SYSTEM IDENTIFICATION BY BAYESIAN LEARNING

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

PATRICK J. DONOGHUE

1968

This is to certify that the

thesis entitled

System Identification by Bayesian Learning

presented by

Patrick J. Donoghue

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in E.E.

Major professor

Date

ABSTRACT

SYSTEM IDENTIFICATION BY BAYESIAN LEARNING

by Patrick J. Donoghue

The problem of system identification is of fundamental importance both from a practical and system-theoretic point of view. As such it has received wide attention in the literature. However, much of the work in this area suffers from the lack of a unifying structure. This is especially true of those identification techniques that treat problems concerned with noise-obscured measurements and unobservable random disturbances.

In this thesis the approach taken to system identification is Bayesian learning. The systems considered are described by a finite set of difference equations relating the system states and inputs, and by a finite set of algebraic equations relating the system states, inputs, and observations. The measurement of all states and all inputs is assumed to be obscured by additive noise. Further, the system states are assumed to be influenced by unobservable additive random disturbances. The object of the identification is to determine the matrices or constants that specify the system.

Using Bayesian learning for identification allows the derivation of a general identification algorithm which is practical, includes many previous results as special cases, and provides a framework for solving new problems. The identification algorithm is iterative and has the form of a general stochastic approximation algorithm. Thus the algorithm operates on the data as it becomes available and produces a sequence of estimates for the parameters which specify the system. These estimates are Bayes-optimal in some cases considered and are sub-optimal in others. The specific identification algorithms for each class of systems considered have one simple structure and are computationally feasible.

The introduction of the concepts of identifiability and strong identifiability provides a workable basis from which convergence of the general identification algorithm can be established. Three theorems on the convergence of the identification algorithm to the true system parameters as the number of observations becomes infinite are proved using these concepts. Two of these theorems are new results in identification. A new proof for the third result follows from a strengthening of the hypotheses of the other two theorems.

The general identification algorithm is used to derive specific algorithms for important classes of systems. Algorithms

for both stationary and time-varying linear systems are obtained, and the derivation of algorithms for nonlinear systems is indicated. Other formulations are given, including systems with state disturbances having an unknown mean and systems with generated inputs.

To demonstrate that implementation of the proposed algorithm is not difficult and to demonstrate that the algorithm does indeed converge, a fourth-order digital control system with eight unknown parameters is identified. Computer-simulated results show the behavior of the algorithm under different initial estimates, noise conditions, and a-priori uncertainty.

SYSTEM IDENTIFICATION BY BAYESIAN LEARNING

By

Patrick J. Donoghue

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering

1968

ACKNOWLEDGEMENTS

The author wishes to thank his thesis advisor, Dr. R. C. Dubes, for his continued assistance and inspiration during the preparation of this work and for his invaluable guidance as a friend and educator.

Thanks are also due to Dr. H. E. Koenig and Dr. G. L. Park for their assistance as academic advisors and to Dr. E. A. Nordhaus and Dr. R. B. Zemach for their interest in this work.

Finally the author owes a great debt of love and gratitude to his wife, Sharon, for her unselfish patience and understanding during his graduate studies and the preparation of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF APPENDICES

I INTRODUCTION

This chapter provides a foundation for the results in this thesis. Section 1.1 indicates in a general way how induction and modeling are related. The role of identification in the general structure of system theory is discussed in section 1.2, and the wide range of practical applications for identification techniques is indicated in section 1.3. Various general identification techniques are listed in section 1.4 to show how Bayesian learning is related to other methods. Section 1.5 discusses the basic results that will be derived in the thesis.

1.1 Modeling and Induction

The fundamental goal of virtually all scientific investigation is to acquire knowledge about the real world (except perhaps for the mathematical formalist's investigation [SCH]). This real world may take on diverse appearances depending upon a particular investigator's background and individual goals.

Information about the real world usually appears in the form of causality relations, and one medium through which these causality

relations are displayed is mathematics. In fact, mathematics is practically the only universally accepted tool for handling the relationships which arise in scientific studies. Mathematics provides a systematic procedure for predicting and analyzing the observable characteristics of an investigator's environment. This is particularly true in those systems that are of practical interest. This prediction and analysis is realized by development of a mathematical model, i.e., a finite set of mathematical relationships which interrelate certain observable or measurable quantities of the system under consideration. The derivation or synthesis of the relationships forming the model is accomplished in terms of the structural features of the system. These features are the characteristics of the individual members of the aggregate and their manner of interaction or interconnection. The individual characteristics are obtained by assumption, a-priori information, or sets of fundamental measurements. The test of the mathematical model's validity rests upon the connection between the variables of the mathematical structure and their physical counterpart. Hopefully, the relation between these entities is isomorphic [AHL]. That is, the values assumed by the variables in the mathematical model are in a one-to-one property correspondence with the values that are measured.

An alternative way of arriving at mathematical models of physical systems is to employ inductive inferential relationships [CAR][AY]. That is, a partially specified mathematical model

is assumed to be an accurate counterpart of the physical system and appropriate measurements are taken to confirm or deny the hypothesis. If all variations of the original model fail to accurately predict or account for the observations, then the model is changed or discarded. If, however, suitable manipulation or further specification of the model results in accurate prediction of the observations for some finite time, then the model is retained. Since the phenomena are only observed for a finite time it is impossible to verify the hypothesis completely and only confirmational support for it is obtained [HE][POP]. This is the nature of inductive logic. The model is retained only as long as it is able to account for all the observations. Even if no contradictory evidence occurs for a very long time, the original hypothesis is not necessarily verified. The approach taken in this thesis, called identification by Bayesian learning,is based on inductive inference.

1.2 System Theory and the Identification Problem

In the foregoing discussion the nature of the problem to be considered here was indicated only in a very general way. In this section the problem will be defined in the context of contemporary system theory.

In one of his treatments of general system theory, Zadeh [Z-1] has given a comprehensive listing of the problems that are included in this theoretic structure. Among these problems are

system analysis, synthesis, control, stability, reliability, learning, and signal theory. In addition three distinct aspects of the theory are singled out for special consideration because of their fundamental importance. These are the problems of system characterization, classification and identification.

The first of these problems deals with the representation of input-output relations. These representations may be expressed as solutions of differential equations, state functions, transfer functions, integral operators or in any other convenient form and may change according to whether the system is time-continuous or discrete, stochastic, causal or finite state among other considerations. The second problem, classification, is concerned with determining class membership when a system is assumed to belong to one of a family of classes. The classification relies on observations of system input and output. Two problems that are in this category are: (1) determination of the order of a system's differential or difference equation, and (2) determination of whether a time-varying system is linear or nonlinear. Lastly, and most important for this discussion and often for practical situations, is the problem of system identification. Generally, the identification problem considers means for determining the specific characteristics of a system through observations of the input and output.

A precise definition of identification is now stated [Z-1]: Given a class of systems S where each member of the class is

completely specified, the identification of a system A consists in finding a system $s \in S$ that is input-output equivalent to A. It is important to note that the definition requires input-output equivalence and does not require $s \in S$ and A to be identical. For a given input-output relation, there is generally no unique system representation [K-1][ARN][Z-2]. In this thesis systems are assumed to be completely specified within a parameter set and the purpose of identification is to determine these parameters.

There are three major complications which normally enter into any real identification problem. The first of these is the absence of knowledge concerning the system's initial condition; the second is the presence of random noises obscuring the observations of inputs and outputs; last is the difficulty in establishing a meaningful and convenient method for estimating the system parameters as a function of the observations.

1.3 Applications of System Identification

Probably the most common use of identification techniques is found in the area of automatic control system design. Here a device such as a servomechanism is identified using sinusoidal or step response data. The identification consists of determining poles and zeros or time constants. Knowledge of these system characteristics aids in the design of devices which may improve total system performance.

In process control situations, such as found in oil refineries or batch production industries, the input-output relations may be very complicated or essentially unknown. Only by identifying the characteristics of the process from operating records can any kind of satisfactory control policy be developed.

Many other systems of interest are describable by sets of equations which have slowly varying parameters. Through continuous monitoring, the system can be approximately identified at each instant in time and effectively controlled. Controllers which utilize continuous on-line identification are called (parameter) adaptive controls [AO-1][SK]. Still other systems may vary in a random manner. By using a system identifier in a learning loop, effective learning control systems can be developed [SK].

Identification techniques can also be applied to communication problems. For example a communication channel may sometimes be characterized by a slowly-varying linear system. Continuous identification of the channel can make more efficient communication possible [KAI][HAN][DAL].

Two more areas where identification techniques are of interest are pattern recognition and reliability. The object in pattern recognition problems is to distinguish one probability distribution from another. If each distribution is assumed to be generated by a different system, then an identification scheme

may be effective in discriminating among these patterns. In reliability problems a system reliability index may depend upon the system parameters which can vary. By identifying the system parameters, a check on the reliability of the system can be maintained.

1.4 General Identification Techniques

A wide variety of techniques have been developed to solve some aspect of the identification problem. Many of these techniques were developed to study specific processes of a practical interest [IFAC]. Others have been developed to attack the problem in a very general way [BAL-1, 2][HE].

Most identification techniques may be classified either as statistical or analytic, depending upon whether or not the problem formulation accounts for random effects. These general areas may be subdivided as follows:

A. Statistical Methods [EY]
   1. Parameter Estimation
      a. Maximum Liklihood
         i) conditional [SMI][RAU]
         ii) unconditional [EY]
      b. Regression Analysis and Least Squares
         i) linear [ALB][STI]
         ii) nonlinear [ALB]
      c. Stochastic Approximation [HO-1][LEE] [SAR][SAK-1, 2]
      d. Bayesian Estimation [FK-3][MA][AO-1]

2. Learning Model Techniques [SK][TSY][FK-1] [FU-1]
3. Spectral Analysis (Wiener-Kolmogorov Theory) [DA][LAN][LEV][AND][WE-1, 2][LE-1, 2]

B. Analytic Methods

1. Functional State Representations [HE][BAL-1, 2]
2. Gradient Techniques [BAL-2][MG]
3. Numerical Methods and Approximations [BX][CU-2] [BEL-1, 2][SG-2]
4. Frequency Domain Techniques [SG-1][PU][BEL-1] [CH][AH].

Most identification techniques in the literature apply only to stationary linear systems. Some work has been done on nonlinear problems, but it is a relatively untouched area even in those problems that are of important practical interest. Relatively few studies are based upon mathematical statistics and probability theory. However, work in this area is significant and treats a wide variety of problems.

1.5 Object of the Thesis

In this thesis system identification, as defined in section 1.2, is approached by a method called Bayesian learning. The systems considered are represented by a finite set of difference equations relating the system states and inputs, and by a finite set of algebraic equations relating the system states, inputs and observations. Identification of such systems when the states are subjected to

unobservable additive random disturbances and when observations are corrupted by additive noise will be examined in this study.

Most of the work in this area lacks a unifying approach. The purpose of this study is to derive a general identification algorithm which is convenient and practical, includes many previous results, extends the applicability of some of them, and allows the derivation of new results. The identification algorithm is iterative. That is, it operates on the data as it becomes available and produce a sequence of estimates for the parameters which specify the system. These estimates are approximations to estimates which are optimal in a sense to be defined, by virtue of an approximation to Bayes' rule. The specific algorithms for each class of systems have one simple structure and are computationally feasible. Furthermore, the algorithms converge (in various senses) to the true parameter values as the number of observations becomes infinite.

The main contribution of the thesis consists in the formulation and solution of a general identification problem in the structure of Bayesian learning. This formulation allows the derivation of algorithms for important classes of systems; namely linear stationary, linear time-varying and nonlinear systems. Three important convergence theorems are proved, two of which are new results in identification. A detailed examination of the properties of the identification algorithm for a specific system is carried out by digital computer simulation.

II FORMULATION OF THE IDENTIFICATION PROBLEM AS BAYESIAN LEARNING

Identification by Bayesian learning is a special method that comes under the general heading of statistical estimation. After the structure of estimators that are useful in identification is indicated in section 2.1, the scope of previous results in this area is given in section 2.2. The basic relations used to derive algorithms are developed in section 2.3. Important implications of using these particular relations are discussed in section 2.4.

2.1 Identification by Estimation and Learning

System identification by methods that are based on statistical estimation have found wide application. These methods can account for the effects of random observation noises as well as unknown random inputs or disturbances. Further, they may be applied to nonlinear as well as linear systems and are applicable to both on-line and off-line problems.

One way of formulating the identification problem of section 1.2 follows. Let the system state equations be

$$x_{k+1} = f_k(x_k, u_k, \xi_k, \alpha) \qquad 2.1.1$$

for $k = 0, 1, \ldots$, where x_k is an n-vector called the system state, u_k is an r-vector of inputs, ξ_k is a sample from an n-dimensional random process*, $f_k(.)$ is a known vector function of its arguments at each time t_k and α is a p-vector of fixed but unknown parameters. The system observation equations are

$$y_k = g_k(x_k, \eta_k^1, \beta) \qquad 2.1.2$$

$$v_k = u_k + \eta_k^2 \qquad 2.1.3$$

where y_k is an m-vector of observations, v_k is an r-vector of input observations, η_k^1 and η_k^2 are samples from m and r-dimensional independent random processes, $g_k(.)$ is a known function of its arguments for each k and β is a q-vector of fixed but unknown parameters. Identification of the system represented by 2.1.1, 2.1.2, 2.1.3, consists in the determination of the parameters α and β.

Since random disturbances enter into the system equations, α and β will in general be impossible to find exactly. Rather, successive estimates of these parameters that converge (in some sense) to the true values of α and β will be sought. In this context an identification algorithm will learn the true parameter values.

---

*In this thesis a random variable and the values it takes on will be denoted by the same letter. The meaning will be clear from the context.

Knowledge of $f_k(.)$ and $g_k(.)$ is required since identification and not classification or characterization is being considered.

Successive estimates of α and β will be derived in the framework of Bayesian estimation. In particular, a sequence of estimators $\{\hat{\psi}_k\}$ of the composite vector $\psi = (\alpha, \beta)$ is sought. This sequence of estimators is to depend upon the observations y_k and v_k, $k = 0, 1, \ldots$. The loss function $L(\hat{\psi}_k, \psi)$ represents the loss incurred when the estimator $\hat{\psi}_k$ based on the first k observations is used and ψ is the true parameter value. This loss function is assumed to be non-negative and to have a relative minimum for $\hat{\psi}_k = \psi$. Denoting by y^k the set of observations $(y_k, y_{k-1}, \ldots, y_o)$, and by v^k the set of observations $(v_k \ldots v_o)$, the conditional risk or expected loss in choosing $\hat{\psi}_k$ as an estimator of ψ is defined by the conditional expectation,

$$E\{L(\hat{\psi}_k, \psi)/\psi\} = \int_{y^k, v^k} L(\hat{\psi}_k, \psi)\, p(y^k, v^k/\psi)\, d(y^k, v^k)$$

where $p(y^k, v^k/\psi)$ represents the joint probability density of the observations conditioned on ψ.* Letting $p(\psi)$ denote the a-priori density of the parameters, the expected risk of choosing $\hat{\psi}_k$ is defined by

---

*For notational convenience, in all of the following development p(.) will be used to denote the density function of a random variable and the argument of the density will specify the random variable.

$$R(\hat{\psi}_k) = E\{E\{L(\hat{\psi}_k, \psi)/\psi\}\}$$

$$= \int_{\psi}\int_{y^k, v^k} L(\hat{\psi}_k, \psi)p(v^k, y^k/\psi)p(\psi)d(y^k, v^k, \psi) \quad .$$

That function $\hat{\psi}_k$ which minimize $R(\hat{\psi}_k)$ is called the minimum risk or Bayes estimator of ψ for each k. Since

$$p(v^k, y^k/\psi)p(\psi) = p(\psi/v^k, y^k)p(v^k, y^k)$$

then

$$R(\hat{\psi}_k) = \int_{y^k, v^k} p(v^k, y^k) \int_{\psi} L(\hat{\psi}_k, \psi)p(\psi/v^k, y^k)d(y^k, v^k, \psi)$$

where $p(\psi/y^k, v^k)$ is the a-posteriori density of the parameters evaluated at ψ conditioned on y^k, v^k and

$$p(v^k, y^k) = \int_{\psi} p(v^k, y^k/\psi)p(\psi)\, d\psi .$$

Since $p(v^k, y^k) \geq 0$, $R(\hat{\psi}_k)$ can be minimized at each k by minimizing the inner integral for each y^k, v^k. This is often done by solving the gradient equation associated with $R(\hat{\psi}_k)$.

To form recursive estimates of ψ as the observations (v_k, y_k) become available, the a-priori density $p(\psi)$ is replaced by $p(\psi/v^{k-1}, y^{k-1})$ and a relation between $p(\psi/v^k, y^k)$ and $p(\psi/v^{k-1}, y^{k-1})$ is established by Bayes' rule. Furthermore, it is possible to show under conditions of convex loss functions and symmetric density functions that this minimum risk is realized by taking $\hat{\psi}_k$ to be the conditional expectation [DE],

$$\hat{\psi}_k = E\{\psi / v^k, y^k\} .$$

These conditions hold for the very important case of gaussian random variables and squared error loss functions. This particular case is used extensively in the development of the system identification algorithms.

2.2 Previous Results in Identification by Estimation and Learning

A large number of publications deal with system identification when random inputs and/or observation noises are present. Only a relative few of these utilize estimation theory and learning procedures to derive recursive estimators for the system parameters. Some recent results in this area are discussed in this section. The basic system structure and conditions for convergence of the identification scheme are considered to be the important factors of the individual investigations. Only time domain techniques for identification of discrete-time systems are considered.

