Pseudo-Linear Identification:
Optimal Joint Parameter and State Estimation
# of Linear Stochastic MIMO Systems

by

Mark A. Hopkins

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Electrical Engineering

APPROVED:

H. F. VanLandingham, Chairman

A. A. (Louis) Beex

M. V. Day

D. W. Luse

K.-B. Yu

March 24, 1988

Blacksburg, Virginia

Pseudo-Linear Identification:

Optimal Joint Parameter and State Estimation

of Linear Stochastic MIMO Systems

by

Mark A. Hopkins

H. F. VanLandingham, Chairman

Electrical Engineering

(ABSTRACT)

This dissertation presents a new method of simultaneous parameter and state estimation for linear, stochastic, discrete-time, multiple-input, multiple-output (MIMO) systems. This new method is called *pseudo-linear identification* (*PLID*), and extends an earlier method to the more general case where system input and output measurements are corrupted by noise. PLID can be applied to completely observable, completely controllable systems with known structure (*i.e.*, known observability indexes) and unknown parameters. No assumptions on pole and zero locations are required; and no assumptions on relative degree are required, except that the system transfer functions must be strictly proper.

Under standard gaussian assumptions on the various noises, for time-invariant systems in the class described above, it is proved that PLID is the optimal estimator (in the mean-square-error sense) of the states and the parameters, conditioned on the output measurements. It is also proved, under a reasonable assumption of persistent excitation, that the PLID parameter estimates converge a.e. to the true parameter values of the unknown system.

For deterministic systems, it is proved that PLID exactly identifies the states and parameters in the minimum possible time, so-called deadbeat identification. The proof

brings out an interesting relation between the estimate error propagation and the observability matrix of the time-varying *extended* system (the extended system incorporates the unknown parameters into the state vector). This relation gives rise to an intuitively appealing notion of persistent excitation.

Some results of system identification simulations are presented. Several different cases are simulated, including a two-input, two-output system with non-minimum-phase zeros, and an unstable system. A comparison of PLID with the widely used extended Kalman filter is presented for a single-input, single-output system with near cancellation of a pole-zero pair.

Results are also presented from simulations of the adaptive control of an unstable, two-input, two-output system. In these simulations, PLID is used in a self-tuning regulator to identify the parameters needed to compute the feedback gain matrix, and (simultaneously) to estimate the system states, for the state feedback.

Acknowledgements

I owe my deepest thanks to my advisor, Professor Hugh F. VanLandingham, for providing the initial idea for this dissertation, and for his interest in and support of this work throughout the process of developing it. It has been a great personal pleasure to work and travel with Dr. VanLandingham, whose equanimity is a natural wonder, and who can surely pun with the best.

To my wife, , I offer inexpressible gratitude and admiration for accepting and performing a task at least equal to my own, namely, loving me, guiding me, and supporting me through this roller-coaster ride, all with a wonderful sense of humor and an astonishingly consistent perspective.

I would be remiss if I failed to thank Dr. D. William Luse for the many hours of illuminating conversation, the uncountable helpful suggestions, and, most important to me, the friendship which he freely gave.

I want to thank the other members of my committee, Dr. A. A. (Louis) Beex, Dr. Martin Day, and Dr. Kai-Bor Yu, who graciously and generously gave of their time, their resources, and their expertise whenever I required it.

I also want to thank Dr. Dan Hodge for channeling financial support to me, all through my years at Virginia Tech, which made life as a graduate student considerably easier; and for his willingness to spend time advising me about a career in academics, giving me very valuable advice, indeed.

Table of Contents

List of Illustrations

1.0 Introduction

As the algorithm developed in this thesis is, essentially, another least-squares method, it is fitting to recapitulate a bit of the history of this type of method. A least-squares method is used to identify parameters of a mathematical model of a system, based on measurements of the system outputs, and, if available, measurements of the system inputs. Karl F. Gauss was probably the first to employ a least-squares calculation, and, because of his clear understanding of what he was doing, must surely be given credit for initiating the entire area of investigation [1].

In 1795, Gauss developed the least-squares method which, based upon a large number of observations, he used to estimate the six parameters needed to completely describe planetary motion. Legendre independently developed the least-squares method, publishing his results in 1806 in his book, *Nouvelles méthodes pour la determination des orbites des comètes.*

Some ill-feelings arose between Gauss and Legendre, because Gauss, who did not publish his results until 1809, claimed to be the first to develop the method, by virtue of the earlier date of his work, and gave only passing reference to Legendre's publication. Historians have substantiated Gauss' claim, and the furor surrounding the original development of the least-squares method has largely been forgotten.

Gauss stated the motivation for the least-squares method quite clearly:

> "If the astronomical observations and other quantities on which the computation of orbits is based were absolutely correct, the elements also, whether deduced from three or four observations, would be strictly accurate (so far indeed as the motion is supposed to take place exactly according to the laws of Kepler) and, therefore, if other observations were used, they might be confirmed but not corrected. But since all our measurements and observations are nothing more than approximations to the truth, the same must be true of all calculations resting

> upon them, and the highest aim of all computations made concerning concrete phenomena must be to approximate, as nearly as practicable, to the truth. But this can be accomplished in no other way than by a suitable combination of more observations than the number absolutely requisite for the determination of the unknown quantities. This problem can only be properly undertaken when an approximate knowledge of the orbit has been already attained, which is afterwards to be corrected so as to satisfy all the observations in the most accurate manner possible." [2]

This is a remarkable quote, not merely because of its clarity, but also because it touches upon several features of least-squares estimation, and of modelling, that have kept researchers busy for nearly 200 years since. For example, he refers to the minimum number of observations necessary to identify the unknown parameters, presaging the questions of the *model order* and the *observability of a system*, and, less directly, the question of *persistency of excitation* of the input.

Gauss clearly saw that redundant measurements could neutralize the effects of error in individual measurements. He indicated the method should yield a model that would agree as well as possible with *all* the measurements, a notion which today would be called *minimizing the residuals.*

Gauss chose not to pursue the computational aspects of the least-squares method beyond what was needed for solving the planetary motion problem. The method of least-squares typically involves an enormous computational burden, which Gauss surely realized, making it very difficult to implement before the advent of computers. Even with computers, it was not until 1960, with the landmark work of Kalman [3], in which a recursive algorithm was described, that the least-squares problem became computationally efficient.

Kalman provided an algorithm that recursively computes the conditional mean estimate of the state of a Gauss-Markov system; it is easy to show that the conditional mean estimate is also the least-square-error estimate. Kalman's work was, in a sense, the culmination of 165 years of research into least-square methods.

In the area of least-squares estimation, it could be argued that there is really nothing to add to Kalman's work, only applications of it. This could explain the ubiquitousness of the Kalman filter. However, one might just as well say that Kalman's work was simply an implementation of the ideas of Gauss. But that is not totally accurate. Gauss described a method of estimating the *parameters* of a system model; on the other hand, Kalman's algorithm assumes the system model is completely known *a priori*, and estimates the *states* of the system.

Of course, the Kalman state estimator has been studied extensively. Some investigators have studied what becomes of the state estimates when the model structure or the model parameters are not perfectly known, such as in [39] and [40]. But, more in line with the problem Gauss studied, many attempts have been made to use the Kalman filter to estimate the parameters of a system model, such as the straightforward least-squares method described by Åström [33].

While state estimators have been used for years in feedback control systems, it is only recently, and still seldom, that parameter estimators have been employed in controllers, in what is generally called *adaptive* control. This state of affairs is partly due to a natural inertia among control system designers, but mostly due to the fact that so few theoretical results are available concerning the incorporation of parameter estimators into the control loop. Typically, for systems whose structure or whose parameters are changing, control designers are more likely to incorporate gain scheduling schemes, which are simple, and have performed adequately in a variety of complex systems since the 1940's. While gain scheduling is a type of adaptive control, adaptive control purists consider it to be "open-loop" adaptive control.

With the advent of "closed-loop" adaptive controllers, some of which require both state and parameter estimates, the natural tendency is to extend the formulation of the Kalman filter, so that it will estimate both the states of the system and the parameters of the system model, simultaneously. This has led to various schemes intended to simultaneously estimate the system parameters and the system states. These schemes are known under such names as *adaptive state estimation, adaptive prediction, joint parameter and state estimation*, and *simultaneous parameter and state estimation.*

There are two main approaches to joint parameter and state estimation of linear systems [4]. In the first, a "bootstrap" arrangement is employed, in which there are separate state and parameter estimators. The most recent state estimates are used in the parameter estimator, and vice versa. Proofs of convergence for this arrangement usually invoke the *separation principle*, which says that state estimation and parameter estimation can be carried out independently. Thus the choice of parameter estimator is independent of the state estimator being used; and there are many from which to choose. Chapter 2 outlines several of the most commonly used methods of parameter estimation. Typically, the states are estimated by a Kalman filter.

In the second main approach, the system state vector is augmented by the vector of unknown parameters. This results in an augmented system that is nonlinear, due to the multiplication of states by parameters, which are now other states of the system, by virtue of their inclusion in the state vector. The augmented state vector is estimated by some nonlinear estimator, or by an Extended Kalman Filter (EKF), which performs a linearization of the system about the estimated state, at each iteration. This approach is somewhat less computationally efficient than the first one, because the matrices in the augmented state estimator are significantly larger than those in the independent parameter and state estimators of the first approach. A typical development of the EKF is presented in Chapter 3.

This thesis presents the development of a joint parameter and state estimator called pseudo-linear identification (PLID). The development is carried out for discrete-time, linear, multiple-input, multiple output (MIMO) stochastic systems. The systems may be stochastic in the following sense: Precise measurements of the input and output are not available; rather, the measurements are corrupted by gaussian white noise of known autocovariance. Furthermore, gaussian white noise is input directly to the states of the system, again with known autocovariance. All the cross-covariances are assumed known, as well.

The PLID algorithm is an extension of earlier work by Salut *et.al* [5]. The work in [5] is actually an extension of Salut's PhD dissertation, which is summarized in [6], presenting a unified treatment of continuous-time and discrete-time systems. In [5], a method of simultaneous parameter and state estimation is presented for linear MIMO systems having a white gaussian state-noise vector with known autocovariance, and for which the input and output are known perfectly. There it is shown how the observable-canonic form can lead to a linear minimum-mean-square-error estimator of the parameters and states of the unknown system, *provided that inputs and outputs are measured exactly.* No theoretical study of the convergence of the joint estimator was presented in either of Salut's publications.

Chen *et.al.* [7] drew heavily on the work in [6] and [5], extending it to the case where the state-noise vector has ***unknown*** autocovariance (similar to the assumptions made in the development of the Generalized Least Squares method of parameter estimation). This leads to a nonlinear representation, with the nonlinearities due to the fact that states are multiplying states. The system is linearized in a typical extended Kalman filter development, yielding a suboptimal filter. A convergence proof is presented, in which it is shown that the parameter subvector (of the joint parameter and state estimate) converges with probability 1 (w.p.1) to a stationary point of a Lyapunov function. Such a stationary point is not necessarily equivalent to the true parameter vector.

The development of the PLID algorithm starts with the observable canonic form, from which is derived a rather interesting extended system form, apparently first noted by Salut [6], and later, independently, by VanLandingham [8]. In deference to Salut's earlier work, this extended canonic form here will be called the Salut form. More generally, this Salut canonic form can be called the *extended state representation.*

The work presented in this thesis extends the work of Salut *et.al.* [5] to the case where, in addition to the state-noise vector of known autocovariance, the input vector and the output vector are also contaminated by gaussian white noise vectors with known autocovariances. The input, output, and state-noise vectors may be cross-correlated, but it is assumed that such cross-covariances are known. Under these assumptions, the conditional mean estimator (PLID) is derived, with conditioning on all the input and output measurements up to the current time. The conditional mean estimator is, of course, optimal in the mean-square-error sense. It is also shown here that the algorithm converges to the exact values of the states and parameters in the minimum possible time whenever the system is completely deterministic. In the stochastic case, martingale theory is used to show that the PLID parameter subvector converges w.p.1 to the exact value of the unknown parameter vector.

The development of the Salut form from the observable-canonic form is shown in Chapter 4 both for single-input, single-output (SISO) systems and for multiple-input, multiple-output (MIMO) systems.

Chapter 5 develops the (stochastic) PLID algorithm from the MIMO Salut form of the previous chapter. By derivation, PLID is the conditional mean estimator (conditioned on input and output data up to the current time), implying that, as an estimator of the states and parameters, PLID is optimal in the mean-square-error sense.

In Chapter 6 the *deterministic* pseudo-linear identification is presented. The algorithm is obtained by setting all the noise terms in the (stochastic) PLID algorithm to zero; therefore, it is actually a subset of the work in [5]. However, Chapter 6 presents a proof of time-optimal convergence that has not been found in the literature. An interesting relationship between the input signal and the convergence of the algorithm is also highlighted there, embodied in Theorem 6.2.5, and Corollaries 6.2.7 and 6.2.8. This relationship can be considered as a *requirement of persistent excitation*, which is a recurring topic in the literature on system identification.

Chapter 7 completes the theoretical convergence study, presenting proof that, under a mild persistency of excitation assumption, the stochastic PLID algorithm converges to the correct parameter values, drawing upon theorems from the study of martingales. (Martingales and the associated definitions in probability theory, are reviewed in the Appendices.)

The convergence proofs of the previous chapters no longer hold when the system parameters vary with time in an unknown way. The theoretical difficulties introduced by time-varying systems are discussed in Chapter 8. However, Section 8.1 presents an analysis of the initial convergence rate of the stochastic PLID algorithm for the small noise case. This analysis makes no assumption about time-invariance of the system, so applies generally. Thus it gives a useful expected bound, at least for SISO systems of reasonably low order, on the parameter and state estimate errors for the small noise case, when the input is persistently exciting.

Chapter 9 presents a variety of PLID simulation results. In Section 9.1, typical results are shown at various noise levels, for a fourth-order, two-input, two-output system. The particular system chosen for these simulations has non-minimum-phase zeros in three out of four subsystems. Because there are no assumptions about the system poles and zeros in the derivation, non-minimum-phase zeros should present no special problem to the PLID algorithm. Clearly, from the simulations, they do not.

Section 9.2 presents results when the assumptions under which the PLID algorithm was derived are violated. In this case, the system is not strictly proper. The estimator was rederived to account for an extra system parameter, which also introduces explicit nonlinearities into the estimator. Thus, the estimator of this section is a slightly altered (and suboptimal) version of the PLID algorithm. This suboptimal variation on the PLID algorithm is compared to a more standard method of joint parameter and state estimation, the Extended Kalman Filter. The two methods were applied to a near-prime system, which has near cancellation of a pole-zero pair. The results indicate that the nice convergence properties of the PLID algorithm carry over to cases that slightly violate the basic assumptions of the PLID derivation.

Section 9.3 presents results for an unstable system, highlighting some of the difficulties introduced if the output becomes unbounded.

Chapter 10 discusses using the PLID algorithm in an adaptive control scheme. Simulation results of such a scheme are presented in Section 10.1, where PLID is used together with a minimum variance set point controller to control a fourth-order, two-input, two-output system having two unstable poles and several non-minimum-phase zeros. Some interesting aspects of the problems of ill-conditioning and persistency of excitation, as they affect the PLID algorithm, are highlighted in these results.

2.0 Previous Methods of Parameter Estimation

Over the years, many methods have been devised for parameter estimation, and a great deal of work has been done developing proofs of their convergence properties. In this chapter several common methods of parameter estimation are briefly described, for on-line estimation of parameters for single-input, single-output (SISO), linear, lumped-parameter, time-invariant, discrete-time systems of known order. The chapter is organized as follows:

2.1. Method of Weighted Least Squares (WLS)

2.2. Method of Stochastic Approximation (SA)
and the Stochastic Newton Method

2.3. Bayesian approach

2.4. Method of Instrumental Variables (IV)

2.5. Method of Extended Least Squares (ELS)
(or Pseudo-Linear Regression (PLR))

2.6. Method of Generalized Least Squares (GLS)

2.7. Model Reference Technique (MR)

2.8. Recursive Maximum Likelihood (RML) Method

For more complete treatments of these methods, see [9] and [10].

2.1 Method of Weighted Least Squares (WLS)

Suppose a system with input sequence $\{u_k\}$ and output sequence $\{y_k\}$ can be represented by the model

$$y_k + a_1 y_{k-1} + \cdots + a_n y_{k-n} = b_1 u_{k-1} + \cdots + b_n u_{k-n} + v_k \tag{2.1.1}$$

where v_k is an unknown noise sequence (measurement noise).

Let the vectors

$$\begin{aligned} \theta &\triangleq [a_1, \dots, a_n, b_1, \dots, b_n]^T \\ \phi_k &\triangleq [-y_{k-1}, \dots, -y_{k-n}, u_{k-1}, \dots, u_{k-n}]^T \end{aligned} \tag{2.1.2}$$

be defined, with θ being the parameter vector, and ϕ_k being the observation data vector.

Then $$y_k = \theta^T \phi_k + v_k \tag{2.1.3}$$

Now denote the vector parameter estimate $\tilde{\theta}_k$, the estimate based on all data to time k. The criterion for judging the estimate is the weighted output error, or "equation error,"

$$V(\tilde{\theta}_k, k) = \frac{1}{k} \sum_{l=1}^{k} \alpha_l (y_l - \tilde{\theta}_k^T \phi_l)^2 \tag{2.1.4}$$

where α_l are the "weights."

The estimate $\tilde{\theta}_k$ is to be selected so as to minimize $V(\tilde{\theta}_k, k)$. Denoting the minimizing estimate by $\hat{\theta}_k$, then

$$\left[\frac{\partial}{\partial \tilde{\theta}_k} V(\tilde{\theta}_k, k)\right]_{\tilde{\theta}_k = \hat{\theta}_k} = 0. \tag{2.1.5}$$

Hence,

$$0 = \frac{1}{k} \sum_{i=1}^{k} \alpha_i \left[2(y_i - \hat{\theta}_k^T \phi_i)(-\phi_i^T) \right], \tag{2.1.6}$$

$$\sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \hat{\theta}_k = \sum_{i=1}^{k} \alpha_i y_i \phi_i , \tag{2.1.7}$$

and, if the inverse exists,

$$\hat{\theta}_k = \left[\sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \right]^{-1} \sum_{i=1}^{k} \alpha_i y_i \phi_i \tag{2.1.8}$$

To recast Equation 2.1.8 as a recursive algorithm, define

$$M_k \triangleq \sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \tag{2.1.9}$$

Hence,

$$M_{k-1} = M_k - \alpha_k \phi_k \phi_k^T . \tag{2.1.10}$$

With that definition, Equation 2.1.7 becomes

$$M_k \hat{\theta}_k = \sum_{i=1}^{k} \alpha_i y_i \phi_i \tag{2.1.11}$$

and from Equation 2.1.8,

$$\begin{aligned}\hat{\theta}_k &= M_k^{-1} \sum_{l=1}^{k} \alpha_l y_l \phi_l \quad = \quad M_k^{-1} \left(\sum_{l=1}^{k-1} \alpha_l y_l \phi_l + \alpha_k y_k \phi_k\right) \\ &= M_k^{-1} \left(M_{k-1} \hat{\theta}_{k-1} + \alpha_k y_k \phi_k\right) \\ &= M_k^{-1} \left[(M_k - \alpha_k \phi_k \phi_k^T)\hat{\theta}_{k-1} + \alpha_k y_k \phi_k\right] \\ &= M_k^{-1} \left[M_k \hat{\theta}_{k-1} + \alpha_k \phi_k (y_k - \phi_k^T \hat{\theta}_{k-1})\right] \\ &= \hat{\theta}_{k-1} + \alpha_k M_k^{-1} \phi_k (y_k - \phi_k \hat{\theta}_{k-1})\end{aligned} \tag{2.1.12}$$

Now, to avoid having to find the inverse of M_k on each iteration, it is useful to apply the matrix inversion lemma. Define $P_{k+1} \triangleq M_{k+1}^{-1}$, for all $k \geq 0$. So,

$$\begin{aligned} M_{k+1} &= M_k + \alpha_{k+1} \phi_{k+1} \phi_{k+1}^T \\ &\Rightarrow \quad P_{k+1}^{-1} = P_k^{-1} + \phi_{k+1} \left(\frac{1}{\alpha_{k+1}}\right)^{-1} \phi_{k+1}^T \end{aligned} \tag{2.1.13}$$

Applying the matrix inversion lemma to Equation 2.1.13 yields

$$P_{k+1} = P_k - \frac{P_k \phi_{k+1} \phi_{k+1}^T P_k}{\frac{1}{\alpha_{k+1}} + \phi_{k+1}^T P_k \phi_{k+1}} \tag{2.1.14}$$

So, from Equation 2.1.12, the recursion on the estimate is

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \alpha_{k+1} P_{k+1} \phi_{k+1} (y_{k+1} - \phi_{k+1}^T \hat{\theta}_k) \tag{2.1.15}$$

Equations 2.1.14 and 2.1.15 define the recursive weighted least squares algorithm. Note that if $\alpha_i = 1$, for all $i = 1, 2, \ldots$, then the equations reduce to the least squares algorithm.

Strictly speaking, the recursion should begin at the time when M_k first becomes invertible, since the derivation of the recursion required it. However, the recursion is usually begun at time zero, by assuming $P_0 = c\,I$, where $c \gg 1$, and some arbitrary $\hat{\theta}_0$. To see the effect of this, denote the new values in the recursion by P_k^*, M_k^*, and $\hat{\theta}_k^*$. Then

$$M_k^* = P_0^{-1} + \sum_{l=1}^{k} \alpha_l \phi_l \phi_l^T = P_0^{-1} + M_k \tag{2.1.16}$$

Equation 2.1.16 can be proven by induction:

$$\begin{array}{lll}
\text{Let} & M_1^* = M_0^* + \alpha_1 \phi_1 \phi_1^T & = P_0^{-1} + M_1 . \\
\text{Assume} & M_{k-1}^* = P_0^{-1} + M_{k-1} . & \\
\text{Then} & M_k^* = M_{k-1}^* + \alpha_k \phi_k \phi_k^T & = P_0^{-1} + M_{k-1} + \alpha_k \phi_k \phi_k^T \\
 & = P_0^{-1} + M_k . &
\end{array}$$

Similarly, $\hat{\theta}_k^* = (P_0^{-1} + M_k)^{-1} \; (P_0^{-1} \hat{\theta}_0 + \sum_{l=1}^{k} \alpha_l \phi_l y_l)$ (2.1.17)

can be shown by induction:

$$\begin{aligned}
\hat{\theta}_1^* &= \hat{\theta}_0 + \alpha_1 P_1^* \phi_1 (y_1 - \phi_1^T \hat{\theta}_0) \\
&= P_1^* \left[(P_1^{*\,-1} - \alpha_1 \phi_1 \phi_1^T) \; \hat{\theta}_0 + \alpha_1 \phi_1 y_1 \right] \\
&= M_1^{*\,-1} \; \left[(M_1^* - \alpha_1 \phi_1 \phi_1^T) \; \hat{\theta}_0 + \alpha_1 \phi_1 y_1 \right] \\
&= (P_0^{-1} + M_1)^{-1} \left[(P_0^{-1} + M_1 - \alpha_1 \phi_1 \phi_1^T) \; \hat{\theta}_0 + \alpha_1 \phi_1 y_1 \right].
\end{aligned}$$

But $M_1 = \alpha_1 \phi_1 \phi_1^T$, so

$$\hat{\theta}_1^* = (P_0^{-1} + M_1)^{-1} \; (P_0^{-1} \hat{\theta}_0 + \alpha_1 \phi_1 y_1).$$

Assume $\hat{\theta}^*_{k-1} = (P_0^{-1} + M_{k-1})^{-1} \quad (P_0^{-1}\,\hat{\theta}_0 + \sum_{l=1}^{k-1} \alpha_l\,\phi_l\,y_l)$

$$\begin{aligned}
\text{Then}\quad \hat{\theta}^*_k &= \hat{\theta}^*_{k-1} + \alpha_k\,P^*_k\,\phi_k\,(y_k - \phi_k^T\,\hat{\theta}^*_{k-1}) \\
&= P^*_k\left[(P^{*\,-1}_k - \alpha_k\,\phi_k\,\phi_k^T)\ \hat{\theta}^*_{k-1} + \alpha_k\,\phi_k\,y_k\right] \\
&= M^{*\,-1}_k\ \left[(M^*_k - \alpha_k\,\phi_k\,\phi_k^T)\ \hat{\theta}_{k-1} + \alpha_k\,\phi_k\,y_k\right] \\
&= (P_0^{-1} + M_k)^{-1}\left[(P_0^{-1} + M_k - \alpha_k\,\phi_k\,\phi_k^T)\ \hat{\theta}^*_{k-1} + \alpha_k\,\phi_k\,y_k\right] \\
&= (P_0^{-1} + M_k)^{-1}\Bigg[(P_0^{-1} + M_{k-1})(P_0^{-1} + M_{k-1})^{-1}\{P_0^{-1}\,\hat{\theta}_0 + \sum_{l=1}^{k-1}\alpha_l\,\phi_l\,y_l\} \\
&\qquad\qquad + \alpha_k\,\phi_k\,y_k\Bigg] \\
&= (P_0^{-1} + M_k)^{-1}\left[P_0^{-1}\,\hat{\theta}_0 + \sum_{i=1}^{k-1}\alpha_l\,\phi_i\,y_l + \alpha_k\,\phi_k\,y_k\right] \\
&= (P_0^{-1} + M_k)^{-1}\left[P_0^{-1}\,\hat{\theta}_0 + \sum_{l=1}^{k}\alpha_l\,\phi_l\,y_l\right]
\end{aligned}$$

Now if $P_0 = c\,I$, $c \gg 1$, then because $M_k = \sum_{i=1}^{k} \alpha_i\,\phi_i\,\phi_i^T$ if $\alpha_i \neq 0$, $\forall\, i = 1, 2, \ldots$, and if the input sequence $\{u_k\}$ is persistently exciting (so that $\phi_k\,\phi_k^T$ does not go to zero), then the effect of P_0^{-1} decreases over time. So

$$\lim_{k\to\infty} M^*_k = M_k, \quad \text{and} \quad \lim_{k\to\infty} P^*_k = P_k. \tag{2.1.18}$$

Thus,

$$\hat{\theta}^*_k = P^*_k\,P_0^{-1}\,\hat{\theta}_0 + P^*_k \sum_{l=1}^{k} \alpha_l\,\phi_l\,y_l \tag{2.1.19}$$

$$\lim_{k\to\infty} \hat{\theta}^*_k = P_k\,P_0^{-1}\,\hat{\theta}_0 + \hat{\theta}_k \tag{2.1.20}$$

But $P_k \to 0$ and $P_0^{-1} \ll 1$. So, $\lim\limits_{k \to \infty} \hat{\theta}_k^* = \hat{\theta}_k$. Therefore the effects of the initial values $P_{0|-1}$ and $\hat{\theta}_0$ diminish over time.

To investigate the convergence properties of the least squares method, it is necessary to assume that the system generating the data really can be modeled by

$$y_k = \theta_0^T \phi_k + v_k . \tag{2.1.21}$$

Thus, Equation 2.1.8 becomes

$$\begin{aligned}
\hat{\theta}_k &= \left[\sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \right]^{-1} \sum_{i=1}^{k} \alpha_i \phi_i (\theta_0^T \phi_i + v_i) \\
&= \left[\sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \right]^{-1} \left[\sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \theta_0 + \alpha_i \phi_i v_i \right] \\
&= \theta_0 + \left[\sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \right]^{-1} \sum_{i=1}^{k} \alpha_i \phi_i v_i \\
&= \theta_0 + \left[\frac{1}{k} \sum_{i=1}^{k} \alpha_i \phi_i \phi_i^T \right]^{-1} \left[\frac{1}{k} \sum_{i=1}^{k} \alpha_i \phi_i v_i \right] .
\end{aligned} \tag{2.1.22}$$

As k gets large, it is hoped that the trailing term of Equation 2.1.22 will vanish. In that term, both factors are weighted sample means. Thus, the second factor can only vanish if ϕ_i and v_i are uncorrelated. That will not be true if $\{v_k\}$ is not a white sequence.

So the weighted least squares method will give biased estimates in the case where $\{v_k\}$ is correlated. There are several approaches to overcoming this problem, some of which are discussed in later sections (ELS, GLS, IV).

2.2 *The Stochastic Approximation Method*

This method was originally proposed by Robbins and Monro in [11]. Their method can be presented as follows:

Suppose there is a sequence of measurements $\{y_k\}$ on a process of x_0,

$$y_k = R_k(x_0) + \xi_k \tag{2.2.1}$$

where $\xi_1, \xi_2, \dots, \xi_n$ are independent, zero-mean random variables. Then taking an arbitrary point $\hat{x}_0$, and an arbitrary sequence of positive numbers (read 'gains') $\{a_k\}$

satisfying $$\sum_{l=1}^{\infty} a_l^{\,2} < \infty \quad \text{and} \quad \sum_{l=1}^{\infty} a_l = \infty, \tag{2.2.2}$$

and letting $$\hat{x}_{k+1} = \hat{x}_k + a_k \left[y_k - R_k(\hat{x}_k) \right] \tag{2.2.3}$$

results in the convergence, $$\lim_{k \to \infty} \hat{x}_k = x_0 \tag{2.2.4}$$

The tendency to converge is easier to see by writing Equation 2.2.3 as

$$\hat{x}_{k+1} - \hat{x}_k = a_k \left[y_k - R_k(\hat{x}_k) \right] \;=\; a_k \left[R_k(x_0) + \xi_k - R_k(\hat{x}_k) \right] \tag{2.2.5}$$

and taking expectation of both sides, given $\hat{x}_k$:

$$E\left[(\hat{x}_{k+1} - \hat{x}_k) \mid \hat{x}_k \right] = a_k \left[R_k(x_0) - R_k(\hat{x}_k) \right]$$

That is, the expected correction of the estimate is in the direction of x_0 . It turns out that the properties of Equation 2.2.2 guarantee convergence, as discussed below.

The requirements of a system identification algorithm make the restrictions on $\{\xi_k\}$ seem a bit stringent. Much work has been done to loosen those restrictions, but the basic form of the algorithm has remained the same. Consider its application to the SISO system of Equation 2.1.3:

$$y_k = \theta^T \phi_k + v_k \tag{2.2.6}$$

where y_k and θ_k are measured and v_k is zero-mean measurement error (uncorrelated with ϕ_k). It is natural to seek the estimate $\tilde{\theta}$ that minimizes the variance of the

"equation error" $$V(\tilde{\theta}, k) \triangleq \frac{1}{2} E\left[(y_k - \tilde{\theta}^T \phi_k)^2\right] \tag{2.2.7}$$

If $\hat{\theta}_k$ is the minimizing value of $\tilde{\theta}$, then $$\left[\frac{\partial}{\partial \tilde{\theta}} V(\tilde{\theta}, k)\right]_{\tilde{\theta} = \hat{\theta}_k} = 0 \tag{2.2.8}$$

which is $$E\left[\phi_k (y_k - \phi_k^T \hat{\theta}_k)\right] = 0 \tag{2.2.9}$$

The solution to Equation 2.2.9 cannot be obtained exactly, since the distribution of $[y_k \mid \phi_k^T]$ is not known. (Although it is interesting to note that if the *sample mean* is substituted into Equation 2.2.9, then

$$\frac{1}{k} \sum_{i=1}^{k} \phi_i (y_i - \phi_i^T \hat{\theta}_k) = 0 \tag{2.2.10}$$

which is the weighted least squares formulation of Equation 2.1.6 with all the weights α_i set to unity.)

So, the stochastic approximation method of Equation 2.2.3 is applied to Equation 2.2.9, yielding

$$\hat{\theta}_k = \hat{\theta}_{k-1} + a_k \phi_k (y_k - \phi_k^T \hat{\theta}_{k-1}) \tag{2.2.11}$$

For a time-invariant system, the positive gains a_k are fairly arbitrary, provided they satisfy Equation 2.2.2. The first part of Equation 2.2.2 implies that $a_k \to 0$, which implies that the effects of measurement noise are eventually eliminated. The second part implies that the initial estimate can be an arbitrary distance from the actual value θ_0, and the estimate will still converge.

The *stochastic Newton method* is a variation on the preceding. Note that Equation 2.2.8 can be rewritten as a stochastic gradient problem,

$$\left[-\frac{\partial}{\partial \tilde{\theta}} V(\tilde{\theta}, k) \right]_{\tilde{\theta} = \hat{\theta}_k} = 0. \tag{2.2.12}$$

The solution is, of course, the same. However, analogous to the deterministic case, the efficiency of the gradient algorithm might be improved by introducing the (approximated) Hessian,

$$\frac{\partial^2}{\partial \tilde{\theta}^2} V(\tilde{\theta}, k) = E(\phi_k \phi_k^T). \tag{2.2.13}$$

Defining the approximation W_k of $E(\phi_k \phi_k^T)$, then the Robbins-Monro method can again be applied to yield

$$W_k = W_{k-1} + b_k (\phi_k \phi_k^T - W_{k-1}). \tag{2.2.14}$$

This is now implemented to yield the stochastic Newton algorithm,

$$\hat{\theta}_k = \hat{\theta}_{k-1} + a_k W_k^{-1} \phi_k (y_k - \phi_k^T \hat{\theta}_{k-1}). \tag{2.2.15}$$

Note the similarity to Equation 2.1.12, which can, in fact, be derived from Equation 2.2.15 by the proper definitions of the gain terms b_k in Equation 2.2.14, and a_k in Equation 2.2.15.

Convergence of the stochastic approximation methods was shown initially for independent, identically distributed random variables ξ_k, but later work showed convergence for identically distributed, possibly correlated ξ_k [12].

2.3 Bayesian Method

In this method, the data are assumed (as in Equation 2.1.3) to be generated by a system which can be modeled as

$$y_k = \theta^T \phi_k + v_k \tag{2.3.1}$$

However, statistics of the measurement noise are *known*; viz., v_k is gaussian with

$$E(v_k) = 0 \text{ , and } \quad E(v_k^2) = R . \tag{2.3.2}$$

Ljung and Söderström provide a detailed proof [13] based on Baye's rule that the gaussian posterior density, $p(\theta_0 \mid y_k, \ldots, y_1, u_k, \ldots, u_1) \triangleq p(\theta_0 \mid Z_k)$ has the following statistics:

$$E(\theta_0 \theta_0^T \mid Z_k) \triangleq P_k = \left[P_{k-1} - \frac{P_{k-1} \phi_k \phi_k^T P_{k-1}}{\phi_k^T P_{k-1} \phi_k + R} \right], \qquad P_0 ; \tag{2.3.3}$$

and $E(\theta_0 \mid Z_k) \triangleq \hat{\theta}_k = \hat{\theta}_{k-1} + \frac{1}{R} P_k \phi_k (y_k - \phi_k^T \hat{\theta}_{k-1}), \qquad \hat{\theta}_0 .$ (2.3.4)

Note that Equations 2.3.3 and 2.3.4 can be obtained from the weighted least squares recursion, Equations 2.1.14 and 2.1.15, by setting the weights $\alpha_i = 1/R$ for all $i = 1, 2, \ldots$.

Equations 2.3.3 and 2.3.4 also bear a close resemblance to Kalman filter equations for the system

$$\begin{aligned} \theta_{k+1} &= \theta_k + w_k \\ y_k &= \phi_k^T \theta_k + v_k \end{aligned} \tag{2.3.5}$$

where w_k and v_k are mutually independent gaussian white noise sequences with zero means and covariances Q and R, respectively. If $w_k \equiv 0$, the Kalman filter can be derived by Bayesian methods similar to that cited above, obtaining equations identical to Equations 2.3.3 and 2.3.4. The same result can also be derived by minimization techniques when first- and second-order noise statistics are known.

The Kalman filter method is thus shown to be applicable to parameter estimation, so Kalman filter stability and convergence results may be expected to apply.

2.4 Method of Instrumental Variables (IV)

As mentioned at the end of the section on WLS, the method of least squares produces biased estimates $\hat{\theta}_k$ if θ_k and v_k are correlated. The method of instrumental variables is a variation of the least squares method, designed to overcome that problem.

To begin with, assume the system model of Equation 2.1.3,

$$y_k = \theta_0^T \phi_k + v_k \tag{2.4.1}$$

with the noise v_k correlated with previous measurements ϕ_k, so that the WLS method yields biased estimates. Further assume that there is available a sequence $\{\xi_k\}$ such that $E(\xi_k v_k) = 0$.

Then if the vectors ξ_k are substituted into the recursion for the estimate (Equation 2.1.8),

$$\hat{\theta}_k = \left[\sum_{l=1}^{k} \alpha_l \xi_l \phi_l^T \right]^{-1} \sum_{l=1}^{k} \alpha_l y_l \xi_l \tag{2.4.2}$$

the convergence result (following from Equation 2.1.22) is

$$\hat{\theta}_k = \theta_0 + \left[\frac{1}{k} \sum_{l=1}^{k} \alpha_l \xi_l \phi_l^T \right]^{-1} \left[\frac{1}{k} \sum_{l=1}^{k} \alpha_l \xi_l v_l \right] \tag{2.4.3}$$

Since the last factor on the right side of Equation 2.4.3 approaches the sample mean of $\xi_k v_k$, then $\hat{\theta}_k \to \theta_0$, provided that one can invert $\left[\frac{1}{k} \sum_{l=1}^{k} \alpha_l \xi_l \phi_l^T \right]$. That provision is important because it requires that ξ_k and v_k be "sufficiently correlated," while maintaining the uncorrelated relationship between ξ_k and ϕ_k.

While there is an infinite number of sequences $\{\xi_k\}$ satisfying those correlation requirements, leading to any number of variations on this algorithm, only one method is presented here, since it is fairly well-known.

In this method, a reference model of the system is created using the current estimate $\hat{\theta}_k$ and the input u_k, which is presumed to be uncorrelated with the output noise v_k. The model generates an estimate $\hat{x}_k$ of the "true" output of the system, i.e., of the system output if there were no measurement noise v_k.

From Figure 1 on page 23, the complete system diagram, it is fairly clear that ξ_k and ϕ_k can be expected to be correlated, while ξ_k and v_k should be uncorrelated. In this implementation

$$\begin{aligned} \xi_k &= [\, -\hat{x}_{k-1}, \dots, -\hat{x}_{k-n}, u_{k-1}, \dots, u_{k-m}]^T \\ \hat{x}_k &= \hat{\theta}_k^T \; \xi_k \end{aligned} \tag{2.4.4}$$

and the main recursions, with a development analogous to that of the SA method, are

$$\begin{aligned} P_{k+1} &= P_k - \frac{P_k \, \xi_{k+1} \, \phi_{k+1}^T \, P_k}{\frac{1}{\alpha_{k+1}} + \phi_{k+1}^T \, P_k \, \xi_{k+1}} \\ \hat{\theta}_{k+1} &= \hat{\theta}_k + \alpha_{k+1} \, P_{k+1} \, \xi_{k+1} \, (y_{k+1} - \phi_{k+1}^T \, \hat{\theta}_k) \end{aligned} \tag{2.4.5}$$

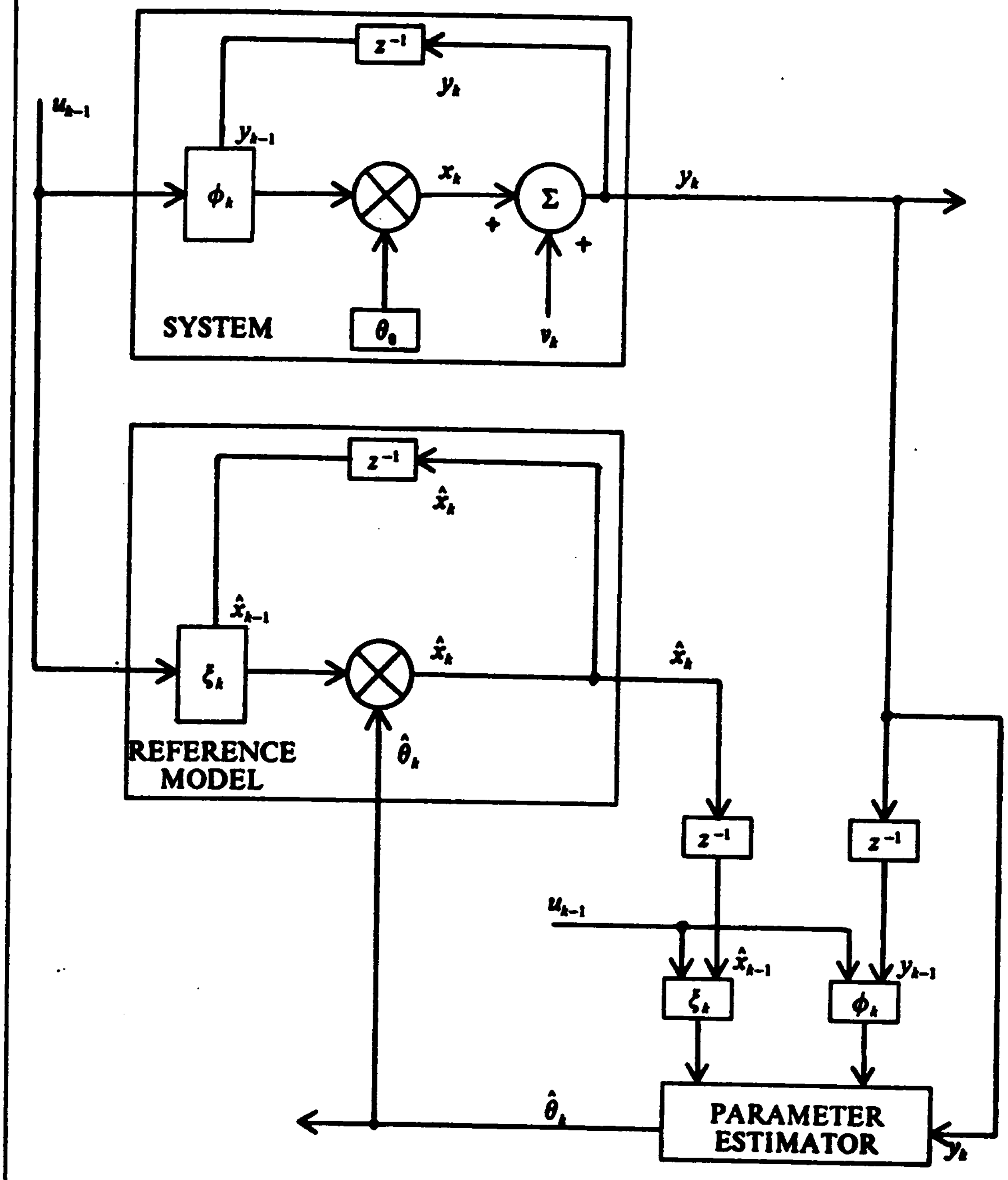

Figure 1. A typical implementation of the Method of Instrumental Variables.

2.5 Extended Least Squares (ELS), or Pseudo-Linear Regression

Another approach to the problem of correlation between ϕ_k and v_k (which essentially results from having v_k a non-white noise sequence) is to model the sequence v_k as the output of a moving average (MA) filter whose parameters and input are unknown. That is, the system model is extended so that the system input is (unknown) white noise; thus, the least squares method will give unbiased estimates.

The extension of the model is implemented simply by increasing the dimension of the vector of unknown parameters, so

$$y_k = [-a_1 y_{k-1} - \cdots -a_n y_{k-n} + b_1 u_{k-1} + \cdots + b_m u_{k-m} \\ + e_k + c_1 e_{k-1} + \cdots + c_r e_{k-r}] \tag{2.5.1}$$

Defining

$$\theta_0 \triangleq [a_1, \dots, a_n, b_1, \dots, b_m, c_1, \dots, c_r]^T \tag{2.5.2}$$

$$\phi_k \triangleq [-y_{k-1}, \dots, -y_{k-n}, u_{k-1}, \dots, u_{k-m}, e_{k-1}, \dots, e_{k-r}]^T \tag{2.5.3}$$

then the "equation error" is $e_k = y_k - \theta_0^T \phi_k$. (2.5.4)

So the problem is cast in the same form as Equation 2.1.3. However, it is not enough to estimate the additional elements c_i, $i = 1, \dots, r$; the equation error e_k must also be estimated in order to fill in the unmeasurable elements of the observation vector ϕ_{k+1}.

Thus, although the problem appears to be a linear regression, it is not; however, the same methods can be applied, with the additional steps of estimating $e_{k|k-1}$. Thus, the method is also called PLR.

From Equation 2.5.4, it is reasonable to estimate

$$\hat{e}_{k+1} = y_{k+1} - \phi_{k+1}^T \hat{\theta}_k \tag{2.5.5}$$

Now the WLS recursion can be extended:

$$(u_k, y_k, \hat{e}_k) \rightarrow \phi_{k+1} \tag{2.5.6a}$$

$$\hat{e}_{k+1} = y_{k+1} - \phi_{k+1}^T \hat{\theta}_k \tag{2.5.6b}$$

$$P_{k+1} = P_k - \frac{P_k \phi_{k+1} \phi_{k+1}^T P_k}{\frac{1}{\alpha_{k+1}} + \phi_{k+1}^T P_k \phi_{k+1}} \tag{2.5.6c}$$

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \alpha_{k+1} P_{k+1} \phi_{k+1} \hat{e}_{k+1} \tag{2.5.6d}$$

2.6 The Generalized Least Squares (GLS) Method

The development of GLS is similar to that of the ELS method. However, where the ELS method assumes that the noise $\{v_k\}$ in the system model

$$y_k = \theta_0^T \phi_k + v_k \tag{2.6.1}$$

can be modeled by white noise filtered through a moving average filter, the GLS method assumes that $\{v_k\}$ can be modeled by white noise filtered through an autoregressive (AR) filter. Thus, for the GLS method, it is assumed that

$$v_k = c_1 v_{k-1} + \cdots + c_r v_{k-r} + e_k , \tag{2.6.2}$$

where $\{e_k\}$ is a white noise sequence.

So, define the extended vectors

$$\left[\theta_0^T \,|\, \mu_0^T\right] \triangleq \left[a_1, \dots, a_n, b_1, \dots, b_m \,|\, c_1, \dots, c_r\right] \tag{2.6.3}$$

$$\left[\phi_k^T \,|\, \xi_k^T\right] \triangleq \left[-y_{k-1}, \dots, -y_{k-n}, u_{k-1}, \dots, u_{k-m} \,|\, v_{k-1}, \dots, v_{k-r}\right]. \tag{2.6.4}$$

As in the ELS method, v_k is unmeasurable, and must therefore be estimated. So the vector $\hat{\xi}$ will be used to estimate the vector of Equation 2.6.4, obtaining the elements by

$$\hat{v}_k = y_k - \phi_k^T \hat{\theta}_k . \tag{2.6.5}$$

Now the problem has been cast in the same form as the original LS problem,

$$y_k = \begin{bmatrix} \theta_0^T & \mu_0^T \end{bmatrix} \begin{bmatrix} \phi_k \\ \xi_k \end{bmatrix} + e_k \tag{2.6.6}$$

It results in a familiar recursion, proceeding as in the SA method:

$$P_{k+1} = P_k - \frac{P_k \begin{bmatrix} \phi_{k+1} \\ \hat{\xi}_{k+1} \end{bmatrix} \begin{bmatrix} \phi_{k+1}^T & \hat{\xi}_{k+1}^T \end{bmatrix} P_k}{\frac{1}{\alpha_{k+1}} + \begin{bmatrix} \phi_{k+1}^T & \hat{\xi}_{k+1}^T \end{bmatrix} P_k \begin{bmatrix} \phi_{k+1} \\ \hat{\xi}_{k+1} \end{bmatrix}}$$

$$\begin{bmatrix} \hat{\theta}_{k+1} \\ \hat{\mu}_{k+1} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_k \\ \hat{\mu}_k \end{bmatrix} + \alpha_{k+1} P_{k+1} \begin{bmatrix} \phi_{k+1} \\ \hat{\xi}_{k+1} \end{bmatrix} \left[y_{k+1} - \begin{bmatrix} \phi_{k+1}^T & \hat{\xi}_{k+1}^T \end{bmatrix} \begin{bmatrix} \hat{\theta}_k \\ \hat{\mu}_k \end{bmatrix} \right]$$

$$\hat{v}_{k+1} = y_{k+1} - \phi_{k+1}^T \hat{\theta}_{k+1} \tag{2.6.7}$$

$$\{y_{k+1}, u_{k+1}, \hat{v}_{k+1}\} \Rightarrow \begin{bmatrix} \phi_{k+2} \\ \hat{\xi}_{k+2} \end{bmatrix}$$

2.7 Model Reference Techniques

The IV, GLS, and GLS methods postulate changes to or extensions of the basic system model. The IV method in particular, might be considered a type of model reference method.

Model reference techniques form one of the two main approaches to adaptive control (the other being based on self-tuning regulators); its "dual," in a sense, can be used in parameter identification. In adaptive control, a reference model is set up to provide the desired (or reference) response to the known input; the controller uses the difference between the actual system output and the reference level to determine any necessary change in the control configuration. That is, the closed-loop system is altered until it conforms to the reference model.

In system identification, it is the reference model that is changed, until it conforms to the actual system. The model reference identification scheme can be visualized as in Figure 2 on page 29. A scheme attributed to Landau for the parameter adjustment of the reference model is described in [14].

