x}(t + \sigma)$. Hence, writing $\delta^2 = \epsilon^2 + \eta^2$,

$$V(\underline{x}(t + \sigma), t + \sigma) - V(\underline{x}(t), t) = \delta^2 \|\underline{x}(t + \sigma)\|^2.$$

Further

$$\eta^2 \|\underline{x}(t + \sigma)\|^2 \geq \int_t^{t+\sigma} \|\underline{\Phi}'(\tau)\underline{\Phi}'(t + \sigma, \tau)(\underline{x}(t + \sigma) + \int_{\tau}^{t+\sigma} \underline{\Phi}'(v, t+\sigma)H'(v)u(v)dv)\|_{R^{-1}(\tau)}^2 d\tau. \quad (16.27)$$

Now if $t \leq \tau \leq t + \sigma$

$$\begin{aligned} &\left\| \int_{\tau}^{t+\sigma} \underline{\Phi}'(v, t + \sigma)H'(v)u(v)dv \right\|^2 \\ &\leq \left[\int_t^{t+\sigma} \|\underline{\Phi}'(v, t + \sigma)H'(v)R^{-1}(v)\| \cdot \|u(v)\|_{R(v)} dv \right]^2 \end{aligned}$$

By Schwarz's inequality,

$$\leq \epsilon^2 [\text{tr } \underline{M}(t, t + \sigma)] \|\underline{x}(t + \sigma)\|^2.$$

Expanding the integrand in (16.27) and using again Schwarz's inequality, we see that

$$\eta^2 \|\underline{x}(t + \sigma)\|^2 = \|\underline{x}(t + \sigma)\|_{\underline{W}(t, t + \sigma)}^2$$

$$- 2\alpha (\text{tr } \underline{W}(t, t + \sigma))^{1/2} [\text{tr } \underline{M}(t, t + \sigma)]^{1/2} \|\underline{x}(t + \sigma)\|^2.$$

By uniform complete observability and controllability,

$$\delta^2 = \epsilon^2 + \eta^2 \geq \alpha - 2 \cdot \epsilon \eta^{3/2} \beta^{3/2} + \epsilon^2$$

Moreover, we have trivially also that

$$\delta^2 \geq \epsilon^2.$$

Combining the two preceding inequalities, we find that

$$\delta^2 \geq \frac{\alpha^2}{4\beta^3\eta^3} > 0.$$

Hence

$$V(\underline{x}(t + \sigma), t + \sigma) - V(\underline{x}(t), t) \leq (\alpha^2 / 4\beta^3\eta^3) V(\underline{x}(t + \sigma), t + \sigma).$$

This establishes (16.26), and completes the proof of the theorem.

Our main theorem has an immediate and important consequence:

(16.28) THEOREM. Suppose the system (I_0) is uniformly completely observable and uniformly completely controllable. If $\underline{X}^{(a)}(t|t)$ and $\underline{X}^{(b)}(t|t)$ are any two solutions of the variance equation (III_0) having a nonnegative definite value at $t = t_0$, then

$$\|\underline{X}^{(a)}(t + \sigma|t + \sigma) - \underline{X}^{(b)}(t + \sigma|t + \sigma)\| \leq c \frac{2\alpha}{\epsilon} e^{-\alpha \sigma} \|\underline{X}^{(a)}(t|t) - \underline{X}^{(b)}(t|t)\|.$$

That is, the effect of the initial state $\underline{\Sigma}(t_0)$ is gradually "forgotten" as $t \rightarrow \infty$. This is important in practical applications, because the value of $\underline{\Sigma}(t_0)$ may not be accurately known.

Numerical integration of the variance equation is facilitated because the effect of round-off errors will not be cumulative.

Proof. Let

$$\delta\underline{\Sigma}(t) = \underline{\Sigma}^{(a)}(t|t) - \underline{\Sigma}^{(b)}(t|t)$$

It is easily verified that $\delta\underline{\Sigma}(t)$ obeys the differential equation

$$\begin{aligned} d\delta\underline{\Sigma}(t)/dt = & [\underline{F}(t) - \underline{\Sigma}^{(a)}(t)\underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)]\delta\underline{\Sigma}(t) \\ & + \delta\underline{\Sigma}(t)[\underline{F}(t) - \underline{\Sigma}^{(b)}(t)\underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t)]' \end{aligned}$$

From this, it follows easily that

$$\delta\underline{\Sigma}(t) = \underline{\Psi}^{(a)}(t, t_0)\delta\underline{\Sigma}(t_0)\underline{\Psi}^{(b)'}(t, t_0) \quad (16.29)$$

which is the analog of (13.18); $\underline{\Psi}^{(a)}$ resp. $\underline{\Psi}^{(b)}$ is the transition matrix of the optimal filter corresponding to $\underline{\Sigma}^{(a)}$ and $\underline{\Sigma}^{(b)}$.

Taking norms in (16.29) and invoking (16.24) proves the theorem.

Consider now the solution of the variance equation corresponding to $\underline{\Sigma}(t_0) = 0$, which we denote by $\underline{\Sigma}(t; \underline{Q}, t_0)$. Under the hypotheses of (16.28), the limit

$$\lim_{t_0 \rightarrow \infty} \underline{\Sigma}(t; \underline{Q}, t_0) = \bar{\underline{\Sigma}}(t) \quad (16.30)$$

exists for all t . To prove this, it is only necessary to note that $\underline{\Sigma}(t; \underline{Q}, t_0)$ is nondecreasing with t_0 , i.e.,

$$\underline{\Sigma}(t; \underline{Q}, t_0) \geq \underline{\Sigma}(t; \underline{Q}, t_1) \quad (16.31)$$

whenever $t_1 \geq t_0$. Then (16.30) follows by standard convergence arguments since by (16.9) $\underline{\Sigma}$ is uniformly bounded from above.

To prove (16.31), let $\underline{Y}^{(0)}$ and $\underline{Y}^{(1)}$ be the transition matrix and $\underline{K}^{(0)}$ and $\underline{K}^{(1)}$ be the gain of the optimal filter corresponding to $\underline{\Sigma}(t; \underline{Q}, t_0)$ and $\underline{\Sigma}(t; \underline{Q}, t_1)$. Then

$$\underline{\Sigma}(t; \underline{Q}, t_1) = \text{cov} \left(\int_{t_0}^t \underline{Y}^{(1)}(t, \tau) [\underline{G}(\tau) \underline{Y}(\tau) - \underline{K}^{(1)}(\tau) \underline{Y}(\tau)] d\tau \right).$$

By optimality,

$$\leq \text{cov} \left(\int_{t_0}^t \underline{K}^{(0)}(t, \tau) [\underline{G}(\tau) \underline{Y}(\tau) - \underline{K}^{(0)}(\tau) \underline{Y}(\tau)] d\tau \right),$$

$$\leq \underline{\Sigma}(t; \underline{Q}, t_0),$$

which was to be proved.

Hence we have, as an immediate corollary of (16.28),

(16.32) **THEOREM.** Suppose that the system (I_c) is uniformly completely observable and uniformly completely controllable. Then every solution of the variance which has a nonnegative-definite value at $t = t_0$ converges uniformly to $\bar{\underline{\Sigma}}(t)$ defined by (16.30).

In view of this theorem, we call $\bar{\underline{\Sigma}}(t)$ the moving equilibrium state of the variance equation. In the case of constant system (I_c) , the solution of the variance equation depends only on the difference $t - t_0$. Hence $\bar{\underline{\Sigma}}(t) = \bar{\underline{\Sigma}} = \text{const.}$

(16.33) **THEOREM.** Suppose the random process $\underline{x}(t)$ is generated by a constant system (I_c) , i.e., $\underline{F}, \underline{G}, \underline{H}, \underline{Q}, \underline{R}$ are constants. Suppose further that (I_c) is completely observable and completely controllable. Then every solution of the variance equation which has a nonnegative definite initial value tends uniformly to a constant matrix $\bar{\underline{\Sigma}}$ in the limit $t = \infty$. This

matrix is the unique positive definite equilibrium state of the variance equation, i.e., it is the unique positive definite solution of the system of simultaneous quadratic algebraic equations,

$$d\bar{\Sigma}(t)/dt = 0$$

Proof. The first part of the theorem follows at once from (16.32). $\bar{\Sigma}$ is positive definite by (16.12). It is unique, because if (III₁) has more than one constant solution, (16.28) is contradicted.

(16.34) **EXAMPLES AND COUNTEREXAMPLES.** Consider Example (14.20). We always have complete observability. If $\det Q > 0$, or $\det Q = 0$ but $f_{11} \neq f_{22}$, then we also have complete controllability, and Theorem (14.48), which was proved by direct methods, shows that $\bar{\Sigma}$ is unique and positive definite. On the other hand, if $\det Q = 0$ and $f_{11} = f_{22}$, then Theorem (14.48) shows that $\bar{\Sigma}$, though possibly unique, will always be singular. Hence the condition of complete controllability cannot be dropped from (16.33).

In Example (14.50) we have complete observability and complete controllability; the equilibrium state $\bar{\Sigma}$ given by (14.51) is indeed positive definite. On the other hand, if complete observability is destroyed by setting $h_{11} = 0$, then there does not even exist an equilibrium state (unless $q_{11} = 0$ which means that the second-order problem is degenerated into a first-order one.) Hence the condition of complete observability cannot be dropped from (16.33).

In Example (14.52), we have complete observability but not complete controllability since $Q = 0$. Indeed, we see from (14.57) that all solutions of the variance equation approach $\bar{\Sigma} = 0$ as $t \rightarrow \infty$. Thus in the absence of complete controllability we cannot guarantee that $\bar{\Sigma} > 0$.

It is not clear a priori whether or not the assumptions of the classical Wiener problem imply complete observability and complete controllability. In fact, the answer is yes.

For if (I_c) is not completely observable, we can introduce special coordinates so that the defining equations assume the form [33]

$$\begin{aligned} d\underline{x}^{(1)}/dt &= \underline{F}^{(11)}\underline{x}^{(1)} + \underline{G}^{(1)}\underline{w}^{(1)}(t), \\ d\underline{x}^{(2)}/dt &= \underline{F}^{(21)}\underline{x}^{(1)} + \underline{F}^{(22)}\underline{x}^{(2)} + \underline{G}^{(2)}\underline{w}^{(2)}(t), \\ \underline{z}(t) &= \underline{H}^{(1)}\underline{x}^{(1)}(t) + \underline{v}(t). \end{aligned}$$

In other words, some of the state variables (namely the components of $\underline{x}^{(2)}$) do not affect $\underline{z}(t)$ -- hence they may be ignored in the statement of the filtering problem.

A similar decomposition holds if (I_c) is not completely controllable:

$$\begin{aligned} d\underline{x}^{(1)}/dt &= \underline{F}^{(11)}\underline{x}^{(1)} + \underline{F}^{(12)}\underline{x}^{(2)} + \underline{G}^{(1)}\underline{w}^{(1)}(t), \\ d\underline{x}^{(2)}/dt &= \underline{F}^{(22)}\underline{x}^{(2)}. \end{aligned}$$

In other words, no random excitation acts on the vector $\underline{x}^{(2)}$. The assumption of stationarity in the Wiener problem implies that we must set $\underline{x}^{(2)} = \underline{0}$.

The statement of the classical Wiener problem does not explicitly involve \underline{x} . Therefore in setting up the presentation (I_c) , there is no loss of generality in assuming that $\underline{x}^{(2)}$ is absent from the preceding equations. This proves

(16.35) THEOREM. The classical (stationary) Wiener problem corresponds to a model (I_c) which is completely observable and completely controllable.

A simple example is provided by the following special case of Example (14.20). Let $\det \underline{Q} = 0$ but $q_{12} \neq 0$, while $f_{11} = f_{22} < 0$. Making the change of coordinates

$$\underline{k} = \underline{T}\underline{x} \quad \text{and} \quad \underline{m} = \underline{T}\underline{v},$$

where

$$\underline{T} = \begin{bmatrix} \sqrt{1 + a_{22}/a_{11}} & 0 \\ + \sqrt{a_{22}/a_{11}} & 1 \end{bmatrix},$$

the equations of (I_c) assume the form

$$\dot{z}_1 = f_{11}z_1 + u_1,$$

$$\dot{z}_2 = f_{11}z_2.$$

Since $f_{11} < 0$, the second equation may be disregarded and the Wiener problem reduces to a first-order one.

It should be noted, however, that in numerous applications (see, e.g., Example (14.32)) the Wiener formulation is not sufficiently general, and in such cases questions of observability and controllability may present some nontrivial problems.

In the classical theory of the Wiener problem, the process of solution involves the spectral factorization of fourier transforms into two components which are analytic in the upper res. lower halves of the complex plane. Intuitively, this procedure is related to the fact that the eigenvalues of the matrix defining the hamiltonian system (V_c) (see Sect. 13) occur in pairs: if λ is an eigenvalue, so is $-\lambda$.

It now show that under the hypotheses of the main theorem, this result can be generalized to nonconstant systems as well.

(16.36) THEOREM. Suppose the system (I_c) is uniformly completely observable and uniformly completely controllable. Then there exists a nonsingular linear transformation

$$\begin{bmatrix} \underline{x} \\ \underline{z} \end{bmatrix} = \underline{T}(t) \begin{bmatrix} \underline{k} \\ \underline{r} \end{bmatrix},$$

such that the hamiltonian equations (V_c) assume the diagonal form

$$\begin{aligned} \frac{d\underline{x}}{dt} &= - \underline{F}'(t)\underline{x}, \\ \frac{d\underline{\pi}}{dt} &= \underline{F}(t)\underline{\pi}. \end{aligned} \quad (V'_c)$$

where $\underline{F}(t)$ is the infinitesimal transition matrix of the optimal filter corresponding to $\underline{\Sigma}(t)$.

Moreover, $\underline{T}(t)$ and its inverse are uniformly bounded for all t .

Proof. Let $\underline{X}(\tau, t)$ (t fixed, τ variable) be the transition matrix of the optimal filter, corresponding to

$$\underline{\bar{F}}(t) = \underline{F}(t) - \underline{X}(t)\underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t).$$

The motions of the optimal filter may be denoted by

$$\underline{x}(\tau) = \underline{X}(\tau, t)\underline{x}(t).$$

The scalar function

$$V(\underline{x}(\tau), \tau) = \|\underline{x}(\tau)\|_{\underline{\Sigma}^{-1}(\tau)}^2$$

tends to 0 with $\tau \rightarrow \infty$ in view of the main theorem and of the lemmas preceding it. Differentiating with respect to τ , we obtain an integral expression for V :

$$V(\underline{x}(t), t) - V(\underline{x}(T), T) = \int_t^T \left[\|\underline{H}(\tau)\underline{x}(\tau)\|_{\underline{R}^{-1}(\tau)}^2 + \|\underline{Q}'(\tau)\underline{\Sigma}^{-1}(\tau|\tau)\underline{x}(\tau)\|_{\underline{Q}(\tau)}^2 \right] d\tau.$$

This proves that

$$\underline{B}(t) = \lim_{T \rightarrow \infty} \int_t^T \|\underline{H}(\tau)\underline{x}(\tau)\|_{\underline{R}^{-1}(\tau)}^2 d\tau$$

exists. Differentiating with respect to t , we see that \underline{B} is a solution of the differential equation

$$d\underline{s}/dt = -\underline{P}'(t)\underline{s} - \underline{Q}\underline{s}(t) + \underline{H}'(t)\underline{R}^{-1}(t)\underline{H}(t). \quad (16.37)$$

Now we define the transformation \underline{T} by

$$\underline{T}(t) = \begin{bmatrix} \underline{I} & \underline{s}(t) \\ \underline{\Sigma}(t) & \underline{I} + \underline{\Sigma}(t)\underline{s}(t) \end{bmatrix}. \quad (16.38)$$

Utilizing (16.37-8), we see that the new variables $(\underline{k}, \underline{\pi})$ satisfy the canonical differential equations

$$\begin{aligned} \begin{bmatrix} d\underline{k}/dt \\ d\underline{\pi}/dt \end{bmatrix} &= \left(-\frac{d\underline{T}}{dt} + \underline{T}^{-1} \begin{bmatrix} -\underline{P}' & \underline{H}'\underline{Q}\underline{H} \\ \underline{Q}\underline{Q}' & \underline{P} \end{bmatrix} \right) \begin{bmatrix} \underline{k} \\ \underline{\pi} \end{bmatrix} \\ &= \begin{bmatrix} -\underline{P}' & 0 \\ 0 & \underline{P}' \end{bmatrix} \begin{bmatrix} \underline{k} \\ \underline{\pi} \end{bmatrix}. \end{aligned}$$

Finally, we note that

$$-\underline{s}(t) \leq \underline{\Sigma}^{-1}(t) \leq \underline{M}(t - \sigma, t) + \underline{M}^{-1}(t - \sigma, t),$$

so that $\|\underline{s}(t)\|$ is uniformly bounded. The proof is complete.