A relatively simple problem in identification is studied by Ho and Whalen [HO-1]. It is given that the system equations are

$$x_{k+1} = Ax_k \qquad k = 0, 1, \ldots \qquad 2.2.1$$

$$y_k = x_k + \eta_k \qquad 2.2.2$$

which represent the state and observation equations, where x_k is an n-vector, A is a constant nxn matrix to be found and η_k is a

gaussian random vector with zero mean and covariance

$$E(\eta_k \eta_j^T) = R, \quad k = j$$
$$= 0 \quad \text{otherwise}.$$

The system is stationary, linear, homogenous and all states are observed with additive white noise. The authors give a recursive algorithm for finding the A matrix that converges if

$$\det[x_k, x_{k+1}, \ldots, x_{k+n-1}] \neq 0$$

and there exist constants λ_1, $\lambda_o > 0$ such that

$$\lambda_1 > \|[x_k, \ldots, x_{k+n-1}][x_k, \ldots, x_{k+n-1}]^T\|^2 > \lambda_o > 0.$$

The identification algorithm converges to the true value of A with probability one (WP1 or A.E., [PAP]). The proof follows from an application of Dvoretzky's theorem on stochastic approximation [DV][WOL].

Another type of identification problem can be formulated as a least squares filtering problem. The solution can be derived either as a least squares or Bayesian estimate. The system equations are

$$x_{k+1} = A x_k + \xi_k \qquad 2.2.5$$

$$y_k = x_k \qquad 2.2.6$$

where ξ_k is a random n-vector with zero mean and known covariance, and A is an unknown matrix.

As shown by Mayne [MA] the problem of forming unbiased least-squares estimates of A can be solved using known results [K-2][K-3]. Specifically, the problem can be formulated as a linear state estimation problem and solved using Kalman filtering techniques. Mayne has also shown how estimates of a randomly or deterministically varying A matrix can be found, and can account for an unknown input-gain matrix B. A serious restriction to this formulation is that all states and any control inputs must be observed without noise. No proof of convergence is given by Mayne.

Another method of solution to this problem is given by Fukao using a Bayesian approach [FK-2, 3]. In his original formulation all states are observed without noise and the input noise ξ_k has an unknown mean. Fukao's proof that his algorithm converges WP1 relies on results from stochastic approximation theory. An alternate proof of his result with appropriate assumptions is given in Appendix I. By intuitively extending his algorithm to account for observation noises, Fukao has also shown how to handle this important case under stated assumptions. In fact he is able to apply the algorithm to some nonlinear systems if the assumptions can be shown to hold.

For identification of nonlinear systems with all states obscured by noise, but having no input noise process, Kirvaitis [KI] [FU-2] has indicated how stochastic approximation may be used to develop convergent algorithms.

2.3 Basic Equations and the A-Posteriori Distribution

For the results described in section 2.2, a different approach is taken for each problem. Although each of the algorithms has a similar form, their derivations, where any are given, are directed solely to the problem at hand. By adopting a Bayesian approach to system identification many of the above results may be combined into a single structure and furthermore, new results can be derived.

Consider the system of equations

$$x_{k+1} = f_k(x_k, u_k, \xi_k, \alpha) \qquad 2.3.1$$

and the observation equations

$$y_k = g_k(x_k, \eta_k^1, \beta) \qquad 2.3.2$$

$$v_k = u_k + \eta_k^2 \qquad 2.3.3$$

where each of the variables is defined in section 2.1. Identification of the system consists of finding a sequence of estimates for α and β that converge to the true values based on observations of v_k and y_k.

As indicated in section 2.1, the conditional expectation $E(\psi/y^k, v^k)$, where $\psi = (\alpha, \beta)$, is the optimal Bayes estimator for ψ based on k observations. Thus, if it is possible to obtain a recursive relation, either approximate or exact, for this estimator as a function of observations and previous estimates, then a solution to the identification problem will have been found. Of course if an approximation is used then convergence of the estimates to the true value must be established.

To obtain the desired conditional mean, the a-posteriori density of the unknown parameter vector ψ is needed. This can be found as follows.

$$p(\psi, y^{k+1}, v^{k+1}) = p(\psi, y^{k}, v^{k}, y_{k+1}, v_{k+1})$$

$$= p(y_{k+1}, v_{k+1}/\psi, y^{k}, v^{k})p(\psi/y^{k}, v^{k})p(y^{k}, v^{k})$$

Also

$$p(\psi, y^{k+1}, v^{k+1}) = p(\psi/y^{k+1}, v^{k+1})p(y^{k+1}, v^{k+1})$$

Thus

$$p(\psi/y^{k+1}, v^{k+1}) = \frac{p(y_{k+1}, v_{k+1}/\psi, y^{k}, v^{k})p(\psi/y^{k}, v^{k})p(y^{k}, v^{k})}{p(y^{k+1}, v^{k+1})}$$

Since

$$p(y^{k+1}, v^{k+1}) = p(y_{k+1}, v_{k+1}/y^{k}, v^{k})p(y^{k}, v^{k})$$

then

$$p(\psi/y^{k+1}, v^{k+1}) = \frac{p(\psi/y^{k}, v^{k})p(y_{k+1}, v_{k+1}/\psi, y^{k}, v^{k})}{p(y_{k+1}, v_{k+1}/y^{k}, v^{k})} \qquad 2.3.4$$

where

$$p(y_{k+1}, v_{k+1}/y^{k}, v^{k}) = \int_{\psi} p(\psi/y^{k}, v^{k})p(y_{k+1}, v_{k+1}/\psi, y^{k}, v^{k})d\psi .$$

Equation 2.3.4 is a statement of Bayes rule and is a recursive relation between the probability densities $p(\psi/y^{k+1}, v^{k+1})$ and $p(\psi/y^{k}, v^{k})$. The form of this relation indicates that as observations y_{k+1}, v_{k+1} become available the conditional density for ψ can be updated. Since the true value of ψ is a constant, if certain conditions

on the densities hold, then [SP] it is possible to show that the conditional density $p(\psi / y^k, v^k)$ approaches an impulse function centered at the value ψ.

The density $p(y_{k+1}, v_{k+1} / \psi, y^k, v^k)$ is required in the relation 2.3.4. This is obtained from the state and observation equations 2.3.1, 2.3.2 and 2.3.3 and from the distributions of η_k^1 and η_k^2. However, this is generally a very difficult task except in special cases.

2.4 Alternate Forms for the A-Posteriori Distribution

There are two alternatives to the direct use of equation 2.3.4. These alternatives will be developed here and shown to be of a less tenable form than equation 2.3.4.

One iterative expression for the a-posteriori density of ψ can be derived as follows [AO-1], for a simpler problem having $u_k = v_k$.

$$p(\psi, x_{k+1}, x_k, y_{k+1} / y^k) = p(\psi, x_k / y^k) p(x_{k+1}, y_{k+1} / \psi, x_k, y^k)$$

$$= p(\psi, x_k / y^k) p(y_{k+1} / \psi, x_k, y^k, x_{k+1}) p(x_{k+1} / \psi, x_k, y^k).$$

Also

$$p(\psi, x_{k+1}, y_{k+1} / y^k) = p(\psi, x_{k+1} / y^{k+1}) p(y_{k+1} / y^k)$$

$$= \int_{x_k} p(\psi, x_k, x_{k+1}, y_{k+1} / y^k) dx_k$$

Combining these relations gives

$$p(\psi, x_{k+1}/y^{k+1}) = \frac{\int_{x_k} p(y_{k+1}/\psi, x_k, y^k, x_{k+1}) p(x_{k+1}/\psi, x_k, y^k) p(\psi, x_k/y^k) dx_k}{p(y_{k+1}/y^k)} \qquad 2.4.1$$

where $p(y_{k+1}/y^k)$ equals the integral of the numerator over (ψ, x_{k+1}). Since

$$p(\psi/y^{k+1}) = \int_{x_{k+1}} p(\psi, x_{k+1}/y^{k+1})\, dx_{k+1} \qquad 2.4.2$$

the relations 2.4.1, 2.4.2 give a recursive relation for the a-posteriori distribution of ψ.

The indicated density functions in 2.4.1 simplify because, from 2.3.1 and 2.3.2 assuming ξ_k's and η_k^1's are independent,

$$p(y_{k+1}/\psi, x_k, y^k, x_{k+1}) = p(y_{k+1}/\beta, x_{k+1})$$

and from 2.3.1

$$p(x_{k+1}/\psi, x_k, y^k) = p(x_{k+1}/\alpha, x_k) .$$

Despite the apparent usefulness of 2.4.1 there is a serious drawback to its application. This is the difficulty of performing the required integration. Even with simple gaussian random variables and a linear system it is not clear that the integration can be carried out in closed form. Even if it were possible to do so, there is no reason to suspect that a recursive relation between conditional means for ψ, which are the optimal Bayes estimators, would be obtained.

Another form for the a-posteriori distribution of ψ can be obtained using 2.3.4, and noting that (again assuming $u_k = v_k$)

$$p(y_{k+1}/\psi, y^k) = \int_{x_{k+1}} p(y_{k+1}/\psi, x_{k+1}, y^k) p(x_{k+1}/\psi, y^k) dx_{k+1} \qquad 2.4.3.$$

Since

$$p(y_{k+1}/\psi, x_{k+1}, y^k) = p(y_{k+1}/\beta, x_{k+1})$$

this would appear to be a useful relation. However, establishing the density $p(x_{k+1}/\psi, y^k)$ presents in effect an optimal nonlinear filtering problem. Even in the linear case with gaussian random variables the problem would be very difficult since it requires the solution of an optimal (Kalman) filtering problem for the mean $\hat{x}_{k+1}$ as a function of ψ and all the past observations y^k. The covariance matrix is given by the solution to a Riccati equation.

To establish that the exact form of $p(y_{k+1}/\psi, x_{k+1}, y^k)$ is indeed difficult to work with and furthermore to establish a basis for approximation, the following linear example is considered. Let

$$x_{k+1} = a\, x_k + \xi_k$$

$$y_k = x_k + \eta_k$$

where all variables are scalars, while ξ_k and η_k are zero-mean, white, gaussian processes. To utilize 2.4.3, the density $p(x_{k+1}/\psi, y^k)$ is required. From well known results [K-2, 3], this density is normal with mean $\hat{x}_{k+1}$ and variance P_{k+1}, where

$$\hat{x}_{k+1} = a\,\hat{x}_k + \frac{P_k}{r}(y_k - \hat{x}_k) \qquad 2.4.4$$

and

$$P_{k+1} = 2a\,P_k + q - \frac{P_k^2}{r} \qquad 2.4.5$$

The terms q and r are the variances of ξ_k and η_k, respectively, (assuming stationarity). The observation equation yields

$$p(y_{k+1}/\psi, x_{k+1}, y^k) \sim N(x_{k+1} : r)^*.$$

Using these relations in equation 2.4.3 it can be shown that the exponent of $p(y_{k+1}/\psi, y^k)$ equals (except for the -1/2 factor)

$$\frac{1}{r + P_{k+1}} \{y_{k+1} - \hat{x}_{k+1}\}^2 .$$

Therefore

$$p(y_{k+1}/\psi, y^k) \sim N(\hat{x}_{k+1} : r + P_{k+1})$$

Using 2.4.4 to write $\hat{x}_{k+1}$ as a function of y^k gives

$$\begin{aligned}\hat{x}_{k+1} = {} & (a - P_k/r)(a - P_{k-1}/r)\ldots(a - P_o/r)\hat{x}_o \\ & + (a - P_k/r)(a - P_{k-1}/r)\ldots(a - P_1/r)\frac{P_o}{r}\,y_o \\ & + (a - P_k/r)\ldots(a - P_2/r)\frac{P_1}{r}\,y_1 \\ & \qquad \vdots \qquad\qquad \vdots \\ & + (a - P_k/r)\frac{P_{k-1}}{r}\,y_{k-1} \\ & + \frac{P_k}{r}\,y_k\end{aligned}$$

*The symbols $\sim N(a : b)$ are used to denote the normal distribution with first argument as mean, and second argument as covariance.

where the P_j's satisfy 2.4.5. This is a rather involved expression for $\hat{x}_{k+1}$, and is inconvenient for two reasons. First, to obtain a recursive relation for a from 2.3.4 would require storing all observations, which is undesirable for practical reasons. Second, the expression in the numerator of 2.3.4 must be maximized as a function of a to find the conditional mean. This would be a difficult task because of the form of $\hat{x}_{k+1}$.

A simple approximation to the density $p(y_{k+1}/\psi, y^k)$ can be obtained from 2.4.4. By taking $\hat{x}_k = y_k$, 2.4.4 becomes

$$\hat{x}_{k+1} = a\, y_k \qquad 2.4.6$$

and the variance may be approximated as

$$P_{k+1} + r = q + r + 2\, a\, P_k - P_k^2/r \leq q + r + s_k' = s_k$$

for some $s_k' > 0$. This is especially convenient because if $p(a/y^k)$ has a normal form in 2.3.4 then the product $p(a/y^k)p(y_{k+1}/a, y^k)$ is easily maximized with respect to a and $p(a/y^{k+1})$ has a normal form also. That is, the density $p(a/y^k)$ reproduces [SP] and only one observation needs to be retained.

The relation 2.4.6 is sometimes exact. If there is no system observation noise ($r \equiv 0$) then 2.4.6 is exact since $y_k = x_k$ and the variance for $p(y_{k+1}/a, y^k)$ is $s = q$. This is the case studied by Mayne [MA] and one of the cases considered by Fukao [FK-2, 3].

With these results in mind it would appear that using 2.3.4 directly is a more logical way to proceed than either of the indirect ways indicated. Equation 2.3.4 gives a very fundamental relationship between the density functions. More detailed information about the density functions can be supplied by the form of the system being considered. As was seen in the simple scalar example, some approximation may be necessary in the utilization of 2.3.4 but the effect of these approximations is more clearly seen in this relationship than in the others described. Furthermore, the algorithms which result from this relation are quite simple and can be realized by conventional computing techniques.

2.5 Summary

Under the general structure of Bayes estimation it has been shown that a general class of identification problems can be formulated. In examining previous results it was seen that there is little uniformity of approach. Ho and Whalen's result for the linear stationary autonomous system is derived ad-hoc and the convergence follows from stochastic approximation. Mayne's work is a reformulation of linear filtering and does not allow any observation noises. Fukao's results apply only to stationary systems with no input observation noise. Further, no derivation of results is given or motivated and only a single restrictive convergence theorem is given. A single point of reference in attacking these problems is

provided by Bayes' rule, which also allows computationally feasible algorithms to be developed. Because of the difficulty in forming the exact density functions required by Bayes' rule, an approximation was seen to be desirable.

III SOLUTION OF THE BAYESIAN IDENTIFICATION PROBLEM

The general form of the problem to be solved has been given in Chap. II. In Sec. 3.1 an iterative equation will be derived that relates the estimates at time k-1 to the kth estimate and observations. The conditions under which this algorithm converges to the true value of the parameters and the sense of convergence are given as convergence theorems I, II and III in Sec. 3.2. This result is related to those given by other authors in Sec. 3.3.

3.1 Derivation of the identification Algorithm

In this section, the identification algorithm for a class of systems is derived using 2.3.4. In particular, for the system of Figure 3.1.1 the equations are now assumed to be of the form

$$x_{k+1} = f_k(x_k, u_k, \xi_k, a) = D_k(x_k, u_k)a + \xi_k \qquad 3.1.1$$

$$y_k = x_k + \eta_k^1 \qquad 3.1.2$$

$$v_k = u_k + \eta_k^2 \qquad 3.1.3$$

where $D_k(.)$ is an n x p matrix, a is a fixed but unknown p-vector

and

$$E(\xi_k) = E(\eta_k^1) = 0, \quad E(\eta_k^2) = 0 \qquad 3.1.4a$$

$$E(\xi_k \xi_j^T) = Q_k, \quad k = j \qquad 3.1.4b$$

$$= 0 \ , \quad k \neq j \qquad 3.1.4c$$

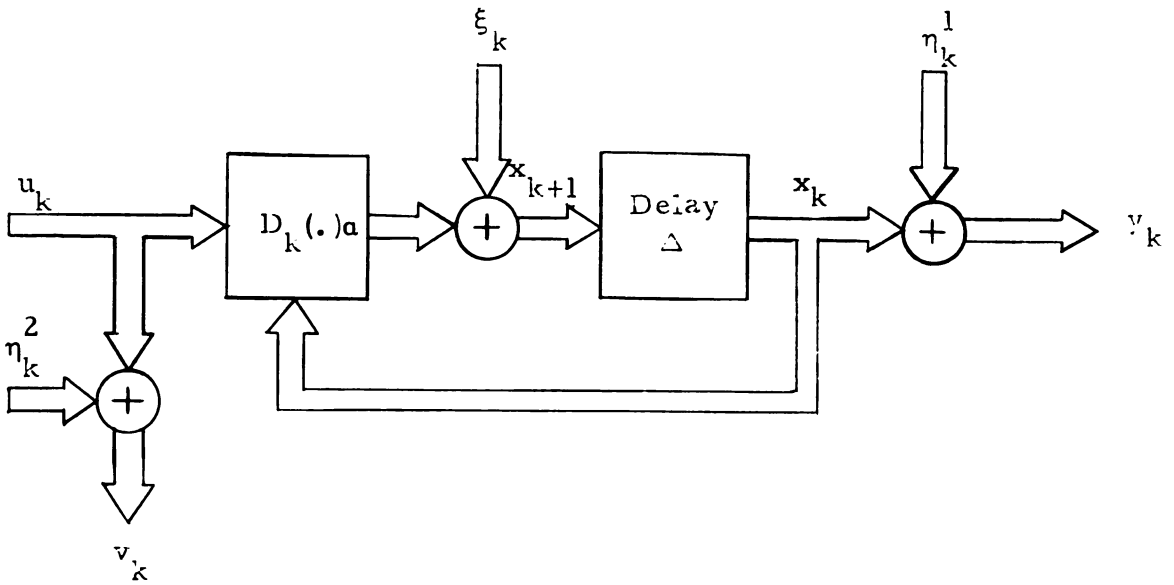

Figure 3.1.1 General System Configuration

$$E(\eta_k^1 [\eta_j^2]^T) = 0 \qquad \text{for all } k, j \qquad 3.1.4d$$

$$E(\eta_k^i [\eta_j^i]^T) = R_k^i \qquad \text{for } k = j \ ; \ i = 1, 2 \qquad 3.1.4e$$

$$= 0 \qquad \text{for } k \neq j \ ; \ i = 1, 2 \qquad 3.1.4f$$

and $E(\xi_k [\eta_j^i]^T) = 0$ for all k, j ; i = 1, 2 3.1.4g

Since β is known in 3.1.2, 2.3.4 becomes

$$p(a/y^{k+1}, v^{k+1}) = \frac{p(a/y^k, v^k) p(y_{k+1}, v_{k+1}/a, y^k, v^k)}{p(y_{k+1}, v_{k+1}/y^k, v^k)} \qquad 3.1.5$$

In 3.1.1 - 3.1.4 assuming v_{k+1} and y_{k+1} are independent and that v_{k+1} is independent of a, v^k and y^k gives

$$p(y_{k+1}, v_{k+1}/a, y^k, v^k) = p(y_{k+1}/a, y^k, v^k)p(v_{k+1}) .$$

Thus $p(v_{k+1})$ can be canceled in the numerator and denominator of 3.1.5. This gives the recurrence relation

$$p(a/y^{k+1}, v^{k+1}) = \frac{p(a/y^k, v^k)p(y_{k+1}/a, y^k, v^k)}{p(y_{k+1}/y^k, v^k)} \qquad 3.1.6$$

where the denominator equals the integral of the numerator over a.