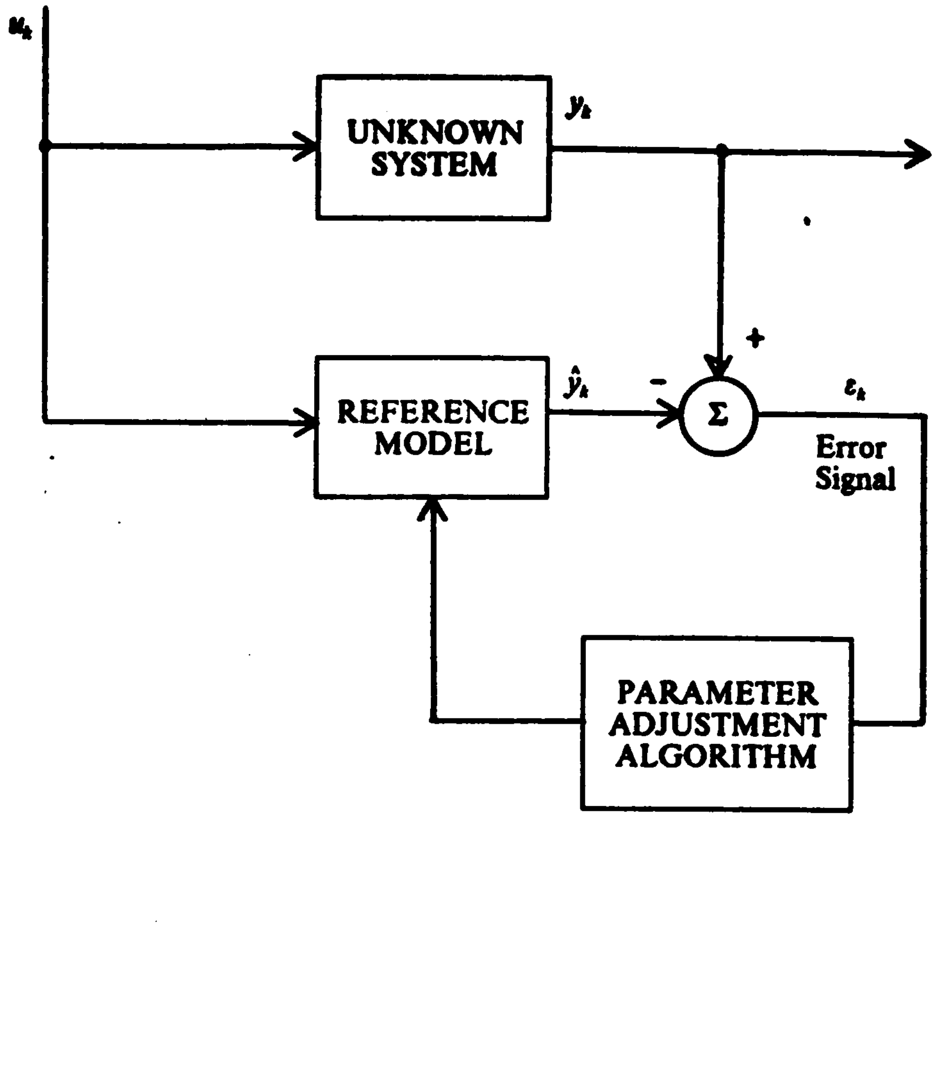

Figure 2. A typical model reference system identification scheme.

2.8 Recursive Maximum Likelihood (RML) Method

The RML method is one of the most successful approaches to obtaining unbiased estimates of the parameters θ of the system of Equation 2.1.3,

$$y_k = \theta^T \phi_k + v_k \tag{2.8.1}$$

when v_k is correlated noise.

As in section 6, the parameter vector is redefined to include the unknown coefficients of a moving average filter, $C(z)$,

$$\theta \triangleq [a_1, \dots, a_n, b_1, \dots, b_m, c_1, \dots, c_r]^T \tag{2.8.2}$$

The input to $C(z)$ is assumed to be an unknown gaussian white noise sequence $\{e_k\}$, and the output taken to be $\{v_k\}$.

The vector of measurements is likewise redefined to include the unmeasurable noise input $e_{k|k-1}$,

$$\phi_k \triangleq [-y_{k-1}, \dots, -y_{k-n}, u_{k-1}, \dots, u_{k-m}, e_{k-1}, \dots, e_{k-r}]^T \tag{2.8.3}$$

Thus,

$$y_k = \theta^T \phi_k + e_k \quad \rightarrow \quad e_k = y_k - \theta^T \phi_k \tag{2.8.4}$$

Now, $e_{k|k-1}$ is white and gaussian; denote its density by

$$p(e_k) = N(0, \sigma_e^2) \tag{2.8.5}$$

Define the vector

$$\underline{e}_r(k) \;\triangleq\; [e_{k-1}, \dots, e_{k-r}]^T \tag{2.8.6}$$

and note that

$$E(\,\underline{e}_r(k)\,\underline{e}_r^T(k)\,) = \sigma_e^2 \; I_r \tag{2.8.7}$$

Thus, if θ were known perfectly, $e_{k|k-1}$ could be computed exactly by Equation 2.8.4, and

$$p(\,\underline{e}_r(k)\;|\,\theta\,) = (2\pi\sigma_e^2\,)^{-r/2} \;\; \exp\left[-\frac{1}{2}\sum_{i=1}^{r}\frac{e_{k-i}^2}{\sigma_e^2}\right] \tag{2.8.8}$$

However, θ is not known; denote its estimate by $\hat{\theta}_k$. So, $e_{k|k-1}$ must be estimated in some way, in order to fill in the elements of ϕ_k. Denote by $\hat{e}_k$ the estimate of $e_{k|k-1}$, and by $\hat{\phi}_k$ the estimate of ϕ_k. Thus, from Equation 2.8.4,

$$\hat{e}_k = y_k - \hat{\theta}_k^T \;\hat{\phi}_k \tag{2.8.9}$$

Now define the "likelihood function" as the negative log of the *estimate* of Equation 2.8.8,

$$L(\,\hat{\underline{e}}_r(k)\;|\hat{\theta}_k\,) = \frac{r}{2}\;2\pi + \frac{r}{2}\log\tilde{\sigma}_e^2 + \frac{1}{2\tilde{\sigma}_e^2}\;\sum_{i=1}^{r}\hat{e}_{k-i}^2 \tag{2.8.10}$$

The estimate $\tilde{\sigma}_e^2$ that maximizes the likelihood function, denoted by $\hat{\sigma}_e^2$, can be found by solving

$$0=\left[\frac{\partial}{\partial\tilde{\sigma}_e^2}\, L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_k)\right]_{\tilde{\sigma}_e^2=\hat{\sigma}_e^2} = \frac{r}{2}\,\frac{1}{\hat{\sigma}_e^2} + (-1)\,\frac{1}{2\hat{\sigma}_e^4}\sum_{i=1}^{r}\hat{e}_{k-i}^2 \tag{2.8.11}$$

$$\hat{\sigma}_e^2 = \frac{1}{r}\sum_{i=1}^{r}\hat{e}_{k-i}^2 \tag{2.8.12}$$

Note that the maximizing value is the sample variance of the estimates $\hat{e}_k$.

Now, given $\hat{\theta}_k$, it is desired to find $\hat{\theta}_{k+1}$ such that

$$L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_{k+1}) > L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_k)\ ,$$

that is, so that the likelihood function montonically increases as the algorithm proceeds.

Define $\delta\theta \triangleq \hat{\theta}_{k+1} - \hat{\theta}_k$ (2.8.13)

and expand the likelihood function in a Taylor series about $\hat{\theta}_k$:

$$\begin{aligned} L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_{k+1}) = {} & L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_k) + \frac{\partial L}{\partial\theta}\Big|_{\theta=\hat{\theta}_k}(\hat{\theta}_{k+1}-\hat{\theta}_k) \\ & + \frac{1}{2}\,(\hat{\theta}_{k+1}-\hat{\theta}_k)^T\,\frac{\partial^2 L}{\partial\theta^2}\Big|_{\theta=\hat{\theta}_k}(\hat{\theta}_{k+1}-\hat{\theta}_k) + \cdots \end{aligned} \tag{2.8.14}$$

where the ellipsis indicates negligible higher-order terms. Substituting Equation 2.8.13,

$$L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_{k+1}) \simeq L(\hat{\underline{e}}_r(k)\,|\,\hat{\theta}_k) + \frac{\partial L}{\partial\theta}\Big|_{\theta=\hat{\theta}_k}\delta\theta + \frac{1}{2}\,\delta\theta^T\,\frac{\partial^2 L}{\partial\theta^2}\Big|_{\theta=\hat{\theta}_k}\delta\theta \tag{2.8.15}$$

Now, assuming that $\frac{\partial}{\partial(\delta\theta)} \; L(\hat{\underline{e}}_r(k) \,|\, \hat{\theta}_k) \simeq 0$

the minimization of Equation 2.8.15 proceeds,

$$0 = \frac{\partial}{\partial(\delta\theta)} \; L(\hat{\underline{e}}_r(k) \,|\, \hat{\theta}_{k+1}) = \frac{\partial L}{\partial\theta}\Big|_{\theta=\hat{\theta}_k} + 2\left(\frac{1}{2}\right) \frac{\partial^2 L}{\partial\theta^2}\Big|_{\theta=\hat{\theta}_k} (\delta\theta) \qquad (2.8.16)$$

$$\text{Hence,} \quad \delta\theta = -\left[\frac{\partial^2 L}{\partial\theta^2}\right]^{-1} \frac{\partial L}{\partial\theta}\Big|_{\theta=\hat{\theta}_k} \qquad (2.8.17)$$

Thus, from Equation 2.8.13,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \left[\frac{\partial^2 L}{\partial\theta^2}\right]^{-1} \frac{\partial L}{\partial\theta}\Big|_{\theta=\hat{\theta}_k} \qquad (2.8.18)$$

The quantities in Equation 2.8.18 must be estimated recursively. Note that $\hat{\sigma}_e^2$ is already determined by Equation 2.8.12. So, the first partial is

$$\begin{aligned}\frac{\partial}{\partial\theta} \; L(\hat{\underline{e}}_r(k) \,|\, \hat{\theta}_k) &= \frac{\partial}{\partial\theta}\left[\frac{r}{2}\, 2\pi + \frac{r}{2}\log\hat{\sigma}_e^2 + \frac{1}{2\hat{\sigma}_e^2}\sum_{i=1}^{r} \hat{e}_{k-i}^2\right] \\ &= \frac{1}{2\hat{\sigma}_e^2}\sum_{i=1}^{r} \frac{\partial}{\partial\theta}\, \hat{e}_{k-i}^2 \\ &= \frac{1}{2\hat{\sigma}_e^2}\sum_{i=1}^{r} \hat{e}_{k-i} \left(\frac{\partial}{\partial\theta}\, \hat{e}_{k-i}\right)\end{aligned} \qquad (2.8.19)$$

From Equation 2.8.19, then the second partial is

$$\frac{\partial^2}{\partial\theta^2} L(\underline{e}_r(k) \mid \hat{\theta}_k) = \frac{1}{\hat{\sigma}_e^2} \sum_{l=1}^{r} \left(\frac{\partial}{\partial\theta} \hat{e}_{k-l} \right)^T \left(\frac{\partial}{\partial\theta} \hat{e}_{k-l} \right) + \hat{e}_{k-l} \left(\frac{\partial^2}{\partial\theta^2} \hat{e}_{k-l} \right)$$

where the last term on the right is assumed negligible, so

$$\frac{\partial^2}{\partial\theta^2} L(\underline{e}_r(k) \mid \hat{\theta}_k) = \frac{1}{\hat{\sigma}_e^2} \sum_{l=1}^{r} \left(\frac{\partial}{\partial\theta} \hat{e}_{k-l} \right)^T \left(\frac{\partial}{\partial\theta} \hat{e}_{k-l} \right) \qquad (2.8.20)$$

Now, using Equations 2.8.6 and 2.8.9, the first partial is clearly

$$\frac{\partial}{\partial\theta}\,\hat{e}_{k-i}=\begin{bmatrix} y_{k-i-1}-\sum_{j=1}^{r}\hat{c}_j\,\frac{\partial}{\partial a_1}\,\hat{e}_{k-i-j}\\ \vdots\\ y_{k-i-n}-\sum_{j=1}^{r}\hat{c}_j\,\frac{\partial}{\partial a_n}\,\hat{e}_{k-i-j}\\ -u_{k-i-1}-\sum_{j=1}^{r}\hat{c}_j\,\frac{\partial}{\partial b_1}\,\hat{e}_{k-i-j}\\ \vdots\\ -u_{k-i-m}-\sum_{j=1}^{r}\hat{c}_j\,\frac{\partial}{\partial b_m}\,\hat{e}_{k-i-j}\\ -\,\hat{e}_{k-i-1}-\sum_{j=1}^{r}\hat{c}_j\,\frac{\partial}{\partial c_1}\,\hat{e}_{k-i-j}\\ \vdots\\ -\,\hat{e}_{k-i-r}-\sum_{j=1}^{r}\hat{c}_j\,\frac{\partial}{\partial c_r}\,\hat{e}_{k-i-j} \end{bmatrix} \qquad (2.8.21)$$

For each element of Equation 2.8.21, the graphs of Figure 3 on page 36 will perform the appropriate iterations. However, note that some concatenation can be performed, so that only three graphs are required, in total, each of the form shown in Figure 4 on page 37.

This algorithm could be initialized by running the extended least squares algorithm first, to get relatively good initial estimates of $\hat{\theta}_k$ and the vector $\frac{\partial}{\partial\theta}\,\hat{e}_k$. One difficulty with it is the required inversion of $\frac{\partial^2}{\partial\theta^2}\,L$, which is a square $(n+m+r)$-dimensional matrix, on each iteration.

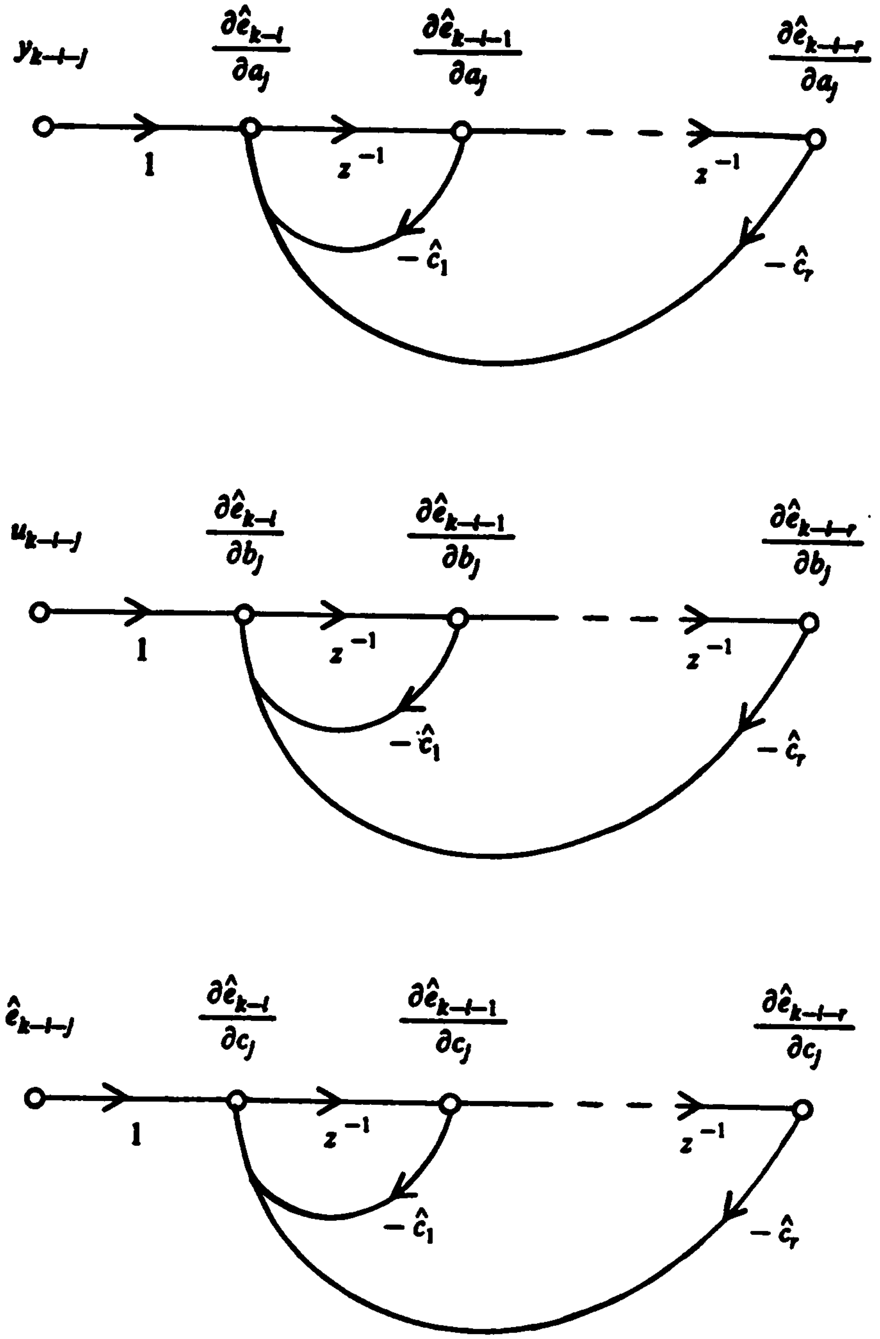
y_{k-l-j}
$\frac{\partial \hat{e}_{k-l}}{\partial a_j}$
$\frac{\partial \hat{e}_{k-l-1}}{\partial a_j}$
$\frac{\partial \hat{e}_{k-l-r}}{\partial a_j}$
1
z^{-1}
z^{-1}
$-\hat{c}_1$
$-\hat{c}_r$
u_{k-l-j}
$\frac{\partial \hat{e}_{k-l}}{\partial b_j}$
$\frac{\partial \hat{e}_{k-l-1}}{\partial b_j}$
$\frac{\partial \hat{e}_{k-l-r}}{\partial b_j}$
1
z^{-1}
z^{-1}
$-\hat{c}_1$
$-\hat{c}_r$
$\hat{e}_{k-l-j}$
$\frac{\partial \hat{e}_{k-l}}{\partial c_j}$
$\frac{\partial \hat{e}_{k-l-1}}{\partial c_j}$
$\frac{\partial \hat{e}_{k-l-r}}{\partial c_j}$
1
z^{-1}
z^{-1}
$-\hat{c}_1$
$-\hat{c}_r$

Figure 3. Implementation of the Equation 2.8.21.

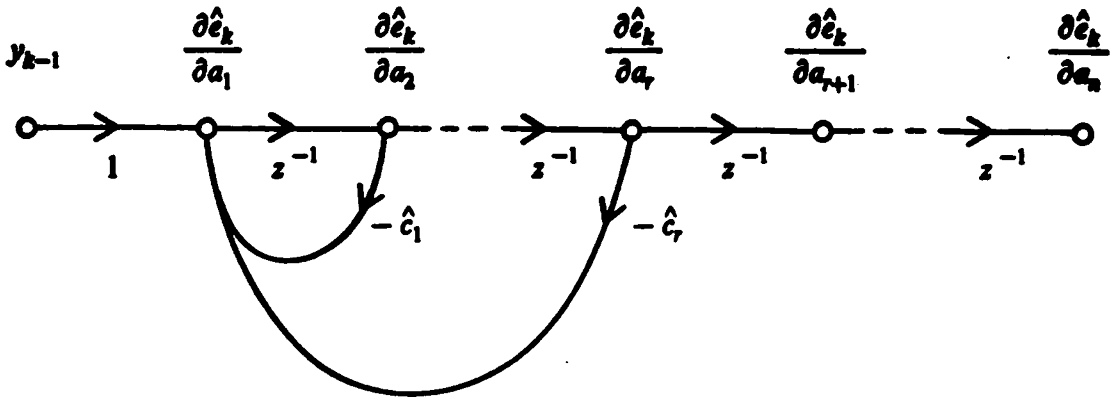
y_{k-1}
$\frac{\partial\hat{e}_k}{\partial a_1}$
$\frac{\partial\hat{e}_k}{\partial a_2}$
$\frac{\partial\hat{e}_k}{\partial a_r}$
$\frac{\partial\hat{e}_k}{\partial a_{r+1}}$
$\frac{\partial\hat{e}_k}{\partial a_n}$
1
z^{-1}
z^{-1}
z^{-1}
z^{-1}
$-\hat{c}_1$
$-\hat{c}_r$

Figure 4. Concatenation-in-time for Equation 2.8.21.

3.0 Previous Methods of Joint Estimation

The joint estimation of states and parameters is a problem that has been investigated by many different researchers. As a result, there is a wide variety of algorithms available in the literature. Typically, a joint estimation algorithm is derived for a very limited class of systems, due to the necessity of enforcing some set of assumptions in order to obtain reasonable theoretical results. After the derivation and associated proofs, a great deal of effort is usually expended to show, at least heuristically, that the class of systems to which the algorithm can be successfully applied is actually larger than the one assumed at the beginning.

One standard approach to joint estimation of parameters and states, which, it is worth noting, is an explicitly nonlinear problem, uses the extended Kalman filter (EKF) to estimate an extended state vector composed of both the states and the parameters. The basic premise of its use is that, in this application, the nonlinear system can be well-approximated by a linearization about the current state. The main difficulty arises from the fact that the EKF is not globally convergent. Thus, when the current state is completely unknown, the linearization most likely proceeds about some state that is quite far from the actual state, perhaps outside the region of convergence of the filter. Thus, the EKF may require some initial estimate that is reasonably accurate in order to converge. Convergence problems of the EKF are illustrated by the simulations in Section 9.2.

Another standard approach is the "bootstrap" algorithm. In this type of algorithm, two tasks are alternated: the state vector is estimated; the state estimate is then used to estimate the parameter vector, which is then used to improve the state estimate, and so on. One problem with this approach lies in trying to incorporate error covariance data from the parameter estimator into the the state estimator, and *vice versa*. That problem is circumvented by methods that use an extended state vector, because they are required to update only one (much larger) error covariance matrix.

In 1980, Salut *et.al.* [5] presented a novel approach that takes advantage of the structure of the observable-canonic form to derive an extended system (*i.e.*, a system whose state vector includes the original unknown states together with all of the unknown parameters) that is explicitly linear. Using a clever replacement of states appearing in the system matrix, by measurements of the system output, the nonlinearities become implicit. This enables the development of an estimator that is linear in the estimates. The work in [5] treats systems that have white gaussian state noise with known autocovariance, and system inputs and outputs that are perfectly known. The resulting estimator was shown to be optimal in the mean-square-error sense. No convergence study was presented.

In 1986, Chen *et.al.* [7] extended the work of Salut to the case where the state noise had ***unknown*** autocovariance. (As in [5], the system inputs and outputs were assumed known exactly.) The unavoidable result is the loss of explicit linearity in the extended system model. The development in [7] therefore requires the use of an extended Kalman filter, a linearization of the nonlinear extended system, resulting in a necessarily suboptimal estimator. The parameter estimates were shown to converge w.p.1 to a stationary point of a Lyapunov function, not necessarily the true parameter values.

An example of the bootstrap method of joint parameter and state estimation is given by the work of Nelson and Stear [15]. Their method involves an alternate state representation that they claimed was canonic; however, Padilla *et.al.* [16] showed that such a state representation cannot, in general, be obtained; it is therefore not canonic.

In this method, the parameters are estimated first, then these estimates are used to update the state estimator, in the typical "bootstrap" arrangement.

The method of [15] reportedly had a tendency toward divergence of the state estimates. Nelson and Stear speculated that there could be a way to modify the state estimate covariance matrix to account for errors in the parameter estimates in order to prevent divergence, thus presaging the PLID algorithm, which does precisely that.

The method of [15] is not detailed here. However, the extended Kalman filter, which is widely used for joint parameter and state estimation, is derived in Section 3.1.

3.1 The Extended Kalman Filter (EKF)

Consider a SISO system, assumed to be completely controllable and observable. To begin, assume the system is in observable-canonic form, which is possible to attain because of the assumption of complete observability. Let v_k be unknown gaussian white noise (variance q) corrupting the input to the system, and let w_k be unknown gaussian white noise (variance r) corrupting the measurement of the system output.

Also assume zero mean white gaussian noise $\Xi_x(k)$ adding directly to the system states, where $Cov\,[\Xi_x(k)\,] \triangleq \Sigma_k^{xx}$ is known for all time k.

So the system can be written

$$\begin{aligned} x_{k+1} &= A\,x_k + B\,(u_k + v_k\,) + \Xi_x(k) \\ z_k &= C\,x_k + w_k \end{aligned}$$

$$\text{where } A = \begin{bmatrix} 0\,0\ldots0\,0 & a_0 \\ 1\,0\ldots0\,0 & a_1 \\ \vdots & \\ 0\,0\ldots0\,1 & a_{n-1} \end{bmatrix}, \qquad B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix}, \tag{3.1.1}$$

$$\text{and} \quad C = [\,0\,0\ldots0\,0 \quad 1 \quad]\,.$$

The parameters appearing in the various matrices are unknown, and must be identified just as x_k must be estimated. Therefore, define the *extended state vector,*

$$s_k = [x_1(k) \;\ldots\; x_n(k) \mid a_0 \;\ldots\; a_{n-1} \mid b_0 \;\ldots\; b_{n-1}\,]^T \tag{3.1.2}$$

Let $\Xi_\theta(k)$, be unknown gaussian white noise adding directly to the parameters in the extended vector s_k. Since the parameters are assumed to be time-invariant, $\Xi_\theta(k)$ must be zero. However, the Kalman filter is, in general, more robust if a non-zero noise is assumed to be added to each state to be estimated [17]. So denote the small non-zero autocovariance of $\Xi_\theta(k)$ as $\Sigma_k^{\theta\theta}$.

Now that the parameters are included in the state vector, the nonlinearities of the *extended system* are obvious. Denoting $\bar{u}_k \triangleq u_k + v_k$, the system equation is

$$s_{k+1} \quad = \quad f(s_k, u_k, v_k, \Xi_k) \tag{3.1.3}$$

$$\begin{bmatrix} s_1(k+1) \\ s_2(k+1) \\ \vdots \\ s_n(k+1) \\ s_{n+1}(k+1) \\ \vdots \\ \vdots \\ s_{3n+1}(k+1) \end{bmatrix} = \begin{bmatrix} s_{n+1}(k)s_n(k) + s_{2n+1}(k)\,\bar{u}_k + \xi_1(k) \\ s_1(k) + s_{n+2}(k)s_n(k) + s_{2n+2}(k)\,\bar{u}_k + \xi_2(k) \\ \vdots \\ s_{n-1}(k) + s_{2n}(k)s_n(k) + s_{3n}(k)\,\bar{u}_k + \xi_n(k) \\ s_{n+1}(k) + \xi_{n+1}(k) \\ \vdots \\ \vdots \\ s_{3n}(k) + \xi_{3n}(k) \end{bmatrix}$$

and the output equation is

$$z_k = h(s_k, u_k, v_k, w_k) = s_n(k) + w_k \tag{3.1.4}$$

Now, following the derivation in Goodwin and Sin [18], the various state noise components in Equation 3.1.3 are placed together in one vector,

$$\eta(k) = [v_k \mid \xi_1(k) \mid \dots \mid \xi_{3n}(k)\,]^T, \tag{3.1.5}$$

and the input/output cross-covariance matrix is written,

$$\begin{aligned} E\left[\begin{bmatrix}\eta_i \\ w_i\end{bmatrix}\begin{bmatrix}\eta_k^T & w_k^T\end{bmatrix}\right] &= \begin{bmatrix} Q & S \\ S^T & R\end{bmatrix}\delta(k-i) \\ &= block\ diag\,[\,q\,, \Sigma_k^{xx}\,, \Sigma_k^{\theta\theta}\,, r\,] \end{aligned} \tag{3.1.6}$$

Now, using a Taylor series expansion and assuming that higher-order terms can be neglected, the system at time k is linearized about some estimate $\hat{s}_k$ of the state s_k, The resulting linearization is

$$\begin{aligned} s_{k+1} &\cong f(\hat{s}_k, u_k, \eta_k)_{\eta_k=0} + F_k[\,s_k - \hat{s}_k\,] + G_k\,\eta_k \\ z_k &\cong h(\hat{s}_k, u_k, w_k)_{w_k=0} + H[\,s_k - \hat{s}_k\,] + J_k\,w_k \end{aligned} \tag{3.1.7}$$

where

$$\begin{aligned} F_k &\triangleq \left[\frac{\partial f(s_k, u_k, \eta_k)}{\partial s_k}\right]_{s_k=\hat{s}_k,\,\eta_k=0} \\ G_k &\triangleq \left[\frac{\partial f(s_k, u_k, \eta_k)}{\partial \eta_k}\right]_{s_k=\hat{s}_k,\,\eta_k=0} \\ H &\triangleq \left[\frac{\partial h(s_k, u_k, w_k)}{\partial s_k}\right]_{s_k=\hat{s}_k,\,w_k=0} \\ J_k &\triangleq \left[\frac{\partial h(s_k, u_k, w_k)}{\partial w_k}\right]_{s_k=\hat{s}_k,\,w_k=0} \end{aligned}$$

Performing the partial derivatives yields the following matrices:

$$F_k = \left[\begin{array}{ccccc|c|c} 0 & 0 & \cdots & 0 & s_{n+1}(k) & & \\ 1 & 0 & \cdots & 0 & s_{n+2}(k) & & \\ \cdot & & & & \vdots & & \\ & \cdot & & & \vdots & s_n(k)\,\mathbf{I}_n & u_k\,\mathbf{I}_n \\ & & \cdot & & \vdots & & \\ 0 & 0 & \cdots & 1 & s_{2n}(k) & & \\ \hline & & & & & & \\ & & \mathbf{0} & & & & \mathbf{I}_{2n} \\ & & & & & & \end{array}\right] \tag{3.1.8}$$

$$G_k = \left[\begin{array}{c|c} s_{2n+1}(k) & \\ s_{2n+2}(k) & \\ \vdots & \\ \vdots & \\ s_{3n}(k) & \mathbf{I}_{3n} \\ \hline & \\ \mathbf{0} & \\ & \end{array}\right] \tag{3.1.9}$$

$$H = [\, 0 \;\cdots\; 0 \; 1 \mid 0 \;\cdots\; 0 \mid 0 \;\cdots\; 0 \,] \tag{3.1.10}$$

$$J_k = 1 \tag{3.1.11}$$

Now, letting $P_{k|k-1}$ denote the prediction error covariance matrix, and $K_{k|k}$ denote the Kalman gains, it is straightforward to write the Kalman equations for a one-step ahead prediction $\hat{s}_{k|k-1}$ for the linearized system:

$$\hat{s}_{k+1|k} = f(\hat{s}_{k|k-1}, u_k, 0) + K_{k|k}\left[z_k - h(\hat{s}_{k|k-1}, 0)\right] \tag{3.1.12}$$

$$K_{k|k} = \left[F_k P_{k|k-1} H^T + G_k S\right]\left[H P_{k|k-1} H^T + R\right]^{-1} \tag{3.1.13}$$

$$\begin{aligned} P_{k+1|k} &= F_k P_{k|k-1} F_k^T + G_k Q G_k^T \\ &\qquad - K_{k|k}\left[H P_{k|k-1} H^T + R\right] K_{k|k}^T \end{aligned} \tag{3.1.14}$$

where some $\hat{s}_{0|-1}$ and $P_{0|-1}$ are assumed in order to start the algorithm.

4.0 *Salut Form: The Extended State Representation*

This chapter develops the extended state representation (ESR) both for the single-input, single-output (SISO) case, and the multiple-input, multiple-output (MIMO) case. The ESR is a canonic form derived from the observable-canonic form through a rather clever algebraic manipulation. The great advantage of the ESR is that the intrinsic nonlinearities become implicit, rather than explicit, so that the estimator based on the ESR will be linear in the estimates. The ESR was apparently first derived by Salut [6], and later independently derived by VanLandingham [8]. As shown in the derivations of this chapter, it is a canonic form because it retains the states of the observable canonic form. Thus, if a system is completely observable, then it can be represented in the ESR form.

The extended state representation is the starting point for the derivation of the PLID algorithm. In addition to being completely observable, a system must also be completely controllable in order for the PLID algorithm to work; otherwise, the input will fail to be *persistently exciting*. The persistent excitation of a system is generally recognized to be a prerequisite for system identification. Essentially, it amounts to exciting all the modes of the system for a length of time sufficient to uniquely identify the parameters (and, possibly, also the states, depending on the identification algorithm). Specific requirements for complete controllability are discussed in connection with the observable-canonic state equations, which are Equation 4.1.4 for the SISO case, and Equation 4.2.4 for the MIMO case.

Specific requirements for complete observability of the extended system are discussed in connection with the time-varying observability matrices, given by Equation

4.1.14 for the SISO case, and by Equation 4.2.12 for the MIMO case. One of the main differences between the observability matrix of a time-varying system and that of a time-invariant system is that the Cayley-Hamilton theorem only applies in the time-invariant case. Thus, in the time-varying case, it is not necessarily sufficient to consider only the first *n* rows (where *n* is the number of states in the system) when determining the observability of the system.

4.1 Extended State Representation of SISO Systems

This section develops the *extended state model*, or Salut form [6], for a single-input single output (SISO) system, allowing the joint estimation problem to be cast as an explicitly linear problem.

Suppose there is an unknown SISO system S for which scalar input data $\{u_k\}$ and scalar output data $\{y_k\}$ are available. If S is assumed to be a linear, n^{th} -order, completely controllable and observable, strictly proper system, where $n > 0$, then there is a transfer function

$$H(z) = \frac{Y(z)}{U(z)} = \frac{b_{n-1}\, z^{n-1} + \cdots + b_1\, z + b_0}{z^n - a_{n-1}\, z^{n-1} - \cdots - a_1\, z - a_0} \tag{4.1.1}$$

where the a_i and b_i are system parameters. Cross-multiplying Equation 4.1.1, then premultiplying both sides by z^{-n},

$$(1 - a_{n-1}\, z^{-1} \; - \cdots - a_0\, z^{-n}\,)\;\; Y(z) = (b_{n-1}\, z^{-1} \; + \cdots + b_0\, z^{-n}\,)\;\; U(z) \tag{4.1.2}$$

Note that the inverse Z-transform of Equation 4.1.2 is

$$y_k \; = a_{n-1}\, y_{k-1} \; + \cdots + a_0\, y_{k-n} \; + b_{n-1}\, u_{k-1} \; + \cdots + b_0\, u_{k-n} \tag{4.1.3}$$

In this form, it is clear that there is no loss of generality in assuming a strictly proper system, because every real system has non-zero delay from the input to the output.

The observable-canonic state diagram follows from Equation 4.1.2, and is shown in Figure 5 on page 52. The observable canonic state model follows from the state diagram:

$$x_{k+1} = A\,x_k + B\,u_k$$
$$y_k = C\,x_k$$
$$\text{where} \quad A = \begin{bmatrix} 0\,0\,\ldots\,0\,0 & a_0 \\ 1\,0\,\ldots\,0\,0 & a_1 \\ \vdots & \\ 0\,0\,\ldots\,0\,1 & a_{n-1} \end{bmatrix}, \qquad B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix}, \tag{4.1.4}$$
$$C = [0\,0\,\ldots\,0\,0 \quad 1 \;].$$

From Equation 4.1.4 it is obvious that the simplest configuration for complete controllability requires that the parameter b_0 must not be zero. This can also be deduced from the observable-canonic state diagram, given in Figure 5 on page 52.

Defining $x_0(k) \equiv 0$, the general form of the rows in Equation 4.1.4 is

$$x_i(k+1) = x_{i-1}(k) + a_{i-1}\,x_n(k) + b_{i-1}\,u_k\,, \quad \forall\, i = 1, \ldots, n, \tag{4.1.5}$$

$$y_k = x_n(k) \tag{4.1.6}$$

Now Equation 4.1.5 can be rearranged as

$$x_i(k+1) = x_{i-1}(k) + x_n(k)\,a_{i-1} + u_k\,b_{i-1}\,, \quad \forall\, i = 1, \ldots, n, \tag{4.1.7}$$

and the identity of Equation 4.1.6 can be applied, yielding

$$x_i(k+1) = x_{i-1}(k) + y_k\,a_{i-1} + u_k\,b_{i-1}\,, \quad \forall\, i = 1, \ldots, n. \tag{4.1.8}$$

From the generalized row form of Equation 4.1.8, the following alternate system description can be deduced:

$$x_{k+1} = \begin{bmatrix} 0\,0 \dots 0\,0 \\ 1\,0 \dots 0\,0 \\ \vdots \\ 0\,0 \dots 1\,0 \end{bmatrix} x_k + \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{bmatrix} y_k + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix} u_k \,. \tag{4.1.9}$$

Define the *extended state vector,*

$$s_k \triangleq \begin{bmatrix} x_k \\ \theta_A \\ \theta_B \end{bmatrix} \in \mathbb{R}^{3n}$$

$$\text{where } \theta_A = \begin{bmatrix} a_0 \\ \vdots \\ a_{n-1} \end{bmatrix}, \text{ and } \theta_B = \begin{bmatrix} b_0 \\ \vdots \\ b_{n-1} \end{bmatrix}. \tag{4.1.10}$$

Note that θ_A is the n^{th} column of system matrix A, and θ_B is the input column matrix B, from Equation 4.1.4.

Now with that definition of the extended state vector, the *extended system*, denoted by $\overline{S}$, can be written:

$$s_{k+1} = F(y_k, u_k)\, s_k \triangleq F_k\, s_k$$
$$y_k = H\, s_k$$

$$\text{where} \quad F_k = \begin{bmatrix} \mathbf{J}_n & y_k \mathbf{I}_n & u_k \mathbf{I}_n \\ 0 & \mathbf{I}_{2n} & \end{bmatrix} \tag{4.1.11}$$

$$\text{and} \quad H = [\ 0 \dots 0\ 1 \mid 0 \dots 0\ 0 \mid 0 \dots 0\ 0\ ]$$

Note that $\mathbf{J}_n$ denotes the $n \times n$ lower Jordan block of zero eigenvalues.

The extended state model $\overline{S}$ of Equation 4.1.11 is time-varying and autonomous. Also, while the original system model S is of order n, the extended state model $\overline{S}$ is of order $3n$. Now, in order to completely identify the original system S, one can identify the vector s_k of system $\overline{S}$, since s_k contains all of the (observable canonic) states and parameters of S. That is, determining s_k is equivalent to identifying S.

Thus, the input sequence $\{u_k\}$ of S affects the ability to identify S, since it shows up in matrix F_k of extended system $\overline{S}$. This is a manifestation of a problem common to all system identification procedures, namely that the input must excite all the modes of the unknown system sufficiently to allow identification. This is discussed further in Chapter 6. For now, assume $\{u_k\}$ is a suitable (*i.e.*, sufficiently exciting) input sequence.

The unique form of $\overline{S}$ allows us to recast the question of the identifiability of S into a question of observability of $\overline{S}$. Note that

$$Y^k \triangleq \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} H \\ H F_0 \\ \vdots \\ H F_{k-1} \ldots F_0 \end{bmatrix} s_0 \triangleq \Phi_k s_0 \qquad (4.1.12)$$

Matrix Φ_k is called the k^{th} *observability matrix* of $\overline{S}$. Although it exhibits obvious differences from the observability matrix of a *time-invariant* system, it serves the same purpose insofar as determining the observability of the system. If *rank*$(\Phi_k) = 3n$, then Equation 4.1.12 can be solved (that is, s_0 can be identified) by obtaining the generalized inverse of Φ_k. The questions surrounding the rank of Φ_k are addressed in Chapter 6.

One major distinction between the observability matrix of a time-varying system and that of a time-invariant system is as follows. In a time-invariant system, it is only necessary to consider the the first n rows in order to determine the observability of the system, a fact that follows from the Cayley-Hamilton theorem. However, in a time-varying system, even if the observability matrix has not reached full column rank with n rows, it may yet reach full column rank with more than n rows, because the Cayley-Hamilton theorem no longer applies.

Since there are $3n$ columns in Φ_k, it is clear that the minimum number of observations (rows of Φ_k) required to identify the initial extended state (*i.e.*, to obtain the generalized inverse) is $3n$. Define the ***extended observability matrix*** Φ^* as the smallest observability matrix Φ_k having rank $3n$. Thus Φ^* exists at the first time k that it is possible to determine s_k uniquely, since

$$s_k \;=\; F_{k-1} \ldots F_0\, s_0 \;=\; F_{k-1} \ldots F_0\, (\,\Phi_k^T \Phi_k\,)^{-1}\, \Phi_k^T\, Y^k \qquad (4.1.13)$$

is not computable before the indicated inverse exists.

Explicitly, the extended observability matrix, where $l \geq 0$, is:

$$\Phi^* = \left[\begin{array}{c|cccc|cccc}
0\,0 \ldots 0\,0\,1 & 0 & 0 & \ldots\ 0 & 0 & 0 & 0 & \ldots\ 0 & 0 \\
0\,0 \ldots 0\,1\,0 & 0 & 0 & \ldots\ 0 & y_0 & 0 & 0 & \ldots\ 0 & u_0 \\
 & & & & \vdots & & & & \\
1\,0 \ldots 0\,0\,0 & 0 & y_0 & \cdots y_{n-3} & y_{n-2} & 0 & u_0 & \cdots\ u_{n-3} & u_{n-2} \\
0\,0 \ldots 0\,0\,0 & y_0 & & \cdots y_{n-2} & y_{n-1} & u_0 & & \cdots\ u_{n-2} & u_{n-1} \\
 & & & & \vdots & & & & \\
0\,0 \ldots 0\,0\,0 & y_{n-1} & & \cdots y_{2n-3} & y_{2n-2} & u_{n-1} & & \cdots\ u_{2n-3} & u_{2n-2} \\
0\,0 \ldots 0\,0\,0 & y_n & & \cdots y_{2n-2} & y_{2n-1} & u_n & & \cdots\ u_{2n-2} & u_{2n-1} \\
 & & & & \vdots & & & & \\
0\,0 \ldots 0\,0\,0 & y_{2n-1} & & \cdots y_{3n-3} & y_{3n-2} & u_{2n-1} & & \cdots\ u_{3n-3} & u_{3n-2} \\
0\,0 \ldots 0\,0\,0 & y_{2n} & & \cdots y_{3n-2} & y_{3n-1} & u_{2n} & & \cdots\ u_{3n-2} & u_{3n-1} \\
 & & & & \vdots & & & & \\
0\,0 \ldots 0\,0\,0 & y_{2n+l} & & \cdots & y_{3n-2+l} & u_{2n+l} & & \cdots & u_{3n-2+l}
\end{array} \right] \qquad (4.1.14)$$

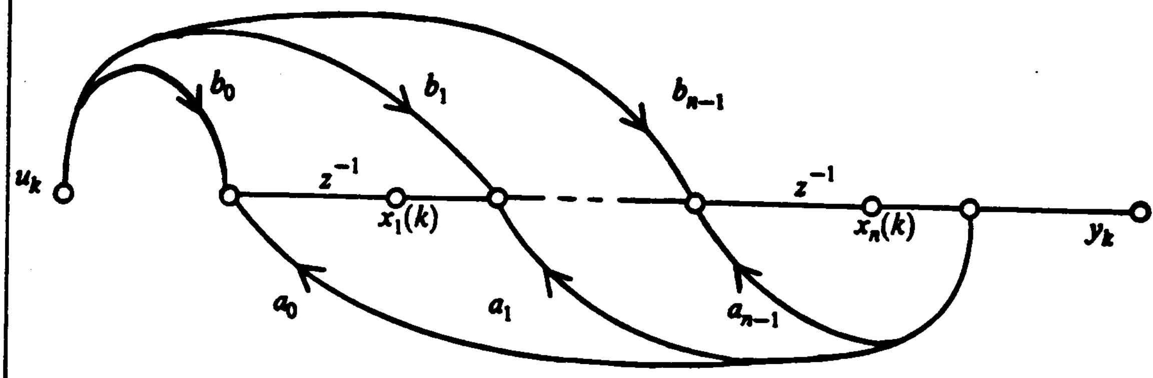

Figure 5. Observable-canonic state diagram for the general SISO system.

4.2 *Extended State Representation of MIMO Systems*

This section extends the SISO result of Section 4.1 to multiple-input, multiple-output (MIMO) systems. This section parallels the SISO development quite closely, resulting in the extended state model for the general observable MIMO system.

Suppose an unknown discrete-time system S is MIMO, linear, time-invariant, deterministic, completely controllable, and completely observable. Let p denote the number of system outputs, $y_1(k), \ldots, y_p(k)$, and let $n_1, \ldots, n_p$ be the known observability indices associated with those outputs, n_i being associated with $y_i(k)$. The total number n of states in the system is, of course, the sum of the observability indices. Also, let m denote the number of system inputs, $u_1(k), \ldots, u_m(k)$.

By the assumption of complete observability and controllability, there is a transfer function for the system:

$$\left[\begin{array}{c|c|c} z^{n_1} - a_{1,1}^{n_1-1}\, z^{n_1-1} - \cdots - a_{1,1}^{0} & \cdots & -a_{p,1}^{n_1-1}\, z^{n_1-1} - \cdots - a_{p,1}^{0} \\ \hdashline \vdots & & \vdots \\ \hdashline -a_{1,p}^{n_p-1}\, z^{n_p-1} - \cdots - a_{1,p}^{0} & \cdots & z^{n_p} - a_{p,p}^{n_p-1}\, z^{n_p-1} - \cdots - a_{p,p}^{0} \end{array}\right] \begin{bmatrix} Y_1(z) \\ \vdots \\ Y_p(z) \end{bmatrix} =$$

$$\left[\begin{array}{c|c|c} b_{1,1}^{n_1-1}\, z^{n_1-1} + \cdots + b_{1,1}^{0} & \cdots & b_{m,1}^{n_1-1}\, z^{n_1-1} + \cdots + b_{m,1}^{0} \\ \hdashline \vdots & & \vdots \\ \hdashline b_{1,p}^{n_p-1}\, z^{n_p-1} + \cdots + b_{1,p}^{0} & \cdots & b_{m,p}^{n_p-1}\, z^{n_p-1} + \cdots + b_{m,p}^{0} \end{array}\right] \begin{bmatrix} U_1(z) \\ \vdots \\ U_m(z) \end{bmatrix} \tag{4.2.1}$$

Premultiply both sides of Equation 4.2.1 by

$$\begin{bmatrix} z^{-n_1} & & 0 \\ & \ddots & \\ 0 & & z^{-n_p} \end{bmatrix} \tag{4.2.2}$$

to obtain the form

$$\begin{bmatrix} 1-a_{1,1}^{n_1-1}\,z^{-1}-\cdots-a_{1,1}^{0}\,z^{-n_1} & \ldots & -a_{p,1}^{n_1-1}\,z^{-1}-\cdots-a_{p,1}^{0}\,z^{-n_1} \\ \hline \vdots & & \vdots \\ \hline -a_{1,p}^{n_p-1}\,z^{-1}-\cdots-a_{1,p}^{0}\,z^{-n_p} & \ldots & 1-a_{p,p}^{n_p-1}\,z^{-1}-\cdots-a_{p,p}^{0}\,z^{-n_p} \end{bmatrix} \begin{bmatrix} Y_1(z) \\ \vdots \\ Y_p(z) \end{bmatrix} =$$

$$\begin{bmatrix} b_{1,1}^{n_1-1}\,z^{-1}+\cdots+b_{1,1}^{0}\,z^{-n_1} & \ldots & b_{m,1}^{n_1-1}\,z^{-1}+\cdots+b_{m,1}^{0}\,z^{n_1} \\ \hline \vdots & & \vdots \\ \hline b_{1,p}^{n_p-1}\,z^{-1}+\cdots+b_{1,p}^{0}\,z^{-n_p} & \ldots & b_{m,p}^{n_p-1}\,z^{-1}+\cdots+b_{m,p}^{0}\,z^{-n_p} \end{bmatrix} \begin{bmatrix} U_1(z) \\ \vdots \\ U_m(z) \end{bmatrix} \qquad (4.2.3)$$

The observable-canonic state diagram of the general MIMO system follows from Equation 4.2.3, and is given in Figure 6 on page 62. The parameter notation used in Equation 4.2.3 is based on the following ideas, which can be verified in the state diagram. Parameters associated with output feedback are *a*-parameters; those associated with input feedforward are *b*-parameters. Each subsystem is numbered according to the output with which it is associated.

Each parameter *subscript* has two numbers. The first number indicates which output (input) is being fed back (forward). The second number indicates the subsystem to which it feeds. The *superscript* indicates which state within the subsystem is the target of the feedback or feedforward.

To parallel the development of Section 4.1, it is now necessary to write the multivariable observable-canonic state equation, that is, the multivariable generalization of Equation 4.1.4. From the state diagram, the system equations are deduced, as shown in the following equations, which are in agreement with the outline of the canonic form, given in Chen [19]:

$$\begin{bmatrix} x_{1,1}(k+1) \\ \vdots \\ x_{1,n_1}(k+1) \\ \vdots \\ \vdots \\ x_{p,1}(k+1) \\ \vdots \\ x_{p,n_p}(k+1) \end{bmatrix} = \left[\begin{array}{l|c|l} 0 \ldots 0\; a^0_{1,1} & \ldots & 0 \ldots 0\; a^0_{p,1} \\ 1 \ldots 0\; a^1_{1,1} & \ldots & 0 \ldots 0\; a^1_{p,1} \\ \quad\cdot & & : \quad\; : \;\; : \\ \quad\;\cdot & & : \quad\; : \;\; : \\ 0 \ldots 1 a^{n_1-1}_{1,1} & \ldots & 0 \ldots 0 a^{n_1-1}_{p,1} \\ \hline \qquad\vdots & & \qquad\vdots \\ \hline 0 \ldots 0\; a^0_{1,p} & \ldots & 0 \ldots 0\; a^0_{p,p} \\ 0 \ldots 0\; a^1_{1,p} & \ldots & 1 \ldots 0\; a^1_{p,p} \\ : \quad : \;\; : & & \cdot \\ : \quad : \;\; : & & \quad\cdot \\ 0 \ldots 0\; a^{n_p-1}_{1,p} & \ldots & 0 \ldots 1\; a^{n_p-1}_{p,p} \end{array}\right] \begin{bmatrix} x_{1,1}(k) \\ \vdots \\ x_{1,n_1}(k) \\ \vdots \\ \vdots \\ x_{p,1}(k) \\ \vdots \\ x_{p,n_p}(k) \end{bmatrix} + \tag{4.2.4}$$

$$\left[\begin{array}{c|c|c} b^0_{1,1} & \ldots & b^0_{m,1} \\ \vdots & & \vdots \\ b^{n_1-1}_{1,1} & \ldots & b^{n_1-1}_{m,1} \\ \hline \vdots & & \vdots \\ \vdots & & \vdots \\ \hline b^0_{1,p} & \ldots & b^0_{m,p} \\ \vdots & & \vdots \\ b^{n_p-1}_{1,p} & \ldots & b^{n_p-1}_{m,p} \end{array}\right] \begin{bmatrix} u_1(k) \\ \vdots \\ \vdots \\ u_m(k) \end{bmatrix}$$

(or, in more compact notation, $x_{k+1} = A\, x_k + B\, u_k$).

$$\begin{bmatrix} y_1(k) \\ \vdots \\ \vdots \\ \vdots \\ y_p(k) \end{bmatrix} = \left[\begin{array}{c|c|c} 0 \dots 0\ 1 & \dots & 0 \dots 0\ 0 \\ \hline \vdots & \vdots & \vdots \\ \hline 0 \dots 0\ 0 & \dots & 0 \dots 0\ 1 \end{array}\right] \begin{bmatrix} x_{1,1}(k) \\ \vdots \\ x_{1,n_1}(k) \\ \vdots \\ \vdots \\ x_{p,1}(k) \\ \vdots \\ x_{p,n_p}(k) \end{bmatrix} \tag{4.2.5}$$

(or, in more compact notation, $y_k = C x_k$).