With the aid of (16.38) we can write the matrix $\underline{\Theta}$ occurring in (13.11) in the canonical form

$$\underline{\Theta}(t, t_0) = \begin{bmatrix} \underline{I} & \underline{s}(t) \\ \underline{\Sigma}(t) & \underline{I} + \underline{\Sigma}(t)\underline{s}(t) \end{bmatrix} \begin{bmatrix} \underline{X}'(t_0, t) & 0 \\ 0 & \underline{X}(t, t_0) \end{bmatrix} \begin{bmatrix} \underline{I} + \underline{s}(t)\underline{\Sigma}(t) & -\underline{s}(t) \\ -\underline{\Sigma}(t) & \underline{I} \end{bmatrix}. \quad (16.39)$$

If \bar{F} , \bar{Q} , \bar{H} , \bar{G} , \bar{R} are constant, then so is $\bar{\Sigma}$ and \bar{S} , and the determination of \bar{S} is reduced to the elementary problem of solving the set of linear equations in the coefficients of \bar{S} obtained by setting the left-hand side of \bar{S} equal to zero. Moreover, in this special case,

$$\bar{\Sigma}(t, t_0) = \exp[(t - t_0)\bar{F}],$$

which can be explicitly computed as a matrix whose elements consist of sums of exponential $e^{\lambda_1(t - t_0)}$, where λ_1 are the eigenvalues of \bar{F} . This shows that the solution of the classical Wiener problem under the Markovian assumption contains in it also the solution of the problem with finite observation interval ($t_0 \neq -\infty$). Thus we have:

(16.40) THEOREM. Under the hypotheses of (16.33) any solution of the variance equation can be expressed in closed form by the following purely algebraic procedure:

- (i) Find $\bar{\Sigma}$ by setting the left-hand side of the variance equation (III_c) equal to zero;
- (ii) find \bar{S} by setting the left-hand side of (16.37) equal to zero;
- (iii) determine the eigenvalues of \bar{F} (which will always have negative real parts);
- (iv) express $\exp[(t - t_0)\bar{F}]$ in terms of the eigenvalues of \bar{F} ;
- (v) compute $\bar{Q}(t, t_0)$ by (16.39);
- (vi) utilize (13.12).

(16.41) EXAMPLE. As an illustration of this theorem, we compute the expression for \bar{Q} in Example (14.1). The constants \bar{a}_{11} and \bar{r}_{11} are given by (14.3) and (14.7); we see that s_{11} obeys the equation

$$0 = 2\bar{r}_{11}s_{11} + 1/\bar{r}_{11} = -2s_{11}\sqrt{\bar{r}_{11}^2} + \bar{a}_{11}/\bar{r}_{11} + 1/\bar{r}_{11}$$

or

$$a_{11} = \frac{1}{2\sqrt{r_{11}^2 r_{11}^2 + q_{11} r_{11}}} > 0.$$

Substituting these values of $\bar{\sigma}_{11}$, a_{11} , \bar{r}_{11} into (16.39), we verify the previously given formula (14.10).

It may be added that Example (14.1) was considered previously by Shubrot [8, Example 2], by his special method of solving the Wiener-Kopf integral equation. With the new method, the solution of the integral equation is avoided and the algebraic nature of the problem (which is the explanation for the possibility of obtaining results in closed form!) is clearly evident.

(16.42) REMARK. All previous considerations can be easily carried over, *mutatis mutandis*, to the case of discrete time. There is only one point which requires caution. In writing down the discrete analog of (16.39), one might be puzzled by the appearance of certain additional terms. But these terms all cancel, by virtue of the pseudo-inverse lemma (A.4).

In matrix calculations there is a frequently recurring difficulty due to the fact that the inverse of a matrix does not always exist. To prove the existence of a given matrix is often cumbersome and difficult. Moreover, in many cases solutions of a set of linear equations exist even when the inverse of the matrix defining these equations does not.

To alleviate some (though not all) such difficulties, it has been found convenient to introduce the notion of the so-called pseudo-inverse of a matrix. Roughly speaking, a pseudo-inverse must possess two properties to be useful: (i) it must always exist; (ii) when used in place of the inverse (which may not exist), it should give the correct answer to such questions as solutions of equations. In general, the pseudo-inverse is not unique; this gives rise to certain complications.

The material which follows provides the main facts needed in this paper. For further details, consult Penrose [39-40] and Kalman [41].

A matrix \underline{A}^\dagger is called a pseudo-inverse of a rectangular (not necessarily square) matrix \underline{A} if it satisfies the following relation

$$(1) \quad \underline{A}\underline{A}^\dagger\underline{A} = \underline{A}$$

If \underline{A} has an inverse, it is equal to \underline{A}^\dagger . For then (1) implies $\underline{A}\underline{A}^\dagger = \underline{I}$ and $\underline{A}^\dagger\underline{A} = \underline{I}$; as is well known [42, p. 62], these two relations imply that $\underline{A}^{-1} = \underline{A}^\dagger$. From (1) we see also that $(\underline{A}')^\dagger = (\underline{A}^\dagger)'$.

It is easy to prove that a pseudo-inverse satisfying (1) exists for any rectangular matrix \underline{A} . We show this first for a nonnegative definite matrix \underline{P} . It is well known in numerical analysis [43] that every nonnegative definite matrix can be transformed to a diagonal form

$$\underline{T}'\underline{P}\underline{T} = \underline{E}, \quad (A.1)$$

where \underline{T} is nonsingular and the matrix \underline{E} is diagonal, having only zeroes or ones on the diagonal. Thus $\underline{E}^2 = \underline{E}$. Then

$$\underline{P}^\dagger = \underline{T}\underline{T}' \quad (A.2)$$

satisfies (i). We can now define a pseudo-inverse of an arbitrary matrix by

$$\left. \begin{aligned} \underline{A}^\dagger &= (\underline{A}'\underline{A})^\dagger \underline{A}', \\ \underline{A}^\dagger &= \underline{A}'(\underline{A}\underline{A}')^\dagger \end{aligned} \right\} \text{A.3}$$

The pseudo-inverses occurring on the right are defined by (A.2).

To show that (A.3a-b) actually satisfy (i), we need a simple lemma, which is the chief tool in applications of the pseudo-inverse as far as this paper is concerned:

(A.4) PSEUDO-INVERSE LEMMA. Let $\{\underline{A}_i\}$, $i = 1, \dots, N$, be arbitrary $m_i \times n$ matrices. Then

$$\underline{A}_i - \underline{A}_i \left[\sum_{j=1}^N \underline{A}_j' \underline{A}_j \right]^\dagger \left[\sum_{j=1}^N \underline{A}_j' \underline{A}_j \right] = \underline{0}$$

for all $i = 1, \dots, N$. Similarly, let $\underline{B}(t)$ be an arbitrary $m \times n$ matrix whose elements are continuous functions of t in the interval $[0, T]$. Then

$$\underline{B}(t) - \underline{B}(t) \left[\int_0^T \underline{B}'(\tau) \underline{B}(\tau) d\tau \right]^\dagger \left[\int_0^T \underline{B}'(\tau) \underline{B}(\tau) d\tau \right] = \underline{0}$$

for all $t \in [0, T]$.

Proof. Let \underline{C}_i respectively $\underline{D}(t)$ denote the matrices on the left-hand sides of the preceding equations. Using (1), we find that

$$\sum_{i=1}^N \underline{C}_i' \underline{C}_i = \underline{0}$$

and

$$\int_0^T \underline{D}'(t)\underline{D}(t)dt = Q.$$

Hence for any \underline{x} ,

$$\sum_{i=1}^N \|\underline{C}_i \underline{x}\|^2 = 0,$$

and

$$\int_0^T \|\underline{D}(t)\underline{x}\|^2 dt = 0.$$

Consequently

$$\|\underline{C}_i \underline{x}\| = \|\underline{D}(t)\underline{x}\| = 0$$

for all \underline{x} , which implies

$$\underline{C}_i = \underline{0} \quad (i = 1, \dots, N) \quad \text{and} \quad \underline{D}(t) = \underline{0} \quad (0 \leq t \leq T);$$

the lemma is proved.

Substituting (A.3a) into (1), using (A.4) with $N = 1$, proves that A^\dagger given by (A.3a) is a pseudo-inverse. Formula (A.3a) is proved similarly, taking transposes. Since (A.3a-b) are in general not the same, we see that the pseudo-inverse is not unique.

For computing the pseudo-inverse of a matrix which is not square, one would naturally choose that one of formulas (A.3a-b) in which the square matrix $\underline{A}'\underline{A}$ or $\underline{A}\underline{A}'$ is smaller. For instance, the pseudo-inverse of a vector (1-column matrix) is given by

$$\underline{x}^\dagger = (\underline{x}'\underline{x})^\dagger \underline{x}' = \underline{x}' / \|\underline{x}\|^2.$$

For many purposes, the lack of uniqueness of the pseudo-inverse is inconvenient. This difficulty can be overcome by adjoining to (i) the following further axioms:

$$(ii) \quad \underline{A}^\dagger \underline{A} \underline{A}^\dagger = \underline{A}^\dagger,$$

$$(iii) \quad (\underline{A} \underline{A}^\dagger)' = \underline{A} \underline{A}^\dagger,$$

$$(iv) \quad (\underline{A}^\dagger \underline{A})' = \underline{A}^\dagger \underline{A}.$$

These axioms were introduced by Penrose [39], who proved the following

(A.5) THEOREM (Penrose). For every rectangular matrix \underline{A} , there exists one and only one matrix \underline{A}^\dagger satisfying simultaneously (i-iv).

To avoid confusion, we shall call the (unique) matrix given by Penrose's Theorem the generalized inverse of \underline{A} and designate it by $\underline{A}^\#$.

Penrose (see [40], with slight modifications) has proved also the following important property of the generalized inverse:

(A.6) THEOREM (Penrose). Consider the equation $\underline{A}\underline{x} = \underline{b}$. Let $\underline{x}^0 = \underline{A}^\# \underline{b}$, and $\underline{x} \neq \underline{x}^0$. Then:

$$(a) \quad \text{either } \|\underline{A}\underline{x} - \underline{b}\| > \|\underline{A}\underline{x}^0 - \underline{b}\|;$$

$$(b) \quad \text{or } \|\underline{A}\underline{x} - \underline{b}\| = \|\underline{A}\underline{x}^0 - \underline{b}\| \quad \text{and} \quad \|\underline{x}\| > \|\underline{x}^0\|.$$

In other words, $\underline{x}^0 = \underline{A}^\# \underline{b}$ gives the solution of $\underline{A}\underline{x} = \underline{b}$ if one exists and the best approximation to a solution where none exists. Thus, with Penrose, we can call \underline{x}^0 the best approximate solution of $\underline{A}\underline{x} = \underline{b}$.

Let $|\underline{A}|$ denote the following norm of the matrix \underline{A} :

$$|\underline{A}|^2 = \text{trace } \underline{A}'\underline{A} = \sum_{i,j} a_{ij}^2. \quad (A.7)$$

(A.8) COROLLARY. Let $\underline{A}^\dagger \neq \underline{A}^\#$ be a pseudo-inverse of \underline{A} ; i.e., any matrix which satisfies (1). Then $|\underline{A}^\dagger| > |\underline{A}^\#|$.

In other words, the generalized inverse is "smaller" (in the sense of the norm (A.7)) than any other pseudo-inverse.

To prove (A.8), it suffices to note that the matrix equation

$$\underline{A}\underline{X}\underline{A} = \underline{A}$$

may be interpreted as a vector equation in the elements of the matrix \underline{X} . By hypothesis, \underline{A}^\dagger is a solution of this equation, so that case (b) of (A.6) is applicable. Q.E.D.

To illustrate this result, consider the matrix

$$\underline{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Two pseudo-inverses are given by

$$\underline{A}_1^\dagger = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \quad \text{and} \quad \underline{A}_2^\dagger = \begin{bmatrix} 1/6 & 1/6 & 0 \\ 1/6 & 1/6 & 0 \\ 0 & 0 & 1/3 \end{bmatrix},$$

and the generalized inverse is given by

$$\underline{A}^\# = \begin{bmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{bmatrix},$$

so that

$$|\underline{A}_1^\dagger| = 1/3 > |\underline{A}_2^\dagger| = 2/9 > |\underline{A}^\#| = 1/9.$$

The generalized inverse is evidently uniquely determined by (1) and the requirement that it is the smallest pseudo-inverse.

Finally, let us mention that the generalized inverse can be determined by a method similar to (A.1) through (A.3). This is done [40] by iterating twice the algorithm which determines \underline{T} satisfying $\underline{T}^2 \underline{T} = \underline{T}$.

Appendix B. Gaussian Random Vectors

Let \underline{x} be an n-dimensional random vector with mean $\underline{\mu} = E\{\underline{x}\}$ and covariance matrix $\underline{\Sigma} = E\{\underline{x}\underline{x}'\}$. It is customary to say [44, p. 17] that \underline{x} is gaussian if its probability density function is

$$p_{\underline{x}}(\underline{z}) = \frac{1}{(2\pi)^{n/2} (\det \underline{\Sigma})^{1/2}} \exp \left[-\frac{1}{2} \|\underline{z} - \underline{\mu}\|_{\underline{\Sigma}^{-1}}^2 \right]. \quad (\text{B.1})$$

This definition of a gaussian random vector does not apply, however, when $\underline{\Sigma}$ is singular. In this case the values \underline{z} taken on by \underline{x} are confined with probability 1 to a hyperplane of dimension less than n, and one cannot express this fact by a formula such as (B.1). Consequently if $\underline{\Sigma}$ is singular, the probability distribution of \underline{x} is defined by first introducing a linear transformation $\underline{x} = \underline{A}\underline{y} + \underline{\mu}$ (where \underline{y} is a m-vector, m being the rank of $\underline{\Sigma}$), such that the covariance matrix of \underline{y} is nonsingular [44, p. 26] and $E\{\underline{y}\} = \underline{0}$. Then the probability density function of \underline{y} can be expressed by a formula analogous to (B.1).

These awkward difficulties caused by the singularity of $\underline{\Sigma}$ can be avoided if one chooses as the basic definition of gaussianity the characteristic function:

$$\chi_{\underline{x}}(\underline{s}) = E\{\exp i\underline{s}'\underline{x}\} = \exp [i\underline{s}'\underline{\mu} - \frac{1}{2} \|\underline{s}\|_{\underline{\Sigma}}^2]. \quad (\text{B.2})$$

In this definition the inverse of Σ is not required.

Since the distribution of a gaussian random vector is uniquely determined by its mean and covariance matrix, it is desirable to calculate as much as possible directly with μ and Σ . For these purposes, (B.2) is better suited than (B.1).

Similarly, a pair of gaussian random vectors x_1, x_2 is defined by their joint characteristic function:

$$\begin{aligned} \chi_{x_1 x_2}(s_1, s_2) &= E(\exp i(s_1' x_1 + s_2' x_2)) = \\ &= \exp [i(s_1' \mu_1 + s_2' \mu_2) - \frac{1}{2}(\|s_1\|_{\Sigma_{11}}^2 + 2(s_1, s_2)_{\Sigma_{12}} + \|s_2\|_{\Sigma_{22}}^2)], \end{aligned} \quad (B.3)$$

where

$$\mu_1 = E(x_1), \quad \Sigma_{1j} = E(x_1 x_j'), \quad 1, j = 1, 2.$$

It follows from (B.3) that x_1 and x_2 are independent if and only if $\Sigma_{12} = 0$. Similarly, if x_1 and x_2 are gaussian, then $\alpha x_1 + \beta x_2 + c$ is also gaussian.

We now proceed to derive explicit expressions for the conditional mean and conditional covariances of a pair of gaussian random vectors. To do this elegantly, we make use of a recent observation of Balakrishnan [45] which relates these quantities to the joint characteristic function. In a slightly modified form, this result is:

The conditional expectation of x_1 given x_2 is a linear function of x_2

$$E(x_1 | x_2) = \mu_{12} + \Sigma_{12} x_2 \quad (B.4)$$

(1) if and (ii) only if

$$\frac{\partial}{\partial s_1} \chi_{x_1 x_2}(s_1, s_2) \Big|_{s_1=0} = [\mu_{12} + \Sigma \frac{\partial}{\partial s_2}] \chi_{x_2}(s_2), \quad (B.5)$$

where $\frac{\partial}{\partial \underline{x}}$ denotes the vector with components $\partial/\partial x_1$.