The vector a is assumed to have a gaussian a-priori distribution, with density

$$p(a/y^o, v^o) = p_o(a) = \frac{e^{-\frac{1}{2}(a-\hat{a}_o)^T P_o^{-1}(a-\hat{a}_o)}}{\sqrt{(2\pi)^p \det(P_o)}}$$

where det(.) denotes determinant.

$$p(a/y^o, v^o) \sim N(\hat{a}_o ; P_o)$$

The form of the density function $p_o(a)$ and the mean and covariance matrices are assumed to be given and reflect the a-priori uncertainty concerning the parameter vector a. If the a-priori information concerning a is small, then the P_o matrix is large in norm (c.f. Sec. 3.2). If the true value of a is known with some certainty then the matrix P_o is small in norm. The initial mean

value $\hat{a}_o$ is taken to be the best available estimate of the true value of a.

The density function $p(y_{k+1}/a, y^k, v^k)$ is now needed. From 3.1.1, 3.1.2, 3.1.3

$$\begin{aligned} y_{k+1} &= x_{k+1} + \eta^1_{k+1} \\ &= D_k(x_k, u_k)\,a + \xi_k + \eta^1_{k+1} \\ &= D_k(y_k - \eta^1_k,\ v_k - \eta^2_k)\,a + \xi_k + \eta^1_{k+1} . \end{aligned}$$

This equation gives a basis for approximating the function $p(y_{k+1}/a, y^k, v^k)$ as discussed in Sec. 2.4. Taking

$$E(y_{k+1}/a, y^k, v^k) \cong D_k(y_k, v_k)\cdot a$$

and

$$\operatorname{cov}\{y_{k+1}/a, y^k, v^k\} = S_k$$

where S_k is larger (in norm) than $Q_k + R_k$, the density is approximated as normal. Thus

$$p(y_{k+1}/a, y^k, v^k) \sim N(D_k(y_k, v_k)\,a\ ;\ S_k) . \qquad 3.1.7$$

At the k th step suppose

$$p(a/y^k, v^k) \sim N(a_k\ ;\ P_k) . \qquad 3.1.8$$

Since $p_o(a)$ is normal, if a recursive relation between $p(a/y^k, v^k)$ and $p(a/y^{k+1}, v^{k+1})$ is obtained having the a-posteriori as normal then a complete algorithm will be established. The a-posteriori mean and covariance will define the algorithm, with the mean being taken as the estimator of a.

Using 3.1.7 and 3.1.8 in 3.1.6 gives

$$p(a/y^{k+1}, v^{k+1}) =$$

$$= \frac{k \exp -\frac{1}{2}\{(a-\hat{a}_k)^T P_k^{-1}(a-\hat{a}_k)+(y_{k+1} - D_k(y_k, v_k)a)^T S_k^{-1}(y_{k+1} - D_k(y_k, v_k)a)\}}{\int_a [\text{numerator}]\, da} \qquad 3.1.9$$

where k is a constant. Since it is known that the exponential family reproduces [SP], the density $p(a/y^{k+1}, v^{k+1})$ is normal. Thus it is only necessary to find the maximum value of the numerator to find the conditional mean. The exponent of the numerator (except for the constant - 1/2) equals

$$\begin{aligned} & a^T\{P_k^{-1} + D_k^T(y_k, v_k)S_k^{-1}D_k(y_k, v_k)\}a \\ & - 2a^T\{P_k^{-1}\hat{a}_k + D_k^T(y_k, v_k)S_k^{-1}y_{k+1}\} \\ & + \{\hat{a}_k^T P_k^{-1}\hat{a}_k + y_{k+1}^T S_k^{-1}y_{k+1}\} . \end{aligned} \qquad 3.1.10$$

To maximize the numerator of 3.1.9, the quadratic form 3.1.10 must be minimized. This is accomplished by taking a, and thus the a-posteriori mean $\hat{a}_{k+1}$, to be

$$\hat{a}_{k+1} = P_{k+1}\{P_k^T\hat{a}_k + D_k^T(y_k, v_k)S_k^{-1}y_{k+1}\} \qquad 3.1.11$$

where

$$P_{k+1}^{-1} = P_k^{-1} + D_k^T(y_k, v_k)S_k^{-1}D_k(y_k, v_k) \qquad 3.1.12$$

The matrix P_{k+1} is the a-posteriori covariance matrix associated with $p(a/y^{k+1}, v^{k+1})$. The details of the argument are found in Appendix II.

Equation 3.1.11 and 3.1.12 can be written in a more convenient form

$$\hat{a}_{k+1} = P_{k+1}P_k^{-1}\hat{a}_k + P_{k+1}D_k^T(y_k, v_k)S_k^{-1}y_{k+1}$$

$$= \hat{a}_k + P_{k+1}(P_k^{-1} - P_{k+1}^{-1})\hat{a}_k + P_{k+1}D_k^T(y_k, v_k)S_k^{-1}y_{k+1}$$

From 3.1.12

$$P_k^{-1} - P_{k+1}^{-1} = -D_k^T(y_k, v_k)S_k^{-1}D_k(y_k, v_k)$$

so that

$$\hat{a}_{k+1} = \hat{a}_k + P_{k+1}\{-D_k^T(y_k, v_k)S_k^{-1}D_k(y_k, v_k)\}\hat{a}_k + P_{k+1}D_k^T(y_k, v_k)S_k^{-1}y_{k+1}$$

$$= \hat{a}_k + P_{k+1}D_k^T(y_k, v_k)S_k^{-1}[y_{k+1} - D_k(y_k, v_k)\hat{a}_k] \qquad 3.1.13$$

Using a matrix inversion lemma [AO-1][HOU], P_{k+1} can be written as, from 3.1.12

$$P_{k+1} = P_k - P_kD_k^T(y_k, v_k)[S_k + D_k(y_k, v_k)P_kD_k^T(y_k, v_k)]^{-1}D_k(y_k, v_k)P_k \qquad 3.1.14$$

Equations 3.1.3 and 3.1.14 represent the identification algorithm for the system 3.1.1, 3.1.2, 3.1.3. This algorithm gives a Bayes-optimal sequence of estimates for the vector a whenever the system is linear and no observation noises are present. Otherwise, the estimates are sub-optimal. The form of 3.1.13 is of interest and has intuitive appeal. From 3.1.1, the term $D_k(y_k, v_k)a_k$ can be interpreted as a prediction of the next observation y_{k+1}, based on the present observations y_k and v_k and on

the present parameter estimate $\hat{a}_k$. Therefore the difference $(y_{k+1} - D_k(y_k, v_k)\hat{a}_k)$ represents an error term, and the present estimate $\hat{a}_k$ is updated by an amount proportional to this error. The algorithm 3.1.13, 3.1.14 is actually a generalized stochastic approximation algorithm [DV][BL][KIE][CHU][ROB] with a gain sequence that depends upon the observations rather than being specified a-priori.

3.2 Convergence Theorems and Assumptions

Convergence of the general algorithm 3.1.13, 3.1.14 will be considered here. Conditions under which this algorithm converges to the true value a, and in what senses will be given. The proofs for the three theorems to be stated are found in Appendix I.

Equation 3.1.13 can be written in error form by defining

$$\epsilon_k = \hat{a}_k - a$$

and writing

$$\hat{a}_{k+1} - a = \hat{a}_k - a + P_{k+1} D_k^T S_k^{-1} y_{k+1} - P_{k+1} D_k^T S_k^{-1} D_k \hat{a}_k$$

or

$$\epsilon_{k+1} = \epsilon_k + P_{k+1} D_k^T S_k^{-1} y_{k+1} - P_{k+1} D_k^T S_k^{-1} D_k \epsilon_k - P_{k+1} D_k^T S_k^{-1} D_k a$$

$$\epsilon_{k+1} = (I - P_{k+1} D_k^T S_k^{-1} D_k)\epsilon_k + P_{k+1} D_k^T S_k^{-1} (y_{k+1} - D_k a)$$

where D_k is assumed to represent $D_k(y_k, v_k)$ for convenience. From 3.1.12

$$I = P_{k+1}P_k^{-1} + P_{k+1}D_k^T S_k^{-1} D_k .$$

Substituting this into the equation for ϵ_{k+1}

$$\epsilon_{k+1} = P_{k+1}P_k^{-1}\epsilon_k + P_{k+1}D_k^T S_k^{-1}(y_{k+1} - D_k a) . \qquad 3.2.1$$

Equation 3.2.1 relates the estimation error at the $(k+1)^{st}$ step to the previous error and the true value of the parameter a.

Conditions under which ϵ_k approaches zero and in what sense are now given. The concepts of system identifiability and strong identifiability are introduced. These concepts are fundamental to the convergence theorems, and indicate whether sufficient observational information about the system is available. Other assumptions required for convergence are also stated.

The norm of a vector X and a square matrix A as used in this discussion are defined by

$$\| X\|^2 = \sum_i^n |x_k|^2$$

and

$$\| A\| = \lambda_{max}(A^* A)$$

where (*) denctes conjugate transpose and $\lambda_{max}(.)$ denotes the largest eigenvalue of the matrix argument.

Definition 3.2.1: The system 3.1.1, 3.1.2, 3.1.3 is said to be identifiable if there exists a finite positive integer q such that the matrix sum

$$\sum_{j=k}^{k+q} D_j^T S_j^{-1} D_j$$

is positive definite with probability one (WP1) for any k.

Assumption 3.2.A1 : $\|\hat{a}_o - a\| < \infty$ and $0 < \|P_o\| < \infty$.

Assumption 3.2.A2 : $E\{(y_{j+1} - D_j a)/y^j, v^j\} = 0$

Assumption 3.2.A3 : $\|E\{ \sum_{j=k-q}^{k} D_j^T S_j^{-1} (y_{j+1} - D_j a) \sum_{j=k-q}^{k} (y_{j+1} - D_j a)^T S_j^{-1}\}\| \leq M < \infty$

Convergence Theorem I: If the system 3.1.1, 3.1.2, 3.1.3 is identifiable and assumptions 3.2.A1, 3.2.A2, 3.2.A3 hold then the algorithm 3.1.13, 3.1.14 converges to the true value of the parameter vector a in the sense that

$$\|\mathrm{cov}(\hat{a}_k - a)\| = \|\mathrm{cov}\, \epsilon_k\| \to 0 \quad \text{as } k \to \infty .$$

i.e., the norm of the error covariance matrix converges to zero.

Assumption 3.2.A4:

$$E\|P_{k+1} P_{k-q}^{-1} \epsilon_{k-q} + P_{k+1} \sum_{j=k-q}^{k} D_j^T S_j^{-1} (y_{j+1} - D_j a)\|^2$$

$$\leq E\|P_{k+1} P_{k-q}^{-1} \epsilon_{k-q}\|^2 + E\|P_{k+1} \sum_{j=k-q}^{k} D_j^T S_j^{-1} (y_{j+1} - D_j a)\|^2$$

Assumption 3.2.A5:

$$E\| \sum_{j=k-q}^{k} D_j^T (y_{j+1} - D_j a)\|^2 \leq M < \infty .$$

Convergence Theorem II: If the system 3.1.1, 3.1.2, 3.1.3 is identifiable and assumptions 3.2.A1, 3.2.A4, 3.2.A5 hold then the algorithm 3.1.13, 3.1.14 converges to the true value of the parameter vector a in the sense that

$$E\{\|\epsilon_k\|^2\} = E\{\|\hat{a}_k - a\|^2\} \to 0 \text{ as } k \to \infty.$$

i.e., the error norm converges to zero in mean square [PAP].

Definition 3.2.2: The system 3.1.1, 3.1.2, 3.1.3 is said to be strongly identifiable if the minimum eigenvalue of $\sum_{j=0}^{k} D_j^T S_j^{-1} D_j$ satisfies

$$\lambda_{min}\{\sum_{j=0}^{k} D_j^T S_j^{-1} D_j\} \geq c\, k^p$$

with probability one (WP1) where $1 < p < \infty$ and $0 < c < \infty$ are constants.

Assumption 3.2.A6: $\|\sum_{j=k-q}^{k} D_j^T S_j^{-1}(y_{j+1} - D_j a)\| \leq M < \infty$ with probability one (WP1), where M is a constant.

Convergence Theorem III: If the system 3.1.1, 3.1.2, 3.1.3 is strongly identifiable and if assumptions 3.2.A1, 3.2.A6 hold then the algorithm 3.1.13, 3.1.14 converges to the true value of the parameter vector a in the sense that

$$\|\epsilon_k\| \equiv \|\hat{a}_k - a\| \to 0 \text{ as } k \to \infty$$

with probability one (WP1).

This theorem is equivalent to one given by Fukao [FK-3], but the proof given in Appendix I is distinct.

Assumption 3.2.A1 is common to the three converg ence theorems and is very reasonable. Assumptions 3.2.A3, 3.2.A5 and 3.2.A6 are constraints on the magnitude of the random variables and their moments, and are easily satisfied. Assumption 3.2.A6 does not hold for gaussian random variables. This is not a stringent condition because in practical situations all variables are bounded.

Assumption 3.2.A2, as required by Theorem I, is true for linear systems, but is not true in general for nonlinear systems. Theorem II is similar to Theorem I. In the former, assumption 3.2.A4 replaces 3.2.A2 and allows convergence in mean square of the error norm as opposed to convergence of the error covariance norm. Theorem II applies to linear systems, but in general will apply to nonlinear systems only if 3.2.A4 can be verified.

The condition of identifiability is required by both Theorem I and Theorem II. Conditions under which a system is identifiable are discussed in Chapter V. The strong identifiability of the system as required by Theorem III is sufficient for identifiability, as shown in Appendix I. Theorem III may be applied to nonlinear systems.

3.3 Derivation of Previous Results as Special Cases

The form of the system equations 3.1.1, 3.1.2, 3.1.3 are sufficiently general to permit derivation of known results as special cases.

The problem studied by Mayne [MA] is considered first. Equation 2.2.5 may be rewritten as

$$x_{k+1} = X_k^T a + \xi_k$$

and

$$y_k = x_k$$

where

$$X_k^T = \begin{bmatrix} x_k^T & & 0 \\ & x_k^T & \\ & & \ddots \\ 0 & & x_k^T \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \qquad 3.3.1$$

and a_i^T is the ith row of the matrix A. In the algorithm 3.1.13, 3.1.14, the $D_k(.)$ matrix is taken as

$$D_k(y_k, v_k) = D_k(y_k) = D(x_k) = X_k^T \qquad 3.3.2$$

Since there are no observation noises, $p(y_{k+1}/a, y^k, v^k)$ need not be approximated since

$$p(y_{k+1}/a, y^k, v^k) = p(y_{k+1}/a, x^k)$$

and

$$y_{k+1} = x_{k+1} = A x_k + \xi_k = X_k^T a + \xi_k$$

Therefore with $S_k = Q_k$

$$p(y_{k+1}/a, x^k) \sim N(X_k^T a \; ; \; Q_k)$$

where Q_k is the covariance matrix of the gaussian random variable ξ_k. The definitions for a and D_k in 3.3.1 and 3.3.2 are then used with the algorithm 3.1.13, 3.1.14 to solve the problem. If the matrix A varies with k then D_k is changed at each step accordingly. If inputs are also applied then $D_k(.)$ must include these noise-free observations.

The results of Fukao [FK-2, 3], for a linear system with no observation noise, are the same as those given by Mayne and can be derived accordingly. When observation noises are present at the output, then $u_k = v_k$ and D_k is taken to be

$$D_k(y_k, v_k) = D(y_k, u_k) = \begin{bmatrix} y_k^T & & 0 & u_k^T & & 0 \\ & \ddots & & & \ddots & \\ 0 & & y_k^T & 0 & & u_k^T \end{bmatrix}$$

The nonlinear problems studied by Fukao are also included in the formulation 3.1.1, 3.1.2, 3.1.3. For his formulation there is no observation noise on the inputs so that $v_k = u_k$. Also, time-varying systems are not considered so that the $D_k(.)$ matrix remains fixed,

$$D_k(y_k, v_k) = D(y_k, u_k) .$$

The identification of the system studied by Ho and Whalen [HO-1] is easily handled. There is no input noise, so Q_k, the covariance of ξ_k, is identically zero. There are no inputs so

$u_k = v_k = 0$, and the matrix A is fixed. All states are observed. Therefore in 3.1.13, 3.1.14

$$D_k(y_k, v_k) = D(y_k) = \begin{bmatrix} y_k^T & & 0 \\ & \ddots & \\ 0 & & y_k^T \end{bmatrix}$$

and this type of system can be identified accordingly.

Kirvaitis [KI][FU-2] considers identification of nonlinear systems with all states obscured by noise and with no state disturbances. If the system is taken to be time-discrete then 3.1.1, 3.1.2, 3.1.3 are sufficiently general to include this case, taking $\xi_k \equiv 0$, and $v_k \equiv u_k$.