From Equation 4.2.4, it can be seen that the simplest requirements for complete controllability of the system are: $m \geq p$, and for all $i = 1, \dots, p$, the coefficients $b^0_{i,i}$, are not zero. This can also be deduced from the state diagram, given in Figure 6 on page 62, where it is equivalent to requiring that the first state in each of the p subsystems is fed directly by a unique input.

Define $x_{i,0}(k) \equiv 0, \ \forall\, i = 1, \dots, p$. In keeping with the development of Section 4.1, then the general form of the rows of Equations 4.2.4 and 4.2.5 is:

$$\begin{aligned} x_{i,j}(k+1) = \big[& x_{i,j-1}(k) + a^{j-1}_{1,i}\, x_{1,n_1}(k) + \dots + a^{j-1}_{p,i}\, x_{p,n_p}(k) \\ & + b^{j-1}_{1,i}\, u_1(k) + \dots + b^{j-1}_{m,i}\, u_m(k) \big], \\ & \forall\, i = 1, \dots, p,\ \forall\, j = 1, \dots, n_i, \end{aligned} \tag{4.2.6}$$

$$y_i(k) = x_{i,n_i}(k),\ \forall\, i = 1, \dots, p. \tag{4.2.7}$$

Substituting the identity of Equation 4.2.7 into Equation 4.2.6,

$$\begin{aligned} x_{i,j}(k+1) = \big[& x_{i,j-1}(k) + a^{j-1}_{1,i}\, y_1(k) + \dots + a^{j-1}_{p,i}\, y_p(k) \\ & + b^{j-1}_{1,i}\, u_1(k) + \dots + b^{j-1}_{m,i}\, u_m(k) \big], \\ & \forall\, i = 1, \dots, p,\ \forall\, j = 1, \dots, n_i. \end{aligned} \tag{4.2.8}$$

The *extended state vector* is now defined by appending the columns of unknown parameters in Equation 4.2.4 to the original state vector,

$$s_k \triangleq \begin{bmatrix} x_k \\ \theta_A^1 \\ \vdots \\ \theta_A^p \\ \theta_B^1 \\ \vdots \\ \theta_B^m \end{bmatrix}, \tag{4.2.9}$$

$$\text{where} \quad \theta_A^i \triangleq \begin{bmatrix} a_{i,1}^0 \\ \vdots \\ a_{i,1}^{n_1-1} \\ \vdots \\ \vdots \\ a_{i,p}^0 \\ \vdots \\ a_{i,p}^{n_p-1} \end{bmatrix}, \text{ and } \theta_B^i \triangleq \begin{bmatrix} b_{i,1}^0 \\ \vdots \\ b_{i,1}^{n_1-1} \\ \vdots \\ \vdots \\ b_{i,p}^0 \\ \vdots \\ b_{i,p}^{n_p-1} \end{bmatrix}. \tag{4.2.10}$$

Note that θ_A^i is the $(n_1 + \cdots + n_i)^{th}$-column of system matrix A; and θ_B^i is the i^{th}-column of input matrix B, in Equation 4.2.4. That is, the columns of parameters are appended to the state vector in the order they occur, from left to right, in Equation 4.2.4. Equations 4.2.9 and 4.2.10 are easily seen to be the multivariable generalization of Equation 4.1.10.

With Equation 4.2.9 defining the extended state vector, and with the general form of the rows identified in Equation 4.2.8, the *extended system* $\bar{S}$ can be written as

$$\begin{aligned} s_{k+1} &= F_k s_k \\ y_k &= H s_k \end{aligned} \tag{4.2.11}$$

where

$$F_k = \begin{array}{c}
\\ 1 \\ \\ 2 \\ 3 \\ \\ 4 \\ 5 \\ \\
\end{array}
\left[\begin{array}{ccccc|cccccc}
1 & & & & n & \overset{5}{n+1} & & \overset{9}{2n} & 2n+2 & 3n+1 & \\
\mathbf{J}_{n_1} & | & \ldots & | & 0 & & & & & & \\
 & & & & & y_1(k)\,\mathbf{I}_n & | \ldots & y_p(k)\,\mathbf{I}_n & u_1(k)\,\mathbf{I}_n & | \ldots & u_m(k)\,\mathbf{I}_n \\
\mathbf{0} & | & \ldots & | & \mathbf{J}_{n_p} & & & & & & \\
\hline
 & & & & & & & & & & \\
 & & \mathbf{0}_{(m+p)n\times n} & & & & & \mathbf{I}_{(m+p)n} & & & \\
 & & & & & & & & & &
\end{array}\right]$$

and

$$H = \left[\begin{array}{ccc|c|ccc|c}
0\ldots0 & 1 & & \ldots & 0\ldots0 & 0 & & \\
: & & & & : & & & \mathbf{0}_{p\times(m+p)n} \\
0\ldots0 & 0 & & \ldots & 0\ldots0 & 1 & &
\end{array}\right]$$

Note that in matrix F_k , $\mathbf{J}_{n_i}$ is an $n_i \times n_i$ lower Jordan block of zero eigenvalues.

Clearly, Equation 4.2.11 is the multivariable generalization of Equation 4.1.11. Note that there are $(m+p+1)n$ states in the extended system $\overline{S}$.

In Section 4.1, the discussion of the observability of the SISO extended system was brought down to a question of solving Equation 4.1.12, using the generalized inverse as in Equation 4.1.13. Both Equations 4.1.12 and 4.1.13 still hold for the MIMO case, after substitution of $p \times 1$ *output vectors* for the scalar outputs y_k, and substitution of the $p \times (m+p+1)n$ *output matrix* for the $1 \times 3n$ output row vector H.

However, Equation 4.1.14 no longer describes the extended observability matrix. Instead, it is described by a much more complex matrix, which is actually a generalization of Equation 4.1.14. This generalization is given in the following equation, for some time k when the matrix has full column rank.

$$\Phi_k = [\ \Phi_1(k)\ \mid\ \Phi_2(k)\ \mid\ \Phi_3(k)\ ] \tag{4.2.12}$$

where

$$\Phi_1(k) = \left[\begin{array}{c}
0 \dots 0\ 1 \mid \dots \mid 0 \dots 0\ 0 \\
\vdots \\
0 \dots 0\ 0 \mid \dots \mid 0 \dots 0\ 1 \\
\hdashline
0 \dots 1\ 0 \mid \dots \mid 0 \dots 0\ 0 \\
\vdots \\
0 \dots 0\ 0 \mid \dots \mid 0 \dots 1\ 0 \\
\hdashline
\vdots \\
\vdots \\
\hdashline
1\ 0 \dots 0 \mid \dots \mid 0 \dots 0\ 0 \\
\vdots \\
\vdots \\
\hdashline
0\ 0 \dots 0 \mid \dots \mid 0 \dots 0\ 0 \\
\vdots \\
\vdots \\
\hdashline
\vdots \\
\vdots \\
\hdashline
0 \dots 0\ 0 \mid \dots \mid 0 \dots 0\ 0 \\
\vdots \\
0 \dots 0\ 0 \mid \dots \mid 0 \dots 0\ 0
\end{array}\right],$$

(continued on the next page)

$\Phi_2(k)\ =$

$$\left[\begin{array}{ccccccccccccccccccccc}
0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & 0 & 0 & | & & | & 0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & 0 & 0 \\
 & & & & & & & & & | & \cdots & | & & & & & & & & & \\
0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & 0 & 0 & | & & | & 0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & 0 & 0 \\
\hdashline
0 & \cdots & 0 & y_1^0 & |\cdots| & 0 & \cdots & 0 & 0 & | & & | & 0 & \cdots & 0 & y_p^0 & |\cdots| & 0 & \cdots & 0 & 0 \\
 & & & & & & & & & | & \cdots & | & & & & & & & & & \\
0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & 0 & y_1^0 & | & & | & 0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & 0 & y_p^0 \\
\hdashline
0 & \cdots & y_1^0 & y_1^1 & |\cdots| & 0 & \cdots & 0 & 0 & | & & | & 0 & \cdots & y_p^0 & y_p^1 & |\cdots| & 0 & \cdots & 0 & 0 \\
 & & & & & & & & & | & \cdots & | & & & & & & & & & \\
0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & y_1^0 & y_1^1 & | & & | & 0 & \cdots & 0 & 0 & |\cdots| & 0 & \cdots & y_p^0 & y_p^1 \\
\hdashline
 & & & & \vdots & & & & & & & & & & & & \vdots & & & & \\
\hdashline
y_1^0 & \cdots & y_1^{n_1-1} & & |\cdots| & 0 & \cdots & 0 & 0 & | & & | & y_p^0 & \cdots & y_p^{n_1-1} & & |\cdots| & 0 & \cdots & 0 & 0 \\
 & & & & & & & & & | & \cdots & | & & & & & & & & & \\
 & & & & & & & & & | & \cdots & | & & & & & & & & & \\
\hdashline
 & & & & \vdots & & & & & & & & & & & & \vdots & & & & \\
\hdashline
y_1^{k-n_1} & \cdots & y_1^{k-1} & & |\cdots| & 0 & \cdots & 0 & 0 & | & & | & y_p^{k-n_1} & \cdots & y_p^{k-1} & & |\cdots| & 0 & \cdots & 0 & 0 \\
 & & & & & & & & & | & \cdots & | & & & & & & & & & \\
0 & \cdots & 0 & 0 & |\cdots| & y_1^{k-n_p} & \cdots & y_1^{k-1} & & | & & | & 0 & \cdots & 0 & 0 & |\cdots| & y_p^{k-n_p} & \cdots & y_p^{k-1} &
\end{array}\right],$$

(continued on the next page)

and

$$\Phi_3(k) =$$

$$\left[\begin{array}{ccccccccccccccccccccc}
0 & \dots & 0 & 0 & |\dots| & 0 & \dots & 0 & 0 & | & & | & 0 & \dots & 0 & 0 & |\dots| & 0 & \dots & 0 & 0 \\
 & & & & & & & & & | & \dots & | & & & & & & & & & \\
0 & \dots & 0 & 0 & |\dots| & 0 & \dots & 0 & 0 & | & & | & 0 & \dots & 0 & 0 & |\dots| & 0 & \dots & 0 & 0 \\
\hline
0 & \dots & 0 & u_1^0 & |\dots| & 0 & \dots & 0 & 0 & | & & | & 0 & \dots & 0 & u_p^0 & |\dots| & 0 & \dots & 0 & 0 \\
 & & & & & & & & & | & \dots & | & & & & & & & & & \\
0 & \dots & 0 & 0 & |\dots| & 0 & \dots & 0 & u_1^0 & | & & | & 0 & \dots & 0 & 0 & |\dots| & 0 & \dots & 0 & u_p^0 \\
\hline
0 & \dots & u_1^0 & u_1^1 & |\dots| & 0 & \dots & 0 & 0 & | & & | & 0 & \dots & u_p^0 & u_p^1 & |\dots| & 0 & \dots & 0 & 0 \\
 & & & & & & & & & | & \dots & | & & & & & & & & & \\
0 & \dots & 0 & 0 & |\dots| & 0 & \dots & u_1^0 & u_1^1 & | & & | & 0 & \dots & 0 & 0 & |\dots| & 0 & \dots & u_p^0 & u_p^1 \\
\hline
 & & & & \vdots & & & & & & & & & & & & \vdots & & & & \\
\hline
u_1^0 & \dots & & u_1^{n_1-1} & |\dots| & 0 & \dots & 0 & 0 & | & & | & u_p^0 & \dots & & u_p^{n_1-1} & |\dots| & 0 & \dots & 0 & 0 \\
 & & & & & & & & & | & \dots & | & & & & & & & & & \\
 & & & & & & & & & | & \dots & | & & & & & & & & & \\
\hline
 & & & & \vdots & & & & & & & & & & & & \vdots & & & & \\
\hline
u_1^{k-n_1} & \dots & & u_1^{k-1} & |\dots| & 0 & \dots & 0 & 0 & | & & | & u_p^{k-n_1} & \dots & & u_p^{k-1} & |\dots| & 0 & \dots & 0 & 0 \\
 & & & & & & & & & | & \dots & | & & & & & & & & & \\
0 & \dots & 0 & 0 & |\dots| & u_1^{k-n_p} & \dots & & u_1^{k-1} & | & & | & 0 & \dots & 0 & 0 & |\dots| & u_p^{k-n_p} & \dots & & u_p^{k-1}
\end{array}\right]$$

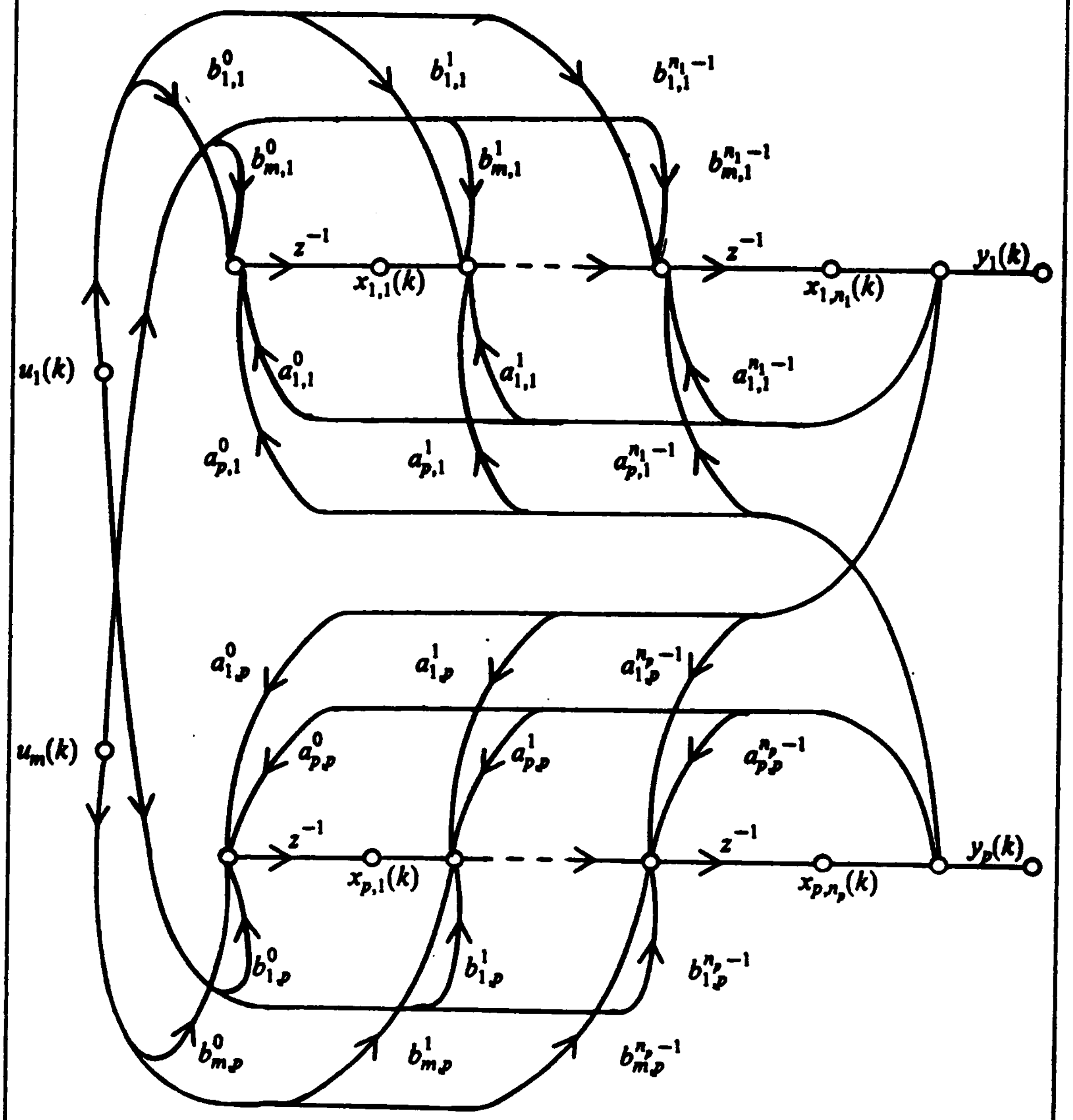

Figure 6. Observable-canonic form for the general MIMO system.

5.0 Stochastic PLID: The Optimal Joint Estimator

In this chapter, the pseudo-linear identification (PLID) algorithm is developed, starting with the ESR (extended state representation, also known as the Salut form), which was derived in Chapter 4.

In Section 5.1, the ESR is extended to the stochastic case. In the work of Salut *et.al.* [5], the ESR was also extended to the stochastic case. However, in that work the system was stochastic only in the following sense. It was assumed that a gaussian white noise vector with known autocovariance was added directly to the states of the system. This will be referred to as the *state noise vector.* No other noise was considered; in particular, it was assumed that the inputs and the outputs of the system were known exactly.

In Section 5.1, the extension of the ESR to the stochastic case includes the state noise vector of [5]. However, it also includes noise vectors corrupting the inputs and the outputs. Thus, the inputs and outputs are no longer known exactly. The input and output noise vectors are likewise assumed to be gaussian and white, with known autocovariances. The various noise vectors are allowed to be cross-correlated, provided that the cross-correlations are known. The complete stochastic ESR is given by Equation 5.1.6.

In Section 5.2, the basic form of the PLID algorithm is derived. First, it is proved that the state vector of the stochastic ESR is a Gauss-Markov process. That ensures the conditional mean estimator is equivalent to the linear minimum-mean-square-error (MMSE) estimator. What follows is a standard error covariance minimization procedure, but the resulting algorithm, given by Equations 5.2.21 through 5.2.23, contains terms not ordinarily seen in the equations for a recursive linear MMSE algorithm.

The unusual terms in Equations 5.2.21 and 5.2.23 are conditional mean estimates of matrices that would ordinarily be known. These matrices are not known because some elements within them are functions of the unknown parameters. In Section 5.3, a method of computing these conditional means is developed, resulting in the complete PLID algorithm, given by Equations 5.3.16 through 5.3.22.

Aside from the development of the algorithm, itself, the most important point of this chapter is that the PLID algorithm is a conditional mean estimator, and is therefore optimal in the mean-square-error sense.

5.1 *The Stochastic Extended State Representation*

The stochastic PLID algorithm derives from the extended state representation developed in Chapter 4. However, it is necessary to introduce several noise terms into that model, resulting in the stochastic extended state representation.

Assume the extended state propagation relation, given by Equation 4.2.4, has unmeasurable noise corrupting the input. Thus, the true input to the system is $u_k + v_k$, where the vector v_k is assumed to be a zero-mean, uncorrelated (white noise), gaussian random vector with auto-covariances Q_k, known for all k.

Also assume that a zero-mean, white, gaussian random vector Ξ_k adds *directly* to the state vector s_k in Equation 4.2.4. The autocovariance Σ_k of this noise is assumed known for each time k, and is assumed to be in block diagonal form, $\Sigma_k = block\ diag\,(\Sigma_k^{xx}, \Sigma_k^{\theta\theta})$. The block diagonal form indicates that the noise $\Xi_\theta(k)$ introduced into the parameters θ_k of the original system is independent of the noise $\Xi_x(k)$ introduced into the states x_k of the original system. Further assume that the parameter noise $\Xi_\theta(k)$ is independent of any other noise in the extended system.

Note that if the parameter noise auto-covariance matrix $\Sigma_k^{\theta\theta}$ is not zero, then the original system is time-varying. However, this model encompasses only the case where the parameter variation can be modeled by an independent, zero-mean, white, gaussian random vector $\Xi_\theta(k)$ adding directly to the parameter vector θ_k in the propagation of the parameter vector from time k to time $k+1$. Extension to the case where a deterministic input is also added to the parameters is straightforward, but is not pursued here.

Explicitly, the extended state noise vector is

$$\Xi_k = \begin{bmatrix} \Xi_x(k) \\ \hline \Xi_\theta(k) \end{bmatrix} = \begin{bmatrix} \xi_1(k) \\ \vdots \\ \xi_n(k) \\ \hline \xi_{n+1}(k) \\ \vdots \\ \xi_{2n}(k) \\ \vdots \\ \vdots \\ \xi_{n(p)+1}(k) \\ \vdots \\ \xi_{n(p+1)}(k) \\ \vdots \\ \vdots \\ \xi_{n(p+m)+1}(k) \\ \vdots \\ \xi_{n(p+m+1)}(k) \end{bmatrix} \tag{5.1.1}$$

Finally, assume that the output measurement, given by Equation 4.2.5, is corrupted by noise. Thus, the output is $z_k = y_k + w_k$, where w_k is assumed to be a zero-mean, white, gaussian random vector with autocovariance R_k, known for all k.

The independence of the *parameter* noise insures that all cross-covariances with $\Xi_\theta(k)$ are zero. However, there are other cross-covariances that may be non zero. These are assumed known, and are denoted as follows:

$$E\,[v_k\,w_k^T] \triangleq S_{vw}(k)\,, \qquad E\,[\Xi_x(k)\,w_k^T] \triangleq S_{\Xi_x w}(k)\,, \text{and} \quad E\,[\Xi_x(k)\,v_k] \triangleq S_{\Xi_x v}(k)\,. \tag{5.1.2}$$

Thus, the extended system of Equation 4.2.11 becomes

$$\begin{aligned} s_{k+1} &= F(y_k\,, u_k + v_k)s_k + \Xi_k \\ z_k &= y_k + w_k = H\,s_k + w_k \end{aligned} \tag{5.1.3}$$

Let $u_i^*(k) = u_i(k) + v_i(k)$, for each $i = 1, \ldots, m$, the noise-corrupted inputs. Thus, Equation 5.1.3 becomes

$$\begin{aligned} s_{k+1} &= F_k^* s_k + \Xi_k \\ y_k &= H s_k + w_k \quad \triangleleft \quad z_k = y_k + N_k \end{aligned} \tag{5.1.4}$$

where

$$F_k^* = \left[\begin{array}{ccc|ccccc} \mathbf{J}_{n_1} & \cdots & 0 & & & & & \\ & & & y_1(k)\,\mathbf{I}_n & \cdots & y_p(k)\,\mathbf{I}_n & u_1^*(k)\,\mathbf{I}_n & \cdots & u_m^*(k)\,\mathbf{I}_n \\ 0 & \cdots & \mathbf{J}_{n_p} & & & & & \\ \hline & & & & & & & \\ \multicolumn{3}{c|}{\mathbf{0}_{(m+p)n\times n}} & \multicolumn{5}{c}{\mathbf{I}_{(m+p)n}} \\ & & & & & & & \end{array}\right]$$

$$H = \left[\begin{array}{c|c|c|c} 0\ldots0\,1 & \ldots & 0\ldots0\,0 & \\ \vdots & & \vdots & \mathbf{0}_{p\times(m+p)n} \\ 0\ldots0\,0 & \ldots & 0\ldots0\,1 & \end{array}\right].$$

By making the substitution $y_i(k) = z_i(k) - w_i(k)$, for all $i = 1, \ldots, p$, and using the parameter vector definitions of Equation 4.2.10, Equation 5.1.4 can be reorganized as follows:

$$\begin{aligned} s_{k+1} &= F_k s_k + \left[\begin{array}{c|c|c|c|c|c} -\theta_A^1 & & -\theta_A^p & \theta_B^1 & & \theta_B^m \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{array}\right] \begin{bmatrix} w_k \\ v_k \end{bmatrix} + \Xi_k \\ y_k &= H s_k + w_k \end{aligned} \tag{5.1.5}$$

$z_k =$

where

$$
F_k = \left[\begin{array}{ccc|c}
\mathbf{J}_{n_1} & \mid \cdots \mid & 0 & \\
 & & & z_1(k)\,\mathbf{I}_n \mid \cdots \mid z_p(k)\,\mathbf{I}_n \mid u_1(k)\,\mathbf{I}_n \mid \cdots \mid u_m(k)\,\mathbf{I}_n \\
0 & \mid \cdots \mid & \mathbf{J}_{n_p} & \\
\hline
 & \mathbf{0}_{(m+p)n} & & \mathbf{I}_{(m+p)n}
\end{array}\right]
$$

Hence, the stochastic extended state representation can be written,

$$
\begin{aligned}
s_{k+1} &= F_k\, s_k + G_k\, \eta_k \\
z_k &= H\, s_k + w_k
\end{aligned}
\qquad (5.1.6)
$$

where

$$
G_k = \left[\begin{array}{c|c}
-\,\theta_A^1(k) \mid \cdots \mid -\,\theta_A^p(k) \mid \theta_B^1(k) \mid \cdots \mid \theta_B^m(k) \mid \mathbf{I}_n & \mathbf{0} \\
\hline
\mathbf{0} & \mathbf{I}_{(m+p)n}
\end{array}\right],
$$

$$
\eta_k = \begin{bmatrix} w_k \\ v_k \\ \Xi_x(k) \\ \Xi_\theta(k) \end{bmatrix}, \quad
Q_k^0 \triangleq E\{\eta_k\, \eta_k^T\} =
\left[\begin{array}{ccc|c}
R_k & S_{v\,w}^T(k) & S_{\Xi_x\,w}^T(k) & \\
S_{v\,w}(k) & Q_k & S_{\Xi_x\,v}^T(k) & \mathbf{0} \\
S_{\Xi_x\,w}(k) & S_{\Xi_x\,v}(k) & \Sigma_k^{xx} & \\
\hline
 & \mathbf{0} & & \Sigma_k^{\theta\theta}
\end{array}\right]
$$

and F_k is the same as in Equation 5.1.5.

It is an important fact that the noise input matrix G_k is a *linear* function of the unknown parameters, easily verified by inspection of Equation 5.1.6.

From Equation 5.1.6, it is also clear that the input noise η_k and the output noise w_k in the stochastic extended state representation are not independent, even if the input noise v_k and the output noise w_k in the *original* unknown system are independent. The cross-covariance is

$$S_k \triangleq E\{\eta_k w_k^T\} = \left[\begin{array}{c} R_k \\ \hline S_{v\,w}(k) \\ \hline S_{\Xi_x\,w}(k) \\ \hline 0_{(m+p)n\times p} \end{array}\right], \tag{5.1.7}$$

which is equivalent to the first p columns of $Q_0(k)$.

5.2 Essential Form of the PLID Algorithm

To develop the PLID algorithm, it is necessary to assume that the unknown system parameters are time-invariant. It is necessary because that assumption results in a Gauss-Markov extended system.

Lemma 5.2.1: Consider the extended state representation $\bar{S}$ of Equation 4.2.11. If

(1) the underlying linear system S of Equation 4.2.4 is time-invariant (*i.e.*, $\Sigma_k^{\theta\theta} \equiv 0$),

(2) input noise vector v_k, output noise vector w_k , and state noise vector $\Xi_x(k)$ are zero mean white gaussian random vectors, and

(3) initial state x_0 of S is gaussian distributed,

then the state vector s_k of the stochastic extended state representation is a jointly gaussian first-order Markov process.

Proof: The proof consists of two parts: first it is shown that the process is first-order Markov; then it is shown that s_k is jointly gaussian.

To show the Markov property, write the expected value of the extended state vector at time $k+1$ conditioned on all the state and input data up to time k:

$$\begin{aligned}
&E\{s_{k+1} \,|s_k\,, u_k\,, s_{k-1}\,, u_{k-1}\,, \dots, s_0\,, u_0\} \\
&\qquad = E\{F_k\, s_k + G\,\eta_k \,|s_k\,, u_k\,, \dots, s_0\,, u_0\} \\
&\qquad = E\{F_k(z_k\,, u_k)\, s_k + G\,(s_k)\,\eta_k \,|s_k\,, u_k\,, \dots, s_0\,, u_0\} \\
&\qquad = E\{F_k(H\, s_k + w_k\,, u_k)\, s_k + G\,(s_k)\eta_k \,|s_k\,, u_k\,, \dots, s_0\,, u_0\} \\
&\qquad = E\{F_k(H\, s_k + w_k\,, u_k)\, s_k + G\,(s_k)\,\eta_k \,|s_k\,, u_k\} \\
&\qquad = E\{s_{k+1} \,|s_k\,, u_k\}
\end{aligned} \tag{5.2.1}$$

with the second to last equality due to the premise that η_k and w_k are independent of the input sequence $\{u_k\}$ and the state trajectory $\{s_k\}$. Therefore, by definition, s_k is a first-order Markov process.

To show that s_k is jointly gaussian, note first that the state vector $\{x_k\}$ is jointly gaussian (because S is a linear time-invariant system with deterministic and gaussian random sequences as input, and a gaussian initial state [20]).

Now, $s_k = [x_k^T \quad \theta^T]^T$, where θ is the vector of constant parameters. So

$$Cov\,[s_k] = \begin{bmatrix} Cov\,[x_k] & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.2.2}$$

The characteristic function is

$$\begin{aligned} E\left[e^{j\omega^T s_k}\right] &= E\left[e^{j\left[\omega_1^T \mid \omega_2^T\right]\left[\begin{smallmatrix} x_k \\ \theta \end{smallmatrix}\right]}\right] = E\left[e^{j\omega_1^T x_k + j\omega_2^T \theta}\right] \\ &= E\left[e^{j\omega_1^T x_k}\right] e^{j\omega_2^T \theta}. \end{aligned}$$

x_k is known to be jointly gaussian, so $E\left[e^{j\omega_1^T x_k}\right] = e^{j\omega_1^T E\,[x_k]}\; e^{(1/2)\,\omega_1^T Cov[x_k]\,\omega_1}$.

Hence,
$$\begin{aligned} E\left[e^{j\omega^T s_k}\right] &= e^{j\omega_1^T E\,[x_k]}\; e^{(1/2)\,\omega_1^T Cov[x_k]\,\omega_1}\; e^{j\omega_2^T \theta} \\ &= e^{j\omega_1^T E\,[x_k] + j\omega_2^T \theta}\; e^{(1/2)\,\omega_1^T Cov[x_k]\,\omega_1} \\ &= e^{j\omega^T E\,[s_k]}\; e^{(1/2)\,\omega_1^T Cov[x_k]\,\omega_1}. \end{aligned} \tag{5.2.3}$$

But, by Equation 5.2.2, $\omega_1^T\, Cov\,[x_k]\,\omega_1 = \omega^T\, Cov\,[s_k]\,\omega$. So

$$E\left[e^{j\omega^T s_k}\right] = e^{j\omega^T E\,[s_k]}\; e^{(1/2)\,\omega^T Cov[s_k]\,\omega}. \tag{5.2.4}$$

Then, by definition, s_k is a jointly distributed gaussian random vector [21]. ■

Knowing that s_k is jointly gaussian distributed, it is obvious that $z_k = H\,s_k + w_k$ is jointly gaussian distributed. It follows that the composite vector $[\; s_{k+1}^T \mid z_k^T \;]^T$ is also a jointly distributed gaussian random vector.

In the development of the PLID algorithm, the jointly gaussian nature of s_k and z_k is crucial, because it implies that the conditional linear minimum-mean-square-error

(MMSE) estimator is equivalent to the conditional mean estimator. It is well known that if S and Z are jointly gaussian, then the *linear* MMSE estimate of S given Z is equivalent to *the* MMSE (or minimum variance) estimate of S given Z [22]. But it is also true, for any joint distribution of S and Z, that the MMSE estimate of S given Z is equivalent to the conditional mean estimate of S given Z, as shown next.

Lemma 5.2.2: Suppose S and Z are jointly distributed random vectors.

The conditional minimum variance estimate of S given Z

is equivalent to the conditional mean estimate, $E[S|Z]$.

Proof: (from Anderson and Moore [23]) Let $\zeta(Z)$ be some estimate of S that depends on Z and not on S. Hence,

$$E\{\|S-\zeta\|_2^2 \; |Z=z\} = \int_{-\infty}^{\infty} (s-\zeta)^T(s-\zeta)\, p_{S|Z}(s|z)\, dx$$

$$= \int_{-\infty}^{\infty} [s^T s - 2\zeta^T x + \zeta^T \zeta\,]\, p_{S|Z}(s|z)\, dx$$

$$= \zeta^T \zeta - 2\zeta^T \int_{-\infty}^{\infty} s\, p_{S|Z}(s|z)\, ds + \int_{-\infty}^{\infty} s^T s\, p_{S|Z}(s|z)\, ds \tag{5.2.5}$$

$$= \left[\zeta^T - \int_{-\infty}^{\infty} s^T p_{S|Z}(s|z)\, ds\right]\left[\zeta - \int_{-\infty}^{\infty} s\, p_{S|Z}(s|z)\, ds\right]$$

$$+ \int_{-\infty}^{\infty} s^T s\, p_{S|Z}(s|z)\, ds \; - \left\|\int_{-\infty}^{\infty} s\, p_{S|Z}(s|z)\, ds\,\right\|^2$$

Equation 5.2.5 gives an expression for the norm of the conditional estimate error covariance. Considering the right side to be a function of ζ, the norm of the covariance is minimized if and only if

$$\zeta = \int_{-\infty}^{\infty} s\, p_{S|Z}(s|z)\, ds = E[\, S \,|Z = z\,] \tag{5.2.6}$$

i.e., if and only if the MMSE estimate is the conditional mean. Furthermore, the minimized covariance is the conditional covariance. ■

Now, we want to compute the linear conditional minimum-mean-square-error recursive prediction $\hat{s}_{k+1|k}$ of the state s_{k+1} of the stochastic extended state representation, given by Equation 5.1.6, using the form $\hat{s}_{k+1|k} = M_k\, \hat{s}_{k|k-1} + \overline{K}_k\, z_k$. Note $\overline{K}_k$ is a gain matrix chosen to minimize the conditional mean square error of the prediction.

The prediction is to be conditioned on the increasing sequence of sub-σ-fields ψ_k generated by the increasing sequence of sets of measurement data $\{\, z_0 \,, \dots , z_k \}$. As shown above, this prediction is also the conditional mean; that is, $\hat{s}_{k+1|k} = E\,(s_{k+1} \,|\, \psi_k\,)$ (also denoted $E_{\psi_k}(s_{k+1}\,)$).

The predictor is required to be unbiased; *i.e.*, the expectations

$$\begin{aligned} E\,[s_{k+1} \,|\, s_k\,] = E\,[\hat{s}_{k+1|k} \,|\, s_k\,] &= E\,[M_k\, \hat{s}_{k|k-1} + \overline{K}_k\, z_k \,|\, s_k\,] \\ F_k\, s_k &= M_k\, s_k + \overline{K}_k\, H\, s_k \end{aligned} \tag{5.2.7}$$

Therefore, $\quad F_k = M_k + \overline{K}_k\, H \quad \Rightarrow \quad M_k = F_k - \overline{K}_k\, H\,.$ (5.2.8)

Equation 5.2.8 implies the unbiased predictor should take the form

$$
\begin{aligned}
\hat{s}_{k+1|k} &= (F_k - \overline{K}_k H)\, \hat{s}_{k|k-1} + \overline{K}_k z_k \\
&= F_k \hat{s}_{k|k-1} + \overline{K}_k (z_k - H \hat{s}_{k|k-1})
\end{aligned}
\tag{5.2.9}
$$

Now define the prediction error, recalling that the extended state representation has the input noise vector $\eta_k = [w_k^T \,|\, v_k^T \,|\, \Xi_k^T]^T$,

$$
\begin{aligned}
e_{k+1|k} &\triangleq \hat{s}_{k+1|k} - s_{k+1} \\
&= (F_k - \overline{K}_k H)\, \hat{s}_{k|k-1} + \overline{K}_k z_k - s_{k+1} \\
&= (F_k - \overline{K}_k H)\, \hat{s}_{k|k-1} + \overline{K}_k (H s_k + w_k) - (F_k s_k + G \eta_k) \\
&= (F_k - \overline{K}_k H) (\hat{s}_{k|k-1} - s_k) + \overline{K}_k w_k - G \eta_k \\
&= (F_k - \overline{K}_k H)\, e_{k|k-1} + \overline{K}_k w_k - G \eta_k
\end{aligned}
\tag{5.2.10}
$$

Define the error covariance, again conditioning on the increasing sequence of sub-σ-fields ψ_k, by

$$
\begin{aligned}
P_{k+1|k} &\triangleq E\Big[e_{k+1|k}\, e_{k+1|k}^T \,|\, \psi_k \Big] \\
&= E\Big[(F_k - \overline{K}_k H)\, e_{k|k-1}\, e_{k|k-1}^T (F_k - \overline{K}_k H)^T + K_{k|k} w_k w_k^T \overline{K}_k^T \\
&\quad + G \eta_k \eta_k^T G^T - \overline{K}_k w_k \eta_k^T G^T - G \eta_k w_k^T \overline{K}_k^T \\
&\quad - (F_k - \overline{K}_k H)\, e_{k|k-1} \eta_k^T G^T - G \eta_k e_{k|k-1}^T (F_k - \overline{K}_k H)^T \\
&\quad + (F_k - \overline{K}_k H)\, e_{k|k-1} w_k^T \overline{K}_k^T + \overline{K}_k w_k e_{k|k-1}^T (F_k - \overline{K}_k H)^T \,|\, \psi_k \Big]
\end{aligned}
\tag{5.2.11}
$$

Taking the expectations of the various terms, using $E_{\psi_k}(\ )$ to indicate conditional expectation relative to ψ_k,

$$\begin{aligned}
&P_{k+1|k} = \\
&\quad \Big[(F_k - \overline{K}_k H)\, P_{k|k-1}\, (F_k - \overline{K}_k H)^T + \overline{K}_k R_k\, \overline{K}_k^T + E_{\psi_k}(G\, Q_k^0\, G^T) \\
&\qquad - \overline{K}_k S_k^T E_{\psi_k}(G^T) - E_{\psi_k}(G) S_k \overline{K}_k^T \\
&\qquad - (F_k - \overline{K}_k H)\, E_{\psi_k}(e_{k|k-1}\, \eta_k^T G^T) - E_{\psi_k}(G\, \eta_k\, e_{k|k-1}^T)(F_k - \overline{K}_k H)^T \\
&\qquad + (F_k - \overline{K}_k H)\, E_{\psi_k}(e_{k|k-1}\, w_k^T) \overline{K}_k^T + \overline{K}_k\, E_{\psi_k}(w_k\, e_{k|k-1}^T)(F_k - \overline{K}_k H)^T \Big]
\end{aligned} \tag{5.2.12}$$

The expectations in the last four terms need to be evaluated:

$$\begin{aligned}
E_{\psi_k}(G\, \eta_k\, e_{k|k-1}^T) &= [\, E_{\psi_k}(e_{k|k-1}\, \eta_k^T G^T)\,]^T \\
&= E_{\psi_k}(G)\, E_{\psi_k}\Big[\eta_k \{ e_{k-1|k-2}^T (F_{k-1} - \overline{K}_{k-1} H)^T \\
&\qquad\qquad + w_{k-1}^T\, \overline{K}_{k-1}^T - \eta_{k-1}^T G \} \Big] \\
&= E_{\psi_k}(G) \Big[\, E_{\psi_k}(\eta_k\, e_{k-1|k-2}^T)(F_{k-1} - \overline{K}_{k-1} H)^T \\
&\qquad\qquad + E(\eta_k\, w_{k-1}^T) \overline{K}_{k-1}^T + E(\eta_k\, \eta_{k-1}^T)\, E_{\psi_k}(G) \Big] \\
&= E_{\psi_k}(G) \times 0 = 0
\end{aligned} \tag{5.2.13}$$

$$\begin{aligned}
E_{\psi_k}(w_k\, e_{k|k-1}^T) &= [\, E_{\psi_k}(e_{k|k-1}\, w_k^T)\,]^T \\
&= E_{\psi_k}\Big[w_k \{ e_{k-1|k-2}^T (F_{k-1} - \overline{K}_{k-1} H)^T \\
&\qquad\qquad + w_{k-1}^T\, \overline{K}_{k-1}^T - \eta_{k-1}^T G \} \Big] \\
&= \Big[\, E_{\psi_k}(w_k\, e_{k-1|k-2}^T)(F_{k-1} - \overline{K}_{k-1} H)^T \\
&\qquad\qquad + E(w_k\, w_{k-1}^T) \overline{K}_{k-1}^T + E(w_k\, \eta_{k-1}^T)\, E_{\psi_k}(G) \Big] \\
&= 0
\end{aligned} \tag{5.2.14}$$

So, applying Equations 5.2.13 and 5.2.14 to Equation 5.2.12,

$$\begin{aligned}
P_{k+1|k} = \Big[&(F_k - \overline{K}_k H)\, P_{k|k-1}\, (F_k - \overline{K}_k H)^T + \overline{K}_k R_k\, \overline{K}_k^T \\
&+ E_{\psi_k}(G\, Q_k^0\, G^T) - \overline{K}_k S_k^T E_{\psi_k}(G^T) - E_{\psi_k}(G) S_k \overline{K}_k^T \Big]
\end{aligned} \tag{5.2.15}$$

The gain $\overline{K}_k$ is to be computed so as to minimize the mean-square-error of the prediction $\hat{s}_{k+1|k}$, so choose the functional

$$L_{k+1|k} \triangleq E[\, e_{k+1|k}^T \, e_{k+1|k} \mid \psi_k \,] = \mathrm{tr}\,(P_{k+1|k}) \tag{5.2.16}$$

The minimizing gain, denoted $K_{k|k}$, is found by solving

$$\left[\frac{\partial}{\partial \overline{K}_k} \, L_{k+1|k} \right]_{\overline{K}_k = K_{k|k}} = 0 \tag{5.2.17}$$

Substituting Equations 5.2.15 and 5.2.16 into 5.2.17,

$$\begin{aligned}
0 = & \Big[\frac{\partial}{\partial \overline{K}_k} \; \mathrm{tr}\,\{F_k \, P_{k|k-1} \, F_k^T + E_{\psi_k}(\, G \, Q_k^0 \, G^T \,) + \overline{K}_k \, (H \, P_{k|k-1} \, H^T + R_k \,) \overline{K}_k^T \\
& - [F_k \, P_{k|k-1} \, H^T + E_{\psi_k}(\, G \,) S_k \,] \overline{K}_k^T - \overline{K}_k \, [H \, P_{k|k-1} \, F_k^T + S_k^T \, E_{\psi_k}(\, G^T \,) \,] \,\} \Big]_{\overline{K}_k = K_{k|k}} \\
= & \Big[\frac{\partial}{\partial \overline{K}_k} \; \{ \mathrm{tr}\,[\overline{K}_k \, (H \, P_{k|k-1} \, H^T + R_k \,) \overline{K}_k^T \,] \\
& \qquad - 2 \, \mathrm{tr}\,[(F_k \, P_{k|k-1} \, H^T + E_{\psi_k}[\, G \,] \, S_k \,) \overline{K}_k^T \,] \,\} \Big]_{\overline{K}_k = K_{k|k}} \\
= & \; 2 \, K_{k|k} \, (H \, P_{k|k-1} \, H^T + R_k \,) - 2 \, [F_k \, P_{k|k-1} \, H^T + E_{\psi_k}(\, G \,) \, S_k \,]
\end{aligned} \tag{5.2.18}$$

Thus,

$$K_{k|k} = [F_k \, P_{k|k-1} \, H^T + E_{\psi_k}(\, G \,) \, S_k \,] \, (H \, P_{k|k-1} \, H^T + R_k \,)^{-1} \tag{5.2.19}$$

Now back-substitution of Equation 5.2.19 into 5.2.15 yields:

$$P_{k+1|k} = F_k \, P_{k|k-1} \, F_k^T + E_{\psi_k}(G \, Q_k^0 \, G^T) - K_{k|k} \, (H \, P_{k|k-1} \, H^T + R_k) K_{k|k}^T \tag{5.2.20}$$

Choosing $P_{0|-1}$ symmetric positive definite and for an arbitrary choice of initial estimate $\hat{s}_{0|-1}$, the predictor algorithm for $k \geq 0$ is:

$$K_{k|k} = [\, F_k \, P_{k|k-1} \, H^T + E(\, G \mid \psi_k)S_k \,](\, H \, P_{k|k-1} \, H^T + R_k \,)^{-1} \tag{5.2.21}$$

$$\hat{s}_{k+1|k} = (\, F_k - K_{k|k} \, H \,)\, \hat{s}_{k|k-1} + K_{k|k} \, z_k \tag{5.2.22}$$

$$\begin{aligned} P_{k+1|k} = {} & F_k \, P_{k|k-1} \, F_k^T + E(\, G \, Q_k^0 \, G^T \mid \psi_k \,) \\ & - K_{k|k}(\, H \, P_{k|k-1} \, H^T + R_k \,) K_{k|k}^T \end{aligned} \tag{5.2.23}$$

Equations 5.2.21 through 5.2.23 are the defining equations of the PLID algorithm. However, more work is required to specify the two expectation terms, $E(\, G \mid \psi_k)$ and $E(\, G \, Q_k^0 \, G^T \mid \psi_k)$ in Equations 5.2.21 and 5.2.23, as discussed in Section 5.3.

5.3 The PLID Algorithm for Time-Invariant Systems

In this section, the final, implementable PLID algorithm is developed for the case where the original unknown system is time-invariant. The development begins from the outline of the PLID algorithm given by Equations 5.2.21 through 5.2.23. Recall that requiring time-invariance in the original system is equivalent to requiring that the parameter noise $\Xi_\theta(k)\equiv 0$ for all k. Therefore, $\Sigma_k^{\theta\theta}\equiv 0$ for all k.

Recall $\{\psi_k\}$ is the increasing sequence of sub-σ-fields generated by the increasing sequence of sets $\{ z_0, \dots ,z_k \}$. This is the same sequence of sub-σ-fields to which the estimates

$$\hat{s}_{k+1|k} = \begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{\theta}_{k+1|k} \end{bmatrix} = \begin{bmatrix} E(x_{k+1} \mid \psi_k) \\ E(\theta \mid \psi_k) \end{bmatrix} \tag{5.3.1}$$

are adapted. Thus,

$$E[G \mid \psi_k] = E[G(\theta) \mid \psi_k] = G(E[\theta \mid \psi_k]) = G(\hat{\theta}_{k+1|k}) \tag{5.3.2}$$

with the second equality due to the fact that $G(\theta)$ is a linear function of θ.

At first glance, this computation does not appear possible because, by Equation 5.2.22, $\hat{s}_{k+1|k}$ depends upon $E(G \mid \psi_k)$, and so depends upon itself. However, due to the unique structure of the matrix G, only those elements of G related to the *state* estimates $\hat{x}_{k+1|k}$ actually depend upon θ, while those elements related to the parameter estimates $\hat{\theta}_{k+1|k}$ are all known constants (see Equation 5.1.6). This allows $\hat{\theta}_{k+1|k}$ to be computed *before* computing the unknown part of $E(G \mid \psi_k)$.

Define the *intermediate* gain matrix

$$\tilde{K}_k = F_k P_{k|k-1} H^T (H P_{k|k-1} H^T + R_k)^{-1} \tag{5.3.3}$$

Decompose Equation 5.3.3 according to those elements related to the state estimates and those related to the parameter estimates:

$$\tilde{K}_k = \begin{bmatrix} \tilde{K}_k^x \\ \tilde{K}_k^\theta \end{bmatrix}. \tag{5.3.4}$$

But, by Equations 5.1.6 and 5.1.7, $G^\theta S_k = 0$. So, from the gain computation, Equation 5.2.21, because $G S_k = G [R_k \mid 0]^T$,

$$K_{k|k}^\theta = \tilde{K}_k^\theta \tag{5.3.5}$$

Now the conditional expectation of the parameters can be computed. Define the *intermediate* extended state estimate,

$$\tilde{s}_{k+1} = \begin{bmatrix} \tilde{x}_{k+1} \\ \tilde{\theta}_{k+1} \end{bmatrix} = F_k \hat{s}_{k|k-1} + \tilde{K}_k (z_k - H \hat{s}_{k|k-1}) \tag{5.3.6}$$

Then, because $K_{k|k}^\theta = \tilde{K}_k^\theta$, we have

$$\hat{\theta}_{k+1|k} = \tilde{\theta}_{k+1} \tag{5.3.7}$$

Now $E(G^x \mid \psi_k) = G^x (\hat{\theta}_{k+1|k})$ is "computed" simply by substituting the various parameter estimates into the appropriate element of the matrix $E(G \mid \psi_k)$, premultiplying by -1 in some cases (see Equations 5.1.6 and 4.2.10).

So the rest of the gain matrix can be computed by

$$K^x_{k|k} = \tilde{K}^x_k + E_{\psi_k}(G^x)\, S_k \;\; (H\, P_{k|k-1}\, H^T + R_k\,)^{-1}\,. \tag{5.3.8}$$

Similarly, the rest of the state estimate is computed by

$$\hat{x}_{k+1|k} = \tilde{x}_{k+1} + E_{\psi_k}(G^x)\, S_k \;\; (H\, P_{k|k-1}\, H^T + R_k\,)^{-1}\,(\,z_k - H\,\hat{s}_{k|k-1}\,)\,. \tag{5.3.9}$$

At this point, the gain matrix of Equation 5.2.21 and the extended state estimate of Equation 5.2.22 have been computed completely. It remains only to compute the conditional error covariance matrix of Equation 5.2.23.