Proof: Since every moment of a gaussian distribution is finite, it follows [27, p. 67 and 89] that we may interchange differentiation with respect to \underline{s} with the expected-value operation. Thus

$$\left. \frac{\partial}{\partial \underline{s}_2} \chi_{\underline{x}_1 \underline{x}_2}(\underline{s}_1, \underline{s}_2) \right|_{\underline{s}_1 = \underline{0}} = E(\underline{x}_1 e^{\underline{s}_2' \underline{x}_2}).$$

Taking the expectation first with respect to the conditional probability distribution of \underline{x}_1 given \underline{x}_2 , and then with respect to \underline{x}_2 alone, we get

$$= E(e^{\underline{s}_2' \underline{x}_2} E(\underline{x}_1 | \underline{x}_2)).$$

Utilizing (B.4),

$$= E((\underline{\mu}_{12} + \underline{K} \underline{x}_2) e^{\underline{s}_2' \underline{x}_2}).$$

Interchanging E and $\partial/\partial \underline{s}_2$,

$$= (\underline{\mu}_{12} + \underline{K} \frac{\partial}{\partial \underline{s}_2}) \chi_{\underline{x}_2}(\underline{s}_2),$$

which proves part (ii).

To prove part (i), we proceed as before, interchanging E and $\partial/\partial \underline{s}$ which leads to the relation

$$E[(\underline{\mu}_{12} + \underline{K} \underline{x}_2 - E(\underline{x}_1 | \underline{x}_2)) e^{\underline{s}_2' \underline{x}_2}] = \underline{0},$$

valid for every \underline{s}_2 . This implies (B.4) by the uniqueness of the Fourier-Stieltjes transform. Q.E.D.

We now state the main result of this section:

(B.6) THEOREM. The conditional expectation of x_1 given x_2 is a gaussian random vector given by $\mu_1 + \Sigma_{12} \Sigma_{22}^{\dagger} (x_2 - \mu_2)$.

This is the formula found in textbooks [44, p. 28], except that here the inverse is replaced by a pseudo-inverse. Note that if $x_2 = \mu_2$ almost surely, i.e., $\Sigma_{22} = 0$, then (B.6) is correct since $0^{\dagger} = 0$.

To prove (B.6), we utilize (B.5). Straightforward differentiation leads to the condition

$$\mu_1 + \Sigma_{12} \Sigma_{22}^{\dagger} x_2 = \mu_1 + K (\mu_2 + \Sigma_{22} \Sigma_{22}^{\dagger} x_2).$$

Since this must hold identically in x_2 , we see that K must satisfy the equation

$$\Sigma_{12} = K \Sigma_{22}. \quad (B.7)$$

We now show that this equation always has a solution, which can be expressed as

$$K = \Sigma_{12} \Sigma_{22}^{\dagger}, \quad (B.8)$$

where Σ_{22}^{\dagger} denotes a pseudo-inverse of Σ_{22} as defined in the previous section. Indeed, if K satisfies (B.7-8), (B.6) follows at once.

Let Σ_{11}^{\dagger} be a nonsingular pseudo-inverse of Σ_{11} . (Such a Σ_{11}^{\dagger} always exists; it can be found, for instance, by means of relations (A.1-2).) The matrix

$$E[(\Sigma_{21} \Sigma_{11}^{\dagger} x_1 - x_2)(\Sigma_{21} \Sigma_{11}^{\dagger} x_1 - x_2)'] = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{\dagger} \Sigma_{12} = \Sigma_{11-2}$$

is clearly nonnegative definite; hence we can write

$$\Sigma_{22} = \Sigma_{12} \Sigma_{11}^{\dagger} \Sigma_{12} + (\Sigma_{11-2})^{\frac{1}{2}} (\Sigma_{11-2})^{\frac{1}{2}}. \quad (B.9)$$

We must show that

$$\Sigma_{12} = \Sigma_{12} \Sigma_{22}^{\dagger} \Sigma_{22} \quad (B.10)$$

Substituting Σ_{22}^{-} given by (B.9) into (B.10) and making use of the pseudo-inverse lemma (A.4), we find that

$$(\Sigma_{11}^{\dagger})^{1/2}(\Sigma_{12} - \Sigma_{12}\Sigma_{22}^{\dagger}\Sigma_{22}) = \underline{0}$$

Since Σ_{11}^{\dagger} was chosen to be nonsingular, this implies (B.10). Q.E.D.

It should be noted that the choice of a nonsingular pseudo-inverse Σ_{11}^{\dagger} was for computational convenience only; Σ_{22}^{\dagger} may be any pseudo-inverse.

From (B.6) we obtain immediately:

$$(B.11) \quad \underline{x}_1 - E(\underline{x}_1 | \underline{x}_2) \text{ is independent of } \underline{x}_2.$$

To prove this, it suffices to compute the cross-covariance matrix of $\underline{x}_1 - E(\underline{x}_1 | \underline{x}_2)$ and \underline{x}_2 . We have

$$\begin{aligned} E\{(\underline{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{\dagger}(\underline{x}_2 - \mu_2))(\underline{x}_2 - \mu_2)'\} \\ = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{\dagger}\Sigma_{22}, \end{aligned}$$

which is 0 by (B.10).

Similarly,

$$(B.12) \quad \text{The covariance matrix of } E(\underline{x}_1 | \underline{x}_2) \text{ is } \Sigma_{12}\Sigma_{22}^{\dagger}\Sigma_{12}.$$

$$(B.13) \quad \text{The conditional covariance matrix of } \underline{x}_1 \text{ given } \underline{x}_2 \text{ is independent of } \underline{x}_2 \text{ and is given by } \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{\dagger}\Sigma_{12}.$$

Finally, we point out a useful fact:

(B.14) If $\underline{x}_1, \underline{x}_2, \underline{x}_3$ are gaussian random vectors and the pair $\underline{x}_2, \underline{x}_3$ is independent, then

$$E(x_1 | x_2, x_3) = E(x_1 | x_2) + E(x_1 | x_3)$$

To prove this, let $x_4 = \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}$.

Then

$$\begin{aligned} E(x_1 | x_2, x_3) &= E(x_1 | x_4) \\ &= \mu_1 + \Sigma_{14} \Sigma_{44}^{-1} (x_4 - \mu_4) \end{aligned} \quad (B.15)$$

But

$$\Sigma_{14} = [\Sigma_{12} \quad \Sigma_{13}] \quad \text{and} \quad \Sigma_{44} = \begin{bmatrix} \Sigma_{22} & 0 \\ 0 & \Sigma_{33} \end{bmatrix}$$

by assumption. Substituting into (B.15) proves (B.14).

We have already remarked in Appendix A that the pseudo-inverse is necessarily equal to the inverse if the latter exists. But if the covariance matrix of the conditioning random variables is singular, the pseudo-inverse of this matrix (and therefore the conditional expectation) will not be unique. This is only a minor complication. For instance, let y_1 and y_2 be two (scalar) conditioning variables and assume that $y_1 = y_2$ (with probability 1). Then the conditional expectation of, say, x , given y_1 and y_2 may be written as

$$E(x | y_1, y_2) = ky_1.$$

But since $y_1 = y_2$, we can also write

$$E(x | y_1, y_2) = \frac{1}{2}(ky_1 + ky_2)$$

In both cases, the conditional expectation is the same random variable although expressed differently; the only difference is that in the first case the square of the coefficient matrix is k^2 , in the second case, it is $E(k/k)^2 = k^2/2$.

In numerical computations it is sometimes of interest to make the norm of the matrix \underline{K} in (B.8) as small as possible. In this case, one can take for the pseudo-inverse the generalized inverse of Penrose, which has the smallest norm (in a certain specific sense) among all pseudo-inverses. See (A.7).

References for Appendices.

- (1) J. L. Laming and R. H. Battin, RANDOM PROCESSES IN AUTOMATIC CONTROL (book), McGraw-Hill, 1956.
- (2) W. B. Davenport Jr. and W. L. Root, AN INTRODUCTION TO THE THEORY OF RANDOM SIGNALS AND NOISE (book), McGraw-Hill, 1958.
- (3) Y. W. Lee, STATISTICAL THEORY OF COMMUNICATION (book), Wiley, 1960.
- (4) R. E. Kalman, "A new approach to linear filtering and prediction problems", J. Basic Engr. (ASME Trans.), 82 D (1960) 37-45.
- (5) R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory", J. Basic Engr. (ASME Trans.), 83 D (1961) (to appear).
- (6) J. L. Doob, STOCHASTIC PROCESSES (book), Wiley, 1953.
- (7) M. Løve, PROBABILITY THEORY (book), 2nd Edition, Van Nostrand, 1960.
- (8) M. Shinbrot, "Optimization of time-varying linear systems with nonstationary inputs", Trans. ASME, 80 (1958) 457-462.
- (9) V. S. Pugachev, THEORY OF RANDOM FUNCTIONS AND ITS APPLICATION TO AUTOMATIC CONTROL PROBLEMS (book, in Russian). 2nd Edition, Gostekhizdat, Moscow, 1960.
- (10) E. Parzen, "Statistical inference on time series by Hilbert-space methods", Tech. Rep. 23, 1959, Appl. Math. and Stat. Lab., Stanford Univ.
- (11) E. Parzen, "A new approach to the synthesis of optimal smoothing and prediction systems", Tech. Rep. 34, 1960, Appl. Math. and Stat. Labs., Stanford Univ.
- (12) E. Parzen, "A survey of time-series analysis", Tech. Rep. No. 37, 1960, Appl. Math. and Stat. Labs., Stanford Univ.
- (13) H. Furstenberg, STATIONARY PROCESSES AND PREDICTION THEORY (book), Ann. of Math. Study No. 44, Princeton Univ. Press, 1960.

- (14) R. E. Kalman and J. E. Bartram, "The 'second method' of Lyapunov in the analysis and optimization of control systems. I. Continuous-time systems. II. Discrete-time systems", J. Basic Engr. (Trans. ASME), 82 D (1960) 371-393, 394-399.
- (15) R. A. Coddington and N. Levinson, THEORY OF ORDINARY DIFFERENTIAL EQUATIONS (book), McGraw-Hill, 1955.
- (16) I. M. Gel'fand, "Generalized stochastic processes (in Russian)", Dokl. Akad. Nauk USSR, 100 (1955) 853-856.
- (17) I. M. Gel'fand and G. E. Shilov, THEORY OF DISTRIBUTIONS (book, in Russian), Vol. 4, Chapter 3, Fizmatgiz, Moscow, 1960.
- (18) H. W. Bodv and C. E. Shannon, "A simplified derivation of linear least squares smoothing and prediction theory", Proc. IRE, 38 (1950) 417-425.
- (19) L. A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of prediction", J. Appl. Phys., 21 (1950) 645-655.
- (20) K. Schwerdtfeger, LES FONCTIONS DES MATRICES. I. LES FONCTIONS UNIVALENTES (book), Act. Sci. Ind. No. 649 (Hermann, Paris), 1958.
- (21) J. L. Doob, "The brownian movement and stochastic equations", Ann. Math., 43 (1942) 351-369.
- (22) M. C. Wang and G. E. Uhlenbeck, "On the theory of brownian motion II", Rev. Mod. Phys., 17 (1945) 323-342.
- (23) H. Cramér, MATHEMATICAL METHODS OF STATISTICS (book), Princeton Univ. Press, 1956.
- (24) S. Sherman, "A theorem on convex sets with applications", Ann. Math. Stat., 26 (1955) 763-767.
- (25) I. M. Gel'fand and A. M. Yaglom, "Calculation of the amount of information about a random function contained in another such function", Izvest. Akad. Nauk, VI (75) (1957) 3-52; English translation in Amer. Math. Soc. Translations, series 2, Vol. 12, 1959.
- (26) R. E. Kalman, "On the duality between energy and information", to appear.
- (27) A. L. Brundno and A. I. Lants, "On the filtering of random time series", Doklady Akad. Nauk USSR, 131 (1960) 485-488.

- (28) J. Radon, "Zum Problem von Lagrange", Abh. Math. Sem. Univ. Hamburg, 6 (1928) 273-299.
- (29) J. J. Levin, "On the matrix Riccati equation" / Proc. Am. Math. Soc., 10 (1959) 519-524.
- (30) N. Wiener, THE EXTRAPOLATION, INTERPOLATION AND SMOOTHING OF STATIONARY TIME SERIES (book), Wiley, 1949.
- (31) R. L. Peterson, "Optimization of multi-input time-varying systems subject to multiple or redundant nonstationary inputs", Proc. First Intern'l Congress on Automatic Control, (Moscow, 1960), Butterworth's, 1961.
- (32) R. E. Kalman, discussion of the preceding paper, *ibid.*
- (33) R. E. Kalman, Y. C. Ho, K. S. Narendra, "Controllability of linear dynamical systems", Contributions to Differential Equations, Vol. 1, 1961, Macmillan.
- (34) S. Kullback, INFORMATION THEORY AND STATISTICS (book), Wiley, 1959.
- (35) B. L. van der Waerden, MATHEMATISCHE STATISTIK (book), Springer, 1957.
- (36) E. Mishkin, "On the computer realization of a given signal function", Proc. IRE, 47 (1959) 1003-1004.
- (37) W. H. Ruggins, "Signal detection in a noisy world", AFCHC-TN-60-360 or RAND Corp. Res. Memo. 2462 (1960).
- (38) P. E. A. Covley, "The application of analog computers to the measurement of process dynamics", Trans. ASME, 79 (1957) 823.
- (39) R. Penrose, "A generalized inverse for matrices", Proc. Cambridge Phil. Soc., 51 (1955) 406-413.
- (40) R. Penrose, "On best approximate solutions of linear matrix equations", Proc. Cambridge Phil. Soc., 52 (1956) 17-19.
- (41) R. E. Kalman, "On the pseudo-inverses of a matrix", to be published.
- (42) P. R. Halmos, FINITE-DIMENSIONAL VECTOR SPACES (book), second edition, Van Nostrand, 1958.

- (43) H. Bodewig, MATRIX CALCULUS (book, second edition), Interscience, 1959.
- (44) T. W. Anderson, AN INTRODUCTION TO MULTIVARIATE STATISTICAL ANALYSIS, (book), Wiley, 1959.
- (45) A. V. Balakrishnan, "On a characterization of processes for which optimal mean-square systems are of specified form", Trans. IRE Prof. Group on Information Theory, IT-6 (1960) 490-500.