3.4 Summary

An identification algorithm 3.1.13, 3.1.14 has been derived for the general system 3.1.1, 3.1.2, 3.1.3. This was accomplished using a reproducing gaussian distribution for the parameters which specify the system. The convergence of this algorithm to the true parameter values is established by convergence Theorems I, II and III in Sec. 3.2. Theorems I and II are new results in identification, and the algorithm 3.1.13, 3.1.14 is more general than any given previously. The algorithm can be specialized to give results of other authors as indicated in Sec. 3.3.

IV SPECIAL FORMS OF THE ALGORITHM

System equations of the form

$$x_{k+1} = D_k(x_k, u_k)a + \xi_k$$

$$y_k = x_k + \eta_k^1$$

$$v_k = u_k + \eta_k^2$$

have been considered in Sec. 3.1. The identification algorithm for these systems has the form 3.1.13, 3.1.14. For a given system structure, implementation of the equations requires that the $D_k(y_k, v_k)$ matrix be found. Special forms of $D_k(.)$ for some important classes of systems are given in this chapter. The convergence theorems of Sec. 3.2 are discussed in relation to each system. The formulations for linear, stationary, and time-varying systems are given in Sec. 4.1 and 4.2, respectively. Nonlinear stationary, and time-varying system formulations are given in Sec. 4.3 and 4.4. Section 4.5 discusses the formulation

for systems with generated inputs, for systems with state disturbances having unknown mean, and systems with indirectly observed states.

4.1 Linear Stationary Systems

When all states are observed the equations for a linear, stationary system have the form

$$x_{k+1} = A x_k + B u_k + \xi_k$$

$$y_k = x_k + \eta_k^1$$

$$v_k = u_k + \eta_k^2$$

where A is an n x n matrix with rows a_i^T and B is an n x r matrix with rows b_i^T. The entries of the matrices A and B are to be identified. The state equations may be written as

$$x_{k+1} = [A \quad B]\begin{bmatrix} x_k \\ u_k \end{bmatrix} + \xi_k$$

$$= \begin{bmatrix} a_1^T & b_1^T \\ \vdots & \\ a_n^T & b_n^T \end{bmatrix}\begin{bmatrix} x_k \\ u_k \end{bmatrix} + \xi_k \quad .$$

Rearranging into the form 3.1.1

$$x_{k+1} = \left[\begin{array}{ccc|ccc} x_k^T & & 0 & u_k^T & & 0 \\ & x_k^T & & & u_k^T & \\ & & \ddots & & & \ddots \\ 0 & & x_k^T & 0 & & u_k^T \end{array}\right] \left[\begin{array}{c} a_1 \\ a_2 \\ \vdots \\ a_n \\ b_1 \\ \vdots \\ b_n \end{array}\right] + \xi_k \qquad 4.1.1$$

$$x_{k+1} = [X_k^T U_k^T]\,a + \xi_k = D(x_k, u_k)a + \xi_k \qquad 4.1.2$$

Thus for the linear, stationary system, the $D_k(.)$ matrix is defined by 4.1.2.

Conditions under which such a system is identifiable can now be given. For $\sum_{j=1}^{q} D_j^T S_j^{-1} D_j$ to be positive definite (WP1)

$$x^T(\sum_j D_j^T S_j^{-1} D_j)x = \sum_j x^T D_j^T S_j^{-1} D_j x = \sum_j (D_j x)^T S_j^{-1} (D_j x) > 0$$

for all $x \neq 0$, where D_j denotes $D_j(y_k, v_j)$. Since S_j^{-1} is positive definite, no $x \neq 0$ can exist for which

$$D_j x = 0 \qquad j = 1, 2, \ldots q$$

or there exists no $x \neq 0$ such that

$$\mathfrak{D}\, x = \left[\begin{array}{c} D_1 \\ \vdots \\ D_q \end{array}\right] x = 0$$

where $\mathfrak{D}$ has nq rows and n(n+r) columns. This equation has only

the trivial solution $x = 0$ if the rank of $\mathfrak{D}$ is equal to $n(n+r)$ so that $q \geq n+r$ is required. The $\mathfrak{D}$ matrix has the form

$$\mathfrak{D}^T = \left[\begin{array}{cc|cc|c|cc} y_1 & 0 & y_2 & 0 & & y_q & 0 \\ & y_1 & & & & & \\ 0 & & 0 & y_2 & \cdots & 0 & y_q \\ v_1 & 0 & v_2 & 0 & & v_q & 0 \\ & & & & & & \\ 0 & v_1 & 0 & v_2 & & 0 & v_q \end{array}\right]$$

so that the rows of $\mathfrak{D}$ (columns of $\mathfrak{D}^T$) are linearly independent if any $(n+r)$ of the pairs $[y_i^T v_i^T]$, $i = 1, 2, \ldots q$, are linearly independent. This is satisfied (WP1) if both y_i and v_i contain additive noise components. The noise process ξ_k guarantees (WP1) the linear independence of the y_i vectors, which is sufficient for the linear independence.

To apply convergence Theorem I of Sec. 3.2 to linear stationary systems, assumptions 3.2.A1, 3.2.A2, 3.2.A3 must also be satisfied. Assumptions A1 and A3 are very reasonable. Assumption A2 is satisfied since, with all random variables having zero mean

$$\begin{aligned} E(y_{j+1} - D_j a) &= E(y_{j+1} - Ay_j - Bv_j) \\ &= E(x_{j+1} - Ax_j - Bu_j) + E(\eta^1_{j+1} - A\eta^1_j - B\eta^2_j) \\ &= E(\xi_j) + E(\eta^1_{j+1} - A\eta^1_j - B\eta^2_j) \\ &= 0 \end{aligned}$$

Therefore Convergence Theorem I applies to the linear stationary case and the algorithm 3.1.13, 3.1.14 converges accordingly. Convergence Theorem II may also be applied if assumptions 3.2.A4 and 3.2.A5 hold.

4.2 Linear, Time-Varying Systems

When all states are observed the equations for a linear, time-varying system have the form

$$x_{k+1} = A_k x_k + B_k u_k + \xi_k$$

$$y_k = x_k + \eta_k^1$$

$$v_k = u_k + \eta_k^2$$

Using

$$A_k = \begin{bmatrix} a_1^T(k) \\ a_2^T(k) \\ \vdots \\ a_n^T(k) \end{bmatrix}, \quad B_k = \begin{bmatrix} b_1^T(k) \\ b_2^T(k) \\ \vdots \\ b_n^T(k) \end{bmatrix}, \quad \gamma(k), \begin{bmatrix} a_1(k) \\ a_2(k) \\ \vdots \\ a_n(k) \\ b_1(k) \\ \vdots \\ b_n(k) \end{bmatrix}$$

where $a_i^T(k)$ and $b_i^T(k)$ are the rows of A_k and B_k respectively, the variation of $\gamma(k)$, and thus the variation of A_k and B_k, is assumed to be governed by

$$\gamma(k+1) = \mathcal{A}\, Z(k+1) \qquad 4.2.1$$

$$Z(k+1) = G\, Z(k) \qquad 4.2.2$$

The matrix $\mathcal{A}$ is unknown, $Z(k)$ is a q-vector, and $\gamma(k)$ is an $n(n+r)$-vector. The matrix G and the initial value Z_o are known.

From 4.2.1 and 4.2.2

$$\gamma(k+1) = \mathcal{A}Z(k+1) = \mathcal{A}G\, Z(k) = \mathcal{A}G^{k+1}Z_o \,.$$

Then

$$A_k x_k + B_k u_k = \left[\begin{array}{ccc:ccc} x_k^T & & 0 & u_k^T & & 0 \\ & \ddots & & & \ddots & \\ 0 & & x_k^T & 0 & & u_k^T \end{array}\right] \gamma(k)$$

$$= [X_k^T U_k^T]\mathcal{A}\, G^k Z_o \,.$$

The vector $\mathcal{A}G^k Z_o$ can be written

$$\mathcal{A}\, G^k Z_o = \mathcal{A}\, h_k = H_k^T a'$$

where

$$a'^T = [a_1'^T \ldots a_p'^T],\ \mathcal{A}^T = [a_1' a_2' \ldots a_p'] \qquad 4.2.3$$

and

$$H_k^T = \left[\begin{array}{ccc} h_k^T & & 0 \\ & h_k^T & \\ & & \ddots \\ 0 & & h_k^T \end{array}\right],\ h_k = G^k Z_o = G \cdot G^{k-1} Z_o \qquad 4.2.4$$

Therefore

$$A x_k + B u_k = [X_k^T U_k^T] H_k^T a' = D_k(x_k, u_k) \cdot a \qquad 4.2.5$$

where a' generally contains both known components and unknown components a which are to be determined. Equations 4.2.3, 4.2.4, 4.2.5 define $D_k(x_k, u_k)$.

Example 4.2.1. The use of 4.2.3, 4.2.4, and 4.2.5, is studied in this example of a time-varying linear system of second order.

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} a_o + a_1 \cdot k & a_2 \\ a_3 k^2 & a_4 k \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 k \end{bmatrix} u_k + \begin{bmatrix} \xi_{1k} \\ \xi_{2k} \end{bmatrix}$$

The a_i and b_i are unknown parameters. The vector γ is then given by

$$\gamma^T(k) = [a_o + a_1 k,\ a_2,\ a_3 k^2,\ a_4 k,\ b_1,\ b_2 k]$$

or

$$\gamma^T(k) = [a_o + a_1 k^{(1)},\ a_2,\ a_3 k^{(2)} + a_3 k^{(1)},\ a_4 k^{(1)},\ b_1,\ b_2 k^{(1)}]$$

where the superscript $(.)$ denotes a factorial polynomial and is defined by [HAM]

$$x^{(n)} = x(x-1)(x-2)\ldots(x-n+1),\quad n \geq 1$$

and $x^{(0)} = 1$. The terms in the second expression for $\gamma^T(k)$ are obtained from Sterling numbers of the second kind.

Equation 4.2.1 is written as

$$\gamma(k) = \mathcal{A} Z(k) = \begin{bmatrix} 0 & a_1 & a_o \\ 0 & 0 & a_2 \\ a_3 & a_3 & 0 \\ 0 & a_4 & 0 \\ 0 & 0 & b_1 \\ 0 & b_2 & 0 \end{bmatrix} \begin{bmatrix} k^{(2)} \\ k^{(1)} \\ k^{(0)} \end{bmatrix}$$

and 4.2.2 is written as

$$Z(k+1) = G Z(k)$$

$$\begin{bmatrix} (k+1)^{(2)} \\ (k+1)^{(1)} \\ (k+1)^{(0)} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} k^{(2)} \\ k^{(1)} \\ k^{(0)} \end{bmatrix}$$

and $Z_o^T = [0 \ 0 \ 1]$. These relations are used with 4.2.3, 4.2.4, and 4.2.5 to determine $D_k(x_k, u_k)$.

From the $\mathcal{A}$ matrix, the a' vector to be found has the form

$$a'^T = [0 a_1 a_o 0 0 a_2 a_3 a_3 0 0 a_4 0 0 0 b_1 0 b_2 0] .$$

Since some of the components of a' are known, further reduction of the problem is possible. For some k,

$$G^k Z_o = W = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}$$

Then the product $H_k{}^T a'$, where $H_k{}^T$ is defined by 4.2.4 can be written as, eliminating the zeros of a'

$$H_k{}^T a' = \begin{bmatrix} W^T & & 0 \\ & W^T & \\ & & \ddots \\ 0 & & W^T \end{bmatrix} \begin{bmatrix} 0 \\ a_1 \\ a_o \\ 0 \\ \vdots \\ 0 \\ b_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \omega_2\omega_3 & & & & 0 \\ & \omega_3 & & & \\ & & \omega_1\omega_2 & & \\ & & & \omega_2 & \\ & & & & \omega_3 \\ 0 & & & & \\ & & & & & \omega_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_o \\ a_2 \\ a_3 \\ a_3 \\ a_4 \\ b_1 \\ b_2 \end{bmatrix}$$

Re-ordering the column vector and eliminating the redundant a_3 gives

$$H_k{}^T a' = \begin{bmatrix} \omega_3\omega_2 & & & & 0 \\ & \omega_3 & & & \\ & & (\omega_1 + \omega_2) & & \\ & & & \omega_2 & \\ & & & & \omega_3 \\ 0 & & & & \\ & & & & & \omega_2 \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ b_1 \\ b_2 \end{bmatrix}$$

Multiplying $H_k{}^T$ on the left by $[X_k{}^T U_k{}^T]$ gives

$$[X_k^T U_k^T] H_k^T a' = \begin{bmatrix} x_k & y_k & 0 & 0 & u_k & 0 \\ 0 & 0 & x_k & y_k & 0 & u_k \end{bmatrix} \begin{bmatrix} \omega_3\omega_2 & & & & 0 & \\ & \omega_3 & & & & \\ & & (\omega_1+\omega_2) & & & \\ & & & \omega_2 & & \\ & 0 & & & \omega_3 & \\ & & & & & \omega_2 \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ b_1 \\ b_2 \end{bmatrix}$$

so that, performing the indicated multiplication,

$$D_k(x_k, u_k)\cdot a = \begin{bmatrix} x_k\omega_3 & x_k\omega_2 & y_k\omega_3 & 0 & 0 & u_k\omega_3 & 0 \\ 0 & 0 & 0 & x_k(\omega_1+\omega_2) & y_k\omega_2 & 0 & u_k\omega_2 \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ b_1 \\ b_2 \end{bmatrix}$$

When the vector a is found, $\gamma(k)$ can be calculated, and from this, A_k and B_k.

Assumptions 3.2.A1 and 3.2.A3 as required by convergence Theorem I of Sec. 3.2 are easily satisfied. Identifiability will be satisfied generally. An indication of why this is true is given by the example of this section. Assumption A2 is also satisfied so that the algorithm will converge. Assumptions A4 and A5 may also hold so that the algorithm would also converge in the sense of Theorem II.

4.3 Stationary Nonlinear Systems

The system equations for a stationary nonlinear system with all states observed are

$$x_{k+1} = f(x_k, u_k, a) + \xi_k$$

$$y_k = x_k + \eta_k^1$$

$$v_k = u_k + \eta_k^2$$

where f(.) is a vector valued function. Its components satisfy

$$f_i(x_k, u_k, a) = [f_{i1}(x_k, u_k), f_{i2}(x_k, u_k), \ldots, f_{ip}(x_k, u_k)] \cdot a$$

for i = 1, 2, ...n. Each of the n·p functions $f_{ij}(.)$ is assumed known. Therefore, the required $D_k(.)$ matrix is

$$D_k(x_k, u_k) = \begin{bmatrix} f_{11}(x_k, u_k) & \cdots & f_{1p}(x_k, u_k) \\ \vdots & & \vdots \\ f_{n1}(x_k, u_k) & \cdots & f_{np}(x_k, u_k) \end{bmatrix} \qquad 4.3.1$$

Example 4.3.1: To illustrate, the third order nonlinear stationary system

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} a_o + a_1 x_k + a_2 x_k y_k \\ a_3 x_k^2 + a_4 y_4 \\ a_5 \sin z_k \end{bmatrix} + \begin{bmatrix} 0 \\ b_1 \\ b_2 \end{bmatrix} u_k + \begin{bmatrix} \xi_{1k} \\ \xi_{2k} \\ \xi_{3k} \end{bmatrix}$$

is considered where all variables are scalars.

This may be written as

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & x_k & x_k y_k & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_k^2 & y_k & 0 & u_k & 0 \\ 0 & 0 & 0 & 0 & 0 & \sin z_k & 0 & u_k \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \xi_{1k} \\ \xi_{2k} \\ \xi_{3k} \end{bmatrix}$$

or

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = D(x_k, y_k, z_k, u_k) \cdot a + \xi_k$$

Convergence Theorem I is not generally applicable to nonlinear systems since assumption A2 may not hold. If assumption A4 can be justified, then Theorem II can be applied. Generally, however, Theorem III will be required for nonlinear systems. Assumptions A1 and A6 hold easily. Identifiability generally holds, as can be conjectured from the form of $D_k(.)$ in the preceding example. Assuming strong identifability gives convergence in the sense of Theorem III.

4.4 Time-Varying Nonlinear Systems

The system equations for time-varying nonlinear systems with all states observed have the general form

$$x_{k+1} = f_k(x_k, u_k, a) + \xi_k$$

$$y_k = x_k + \eta_k^1$$

$$v_k = u_k + \eta_k^2 \quad .$$

By analogy with the stationary nonlinear case, the $D_k(.)$ matrix has the form

$$D_k(x_k, u_k) = \begin{bmatrix} f_{11}(k, x_k, u_k) & \cdots & f_{1p}(k, x_k, u_k) \\ \vdots & & \vdots \\ f_{n1}(k, x_k, u_k) & \cdots & f_{np}(k, x_k, u_k) \end{bmatrix}$$

where all the $f_{ij}(k, x_k, u_k)$ are known.

Example 4.4.1: To illustrate, the second order system

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} a_o k x_k^3 + a_1 y_k \\ a_2 y_k \sin k^2 \end{bmatrix} + \begin{bmatrix} \xi_{1k} \\ \xi_{2k} \end{bmatrix}$$

where all variables are scalars is considered. This may be written as

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} k x_k^3 & y_k & 0 \\ 0 & 0 & y_k \sin k^2 \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} \xi_{1k} \\ \xi_{2k} \end{bmatrix}$$

$$= D_k(x_k, u_k)a + \xi_k$$

As in the stationary case, if assumptions A1, A6 are satisfied then convergence Theorem III can be applied.

4.5 Miscellaneous

Simple generalizations of the above results are considered in this section. Sec. 4.5a considers systems with generated inputs. Sec. 4.5b considers systems having state disturbances with unknown means, while 4.5c considers systems with indirectly observed states.

4.5a. If the input process u_k is generated by the equation

$$u_{k+1} = C\, u_k$$

where C is an unknown matrix the function u_k can be identified by the matrix C. The observations of u_k are obscured by additive noise.

$$v_k = u_k + \eta_k^2$$

Then, for example, if the system is linear

$$\begin{bmatrix} x_{k+1} \\ u_{k+1} \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} \xi_k \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} y_k \\ v_k \end{bmatrix} = \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} \eta_k^1 \\ \eta_k^2 \end{bmatrix}$$

To identify this system, having unknown matrices A, B and C the results of Sec. 4.1 can be applied and the input generating mechanism can be identified. Convergence Theorem I applies to this case.