But every term in Equation 5.2.23 is now specified except $E(G\,Q^{\theta}_k G \mid \psi_k)$. Before proceeding with the method of computing this term, let us define some useful notation. Decompose the state and parameter noise covariance matrix Σ_k (which is assumed known), and the conditional error covariance matrix $P_{k+1|k}$, into

$$\Sigma_k = [\sigma_{i,j}(k)\,] = \begin{bmatrix} \Sigma^{xx}_k & 0 \\ 0 & \Sigma^{\theta\theta}_k = 0 \end{bmatrix} \text{ and } P_{k+1|k} = [\,p_{i,j}(k+1|k)\,] = \begin{bmatrix} P^{xx}_{k+1|k} & P^{x\theta}_{k+1|k} \\ P^{\theta x}_{k+1|k} & P^{\theta\theta}_{k+1|k} \end{bmatrix}$$

Further, decompose the lower right submatrix $P^{\theta\theta}_{k+1|k}$ into $n \times n$ submatrices,

$$P^{\theta\theta}_{k+1|k} =$$

$$\begin{bmatrix} P^{\theta\theta}_{1,1}(k+1|k) & & P^{\theta\theta}_{1,p}(k+1|k) & & P^{\theta\theta}_{1,p+m}(k+1|k) \\ \vdots & & \vdots & & \vdots \\ P^{\theta\theta}_{p,1}(k+1|k) & & P^{\theta\theta}_{p,p}(k+1|k) & & P^{\theta\theta}_{p,p+m}(k+1|k) \\ P^{\theta\theta}_{p+1,1}(k+1|k) & \cdots & P^{\theta\theta}_{p+1,p}(k+1|k) & \cdots & P^{\theta\theta}_{p+1,p+m}(k+1|k) \\ \vdots & & \vdots & & \vdots \\ P^{\theta\theta}_{p+m,1}(k+1|k) & & P^{\theta\theta}_{p+m,p}(k+1|k) & & P^{\theta\theta}_{p+m,p+m}(k+1|k) \end{bmatrix}. \tag{5.3.10}$$

Appendix F gives the rather messy algebraic expansion of $E_{\psi_k}[G\, Q_k^0\, G^T]$, resulting in the following equation:

$$E_{\psi_k}\left[(G\eta_k)(G\eta_k)^T\right] = \left[\begin{array}{c|c} \bigstar_{(n\times n)} & 0_{n\times(m+p)n} \\ 0_{(m+p)n\times n} & 0_{(m+p)n\times(m+p)n} \end{array}\right] \tag{5.3.11}$$

where

$$\begin{aligned} \bigstar_{(n\times n)} = {} & \sum_{i=1}^{p}\sum_{j=1}^{p} P_{i,j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} P_{p+i,p+j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{p+i,p+j} \\ & -\sum_{i=1}^{p}\sum_{j=1}^{m} P_{i,p+j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{i,p+j} - \sum_{i=1}^{m}\sum_{j=1}^{p} P_{p+i,j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{p+i,j} \\ & + E_{\psi_k}(G^x)\,Q_k^0\,E_{\psi_k}[(G^x)^T] \end{aligned} \tag{5.3.12}$$

Just as in the computation of the gain matrix, at first glance it appears that $P_{k+1|k}$ depends upon itself, and is therefore not computable. However, what really occurs is $P_{k+1|k}^{xx}$ depends upon $P_{k+1|k}^{\theta\theta}$, just as in the computation of $K_{k|k}$, it happens that $K_{k|k}^x$ depends upon $\hat{\theta}_{k+1|k}$. So it is necessary to compute $P_{k+1|k}^{\theta\theta}$ *first.*

Define the *intermediate* covariance matrix

$$\begin{aligned} \tilde{P}_{k+1|k} = {} & F_k\, P_{k|k-1}\, F_k^T + E[G \mid \psi_k]\, Q_k^0\, E[G^T \mid \psi_k] \\ & - K_{k|k}\,(H\, P_{k|k-1}\, H^T + R_k\,)K_{k|k}^T \end{aligned} \tag{5.3.13}$$

Because of the form of the matrix composing Equation 5.3.11, it is clear from Equation 5.2.23 that

$$\begin{aligned} & P_{k+1|k}^{\theta\theta} = \tilde{P}_{k+1|k}^{\theta\theta} \\ \text{and}\quad & [\,P_{k+1|k}^{x\theta}\,]^T = P_{k+1|k}^{\theta x} = \tilde{P}_{k+1|k}^{\theta x}\ . \end{aligned} \tag{5.3.14}$$

The rest of the conditional error covariance matrix can be computed

$$\begin{aligned}P^{xx}_{k+1|k} = \widetilde{P}^{xx}_{k+1|k} &+ \sum_{i=1}^{p}\sum_{j=1}^{p} P^{\theta\theta}_{i,j}(k+1|k)\,(Q^0_k)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} P^{\theta\theta}_{p+i,p+j}(k+1|k)\,(Q^0_k)_{p+i,p+j} \\ &- \sum_{i=1}^{p}\sum_{j=1}^{m} P^{\theta\theta}_{i,p+j}(k+1|k)\,(Q^0_k)_{i,p+j} - \sum_{i=1}^{m}\sum_{j=1}^{p} P^{\theta\theta}_{p+i,j}(k+1|k)\,(Q^0_k)_{p+i,j}\end{aligned} \tag{5.3.15}$$

5.3.1 Summary of the PLID Algorithm

A summary of the complete stochastic PLID algorithm is given by Equations 5.3.16 through 5.3.22. The summary simply restates the following equations, with a slight change of notation to indicate the computer implementation: Equation 5.3.3; Equation 5.3.6; the computation of $E(G \mid \psi_k)$ discussed just after Equation 5.3.7; Equation 5.3.8; Equation 5.3.9; Equation 5.3.13; and Equation 5.3.15.

With no *a priori* knowledge of the parameters and states or their probability density functions, the best initial estimate of the extended state vector is $\hat{s}_{0|-1} = 0$. Similarly, lacking any foreknowledge, a reasonable choice of the initial error covariance matrix is $P_{0|-1} = \alpha_0 I$, where $\alpha_0 \gg 1$.

$$K_{k|k} = \begin{bmatrix} K^{x}_{k|k} \\ K^{\theta}_{k|k} \end{bmatrix} = F_k P_{k|k-1} H^T \left(H P_{k|k-1} H^T + R_k \right)^{-1} \qquad (5.3.16)$$

$$\hat{s}_{k+1|k} = \begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{\theta}_{k+1|k} \end{bmatrix} = F_k \hat{s}_{k|k-1} + K_{k|k} \left(z_k - H \hat{s}_{k|k-1} \right) \qquad (5.3.17)$$

$$E(G \mid \psi_k) = \qquad (5.3.18)$$

$$\left[\begin{array}{c|c|c|c|c|c|c} -\hat{\theta}^1_A(k+1|k) & & -\hat{\theta}^p_A(k+1|k) & \hat{\theta}^1_B(k+1|k) & & \hat{\theta}^m_B(k+1|k) & \\ & \dots & & & \dots & & I_{(m+p+1)n} \\ 0 & & 0 & 0 & & 0 & \end{array} \right]$$

$$K^{x}_{k|k} = K^{x}_{k|k} + E_{\psi_k}(G^x) \, S_k \, \left(H P_{k|k-1} H^T + R_k \right)^{-1} \qquad (5.3.19)$$

$$\hat{x}_{k+1|k} = \hat{x}_{k+1|k} + E_{\psi_k}(G^x) \, S_k \left(H P_{k|k-1} H^T + R_k \right)^{-1} \left(z_k - H \hat{s}_{k|k-1} \right) \qquad (5.3.20)$$

$$\begin{aligned} P_{k+1|k} &= \begin{bmatrix} P^{xx}_{k+1|k} & P^{x\theta}_{k+1|k} \\ P^{\theta x}_{k+1|k} & P^{\theta\theta}_{k+1|k} \end{bmatrix} \\ &= F_k \, P_{k|k-1} \, F_k^T + E[G \mid \psi_k] \, Q_k^0 \, E[G^T \mid \psi_k] \\ &\qquad - K_{k|k} (H \, P_{k|k-1} \, H^T + R_k) K_{k|k}^T \end{aligned} \tag{5.3.21}$$

$$\begin{aligned} P^{xx}_{k+1|k} = P^{xx}_{k+1|k} &+ \sum_{i=1}^{p} \sum_{j=1}^{p} \; P^{\theta\theta}_{i,j}(k+1|k) \, (Q_k^0)_{i,j} + \sum_{i=1}^{m} \sum_{j=1}^{m} \; P^{\theta\theta}_{p+i,p+j}(k+1|k) \, (Q_k^0)_{p+i,p+j} \\ &- \sum_{i=1}^{p} \sum_{j=1}^{m} \; P^{\theta\theta}_{i,p+j}(k+1|k) \, (Q_k^0)_{i,p+j} - \sum_{i=1}^{m} \sum_{j=1}^{p} \; P^{\theta\theta}_{p+i,j}(k+1|k) \, (Q_k^0)_{p+i,j} \end{aligned} \tag{5.3.22}$$

where the partition of $P^{\theta\theta}_{k+1|k}$ is given by Equation 5.3.10.

Salut, *et. al.*, presented a result that is related to this chapter's result in [5]. The primary difference is that [5] treats only the case where there is no noise on the input or output (corresponding to $w_k \equiv 0$ and $v_k \equiv 0$), only noise introduced directly into the states (corresponding to $\Xi_x(k)$) by a known matrix (corresponds to knowing Σ_k^{xx}). Allowing no input or output noise seems to be a major omission, since quite often output measurements and input signals are corrupted by significant levels of additive noise.

In case there is noise at the input and/or output, the method of [5] yields biased predictions. The bias is most noticeable in the parameter estimates, increasing as the noise power increases, and resulting in significant steady-state parameter estimate error even with moderate amounts of noise, and degradation of the estimates of the system states, as well. Thus the formulation of Equation 5.3.6 is a significant improvement over previous similar work.

6.0 *The Deterministic PLID Algorithm*

The deterministic (noise-free) PLID algorithm developed in this chapter is essentially a subset of work presented by Salut, *et.al.* in [5], which is, itself, a subset of the stochastic PLID algorithm. However, it is important to present the development here because [5] does not explicitly present the deterministic MIMO discrete-time case, and, in any event, uses quite different notation. More importantly, Section 6.2 gives a proof of the time-optimal convergence of the deterministic PLID algorithm that has not been found elsewhere in the literature.

6.1 *The Deterministic Algorithm*

The complete *stochastic* PLID algorithm is given by Equations 5.3.16 through 5.3.22. To obtain the deterministic algorithm, it is only necessary to consider what happens to the Equations 5.3.16 through 5.3.22. as all the noise covariance terms go to zero. Substituting the extended system input autocovariance $Q_0(k) \equiv \mathbf{0}$, the extended system output autocovariance $R_k \equiv \mathbf{0}$, and the extended system input/output cross-covariance $S_k \equiv \mathbf{0}$, the stochastic PLID algorithm reduces to

$$K_{k|k} = F_k \, P_{k|k-1} \, H^T (\, H \, P_{k|k-1} \, H^T)^{\#} \tag{6.1.1}$$

$$\hat{s}_{k+1|k} = (\, F_k - K_{k|k} \, H \,) \, \hat{s}_{k|k-1} + K_{k|k} \, y_k \tag{6.1.2}$$

$$P_{k+1|k} = F_k \, P_{k|k-1} \, F_k^T - K_{k|k} \, (H \, P_{k|k-1} \, H^T) K_{k|k}^T \tag{6.1.3}$$

where "#" in Equation 6.1.1 denotes the pseudo-inverse.

The next section shows that if $P_{0|-1}$ is chosen to be symmetric positive definite, then convergence of the estimator is time-optimal, regardless of the choice of the initial estimate $\hat{s}_{0|-1}$. This is not altogether an unexpected result, because it is reasonable that the optimality of the PLID algorithm should hold even as the noise covariances tend to zero.

Also, because the extended system input noise $\eta_k \equiv \mathbf{0}$, and the extended system output noise $w_k \equiv \mathbf{0}$, the error propagation, given by Equation 5.2.10, reduces to

$$e_{k+1|k} = (F_k - K_{k|k} H) e_{k|k-1} \tag{6.1.4}$$

The deterministic algorithm is clearly a linear minimum-mean-square-error recursion. However, the convergence properties of the algorithm need to be investigated, because the input and output measurements appear in the system matrix F_k. Although it is not necessarily obvious, system identification is only possible if the input sequence persistently excites the system. For example, the sequence $u_k \equiv \mathbf{0}$ cannot be expected to work in a system identification scheme.

These ideas are made more concrete in the next section, where the convergence of the deterministic PLID algorithm is shown to depend upon attaining an observability matrix (see Equation 4.2.12) with full column rank. The type of input required to achieve full column rank is explored in Chapter 7, where the stochastic PLID parameter estimates are proved to converge w.p.1, given a "persistently exciting" input.

It is interesting to note that Equation 6.1.2 is a generalization of the Luenberger "identity" observer to time-varying systems [24]. The corresponding observability matrix is given by Equation 4.2.12.

6.2 *Proof of Time-Optimal Convergence*

As in Chapter 4, m denotes the number of inputs; p denotes the number of outputs; $n_1, \dots, n_p$ are the observability indices associated with the outputs, and their sum is n, the total number of states. Thus, matrix F_k is square, with dimension $(m+p+1)n$.

Let us start by defining some notation for the output matrix H.

Denote the **rows** of the output matrix by $H = \begin{bmatrix} h_0 \\ \vdots \\ h_{p-1} \end{bmatrix}$

Lemma 6.2.1: Suppose $P_{0|-1}$ is symmetric positive definite.

For all $k \geq 0$, and $\forall\, j \in \{0, \dots, p-1\}$,

$$P_{k|k-1}\, h_j^T \neq \mathbf{0} \quad \Leftrightarrow \quad h_j P_{k|k-1}\, h_j^T \neq 0.$$

Proof: $'\Rightarrow'$ Suppose $P_{k|k-1}\, h_j^T \neq \mathbf{0}$.

Obviously, $P_{k|k-1} \neq \mathbf{0}$. Let $\rho \triangleq rank(P_{k|k-1}) > 0$. Since $P_{k|k-1}$ is symmetric non-negative definite, it can be written as a sum of ρ vector outer products.

That is, $$0 \neq P_{k|k-1}\, h_j^T = \left[\sum_{i=1}^{\rho} c_i(k)\, c_i^T(k) \right] h_j^T. \qquad (6.2.1)$$

Therefore, for some $i \in \{1, \dots, \rho\}$, $0 \neq c_i^T(k)\, h_j^T$. It follows that

$$0 \neq \sum_{i=1}^{\rho}[h_j\, c_i(k)][c_i^T(k)\, h_j^T] = h_j \left[\sum_{i=1}^{\rho} c_i(k)\, c_i^T(k) \right] h_j^T \tag{6.2.2}$$
$$= h_j\, P_{k|k-1}\, h_j^T.$$

' ⇐ ' (by contradiction) Suppose $h_j P_{k|k-1} h_j^T \neq 0$.

Assume that $P_{k|k-1} h_j^T = \mathbf{0}$.

Then $h_j [P_{k|k-1} h_j^T] = 0$, which is a contradiction. ∎

Lemma 6.2.2: Suppose $P_{0|-1}$ is symmetric positive definite.

For all $k \geq 0$, and $j \in \{0, \ldots, p-1\}$,

$P_{k|k-1} h_j^T$ is in the null space of $(F_k - K_{k|k} H)$.

Proof: The proof is carried out for the complete output matrix H.

Case 1: $H P_{k|k-1} H^T = \mathbf{0}$.

By Lemma 6.2.1, $P_{k|k-1} H^T = \mathbf{0}$, hence $(F_k - K_{k|k} H) P_{k|k-1} H^T = \mathbf{0}$.

Case 2: $H P_{k|k-1} H^T \neq \mathbf{0}$.

$$\begin{aligned}
(F_k - K_{k|k} H)\, P_{k|k-1}\, H^T &= \left[F_k - F_k P_{k|k-1} H^T [H P_{k|k-1} H^T]^{\#} H\right] P_{k|k-1} H^T \\
&= F_k P_{k|k-1} H^T - F_k P_{k|k-1} H^T [H P_{k|k-1} H^T]^{\#} [H P_{k|k-1} H^T] \\
&= F_k P_{k|k-1} H^T \left[\mathbf{I} - (H P_{k|k-1} H^T)^{\#} (H P_{k|k-1} H^T)\right] \\
&= F_k \times \left[P_{k|k-1} h_0^T \mid \ldots \mid P_{k|k-1} h_{p-1}^T \right] \\
&\quad \times \begin{bmatrix} 1 - (h_0 P_{k|k-1} h_0^T)^{\#} (h_0 P_{k|k-1} h_0^T) & & \\ & \ddots & \\ & & 1 - (h_{p-1} P_{k|k-1} h_{p-1}^T)^{\#} (h_{p-1} P_{k|k-1} h_{p-1}^T) \end{bmatrix} \\
&= F_k \times (\mathbf{0}) = \mathbf{0}
\end{aligned} \tag{6.2.3}$$

with the last line due to Lemma 6.2.1, because

$$P_{k|k-1}\, h_l^T \neq \mathbf{0} \quad \Rightarrow \quad [\, 1 - (h_l\, P_{k|k-1}\, h_l^T)^{\#}\; h_l\, P_{k|k-1}\, h_l^T \,] = 0,$$

and the fact that $(H\, P_{k|k-1}\, H^T)^{\#}\,(H\, P_{k|k-1}\, H^T)$ is a diagonal matrix. Since the lemma holds for H, then it holds for each row, h_j. ■

The next lemma develops a useful equivalent expression for $P_{k+1|k}$.

Lemma 6.2.3: The error matrix

$$P_{k+1|k} = (\, F_k - K_{k|k}\, H\,)(\, F_{k-1} - K_{k-1|k-1}\, H\,) \;\; \dots \;\; (\, F_0 - K_{0|0}\, H\,)\, P_{0|-1}\, F_0^T \;\; \dots \;\; F_k^T .$$

Proof: (by induction) If $k = 0$, then from Equation 6.1.3,

$$\begin{aligned} P_{1|0} &= F_0 \Big[P_{0|-1} - P_{0|-1}\, H^T (\, H\, P_{0|-1}\, H^T)^{\#}\, H\, P_{0|-1} \Big] F_0^T \\ &= \Big[F_0 - F_0\, P_{0|-1}\, H^T (\, H\, P_{0|-1}\, H^T)^{\#}\, H \Big] P_{0|-1}\, F_0^T \qquad (6.2.4) \\ &= (\, F_0 - K_{0|0}\, H\,)\, P_{0|-1}\, F_0^T \end{aligned}$$

For the induction, assume that for some $k \geq 0$

$$P_{k|k-1} = (\, F_{k-1} - K_{k-1|k-1}\, H\,) \;\; \dots \;\; (\, F_0 - K_{0|0}\, H\,)\, P_{0|-1}\, F_0^T \;\; \dots \;\; F_{k-1}^T \qquad (6.2.5)$$

Then for $k + 1$,

$$\begin{aligned} P_{k+1|k} &= F_k \Big[P_{k|k-1} - P_{k|k-1}\, H^T (\, H\, P_{k|k-1}\, H^T)^{\#}\, H\, P_{k|k-1} \Big] F_k^T \\ &= \Big[F_k - F_k\, P_{k|k-1}\, H^T (\, H\, P_{k|k-1}\, H^T)^{\#}\, H \Big] P_{k|k-1}\, F_k^T \qquad (6.2.6) \\ &= (\, F_k - K_{k|k}\, H\,)\, P_{k|k-1}\, F_k^T \end{aligned}$$

Substitute Equation 6.2.5 into 6.2.6 to obtain the lemma. ■

Now, define the *observation vectors* ϕ_j , $j = 0, \dots ,(kp + p - 1)$, to be the *rows* of the observability matrix Φ_k (see Equation 4.2.12). Thus the rows ϕ_j of Φ_k are

$$
\begin{aligned}
\phi_0 &\triangleq h_0 \\
&\vdots \\
\phi_{p-1} &\triangleq h_{p-1} \\
\phi_p &\triangleq h_0\, F_0 \\
&\vdots \\
\phi_{p+(p-1)} &\triangleq h_{p-1}\, F_0 \\
&\vdots \\
\phi_{kp} &\triangleq h_0\, F_{k-1}\, \dots F_0 \\
&\vdots \\
\phi_{kp+j} &\triangleq h_j\, F_{k-1}\, \dots F_0 \\
&\vdots \\
\phi_{kp+(p-1)} &\triangleq h_{p-1}\, F_{k-1}\, \dots F_0
\end{aligned}
\tag{6.2.7}
$$

Lemma 6.2.4: Suppose $P_{0|-1}$ is symmetric positive definite.

The observation vectors $\{\phi_0^T, \dots, \phi_{kp+p-1}^T\}$ are in the null space of

$$(F_k - K_{k|k} H)(F_{k-1} - K_{k-1|k-1} H) \dots (F_0 - K_{0|0} H) P_{0|-1} .$$

Proof: By Lemma 6.2.2, $0 = (F_k - K_{k|k} H) P_{k|k-1} H^T$.

Therefore, for each $j = 0, \dots, p-1$,

$$0 = (F_k - K_{k|k} H) P_{k|k-1} h_j^T. \tag{6.2.8}$$

Expanding $P_{k|k-1}$ by Lemma 6.2.3 and the definitions of Equation 6.2.7,

$$
\begin{aligned}
0 &= (F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1} F_0^T F_1^T \dots F_{k-1}^T h_j^T \\
&= (F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1} \phi_{pk+j}^T
\end{aligned}
\tag{6.2.9}
$$

$$\forall j = 0, \dots, p-1.$$

Generalizing Equation 6.2.9, for any $i = 0, \dots, k$,

$$0 = (F_l - K_{l|l} H) \dots (F_0 - K_{0|0} H) P_{0|-1} \phi_{lp+j} . \qquad (6.2.10)$$

Hence (premultiplying appropriately), for all $i = 0, \dots ,k$, and all $j = 0, \dots ,p - 1$,

$$0 = (F_k - K_{k|k} H) \dots \left[(F_l - K_{l|l} H) \dots (F_0 - K_{0|0} H) P_{0|-1} \phi_{ip+j} \right]. \qquad (6.2.11)$$

■

Lemma 6.2.4 implies that

$$nullity\left[(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1} \right] \geq rank[\Phi_k]. \qquad (6.2.12)$$

The next theorem proves that the *opposite* inequality also holds in Equation 6.2.12.

Lemma 6.2.5: Suppose $P_{0|-1}$ is symmetric positive definite.

Let $N \triangleq (m + p + 1)n$, the number of states in the extended state vector s_k.

Let A be the *ordered set* of indices α_i, $i \in \{0, \dots , N - 1\}$, created by including each index $kp+j$ for which $k \geq 0$ and $j \in \{0, \dots ,p - 1\}$ satisfy $h_j P_{k|k-1} h_j^T \neq 0$.

(1) The set $\{\phi_{\alpha_0} , \dots , \phi_{\alpha_{N-1}} \}$ is linearly independent.

(2) If $rank(\Phi_k) = m < N$, then the observation vectors $\{\phi^T_{\alpha_m} , \dots , \phi^T_{\alpha_{N-1}} \}$ are in the range space of $(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1}$.

Proof: In this proof, let k_i be such that $pk_i \leq \alpha_i < p(k_i + 1)$.

Note that, by the construction of the set A, $h_{\alpha_i - pk_i} P_{k_i|k_i-1} h^T_{\alpha_i - pk_i} \neq 0$.

Part (1) will be shown by contradiction:

Suppose the set $\{\phi_{\alpha_0}, \dots , \phi_{\alpha_{N-1}}\}$ is not linearly independent. That is,

suppose that, for some $i \in \{1, \dots , N - 1\}$, $\phi_{\alpha_i} = \sum_{j=0}^{i-1} \beta_j \phi_{\alpha_j}$. (6.2.13)

Then, by Theorem 6.2.4,

$$0 = (F_{k_{i-1}-1} - K_{k_{i-1}-1|k_{i-1}-1} H) \dots (F_0 - K_{0|0} H) P_{0|-1} \phi^T_{\alpha_i} . \tag{6.2.14}$$

So, premultiplying appropriately, and expanding $\phi^T_{\alpha_i}$ using Equation 6.2.7,

$$\begin{aligned} 0 &= (F_{k_i-1} - K_{k_i-1|k_i-1} H) \dots (F_0 - K_{0|0} H) P_{0|-1} F_0^T \dots F^T_{k_i-1} h^T_{\alpha_i-pk_i} . \\ &= P_{k_i|k_i-1} h^T_{\alpha_i-pk_i} \end{aligned} \tag{6.2.15}$$

Therefore, $0 = h_{\alpha_i-pk_i} P_{k_i|k_i-1} h^T_{\alpha_i-pk_i} \Rightarrow \alpha_i \notin A.$ (6.2.16)

Therefore, the assumption of Equation 6.2.13 contradicts a premise of the theorem.

Now part (2) of the theorem will be proved, again by contradiction:

Suppose part (2) of the theorem is not true. That is, suppose that, for some $d \in \{1, \dots, N\text{-}1\text{-}i\}$ it is true that

$$0 = (F_{k_i-1} - K_{k_i-1|k_i-1} H) \dots (F_0 - K_{0|0} H) P_{0|-1} \phi^T_{\alpha_{i+d}} . \tag{6.2.17}$$

Then, premultiplying appropriately, and expanding $\phi^T_{\alpha_{i+d}}$ using Equation 6.2.7,

$$\begin{aligned} 0 &= (F_{k_{i+d}-1} - K_{k_{i+d}-1|k_{i+d}-1} H) \dots (F_0 - K_{0|0} H) P_{0|-1} F_0^T \dots \\ &\qquad \dots F^T_{k_{i+d}-1} h^T_{\alpha_{i+d}-pk_{i+d}} \\ &= P_{k_{i+d}|k_{i+d}-1} h^T_{\alpha_{i+d}-pk_{i+d}} . \end{aligned} \tag{6.2.18}$$

Hence, $0 = h_{\alpha_{i+d}-pk_{i+d}} P_{k_{i+d}|k_{i+d}-1} h^T_{\alpha_{i+d}-pk_{i+d}} \Rightarrow \alpha_{i+d} \notin A.$ (6.2.19)

Thus, the assumption of Equation 6.2.17 contradicts a premise of the theorem.

■

Theorem 6.2.6: If $P_{0|-1}$ is symmetric positive definite, then for all $k = 0, 1, \dots$,

$$nullity\left[(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1}\right] = rank\left[\Phi_k\right]. \qquad (6.20)$$

Proof: By Lemma 6.4, $nullity[(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1}] \geq rank[\Phi_k]$.

By Lemma 6.5, $nullity[(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1}] \leq rank[\Phi_k]$.

The theorem follows. ■

Corollary 6.2.7: Suppose $P_{0|-1}$ is symmetric positive definite.

At time k, if the observability matrix Φ_k has rank $(m+p+1)n$,

then the error outer product matrix $P_{k+1|k}$ is zero.

Proof: The matrix $P_{k+1|k}$ is a square matrix with dimension $[(m+p+1)n]$.

Now, by lemma 6.2.2,

$$P_{k+1|k} = (F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1} F_0^T \dots F_k^T. \qquad (6.2.21)$$

But, by Theorem 6.2.6,

$$nullity\left[(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1}\right] = (m+p+1)n. \qquad (6.2.22)$$

Therefore,

$$(F_k - K_{k|k} H) \dots (F_0 - K_{0|0} H) P_{0|-1} = 0 \quad \Rightarrow \quad P_{k+1|k} = 0. \qquad (6.2.23)$$

■

Corollary 6.2.7 makes it clear that the input sequence $\{u_k\}$ must be *persistently exciting of order* $(m+p+1)n$ in order that the error outer product matrix converges to zero.

Finally, consider the implication of Theorem 6.2.4 for the prediction error $e_{k+1|k}$.

Corollary 6.2.8: Suppose $P_{0|-1}$ is symmetric positive definite.

If at time k the observability matrix has rank $(m+p+1)n$,

then the prediction error $e_{k+1|k}$ is zero.

Proof: From Equation 6.1.4,

$$\begin{aligned} e_{k+1|k} &= (F_k - K_{k|k}H) \ldots (F_0 - K_{0|0}H)\, e_{0|-1} \\ &= (F_k - K_{k|k}H) \ldots (F_0 - K_{0|0}H)\, P_{0|-1} (P_{0|-1})^{-1} e_{0|-1} \\ &= \mathbf{0} \end{aligned} \tag{6.2.24}$$

with the last equality due to Theorem 6.2.6. ■

As shown by Equation 4.1.13, the extended state cannot be uniquely identified before the generalized inverse of the observability matrix Φ_k exists, *i.e.*, not before *rank* $(\Phi_k) = (m+p+1)n$. Therefore, the PLID algorithm is time-optimal (deadbeat), because it is precisely when the observability matrix attains rank equal to $(m+p+1)n$ that the error outer product matrix $P_{k|k-1}$ and the error vector $e_{k|k-1}$ become zero. Another implication of Theorem 6.2.6 is that the PLID algorithm will converge even if there is some delay before the input becomes sufficiently exciting; this is borne out by simulations.

A simple way to determine when the prediction error reaches zero is to count the number of times that diagonal elements of $H\, P_{k|k-1}\, H^T$ are non-zero. (The diagonal elements are $h_i\, P_{k|k-1}\, h_i^T$.) When the count reaches $(m+p+1)n$, the prediction error will be zero, because, by Lemma 6.2.5, there must be $(m+p+1)n$ independent observation vectors ϕ_i in Φ_k.

6.3 Linear Dependence Within the Error Covariance Matrix

As shown in Section 6.2, the rank of the error covariance matrix $P_{k+1|k}$ decreases every time a linearly independent observation of the system is made. But what is the precise mechanism that causes the matrix to lose rank? That is the question this section answers.

The answer can be given by presenting the eigenvectors of the zero eigenvalues of $P_{k+1|k}$. The elements of these eigenvectors can be considered as coefficients specifying the linear dependence among the columns (or the rows) of $P_{k+1|k}$.

To find the eigenvectors associated with the zero eigenvalues of the matrix $P_{k+1|k}$, consider the equivalent representation given in Lemma 6.2.3:

$$\begin{aligned} P_{k+1|k} = & (F_k - K_{k|k} H) (F_{k-1} - K_{k-1|k-1} H) \; \dots \\ & \dots \; (F_0 - K_{0|0} H) P_{0|-1} F_0^T \; \dots \; F_k^T . \end{aligned} \tag{6.3.1}$$

From Equation 6.3.1, it is clear that if we find the eigenvectors associated with the zero eigenvalues of F_k^T, then these will also be eigenvectors for zero eigenvalues of $P_{k+1|k}$.

Indeed, eigenvectors associated with the zero eigenvalues of $F_{k-i}^T \dots F_k^T$, $\forall$ $i \in \{1, \dots ,k\}$ are also eigenvectors of zero eigenvalues of $P_{k+1|k}$. It turns out these eigenvectors are not hard to find. There are still others, of course, but let us enumerate first the ones just mentioned.

It turns out that the matrix $P_{k+1|k}$ loses rank for the first few iterations, regardless of the persistency of excitation of the input. It is, quite simply, a function of the rank deficiency of F_k . The nullity of F_k is always equal to p, independent of the inputs and the outputs. That is true because there are p columns that are completely void in the matrix, while all remaining columns have independent unity elements appearing within them (see Equation 4.2.11).

As might be expected, there is a maximum number of states and parameters that are observable through any particular output. Recall that there are p subsystems, one associated with each output $z_1(k), \dots ,z_p(k)$. The observability index of the i^{th} subsystem is denoted by n_i. The maximum number β_i of states of the extended system (*i.e.*, states and parameters of the original system) that can be observed by way of $z_i(k)$ is given by

$$\beta_{i_{\max}} = (m + p + 1)\, n_i \tag{6.3.2}$$

where m is the number of system inputs, and p is the number of system outputs.

Equation 6.3.2 implies that all the subsystems must be persistently excited in order to identify the entire extended system vector s_k, which has $(m+p+1)(n_1 + \cdots + n_p)$ elements. That is perfectly reasonable.

However, of the $(m+p+1)\, n_i$ zero eigenvalues associated with the output z_i, the first n_i arise from the nullity of F_k, regardless of the system excitation. This follows because the associated submatrix of the product $F_k F_{k-1} \dots F_0$ is $\mathbf{J}_{n_i}^{k+1}$, where $\mathbf{J}_{n_i}$ is nilpotent of order $n_i - 1$ because it is an $n_i \times n_i$ lower Jordan block of zero eigenvalues.

The eigenvectors associated with the increasing nullity of the product $F_0^T \dots F_k^T$, for $k = 0, \dots, n_i - 1$, are rather difficult to notate. However, the following development yields one of the most concise forms.

Let $\bar{n}_0 \triangleq 0$; for $1 \le i \le p$, let $\bar{n}_i \triangleq \sum_{r=1}^{i} n_r$.

For all times $k \ge 0$, let $\beta_i(k) \triangleq min\,(k+1, n_i)$ denote number of eigenvectors of zero eigenvalues that are due to the nullity of $F_0^T \dots F_k^T$ and *that are associated with the output* z_i. The j^{th} such eigenvector $\alpha_i(j, k)$, where $1 \le j \le \beta_i(k)$, is given by

$$\alpha_i(j,k) = \left[\begin{array}{c} \alpha_{i,1}(j,k) \\ \hline \alpha_{i,2}(j,k) \\ \vdots \\ \alpha_{i,1+p}(j,k) \\ \hline \alpha_{i,1+p+1}(j,k) \\ \vdots \\ \alpha_{i,1+p+m}(j,k) \end{array}\right], \tag{6.3.3}$$

$$\begin{aligned} \text{where} \quad \alpha_{i,1}(j,k) &= -\,\mathbf{e}_{j+\bar{n}_{i-1}}\,; \\ \alpha_{i,1+r}(j,k) &= \sum_{t=1}^{j} z_r(k-j+t)\,\mathbf{e}_{nr+t+\bar{n}_{i-1}}\,, \qquad 1 \le r \le p\,; \\ \alpha_{i,1+p+r}(j,k) &= \sum_{t=1}^{j} u_r(k-j+t)\,\mathbf{e}_{n(p+r)+t+\bar{n}_{i-1}}\,, \qquad 1 \le r \le m\,; \end{aligned}$$

in which $\mathbf{e}_i$ denotes the i^{th} unit vector in $\mathbb{R}^n$.

Equation 6.3.3 gives the general form for the first n eigenvectors of zero eigenvalues of $P_{k+1|k}$. Increasing nullity of $F_0^T \ldots F_k^T$ is guaranteed to occur in the i^{th} subsystem for the first n_i iterations. Therefore, this is true for the nullity of $P_{k+1|k}$, as discussed before. However, after that point, increasing nullity of $P_{k+1|k}$ requires persistent excitation of the system.

The eigenvectors associated with the persistent excitation of the system are much easier to specify than those in Equation 6.3.3, because they are just the most recent set of independent rows (transposed, of course) of the extended observability matrix Φ_k. This fact is actually a corollary of the following theorem.

Theorem 6.3.9: Denote the rows of the output matrix H by h_i, for $i = 1, \ldots, p$. Denote the rows of the observability matrix Φ_k by ϕ_i, for $i = 0, \ldots, p(k+1) - 1$. Then

$$P_{k|k-1}\, h_l^T = P_{k|k-1}\, \phi_{pk+l}^T, \quad \forall k \geq n_l .$$

Proof: First, note that for any $j \geq 0$, and for all $k \geq n_l - 1$,

$$F_j^T \left[F_0^T ... F_k^T\, h_l^T \right] \;=\; F_0^T ... F_k^T\, h_l^T . \tag{6.3.4}$$

Equation 6.3.4 is due to the fact that $k \geq n_l - 1$ implies that $\mathbf{J}_{n_l}^{k+1} = \mathbf{0}$, because $\mathbf{J}_{n_l}$ is nilpotent of order n_l. From Equation 6.3.4 and the definition of the rows ϕ_l of the observability matrix Φ_k, given in Equation 6.2.7, for any $j \geq 0$, it follows that

$$\begin{aligned} F_j^T\, \phi_{kp+l}^T \;&=\; F_j^T\, F_0^T ... F_{k-1}^T\, h_l^T \\ &=\; F_0^T ... F_{k-1}^T\, h_l^T \;=\; \phi_{kp+l}^T , \qquad \forall\, k \geq n_l . \end{aligned} \tag{6.3.5}$$

Now, using the equivalent expression for $P_{k|k-1}$ developed in Lemma 6.2.3, and repeated application of the Equation 6.3.5,

$$\begin{aligned} P_{k|k-1}\, \phi_{kp+l}^T &= \left[(F_{k-1} - K_{k-1|k-1}\, H) ... (F_0 - K_{0|0}\, H)\, P_{0|-1}\, F_0^T ... F_{k-1}^T \right] \left[F_0^T ... F_{k-1}^T\, h_l^T \right] \\ &= (F_{k-1} - K_{k-1|k-1}\, H) ... (F_0 - K_{0|0}\, H)\, P_{0|-1}\, F_0^T ... F_{k-1}^T\, h_l^T \\ &= P_{k|k-1}\, h_l^T , \\ &\qquad \forall\, k \geq n_l . \qquad \blacksquare \end{aligned} \tag{6.3.6}$$

Now, by Lemma 6.2.2, for all $k \geq 0$, it is true that $P_{k|k-1}\, h_l^T$ is in the null space of $(F_k - K_{k|k}\, H)$. Therefore,

$$\begin{aligned} P_{k+1|k}\, \phi_{kp+l}^T &= (F_k - K_{k|k}\, H)\, P_{k-1|k-2}\, F_k^T\, \phi_{kp+l}^T \\ &= (F_k - K_{k|k}\, H)\, P_{k|k-1}\, \phi_{kp+l}^T \\ &= (F_k - K_{k|k}\, H)\, P_{k|k-1}\, h_l^T = \mathbf{0}, \qquad \forall\, k \geq n_l . \end{aligned} \tag{6.3.7}$$

That is, ϕ^T_{kp+i} is in the null space of $P_{k+1|k}$ for all $k \geq n_i$. In a similar manner, it is easy to show that ϕ^T_{rp+i} is in the null space of $P_{k+1|k}$, for $n_i \leq r \leq k$, and $1 \leq i \leq p$.

From the explicit form of the observability matrix Φ_k, given in Equation 4.2.12, and from the value of $\beta_{i_{\max}}$, it is clear that, at most, only $(m+p)\,n_i$ of the rows ϕ_{kr+i} can be linearly independent, for all $k \geq n_i$. Taking these (or, rather, their transposes) together with the n_i independent eigenvectors $\alpha_i(1,k)$ through $\alpha_i(n_i,k)$, described by Equation 6.3.3, gives at most $(m+p+1)n_i$ eigenvectors of zero eigenvalues associated with the i^{th} output.

7.0 Convergence of the Stochastic PLID Parameter Estimates

In this chapter it is proved that, under standard gaussian assumptions, and for time-invariant parameters, the PLID parameter estimates *converge almost everywhere* to the true parameter values.

There is a wide variety of convergence proofs available for (time-invariant system) identification algorithms which take the form of a stochastic approximation,

$$X_{k+1} = X_k - a_k Q(X_k, \zeta_k)$$

where $Q(\,)$ is some function of the previous estimate X_k and a measurement ζ_k, and a_k is a positive scalar. Examples of such proofs can be seen in [25], [26], [27], and [28].

However, none of these convergence proofs is quite applicable to the PLID algorithm because, with the inclusion of the state estimates, it does not fit the stochastic approximation form. Therefore, a different approach is required to prove convergence of the PLID algorithm. It turns out that martingale theory can be useful here.

The convergence proof is given in Section 7.1. The theorems of that section call upon other theorems from martingale theory, reproduced in the Appendixes, with proofs supplied for those that are not well known. A flowchart of the dependencies among the various theorems in Section 7.1 and the Appendixes is given in Figure 7 on page 102.

It must be remembered that the convergence result is a theoretical result. In short, Section 7.1 proves that the parameter estimates will converge almost everywhere *given enough time*. One can construct examples where the algorithm will not converge due to practical considerations, such as numerical ill-conditioning. For example, into this category fall systems whose output becomes unbounded for bounded input.

Also, if the signal-to-noise ratio is low enough, convergence may not occur. Numerical problems, such as accumulated error, may prevent convergence; or convergence may be so slow that, for practical purposes, the algorithm does not converge.

The problem of convergence rate is considered in Chapter 8, which discusses the more general problem of systems with time-varying parameters, and presents a convergence rate estimate for the small noise case.

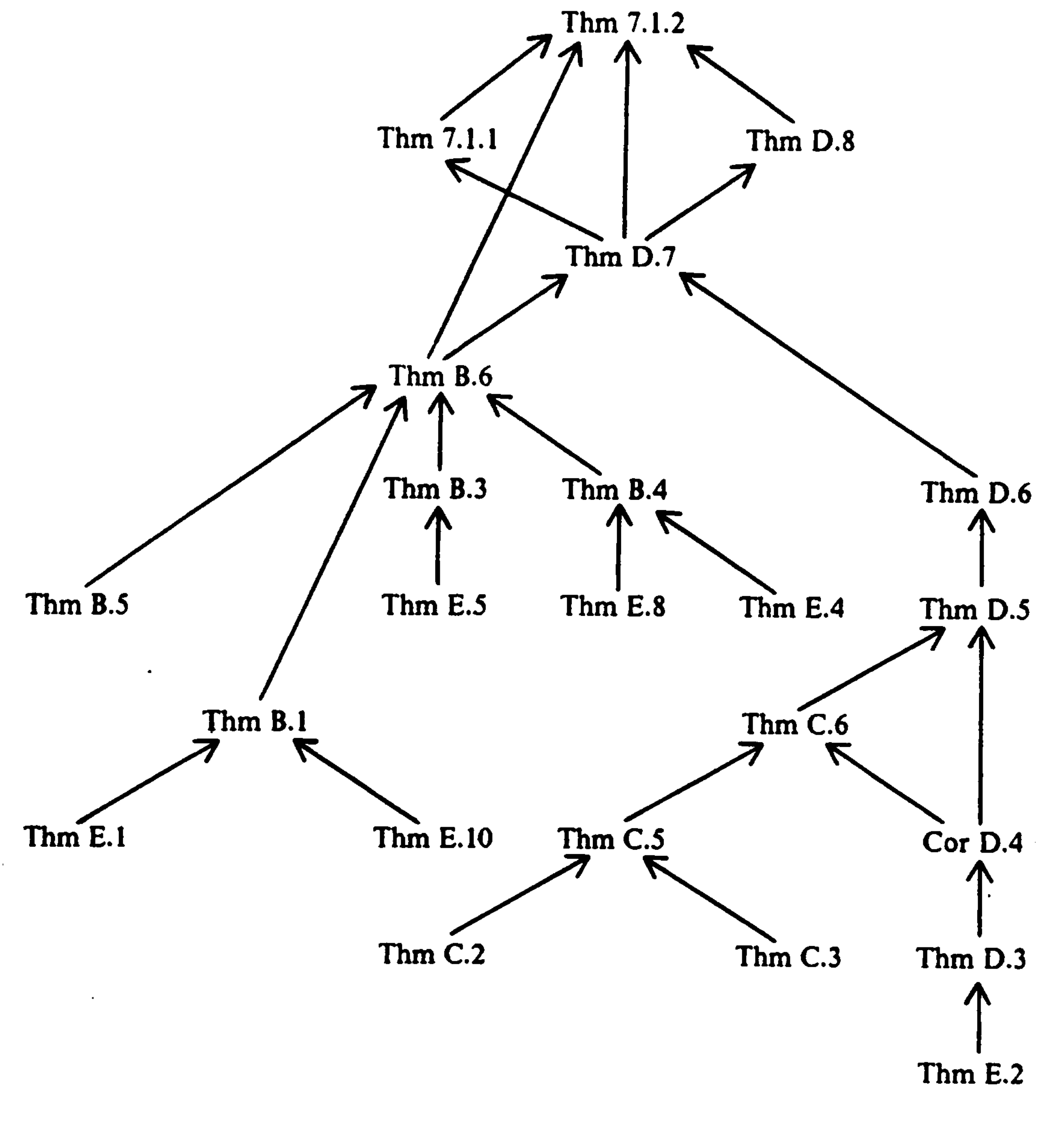

Figure 7. Flowchart of proofs for the theorems of Chapter 7: Dependencies among the various theorems of this chapter and those in the appendixes.

7.1 Proof of Convergence

Under the assumption that the true parameter vector θ is an integrable, constant random vector, and the standard gaussian assumptions on the noise, this section proves that the parameter estimates $\hat{\theta}_{k+1|k}$ converge to θ, where $\hat{\theta}_{k+1|k}$ is a subvector of $\hat{s}_{k+1|k} = [\,\hat{x}^T_{k+1|k} \mid \hat{\theta}^T_{k+1|k}\,]^T$, the stochastic PLID estimate of the extended state vector s_{k+1}.

Recall from the development of Chapter 5 that, under the standard gaussian assumptions on the noise, each element of the PLID estimate is a conditional mean. Thus, for integrable, constant random variables θ, it is fairly simple to show that the conditional mean estimate $\hat{\theta}_{k|k-1}$ of the parameters is a Doob's martingale. (Martingales and associated terms are defined in Appendix A.)

By assumption, the unknown parameter vector $\theta(\omega)$ is a sample of an integrable, constant random vector. Recall ψ_k denotes the increasing sequence of sub-σ-fields generated by the increasing sets of measurements, $\{\, z_0 , \dots , z_k \}$. Then the extended state vector estimates

$$\hat{s}_{k|k-1} = E\,[\, s_k \,|\, \psi_{k-1}\,] = E\left[\begin{bmatrix} x_k \\ \theta \end{bmatrix} \middle| \psi_{k-1} \right], \quad k = 1,2, \dots , \qquad (7.1.1)$$

and, in particular, the subvector of parameter estimates

$$\hat{\theta}_{k|k-1} = E\,[\,\theta \mid \psi_{k-1}\,], \quad k = 1,2, \dots , \qquad (7.1.2)$$

are sequences of random variables adapted to the increasing sequence of sub-σ-fields ψ_k. Note that, by Theorem D.2,

$$
\begin{aligned}
E\left[\hat{\theta}_{k+1|k} \,|\psi_{k-1}\right] &= E\left[E\{\theta \,|\psi_k\} \,|\psi_{k-1}\right] \\
&= E\left[\theta \,|\psi_{k-1}\right] \\
&= \hat{\theta}_{k|k-1}
\end{aligned}
\tag{7.1.3}
$$

Thus, by definition, $\hat{\theta}_{k+1|k}$ is a martingale relative to the sequence ψ_k, or more succinctly, $\{\hat{\theta}_{k+1|k}, \psi_k\}$ is a martingale. This type of martingale is known as a Doob's martingale.

Theorem 7.1.1: If $\hat{\theta}_n = E(\theta \mid \psi_n)$, where θ is an integrable random variable, then $\{\hat{\theta}_n, \psi_n\}$ is a martingale, and

(1) $\lim_{n\to\infty} E(\theta \mid \psi_n) = E(\theta \mid \psi_\infty)$, and

(2) $\lim_{n\to\infty} E(|\theta| \mid \psi_n) = E(|\theta| \mid \psi_\infty)$.

Proof: (as in Chung [29]) Using the definition of martingale, it is easy to verify that the sequence is, indeed, a martingale (as done above). So, adopting the premise of this theorem is equivalent to assuming condition (4) of Theorem D.7 (Appendix D). Therefore, conditions (2) and (3) of Theorem D.7 follow, because (2), (3), and (4) are shown there to be equivalent conditions. Condition (2) states that $\hat{\theta}_n$ converges in $\mathscr{L}^1$, which proves condition (2) of this theorem.

Condition (3) of Theorem D.7 is that $\hat{\theta}_n$ converges a.e. to an integrable $\hat{\theta}_\infty$ such that $\{\hat{\theta}_n, \psi_n\}_{n\in\bar{N}}$ is a martingale.

Now, to identify the a.e. limit referred to in condition (3) of Theorem D.7, note that for each $\Lambda \in \psi_n$,

$$\int_\Lambda \theta \, dP = \int_\Lambda \hat{\theta}_n \, dP = \int_\Lambda \hat{\theta}_\infty \, dP \tag{7.1.4}$$

The same statement can be made for every $\Lambda \in \psi_\infty$, which proves condition (1).

■

Theorem 7.1.1 shows that $E[\,\theta\,|\psi_k\,]$, *i.e.*, the expected value of an integrable random variable conditioned on an increasing sequence of sub-σ-fields, is a martingale that converges almost everywhere to $E[\,\theta\,|\psi_\infty\,]$, where ψ_∞ is the closure of $\bigcup_{k=0}^{\infty}\psi_k$. Sternby [30] examined this limit for the case of conditional mean estimates of a time-invariant parameter vector θ, and proved that $E[\,\theta\,|\,\psi_\infty\,] = \theta$ almost everywhere, provided that the error covariance matrix tends to zero. Sternby's theorem and proof are reproduced here, for the sake of completeness.

Theorem 7.1.2: Suppose θ is a square-integrable random vector, and $\hat{\theta}_n = E(\,\theta\,|\,\psi_n\,)$, where $\{\,\psi_n\,\}$ is an increasing sequence of sub-σ-fields. Let $P_n = E[(\hat{\theta}_n - \theta\,)(\hat{\theta}_n - \theta\,)^T]$, the error covariance of the estimate $\hat{\theta}_n$.

If Λ is the set $\{\,\omega : P_\infty = 0\,\}$, then $[\,1_\Lambda(\omega)\,]\,[\hat{\theta}_\infty(\omega)\,] = [\,1_\Lambda(\omega)\,]\,[\,\theta\,(\omega)\,]$ a.e.,

where 1_Λ is the indicator function of Λ.

Furthermore, if $P(\Lambda) = 1$, then $\hat{\theta}_n \rightarrow \theta$ in $\mathscr{L}^2$.

Proof: P_n is non-negative definite. Therefore, it will suffice to consider the scalar case $\theta \in \mathbb{R}$, because $P_n \rightarrow 0$ implies that all the diagonal elements go to zero.