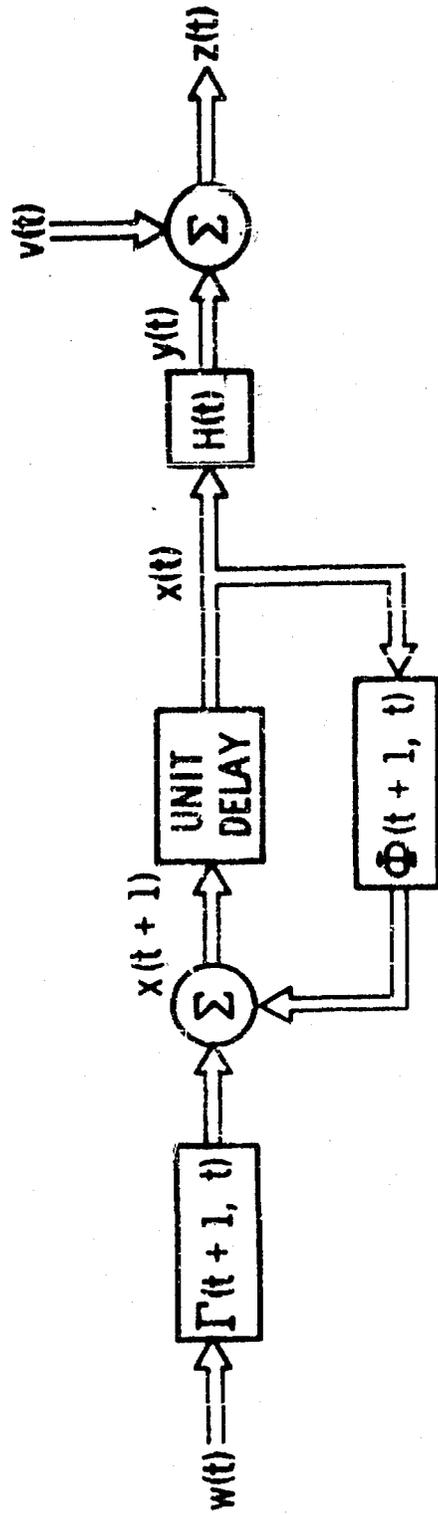


FIG. 1 REPRESENTATION OF GAUSS-MARKOV SEQUENCE

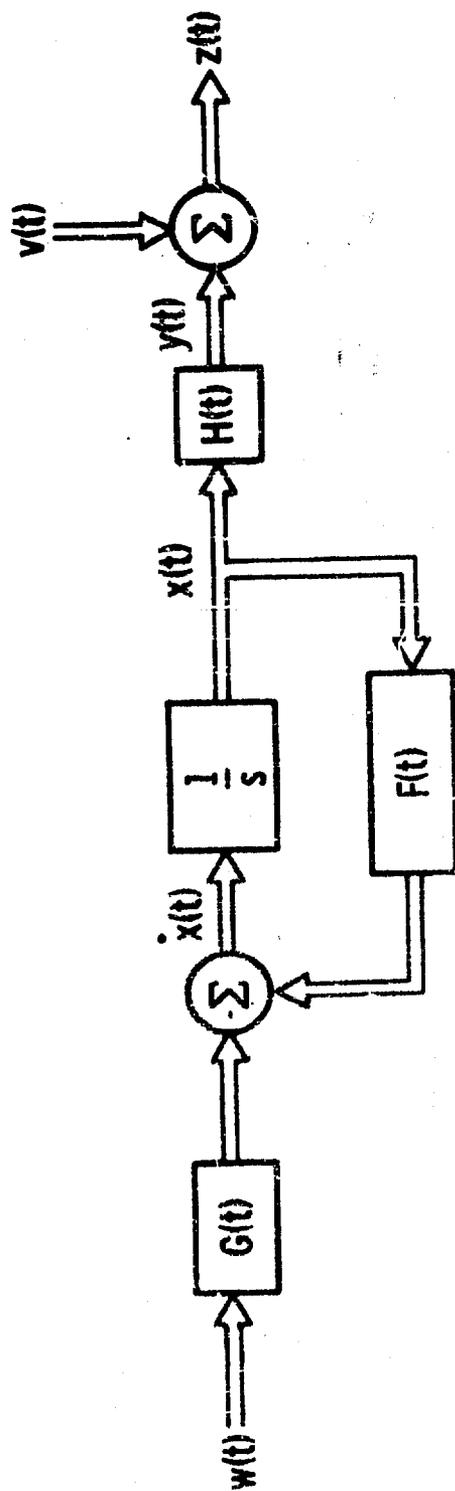


FIG. 2 REPRESENTATION OF GAUSS-MARKOV PROCESS

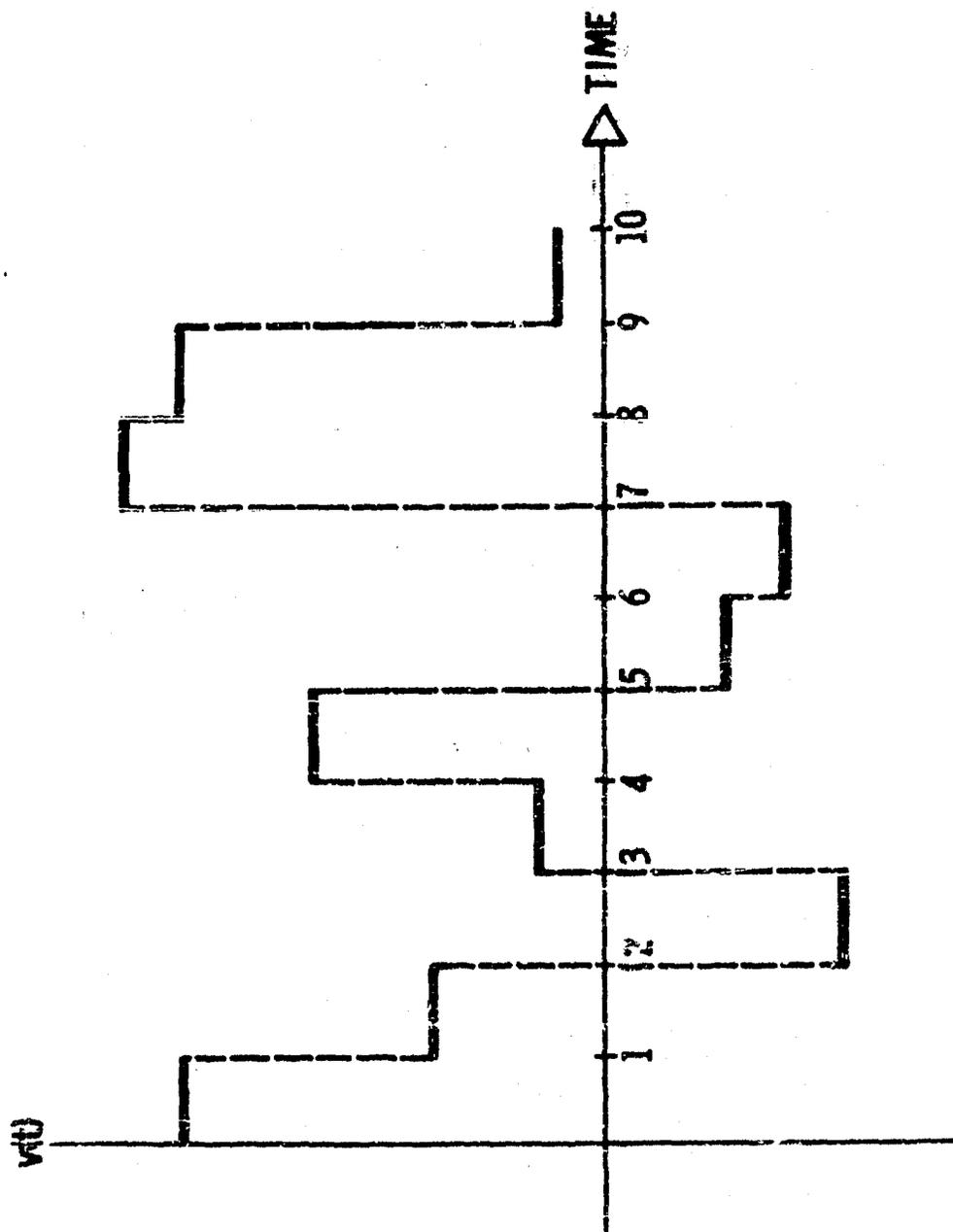


FIG. 3 SAMPLE FUNCTION OF (6.2)

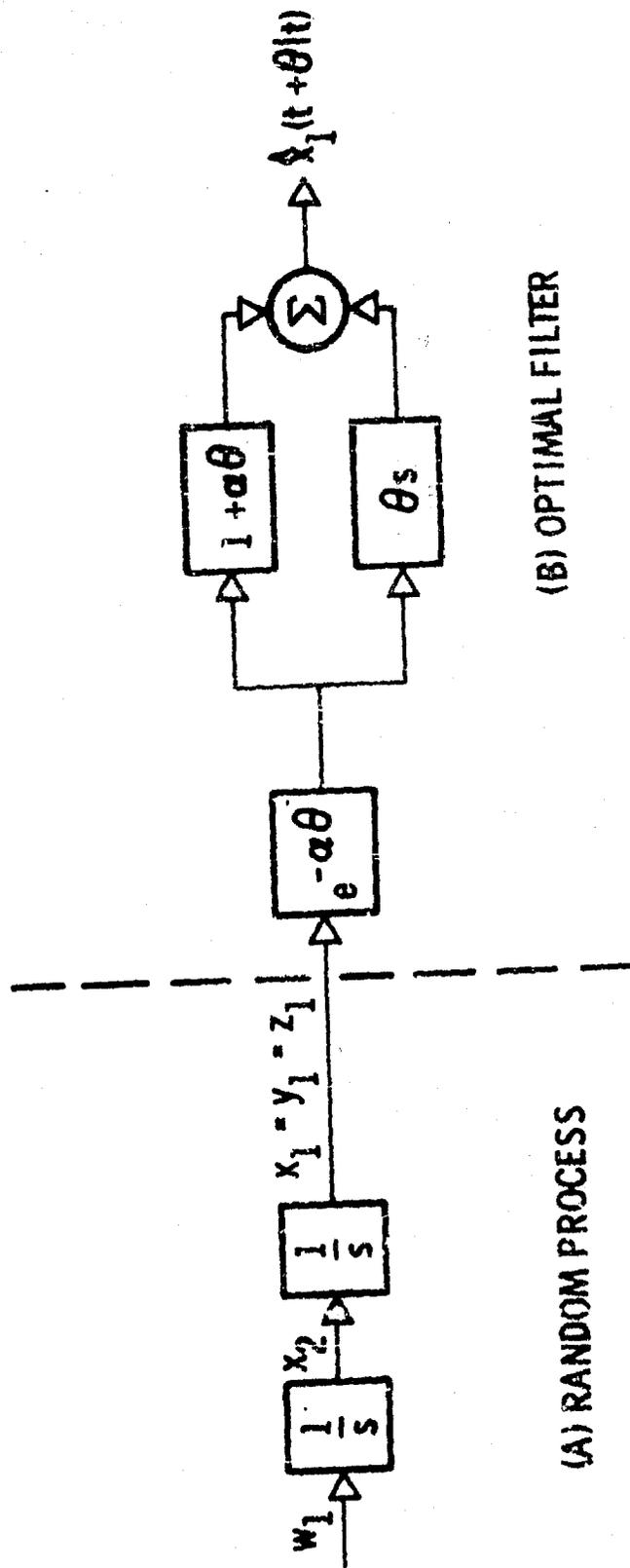


FIG. 4 SIMPLE PREDICTION PROBLEM

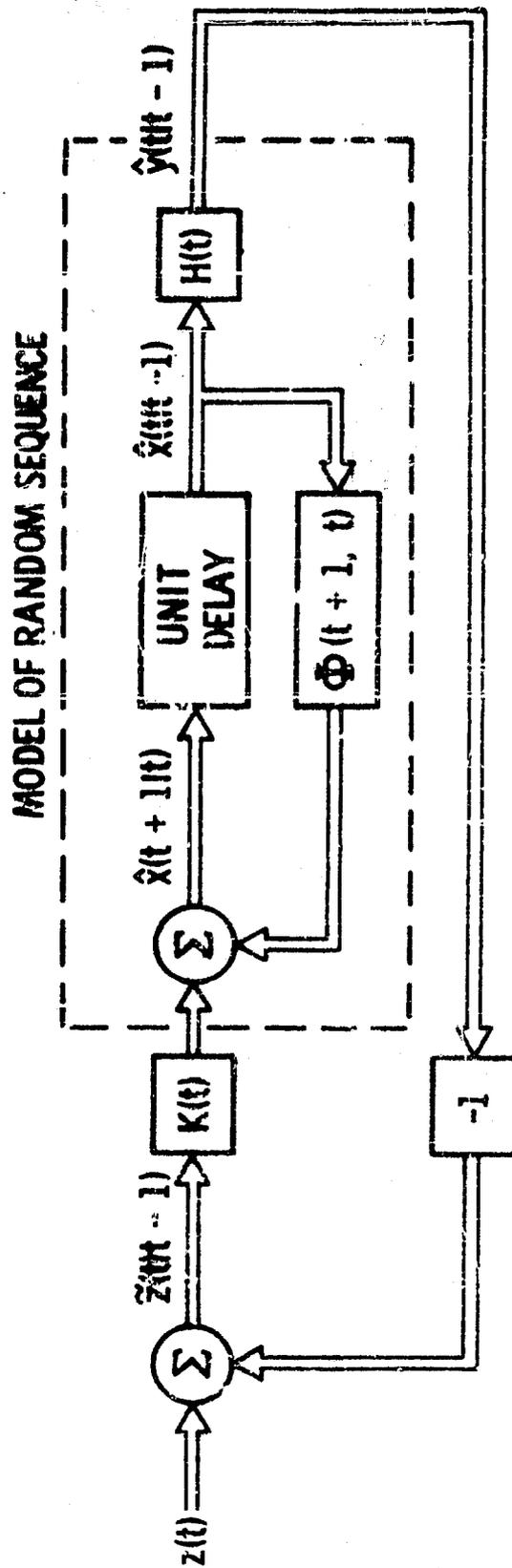


FIG. 5 OPTIMAL FILTER (DISCRETE-TIME)

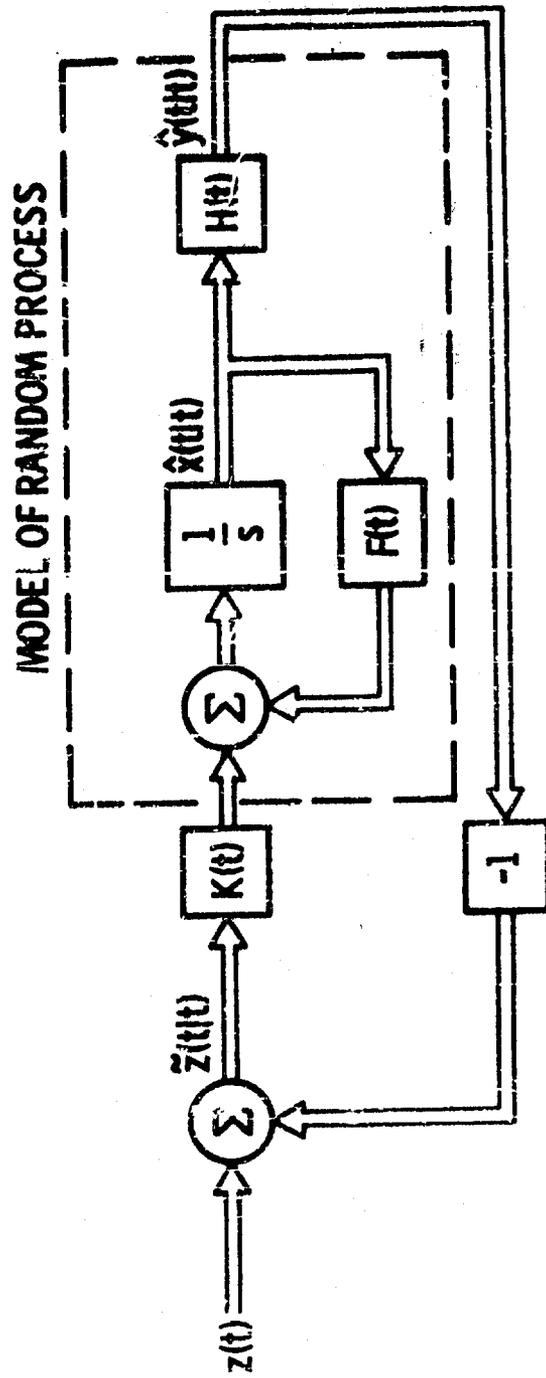
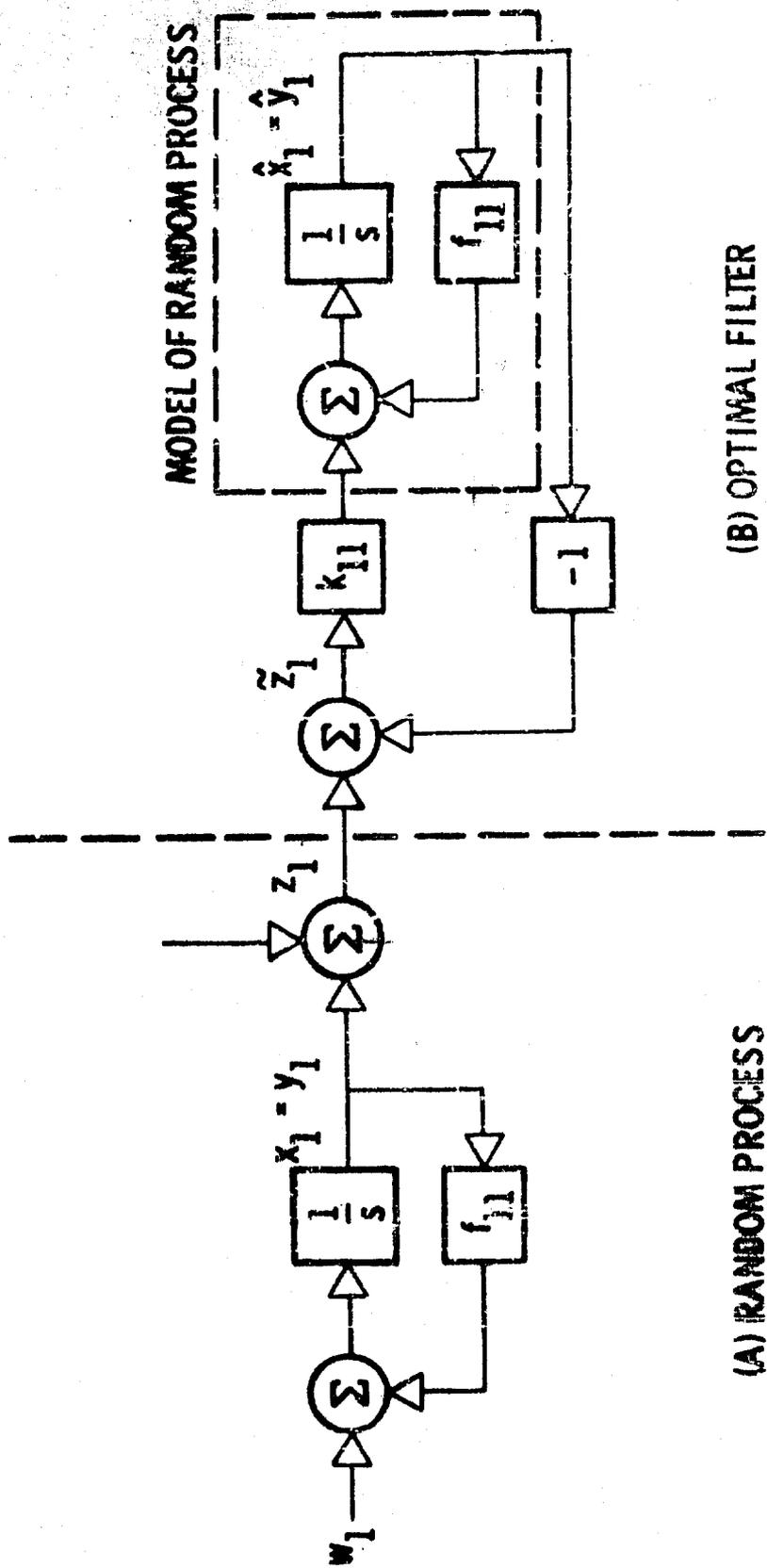