If the input mechanism is a random process generated by

$$u_{k+1} = C\, u_k + \xi_k^2$$

and the observations are

$$v_k = u_k + \eta_k^2$$

where C is unknown and ξ_k^2 is a zero mean white process, then with a linear system

$$\begin{bmatrix} x_{k+1} \\ u_{k+1} \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} \xi_k^1 \\ \xi_k^2 \end{bmatrix}$$

The $D_k(.)$ matrix is formed as in Sec. 4.1 and the algorithm identifies the input noise process by its matrix C, as well as A and B.

By adjoining the input generating mechanism to the state equations, similar generalizations are possible for the other cases considered in this chapter. The corresponding convergence theorems will still apply.

4.5b. If the state disturbance process ξ_k has a non-zero unknown mean m, then the process

$$\xi'_k = \xi_k - m$$

has zero mean. For a linear system

$$\begin{aligned} x_{k+1} &= A x_k + B u_k + \xi_k \\ &= A x_k + B u_k + m + \xi'_k \\ &= [A \;\; B \;\; m] \begin{bmatrix} x_k \\ u_k \\ 1 \end{bmatrix} + \xi'_k \end{aligned}$$

The corresponding observation equation is

$$\begin{bmatrix} y_k \\ v_k \\ 1 \end{bmatrix} = \begin{bmatrix} x_k \\ u_k \\ 1 \end{bmatrix} + \begin{bmatrix} \eta_k^1 \\ \eta_k^2 \\ 0 \end{bmatrix}$$

This formulation allows the unknown mean m to be considered as a system parameter and estimated accordingly. In a similar manner, nonlinear or time-varying systems having unknown mean input processes ξ_k can be identified.

4.5c. If the observation equations have the form

$$y_k = H x_k + \eta_k$$

where H is unknown and H^{-1} exists, then for a linear system, a new basis in the state space may be found. Letting

$$Z_k = H x_k$$

gives

$$Z_{k+1} = HAH^{-1}Z_k + HBu_k + H\xi_k$$

and

$$y_k = Z_k + \eta_k$$

Thus the system characterized by the matrices HAH^{-1} and HB may be identified, even though the covariance matrix of the process $H\xi_k$ is unknown.

4.6 Summary

The generality of the algorithm 3.1.13, 3.1.14 has been clearly established by the formulations of this chapter. Equations 4.1.1 and 4.1.2 define the algorithm for the stationary linear case, while 4.2.3, 4.2.4, and 4.2.5 complete the time-varying case. The relations between the convergence theorems of Sec. 3.2 and these systems is also established. For nonlinear systems, 4.3.1 and 4.4.1 define the algorithm and are used to establish the applicability of the convergence theorems. Sec. 4.5 considers other simple generalizations.

V AN EXAMPLE

In order to show how a specific identification problem may be formulated, and to show that implementation of the proposed algorithm is not always difficult, the identification of a fourth-order digital control system with eight unknown parameters is considered in this chapter. Formulation of the problem is given in Sec. 5.1. Computer simulation results for the behavior of the algorithm under different initial estimates, noise conditions, and a-priori uncertainty are shown in Sec. 5.2.

5.1 Problem Statement and Formulation

An interesting example of a situation where identification techniques can be useful is depicted in Fig. 5.1.1. The system consists of a digital controller $\mathcal{C}$ and an unknown plant $\mathcal{P}$. The plant is assumed to be describable by a second-order, linear differential equation with constant coefficients. The digital controller, which is completely specified, is described by a pair of first-order linear difference equations with constant coefficients. The analog-digital (A/D) converter transforms the

continuous signals $e_1(t)$ and $e_2(t)$ into digital inputs for the controller. The D/A converter accepts the controller's output and transforms it to a zero-order-hold (ZOH) signal. The inputs u_1 and u_2 and controller outputs x_3 and x_4 are observed without noise. The plant $\mathcal{P}$ receives inputs from the digital controller $\mathcal{C}$ that are corrupted by additive noise $\xi(t)$. Observations of the plant outputs are corrupted by the additive noise $\eta(t)$ and are used as feedback signals. This situation may arise when the plant is located at a remote site. For example, the plant may be a moon probe while the controller is on the earth. The identification of this system consists in determining the coefficients of the plant's differential equation.

The identifier I operates on the observations u_1, u_2, y_1, y_2, x_3, and x_4 and calculates the plant parameters. Knowledge of these plant parameters may then be used to alter the digital controller so that system performance is improved. For example, the controller may be altered so that the overall system is asymptotically stable.

The digital controller $\mathcal{C}$ is described by 5.1.1,

$$\begin{bmatrix} x_3[(k+1)T] \\ x_4[(k+1)T] \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} x_3(kT) \\ x_4(kT) \end{bmatrix} + \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \begin{bmatrix} e_1(kT) \\ e_2(kT) \end{bmatrix} \qquad 5.1.1$$

where c_{ij} and d_{ij} are known for $i, j = 1, 2$. The constant T represents the length of time between samples.

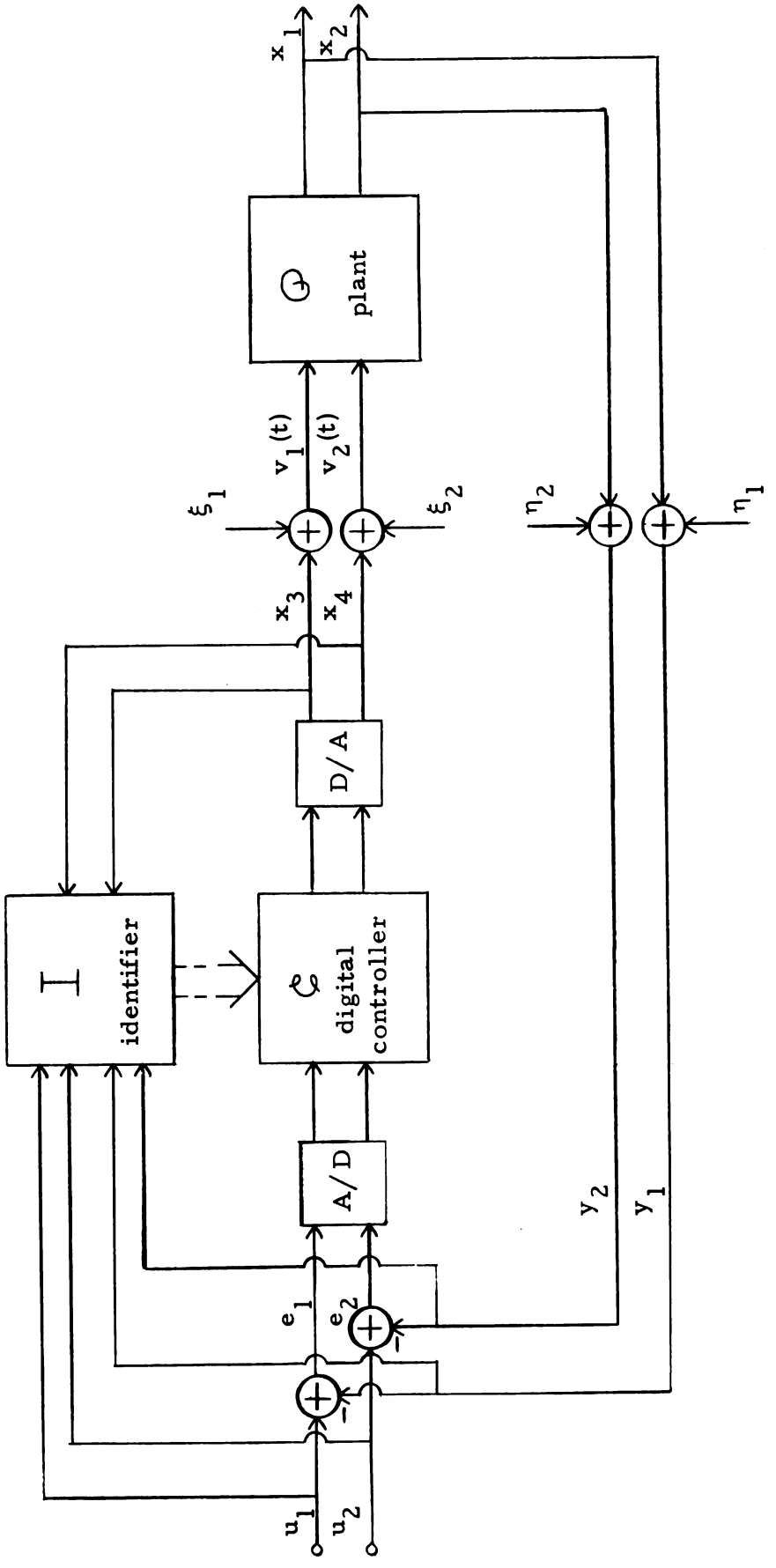

Figure 5.1.1. Digital Control System with Noisy Transmission and Feedback Paths

The plant ⊕ is described by 5.1.2.

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}\begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} \qquad 5.1.2$$

where a_{ij} and b_{ij} are unknown for i, j = 1, 2, and all variables are scalars.

The gaussian processes $\xi(t)$ and $\eta(t)$ are white, have zero mean and are independent. The processes $\xi(t)$ is assumed to behave as a ZOH signal.

From the plant equations and system configuration

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \bar{A}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \bar{B}\begin{bmatrix} x_3(t) \\ x_4(t) \end{bmatrix} + \bar{B}\,\xi(t)$$

where A and B have entries $\bar{a}_{ij}$ and $\bar{b}_{ij}$. Since $x_3(t)$, $x_4(t)$ and $\xi(t)$ are assumed to be ZOH signals, at the sampling times

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = A\begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + B\begin{bmatrix} x_3(k) \\ x_4(k) \end{bmatrix} + B\,\xi(k) \qquad 5.1.3$$

where A and B have entries a_{ij} and b_{ij}, T = 1 is used for convenience and [KOE]

$$A = e^{\bar{A}T}, \quad B = \int_0^T e^{\bar{A}(T-\tau)}\bar{B}\,d\tau \qquad 5.1.4$$

If A and B can be found from the identification algorithm then identification of the continuous system is possible using

$$\overline{A} = \frac{1}{T} \ln A \qquad 5.1.5$$

and

$$\overline{B} = \left[\int_0^T e^{-A\tau} d\tau \right]^{-1} A^{-1} B \qquad 5.1.6$$

Writing the direct sum of 5.1.1 and 5.1.3

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \end{bmatrix} = X_{k+1} = \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} X_k + \begin{bmatrix} B & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} x_3(k) \\ x_4(k) \\ e_1(k) \\ e_2(k) \end{bmatrix} + \begin{bmatrix} B\,\xi(k) \\ 0 \end{bmatrix}$$

Using the relations

$$\begin{bmatrix} x_3(k) \\ x_4(k) \\ e_1(k) \\ e_2(k) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ u_1(k) \\ u_2(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \eta(k) \end{bmatrix}$$

the above becomes

$$X_{k+1} = \begin{bmatrix} A & B \\ -D & C \end{bmatrix} X_k + \begin{bmatrix} 0 \\ D \end{bmatrix} u(k) + \begin{bmatrix} B\,\xi_k \\ -D\,\eta_k \end{bmatrix} \qquad 5.1.7$$

where C and D have entries c_{ij} and d_{ij} as in 5.1.1.

The observation equations are

$$Y_k = \begin{bmatrix} y_1(k) \\ y_2(k) \\ y_3(k) \\ y_4(k) \end{bmatrix} = \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{bmatrix} + \begin{bmatrix} \eta(k) \\ \\ 0 \\ 0 \end{bmatrix} \qquad 5.1.8$$

Equations 5.1.7 and 5.1.8 are of the general form discussed in Sec. 4.1, and the corresponding algorithm can be applied for identification. Since the matrices C and D are known, and x_3 and x_4 are observed, the equations

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = A \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + B \begin{bmatrix} x_3(k) \\ x_4(k) \end{bmatrix} + B\,\xi(k)$$

and

$$\begin{bmatrix} y_1(k) \\ y_2(k) \end{bmatrix} = \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \eta(k)$$

indicate more clearly how the algorithm can be applied.

Defining

$$a^T = [a_{11}a_{12}a_{21}a_{22}b_{11}b_{12}b_{21}b_{22}]$$

then

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} x_1(k) & x_2(k) & 0 & 0 & x_3(k) & x_4(k) & 0 & 0 \\ 0 & 0 & x_1(k) & x_2(k) & 0 & 0 & x_3(k) & x_4(k) \end{bmatrix} a + B\,\xi_k .$$

From Sec. 4.1,

$$D_k(y_k, x_k) = \begin{bmatrix} y_1(k) & y_2(k) & 0 & 0 & x_3(k) & x_4(k) & 0 & 0 \\ 0 & 0 & y_1(k) & y_2(k) & 0 & 0 & x_3(k) & x_4(k) \end{bmatrix}$$

The identification algorithm takes the form

$$\hat{a}_{k+1} = \hat{a}_k + P_{k+1} D_k^T S_k^{-1} \left\{ \begin{bmatrix} y_1(k+1) \\ y_2(k+1) \end{bmatrix} - D_k \cdot \hat{a}_k \right\} \qquad 5.1.9$$

and $$P_{k+1} = P_k - P_k D_k^T (S_k + D_k P_k D_k^T)^{-1} D_k P_k \, . \qquad 5.1.10$$

The S_k matrix may be taken as

$$S_k = Q_k + R_k + S_k'$$

where S_k' is some positive definite symmetric matrix and

$$Q_k = \text{cov}[\xi(k)]$$

$$R_k = \text{cov}[\eta(k)]$$

5.2 Computer Simulation Results

The particular example considered in this section uses the matrices

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 0.9 & 0 \\ 0 & 1 \end{bmatrix} \qquad 5.2.1$$

$$C = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix} \qquad D = \begin{bmatrix} 40/9 & 0 \\ 0 & 1/2 \end{bmatrix} \qquad 5.2.2$$

The plant is unstable since A has eigenvalues of 1 and 2. Also, the closed loop system in the absence of a digital controller is unstable since

$$(A - B) = \begin{bmatrix} 1.1 & 0 \\ 0 & 0 \end{bmatrix}$$

has an eigenvalue greater the unity. However, the complete system with matrix

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

is stable, with eigenvalues of 0, 0, $+1/\sqrt{2}$, $-1/\sqrt{2}$. It is also completely controllable since

$$\begin{bmatrix} 0 & BD \\ D & CD \end{bmatrix}$$

is nonsingular.

The example was programmed on the Control Data 3600 digital computer. Ordinary matrix routines were used to simulate the system. The gaussian noises were simulated using the sum of nine samples from a uniform distribution [SC].

Figure 5.2.1 shows the simulation results for a typical parameter. The average value of $\hat{b}_{22}$ over ten runs is plotted against the number of samples. In this figure, $P_o = 5I$, $S = 3I$, $Q = R = I$ and the initial estimate of all parameters was taken as zero. It can be seen that the initial response of the estimator is good, while as the number of estimates increases, the convergence becomes slow. This is typical of stochastic approximation algorithms.

The effect of different noise levels is depicted in Fig. 5.2.2. Here, the normalized error $\|\epsilon_k\|^2/\|\epsilon_o\|^2$ averaged over ten runs is plotted against the number of samples. In both cases indicated, $P_o = 5I$ and $\hat{a}_o = 0$. In the lower curve, $Q = R = I$ and $S = 3I$, while for the upper curve $Q = R = 25I$ and $S = 50I$. With the increased noise level, $Q = R = 25I$, it can be seen that the convergence rate is slowed considerably. The final value for this case is 1.44×10^{-2} as opposed to 4.34×10^{-4} for the lower noise level.

Figure 5.2.3 shows the effect of different initial estimates. Here, $P_o = 5I$, $Q = R = I$ and $S = 3I$. The error norm $\|\epsilon_k\|^2$ averaged over ten runs is plotted against the number of samples. For the curve having Δ points, $\|\epsilon_o\|^2 = 1,300$, while for the 0-point curve, $\|\epsilon_o\|^2 = 1.00$. As the number of estimates increases, the effect of initial errors disappears and the two estimators have similar asymptotic properties. This indicates that the algorithm's convergence is unaffected by initial errors.