Now, because $\Lambda \in \psi_\infty$, Lévy's zero-or-one law (Corollary D.8) allows us to say $E\{\,1_\Lambda\,|\,\psi_n\,\} \rightarrow 1_\Lambda$ a.e. Therefore,

$$\left[E\{\,1_\Lambda\,|\,\psi_n\,\}^2\,\right][P_n\,] \rightarrow 0 \quad \text{a.e.} \tag{7.1.5}$$

But

$$\begin{aligned}0 &\leq \left[E\{1_\Lambda \mid \psi_n\}^2\right][P_n] \\ &= \left[E\{1_\Lambda \mid \psi_n\}^2\right]\left[E(\theta^2 \mid \psi_n)\right] - \left[E\{1_\Lambda \mid \psi_n\}^2\right]\left[\hat{\theta}_n^2\right]\end{aligned} \tag{7.1.6}$$

Now, $E(\theta^2 \mid \psi_n)$ is the conditional mean of θ^2, which is an integrable random variable. It is easily shown to be a martingale, so condition (4) of Theorem D.7 is satisfied. By that theorem, the following equivalent conditions also hold:

(1) $E(\theta^2 \mid \psi_n)$ is uniformly integrable,

(2) $E(\theta^2 \mid \psi_n)$ converges in $\mathscr{L}^1$, and

(3) $E(\theta^2 \mid \psi_n)$ converges a.e. to an integrable $E(\theta^2 \mid \psi_\infty)$ such that $\{E(\theta^2 \mid \psi_n)\}_{n \in \bar{\mathbb{N}}}$ is a martingale.

Therefore, because both terms on the right side of Equation 7.1.6 are less than $E(\theta^2 \mid \psi_n)$a.e., both of those terms are also uniformly integrable. Furthermore, the *last* term on the right side of Equation 7.1.6 converges a.e., because the term to the left of the equality converges a.e. by premise, and the *first* term on the right side converges a.e. by condition (3) of Theorem D.7 just cited.

Now, because each term on the right side of Equation 7.1.6 converges a.e. and is uniformly integrable in $\mathscr{L}^1$, condition (iii) of Theorem B.6 is satisfied for each term. Therefore the equivalent condition (ii) holds, which is that each term also converges in $\mathscr{L}^1$. It follows that the term to the left of the equality in Equation 7.1.6 also converges in $\mathscr{L}^1$; furthermore the limits to which both sides of the equation converge must be equal, namely zero. Therefore,

$$\begin{aligned}&E\left[\{E(1_\Lambda \mid \psi_n)\}^2 E\{(\hat{\theta}_n - \theta)^2 \mid \psi_n\}\right] \\ &\quad = E\left[E\{[E(1_\Lambda \mid \psi_n)]^2 (\hat{\theta}_n - \theta)^2 \mid \psi_n\}\right] \\ &\quad = E\left[\{E(1_\Lambda \mid \psi_n)(\hat{\theta}_n - \theta)\}^2\right] \to 0\end{aligned} \tag{7.1.7}$$

The last line of Equation 7.1.7 implies that

$$E(\, 1_\Lambda \mid \psi_n\,)(\hat{\theta}_n - \theta\,) \rightarrow 0 \ \text{ in } \mathscr{L}^2 . \tag{7.1.8}$$

Therefore, if $P(\Lambda) = 1$, then the first factor on the left side of Equation 7.1.8 is unity, proving the last statement of the theorem.

Now, By Lévy's zero-or-one law (Corollary D.8), $\lim_{n\to\infty} P(\Lambda \mid \psi_n) = 1_\Lambda$ a.e. But also, by Theorem 7.1.1,

$$E(\, 1_\Lambda \mid \psi_n\,)(\,\hat{\theta}_n - \theta\,) \rightarrow 1_\Lambda(\,\hat{\theta}_\infty - \theta\,) \ \text{ a.e.} \tag{7.1.9}$$

But this limit must be zero, because that is the limit in $\mathscr{L}^2$. Therefore, we also have

$$1_\Lambda(\,\hat{\theta}_n - \theta\,) \rightarrow 0 \ \text{ a.e.,} \tag{7.1.10}$$

which proves the theorem. ■

Sternby's result can be applied to the parameter estimates of the PLID algorithm, as follows. The error covariance matrix $P_{k|k-1}$ of Equation 5.1.23 can be partitioned as in Equation 5.3.1,

$$P_{k|k-1} = \left[\begin{array}{c|c} P^{xx}_{k|k-1} & P^{x\theta}_{k|k-1} \\ \hline P^{\theta x}_{k|k-1} & P^{\theta\theta}_{k|k-1} \end{array}\right] \tag{7.1.11}$$

Because $P_{k|k-1}$ is positive semidefinite, the diagonal elements of $P^{\theta\theta}_{k|k-1}$ are non-negative and, by inspection of the recursion 5.3.21, clearly non-increasing. However, that is not enough.

In order to use Sternby's result, it is necessary to guarantee that $P^{\theta\theta}_{k|k-1} \rightarrow 0$, in some sense. Let us obtain the recursion for a single diagonal element of $P^{\theta\theta}_{k|k-1}$, as follows.

Let the gain matrix be denoted by rows, $K_{k|k} \triangleq \begin{bmatrix} \kappa_1^T(k) \\ : \\ \kappa_{(m+p+1)n}^T(k) \end{bmatrix}$.

Then the (i,i)-element of $P^{\theta\theta}_{k+1|k}$ (or, equivalently, the $(n+i,n+i)$-element of $P_{k+1|k}$) is

$$p_{n+i,n+i}(k+1) = p_{n+i,n+i}(k) - \kappa_{n+i}^T(k)\,[H\,P_{k|k-1}\,H^T + R_k\,]\,\kappa_{n+i}(k) \leq p_{n+i,n+i}(k) \qquad (7.12)$$

where the inequality is due to the fact that $HP_{k|k-1}\,H^T + R_k$ is positive definite symmetric, making the difference term a "weighted norm" of $\kappa_{n+i}(k)$.

The following lemma gives a condition guaranteeing that if $p_{n+i,n+i}(k)$ (*i.e.*, the i^{th} diagonal element of $P^{\theta\theta}_{k|k-1}$) is not zero, then $p_{n+i,n+i}(k+1) < p_{n+i,n+i}(k)$, almost surely.

Lemma 7.1.3: If the input u_k is a pseudo-random vector with a probability density function (p.d.f.) that is continuous everywhere in $\mathbb{R}^m$, except possibly at a finite number of jump discontinuities, and if $p_{n+i,n+i}(k) > 0$, then $\kappa_{n+i}^T(k)\,[HP_{k|k-1}\,H^T + R_k\,]\,\kappa_{n+i}(k) > 0$, almost surely.

Proof: By algebraic expansion,

$$\kappa_{n+i}^T(k)\,[\,H\,P_{k|k-1}\,H^T + R_k\,]\,\kappa_{n+i}(k) = \sum_{r=1}^{p}\sum_{s=1}^{p} p_{n+i,\,\bar{n}_r}(k)\,m_{r,s}(k)\,p_{n+i,\,\bar{n}_s}(k)\,, \qquad (7.1.13)$$

where $m_{r,s}(k)$ is the (r,s)-element of $[\,H\,P_{k|k-1}\,H + R_k\,]^{-1}$, and $\bar{n}_j = n_1 + \cdots + n_j$. (Note, due to the positive semidefinite nature of $P_{k|k-1}$, from Equation 7.1.13, if $p_{n+i,n+i}(k) = 0$ then $p_{n+i,n+i}(k+1) = 0$.) Defining

$$\widetilde{p}_{\bar{n}_r-1,\bar{n}_s}(k) = \begin{cases} p_{\bar{n}_r-1,\bar{n}_s}(k)\,, & \text{if } n_r > 1\,, \\ 0\,, & \text{if } n_r = 1\,, \end{cases}$$

the factors in Equation 7.1.13 can be expanded:

$$\begin{aligned} p_{n+i,\bar{n}_j}(k) &= \widetilde{p}_{n+i,\bar{n}_j-1}(k-1) - \sum_{r=1}^{p}\sum_{s=1}^{p} p_{n+i,\bar{n}_r}(k)\, m_{r,s}(k)\, \widetilde{p}_{\bar{n}_r-1,\bar{n}_s}(k) \\ &\quad + \sum_{r=1}^{p} z_r(k-1)\, p_{n+i,nr+\bar{n}_j}(k) + \sum_{r=1}^{m} u_r(k-1)\, p_{n+i,n(p+r)+\bar{n}_j}(k) \end{aligned} \tag{7.1.14}$$

The expression resulting from substituting Equation 7.1.14 into Equation 7.13 is extremely messy. But the point is, it shows that $\kappa^T_{n+i}(k)\,[\,H\,P_{k|k-1}\,H^T + R_k\,]\,\kappa_{n+i}(k)$ is a random variable with a continuous p.d.f.

Hence, $\quad \Pr\left[\kappa^T_{n+i}(k)\,[H\,P_{k|k-1}\,H^T + R_k\,]\,\kappa_{n+i}(k) = 0\right] = 0\,.$

Therefore, $\quad \Pr\left[\kappa^T_{n+i}(k)\,[H\,P_{k|k-1}\,H^T + R_k\,]\,\kappa_{n+i}(k) > 0\right] = 1\,.$ ■

It follows from Lemma 7.1.3 that $P^{\theta\theta}_{k|k-1} \to 0$ almost surely, provided that the inputs u_k are pseudo-random sequences with p.d.f.'s. as specified. This can be viewed as a requirement of *persistent excitation.* It indicates that something like a *dither signal* is required in order to improve the estimate at any point, where the dither signal is not allowed to dwell on any particular value. The degree of improvement is related to the dither signal-to-noise ratio, which is in turn related to the convergence rate.

Thus we have already proved the following theorem:

Theorem 7.1.4: For a time-invariant linear system, under standard gaussian assumptions, if the input is persistently exciting then the parameter estimates $\hat{\theta}_{k+1|k}$ of the PLID algorithm converge a.e. to the true parameter vector θ.

Of course, this result does not address the problem of convergence *rate*, or the equally significant problem of numerical ill-conditioning. For example, if the system S has unstable poles, and if the dither signal-to-noise ratio at the input and/or the output is low enough, it may happen that some of the matrices in the PLID algorithm become ill-conditioned before convergence to the correct parameter values can occur. In such a case, the algorithm may appear to converge, but will in fact converge to incorrect values of the numerator parameters. An example of PLID applied to an unstable system is given in Section 9.3. The topic of convergence rate is discussed further in Section 8.1, in which a rough measure of initial convergence rate is presented.

8.0 The Convergence Question for Time-Varying Parameters

The concept of convergence on which the proof of Chapter 7 is based, must be modified somewhat for the case when the parameters of the unknown system are time-varying. One can no longer reasonably expect the parameter estimates to converge to some final value, because the parameters themselves do not converge to any final value. Instead, the PLID algorithm must be judged by whether it converges to within some small region of error around the actual parameter and state values. The "size" of this region depends upon the signal-to-noise ratio. This is the same concept of convergence one would apply to a Kalman filter estimating the states of a stochastic system.

The Gaussian nature of the extended state vector is lost once the parameters begin to vary in a non-deterministic way. That happens because the gaussian input noise is multiplied by the parameters on the way in to the states. Even if the parameters are gaussian, the resulting state values have a probability density function that is not gaussian. Therefore, the PLID algorithm becomes suboptimal for this case.

Section 8.1 develops an expected upper bound (in the ∞-norm sense) on the prediction error, as a function of the noise power, at the time the observability matrix attains full rank. This is, in a sense, a measure of the initial convergence rate of the PLID algorithm, at least for the small noise case.

8.1 A Result on Initial Convergence Rate (Small Noise Case)

This section develops an approximation of a bound on the PLID estimate error at the time the observability matrix attains full rank. The interesting point about the development here is that the assumption of gaussian white noise is dropped; in fact, no assumptions are made about the noise except that it is zero mean, and the norm of its autocovariance is bounded by a known constant.

Furthermore, the assumption of time-invariant parameters is dropped. The variation of the parameters is modeled as another noise input to the system. The only requirement on this parameter noise is that it is zero mean and the norm of its autocovariance is bounded by a known constant.

Thus, this section determines a rough measure on the convergence rate of the PLID algorithm when some of the assumptions in its derivation are violated. It turns out that the PLID algorithm is nearly deadbeat (*i.e.*, time-optimal) for the small noise case, where "noise" now includes some other density functions besides gaussian, and includes parameter variations. Indeed, this is not surprising because the small noise case is essentially a perturbation of the deterministic case, which was shown to be deadbeat in Section 6.2.

The results of this section are based on the assumption that, although the PLID algorithm is suboptimal when the parameters are time-varying, or when the noises are not gaussian, it will still usually provide a better estimate than a linear fit (deterministic PLID) of the first $N \triangleq (m+p+1)n$ data points. The last statement is certainly true if the parameters are constant and the noises are gaussian, because then the PLID algorithm is optimal.

The method, then, is to apply the deterministic PLID algorithm to $\{ u_0 , \dots , u_{\rho-1} , z_0 , \dots , z_{\rho-1} \}$, where $\rho = (N/p)^{+}$ (with the superscript "+" indicating that the division is rounded upward). Under the assumption that the input is persistently exciting, there will be a total of N independent observations in the first ρ measurements. Therefore, it should be possible to compute an approximate bound on the error $\bar{e}_\rho$ of the estimate $\bar{s}_{\rho|\rho-1}$ of the extended state vector s_ρ. The notation of the estimate now shows an overbar, to indicate that it is definitely suboptimal.

The deterministic PLID algorithm is given by Equations 6.1.1 through 6.1.3. Selecting the initial error covariance $\bar{P}_{0|-1}$ to be symmetric positive definite, the error resulting from the use of this algorithm is

$$\begin{aligned}
\bar{e}_{k+1} &= \bar{s}_{k+1|k} - s_{k+1} \\
&= (F_k - \bar{K}_k H) \bar{s}_{k|k-1} + \bar{K}_k z_k - (F_k s_k + G \eta_k) \\
&= (F_k - \bar{K}_k H) \bar{s}_{k|k-1} + \bar{K}_k (H s_k + w_k) - (F_k s_k + G \eta_k) \\
&= (F_k - \bar{K}_k H) (\bar{s}_{k|k-1} - s_k) + \bar{K}_k w_k - G \eta_k \\
&= (F_k - \bar{K}_k H) \bar{e}_k + \bar{K}_k w_k - G \eta_k .
\end{aligned} \tag{8.1.1}$$

Therefore,

$$\begin{aligned}
\bar{e}_\rho = \Big[& (F_{\rho-1} - \bar{K}_{\rho-1|\rho-1} H) \dots (F_0 - \bar{K}_0 H) \bar{e}_0 \\
& + (F_{\rho-1} - \bar{K}_{\rho-1|\rho-1} H) \dots (F_1 - \bar{K}_{1|1} H) \bar{e}_0 (\bar{K}_{0|0} w_0 - G \eta_0) \\
& + \cdots \\
& + (F_{\rho-1} - \bar{K}_{\rho-1|\rho-1} H) (\bar{K}_{\rho-2|\rho-2} w_{\rho-2} - G \eta_{\rho-2}) \\
& + (\bar{K}_{\rho-1|\rho-1} w_{\rho-1} - G \eta_{\rho-1}) \Big]
\end{aligned} \tag{8.1.2}$$

But the assumption of persistent excitation means that Φ_ρ, the ρ^{th} observability matrix, is full rank. Therefore, by Theorem 6.2.6, the first term of Equation 8.1.2 is zero. So, the error $\bar{e}_\rho$ is independent of the initial estimate error, indicating that the choice of initial estimate is arbitrary.

Thus, it only remains to compute approximations of the trailing terms of Equation 8.1.2. Let us start by making several assumptions concerning the various system noises. Suppose the autocovariances of the input, output, state, and parameter noises in the original unknown MIMO system satisfy

$$\begin{aligned} &\|R_k\|_\infty \le \varepsilon^2 \\ &\|Q_0(k)\|_\infty \le \varepsilon^2 \\ &\|\Sigma_k^{xx}\|_\infty \le \varepsilon^2 \\ &\|\Sigma_k^{\theta\theta}\|_\infty \le \varepsilon^2 \\ &\qquad\qquad \text{with } 0 < \varepsilon \ll 1, \end{aligned} \tag{8.1.3}$$

where $\| \;\; \|_\infty$ indicates the infinity norm, which, in matrices, is the maximum row sum.

Next, assume that the gain matrix $\overline{K}_k$ satisfies $\|\overline{K}_k\|_\infty \le p \; \forall \; k = 0, \ldots, \rho - 1$. This is a fairly reasonable assumption, based on extensive simulations of the PLID algorithm. Occasionally, individual elements will exceed unity slightly, especially when the observability matrix is nearly full rank, but for the most part the elements of $\overline{K}_k$ do satisfy the assumption.

Assume that the inputs u_k are kept small enough that the following is true

$$\| u_k \|_\infty \le \varepsilon \text{ and } \quad \| z_k \|_\infty \le \varepsilon, \tag{8.1.4}$$

where $\varepsilon > 0$ is the square root of the limit in Equation 8.1.3. Later, an argument will be put forward to relax this restriction without losing the results.

In the rest of the development, the symbols $\underset{\approx}{<}$ and $\underset{\approx}{>}$ will be used to denote approximated inequalities.

Finally, assume that the various noise terms w_t and η_t do not generally exceed three standard deviations of the limiting covariance ε^2. That is, $\|w_t\|_\infty \underset{\approx}{<} 3\,\varepsilon$, and $\|\eta_t\|_\infty \underset{\approx}{<} 3\,\varepsilon$. Obviously, this is a very rough approximation, but useful.

Denote the maximum element of the parameter vector $\| \theta \|_\infty = M_\theta$.

An approximation can now be computed for the common factors in Equation 8.1.2. From the structures of the matrices $\overline{K}_k$ and G, and recalling that there are m inputs and p outputs,

$$\begin{aligned} \| \overline{K}_{l|l}\, w_l - G\, \eta_l \|_\infty &\lesssim p\,(3\,\varepsilon) + [(p+m)\, M_\theta\ + 1]\,(3\,\varepsilon) \\ &= 3\,[(p+m)\, M_\theta\ + p + 1]\,\varepsilon \end{aligned} \tag{8.1.5}$$

The other factors are approximated by considering the matrix structure,

$$\| F_l - \overline{K}_{l|l}\, H \|_\infty \ \lesssim\ 1 + p + (p+m)\,\varepsilon \tag{8.1.6}$$

Applying the approximations of Equations 8.1.5 and 8.1.6 to the nonzero terms in Equation 8.1.2,

$$\begin{aligned} \| \bar{e}_\rho \|_\infty &\lesssim \sum_{j=0}^{\rho-1} [\ 1 + p + (p+m)\,\varepsilon\ ]\ 3\,[(p+m)\, M_\theta\ + p + 1]\,\varepsilon \\ &= \left[\frac{1 - [1 + p + (p+m)\,\varepsilon\,]^\rho}{1 - [1 + p + (p+m)\,\varepsilon\,]} \right] 3\,[(p+m)\, M_\theta\ + p + 1]\,\varepsilon \\ &\approx \frac{(1+p)^{N/p}\ -1}{p}\ \ 3\,[(p+m)\, M_\theta\ + p + 1]\,\varepsilon \end{aligned} \tag{8.1.7}$$

For example, for a second-order SISO system, $\rho = N = 6$; hence,

$$\| \bar{e}_6 \|_\infty \lesssim [(1+1)^6 - 1]\,(3)\,[2\, M_\theta\ + 2]\,\varepsilon = 378\,(1 + M_\theta\,)\,\varepsilon \tag{8.1.8}$$

It turns out that Equation 8.1.8 is a fairly loose approximation for a wide range of S/N ratios.

The argument for loosening the restriction on the inputs and outputs, given in Equation 8.1.4, is the following. If the estimate error is bounded by a particular level with a certain S/N ratio, the estimate error should be even smaller if the S/N ratio is increased, because the system becomes more nearly deterministic. As the norm of the input increases, the S/N ratio increases; therefore, the limit computed for the "small" input should hold for any larger input.

The latter observation does not hold in case the output magnitude becomes very large while the input magnitude remains small (such as in an unstable system). The reason is that such an occurrence causes several of the PLID algorithm's matrices to become ill-conditioned. More intuitively, in that case even very small amounts of input noise cause large variations at the output, which cannot rightly be attributed to noise by the algorithm. The usual result is that PLID determines the denominator (autoregressive) parameters very accurately, but is unable to determine the numerator (moving average) parameters.

Equation 8.1.7 might be called the *expected upper bound* on the estimate error. Clearly, it is most useful as a bound when the system has a single output, because the factor $(1+p)^{N/p}$ becomes large very fast as p increases above unity.

9.0 PLID Simulation Results

This chapter presents simulation results that verify and clarify the theoretical convergence findings of Chapters 6 and 7.

Section 9.1 presents results from simulation of a fourth-order, two-input, two-output system, at various levels of noise at the inputs and outputs. The system chosen for the examples of this section is well-conditioned in the sense that all the poles and zeros are well separated from each other, and the system is stable. Of the four subsystems, three have non-minimum-phase zeros; these present no special problem to the PLID algorithm, because it is derived with no assumptions about the locations of poles and zeros. These examples illustrate the degradation of performance that results from increasing noise levels.

Theoretically, the PLID parameter estimates should converge regardless of the noise level. However, in practice, that is not true. Probably the best explanation is that the theoretical convergence results only hold for computations that have infinite precision. Thus, in a real computer, the very slow convergence rate one expects with very high noise levels, is swamped by the computation error. It is easy to simulate this problem, of course. The noise levels at which the PLID algorithm fails to converge depend upon a number of things. Among them are the number of inputs and outputs, and the separation between poles and zeros.

Section 9.2 presents the situation where the PLID algorithm is to be used, even though one or more of the basic assumptions in the derivation of PLID are violated. In this case, the system is not strictly proper, giving rise to some explicit nonlinearities in the estimator. Clearly, this variation of the PLID algorithm is not optimal; the question is, how does it compare to other common methods? It is compared to the Extended Kalman Filter.

A second-order SISO system with a near cancellation of a pole/zero pair was chosen for the comparison, because it is a fairly difficult identification problem. So, because the system is also non-proper, a total of five unknown parameters and two states must be estimated. Comparisons are carried out for various levels of initial estimate error, and consistently show the superior convergence properties of the PLID algorithm, even in this suboptimal form.

Finally, Section 9.3 presents results from simulation of a third-order SISO system with three unstable poles and two non-minimum-phase zeros. The only problem for the PLID algorithm in such a case is that the output of the system can become so large that the error covariance matrix becomes numerically ill-conditioned before convergence of the parameters has occurred. Indeed, that happens in the simulation of Section 9.3. (In Section 10.1, an adaptive feedback controller is used to prevent divergence of the output of an unstable system, thereby simulating the case where the matrices in the PLID algorithm remain well-conditioned; problems arising from the lack of persistent excitation are demonstrated there.)

Ill-conditioning can occur as follows. All the elements of $P^{\theta\theta}_{k+1|k}$, which are related to the parameter estimates, are non-increasing. However, the elements of $P^{xx}_{k+1|k}$, which are related to the states, are essentially linear combinations of the elements of $P^{\theta\theta}_{k+1|k}$, where the input and output measurements appear as coefficients. Thus, if the output blows up, so do the elements of $P^{xx}_{k+1|k}$.

So, some eigenvalues of $P_{k+1|k}$ are decreasing towards zero while others are increasing towards infinity (or, rather, to the point of computer overflow). Hence the numerical ill-conditioning. If the parameter estimates have converged before the ill-conditioning occurs, then the state estimates will be good up to the point of overflow.

The convergence rate is, as usual, affected by the noise level; thus, if the noise level is low enough, PLID will achieve an accurate estimate of the parameters before ill-conditioning occurs.

9.1 Simulation of a Two-Input, Two-Output System

The purpose of this section is to provide a look at typical PLID algorithm performance. The main problem in assessing its performance is the overwhelming quantity of the data associated with the estimator. There are three main quantities of interest, (1) the estimate error vectors, (2) the gain matrices, and (3) the error covariance matrices.

The particular system being simulated has two inputs, two outputs, and a total of four states. Therefore, the estimate error vector has 20 elements, the gain matrix has 40 elements, and the error covariance matrix has 400 elements. That is altogether too much data to present in detail. Some essential feature of the various matrices must be culled from the data.

One essential feature of each matrix mentioned above is its norm. The norm is much more revealing if the various matrices are first partitioned into those elements related to the states x_k, those related to the autoregression parameters θ_A, and those related to the moving average parameters θ_B; that is what has been done. The graphs are presented in logarithmic form, because the norms typically span several orders of magnitude as the algorithm proceeds.

In most cases, the quantitative aspects of the graph are not nearly as important as the qualitative aspects. For example, does the parameter estimate error tend to decrease? Does the error covariance tend to decrease?

The following system was simulated:

$$\begin{bmatrix} Y_1(z) \\ Y_2(z) \end{bmatrix} = \frac{\left[\frac{(-1)(z-2.66997)(z+0.31186)(z+0.05139)}{(z+1.623)(z-0.558-j0.198)(z-0.558+j0.198)} \quad \frac{(z-1.55005)(z+0.78359)(z-0.04026)}{(-1)(z-0.775)(z-0.109-j0.910)(z-0.109+j0.910)} \right]}{(z-0.5)(z+0.9)(z+0.7+j0.7)(z+0.7+j0.7)} \begin{bmatrix} U_1(z) \\ U_2(z) \end{bmatrix}$$

For the simulation, a pseudo-random gaussian white input sequence $\{u_k\}$, was corrupted by a (pseudo) random gaussian white noise sequence $\{v_k\}$ before being applied to the system inputs. A (pseudo) random gaussian white noise sequence $\{w_k\}$ was added to the system output sequence $\{y_k\}$ to obtain the measurement sequence $\{z_k\}$. The only data available to the PLID algorithm were $\{u_k\}$ and $\{z_k\}$. Furthermore, an (ostensibly) unknown gaussian white state noise sequence $\{\Xi_k\}$ is added directly to the system states during the simulation.

Signal-to-noise ratios at the input were fairly simple to determine, because both the signal and the noise are gaussian white sequences. However, the system output is a highly correlated sequence, which is to be expected. The output signal power was estimated by averaging over several hundred samples. Presenting a single figure for the output signal-to-noise ratio is somewhat misleading because, due to the highly correlated nature of the output sequence, the signal may be small for fairly long periods of time, while during other periods it may be much larger. At the same time, the output noise power remains fairly constant, because for the simulations, the autocovariance was taken to be constant, and the sequence is white.

While some state noise was generated and added in to the states, it was kept at essentially negligible levels in these simulations.

For the simulations, the input signal sequence, input noise sequence, state noise sequence, and output noise sequence were all generated by the International Mathematics and Statistics Library (IMSL) double precision routine "GGNML," which generates a pseudo-random sequence with standard gaussian statistics (*i.e.*, unity variance, zero mean) based upon some seed value supplied by the user. The following seeds were specified in the simulations:

For $\{u_1(k)\}$: 123457.0
For $\{u_2(k)\}$: $\pi \times 123457.0$
For $\{v_1(k)\}$: 372845.0
For $\{v_2(k)\}$: $\pi \times 372845.0$
For $\{w_1(k)\}$: 564213.0
For $\{w_2(k)\}$: $\pi \times 564213.0$
For $\{\xi_1(k)\}$: 564213.0
For $\{\xi_i(k)\}$: $(\pi)^{i-1} \times 564213.0$ for $i = 1, \dots ,n$.

There was no noise input to the parameters, because they are constant.

The input sequence was left at unity variance. The various noise levels were established by appropriately scaling the sequences from the IMSL routine. The output noise scaling was determined by first running the system simulation without any output noise to obtain the average output power over several hundred iterations. Using the same seed, the IMSL routine will always generate precisely the same sequence (hence the name *pseudo*-random sequence). Thus, the same input was applied on a second simulation, this time adding in the appropriate level of output noise, and running the PLID algorithm.

Five complete simulations were run, with 500 iterations each, with noise levels adjusted to simulate the cases where all the signal-to-noise levels were either 20 dB, 40 dB, 60 dB, 80 dB, or 100 dB. The input sequence was the same for each iteration. In each figure in this section, results from all five simulations are presented simultaneously, thereby showing the effects of increasing noise on the performance of the PLID algorithm.

Figure 8 compares the state estimate error at the various noise levels. For each 20 dB increase in the noise level, the state estimate error increases, roughly, by slightly less than 10 dB. In the worst case (20 dB signal-to-noise ratio) the noises at the output have greater than unity variance.

There is surprisingly little variation in the state gains, as seen in Figure 9, which tend to achieve steady-state value fairly soon. However, the state estimate error covariances, given in Figure 10, correspond quite well with the estimate error norms. At each noise level, the trace of the state estimate error covariance is roughly equal to the square of the error norm (or double, on the *logarithmic* scale shown).

The autoregression parameter estimate error is affected in the same way as the state estimate error by increasing noise levels, as shown in Figure 11. The characteristic sharp decrease in error when the observability matrix Φ_k achieves full rank is evident just after iteration ten in all but the worst noise case. (This occurs after iteration ten because there are twenty unknowns, but there are *two* observations at each time.) Very similar comments can be made about the moving average parameter estimate error, which is shown in Figure 12.

The gains associated with the parameter estimates are shown in Figure 13 and Figure 14. These figures are nearly identical, showing the progressively smaller weight given to successive observations. In each of these figures, the similarity of the five curves is remarkable. In part, it is due to the fact that the (known) input sequence $\{u_k\}$ is the same, and the output sequences $\{z_k\}$ are quite similar, modulo some effects of the increasing noise levels. It is somewhat surprising that the increasing values of the elements in matrices Q_0 and R have so little effect on the gain computation, although it should be noted that there are some significant differences in the gains during the first 100 iterations, or so.

The error covariances associated with the parameter estimates are shown in Figure 15 and Figure 16. By definition, these must be non-increasing. The fact that they are strictly decreasing indicates that the input is, indeed, persistently exciting, as discussed in Section 7.1. It is this continual decrease that fuels the decrease in the parameter gains.

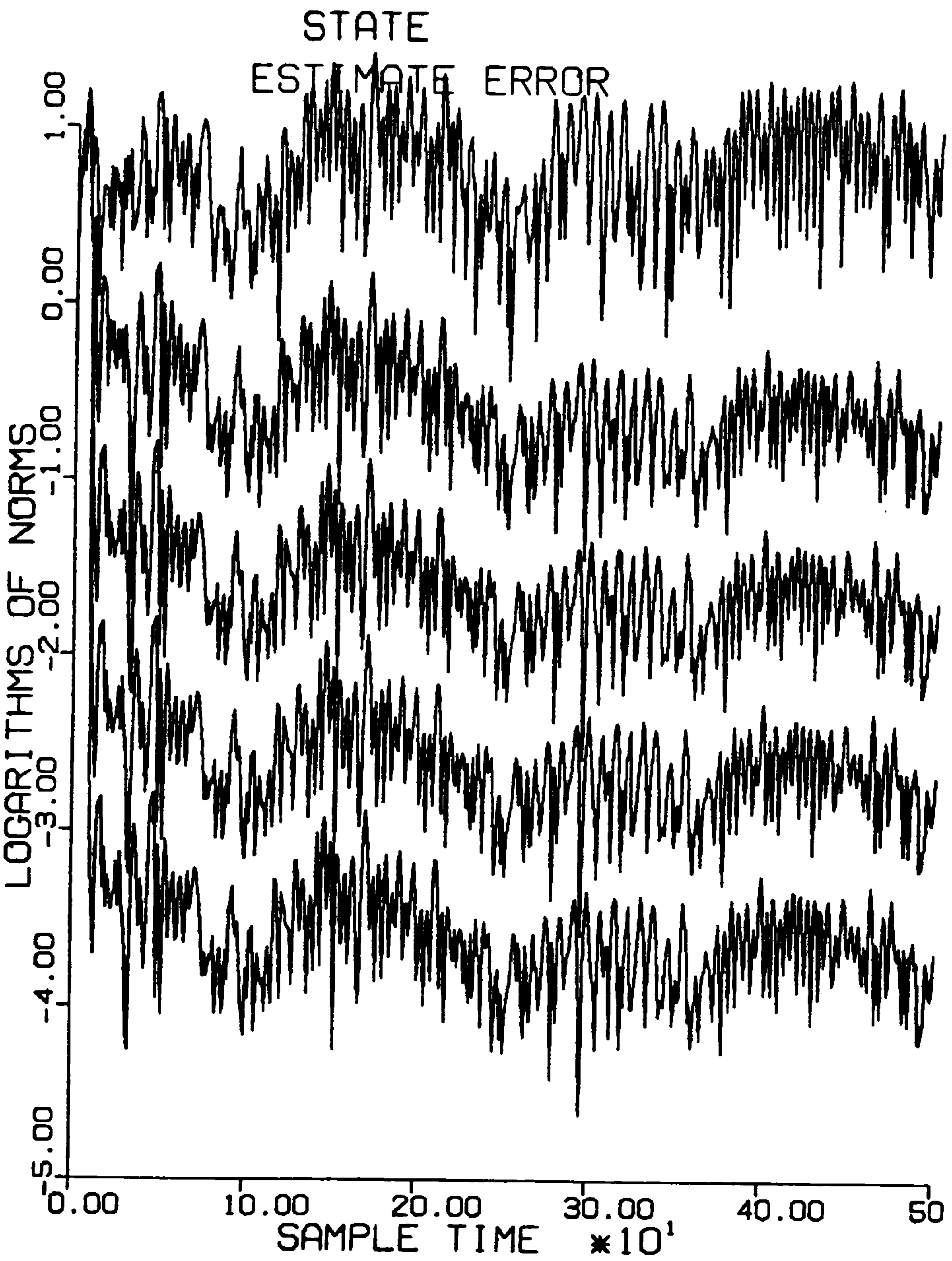

Figure 8. State estimate error for a non-minimum-phase system: Log of the norm of the state estimate error subvector when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

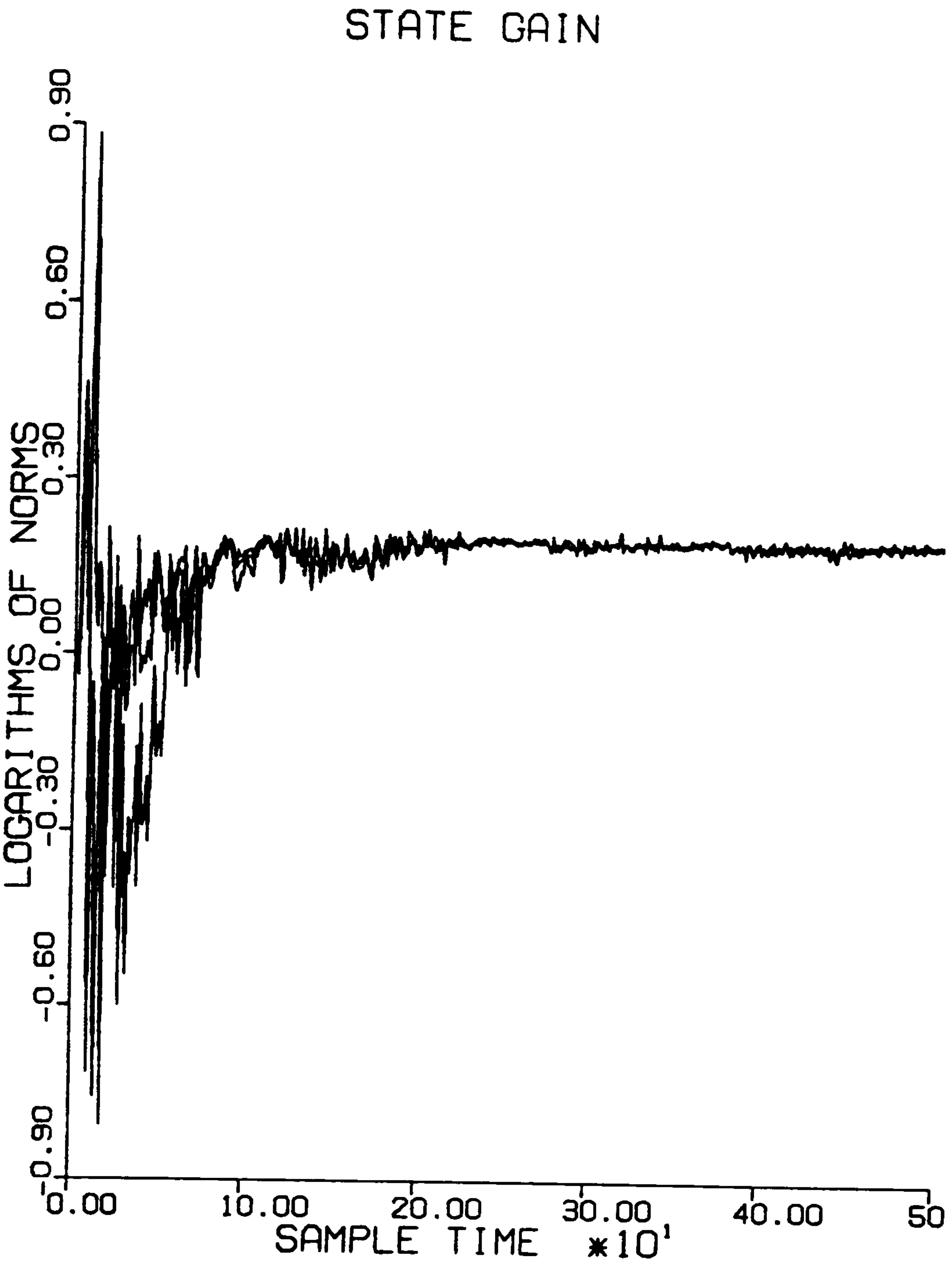

Figure 9. State gain for a non-minimum-phase system: Log of the norm of the state gain submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

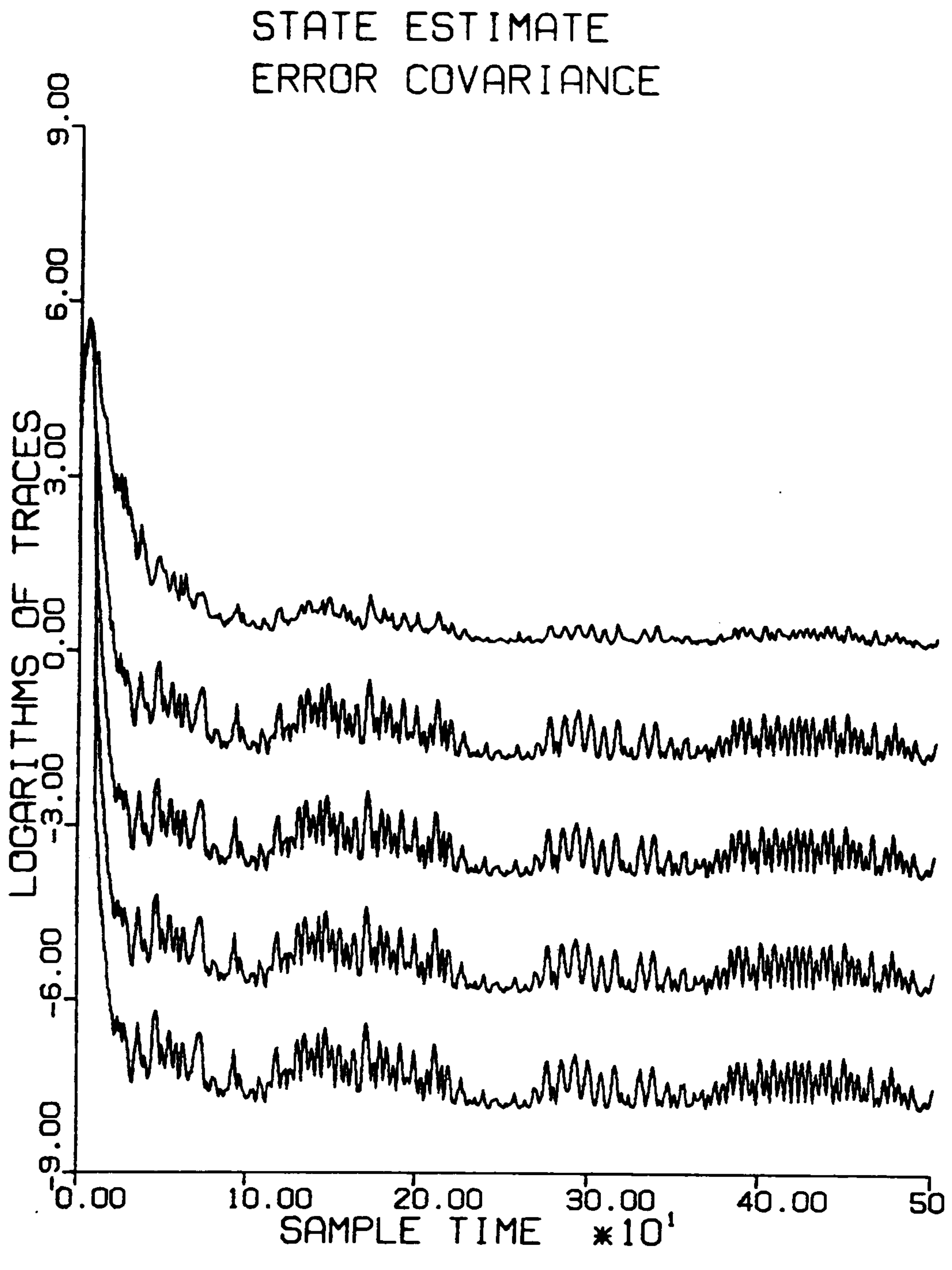

Figure 10. State error covariance for a non-minimum-phase system: Log of the norm of the state error covariance submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

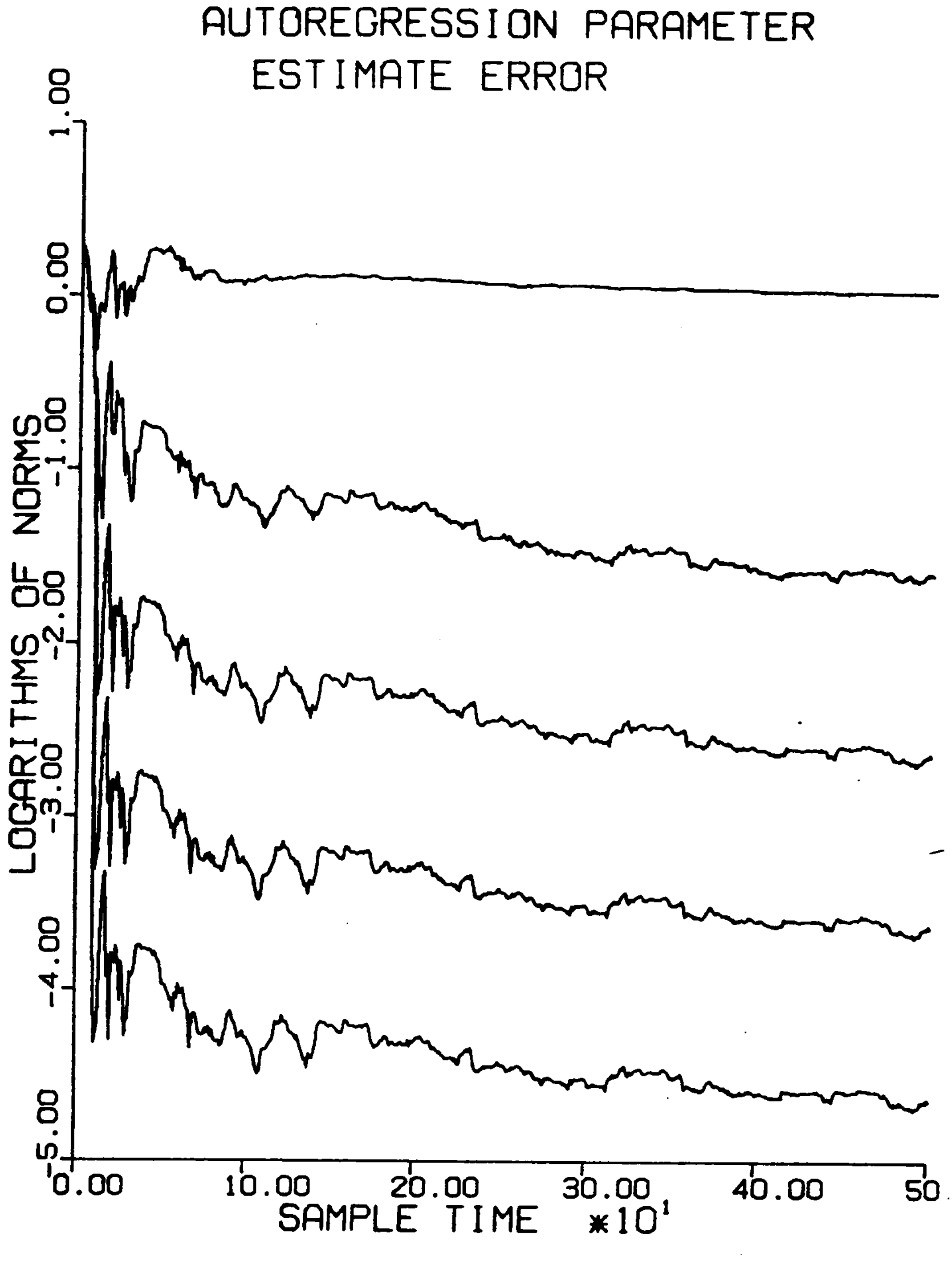

Figure 11. Autoregression-parameter estimate error for a non-minimum-phase system: Log of the norm of the autoregression parameter estimate error subvector when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

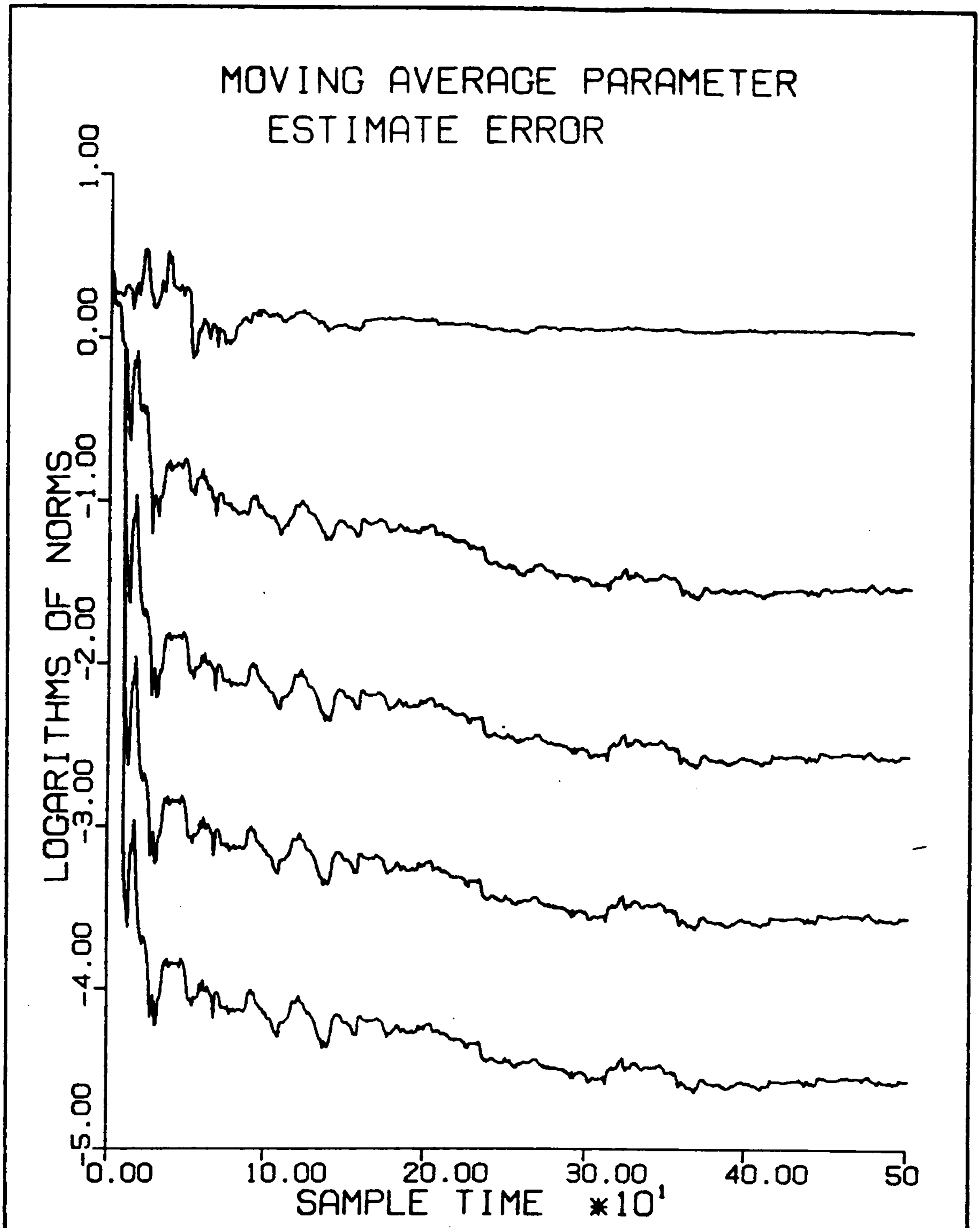

Figure 12. Moving-average-parameter estimate error for a non-minimum-phase system: Log of the norm of the moving-average-parameter estimate error subvector when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