FIG. 6 OPTIMAL FILTER (CONTINUOUS-TIME)



(A) RANDOM PROCESS

(B) OPTIMAL FILTER

FIG. 7 EXAMPLES (14. I) AND (14. II)

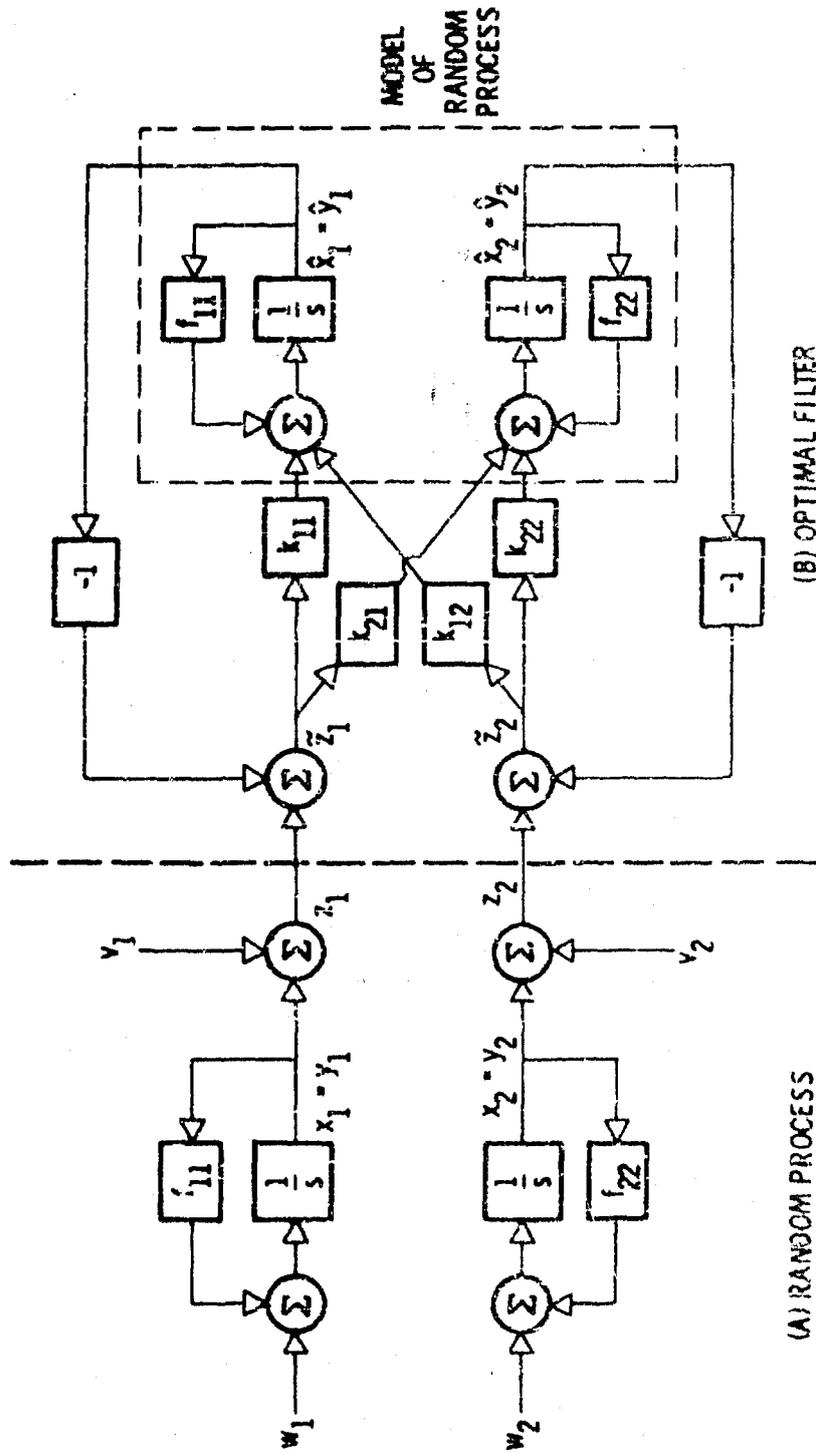


FIG. 8 EXAMPLE (14.20)

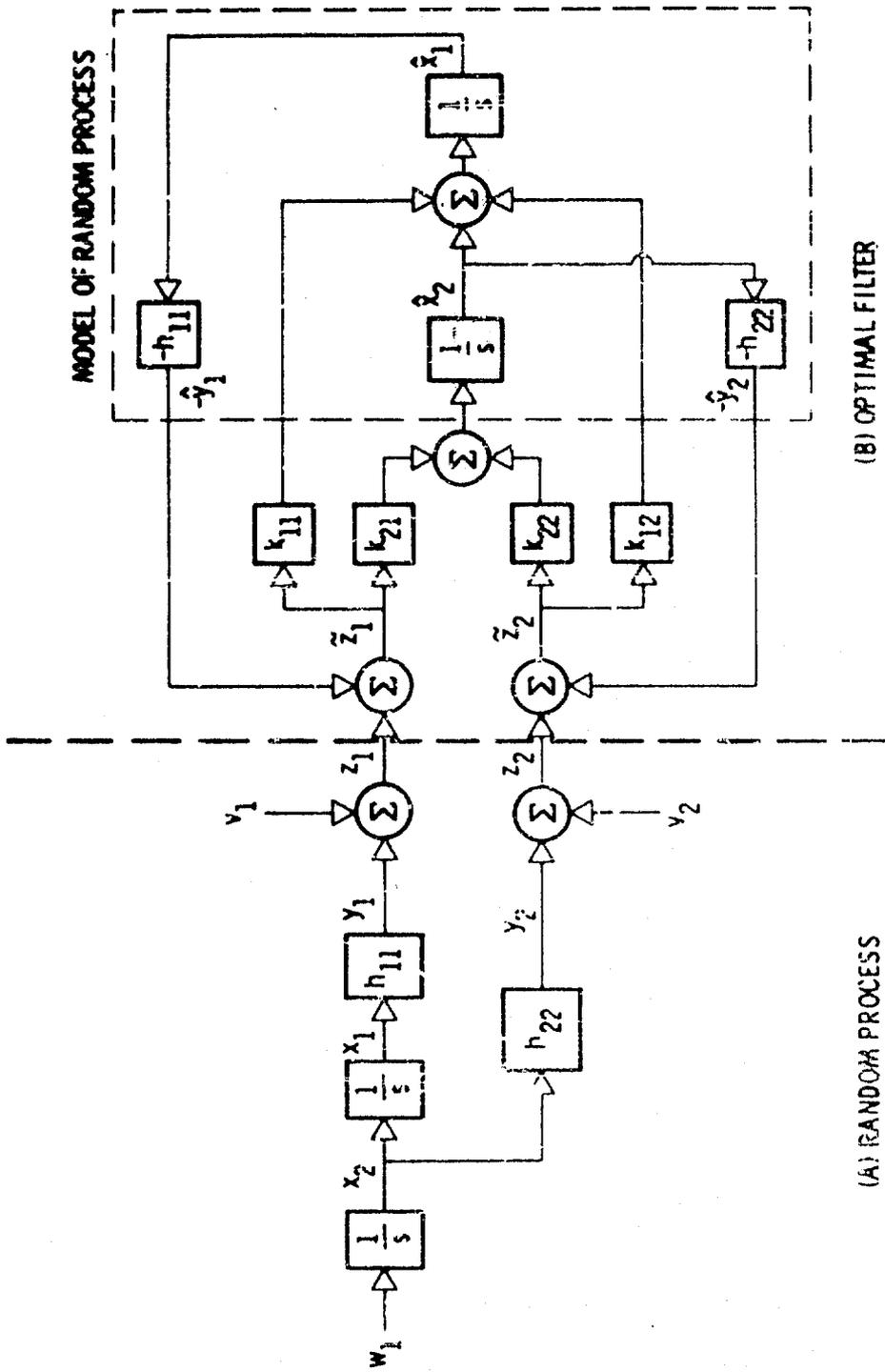


FIG. 9 EXAMPLE (14.50)

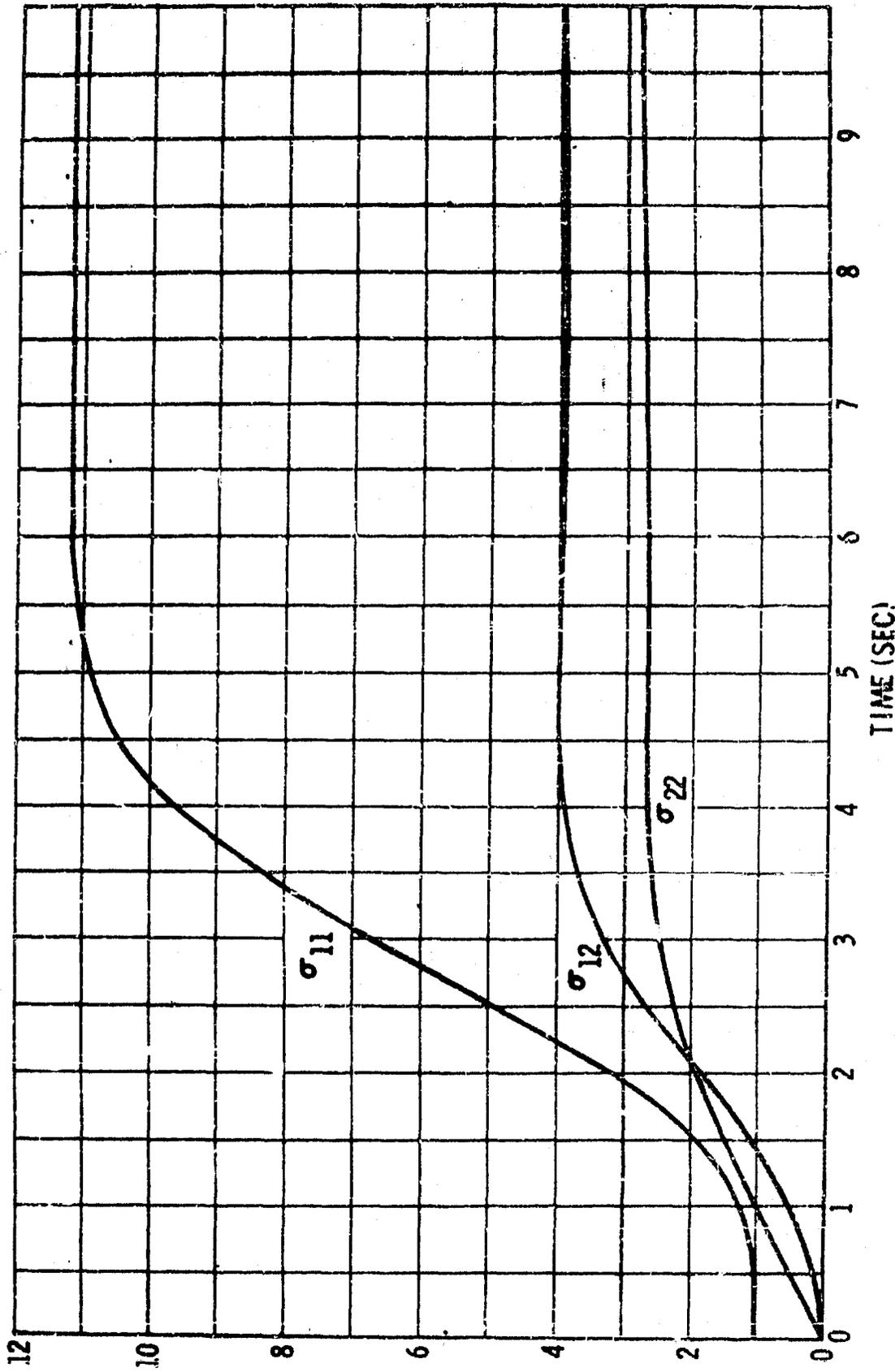


FIG. 10 SOLUTION OF THE VARIANCE EQUATION, CASE A, EXAMPLE (14.50)

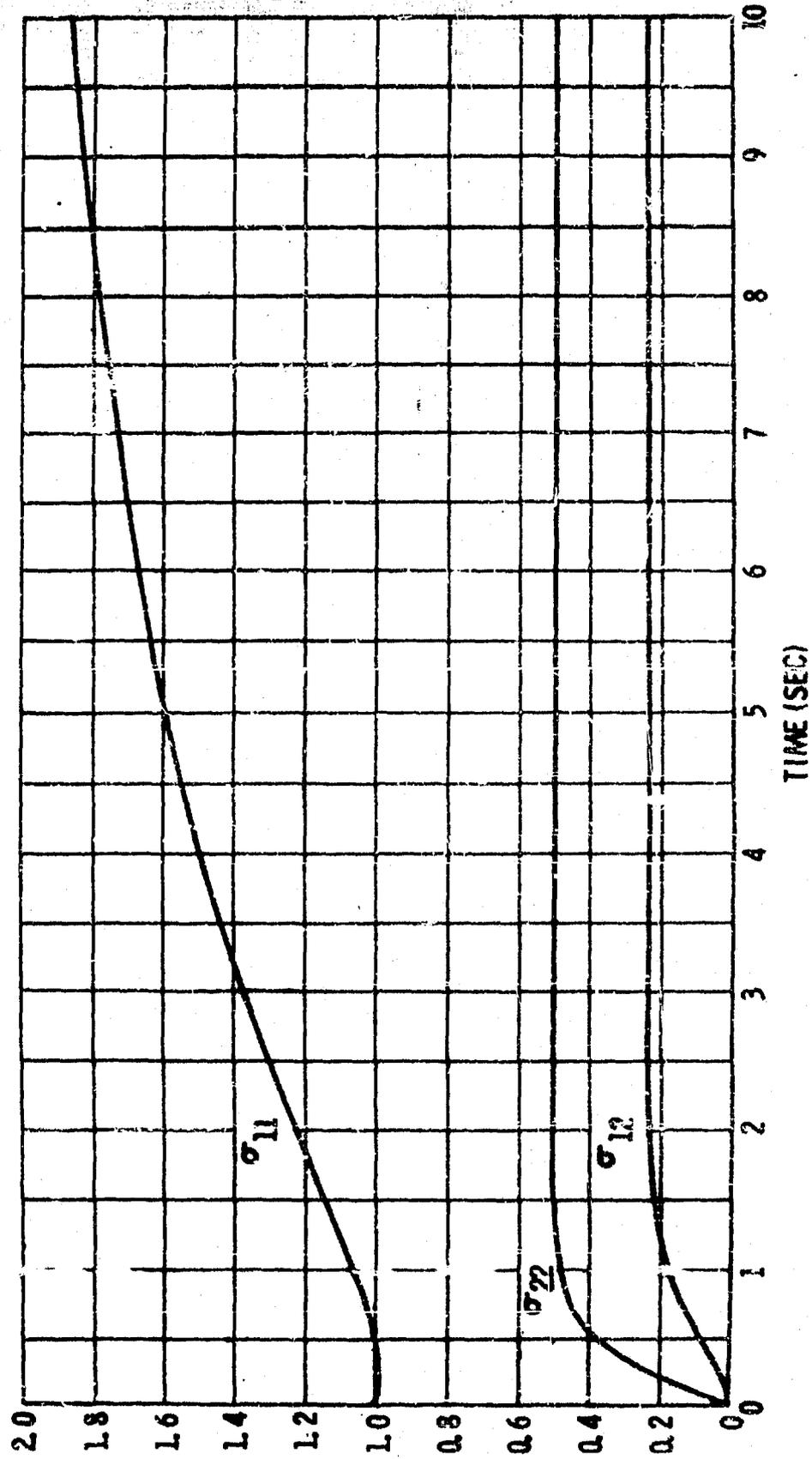


FIG. 11 SOLUTION OF THE VARIANCE EQUATION, CASE B, EXAMPLE (14.50)