The effect of a-priori uncertainty is depicted in Fig. 5.2.4. In both cases indicated, $Q = R = I$, $S = 3I$, and $\hat{a}_o = 0$. A ten run

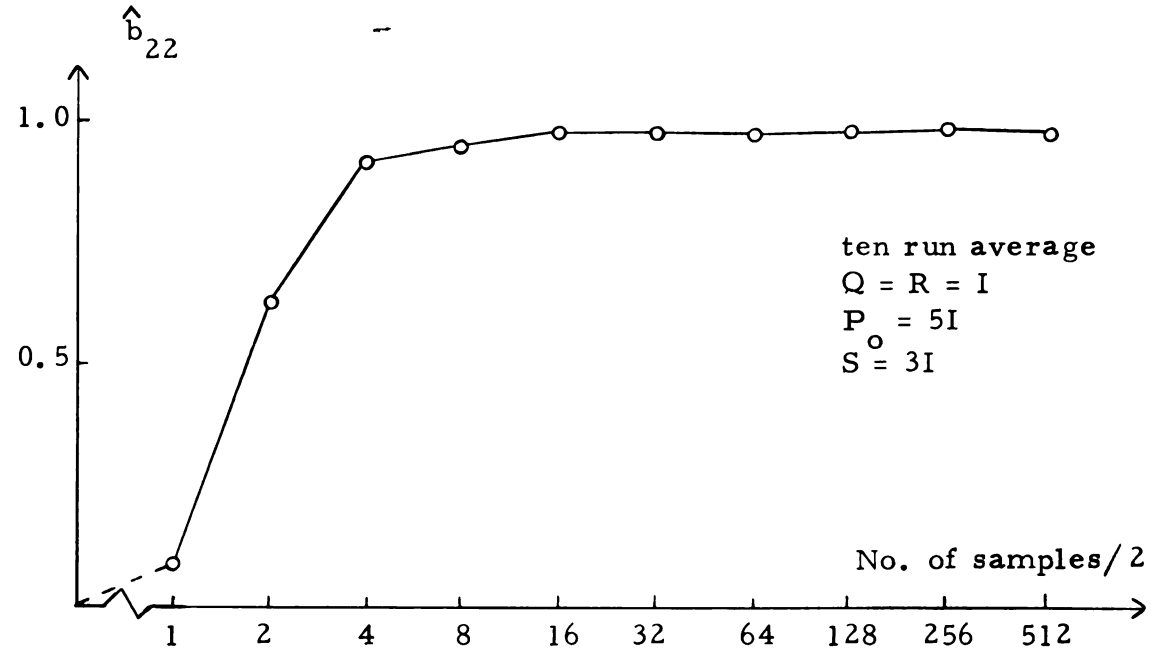

Figure 5.2.1. Behavior of a Typical Estimator

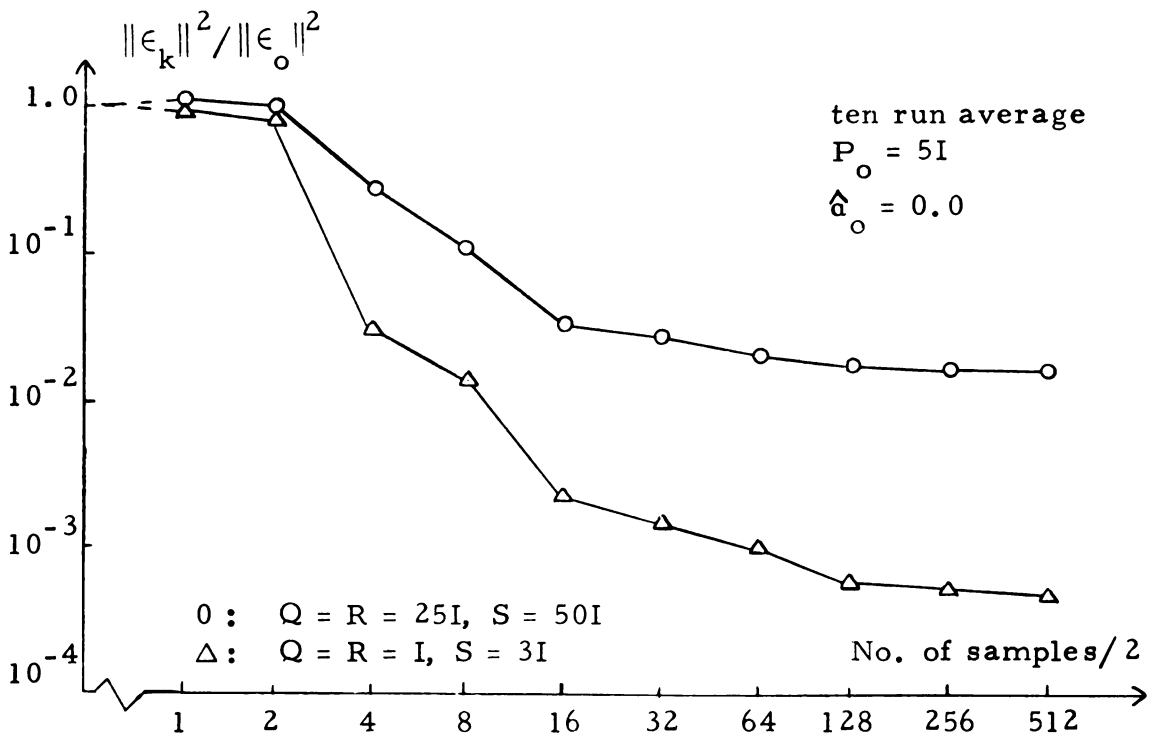

Figure 5.2.2. Effect of Noise Power on Convergence

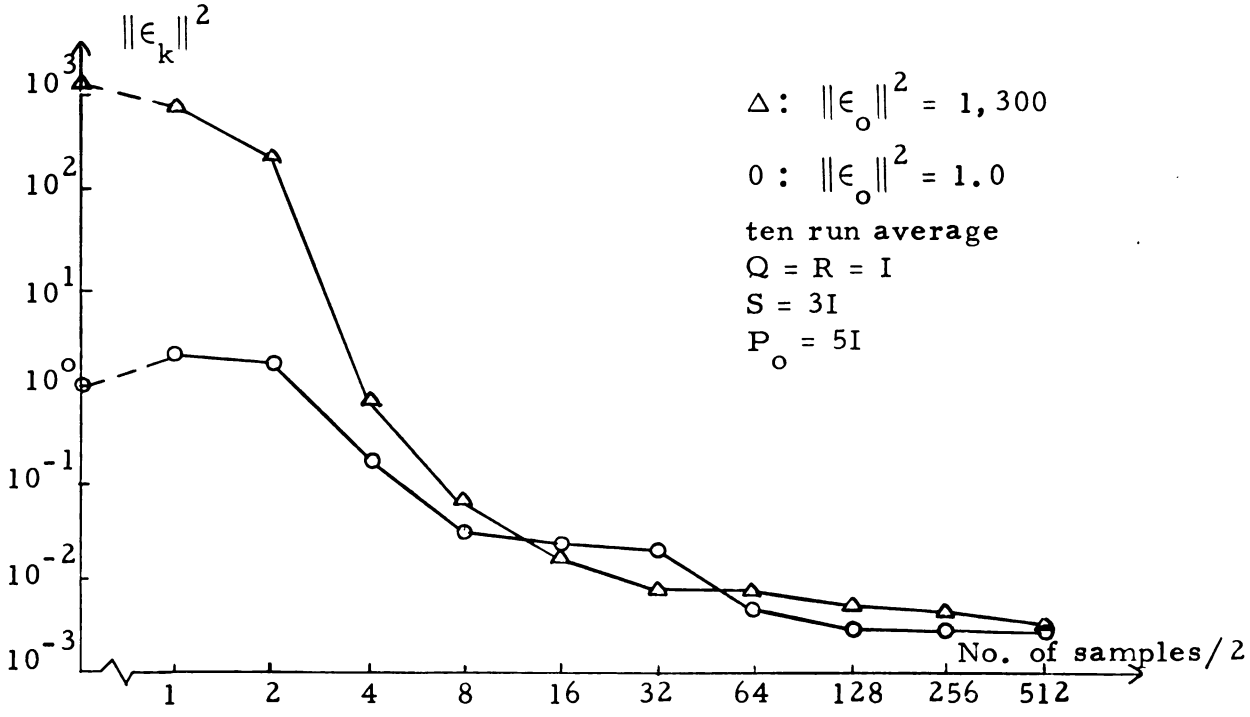

Figure 5.2.3. Effect of Initial Estimates on Convergence

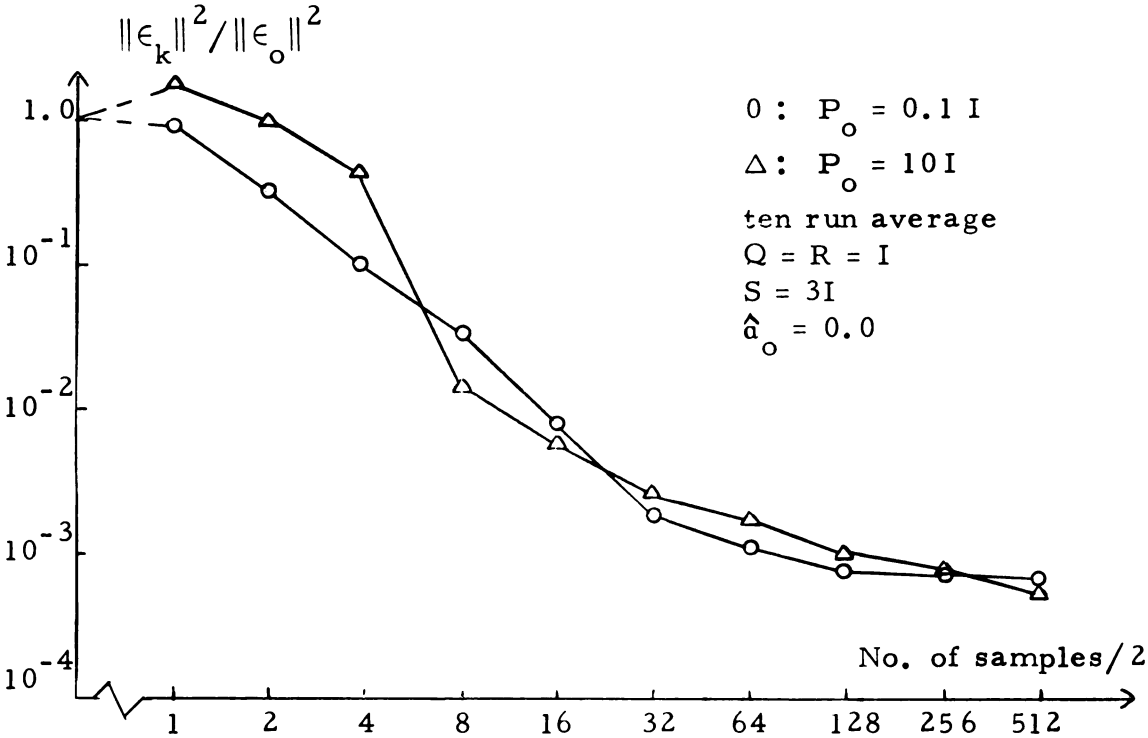

Figure 5.2.4. Effect of Uncertainty on Convergence

average of the normalized error $\|\epsilon_k\|^2/\|\epsilon_o\|^2$ is plotted against the number of estimates. The uncertainty associated with initial estimates is reflected in the a-priori covariance matrix P_o. If the initial estimates are thought to be close to the true values, then P_o will be small (in norm). If initial uncertainty is large, the P_o will be large (in norm). In the curve with points Δ, $P_o = 10I$ and for the 0-point curve $P_o = 0.1I$. In both cases the initial error norm is small, having a value of 2.61. As indicated by Fig. 5.2.4 the effect of initial uncertainty tends to produce large changes in the initial estimates which are close to the true values. This results in errors that are large relative to those produced by the same initial estimate with a smaller P_o. As more observations become available, the effects of the a-priori uncertainty become small and both estimates have similar asymptotic properties.

5.3 Summary

The identification of a fourth-order digital control system with eight unknown parameters has been studied in this chapter. The formulation of the identification algorithm in Sec. 5.1 was shown to be a simple application of the results of Sec. 4.1. Various properties of the algorithm were demonstrated by a computer simulation in Sec. 5.2. It was shown that a-priori uncertainty

and initial estimate errors do not affect the asymptotic properties of the algorithm. It was also shown that increased noise power slows the algorithm's rate of convergence considerably.

VI CONCLUSION

The major results of the thesis are listed in Sec. 6.1 and possible extensions of this work are discussed in Sec. 6.2.

6.1 Results of the Thesis

The formulation of a general system identification problem as a problem in Bayesian learning is carried out in Chap. II. This formulation is significant because it provides a unifying structure through which a wide variety of system identification problems can be solved. In particular, Sec. 2.3 shows the basic relations which exist for the general system 2.3.1, 2.3.2., 2.3.3.

Chapter III gives the solution of the Bayesian identification problem for the general system 3.1.1, 3.1.2, 3.1.3. The general algorithm 3.1.13, 3.1.14 is derived under gaussian assumptions with an approximation to the optimal Bayesian estimator, and is optimal for linear systems having no observation noises. This algorithm is more general than any previous results, and can be specialized to many important cases as shown in Sec. 3.3.

The proofs of the three convergence theorems of Sec. 3.2 are new, and are found in Appendix I. Convergence Theorems I and II are new and can be more readily applied than those given by other authors. These theorems are important to the thesis because they establish that the algorithm does indeed yield the true system parameters.

Chapter IV shows how the algorithm 3.1.13, 3.1.14 can be specialized to important classes of systems. New results or generalizations are obtained for linear and nonlinear systems in each section. The formulation for stationary linear systems is new because of the presence of input observation noises. The time-varying linear formulation is significant because observation noises have not been considered previously in this case. The nonlinear formulations are new because of the input observation noises and non-stationarity of the system. The generalizations of Sec. 4.5 also contribute to the results of this thesis.

The feasibility of the identification algorithm is demonstrated by an example of a fourth-order system with eight unknown parameters in Chap. V. A computer simulation shows that increased noise power slows the algorithm's rate of convergence and that the effect of initial estimates becomes negligible as the number of observations becomes large. Also, when a-priori uncertainty is large, the effect of initial observations is weighted heavily, resulting in initial transients in the algorithm.

6.2 Possible Extensions

The class of systems to which the identification algorithm is applicable is very general. However, the formulation in this thesis applies only to discrete-time systems. A very natural extension of the results given in Chaps. II and III is to continuous-time systems. Conceptually, this is not a difficult problem, although the details of the limiting arguments may be difficult. In the continuous case, an equation analgous to 3.1.13 would be

$$\frac{d}{dt}\alpha(t) = P(t)D^T(t)S(t)^{-1}\{\frac{d}{dt}y(t) - D(t)\alpha(t)\} \qquad 6.2.1$$

and for 3.1.14

$$\frac{d}{dt}P(t) = -P(t)D^T(t)\{S(t)^{-1} + D(t)P(t)D^T(t)\}^{-1}D(t)P(t) \qquad 6.2.2a$$

or

$$\frac{d}{dt}P^{-1}(t) = D^T(t)S(t)^{-1}D(t) \qquad 6.2.2b$$

Implementation of these equations for analog computer simulations would be difficult, although hybrid techniques would reduce these difficulties considerably.

Another useful extension of the results studied here would be the elimination of 3.1.14. A deterministic sequence of matrices P_k would reduce the computational requirements. Furthermore, the convergence proofs would be simpler, although the sequence of estimates would be less optimal. For example, $(1/k)I$ might replace P_k. Other more general forms for P_k might be considered and the rate of convergence optimized over a class of

The identification algorithm could also be used to study adaptive controller systems. As indicated in Chap. V, the identifier could be used to track system parameters and the digital controller could be modified accordingly. Other adaptive learning loops could be studied. This same philosophy can be applied to communication problems. By utilizing a channel identifier, an adaptive detector could be designed.

The convergence proofs of Appendix I could be extended. Some of the techniques used in these proofs may be useful in establishing multi-dimensional stochastic approximation algorithms.

Finally, the study of canonical forms would be very useful when all system states are not observable.

APPENDIX I

CONVERGENCE PROOFS FOR THE IDENTIFICATION ALGORITHM

The three convergence theorems for the identification algorithm 3.1.13, 3.1.14 were stated with assumptions in Sec. 3.2. In this appendix these theorems will be proved.

For convenience of reference the algorithm equations are restated here.

$$\hat{a}_{k+1} = \hat{a}_k + P_{k+1} D_k^T(y_k, v_k) S_k^{-1} [y_{k+1} - D_k(y_k, v_k)\hat{a}_k] \quad \text{A1.1}$$

$$P_{k+1} = P_k - P_k D_k^T(y_k, v_k)[S_k + D_k(y_k, v_k) P_k D_k^T(y_k, v_k)]^{-1} D_k(y_k, v_k) P_k. \quad \text{A1.2}$$

With $\epsilon_k = a_k - a$,

$$\epsilon_{k+1} = P_{k+1} P_k^{-1} \epsilon_k + P_{k+1} D_k^T(y_k, v_k) S_k^{-1} [y_{k+1} - D_k(y_k, v_k) \cdot a] \quad \text{A1.3}$$

In the following discussion D_k will denote $D_k(y_k, v_k)$ as opposed to $D_k(x_k, u_k)$. Since P_i is a covariance matrix it is real, positive definite and symmetric.

From the definition of Sec. 3.2,

$$\|P_i\| = \sqrt{\lambda_{max}(P_i^* P_i)} = \sqrt{\lambda_{max}(P_i^2)} = \lambda_{max}(P_i)$$

Also,

$$\lambda_{max}(P_i) = \frac{1}{\lambda_{min}(P_i^{-1})}$$

Lemma A1.1: The product of two positive definite symmetric matrices has positive eigenvalues.

Proof: For any square matrix F and any nonsingular G, $(FG)^T FG$ and $F^T F\, GG^T$ have the same eigenvalues since by the similarity transform G^T [ROS]

$$G^T[F^T FGG^T]\ (G^T)^{-1} = G^T F^T FG\ .$$

Also, for any nonsingular H, $H^T H$ is positive definite since $\forall\, x \neq 0$

$$x^T(H^T H)x = (Hx)^T(Hx) > 0 \text{ since } Hx \neq 0.$$

For two positive definite symmetric matrices P and Q, $P^{\frac{1}{2}}$ and $Q^{\frac{1}{2}}$ are positive definite and symmetric. Thus

$$PQ = P^{\frac{1}{2}}P^{\frac{1}{2}}Q^{\frac{1}{2}}Q^{\frac{1}{2}} = (P^{\frac{1}{2}})^T P^{\frac{1}{2}}Q^{\frac{1}{2}}(Q^{\frac{1}{2}})^T\ ,$$

has the same eigenvalues as $(P^{\frac{1}{2}}Q^{\frac{1}{2}})^T(P^{\frac{1}{2}}Q^{\frac{1}{2}})$. But $(P^{\frac{1}{2}}Q^{\frac{1}{2}})$ is nonsingular since P and Q are nonsingular, so that $(P^{\frac{1}{2}}Q^{\frac{1}{2}})^T(P^{\frac{1}{2}}Q^{\frac{1}{2}})$ is positive definite and has positive eigenvalues. Therefore the product PQ has positive eigenvalues.

Lemma A1.2: If the system is identifiable then the norm of $P_{k+1}P_{k-q}^{-1}$ satisfies

$$\|P_{k+1}P_{k-q}^{-1}\| < 1$$

and the norm of $P_{k-q}^{-1}P_{k+1}$ satisfies

$$\|P_{k-q}^{-1}P_{k+1}\| < 1$$

with probability one (WP1).