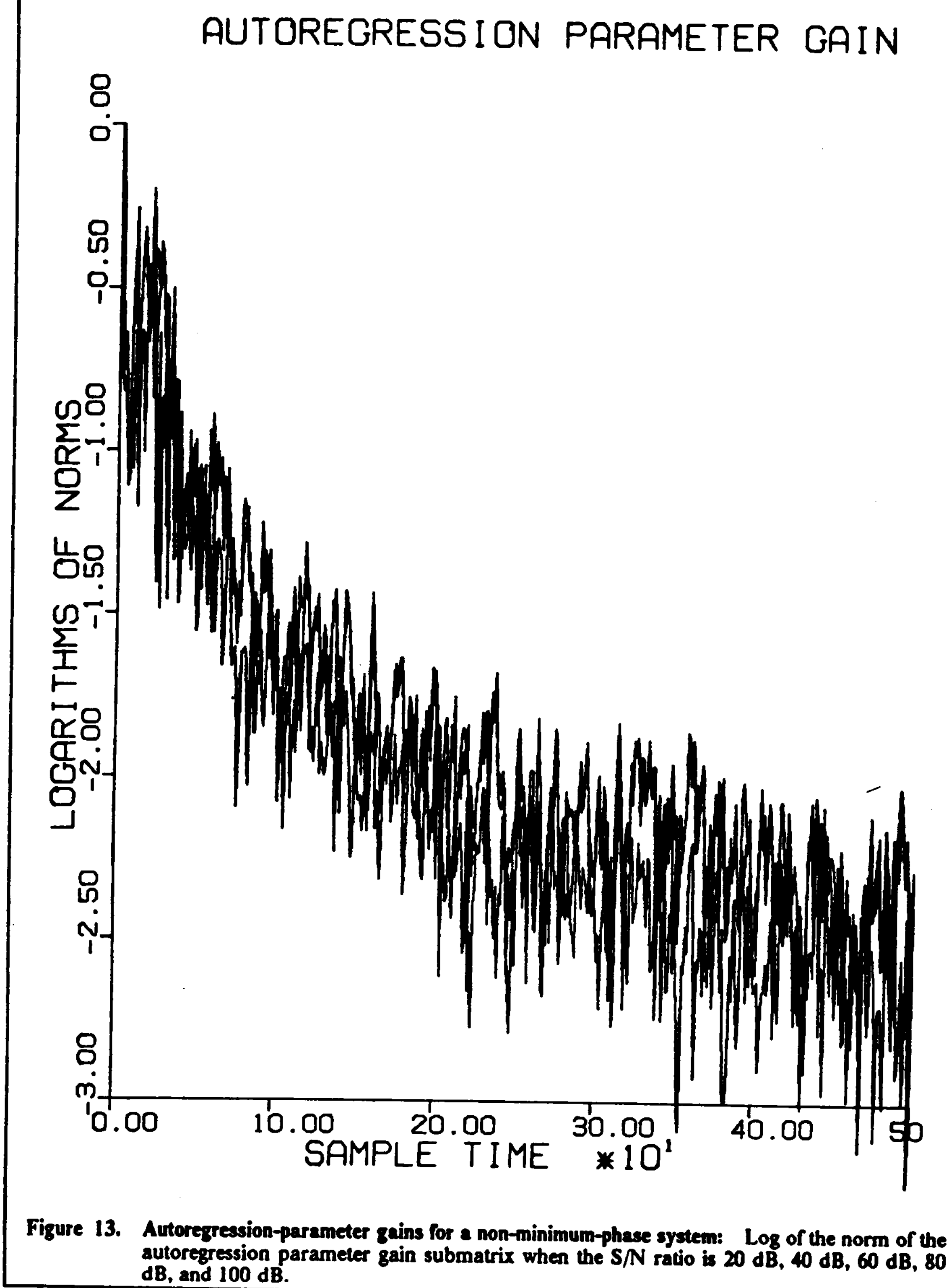

Figure 13. **Autoregression-parameter gains for a non-minimum-phase system:** Log of the norm of the autoregression parameter gain submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

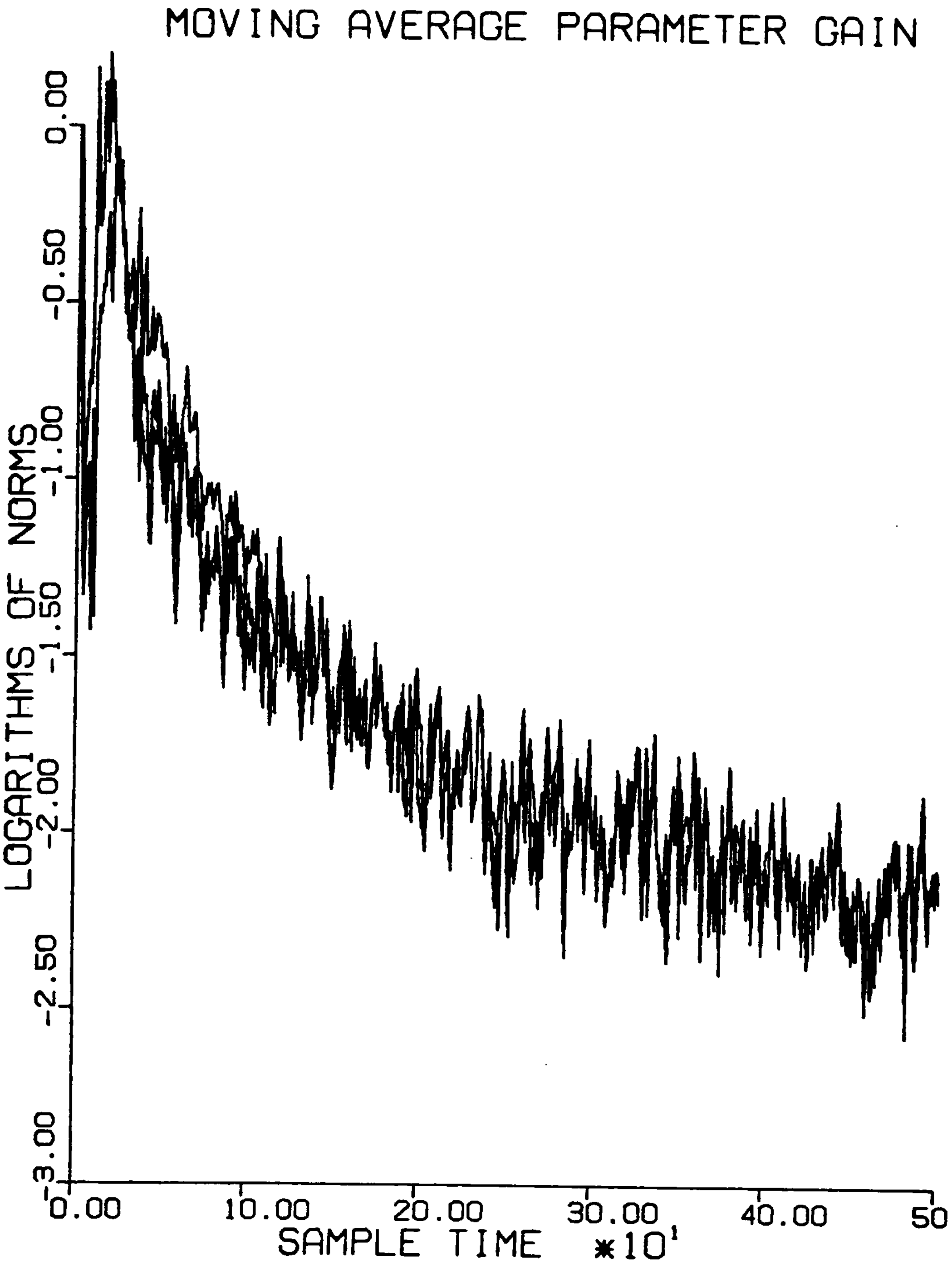

Figure 14. Moving-average-parameter gain for a non-minimum-phase system: Log of the norm of the moving-average-parameter gain submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

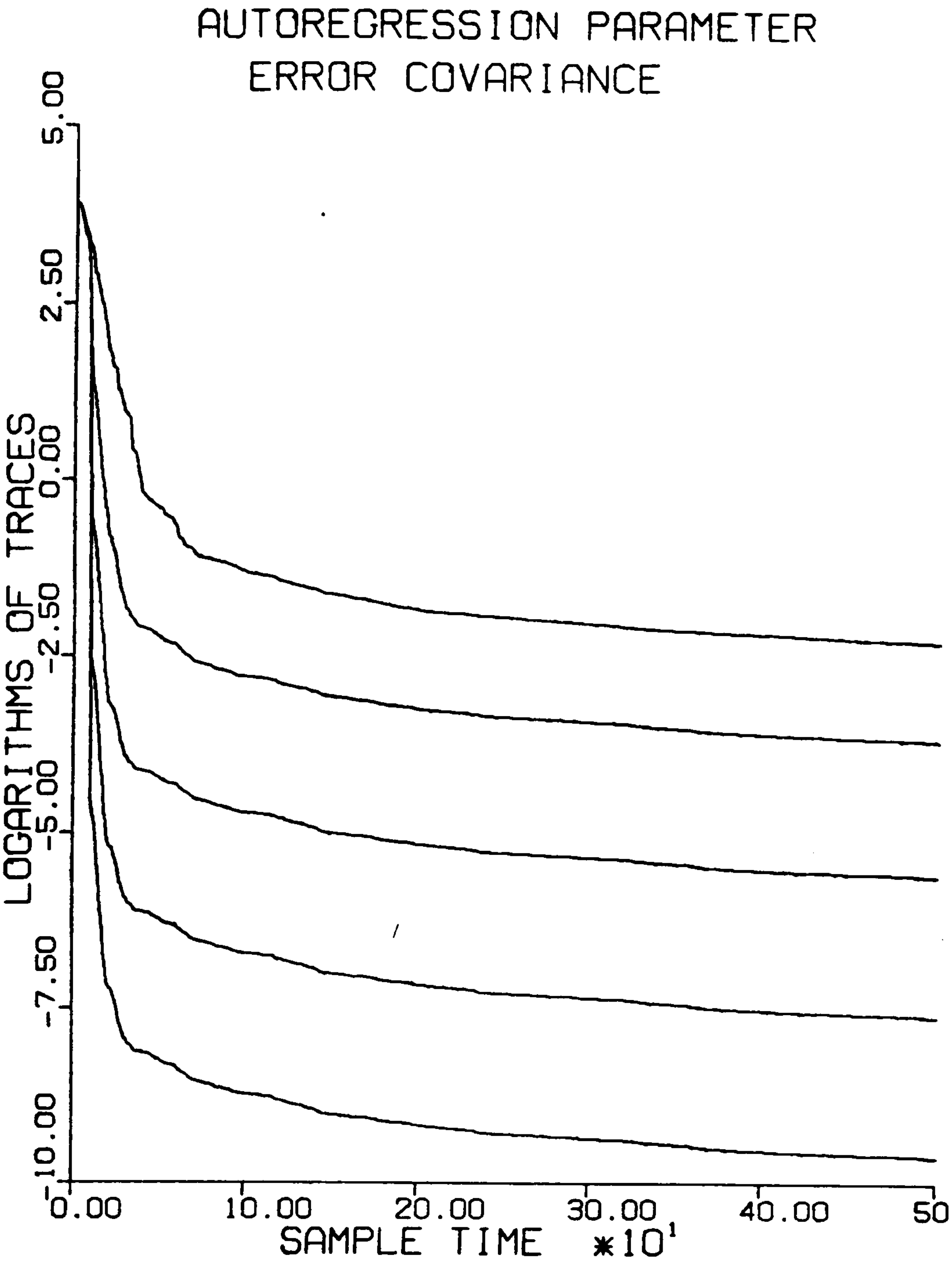

Figure 15. Autoregression-parameter error covariance for a non-minimum-phase system: Log of the norm of the autoregression parameter error covariance submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

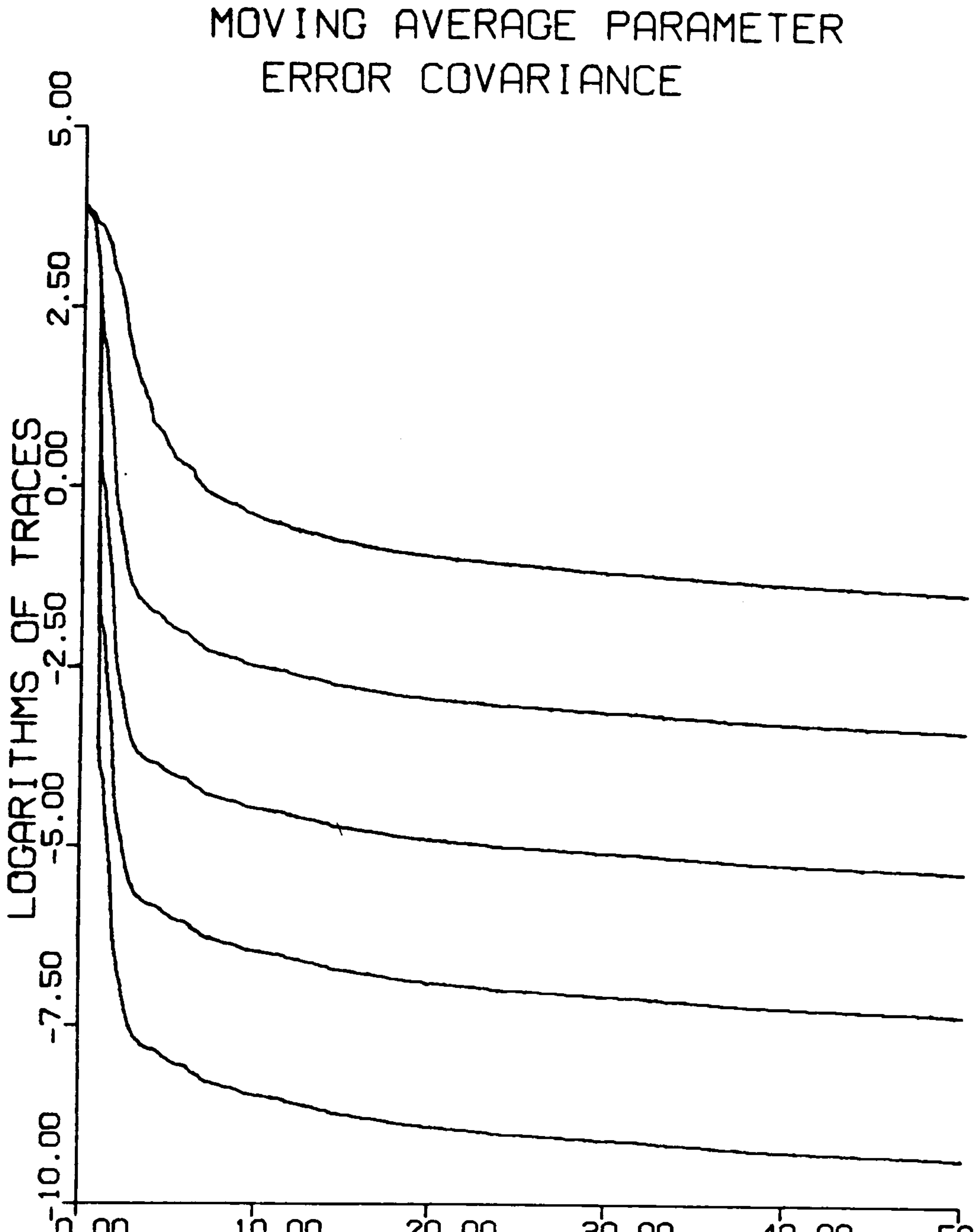

Figure 16. **Moving-average-parameter error covariance for a non-minimum-phase system: Log of the norm of the moving-average-parameter error covariance submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.**

9.2 Comparison of Convergence of PLID and the EKF

This section presents a case where the assumptions of the PLID derivation are slightly violated. While the PLID algorithm is optimal for systems conforming to the assumptions under which it is derived, it is not completely clear what becomes of the algorithm when the assumptions are violated.

The particular system chosen for this section is not strictly proper, so the PLID is not, strictly speaking, applicable. However, only slight modifications of the algorithm are required to accomodate this case. In particular, in estimating the noise input matrix G, there are explicit nonlinearities in the estimator. Hence, it is suboptimal. But just how bad is the resulting estimator?

In order to get an idea of its performance, a comparison is made between the suboptimal version of PLID and the Extended Kalman Filter. To make the comparison more interesting, a system having a near cancellation of a pole/zero pair was chosen. This makes the system hard to identify because it is difficult to persistently excite the mode associated with the nearly cancelled pole.

The following time-invariant SISO system, ostensibly unknown, was used to generate an output sequence from a known input sequence. Note the near pole/zero cancellation, and the presence of one non-minimum-phase zero. The input was corrupted by unknown gaussian white noise before being applied to the system. Similarly, the output measurements were corrupted by another (independent) sequence of gaussian white noise.

The true system transfer function is

$$H(z) = \frac{z^2 + 0.13\,z - 1.067}{z^2 - 1.471\,z + 0.4855} = \frac{(z - 0.970)(z + 1.100)}{(z - 0.971)(z - 0.500)} \tag{9.2.1}$$

According to the notation of Equation 4.1.1, the true parameter values are

$$\begin{aligned} a_0 &= -0.4855 \\ a_1 &= 1.4710 \\ b_0 &= -1.0670 \\ b_1 &= 0.1300 \\ b_2 &= 1.0000 \end{aligned}$$

The main point of the simulations was to compare the convergence properties of the two algorithms, particularly with respect to convergence of the parameter estimates. Both the *rate* of convergence and the *region* of convergence are of interest.

There are seven states in the extended system: five parameters and two states from the original unknown system. Both algorithms were started up with a least-squares fit of the first $3n+1 = 7$ data points. Due to the near-cancellation in the system, the fit of the first seven data points had considerable error. By adjusting the initial error covariance, the error in the least-squares fit of the first seven points could be adjusted. This provided a means of testing the convergence properties of the respective algorithms for various levels of initial estimate error.

The slowness of convergence to the correct values is due to the near-cancellation of the pole/zero pair, which makes persistent excitation quite difficult. Since the algorithms start with an estimate that is a reasonable linear fit of the initial data, it is difficult for them to detect error in the initial estimate. Both algorithms were run with exactly the same sequences of input and output data, which had noise levels of -37 dB and -40 dB, respectively. As in the other simulations, the state noise was negligible in these simulations.

The results indicate that some of the proven convergence properties of the PLID algorithm are retained even when the assumptions used to derive it are violated slightly.

As with any Kalman-type filter, there are some parameters that must be selected to start the PLID algorithm. The elements of the matrix Q were chosen as follows: The variances q of the input noise v_k and r of the output noise w_k were known. The variances σ_i^2 of the additive noise ζ_i on the unknown system parameters were assumed very small, anticipating that the unknown system is time-invariant.

The system was driven by a known input sequence $\{u_k\}$, consisting of unity variance white noise corrupted by an (ostensibly unknown) white noise sequence $\{v_k\}$, with a variance of 0.01. Thus the input signal-to-noise ratio was 40 dB.

The output measurement sequence $\{z_k\}$ was generated by corrupting the sequence $\{y_k = x_n(k) + b_n(u_k + v_k)\}$ with the (ostensibly unknown) white noise sequence $\{w_k\}$, with known variance. As with the previous simulation, the system was run first with no output noise added, to allow the average output power to be calculated. Based upon that average, the average output signal-to-noise ratio was established at 37 dB.

Both the Extended Kalman Filter (EKF) and the PLID algorithm were implemented using a square-root type of algorithm (*UD*-decomposition) to propagate the error covariance matrix. This minimized problems due to computer inaccuracy, and avoided such embarassments as having the error covariance matrix attain negative eigenvalues.

In the *UD*-decomposition method of propagating the error covariance matrix, one starts by assuming the initial error covariance in the form $P_{0|-1} = U_0 D_0 U_0^T$, where U_0 is an upper triangular unitary matrix, and D_0 is a positive definite diagonal matrix. The easiest choices are, of course, $U_0 = \mathbf{I}$, and $D_0 = \alpha_0 \mathbf{I}$, where typically $\alpha_0 \gg 1$.

The error covariance is then propagated in two steps,

$$\begin{aligned}\overline{U}_{k+1}\,\overline{D}_{k+1}\,\overline{U}^T_{k+1} &= F_k\,P_{k|k-1}\,F_k^T + G_k\,Q\,G_k^T \\ &= F_k\,U_k\,D_k\,U_k^T\,F_k + G_k\,Q\,G_k^T\end{aligned} \tag{9.2.2}$$

and

$$\begin{aligned}P_{k+1|k} &= U_{k+1}\,D_{k+1}\,U^T_{k+1} \\ &= \overline{U}_{k+1}\,\overline{D}_{k+1}\,\overline{U}^T_{k+1} - K_{k|k}\left[H\,P_{k|k-1}\,H^T + R_k\right]K^T_{k|k}\end{aligned} \tag{9.2.3}$$

Equation 9.2.2 is implemented by the modified weighted Gram-Schmidt orthogonalization and matrix factorization given in [31], which starts by adopting the form

$$\overline{U}_{k+1}\,\overline{D}_{k+1}\,\overline{U}^T_{k+1} = \left[F_k\,U_k\,|G_k\right]\begin{bmatrix} D_k & 0 \\ 0 & Q\end{bmatrix}\begin{bmatrix} U_k^T F_k^T \\ G_k^T\end{bmatrix} \tag{9.2.4}$$

in which Q is assumed to be a diagonal matrix.

Equation 9.2.3 is implemented by the Agee-Turner positive definite factorization update given in [32], which takes advantage of the equation's form,

$$P_{k+1|k} = \overline{U}_{k+1}\,\overline{D}_{k+1}\,\overline{U}^T_{k+1} + c_k\,\xi_k\,\xi_k^T \tag{9.2.5}$$

where c_k is a scalar, and ξ_k is a vector.

Figure 17 shows the EKF parameter estimates converging to the exact parameter values, which appear as straight lines, from an initial estimate that is reasonably accurate. Convergence occurs after about 1100 iterations. The remainder of the figures in this section present comparisons of the parameter estimates of the two algorithms, where each algorithm starts with similar initial estimate error. The comparisons show that as the initial estimate error becomes worse, the EKF has a more difficult time converging, while convergence of the PLID algorithm is only delayed somewhat.

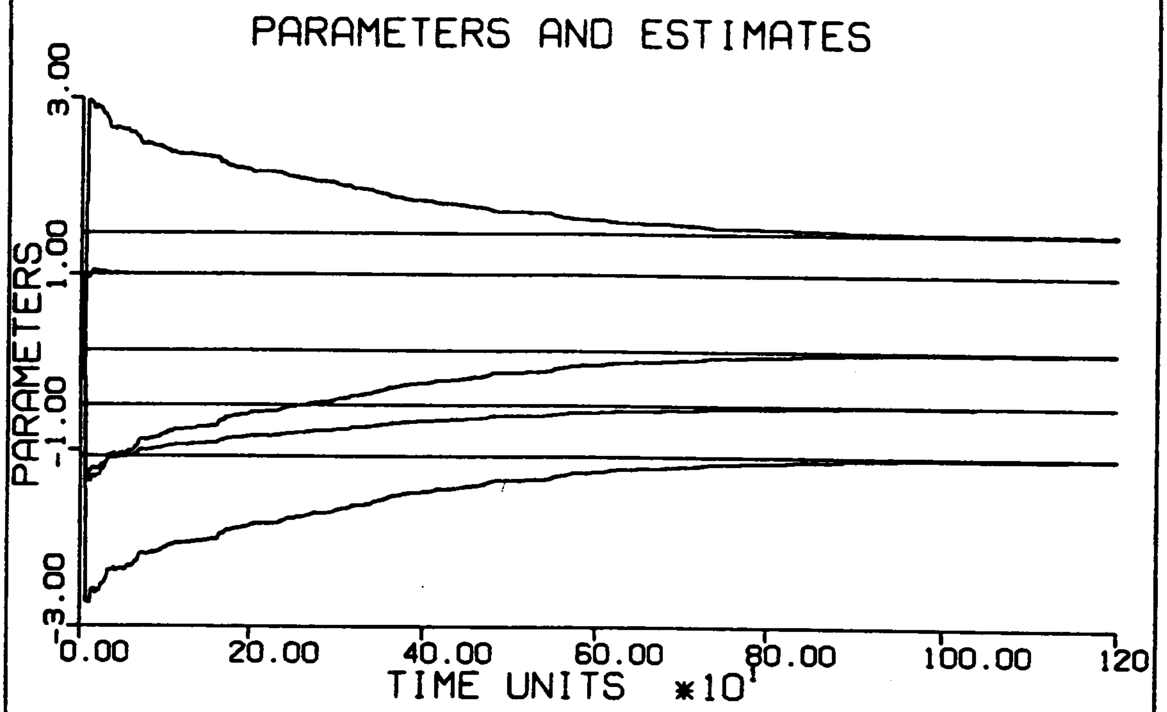

Figure 17. Showing convergence of the EKF with initial error about 2.0: EKF parameter estimates converge to correct values after about 1100 iterations.

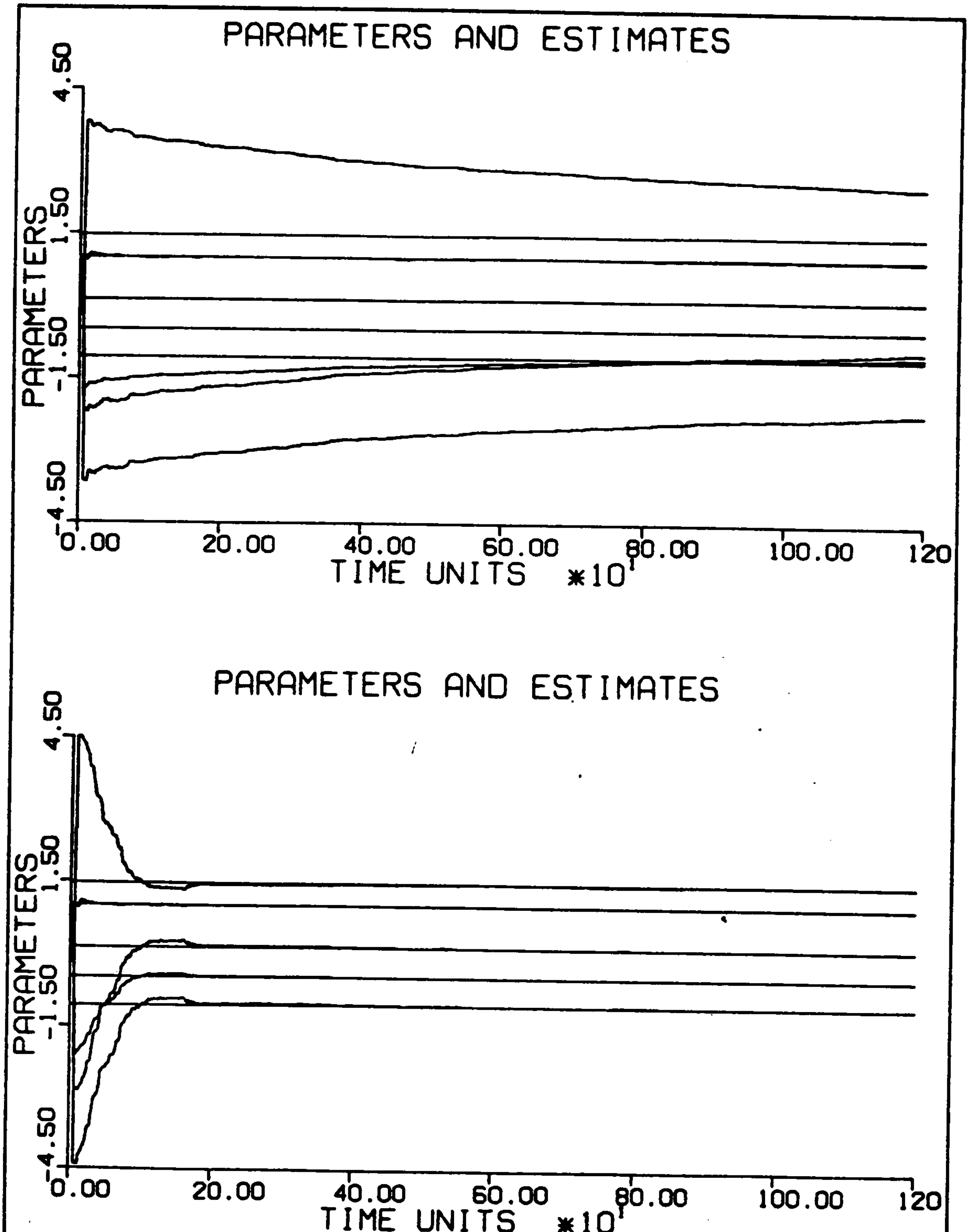

Figure 18. Comparison of convergence rates with initial error about 3.0: EKF converges somewhere just beyond 1200 iterations, while PLID converges in about 200 iterations.

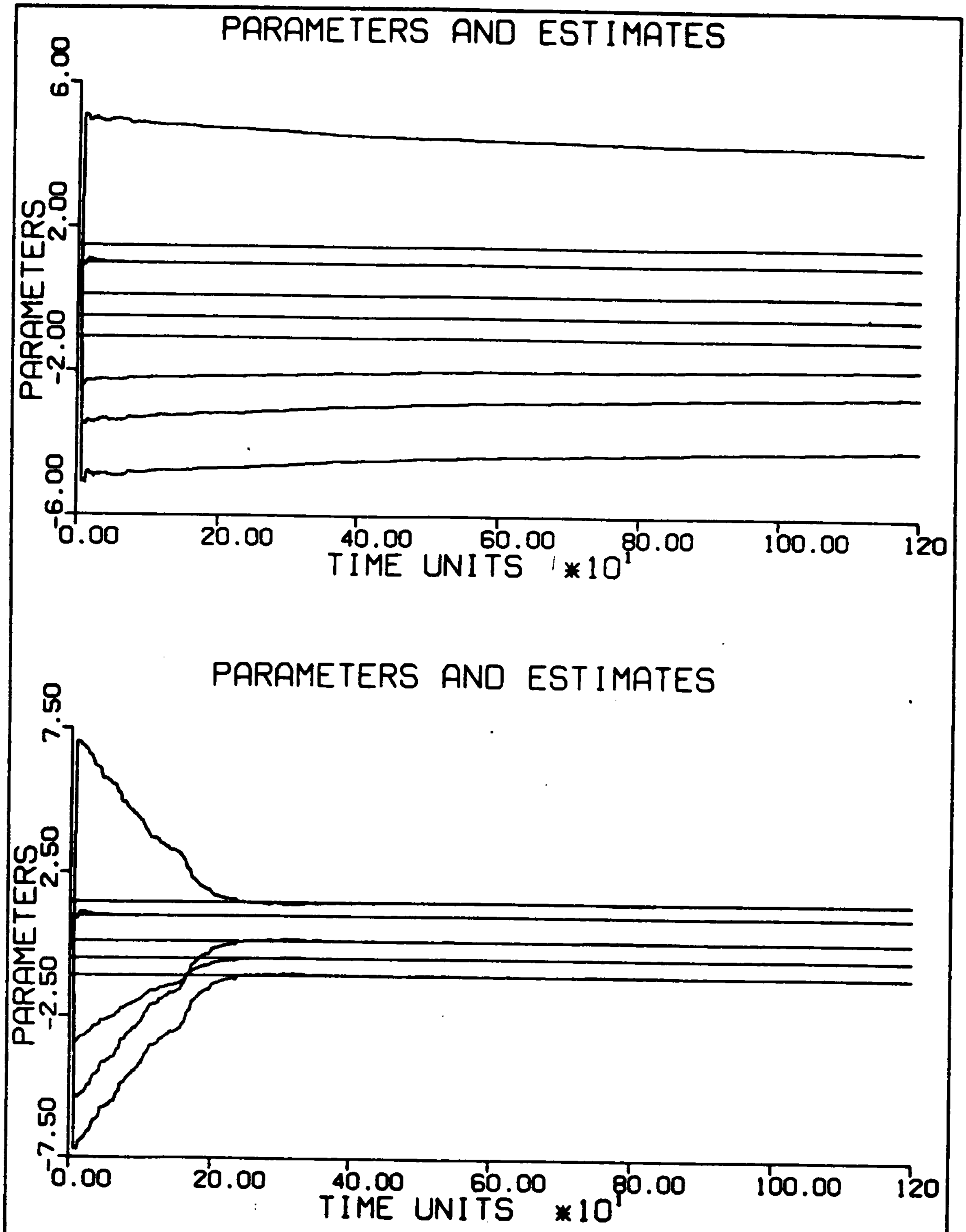

Figure 19. Comparison of convergence rates with initial error about 5.0: EKF convergence rate is very slow, while PLID converges after about 400 iterations.

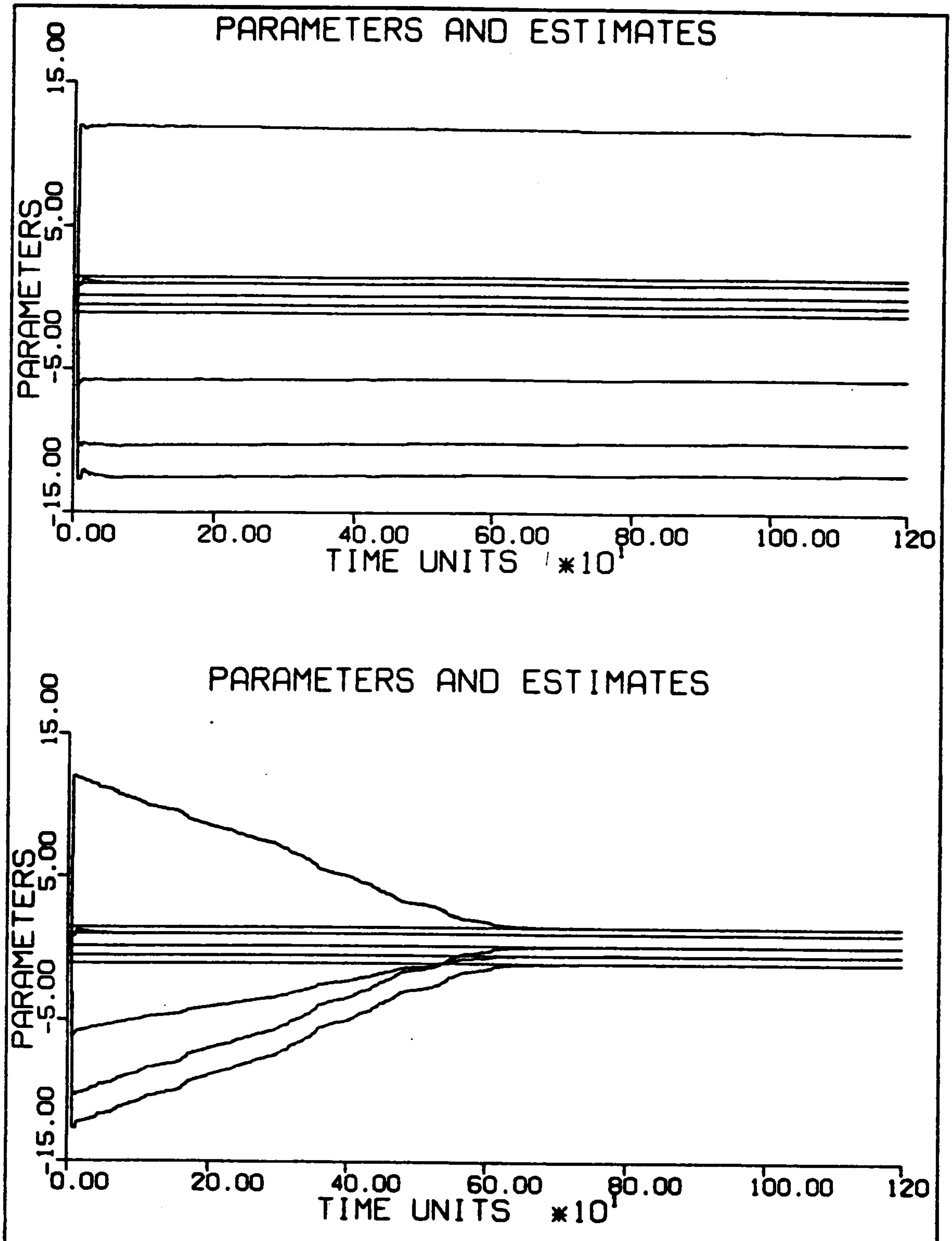

Figure 20. Comparison of convergence rates with initial error about 12.0: EKF convergence is in doubt, while PLID converges after about 700 iterations.

9.3 Simulation of an Unstable System

In this section, results are presented from the simulation of PLID identifying the states and parameters of a single-input, single-output, unstable, non-minimum-phase, second-order system. The theoretical basis of the PLID algorithm indicates it should be able to identify such systems. However, a new set of problems arises in the implementation of PLID in the case of unstable systems.

The problems that arise are all related to numerical ill-conditioning. In an unstable system, bounded inputs can give rise to unbounded outputs, the source of the numerical difficulties. The simplest problem occurs if the system output value exceeds the upper limit of the computer's numerical range, causing an *overflow* condition. Usually, other problems are evident long before overflow occurs.

Typically, when the output is "blowing up," the first problem to occur is numerical ill-conditioning in the error covariance matrix. This can be explained as follows. The error covariance submatrix $P^{\theta\theta}_{k+1|k}$ related to the parameters is non-increasing (strictly decreasing when there is persistent excitation). However, the error covariance submatrix $P^{xx}_{k+1|k}$ related to the states is very nearly linearly dependent upon $P^{\theta\theta}_{k+1|k}$ (modulo some noise covariance terms), where all the coefficients of the linear relation are previous inputs and outputs. Thus, as the outputs become very large, so do the elements of $P^{xx}_{k+1|k}$.

This divergence of the submatrices within $P_{k+1|k}$ is reflected by divergence of the eigenvalues; some of them are becoming very large, while others are tending to zero (or, at best, remaining constant). Hence, the numerical ill-conditioning.

There is also a noticeable effect on the estimate error. As the output becomes very large, it also becomes the dominant force in the determination of the states. Therefore, unless the input also becomes very large, the input has less and less effect on the states as the output blows up. The typical result is that the parameters associated with autoregression, or output feedback, are estimated with very little error. On the other hand, the estimates of the parameters associated with the moving average, or input feedforward, do not converge to the correct values.

However, as the output blows up, the feedforward parameter estimates usually *appear* to converge; that is, they stop changing. The reason is that the bulk of the system excitation is provided by the output feedback, a condition which, although sufficient for the identification of the autoregression parameters, is not sufficient for the identification of the moving average parameters. Thus, the elements of the gain matrix associated with the moving average parameters tend to zero.

It seems that this problem would be avoided in an unstable system with a feedback loop, because then the inputs would become large whenever the outputs do. But that is not necessarily the case, because ill-conditioning can still occur in the error covariance matrix as the output, and now the input, become very large. (In Section 10.1, a two-input, two-output unstable system with an adaptive controller using PLID is simulated. Ill-conditioning is, indeed, avoided, because the controller prevents the output from blowing up.)

The following system transfer function was simulated:

$$
\begin{aligned}
H(z) &= \frac{2\,(z-0.75-j0.75)(z-0.75+j0.75)}{(z-0.90-j0.45)(z-0.90+j0.45)(z-1.01)} \\
&= \frac{2\,z^2-3.00z+2.25}{z^3-2.81z^2+2.8305z-1.022625}
\end{aligned}
\tag{9.3.1}
$$

This strictly proper third-order system has three unstable poles, and two nonminimum-phase zeros. The associated parameters are

$$
\begin{aligned}
a_0 &= 1.022625 & b_0 &= 2.25 \\
a_1 &= -2.8305 & b_1 &= -3.00 \\
a_2 &= 2.81 & b_2 &= 2.00
\end{aligned}
\qquad (9.3.2)
$$

A series of five simulations was carried out, for input signal-to-noise ratios varying from 20 dB to 100 dB. The output noise power was the same as that at the input. State noise was essentially negligible. In each of the graphs of this section, results from all five simulations are displayed simultaneously.

Figure 21 shows the log of the norm of the state estimate error. In each simulation, exponential increase in the estimate error is unavoidable. Notwithstanding this increase, the state estimate actually becomes quite accurate, because the states, themselves, are around five to seven orders of magnitude larger than the error, after 1000 iterations. This happens because after the output becomes large enough, the input has negligible effect on the states compared to the effect of the output feedback. The corresponding effect on the parameter estimates is that the autoregression parameters are always estimated with a high degree of accuracy (see Figure 24), while the moving average parameter estimates only improve until the output begins to blow up (see Figure 27).

Figure 22 shows the log of the norm of the state gain submatrix. As the output blows up, the various noise cases all converge, indicating that the noise becomes negligible. The output noise becomes negligible because its autocovariance remains constant even while the output itself is increasing exponentially. The input noise, and the input itself, become negligible because the system excitation is increasingly dominated by the output feedback. A steady-state, of sorts, occurs. The state estimates do not decouple from the output measurements.

On the other hand, decoupling can be observed in the graphs of the parameter gains, Figure 25 and Figure 28. Decoupling of the autoregression parameter estimates occurs because, as the output becomes increasingly deterministic, there is no new information to be obtained from it. Decoupling of the moving average parameter estimates occurs because excitation due to the input becomes negligible compared to excitation due to the output feedback.

Oscillation of the gain matrices is due to oscillation of the output, caused by the conjugate pair just outside the unit circle.

In Figure 23, the state estimate error covariance submatrix $P^{xx}_{k+1|k}$ is seen to increase exponentially, corresponding to the exponential increases at the outputs. This results because as the outputs increase, $P^{xx}_{k+1|k}$ becomes predominantly just a linear function of the parameter error covariance submatrix $P^{\theta\theta}_{k+1|k}$, where the coefficients of the linear relation are the outputs, themselves. Meanwhile, $P^{\theta\theta}_{k+1|k}$ becomes virtually constant because the system is not being persistently excited. The linear dependence of $P^{xx}_{k+1|k}$ on $P^{\theta\theta}_{k+1|k}$ is described by Equation 6.3.3. The tendency of $P^{\theta\theta}_{k+1|k}$ to become constant can be seen in Figure 26 and Figure 29.

In Figure 24, the norm of the autoregression parameter estimate error is seen to decrease to quite low levels; even the worst case is still less than 10^{-4}. This remarkable accuracy can be explained as follows. In the simulation, the output noise power remained constant even while the output signal increased to extremely high levels. In effect, the output becomes deterministic. At the same time, the input power remained constant, so the system is being driven almost entirely by feedback of a deterministic signal. Thus, the parameters of the feedback paths (*i.e.*, the autoregression parameters) can be determined quite accurately.

Eventually, however, lack of persistent excitation affects the autoregression parameter estimates, because the estimator decouples from the system. Decoupling of the autoregression parameter estimates can be seen in Figure 25, where the gains decrease exponentially. The lack of persistent excitation also affects the associated error covariance matrix, which stops decreasing, as shown in Figure 26. The increase seen in the lowest noise cases must be attributed to numerical ill-conditioning, because it begins to occur about the same time the state estimate error covariance starts to increase exponentially. The ill-conditioning is brought about by having very small elements in the submatrix related to the autoregression parameter estimates, and at the same time very large elements in the submatrix related to the state estimates.

In Figure 27, the norm of the moving average parameter estimate error quickly achieves a good level, because the signal-to-noise ratio at the input is fairly high. This is typical behavior in the PLID algorithm. However, after the initial strong improvement, there is usually some incremental improvement as the input continues to stimulate the system, except when the input signal-to-noise ratio is so bad that convergence is really in doubt.

Such incremental improvement is not seen here, because the input becomes negligible as the output "blows up." That is, the input *effectively* tends to become zero, compared to the output feedback. The result is the decoupling of the parameter estimates from the system, evidenced by the exponential decrease of the associated gains, seen in Figure 28, and by the tendency of the associated error covariance submatrix to become constant (that is, to stop decreasing), seen in Figure 29.

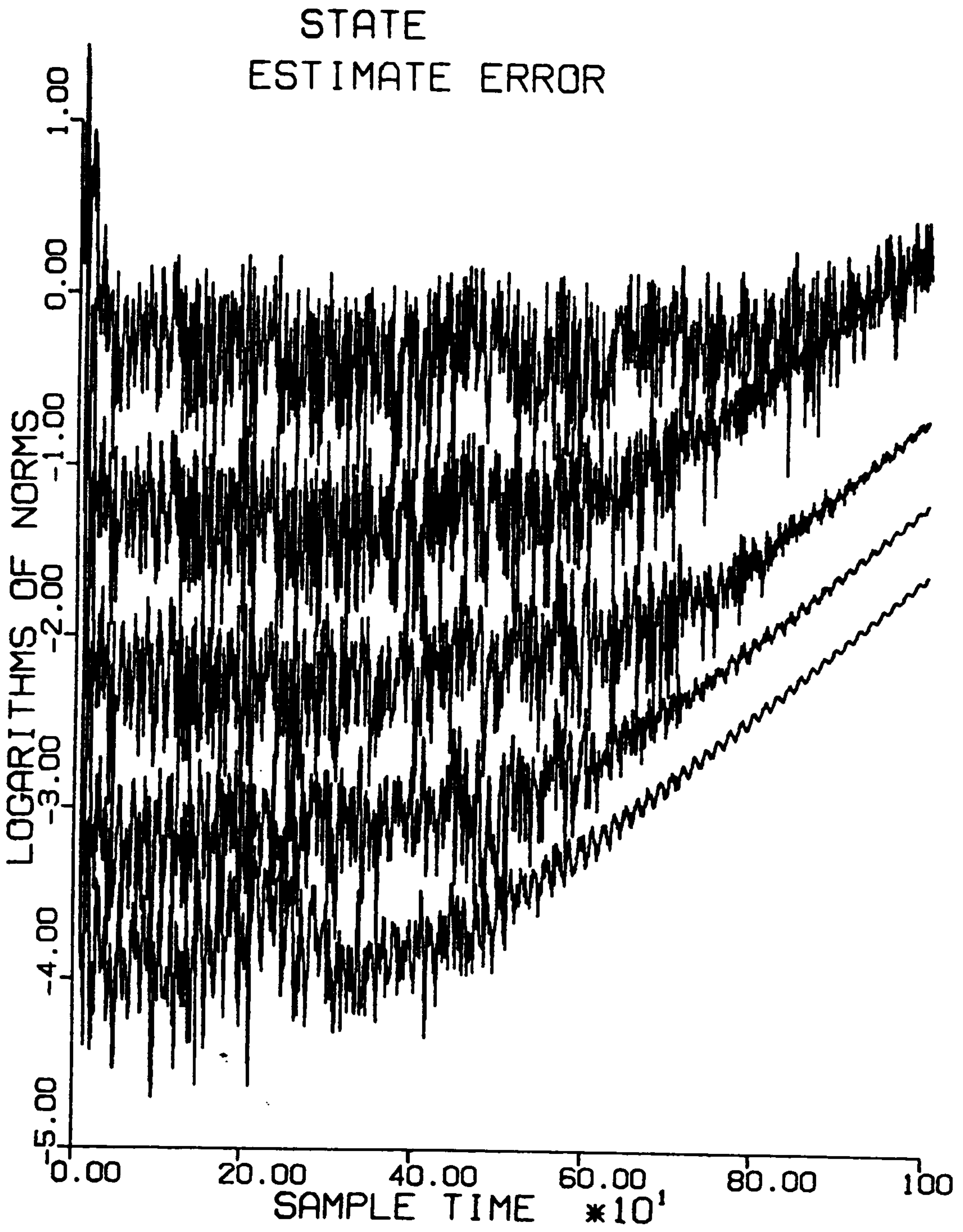

Figure 21. State estimate error for an unstable system: Log of the norm of the state estimate error subvector when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

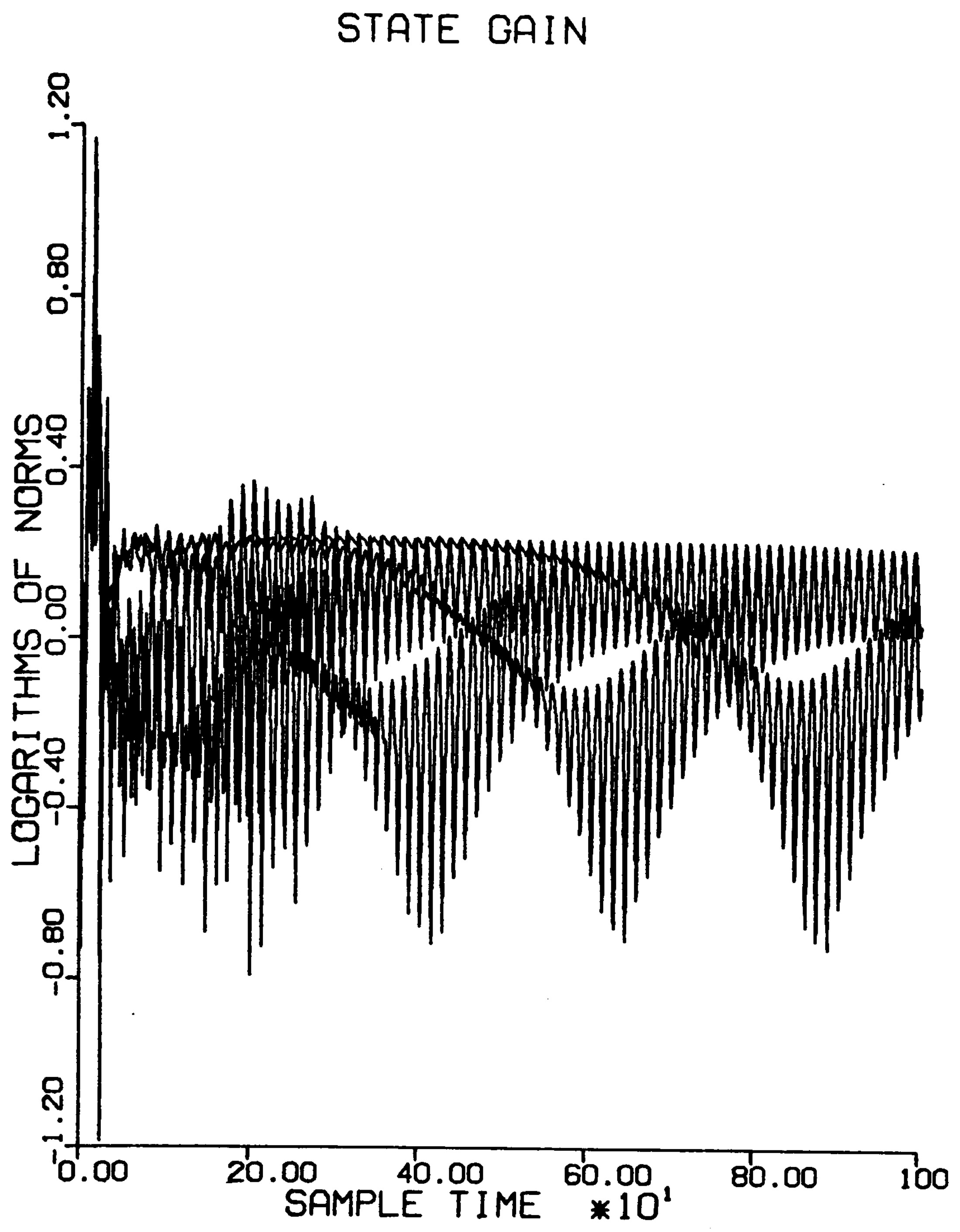

Figure 22. State gain for an unstable system: Log of the norm of the state gain submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