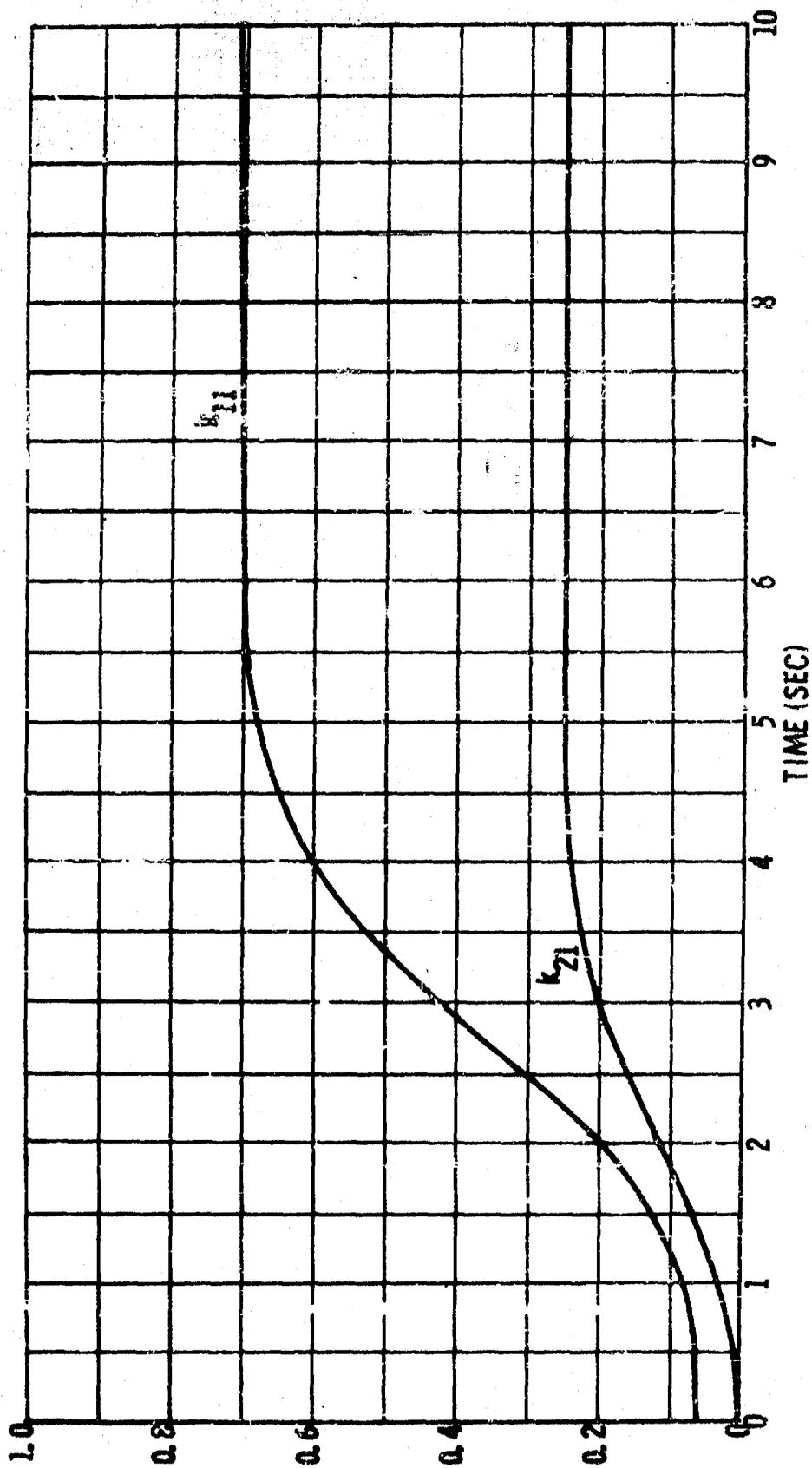


FIG. 12 OPTIMAL GAINS, CASE A, EXAMPLE (14.50)

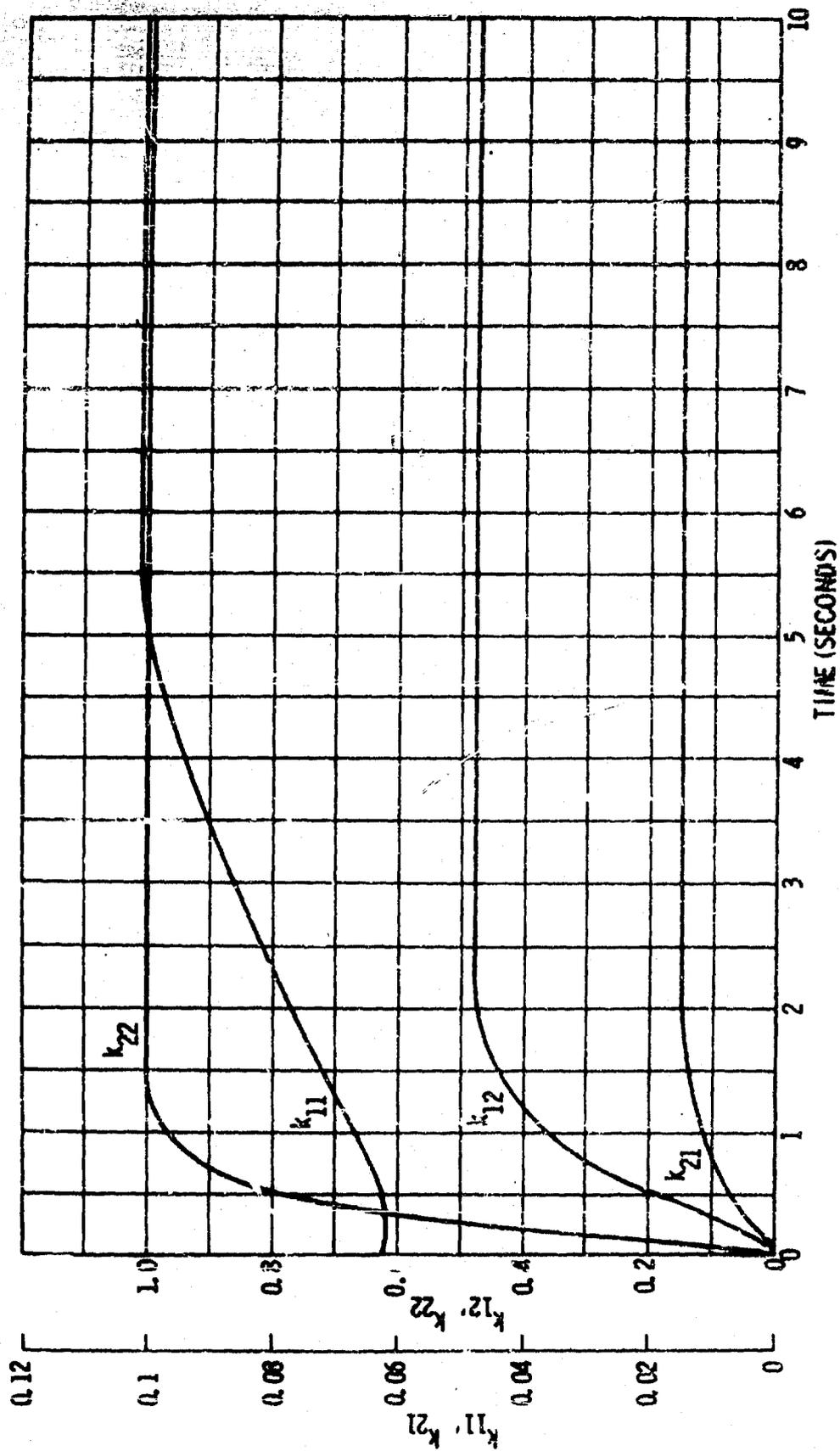


FIG. 13 OPTIMAL GAINS, CASE B, EXAMPLE (14.50)

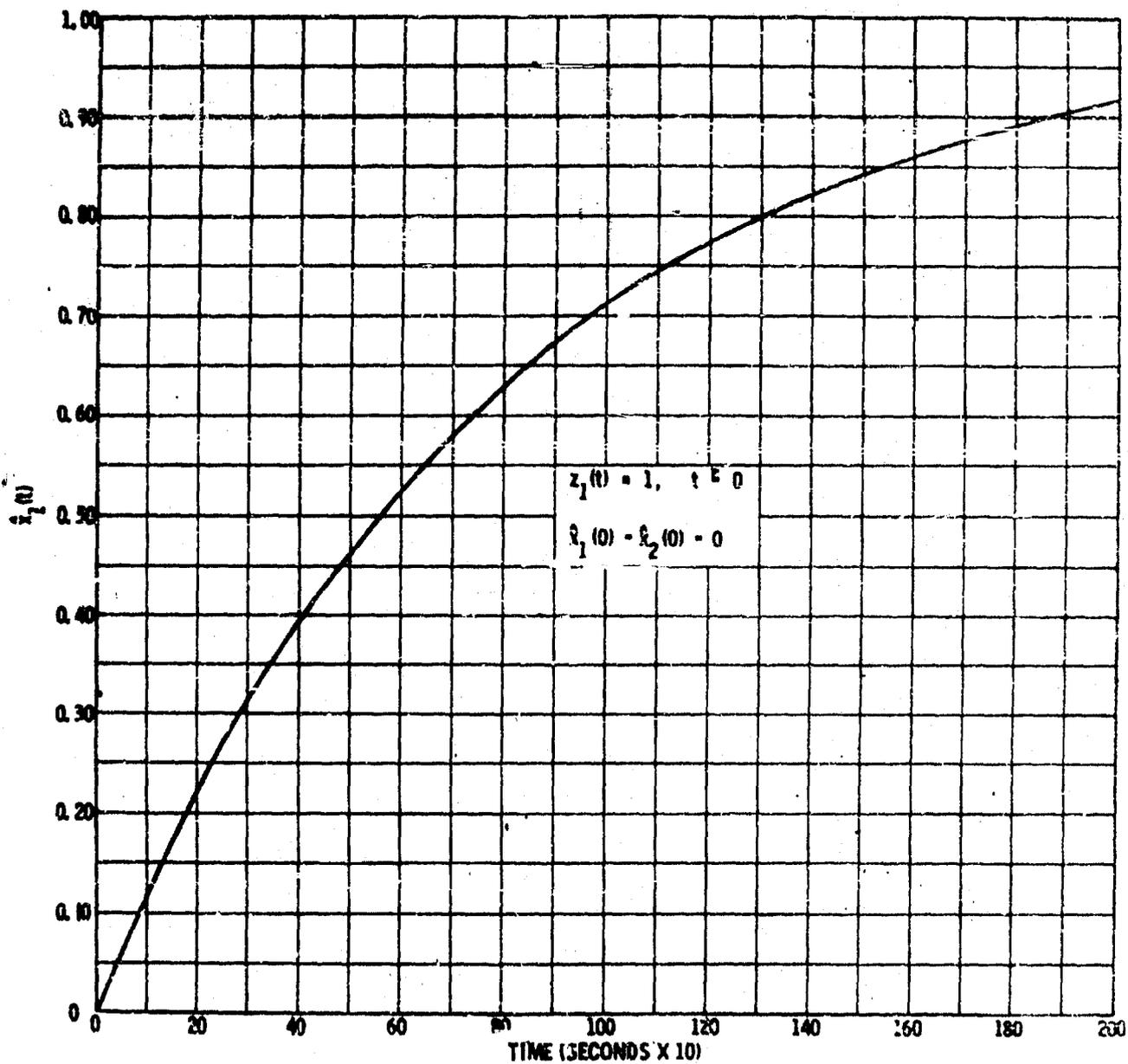


FIG. 14 UNIT STEP RESPONSE OF OPTIMAL FILTER, CASE A, EXAMPLE (14.59)

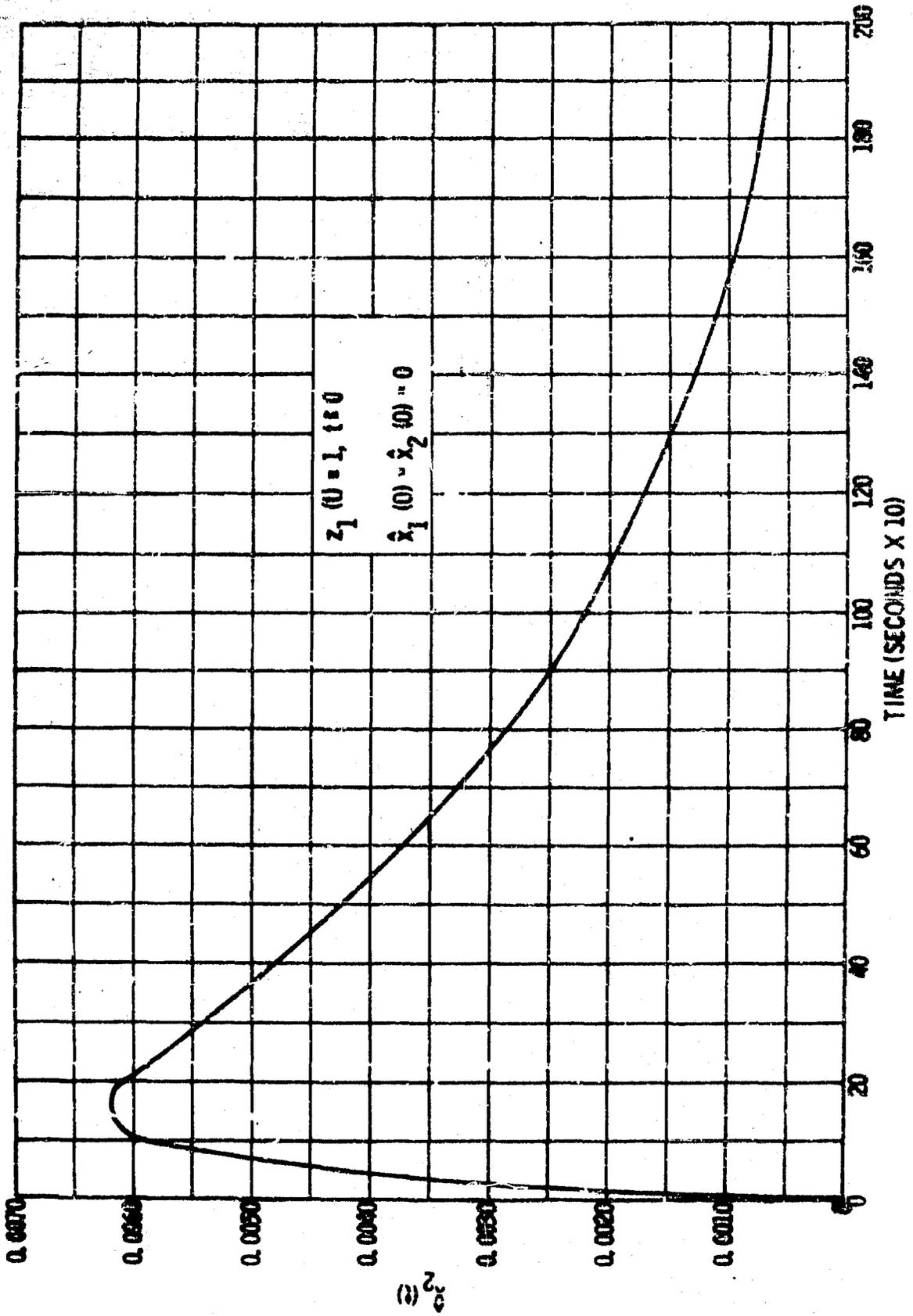


FIG. 15 UNIT STEP RESPONSE OF OPTIMAL FILTER, CASE A, EXAMPLE (14.50)