Proof: Only the first inequality will be proved. The second follows from a similar development. By lemma A1.1, $P_{k+1}P_{k-q}^{-1}$ has positive eigenvalues since P_j is positive definite and symmetric for any j. By identifiability, $\sum_{j=k-q}^{k} D_j^T S_j^{-1} D_j$ is positive definite and symmetric (WP1), so

$$P_{k+1} \sum_{j=k-q}^{k} D_j^T S_j^{-1} D_j$$

has positive eigenvalues (WP1).

From A1.2

$$P_{k+1}^{-1} = P_{k-q}^{-1} + \sum_{j=k-q}^{k} D_j^T S_j^{-1} D_j .$$

Therefore, pre-multiplying by P_{k+1}

$$P_{k+1}P_{k-q}^{-1} = I - P_{k+1} \sum_{j=k-q}^{k} D_j^T S_j^{-1} D_j .$$

Both product terms have positive eigenvalues (WP1) so that the Jordan form

$$J_L = I - J_R$$

with eigenvalues on the diagonal shows that the eigenvalues of J_L or those of $P_{k+1}P_{k-q}^{-1}$ are strictly less than one (WP1). Thus the maximum eigenvalue is less than one, and

$$\|P_{k+1}P_{k-q}^{-1}\| < 1 \qquad \text{(WP1)}.$$

An interesting and important property of the matrix P_{k+1} will now be derived through a sequence of lemmas.

Lemma A1.3. The minimum eigenvalue of a symmetric matrix F satisfies

$$\lambda_{min}(F) = \inf_{x \neq 0} \frac{x^T F x}{x^T x} .$$

Proof: It is first necessary to show that

$$x^T x \lambda_{min}(F) \leq x^T F x$$

or that

$$x^T [F - \lambda_{min}(F) \cdot I] x \geq 0$$

i.e., $[F - \lambda_{min}(F) \cdot I]$ is positive semi-definite. Letting J_F denote the Jordan form of F, and J the Jordan form of the difference gives

$$J = J_F - \lambda_{min}(F) \cdot I .$$

Since all eigenvalues of F are on the diagonal of J_F and $\lambda_{min}(F) \leq \lambda(F)$ $\forall$ $\lambda(F)$ where $\lambda(.)$ denotes an eigenvalue, then the eigenvalues of J satisfy $\lambda(J) \geq 0$. Therefore $F - \lambda_{min}(F) \cdot I$ is positive semi-definite. Thus

$$\lambda_{min}(F) \leq \frac{x^T F x}{x^T x} \qquad \forall \; x \neq 0 .$$

Furthermore, since $\lambda_{min}(F)$ is an eigenvalue there exists $x' \neq 0$ such that

$$F x' = \lambda_{min}(F) x'$$

so $$x'^T F x' = x'^T \lambda_{min}(F) x'$$

and $$\lambda_{min}(F) = \frac{x'^T F x'}{x'^T x'} \leq \frac{x^T F x}{x^T x} \quad \forall\, x \neq 0 .$$

Therefore by definition of infimum [ROY]

$$\lambda_{min}(F) = \inf_{x \neq 0} \frac{x^T F x}{x^T x} .$$

Lemma A1.4. All eigenvalues of P_{k+1}^{-1} approach infinity as $k \to \infty$, WP1, and do so at a rate greater or equal to $c \cdot k$, for some constant $c > 0$.

Proof: Iterating equation A1.2 gives

$$P_{k+1}^{-1} = P_o^{-1} + \sum_{i=1}^{k/q} \sum_{j=q(i-1)}^{q \cdot i - 1} D_j^T S_j^{-1} D_j, \quad k = q, 2q, \ldots .$$

By identifiability, $\sum_j D_j^T S_j^{-1} D_j$ is positive definite (WP1) and has all positive eigenvalues. Therefore there exists a constant $c' > 0$ such that for all eigenvalues λ

$$\lambda \geq c' > 0$$

Using this, Lemma A1.3, and the fact that for all realizations of $\{z_i\}$

$$\inf \sum_i z_i \geq \sum_i \inf z_i$$

it follows that

$$\lambda_{min}\Big[\sum_{j=0}^{k} D_j^T S_j^{-1} D_j\Big] = \lambda_{min}\Big[\sum_{i=1}^{k/q} \sum_{j=q(i-1)}^{q \cdot i - 1} D_j^T S_j^{-1} D_j\Big]$$

$$= \inf_{x \neq 0} \frac{x^T \sum_i (\sum_j D_j^T S_j^{-1} D_j) x}{x^T x}$$

$$\geq \sum_i \inf_{x \neq 0} \frac{x^T (\sum_j D_j^T S_j^{-1} D_j) x}{x^T x}$$

$$\geq \sum_i \lambda_{min} (\sum_j D_j^T S_j^{-1} D_j)$$

$$\geq \sum_i c' \geq c'' \cdot k, \qquad c'' > 0 .$$

Therefore

$$\lambda_{min}[P_{k+1}^{-1}] \geq \lambda_{min}[P_o^{-1}] + \lambda_{min}[\sum_{j=0}^{k} D_j^T S_j^{-1} D_j]$$

$$\geq c_o + c_1 k, \qquad c_o, c_1 > 0$$

and all eigenvalues of P_{k+1}^{-1} approach infinity at least as fast as k (WP1).

Lemma A1.5. The norm of P_{k+1} satisfies

$$\| P_{k+1} \| \leq c_2 / (k + c_3) ,$$

(WP1) where $c_2, c_3 > 0$ are constants.

Proof: $\|P_{k+1}\| = \lambda_{max}(P_{k+1}) = 1/\lambda_{min}(P_{k+1}^{-1})$

$$\leq 1/(c_o + c_1 k) \leq c_2/(k + c_3)$$

by lemma A1.4.

From lemmas A1.2 and A1.5

$$\|P_{k+1}P_{k-q}^{-1}\| < 1 \qquad \text{(WP1)}$$

and

$$\|P_{k+1}\| \leq c_2/(k + c_3) \quad \text{(WP1)}.$$

Only system identifiability was used to establish these properties. These results will be useful in establishing convergence of the identification algorithm.

Equation A1.3 may be rewritten as

$$\epsilon_{k+1} = P_{k+1}P_{k-q}^{-1}\epsilon_{k-q} + P_{k+1}\sum_{j=k-q}^{k} D_j^T S_j^{-1}(y_{j+1} - D_j\alpha)$$

by recursive substitution. From this equation the product $\epsilon_{k+1}\epsilon_{k+1}^T$ is written as

$$\epsilon_{k+1}\epsilon_{k+1}^T = P_{k+1}P_{k-q}^{-1}\epsilon_{k-q}\epsilon_{k-q}^T P_{k-q}^{-1}P_{k+1} + 2P_{k+1}\sum_{j=k-q}^{k} D_j^T S_j^{-1}(y_{j+1} - D_j\alpha)\epsilon_{k-q}^T P_{k-q}^{-1}P_{k+1} + P_{k+1}\sum_{j=k-q}^{k} D_j^T S_j^{-1}(y_{j+1} - D_j\alpha) \cdot \sum_{j=k-q}^{k}(y_{j+1} - D_j\alpha)^T S_j^{-1} P_{k+1} \qquad \text{A1.5}$$

Utilizing assumption 3.2.A2, the expectation of the cross product term in A1.5 vanishes since

$$E\{2P_{k+1}\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j\alpha)\epsilon_{k-q}^T P_{k-q}^{-1}P_{k+1}\}$$

$$= 2E\{P_{k+1}E[\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j\alpha)\epsilon_{k-q}^T P_{k-q}^{-1}/y^k, v^k]P_{k+1}\}$$

$$= 2E\{P_{k+1}\sum_j E\{D_j^T S_j^{-1}E[(y_{j+1} - D_j\alpha)/y^j, v^j]\epsilon_{k-q}^T P_{k-q}^{-1}\}P_{k+1}\}$$

$$= 0.$$

Therefore

$$E\{\epsilon_{k+1}\epsilon_{k+1}^T\} = E\{P_{k+1}P_{k-q}^{-1}\epsilon_{k-q}\epsilon_{k-q}^T P_{k-q}^{-1}P_{k+1}\}$$

$$+ E\{P_{k+1}\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\sum_j (y_{j+1} - D_j a)^T S_j^{-1} P_{k+1}\}$$

Taking the norm of both sides of this equation and using the triangle inequality

$$\|E(\epsilon_{k+1}\epsilon_{k+1}^T)\| \leq \|E\{P_{k+1}P_{k-q}^{-1}\epsilon_{k-q}\epsilon_{k-q}^T P_{k-q}^{-1}P_{k+1}\}\|$$

$$+ \|E\{P_{k+1}\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\sum_j (y_{j+1} - D_j a)^T S_j^{-1} P_{k+1}\}\|$$

Applying this result, Lemma A1.2 and Lemma A1.5,

$$\|E(\epsilon_{k+1}\epsilon_{k+1}^T)\| \leq c_1^2\|E(\epsilon_{k-q}\epsilon_{k-q}^T)\| + \frac{c_2^2}{(k+c_3)^2}\|E\{\sum_j D_j^T S_j^{-1}(y_{j+1}$$

$$- D_j a)\sum_j (y_{j+1} - D_j a)^T S_j^{-1}\| \qquad \text{A1.6}$$

where $0 < c_1 < 1$ and $0 < c_2, c_3 < \infty$.

Lemma A1.6. For the scalars x_i, a_i, β_i if $x_{i+1} \leq a_i x_i + \beta_i$ and $\prod_{i=0}^{\infty} a_i = 0$ with all partial products uniformly bounded, $\sum_0^{\infty} \beta_i < \infty$, $x_i \geq 0$, $x_o < \infty$, $\beta_i \geq 0$ then $x_i \to 0$ as $i \to \infty$.

Proof: Recursively substituting into the inequality for x_{i+1} gives

$$x_{i+1} \leq \alpha_i x_i + \beta_i$$

$$\leq \alpha_i \alpha_{i-1} x_{i-1} + \alpha_i \beta_{i-1} + \beta_i$$

$$\vdots \qquad\qquad \vdots$$

$$x_{i+1} \leq \alpha_i \alpha_{i-1} \cdots \alpha_o x_o + \alpha_i \alpha_{i-1} \cdots \alpha_1 \beta_o + \cdots$$

$$+ \alpha_i \alpha_{i-1} \beta_{i-2} + \alpha_i \beta_{i-1} + \beta_i$$

$$x_{i+1} \leq \prod_{j=0}^{i} \alpha_j x_o + \sum_{j=0}^{i-1} \prod_{k=j+1}^{i} \alpha_k \beta_j + \beta_i$$

Since $\prod_{i}^{\infty} \alpha_i = 0$ and all partial products are bounded, for any $m > 0$

$$\sum_{j=0}^{m-1} \prod_{k=j+1}^{i} \alpha_k \beta_j \leq \sum_{j=0}^{m-1} \beta_j \max_{0 \leq j \leq m-1} \prod_{k=j+1}^{i} \alpha_k$$

and

$$\sum_{j=m}^{i} \prod_{k=j+1}^{i} \alpha_k \beta_j \leq \sum_{j=m}^{i} \beta_j \max_{m \leq j \leq i} \prod_{k=j+1}^{i} \alpha_k$$

Given any $\epsilon > 0$ there exists m sufficiently large so that

$$\sum_{j=m}^{i} \beta_j \max_{m \leq j \leq i} \prod_{k=j+1}^{i} \alpha_k \leq \sum_{m}^{i} \beta_j \cdot M < \epsilon/3$$

since $\sum_j \beta_j < \infty$. Thus for m sufficiently large

$$\sum_{j=0}^{i} \prod_{k=j+1}^{i} \beta_j \alpha_k < \epsilon/3 + \sum_{j=0}^{m-1} \beta_j \max_{0 \leq j \leq m-1} \prod_{k=j+1}^{i} \alpha_k + \beta_i$$

For fixed m,

$$\sum_{j=0}^{m-1} \beta_j \max_{0 \leq j \leq m-1} \prod_{k=j+1}^{i} \alpha_k < \epsilon/3$$

for i sufficiently large, since $\prod a_j = 0$. Also $\beta_i < \epsilon/3$ since $\sum \beta_i < \infty$. Therefore

$$\prod_{j=0}^{i} a_j x_o + \sum_{j=0}^{i-1} \prod_{k=j+1}^{i} a_k \beta_j + \beta_i < \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon \text{ for large } i$$

implies $x_i \to 0$ as $i \to \infty$ since $x_i \geq 0$. [DV][WOL].

Assumption 3.2.A3 and the previous results give the first convergence theorem.

Convergence Theorem I: If the system is identifiable and assumptions 3.2.A1, 3.2.A2, 3.2.A3 hold, then the algorithm A1.1, A1.2 converges to the true value of the parameter vector a in the sense that

$$\|\mathrm{cov}(\epsilon_k)\| = \|\mathrm{cov}(\hat{a}_k - a\| \to 0 \text{ as } k \to \infty.$$

Proof: From A1.6 and 3.2.A3 $\|E(\epsilon_{k+1}\epsilon_{k+1}^T)\| \leq c_1^2 \|E(\epsilon_{k-q}\epsilon_{k-q}^T)\| + c'_2/(k + c_3)^2$ where $0 < c_1 < 1$, $0 < c_3 < \infty$, and $0 < c'_2 < \infty$. The following associations are made in lemma A1.6

$$x_i \longleftrightarrow \|E(\epsilon_{k-q}\epsilon_{k-q}^T)\|$$

$$x_{i+1} \longleftrightarrow \|E(\epsilon_{k+1}\epsilon_{k+1}^T)\|$$

$$a_i \longleftrightarrow c_1^2$$

$$\beta_i \longleftrightarrow c_2/(k + c_3)^2$$

and noting that

$$\prod_{0}^{\infty} c_1^2 = 0 \quad \text{since } c_1 < 1$$

and

$$\sum_{k=0}^{\infty} c_2/(k+c_3)^2 < \infty$$

gives

$$\|E(\epsilon_{k+1}\epsilon_{k+1}^T)\| = \|\text{cov}\,\epsilon_{k+1}\| = \|\text{cov}(\hat{a}_{k+1} - a)\| \to 0 \text{ as } k \to \infty .$$

Convergence theorem II is proved from assumptions 3.2.A1, 3.2.A3, 3.2.A4. Rewriting A1.3 and taking the expected value of the norm squared gives

$$E\{\|\epsilon_{k+1}\|^2\} = E\{\|P_{k+1}P_{k-q}^{-1}\epsilon_{k-q} + P_{k+1}\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\|^2\}$$

From 3.2.A4

$$E\{\|\epsilon_{k+1}\|^2\} \leq E\{\|P_{k+1}P_{k-q}^{-1}\epsilon_{k-q}\|^2\} + E\{\|P_{k+1}\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\|^2\}$$

$$\leq E\{\|P_{k+1}P_{k-q}^{-1}\|^2\|\epsilon_{k-q}\|^2\} + E\{\|P_{k+1}\|^2\|\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\|^2\}$$

Applying Lemmas A1.2 and A1.5

$$E\{\|\epsilon_{k+1}\|^2\} \leq c_1^2\, E\{\|\epsilon_{k-q}\|^2\} + \frac{c_2}{(k+c_3)^2}\, E\{\|\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\|^2\} \qquad \text{A1.7}$$

Convergence Theorem II: If the system is identifiable and assumptions 3.2.A1, 3.2.A3, 3.2.A4 hold then the algorithm A1.1, A1.2 converges to the true value of the parameter vector a in the sense that

$$E\{\|\epsilon_k\|^2\} = E\{\|\hat{a}_k - a\|^2\} \to 0 \text{ as } k \to \infty$$

i.e., the error norm converges to zero in mean square.

Proof: From A1.7 and assumption 3.2. A5

$$E\{\|\epsilon_{k+1}\|^2\} \leq c_1^2 E\{\|\epsilon_{k-q}\|^2\} + \frac{c_2}{(k+c_3)^2} E\{\|\sum_j D_j^T S_j^{-1}(y_{j+1} - D_j a)\|^2\}$$

$$E\{\|\epsilon_{k+1}\|^2\} \leq c_1^2 E\{\|\epsilon_{k-q}\|^2\} + c_4/(k+c_3)^2, \quad 0 < c_1 < 1, \quad 0 < c_4 < \infty .$$

Since $\prod_{j=0}^{\infty} c_1^2 = 0$, and $\sum_0^{\infty} c_4/(k+c_3)^2 < \infty$, Lemma A1.6 implies

$$E\{\|\epsilon_{k+1}\|^2\} \to 0 \quad \text{as } k \to \infty ,$$

$$E\{\|\hat{a}_{k+1} - a\|^2\} \to 0 \quad \text{as } k \to \infty .$$

Theorem III is proved using assumptions 3.2.A1, 3.2.A6, and strong identifiability. This latter condition is stronger than identifiability and follows from lemma A1.4, which states that the minimum eigenvalue of $\sum_{j=0}^{k} D_j^T S_j^{-1} D_j$ approaches infinity at a rate greater than or equal to $c \cdot k$. Strong identifiability requires that this rate be strictly greater than $c \cdot k$, or at least equal to $c \cdot k^p$, $p > 1$. Considering the inequalities used in the proof of lemma A1.4 and the random nature of the system variables, this is not a stringent requirement.

Lemma A1.7: If the system is strongly identifiable,

$$\|P_{k+1}\| \leq c_4/(k^p + c_3) , \qquad \text{(WP1)}$$

where

$$0 < c_4, \; c_3 < \infty, \; 1 < p < \infty .$$

Proof: $\|P_{k+1}\| = \lambda_{max}(P_{k+1}) = 1/\lambda_{min}(P_{k+1}^{-1})$

$$\leq 1/[\lambda_{min}(P_o^{-1}) + \lambda_{min}(\sum_{j=0}^{k} D_j^T S_j^{-1} D_j)]$$

$$\leq 1/[c_1 + c_2 k^p] \leq c_4/(k^p + c_3),$$

$1 < p < \infty$, $0 < c_3, c_4 < \infty$ with probability one, by assumption.