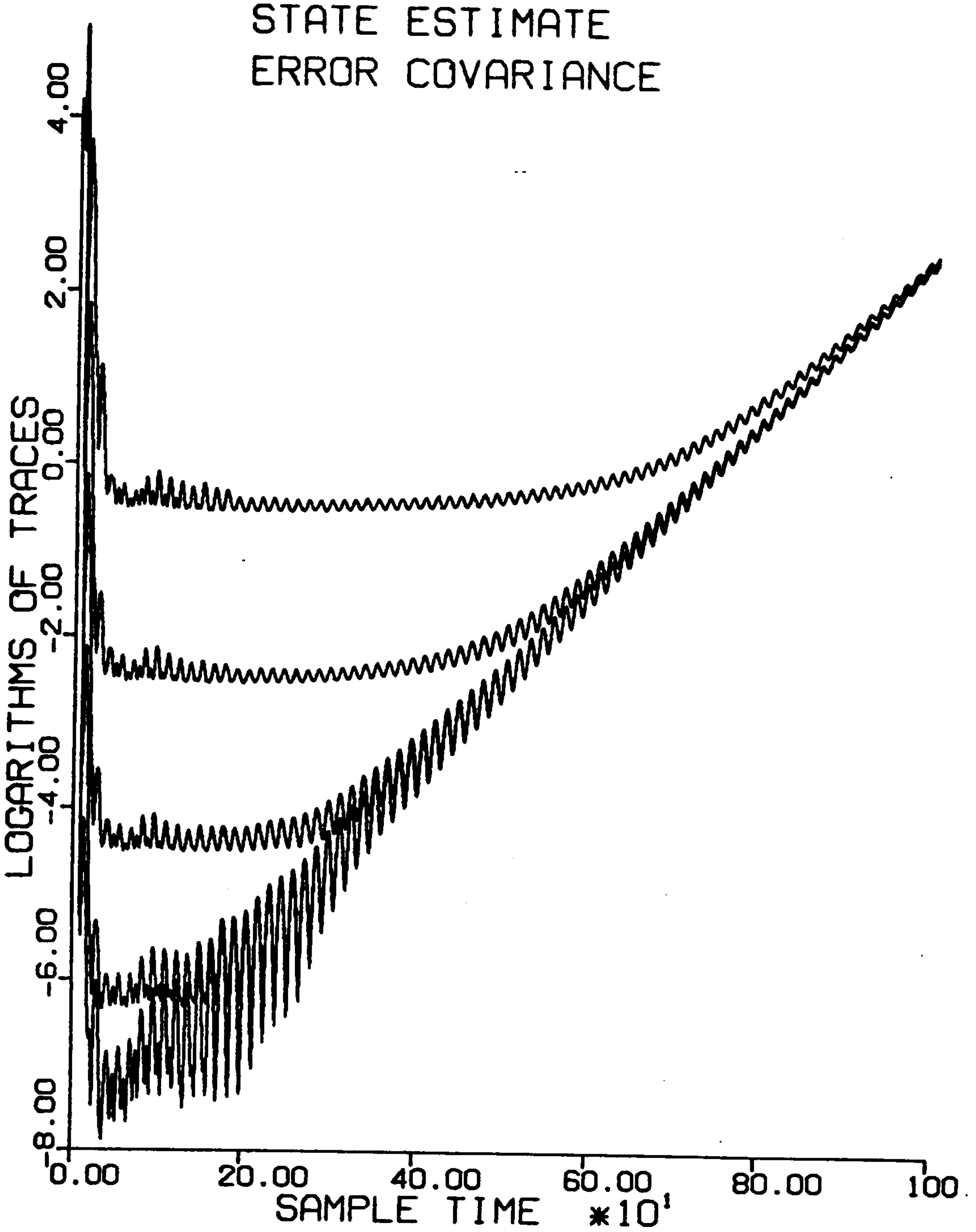

Figure 23. State error covariance for an unstable system: Log of the norm of the state error covariance submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

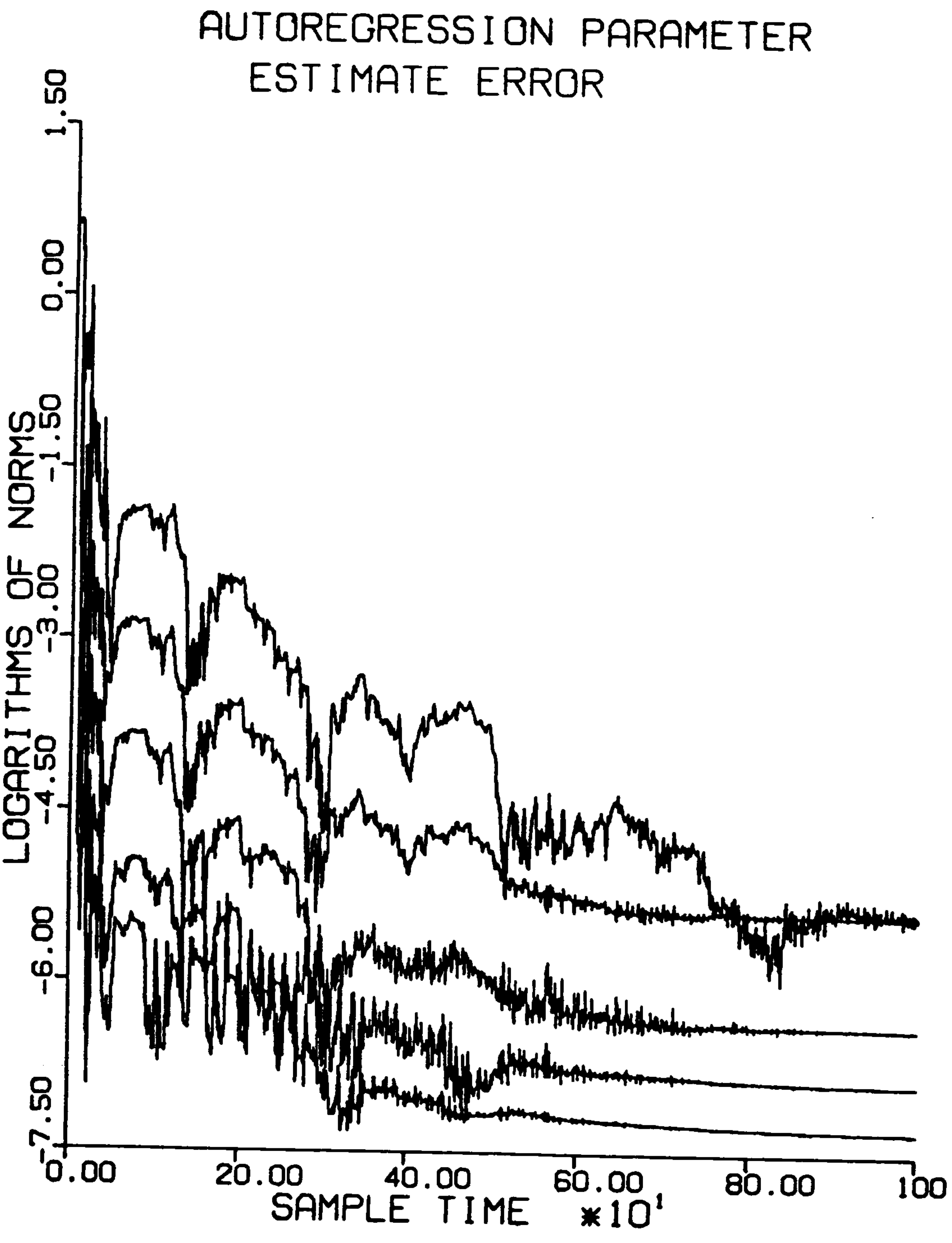

Figure 24. Autoregression-parameter estimate error for an unstable system: Log of the norm of the autoregression parameter estimate error subvector when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

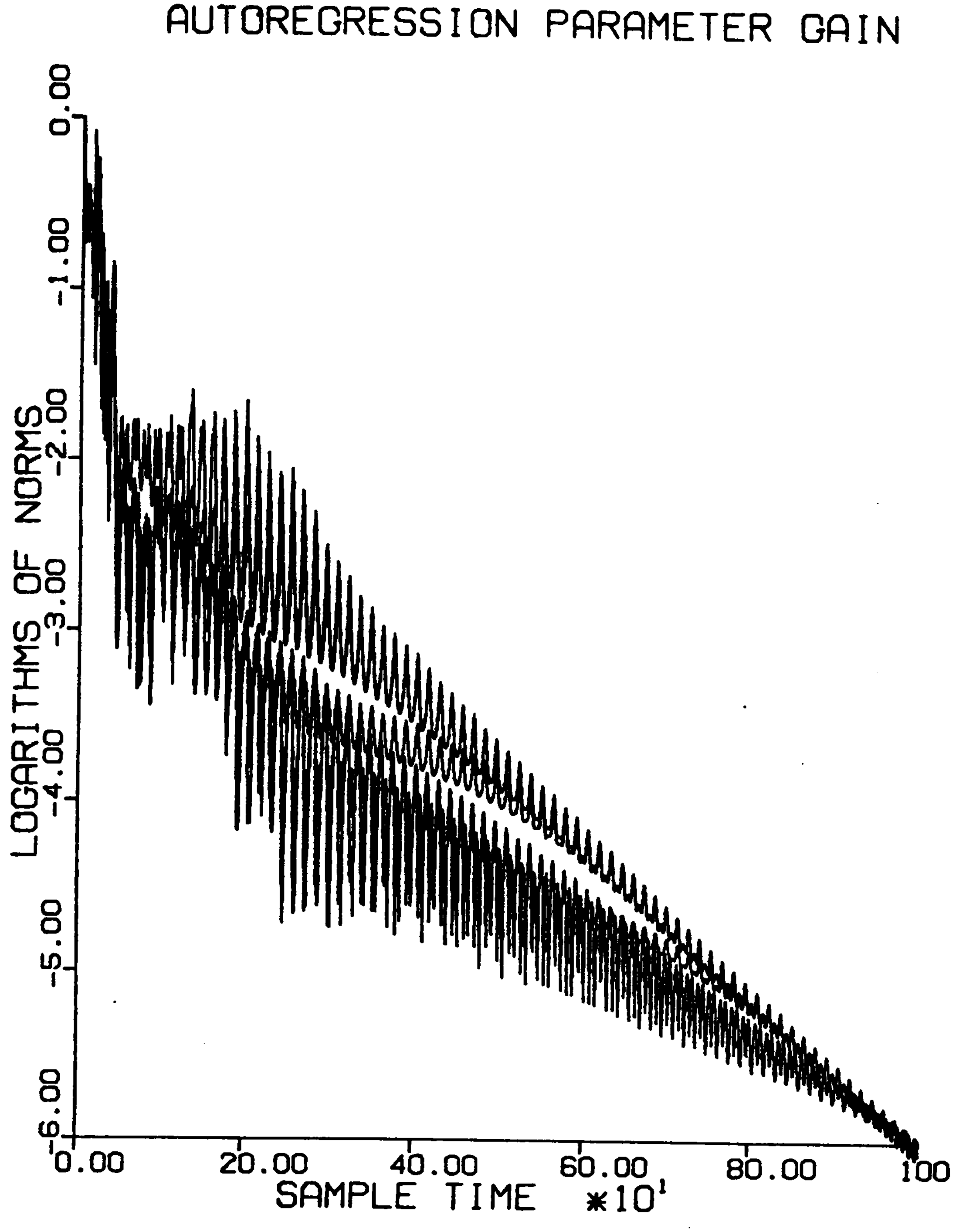

Figure 25. **Autoregression-parameter gains for an unstable system: Log of the norm of the autoregression parameter gain submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.**

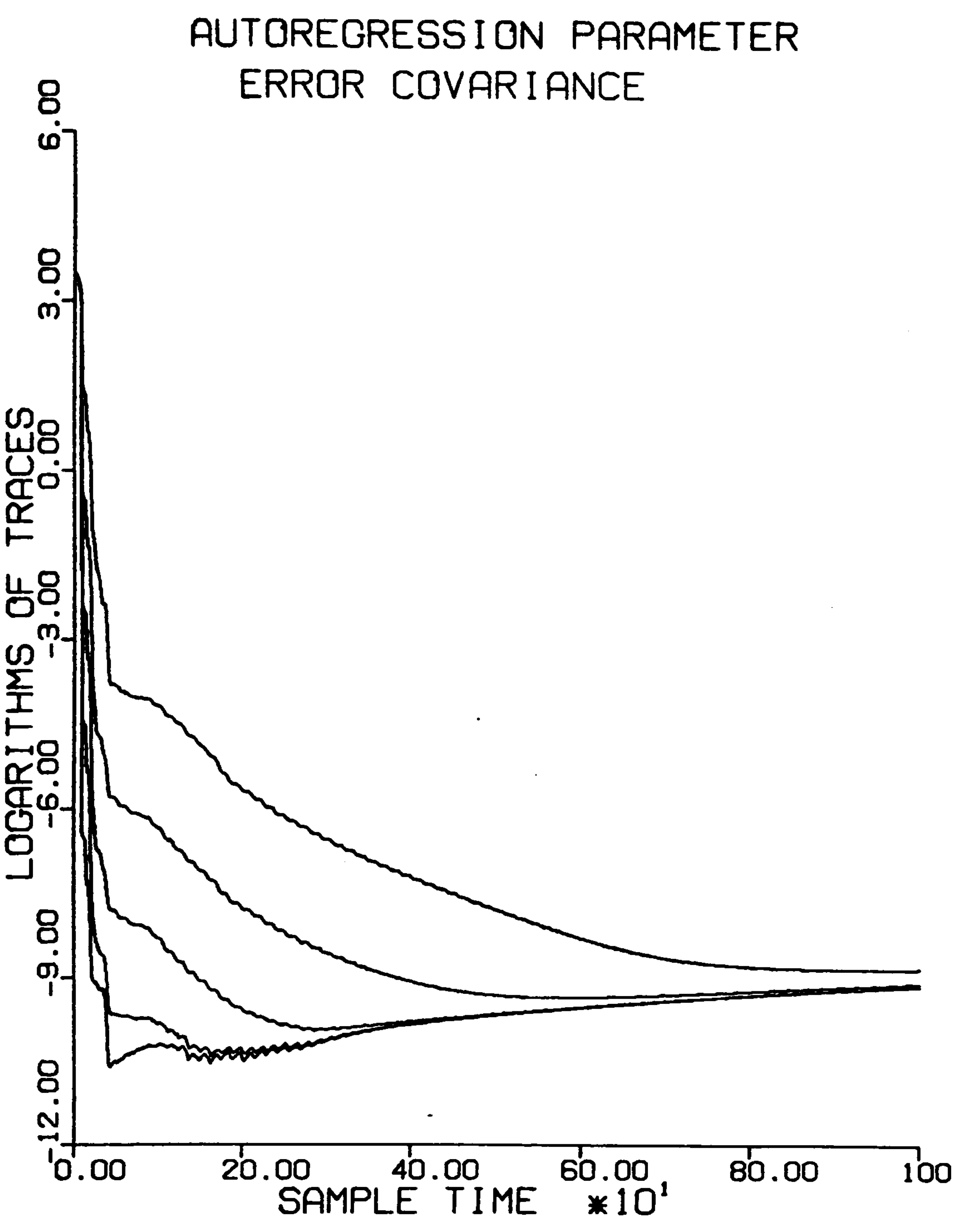

Figure 26. Autoregression-parameter error covariance for an unstable system: Log of the norm of the autoregression parameter error covariance submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

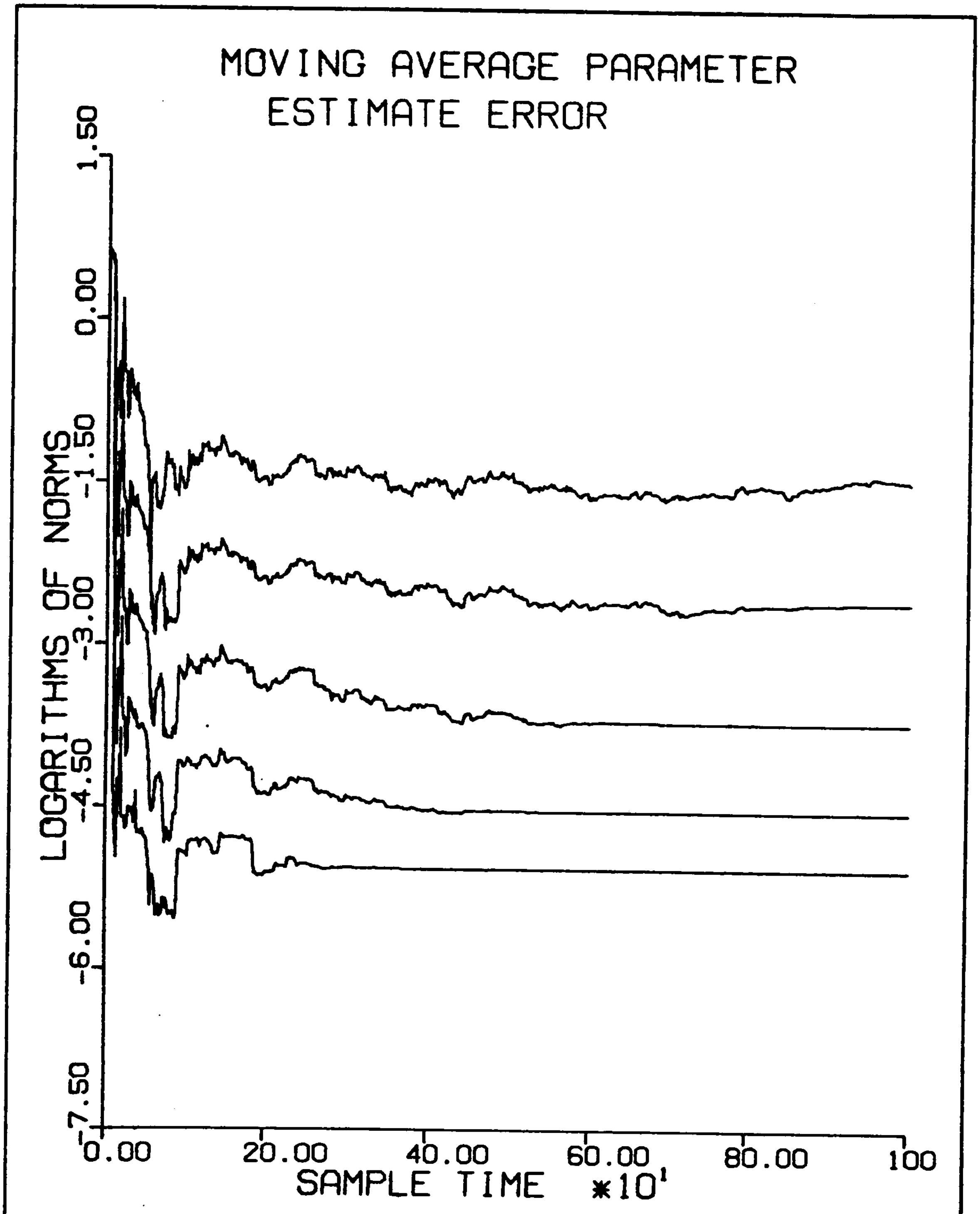

Figure 27. Moving-average-parameter estimate error for an unstable system: Log of the norm of the moving-average-parameter estimate error subvector when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

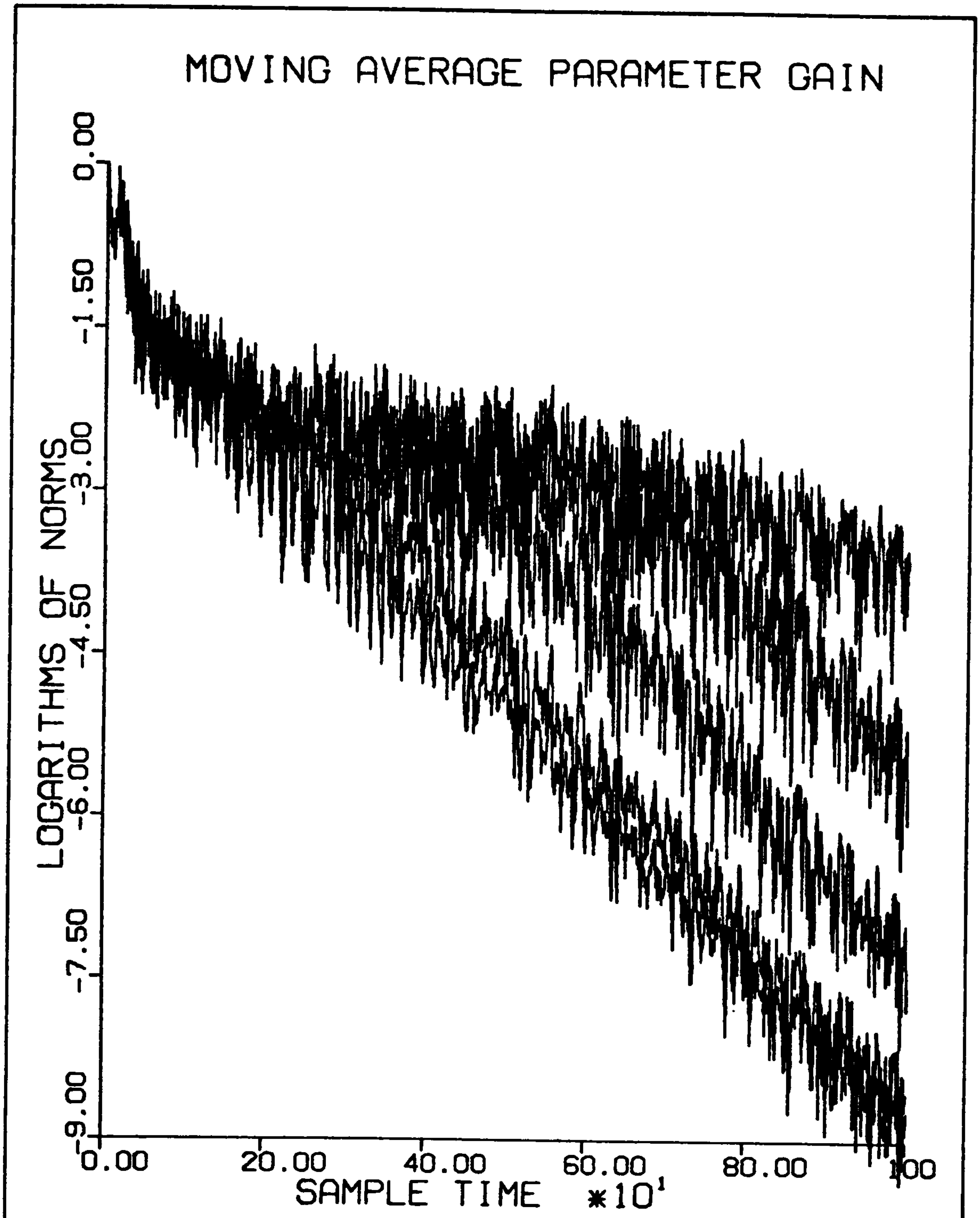

Figure 28. **Moving-average-parameter gain for an unstable system: Log of the norm of the moving-average-parameter gain submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.**

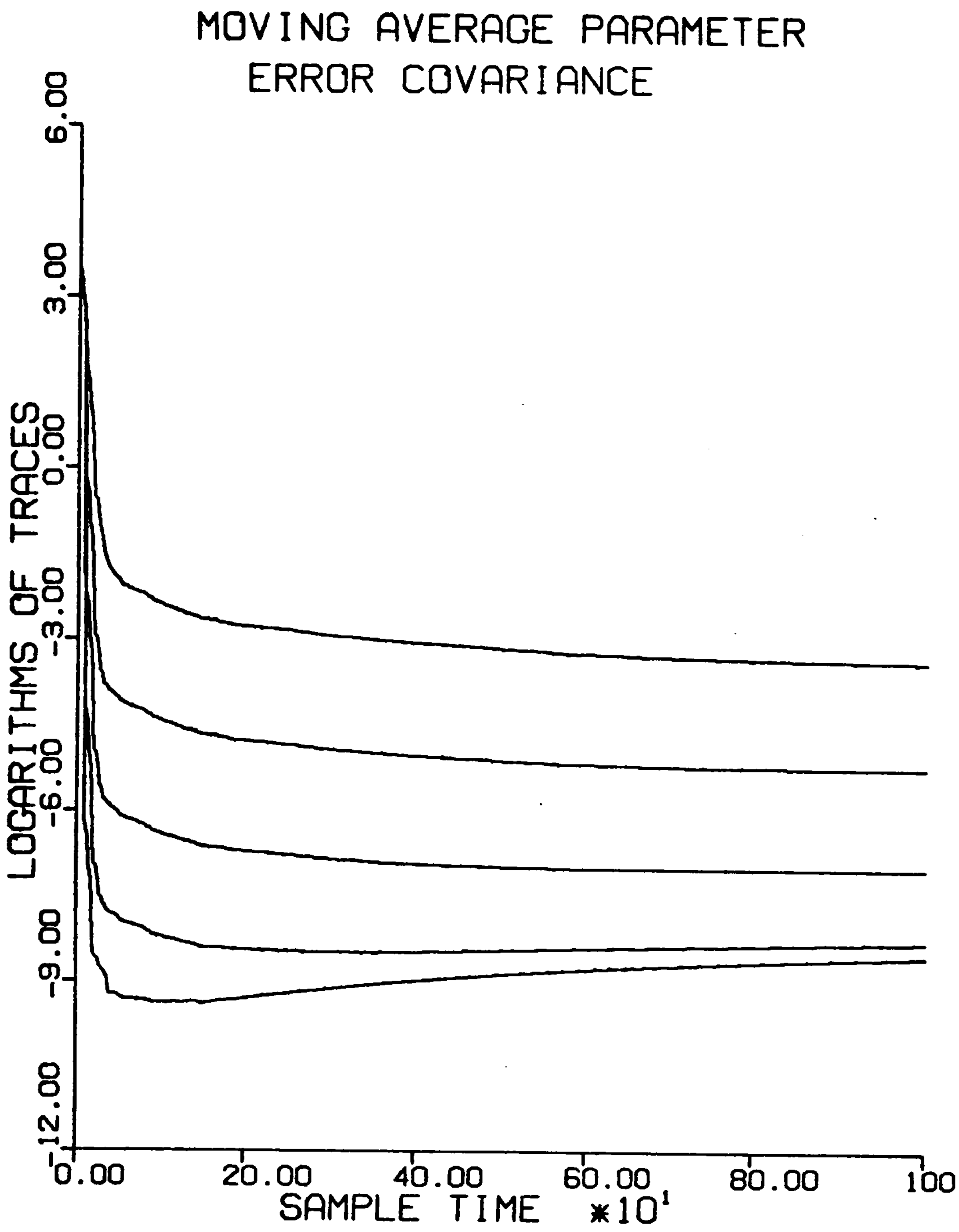

Figure 29. Moving-average-parameter error covariance for an unstable system: Log of the norm of the moving-average-parameter error covariance submatrix when the S/N ratio is 20 dB, 40 dB, 60 dB, 80 dB, and 100 dB.

10.0 PLID in Adaptive Control

10.1 An Overview of Adaptive Control Schemes

In the literature on adaptive control, the only consensus on the *definition* of adaptive control is that there is no consensus. Various authors have defined adaptive control in ways that suit their own purposes. Most agree, however, that constant gain feedback is not adaptive control.

One of the most general characterizations of adaptive control is given by Åström, in which adaptive control is viewed simply as a special type of nonlinear feedback control, and in which "the states of the process can be separated into two categories, which change at different rates. The slowly changing states are viewed as parameters." [33] The concept of an extended state vector, incorporating both states and parameters, fits in very well with Åström's characterization.

A general definition of *adaptation* is give by Tsypkin:

> "By adaptation we will mean the process of changing the parameters structure and possibly the controls of a system on the basis of information obtained during the control period so as to optimize (from one point of view or another) the state of a system when the operating conditions are either incompletely defined initially or changed." [34]

Not all authors agree that control which self-adjusts to an unknown *time-invariant* system is adaptive (some call it initial self-tuning). However, many use precisely this case to show convergence or stability properties of adaptive control schemes that are designed to work on time-varying systems. The extension of the proofs to time-varying systems may rest on the assumption that the system parameters vary slowly compared to the time required for parameter estimation.

Although the concept of adaptive control has been known since sometime in the late 1940's, and notwithstanding the great amount of research in the 1950's aimed at achieving it (mostly *ad hoc* autopilots for high-performance aircraft), meaningful results were not really possible until the 1960's, when state space and stability theory were introduced, and important results in stochastic control theory were obtained. The 1960's also were a period of major development in the area of system identification and parameter estimation. These developments marked the beginning of adaptive control theory.

Adaptive control schemes can be broadly separated into three categories: gain scheduling, model reference adaptive control, and self-tuning regulators. *Gain scheduling* differs from the other categories in that it is, in an adaptive sense, a type of open-loop control. If, for one reason or another, the scheduled gains become unsuitable, there is no way to compensate for the deterioration in system performance. On that basis, some authors exclude it from the class of adaptive control schemes.

While some authors still view gain scheduling as adaptive control, very little theoretical work on it appears in the adaptive control literature. It seems to be regarded as a vestige of the early work (circa 1950's) in the area, although it has been noted that certain proofs that are difficult for the other categories of adaptive control might be forthcoming for gain scheduled systems. Interestingly, the applications of gain scheduling far outnumber those of the other categories; furthermore, they have been quite successful, from an engineering standpoint, in systems where more computation-intensive approaches may fail due to delay (e.g., in flight control systems).

Model reference adaptive control (MRAC) was first presented by Whitaker, Yamron, and Kezer of MIT in 1958. The distinctive feature of this method is the construction of a reference model that operates in parallel with the plant. The reference model produces the "desired output," that is, the output which the controller should induce from the plant.

A typical MRAC system is shown in block diagram form in Figure 30 on page 163 The parameters for the regulator will be determined from a function of the input, the plant output, and the *difference* between the plant output and the reference output. The original parameter adjustment function given by Whitaker, *et.al.*, now known as the MIT rule, has been shown to result in unstable closed-loop systems for certain plants. Various modifications of the MIT rule that were proposed in the mid-1960's and later, overcame the stability problem. MRAC is currently the focus of attention of many workers in adaptive control, and some theoretical results on global convergence and convergence rate have been presented for certain restricted classes of systems.

The self-tuning regulator (STR) is more obviously a type of non-linear feedback controller. The essential feature is an observer/estimator that determines the system model, from which parameters for the controller are computed. Figure 31 on page 164 shows this in block diagram form. Actually, this represents only one type of STR, known as the *explicit* type, since the model of the system is first determined explicitly. In an *implicit* STR, the entire algorithm is re-formulated so that only the regulator parameters are calculated; hence the system model is implicit.

The explicit STR is also called the *indirect method* because the controller parameters are computed by way of the system model parameters. Similarly, implicit STR is also called the *direct method* because the controller parameters are computed directly.

The STR was originally proposed in 1958 by Kalman. In a landmark paper [35], he presented a dedicated computer to implement the scheme in a deterministic system. However, widespread interest in the approach was delayed until the advent of microprocessors and the more complete development of recursive identification methods; that is, until the mid-1970's. Several authors have pointed out essential similarities between MRAC and STR [33].

The STR approach is appealing to many investigators since it does not specify a particular identification/parameter estimation algorithm, nor does it specify any particular type of controller. As a result, many different combinations of estimator and controller have been studied.

A typical approach to adaptive control of *stochastic* systems uses the explicit STR. This allows the utilization of proven estimation techniques in determining the system model parameters. The controller could be, for example, the minimum variance type. The main theoretical problems one addresses in any type of adaptive control scheme are stability, convergence, rate of convergence, and performance of the adaptive system. For the explicit STR, none of these can be addressed until some convergence results are known for the parameter estimation technique that is being used.

The strongest results for stochastic explicit STR's are achieved when the *certainty equivalence principle* holds (or is assumed to hold). By definition, it holds when one can use the parameter estimates as if they were the true parameters for designing the controller. Certainty equivalence is known to hold, for example, in time-invariant linear-quadratic-gaussian control problems [36].

More generally, certainty equivalence is a good assumption when it can be shown that the parameter estimates converge to the true parameter values with probability one. In stochastic systems, this is not a trivial thing to prove. Typically it is assumed that the parameters are nearly time-invariant, since the time-varying stochastic case is so difficult to analyze.

Certainty equivalence is related to the *separation principle*; in fact, the former is a stronger condition than the latter. The separation principle holds if estimating the process parameters can be done separately from determining the controller parameters.

The controller parameters may be functions of the uncertainties in the process parameter estimates, in which case certainty equivalence does not hold.

For any given parameter estimation technique, the validity of the certainty equivalence assumption may be affected by the choice of controller type [37]. As one might suppose, controllers obtained by enforcing the certainty equivalence principle are called *certainty equivalence controllers.* Controllers that are functions of the estimation uncertainty (i.e., in which only the separation principle is assumed) are called *cautious controllers.*

A major factor in the stability and convergence studies of explicit STR's is the need for *persistent excitation* of the unknown plant. Most, if not all, system identification techniques require all modes of the unknown system to be persistently excited to carry out the identification. However, once the feedback loop is closed, the plant input (controller output) is no longer under external control. In fact, it may tend to zero, in which case the identification algorithm cannot detect changes in the plant parameters.

Various approaches have been used to overcome the persistent excitation problem. For example, one adds a small white noise signal to the reference input to avoid reaching steady-state. Many of the theoretical results for adaptive control systems presuppose persistent excitation.

Another aspect of adaptive systems that is under study by many investigators, and that has yielded very few solid results, is the effect of unmodeled dynamics on stability and convergence. This is important, since in real systems it is almost always the case that the model order is lower than the order of the plant (which may in fact have infinite order, such as in distributed systems).

The pseudo-linear identification (PLID) technique of joint state and parameter estimation can be used in an adaptive control scheme for a stochastic linear MIMO system. An adaptive control system using PLID would, of necessity, be an explicit STR.

Furthermore, the controller should require state (observer) feedback; otherwise it is pointless to use PLID since its joint state and parameter estimation is more "expensive" than parameter estimation, alone. If the system is linear, time-invariant, and Gaussian, then the PLID algorithm is optimal and the parameter estimates converge w.p.1 to the true parameter values. Thus, the certainty equivalence principle can be invoked.

An appropriate controller type to use in conjunction with PLID is, for example, an optimal controller. The cost functional could be set up to minimize the control effort, to minimize the variance of the output, or to minimize some linear combination of the control effort and the output variance.

For the class of systems in which it is optimal, the use of PLID in an adaptive control system represents a significant improvement over other methods of joint state and parameter estimation described in the literature. Other methods involve explicitly nonlinear forms, and hence require more difficult implementations, such as the extended Kalman filter [36].

To obtain certainty equivalence results, it is necessary to assume that the unknown plant is time invariant. For the time-varying case, the most general result for PLID would likely only support the separation principle. However, a typical "slowly varying system" assumption could allow enforcement of certainty equivalence.

Typically, a Kalman filter is robust in the presence of improper noise statistics, provided that the noise power is "overestimated" [38]. Of course, it is no longer the optimal linear filter if the noise statistics are incorrect. Some results on robustness in the face of mismodeling and improper noise statistics and *a priori* distribution of the initial state are available in [39] and [40], and appear to have some applicability for the PLID algorithm.

The pole-placement strategy is well known from deterministic control theory. The minimum variance controller strives to reduce the variance of the output error, $y(k)$ -

$r(k)$, where $r(k)$ is a reference input (usually a set point). There is a strong relationship to MRAC in the latter scheme.

Stochastic optimal control minimizes a quadratic cost functional. It is usually only practicable for the linear quadratic gaussian (LQG) case; if the system is not LQG this control approach becomes suboptimal, but may still yield good results [41].

Major problems to deal with in setting up the adaptive system are the same ones that other investigators have faced, namely, the problems of persistent excitation, unmodeled dynamics, and incompletely known noise statistics. Obtaining theoretical results for the entire system is made more difficult by these problems, but is difficult even without them due to the nonlinear stochastic nature of the system.

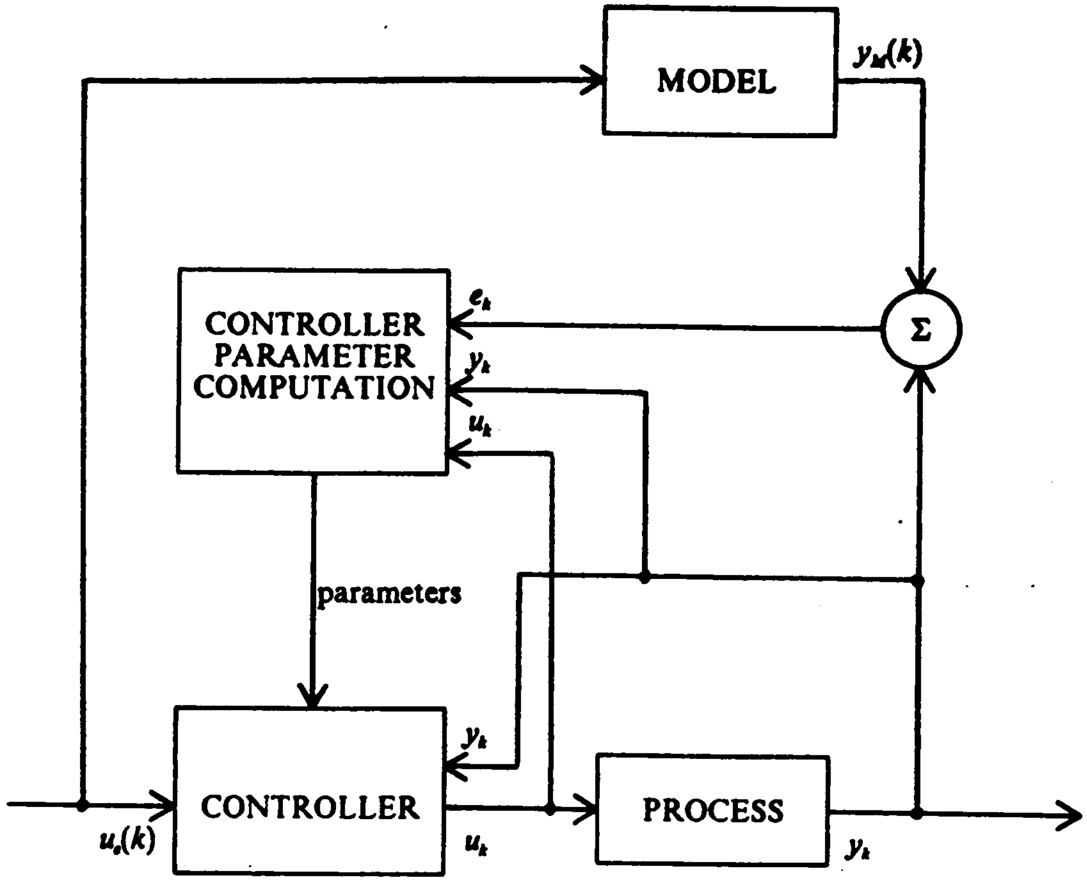

Figure 30. A typical Model Reference Adaptive Control configuration.

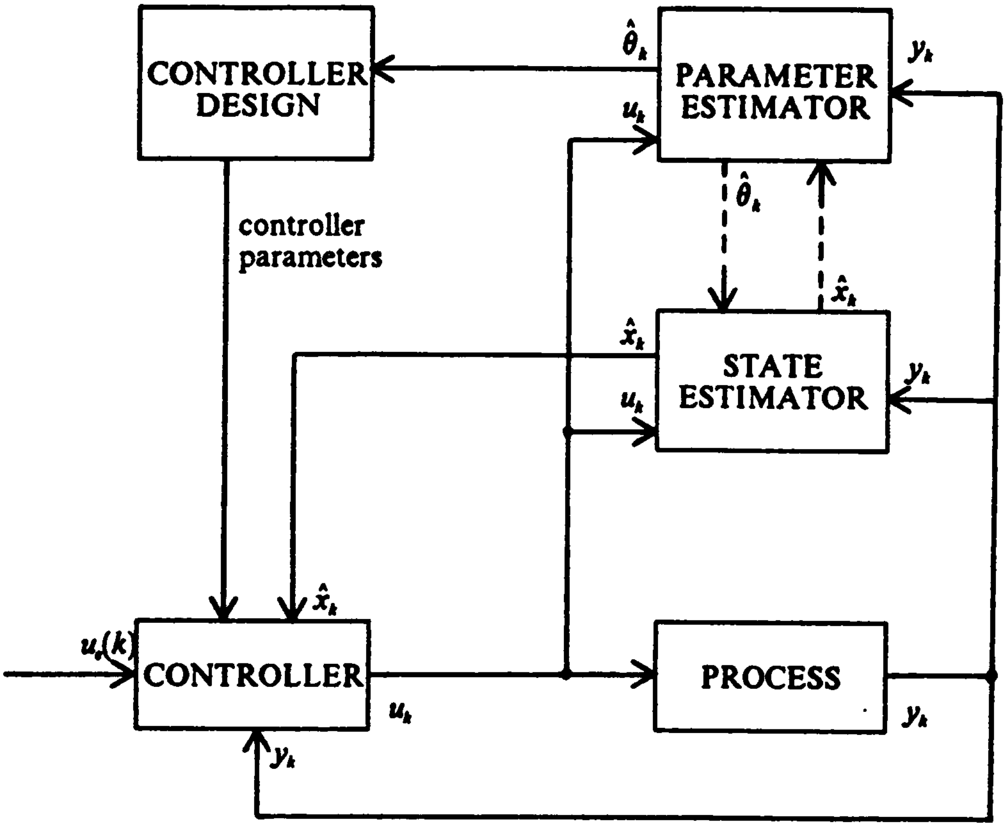

Figure 31. A typical Self-Tuning Regulator configuration.

10.2 Adaptive Control Simulation

The application of PLID in an adaptive control scheme is of considerable practical interest. PLID represents a distinct improvement over algorithms available in the past for estimating the states and parameters of linear stochastic MIMO systems with noisy input and output vectors, yet previous estimators have been successfully integrated into adaptive control systems.

Therefore, it is appropriate to include a section detailing the results of simulations of an adaptive control system. The type of adaptive controller that requires parameter estimates is generally known as a self-tuning regulator. To motivate the incorporation of PLID, the controller should also require state estimates, which means it should be a pole-placement algorithm, or some type of optimal controller.

In this section, simulations of an adaptive optimal self-tuning regulator are presented. The controller seeks to minimize a cost functional that includes both the modulus of the control effort and the variance of the output from the set point being specified.

The system chosen for the simulation is a two-input and two-output system with the following (ostensibly unknown) transfer function, in pole/zero form:

$$\begin{bmatrix} Y_1(z) \\ Y_2(z) \end{bmatrix} = \frac{\begin{array}{c} \left[\dfrac{-(z-2.79863)(z+1.02111)(z+0.12248)}{(z+1.604)(z-0.352-j0.341)(z-0.352+j0.341)} \right| \\ \left. \dfrac{(z-1.81266)(z+1.23381)(z+0.17885)}{-(z-0.509)(z-0.045-j0.928)(z-0.045+j0.928)} \right] \end{array}}{(z-1.05625\mathit{MATHF}0 \,{*}(z+1.0560)(z-0.5+j0.596)(z-0.5+j0.596)} \begin{bmatrix} U_1(z) \\ U_2(z) \end{bmatrix}$$

The corresponding parameters are

$$\begin{array}{llll}
a_{1,1}^{0} = -0.25 & a_{2,1}^{0} = 1.00 & b_{1,1}^{0} = 0.70 & b_{2,1}^{0} = 0.80 \\
a_{1,1}^{1} = 0.50 & a_{2,1}^{1} = 0.70 & b_{1,1}^{1} = -1.00 & b_{2,1}^{1} = 1.00 \\
a_{1,2}^{0} = 0.30 & a_{2,2}^{0} = 1.50 & b_{1,2}^{0} = 0.70 & b_{2,2}^{0} = 0.80 \\
a_{1,2}^{1} = -0.70 & a_{2,2}^{1} = 0.50 & b_{1,2}^{1} = 1.00 & b_{2,2}^{1} = -1.00
\end{array}$$

Note that two of the four system poles are unstable, and three of the four subsystems have non-minimum-phase zeros. From the theoretical point of view, this ought to present no problem to the PLID algorithm. However, as shown in Section 9.3, there are some computational practicalities that can interfere with the convergence of PLID parameter estimates when the output becomes very large. But if a controller prevents the output from becoming large, then it would seem that the problems associated with large output would be avoided. That is one of the things the simulation of this section is intended to test.

The simulation was run with a dither signal having covariance of 0.1 at each input; the input noise was 20 dB below that, having covariance of 0.001. At each output, unity covariance noise was added; initially, the noise is of the same order as the states, themselves. The reference input signal was initially set to zero, so the system was being driven only by the dither signal.

The system was allowed to run uncontrolled for about 120 iterations, by which time the outputs were beginning to "blow up." Then the PLID parameter estimates were used to construct a controller. The controller used in the simulation is a set-point controller, which inherits much of its solution from the linear quadratic regulator solution. (This type of controller is described in detail in VanLandingham [42].) It requires both parameter estimates and state estimates, making PLID a suitable choice as the companion estimator. The set point controller is intended to drive the outputs to the same levels as the corresponding set points, which, at this point, are still zero.

The controller gains were recomputed based on the most recent parameter estimates at iterations 130, 150, 175, and 200, and every 100 iterations thereafter. The set point was left at zero until time 650, when a series of changes was begun. The two input set points were changed, alternately, every 50 iterations, varying within a range of -200 to 200.

The data most obviously of interest are a comparison of the input set points to the actual outputs. These are given in Figure 32 for the first 1000 iterations, and in Figure 33, which presents a closer look at the last 400 iterations, when the set point is being changed from zero. In Figure 32, the instabililties in the system are quite apparent, as the outputs begin to blow up around iteration 100. When the controller is switched on at iteration 120, there is a rather dramatic effect as the outputs are brought in line with the set points, which are both at zero until iteration 650.

Note that the inputs are not decoupled in the system, so changing one set point tends to change the output of *both* subsystems. By design, the controller acts as a decoupler of the inputs, but it cannot react instantaneously to changes in the input. The non-minimum-phase zeros are evident from inverted reactions at the output.

Some quite interesting aspects of PLID estimator performance can be observed in this simulation example. These are revealed in the remaining figures, which present norms of the estimate errors, the gains and the error covariances. During the period before the controller is switched on, the problems PLID has with unstable systems that are blowing up start to occur, just as they did in the graphs of Section 9.3. However, after the controller is switched on, those problems disappear, and the effects of system excitation on the PLID algorithm are displayed quite clearly.

Figure 34 and Figure 35 present norms of the state estimate error and the parameter estimate error, respectively, on logarithmic scale. In Figure 34, the state estimate error begins to blow up, just as it does in Figure 21 on page 147 However, after the controller is turned on, the norm of the state estimate error quickly falls to, and stays within, a region around 10^0.

In Figure 35, the change is just as startling. When the controller is turned on, the norm of the parameter estimate estimate error drops almost two orders of magnitude. The cause of this is the large control signal being applied, which provides a great deal of excitation to the system, and simultaneously improves the signal-to-noise ratio at the input, which had been languishing around 20 dB above the noise level. But interesting things continue to happen.

After achieving such a good parameter estimate, between time 200 and time 650, no improvement occurs, because the system is not being sufficiently excited. During that period, the set point is zero, and the dither signal is still only 20 dB above the input noise level. However, at time 650 the set point begins a series of changes, exciting the system and improving the input signal-to-noise ratio as before. Consequently, the parameter estimate error decreases again after time 650, in a series of improvements roughly corresponding to the step changes in the set point.

Figure 36 and Figure 37 present the norms of the state estimator gains and the parameter estimator gains, respectively. In Figure 36, the state gain matrix initially tends to increase, until the controller is switched on. Then it quickly achieves a steady-state condition. The history of the parameter gains, shown in Figure 37, is much more interesting. Initially, as is typical, there is a period of slow decrease, starting at a fairly high level.

However, during the period of weak excitation following the startup of the controller, the gains drop to very low levels, indicating that there is very little new information available with which to alter the current parameter estimates. The series of spikes in the gain levels, starting at time 650, correspond to those periods of transition just after each set point change. The algorithm "wakes up," gathering as much information as possible from these periods of strong excitation.

Figure 38 and Figure 39, which present the traces of the state error covariance submatrix and the parameter error covariance submatrix, correlate highly with the previous two figures. In Figure 38, initial exponential increase is followed by steady-state, once the controller is switched on. When the output levels change from zero to magnitudes of around 100 (starting at time 650), the state error covariance increases slightly. In Figure 39, the strongest decreases in the norm, which by definition is non-increasing, are seen to occur during times of strong excitation, and correspond to decreases in the parameter estimate error.

There are two main points in this section. First, the PLID algorithm can be used in adaptive control loops, even in cases where the system is really "nasty," as it was here. Second, the PLID algorithm requires persistent excitation to converge.