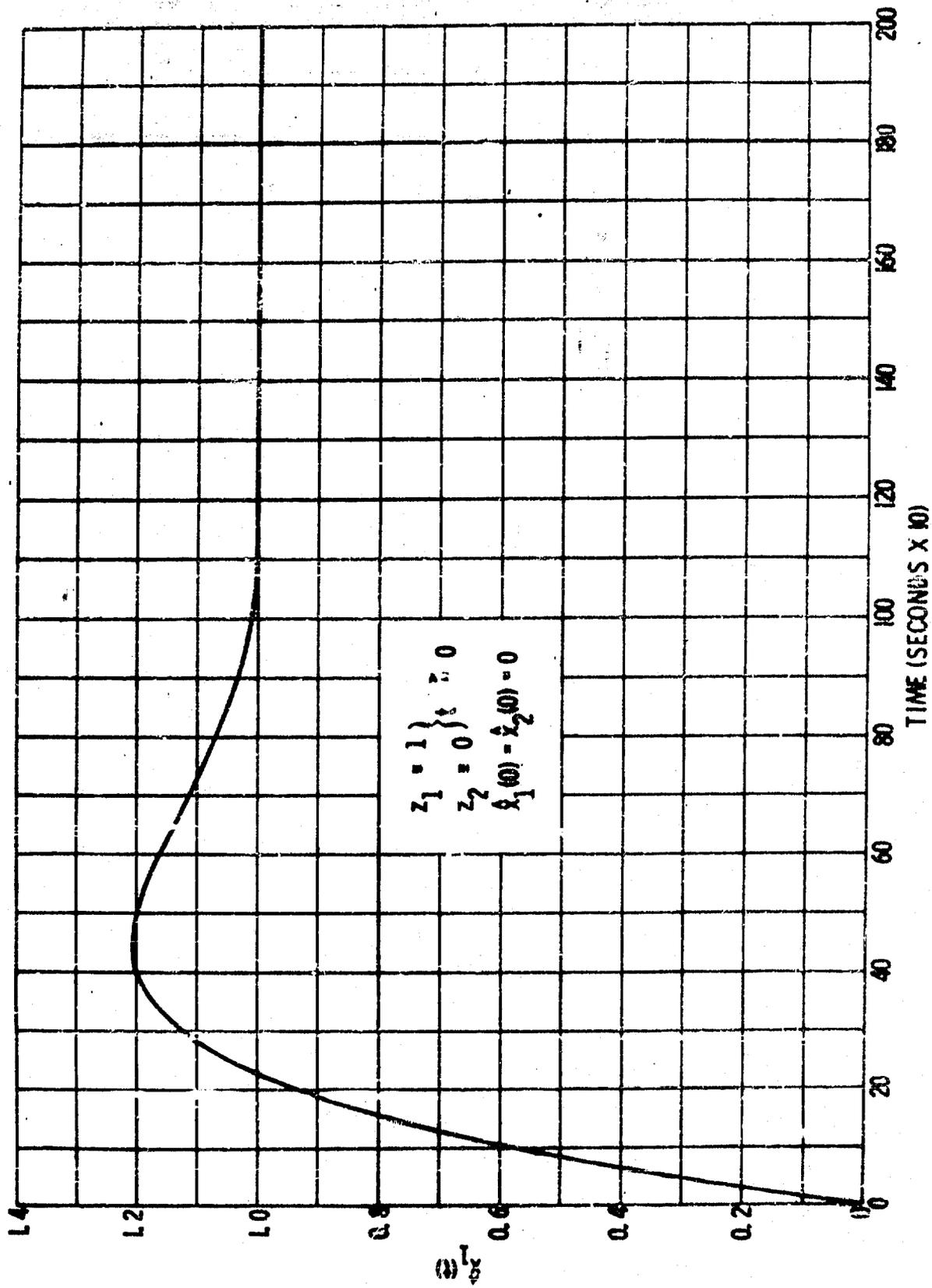


FIG. 16 UNIT STEP RESPONSE OF OPTIMAL FILTER, CASE B, EXAMPLE (14.50)

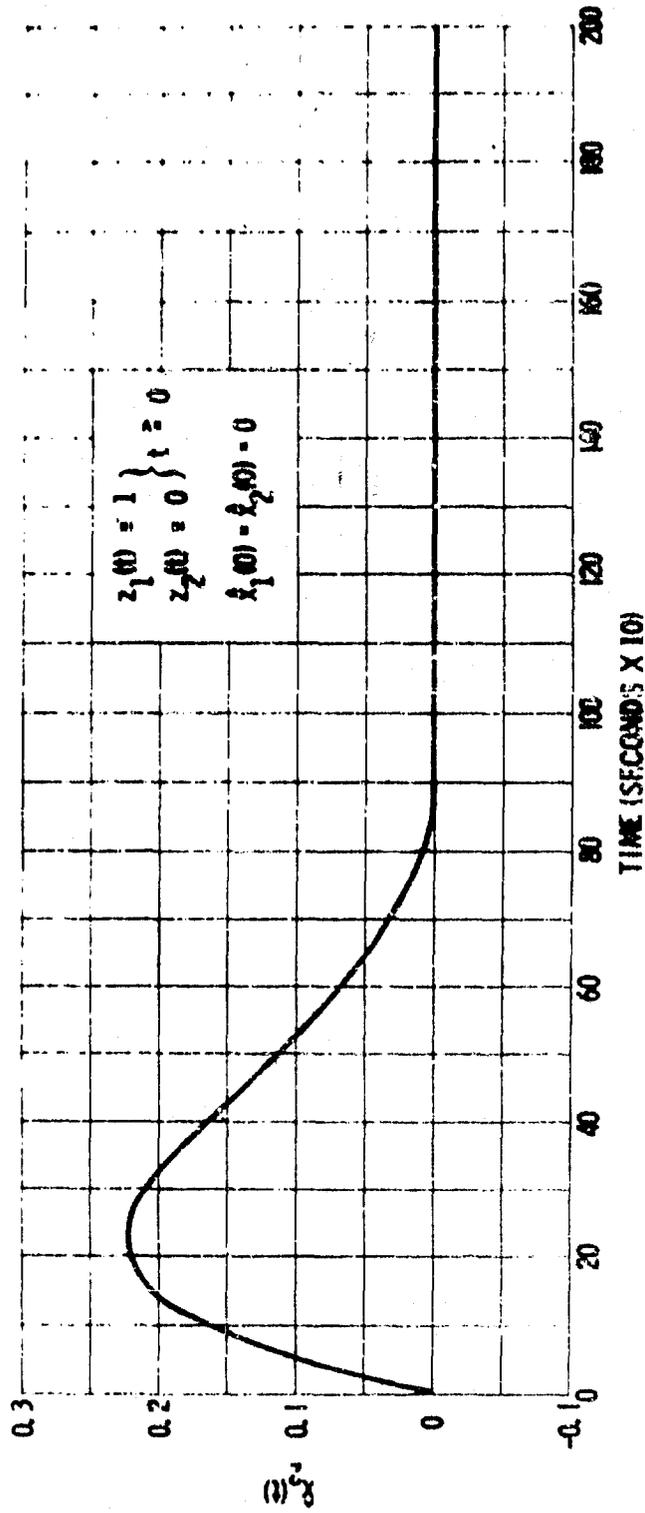


FIG. 17 UNIT STEP RESPONSE OF OPTIMAL FILTER, CASE B, EXAMPLE (14.50)

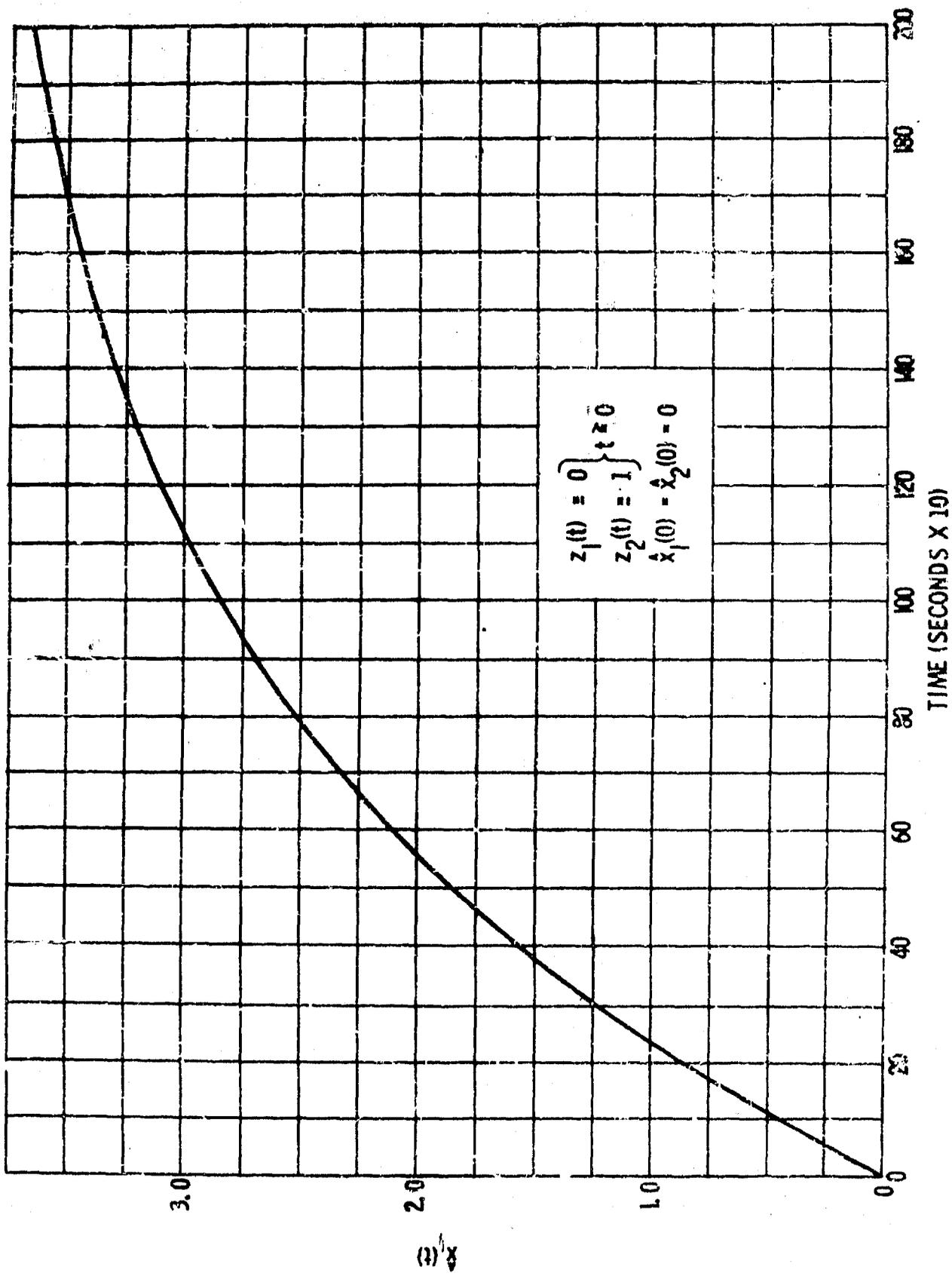


FIG. 18 UNIT STEP RESPONSE OF OPTIMAL FILTER, CASE B,

EXAMPLE (IL 50)

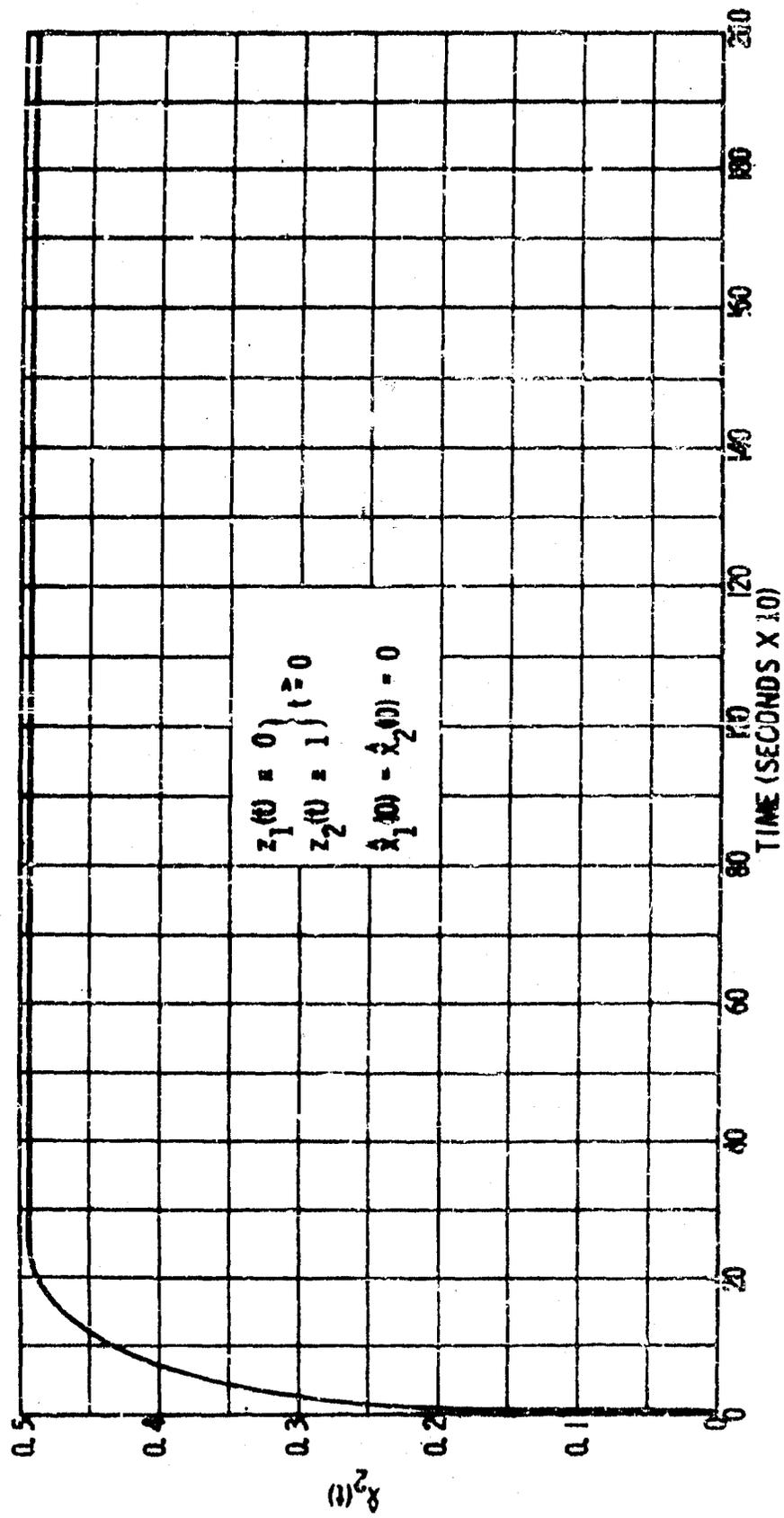


FIG. 19 UNIT STEP RESPONSE OF OPTIMAL FILTER, CASE B, EXAMPLE (14.50)