Lemma A1.8: If $x_{k+1} \leq a_k x_k + \beta_k$ and $0 \leq x_o < \infty$, $\prod_{k=0}^{\infty} a_k = 0$ with all partial products uniformly bounded, $\sum_{k=0}^{\infty} \beta_k < \infty$, $\beta_k \geq 0$, all with probability one, then $x_k \to 0$ as $k \to \infty$ WP1.

Proof: This proof follows that of lemma A1.6 except that statements hold with probability one (WP1). This lemma is stated by Fukao [FK-3] without proof.

Convergence Theorem III: If the system is strongly identifiable and if assumptions 3.2.A1, 3.2.A6 hold then the algorithm A1.1, A1.2 converges to the true value of the parameter vector a in the sense that

$$\|\epsilon_{k+1}\| = \|\hat{a}_{k+1} - a\| \to 0 \text{ as } k \to \infty \quad \text{(WP1)}.$$

Proof: Using the triangle inequality for norms in A1.3

$$\|\epsilon_{k+1}\| \leq \|P_{k+1} P_{k-q}^{-1}\| \cdot \|\epsilon_{k-q}\| + \|P_{k+1}\| \cdot \|\sum_{j=k-q}^{k} D_j^T S_j^{-1} (y_{j+1} - D_j a)\|$$

Lemma A1.2, strong identifiability, and 3.2.A6 give

$$\|\epsilon_{k+1}\| \leq c_1\|\epsilon_{k-q}\| + c_5/(k^p + c_3)$$

where $0 < c_1 < 1$, $1 < p < \infty$, $0 < c_5 < \infty$ (WP1).

Since

$$\sum_{k=0}^{\infty} c_5/(k+c_3) < \infty \quad \text{and} \quad \prod_{k=0}^{\infty} c_1 = 0 \quad \text{(WP1)},$$

applying Lemma A1.8 shows that

$$\|\epsilon_{k+1}\| = \|\hat{a}_{k+1} - a\| \to 0 \quad \text{(WP1) as } k \to \infty .$$

This theorem is equivalent to one given by Fukao [FK-3] but the proof is distinct.

APPENDIX II

THE REPRODUCING NORMAL AND ITS QUADRATIC FORM

In Sec. 3.1, the a-posteriori distribution of the parameter vector a was given as gaussian with mean value and covariance given by 3.1.11 and 3.1.12, respectively, and followed from the reproducing property of the gaussian distribution and minimization of a quadratic form. The details of this derivation are given in this appendix.

From 3.1.9 and 3.1.10 the exponent of the numerator equals (within a -1/2 factor)

$$a^T\{P_k^{-1} + D_k^T S_k^{-1} D_k\} a - 2a^T\{P_{k+1}^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}\} + \{\hat{a}_k^T P_k^{-1}\hat{a}_k + y_{k+1}^T S_k^{-1} y_{k+1}\} \quad A2.1$$

Since the integration in the denominator 3.1.9 is with respect to a, the third term in this quadratic form cancels. Completing the square of the remaining part of A2.1 gives

$$a^T\{P_k^{-1} + D_k^T S_k^{-1} D_k\} a - 2a^T\{P_{k+1}^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}\} =$$

$$(a - [P_k^{-1} + D_k^T S_k^{-1} D_k]^{-1}[P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}])^T [P_k^{-1} + D_k^T S_k^{-1} D_k]$$

$$\times (a - [P_k^{-1} + D_k^T S_k^{-1} D_k]^{-1}[P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}]$$

$$- [P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} D_k]^T [P_k^{-1} + D_k S_k^{-1} D_k]^{-1}[P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}]$$

A2.2

which is the exponent of the numerator and the exponent under the integral sign of 3.1.9. Integrating the perfect square in a leaves the second term of A2.2, which cancels with an equal term in the numerator. The resulting exponent of the a-posteriori distribution is the quadratic form

$$(a - [P_k^{-1} + D_k^T S_k^{-1} D_k]^{-1}[P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}])^T [P_k^{-1} + D_k^T S_k^{-1} D_k]$$

$$\times (a - [P_k^{-1} + D_k^T S_k^{-1} D_k]^{-1}[P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}])$$

and is a perfect square. Therefore the a-posterior distribution is gaussian with mean

$$\hat{a}_{k+1} = [P_k^{-1} + D_k^T S_k^{-1} D_k]^{-1}[P_k^{-1}\hat{a}_k + D_k^T S_k^{-1} y_{k+1}] \quad \text{A2.3}$$

and covariance

$$P_{k+1}^{-1} = P_k^{-1} + D_k^T S_k^{-1} D_k . \quad \text{A2.4}$$

Equation A2.3 is the same as 3.1.11 and can be written in the form 3.1.13. Using a matrix inversion lemma [HOU][AO-1], A2.4 can be written in the form 3.1.14.

The mean $\hat{a}_{k+1}$ can also be obtained directly from A2.1 by minimization of that quadratic form with respect to a, once the above gaussian reproducing property is established.

BIBLIOGRAPHY

[AH] Ahmed, N. and S. Karni, "On Obtaining Transfer Functions from Gain Function Derivatives," IEEE Trans. on Automatic Control, AC-12, No. 2, p. 229, April 1967.

[AHL] Ahlfors, L. V., Complex Analysis, 2nd Ed., McGraw-Hill Book Co., Inc., New York, 1966.

[AND] Anderson, G. W., R. N. Buland, and G. R. Cooper, "Use of Cross Correlation in an Adaptive Control System," Proc. Nat'l. Elec. Conf., Vol. 15, Oct. 1959.

[AO-1] Aoki, M., Optimization of Stochastic Systems, Academic Press, New York, 1967.

[AO-2] Aoki, M., "On Some Convergence Questions in Bayesian Optimization Problems," IEEE Trans. on Automatic Control, Vol. 10, No. 2, pp. 180-182, April 1965.

[ALB] Albert, A., "Nonlinear Regression and Stochastic Approximation," IEEE Conv. Rec'd, 1967.

[ARN] Arnold, C. R. and K. Narendra, "The Characterization and Identification of Systems," Tech. Rept. No. 471, Cruft Lab., Harvard Univ., Cambridge, Mass., June 18, 1965.

[AT] Athans, M. and P. Falb, Optimal Control, McGraw-Hill Book Co., Inc., New York, 1966.

[AY] Ayer, Alfred J., "The A PRIORI," in Philosophy of Mathematics: Selected Readings, p. 289-301, Eds. P. Benecerraf and H. Putman, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1964.

[BAL-1] Balakrishnan, A. V., "Identification of Control Systems from Input-Output Data," Proc. IFAC Conf. on Identification in Automatic Control Systems, Prague, Czechoslovakia, June 1967. Also in SIAM J. Control, Vol. 5, No. 3, Aug. 1967.

[BAL-2] Balakrishnan, A. V., "Determination of Nonlinear Systems from Input-Output Data," Proc. Princeton Conf. on Identification Problems in Communication and Control, Mar. 1963.

[BEL-1] Bellman, R., B. Gluss, and R. Roth, "Identification of Differential Systems with Time Varying Coefficients," RAND Corp., RM-4288-PR, Nov. 1964.

[BEL-2] Bellman, R., R. Kalaba, and R. Sridhar, "Adaptive Control via Quasilinearization and Differential Approximation," RAND Corp., RM-3928-PR, Nov. 1963.

[BLU] Blum, J. R., "Approximation Methods which Converge with Probability One," Annals. Math. Stat., Vol. 25, pp. 382-386, 1954.

[BX] Boxer, R. and S. Thaler, "A Simplified Method of Solving Linear and Nonlinear Systems," Proc. IRE, Jan. 1956.

[CAR] Carnap, R., Logical Foundations of Probability, Chicago, Univ. of Chicago Press, 1950.

[CH] Chen, C. F. and B. L. Philips, "Accurate Determination of Complex Root Transfer Functions from Frequency Response Data," IEEE Trans. on Automatic Control, AC-10, No. 3, p. 356, July 1965.

[CHU] Chung, K. L., "On a Stochastic Approximation Method," Annals of Math. Stat., Vol. 25, pp. 463-483, 1954.

[CU-1] Cuenod, M. and A. P. Sage, "Comparison of Some Methods Used for Process Identification in Automatic Control Systems, Proc. IFAC Symposium on Identification in Automatic Control Systems, Prague, June 1967.

[CU-2] Cuenod, M. and A. E. Durling, An Introduction to Impulse Analysis, Academic Press, New York 1967.

[DA] Davenport, W. B. and W. L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill Book Co., Inc., New York, 1958.

[DAL] Daly, R. F., "The Adaptive Binary Detection Problem on the Real Line," Stanford Electron Lab. Rept., TR 2003-3, Feb. 1962.

[DE] Deutsch, R., Estimation Theory, Prentice-Hall, Englewood Cliffs, New Jersey, 1965.

[DU] Dubes, R. C., Theory of Applied Probability, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

[DV] Dvoretzky, A., "On Stochastic Approximation," Proc. Third Berkeley Symposium on Mathematical Statistics and Probability I, pp. 39-55, Univ. of Cal. Press., Berkeley, Cal., 1956.

[EY] Eykhoff, P., "Process Parameter and State Estimation," Proc. IFAC Symposium on Identification in Automatic Control Systems, Prague, 1967.

[FU-1] Fu, K. S., "Learning Control Systems," Computer and Information Sciences, pp. 318-343, J. T. Tou and R. H. Wilcox, eds., Spartan Books, Washington, D. C., 1964.

[FU-2] Fu, K. S. and K. Kirvaitis, "Identification of Nonlinear Systems by Stochastic Approximation," Proc. 1966 JACC, pp. 255-264.

[FK-1] Fukao, T., "Some Fundamental Properties of Adaptive Control Processes (I), "Bull. of Electrotech. Lab., Vol. 28, No. 1, pp. 1-19, Jan. 1964.

[FK-2] Fukao, T., "System Identification by Bayesian Learning," I, II, Bull. Electrotech. Lab. 29, No. 5, p. 364, Tokyo, 1965.

[FK-3] Fukao, T., "System Identification by Bayesian Learning," Proc. Tokyo IFAC Symposium, pp. 137-146, 1965.

[FR] Frame, J. S., "Matrix Functions and Applications," IEEE Spectrum, vol. 1, March-July 1964.

[HAM] Hamming, R. W., Numerical Methods for Scientists and Engineers, McGraw-Hill Book Company, Inc., New York, 1962.

[HAN] Hancock, J. C. and P. A. Wintz, Signal Detection Theory, McGraw-Hill Book Co., Inc., New York, 1966.

[HE] Hempel, C. G., Fundamentals of Concept Formation in Empirical Science, Chicago: Univ. of Chicago Press, 1952.

[HS] Hsieh, H. C., "Least Squares Estimation of Linear and Nonlinear System Weighting Function Matrices," J. Info. and Control, Vol. 7, March 1964. Also in Computing Methods in Optimal Control, A. Balakrishnan, and L. Neustadt, eds., Academic Press, New York, 1966.

[HO-1] Ho, Y. C. and B. H. Whalen, "An Approach to the Identification and Control of Linear Dynamic Systems with Unknown Parameters," IEEE Trans. on Automatic Control, Vol. 8, No. 3, pp. 225-256, July 1963.

[HO-2] Ho, Y. C. and R. C. K. Lee, "Identification of Linear Dynamic Systems," Jrnl. of Information and Control, Vol. 8, pp. 93-110, Feb. 1965.

[HO-3] Ho, Y. C. and R. C. K. Lee, "Identification of Linear Dynamic Systems," Proc. N.E.C., 1964.

[HO-4] Ho, Y. C., "On Optimal Filtering and Stochastic Approximation," J. Math. Analysis and Applications, Vol. 6, No. 1, pp. 152-154, 1963.

[HOU] Householder, A. S., The Theory of Matrices in Numerical Analysis, Blaisdell Publishing Co., New York, 1964.

[IFAC] Proc. of Prague IFAC Conference on Identification in Automatic Control Systems, Prague, Czechoslovakia, June 1967.

[K-1] Kalman, R., "Mathematical Description of Linear Dynamical Systems," S.I.A.M. Control, Vol. 1, No. 2, 1963.

[K-2] Kalman, R., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, ser. D, J. Basic Eng., 82, pp. 35-45, 1960.

[K-3] Kalman, R. and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. ASME ser. D, J. Basic Eng. 83, pp. 95-108, Mar. 1961.

[KAI] Kailath, T., "Correlation Detection of Signals Perturbed by a Random Channel, IRE Trans., Vol. IT-6, No. 3, pp. 361-366, June 1960.

[KIE] Kiefer, J. and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," Annals. Math. Stat., Vol. 23, pp. 462-466, 1952.

[KIR] Kirvaitis, K., "Identification of Nonlinear Systems by Stochastic Approximation," Ph.D. Thesis, Purdue Univ., Aug. 1965.

[LAN] Lanning, J. H. and R. H. Battin, Random Processes in Automatic Control, McGraw-Hill Book Co., Inc., New York, 1956.

[LE-1] Lee, Y. W. and M. Schetzen, "Measurement of the Kernels of a Nonlinear System by Cross Correlation," Int. Jrnl. of Control, 1965.

[LE-2] Lee, Y. W. and M. Schetzen, "Some Aspects of the Wiener Theory of Nonlinear Systems," Proc. N.E.C., 1965.

[LEE] Lee, R. C. K., Optimal Estimation, Identification and Control, Research Monograph, No. 28, M.I.T. Press, Cambridge, Mass., 1964.

[LEV] Levin, M. J., "Estimation of a System Pulse Transfer Function in the Presence of Noise," IEEE Trans. on Automatic Control, AC-9, No. 3, July 1964.

[MA] Mayne, D. Q., "Optimal Non-Stationary Estimation of the Parameters of a Linear System with Gaussian Inputs," J. Electr. Control, pp. 101-112, Jan. 1963.

[MG] McGhee, R. B., "Identification of Non-Linear Dynamic Systems by Regression Analysis Methods," Ph.D. Thesis, Univ. of So. Cal., June 1963.

[POP] Popper, K. R., The Logic of Scientific Discovery, Routeledge and Kegan Paul, London, 1951.

[PAP] Papoulis, A., Probability, Random Variables and Stochastic Processes, McGraw-Hill Book Co., Inc., New York, 1965.

[PU] Puri, N. W. and C. N. Weygandt, "Transfer Function Tracking of a Linear Time-Varying System by Means of Auxiliary Simple Lag Networks," IEEE Trans. on Appl. and Ind., vol. 83, pp. 70-72, Jan. 1964.

[Q] Quine, W. V., "On What There Is," Philosophy of Mathematics: Selected Readings, pp. 183-196, P. Benecerraf and H. Putnam eds., Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

[RAU] Rauch, Tung, and Striebel, "On Maximum Likelihood Estimation of Linear Systems," Lockheed Missile and Space Co. Tech. Rpt., June 5, 1963.

[ROB] Robbins, H. and S. Monro, "A Stochastic Approximation Method," Annals. of Math. Stat., Vol. 22, pp. 400-407, 1951.

[ROS] Rosen, J. B., "Stability and Bounds for Nonlinear Systems of Difference and Differential Equations," Jrnl. Math. Anal. and Appl., Vol. 2, pp. 370-393, 1961.

[SAK-1] Sakrison, D. J., "The Use of Stochastic Approximation to Solve the System Identification Problem," IEEE Trans. on Automatic Control, AC-12, No. 5, pp. 563-567, Oct. 1967.

[SAK-2] Sakrison, D. J., "Stochastic Approximations: A Recursive Method for Solving Regression Problems," Advances in Communication Systems, Vol. 2, 1966., Academic Press, New York, 1966.

[SAR] Saridis, G. N. and G. Stein, "Stochastic Approximation Algorithms for Linear Discrete-Time System Identification," Proc. NEC, 1967.

[SC] Schreider, Methods of Statistical Testing, Elsevier Publishing Co., New York, 1964.

[SCH] Scheffler, I., The Anatomy of Inquiry: Philosophical Studies in the Theory of Science, A. Knopf Publ. Co., New York, 1963.

[SG-1] Sage, A. B. and W. C. Choate, "Minimum Time Identification of Non-Stationary Dynamic Processes," Proc. N.E.C., 1965.

[SG-2] Sage, A. P. and B. R. Eisenberg, "Experiments in Nonlinear and Non-Stationary System Identification via Quasilinearization and Differential Approximation," Proc. J.A.C.C., 1966.

[SK] Sklansky, J., "Learning Systems for Automatic Control," IEEE Trans. on Automatic Control, Vol. AC-11, No. 1, Jan. 1966.

[SMI] Smith. B. W., "Parameter Estimation in the Presence of Measurement Noise," Int. Jrnl. of Control, Vol. 3, No. 4, pp. 297-312, 1966.

[SP] Spragins, J. D., "A Note on the Iterative Application of Bayes' Rule," IEEE Trans. on Info. Theory, IT-11, No. 4, pp. 544-549, Oct. 1965.

[STI] Steiglitz, and Mcbridge, "A Technique for Identification of Linear Systems," IEEE Trans. Automatic Control, AC-10, Oct. 1965.

[TSY] Tsypkin, Y. Z., "Adaptation, Learning and Selflearning in Control Systems," Third IFAC Congress, Buttenworths, London, June 1966.

[WE-1] Wiener, N., The Extrapolation, Interpolation and Smoothing of Time Series, John Wiley and Sons, New York, 1949.

[WE-2] Wiener, N., Nonlinear Problems in Random Theory, John Wiley and Sons, New York, 1958.

[W] Wilde, D. J., Optimum Seeking Methods, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

[WOL] Wolverton, C. T., and J. T. Rawgen," A Counter-example to Dvortetzky's Stochastic Approximation Theorem," I.E.E.E. Trans., IT-14, No. 1, pp. 157-158, Jan. 1968.

[Z-1] Zadeh, L. A., "From Circuit Theory to System Theory," Proc. IRE, 50, pp. 856-865, May 1962.

[Z-2] Zadeh, L. A. and C. A. Desoer, Linear System Theory: The State Space Approach, McGraw-Hill Book Co., Inc., New York, 1963.