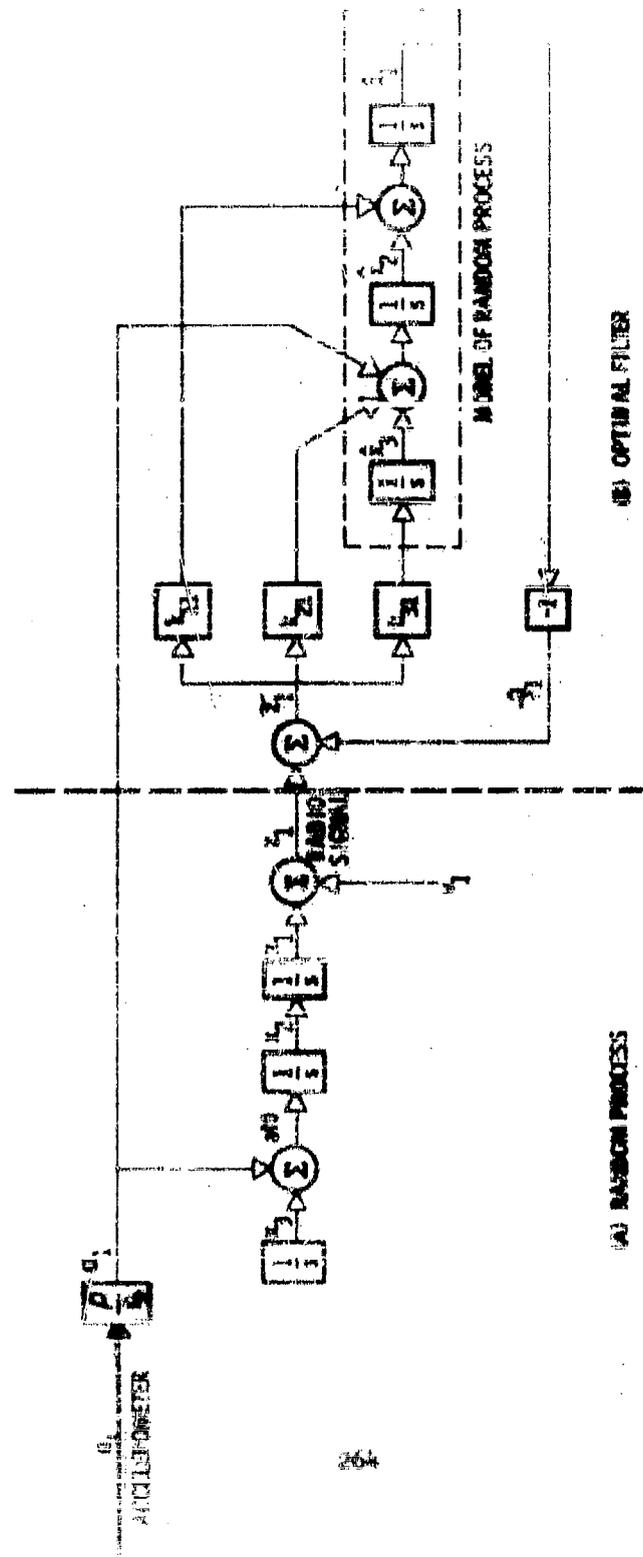


Fig. 20. Example 14.52

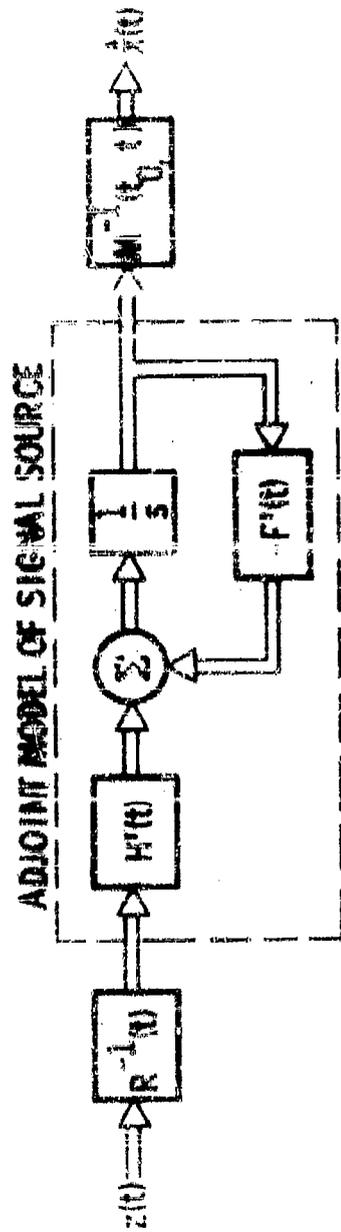
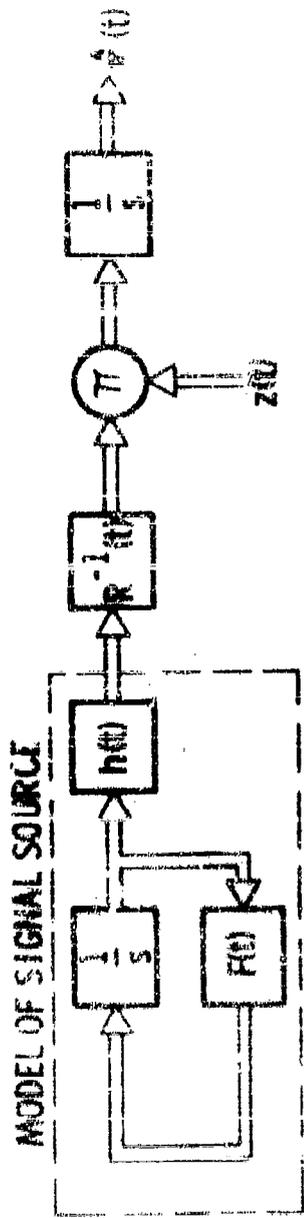
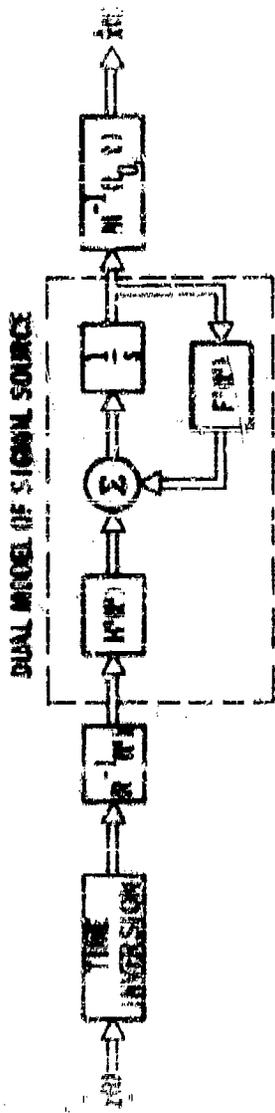
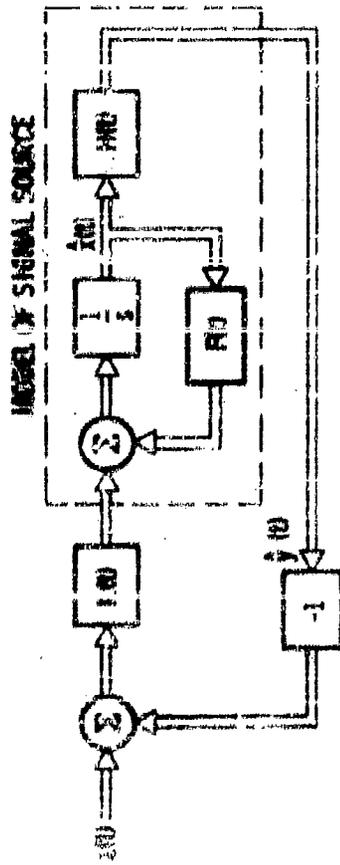


FIG. 21 PHYSICAL REALIZATION OF OPTIMAL ESTIMATOR



(C) REALIZATION BY DUAL



(D) REALIZATION BY FEEDBACK

FIG. 21 PHYSICAL REALIZATION OF OPTIMAL ESTIMATOR (CONTINUED)

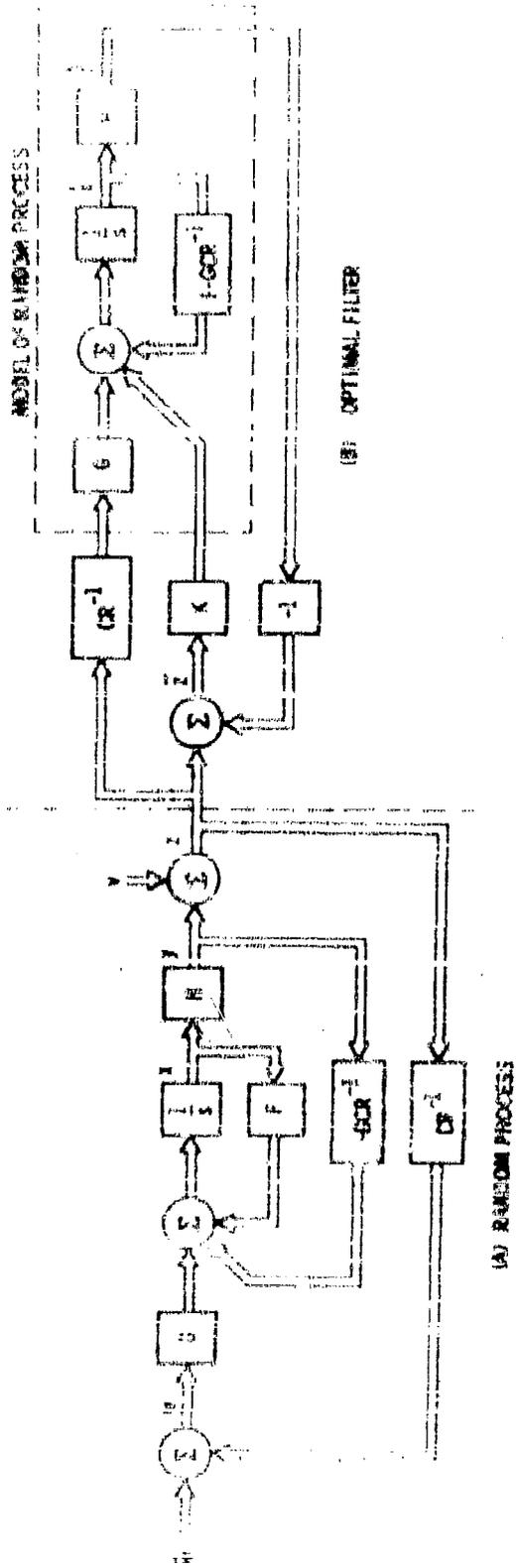


FIG. 2: TRANSFORMATION TO A GAUSSIAN FORM

Aeronautical Systems Division, Dir/Aero-
mechanics, Flight Control Laboratory,
Wright-Patterson Air Force Base, Ohio.
Rpt Nr ASD-TR-61-27, Vol I. FUNDA-
MENTAL STUDY OF ADAPTIVE CONTROL
SYSTEMS. Final report, Apr 62, 271 p.
incl illus., 64 refs.

Unclassified Report

This is the first detailed report on Contract
AF 33(616)-6952, concerned with the funda-
mental investigation of adaptive control
systems.
A general survey of modern analytical
methods of control theory is presented, with
emphasis on special topics relevant to adap-
tive system problems. In addition, it is

(over)

shown how these methods are implemented
by means of digital computers. A set of
new matrix sub-routines is described in
detail.

To render this report as nearly self-
contained as was considered feasible, a
comprehensive appendix has been included.
This appendix is referred to in the body of
the report as [Kalman, 1961 C].

1. Control systems
 2. Optimal control theory
 3. Adaptive control theory
- I. AFSC Project 8225, Task 82181
 - II. Contract AF 33(616)-6952
 - III. RIAS Div., The Martin Company, Baltimore, Md.
 - IV. R.E. Kalman, et al.
 - V. In ASTIA collection
 - VI. Not avail fr OTS

Aeronautical Systems Division, Dir/Aero-
mechanics, Flight Control Laboratory,
Wright-Patterson Air Force Base, Ohio.
Rpt Nr ASD-TR-61-27, Vol I. FUNDA-
MENTAL STUDY OF ADAPTIVE CONTROL
SYSTEMS. Final report, Apr 62, 271 p.
incl illus., 64 refs.

Unclassified Report

This is the first detailed report on Contract
AF 33(616)-6952, concerned with the funda-
mental investigation of adaptive control
systems.
A general survey of modern analytical
methods of control theory is presented, with
emphasis on special topics relevant to adap-
tive system problems. In addition, it is

(over)

shown how these methods are implemented
by means of digital computers. A set of
new matrix sub-routines is described in
detail.

To render this report as nearly self-
contained as was considered feasible, a
comprehensive appendix has been included.
This appendix is referred to in the body of
the report as [Kalman, 1961 C].

1. Control systems
 2. Optimal control theory
 3. Adaptive control theory
- I. AFSC Project 8225, Task 82181
 - II. Contract AF 33(616)-6952
 - III. RIAS Div., The Martin Company, Baltimore, Md.
 - IV. R.E. Kalman, et al.
 - V. In ASTIA collection
 - VI. Not avail fr OTS

Best Available Copy

Aeronautical Systems Division, Dir/Aeromechanics, Flight Control Laboratory, Wright-Patterson Air Force Base, Ohio. Rpt Nr ASD-TR-61-27, Vol I, FUNDAMENTAL STUDY OF ADAPTIVE CONTROL SYSTEMS. Final report, Apr 62, 271 P. incl illus., 64 refs.

Unclassified Report
 This is the first detailed report on Contract AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.
 A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is

(over)

shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.
 To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

Best Available Copy

Control systems
 Optimal control theory
 Adaptive control theory
 AFSC Project 8225, Task 82181

Contract AF 33(616)-6952
 III. RIAS Div., The Martin Company, Baltimore, Md.
 IV. R. E. Kalman, et al.
 V. In ASTIA collection
 VI. Not avail fr OTS

Aeronautical Systems Division, Dir/Aeromechanics, Flight Control Laboratory, Wright-Patterson Air Force Base, Ohio. Rpt Nr ASD-TR-61-27, Vol I, FUNDAMENTAL STUDY OF ADAPTIVE CONTROL SYSTEMS. Final report, Apr 62, 271 P. incl illus., 64 refs.

Unclassified Report
 This is the first detailed report on Contract AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.
 A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is

(over)

shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.
 To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

Control systems
 Optimal control theory
 Adaptive control theory
 AFSC Project 8225, Task 82181

Contract AF 33(616)-6952
 III. RIAS Div., The Martin Company, Baltimore, Md.
 IV. R. E. Kalman, et al.
 V. In ASTIA collection
 VI. Not avail fr OTS

Aeronautical Systems Division, Dir/Aero-mechanics, Flight Control Laboratory, Wright-Patterson Air Force Base, Ohio. Rpt Nr ASD-TR-61-27, Vol I. FUNDAMENTAL STUDY OF ADAPTIVE CONTROL SYSTEMS. Final report, Apr 62, 271 p. Incl illus., 64 refs.

Unclassified Report

This is the first detailed report on Control AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.

A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is

(over)

shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.

To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

1. Control systems
2. Optimal control theory
3. Adaptive control theory
I. AFSC Project 8225, Task 82181
II. Contract
AF 33(616)-6952
III. RIAS Div., The Martin Company, Baltimore, Md.
IV. R. E. Kalman, et al.
V. In ASTIA collection
VI. Not avail fr OTS

Unclassified Report

This is the first detailed report on Contract AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.

A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is

(over)

shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.

To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

Aeronautical Systems Division, Dir/Aero-mechanics, Flight Control Laboratory, Wright-Patterson Air Force Base, Ohio. Rpt Nr ASD-TR-61-27, Vol I. FUNDAMENTAL STUDY OF ADAPTIVE CONTROL SYSTEMS. Final report, Apr 62, 271 p. Incl illus., 64 refs.

Unclassified Report

This is the first detailed report on Contract AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.

A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is

(over)

shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.

To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

1. Control systems
2. Optimal control theory
3. Adaptive control theory
I. AFSC Project 8225, Task 82181
II. Contract
AF 33(616)-6952
III. RIAS Div., The Martin Company, Baltimore, Md.
IV. R. E. Kalman, et al.
V. In ASTIA collection
VI. Not avail fr OTS

Unclassified Report

This is the first detailed report on Contract AF 33(616)-6952, concerned with the fundamental investigation of adaptive control systems.

A general survey of modern analytical methods of control theory is presented, with emphasis on special topics relevant to adaptive system problems. In addition, it is

(over)

shown how these methods are implemented by means of digital computers. A set of new matrix sub-routines is described in detail.

To render this report as nearly self-contained as was considered feasible, a comprehensive appendix has been included. This appendix is referred to in the body of the report as [Kalman, 1961 C].

Best Available Copy

Aeronautical Systems Division, Dir/Aero-
mechanics, Flight Control Laboratory,
Wright-Patterson Air Force Base, Ohio.
Rpt Nr ASD-TR-61-27, Vol I. FUNDA-
MENTAL STUDY OF ADAPTIVE CONTROL
SYSTEMS. Final report, Apr 62, 271 p.
incl illus., 64 refs.

Unclassified Report

This is the first detailed report on Contract
AF 33(616)-6952, concerned with the funda-
mental investigation of adaptive control
systems.
A general survey of modern analytical
methods of control theory is presented, with
emphasis on special topics relevant to adap-
tive systems problems. In addition, it is

(over)

shown how these methods are implemented
by means of digital computers. A set of
new matrix sub-routines is described in
detail.

To render this report as nearly self-
contained as was considered feasible, a
comprehensive appendix has been included.
This appendix is referred to in the body of
the report as [Katzman, 1961 C].

1. Control systems
2. Optimal control
theory
3. Adaptive control
theory
I. AFSC Project 8225,
Task 82181
II. Contract
AF 33(616)-6952

III. RIAS Div., The
Mardin Company,
Baltimore, Md.
IV. R. E. Kalman, et al.
V. In ASTIA collection
VI. Not avail fr OTS

Aeronautical Systems Division, Dir/Aero-
mechanics, Flight Control Laboratory,
Wright-Patterson Air Force Base, Ohio.
Rpt Nr ASD-TR-61-27, Vol I. FUNDA-
MENTAL STUDY OF ADAPTIVE CONTROL
SYSTEMS. Final report, Apr 62, 271 p.
incl illus., 64 refs.

Unclassified Report

This is the first detailed report on Contract
AF 33(616)-6952, concerned with the funda-
mental investigation of adaptive control
systems.
A general survey of modern analytical
methods of control theory is presented, with
emphasis on special topics relevant to adap-
tive systems problems. In addition, it is

(over)

shown how these methods are implemented
by means of digital computers. A set of
new matrix sub-routines is described in
detail.

To render this report as nearly self-
contained as was considered feasible, a
comprehensive appendix has been included.
This appendix is referred to in the body of
the report as [Katzman, 1961 C].

1. Control systems
2. Optimal control
theory
3. Adaptive control
theory
I. AFSC Project 8225,
Task 82181
II. Contract
AF 33(616)-6952

III. RIAS Div., The
Mardin Company,
Baltimore, Md.
IV. R. E. Kalman, et al.
V. In ASTIA collection
VI. Not avail fr OTS

Best Available Copy