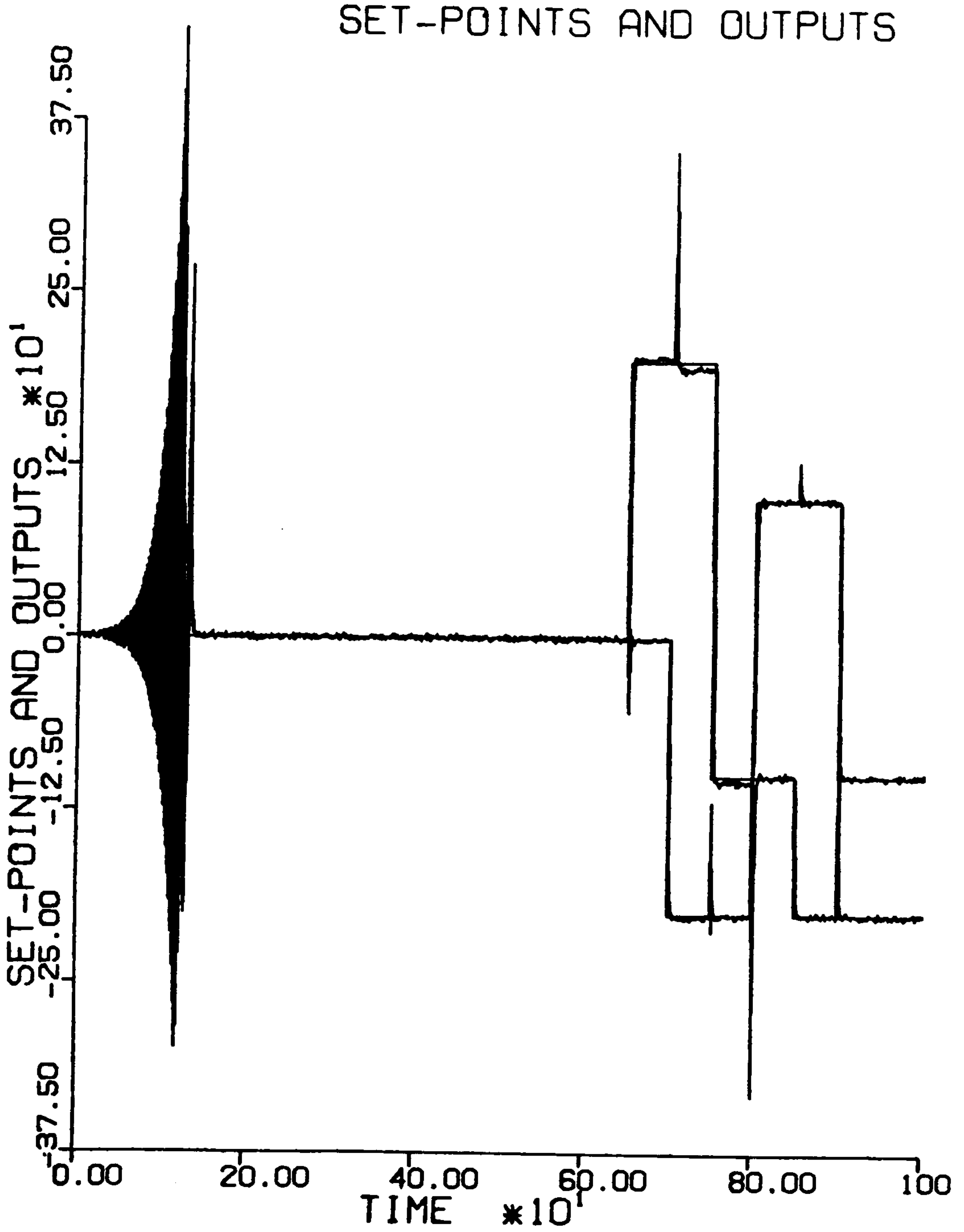

Figure 32. A comparison of set points and outputs under adaptive control: The first 120 iterations are uncontrolled; the set point is zero for the first 650 iterations, then is varied between -200 and 200.

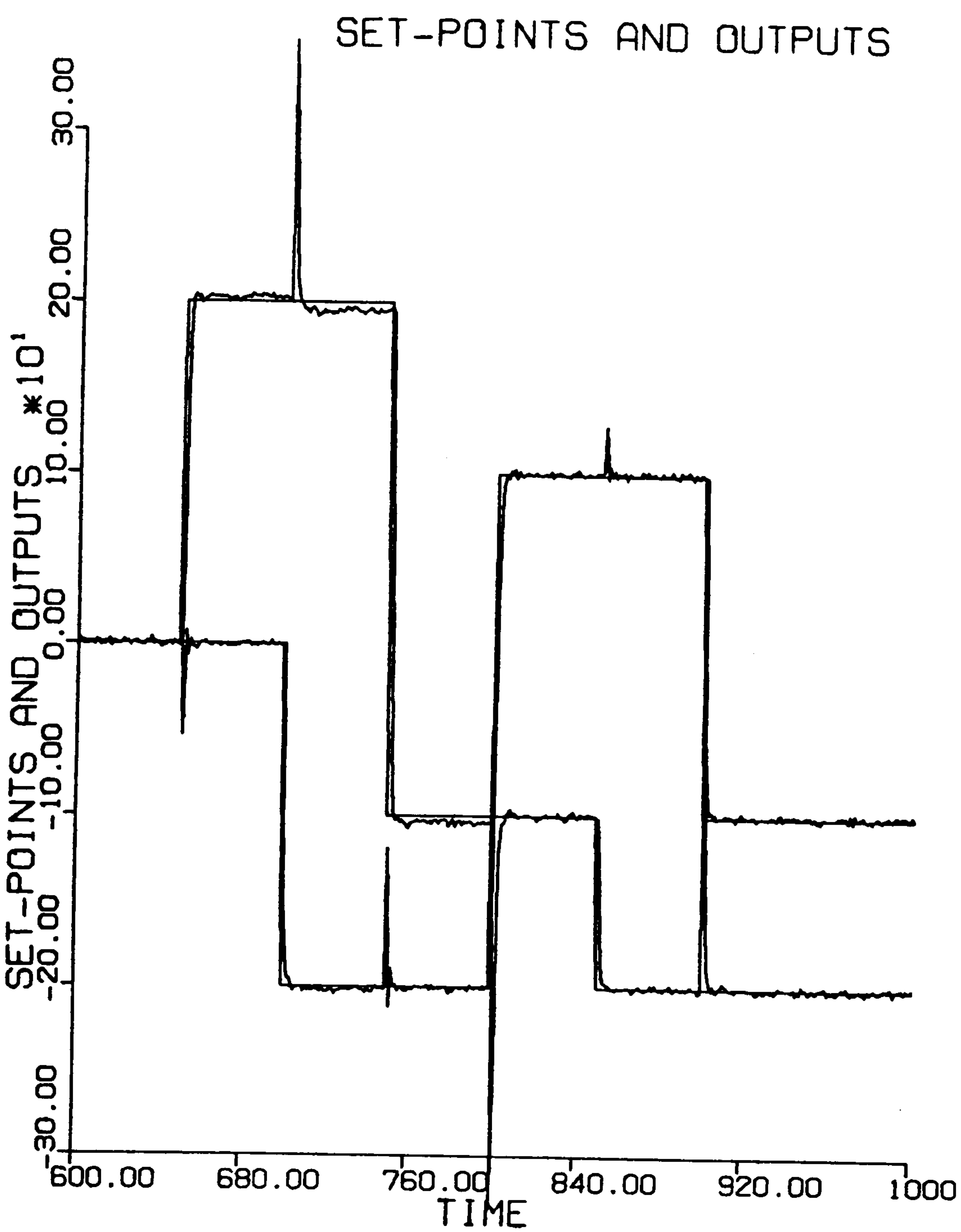

Figure 33. A comparison of set points and outputs under adaptive control: A closer view of the last 400 iterations, during which the set point is varied between -200 and 200. The effect of input coupling is evident.

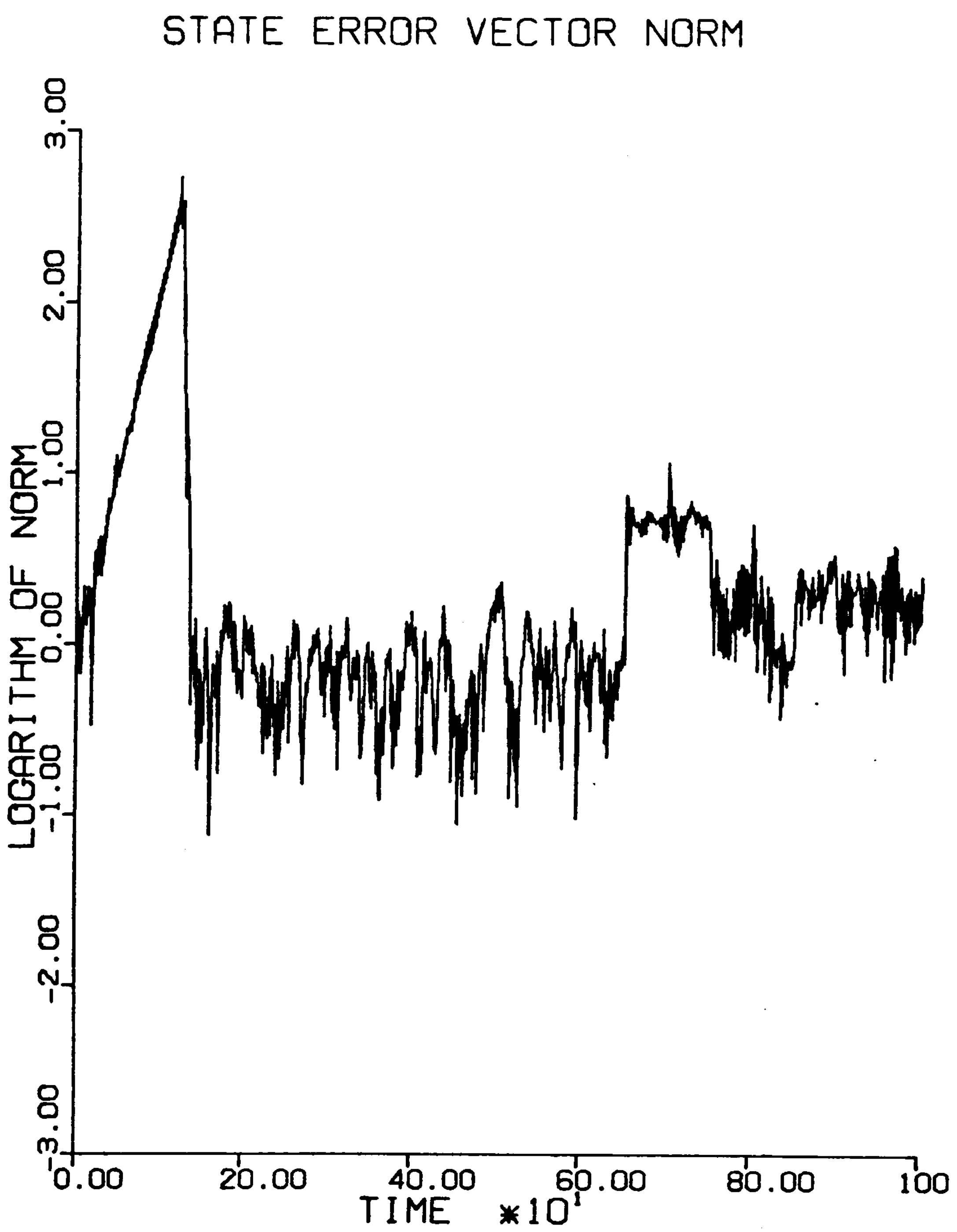

Figure 34. The state error vector, with adaptive control: Estimate error starts to run away, typical of unstable systems, until the controller is turned on at time 120.

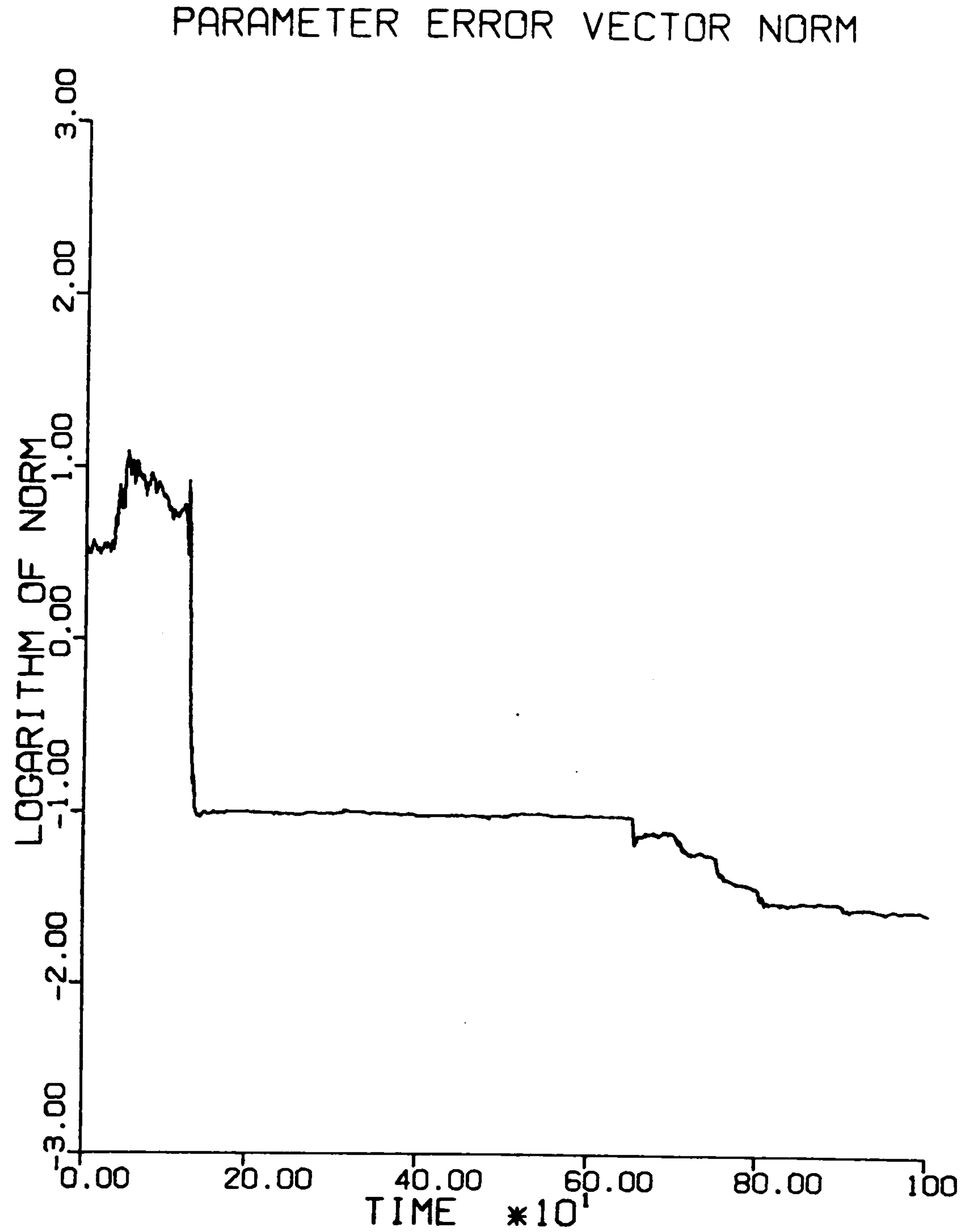

Figure 35. The parameter error vector, with adaptive control: Parameter error remains high until the controller creates large signal-to-noise ratios. During periods of set point stability, error creeps in.

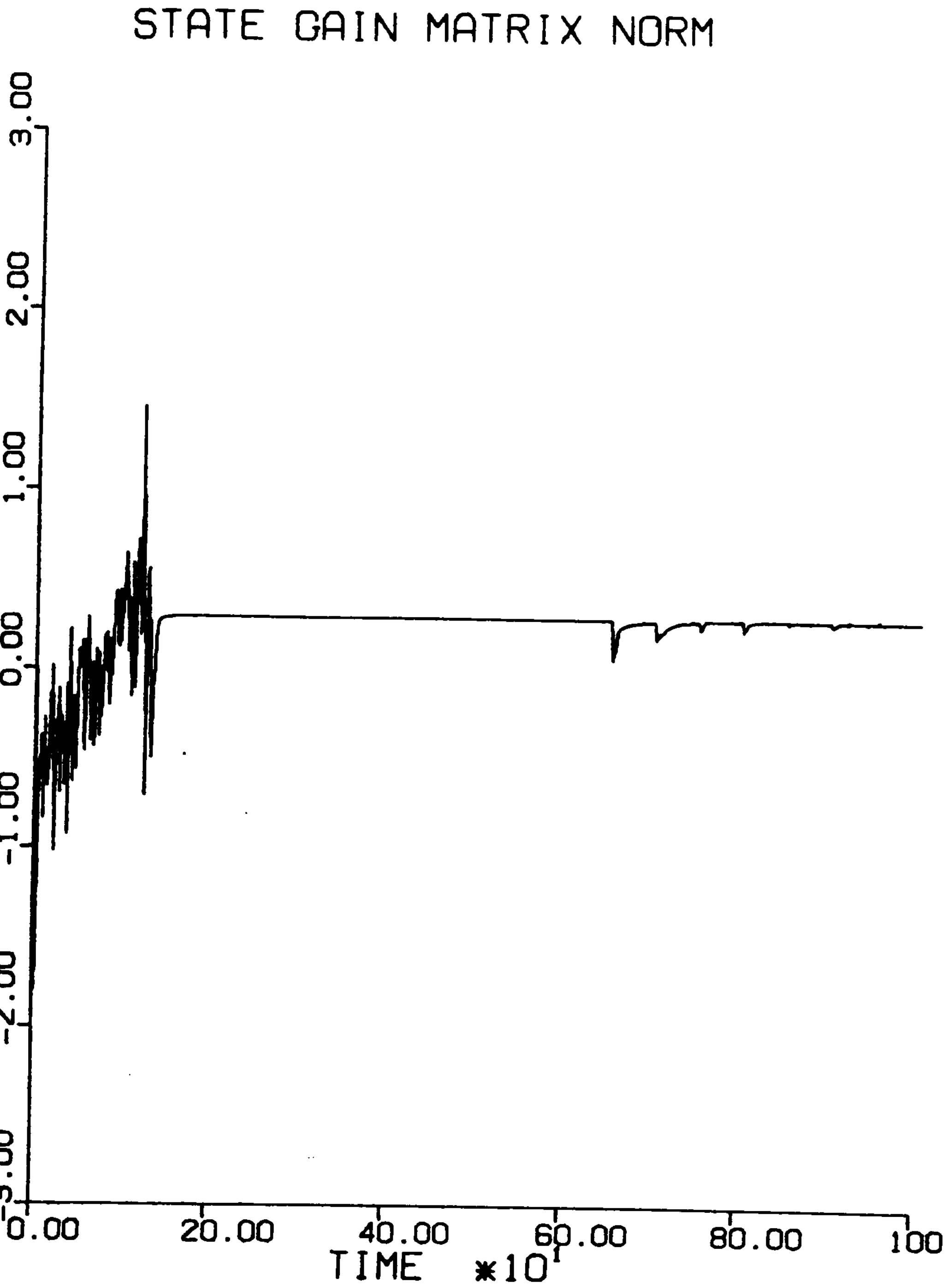

Figure 36. **The norm of the state gain matrix, with adaptive control:** State gain starts to "blow up" as the states, themselves, do. After the controller is turned on, a steady-state of sorts ensues.

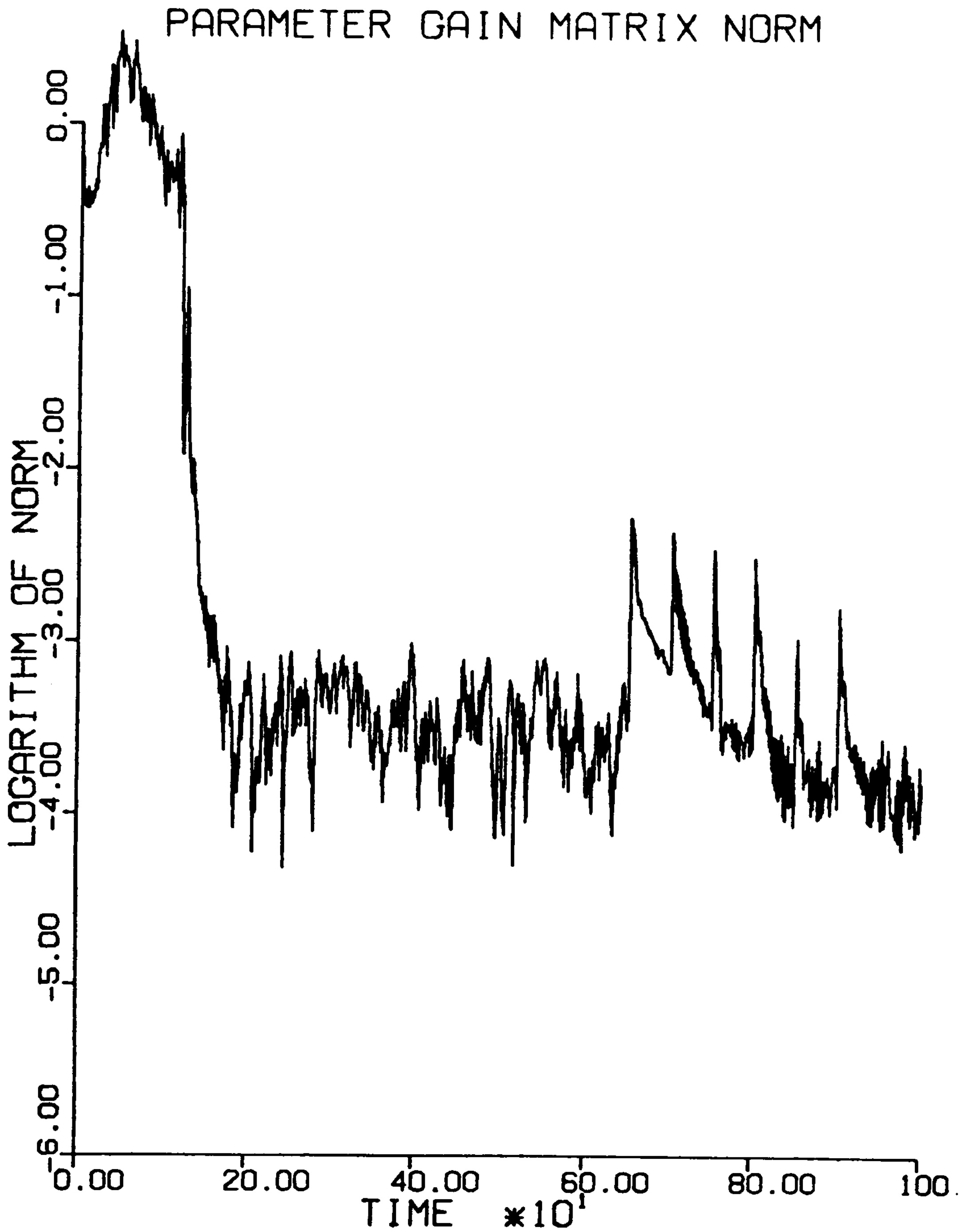

Figure 37. The parameter gain matrix, with adaptive control: A slow decrease is evident before the controller is turned on. Under control, the gains become large only during periods of change (excitation).

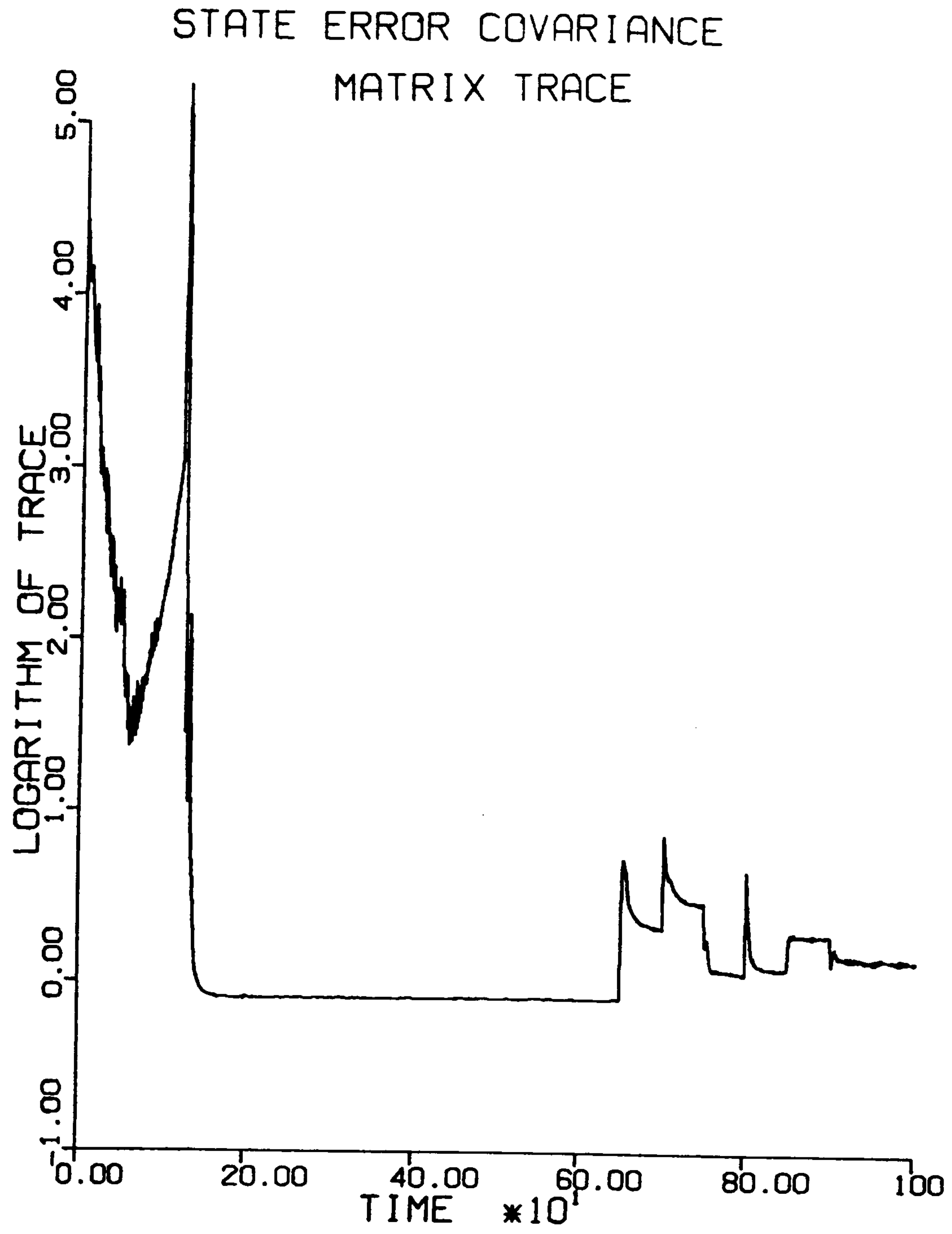

Figure 38. The state error covariance matrix, with adaptive control: Runaway increase begins before the controller is turned on. Under control, a steady-state is achieved.

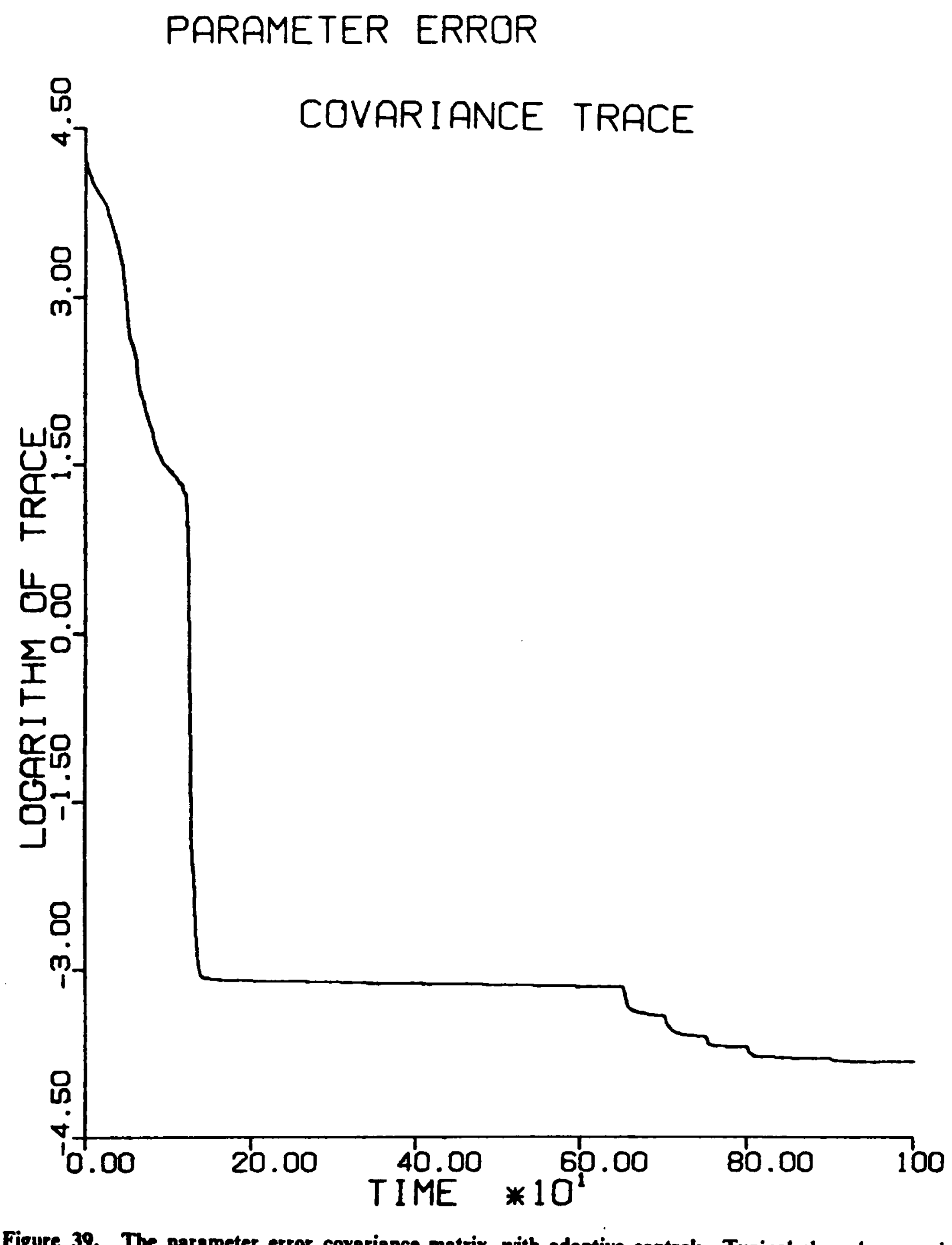

Figure 39. The parameter error covariance matrix, with adaptive control: Typical slow decrease is evident before controller is turned on. Under control, initial decrease is dramatic, but further decrease occurs only at times of excitation.

11.0 Conclusions and Areas for Further Study

In spite of the inherently nonlinear nature of the simultaneous parameter and state estimation problem for a linear time-invariant system, it turns out that a linear minimum variance estimator can be derived for it, providing that the system is transformed to observable-canonic form to begin with. That much was shown by Salut *et.al.* [5] in 1980, with the further proviso that the system inputs and outputs must be known exactly. In the work presented here, the latter proviso was eliminated, and a minimum variance estimator (PLID) generalizing the one in [5] was derived.

Unlike [5], in which no convergence result was presented, it was shown here, under some standard gaussian assumptions, and the assumption of system time-invariance, that the PLID parameter estimates converge a.e. to the exact parameter values.

These are strong results, but there are some inherent limitations due to the primary assumptions. Presumably, the object of employing the PLID algorithm is to incorporate it into an adaptive control scheme. The assumption of system time-invariance means that it can only be used in initial self-tuning regulators, but some adaptive control purists argue that initial self-tuning is not even adaptive control.

However, violating the assumption of system time-invariance, so that the unknown system varies in an unknown way, means that the extended system state vector s_k is no longer a gaussian process. Thus, the entire basis of the development of PLID is lost. One could argue that if the system is only slowly varying, then s_k is *nearly* gaussian, so the PLID algorithm is *almost* the minimum variance estimator.

No one has found the minimum variance estimator for the time-varying case, so it seems reasonable to expect that some variant of the PLID algorithm (such as incorporating a forgetting factor into the parameter error covariance computation) could be the optimal *linear* estimator for the slowly time-varying case. Certainly this is true in the limit, as the time variations tend to zero.

Another assumption that begs to be violated is the assumption that the system structure is known *a priori*. This line of thinking leads to some very interesting possible applications of the PLID algorithm. Can PLID be used to find a linear model for a nonlinear system? Can PLID be useful in model-order reduction?

Consider a nonlinear finite-order system operating at a set point. The system could be linearized by the usual Taylor series expansion about the operating point, obtaining a linear model of the same order. It seems reasonable to expect that the PLID algorithm would yield the same linearization, given the order of the system at the start. If the set point changed, one could re-initialize the PLID algorithm to obtain the new linearization. This concept certainly warrants investigation.

If the system order is underestimated, what is the result? Of course, it is tempting to say that the PLID algorithm will give the best lower-order linear system description to account for the measured behavior. It seems likely that a rigorous investigation of this case would be considerably easier if the system were assumed to be operating at some set point.

One problem with identifying a system operating at a set point is excitation of the system modes. If the system really is at a set point, it may not be possible to uniquely identify it. Therefore, a dither signal must be applied, disturbing the set point. In a system with very low noise, this is not much of a problem, because very little dither is required; but in a system with high noise levels, significant deviation from the set point may be necessary to identify the system. In practice, it may be very undesirable to introduce such deviation.

Cases of linearization and/or model-order reduction involving time-invariant systems operating at a set point may prove to be easier than the converse case, in which the order is known, but the parameters really are varying. In the former case, it is clear that a covariance reset should accompany each change in the set point. In the latter, the course of action is not as clear. Kemp [43] developed a method of applying a forgetting factor only to that part of the error covariance matrix related to the parameter estimates. Elaborate simulations of an F-15 jet fighter with an adaptive control loop employing this method (reported in [43]), showed it compares favorably to gain scheduling, which is the method actually used in the fighter. Of course, gain scheduling has a very low computation time; on the other hand, it involves a relatively large computer storage of tabular data.

The comparison in [43] raises a question concerning the implementation of the PLID algorithm, namely, can the computation time be cut substantially by some hardware implementation? This is an important question, in terms of the usability of the algorithm in real- time applications, because hardware implementations have a distinct speed advantage over implementations in software.

Many fertile areas of investigation have been opened by development of the PLID algorithm.

Appendix A. Basic Definitions in Statistics

This appendix presents definitions relevant to the understanding of martingales and the theorems in the remaining appendices.

Definition A.1: The **sample space Ω** is the set or space of all possible results or outcomes ω of an experiment or observation.

Definition A.2: An **event** is a subset of the sample space Ω.

Definition A.3: A **random variable** $X(\omega)$ is a function of the sample space Ω .

Example: For the vector random variable $X \in \mathbb{R}^m$, let A be the event

$$A = \{x_1 = a_1 , \dots , x_m = a_m\} = \bigcap_{j=1}^{m} \{\omega : x_j(\omega) = a_j\}$$

Clearly, $A \in \Omega$.

Definition A.4: Suppose $M \subseteq \Omega$. The **indicator function 1_M** is a random variable defined by

$$1_M(\omega) = \begin{bmatrix} 1 & \text{if } \omega \in M \\ 0 & \text{if } \omega \in \Omega \backslash M \end{bmatrix}$$

Definition A.5: A **σ-field on Ω** is a class of subsets of Ω that

1) contains Ω itself,
2) is closed under complementation, and
3) is closed under the formation of countable unions.

Definition A.6: A **sub-σ-field** is a σ-field that is a subset of another σ-field.

Definition A.7: Let $P(\Psi)$ denote a ***probability measure on a σ-field Ψ***; that is, P is a set function satisfying

1) $0 \leq P(A) \leq 1$ for any set $A \in \Psi$,
2) $P(\emptyset) = 0$, $P(\Omega) = 1$, and
3) P is countably additive; *i.e.*, if $A_k \in \Psi$, and $A_i \cap A_j = \emptyset$ if $i \neq j$, then

$$P\left[\bigcup_{k=1}^{\infty} A_k\right] = \sum_{k=1}^{\infty} P[A_k].$$

Definition A.8: The event $A \subset \Omega$ is true **for almost all (a.a.) ω**, or **almost everywhere (a.e.)**, or **almost surely (a.s.)**, or **almost always (a.a.)**, if $P\{\omega : \omega \notin A\} = 0$.

Definition A.9: Let (Ω, Ψ, P) denote a **probability space**, where Ω is a space of outcomes ω , Ψ is a σ-field of subsets of Ω , and P is a probability measure on Ψ.

Definition A.10: The random variable $X(\omega)$ is **measurable with respect to the sub-σ-field ψ** (or more simply, is **measurable-ψ**) if $\{\omega \mid X(\omega) = x\} \in \psi$ for almost all ω.

Definition A.11: The **real vector space of equivalence classes of finite real-valued measurable functions defined on (Ω, Ψ, P)** is denoted by $\mathscr{L}(\Omega, \Psi, P)$, or simply by $\mathscr{L}$, if it is obvious what underlying probability space is assumed. This vector space also contains the *Banach spaces*, which are normed vector spaces that are complete with respect to the norm metric; the Banach spaces will be denoted by $\mathscr{L}^p$, $1 \leq p \leq \infty$, where p indicates the norm, according to

$$\|X\|_p \triangleq \left[\int_\Omega |X|^p \, dP\right]^{1/p} \leq \infty.$$

Definition A.12: Let $\mathscr{L}(\psi)$ denote the **equivalence classes in $\mathscr{L}$ containing at least one ψ-measurable function.** Denote $\mathscr{L}^p(\psi) = \mathscr{L}^p \cap \mathscr{L}(\psi)$.

Definition A.13: The random variable $X(\omega)$ is said to be an **integrable random variable** if $E[\,|X(\omega)|\,] < \infty$, almost surely.

Definition A.14: The **σ-field generated by the random variable** X, denoted $\sigma(X)$, is the smallest σ-field with respect to which X is measurable. A σ-field can also be generated by a measurable subset of Ω.

Definition A.15: A sequence of sub-σ-fields $\psi_k \subset \Psi$ is said to be **adapted to the sequence of measurable subsets** $A_k \subset \Omega$ if $\psi_k = \sigma(A_k)$ for all k. A sequence of random variables X_k is said to be **adapted to the sequence ψ_k** if X_k is measurable-ψ_k for all k.

Definition A.16: Denote by $\psi_0, \ldots, \psi_k$ the **increasing sequence of sub-σ-fields of Ψ** generated by the increasing sequence of sets $A_0 = \{Z_0\}, \ldots, A_k = \{Z_k, \ldots, Z_0\}$.

Definition A.17: A sequence $\{X_k\}_{k=0}^{\infty}$ of real-valued random variables adapted to an increasing sequence of sub-σ-fields $\{\psi_k\}_{k=0}^{\infty}$ of Ψ is called a **martingale** if

$$(1)\ E(|X_k|) < \infty \ \forall\, k \in \mathbb{N}$$

and $$(2)\ X_k = E[X_{k+1} \mid \psi_k] \text{ a.s., } \forall\, k \in \mathbb{N}$$

In particular, if $X_k \triangleq E[Z \mid \psi_k]$, then X_k is a martingale if

$$X_k \triangleq E[Z \mid \psi_k] = E\{\, E[Z \mid \psi_{k+1}] \mid \psi_k \,\} \triangleq E\{X_{k+1} \mid \psi_k\}$$

almost surely, for all $k \geq 0$.

Definition A.18: If the equality in part (2) of Definition A.17 is replaced by '$\geq$', then the adapted sequence is called a **supermartingale**. If it is replaced by '$\leq$', then the adapted sequence is called a **submartingale.**

Notation: Martingales are denoted in several different ways. Some of the most common notations are shown below.

(i) $\{X_n, \psi_n\}_{n \in \Lambda}$, where Λ is some index set,

(ii) $\{X_n, \psi_n\}$,

(iii) $\{X_n\}$ adapted to $\{\psi_n\}$.

Appendix B. Some Convergence Concepts

This appendix defines three types of convergence commonly encountered in probability studies, and an additional type not often encountered, but required in the proof of Theorem B.6. Some of the relationships among the various convergence types are also presented, in the form of theorems and proofs. In each case, the theorem presented is required in the proof(s) of some other theorem(s) in this work.

Definition B.1: The sequence of random variables $\{X_n(\omega)\}$ **converges in probability** to the random variable $X(\omega)$ if and only if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P[\omega : \, | X_n(\omega) - X(\omega) | > \varepsilon] = 0 .$$

Definition B.2: The sequence of random variables $\{X_n(\omega)\}$ **converges almost everywhere (a.e.)**, or **with probability 1 (w.p.1)**, to the random variable $X(\omega)$ if and only if there exists a null set N (*i.e.*, $P[\mathrm{N}] = 0$) such that

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) < \infty , \quad \forall \, \omega \in \Omega \backslash \mathrm{N}.$$

Definition B.3: The sequence of random variables $\{X_n(\omega)\}$ **converges in $\mathscr{L}^p$** to the random variable $X(\omega)$ if and only if

$$\lim_{n\to\infty} E[\, | X_n(\omega) - X(\omega) |^p] = 0.$$

The last type of convergence we consider is *vague convergence*, a type not commonly encountered, but required in the proof of Theorem B.6. A preliminary definition is required, however.

Definition B.4: A measure μ on $(\mathbb{R}^1, \Psi)$, with $\mu(\mathbb{R}^1) \leq 1$, is called a **sub-probability measure (s.p.m.).**

Definition B.5: A sequence $\{\mu_n, n \geq 1\}$ of s.p.m.'s **converges vaguely** to an s.p.m. μ if and only if there exists a dense subset D of $\mathbb{R}^1$ such that for all $a,b \in D$, $a < b$, $\mu_n\{(a,b]\} \rightarrow \mu\{(a,b]\}$.

The rest of this section presents theorems detailing relationships among the various types of convergence.

Theorem B.1: If X_n converges in $\mathscr{L}^p$ to X, then X_n converges in probability to X. The converse is true, provided that $\{X_n\}$ is dominated by some $Y \in \mathscr{L}^p$.

Proof: By Chebyshev's inequality (see Theorem E.1), with $\phi(z) \equiv |z|^p$ and $\varepsilon > 0$,

$$P[\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon] \leq \frac{E[|X_n(\omega) - X(\omega)|^p]}{\varepsilon^p} . \tag{B.1}$$

By hypothesis, $E[|X_n(\omega) - X(\omega)|^p] \rightarrow 0$. Hence, $P[|X_n(\omega) - X(\omega)| \geq \varepsilon] \rightarrow 0$. So, by definition, X_n converges in probability to X, proving the first assertion.

To prove the converse, the hypotheses are $P[|X_n(\omega) - X(\omega)| \geq \varepsilon] \rightarrow 0$, for all $\varepsilon > 0$; and $|X_n| \leq Y$ a.e., where $E[Y^p] < \infty$. Note that $|X_n - X| \leq 3Y \in \mathscr{L}^p$. Hence,

$$E\left[|X_n(\omega)-X(\omega)|^p\right] = \int_{\{|X_n - X| \le \varepsilon\}} |X_n(\omega)-X(\omega)|^p\, dP + \int_{\{|X_n - X| > \varepsilon\}} |X_n(\omega)-X(\omega)|^p\, dP$$

$$\le \varepsilon^p + \int_{\{|X_n - X| > \varepsilon\}} (3Y)^p\, dP \tag{B.2}$$

But, by hypothesis, $P\left[|X_n(\omega)-X(\omega)| \ge \varepsilon\right] \to 0$, for all $\varepsilon > 0$. Therefore, by Theorem E.10, the limit, as n goes to infinity, of the last integral on the right is zero. That leaves

$$\lim_{n\to\infty} E\left[|X_n(\omega)-X(\omega)|^p\right] \le \varepsilon^p \tag{B.3}$$

for any $\varepsilon > 0$. Letting $\varepsilon \to 0$, the limit in Eqn. B.3 is zero. Therefore, $X_n \to X$ in $\mathscr{L}^p$. ■

Theorem B.2: If X_n converges a.e. to X, then X_n converges in probability to X.

Proof: Since X_n converges a.e. to X, let N denote the null set on which it does not converge; define $\Omega_0 = \Omega \backslash \mathrm{N}$. For $m \ge 1$ and $\varepsilon > 0$, define the event

$$A_m = \cap_{n=m}^{\infty} \{\omega : |X_n - X| \le \varepsilon\}. \tag{B.4}$$

For a fixed value of ε, A_m is increasing with m. For each $\omega_0 \in \Omega_0$, $X_n(\omega_0) \to X(\omega_0)$; hence, there exists $m(\omega_0, \varepsilon)$ such that $n \ge m(\omega_0, \varepsilon) \Rightarrow |X_n(\omega_0) - X(\omega_0)| \le \varepsilon$. Therefore, $\omega_0 \in A_{m(\omega_0,\varepsilon)}$; it follows that $\Omega_0 \subset \cup_{m=1}^{\infty} A_m$. Because A_m is increasing,

$$\Omega_0 \subset \lim_{m\to\infty} A_m \Rightarrow \Omega_0 \subset \lim_{k\to\infty} \bigcup_{k=1}^{\infty} \bigcap_{m=k}^{\infty} A_m = \{\omega : \omega \in A_m \ a.a.\} \tag{B.5}$$

Because A_m increases at least until it includes Ω_0, and $P(\Omega_0) = 1$, then by the monotone convergence property of measures, $\lim_{m\to\infty} P(A_m) = 1$.

■

The next theorem is not quite the converse of Theorem B.2. It shows that convergence in probability implies convergence a.e. along some subsequence.

Theorem B.3: If X_n converges to X in probability, then there exists a countably infinite sequence of integers $\{ n_k \}$, $k = 0,1, \ldots$, such that the subsequence X_{n_k} converges to X a.e..

Proof: The hypothesis can be restated as follows. For all $k \geq 0$,

$$\lim_{n\to\infty} P\left[\mid X_n - X \mid > \frac{1}{2^k} \right] = 0. \tag{B.6}$$

Therefore, for each k there is a value n_k such that

$$P\left[\mid X_{n_k} - X \mid > \frac{1}{2^k} \right] \leq \frac{1}{2^k} . \tag{B.7}$$

Choosing the entire subsequence $\{ n_k \}$ by that criterion, it follows that

$$\sum_{k=0}^{\infty} P\left[\mid X_{n_k} - X \mid > \frac{1}{2^k} \right] \leq \sum_{k=0}^{\infty} \frac{1}{2^k} < \infty. \tag{B.8}$$

So by the first Borel-Cantelli lemma (see Theorem E.5),

$$P\left[\mid X_{n_k} - X \mid > \frac{1}{2^k} \ \ i.o. \right] = 0. \tag{B.9}$$

Therefore,

$$P\left[\mid X_{n_k} - X \mid \leq \frac{1}{2^k} \ \ a.a. \right] = 1. \tag{B.10}$$

So $| X_{n_k} - X | \to 0$ a.e., implying that $X_{n_k} \to X$ a.e.. ∎

The next theorem considers the relation between the convergence of a random sequence and the convergence of its higher-order moments, for two types of convergence.

Theorem B.4: (1) If X_n converges to X a.e., then for every $p > 0$,

$$E[| X |^p] \leq \liminf_n E[| X_n |^p].$$

(2) If X_n converges to X in $\mathscr{L}^p$, where $1 \leq p < \infty$, and $X \in \mathscr{L}^p$, then

$$E[| X_n |^p] \to E[| X |^p].$$

Proof: The first part is just a special case of Fatou's Lemma (see Theorem E.8),

$$\int_\Omega | X |^p \, dP = \int_\Omega \liminf_n | X_n |^p \, dP \leq \liminf_n \int_\Omega | X_n |^p \, dP. \tag{B.11}$$

For the second part of the theorem, where X_n converges in $\mathscr{L}^p$ to X, using the fact that $X_n = X - (X - X_n)$, Minkowski's inequality (see Theorem E.4) can be used to write

$$E^{1/p}[| X_n |^p] \leq E^{1/p}[| X |^p] + E^{1/p}[| X_n - X |^p] \tag{B.12}$$

Similarly,

$$E^{1/p}[| X |^p] \leq E^{1/p}[| X_n |^p] + E^{1/p}[| X_n - X |^p] \tag{B.13}$$

Hence,

$$E^{1/p}[\,|X_n|^p] - E^{1/p}[\,|X_n - X|^p] \\ \leq E^{1/p}[\,|X|^p] \leq E^{1/p}[\,|X_n|^p] + E^{1/p}[\,|X_n - X|^p] \tag{B.14}$$

Letting $n \to \infty$, second part of the theorem is proved. ■

Before presenting the last theorem relating different convergence types, it is necessary to define uniform integrability and to present a theorem characterizing it.

Definition B.6: A family of random variables $\{X_t\}_{t\in T}$, where T is an arbitary index set, is **uniformly integrable** if and only if

$$\lim_{M\to\infty} \int_{|X_t|>M} |X_t| \; dP = 0 \;\; \text{uniformly in } t\in T.$$

Theorem B.5: The family of random variables $\{X_t\}$ is uniformly integrable if and only if the following conditions are both satisfied:

(1) $E[\,|X_t|\,]$ is bounded in $t\in T$;

(2) For each $\varepsilon > 0$, there exists $\delta(\varepsilon)$ such that, for any $A \subset \Psi$,

$$P[A] < \delta \;\; \Rightarrow \;\; \int_A |X_t| \; dP < \varepsilon \quad \text{for every } t\in T.$$

Proof: The definition of uniform integrability implies condition (1). To see that uniform integrability implies condition (2), let $A\subset\Psi$, and denote $A_t = \{\omega : \; |X_t| > M\}$. Hence,

$$\int_A |X_t|\,dP = \int_{A\cap A_t} |X_t|\,dP + \int_{A\setminus A_t} |X_t|\,dP \le \int_{A_t} |X_t|\,dP + M\,P[\,A\,]. \qquad \text{(B.15)}$$

Now, given $\varepsilon > 0$, by hypothesis there exists $M = M(\varepsilon)$ such that the last integral on the right is less than $\varepsilon/2$ for every t. Therefore, condition (2) follows if $\delta = \varepsilon/2M$, from the definition of uniform integrability.

Conversely, suppose that conditions (1) and (2) are both true. Then by Chebyshev's inequality (see Theorem E.1),

$$P\{\,|X_t| > M\,\} \le \frac{E[\,|X_t|\,]}{M} \le \frac{B}{M}\,,\ \forall\, t\in T, \qquad \text{(B.16)}$$

where B is the bound imposed by condition (1). For hypothesis (2), defining $\varepsilon = 1/M$ and $\delta(\varepsilon) = B(\,1/M\,)$, then from Equation B.16,

$$\int_{A_t} |X_t|\ dP = \int_{|X_t|>M} |X_t|\ dP < \varepsilon = 1/M \qquad \text{(B.17)}$$

In the limit, as $M \to \infty$, the defining equation for uniform integrability is obtained from Equation B.17. ■

The last theorem in this appendix presents a relation between convergence in probability, uniform integrability, and convergence in $\mathscr{L}^p$.

Theorem B.6: Suppose $X_n \in \mathscr{L}^1$, and X_n converges to X in probability. Then the following three propositions are equivalent:

(i) $\{\,|X_n|\,\}$ is uniformly integrable;

(ii) X_n converges to X in $\mathscr{L}^1$;

(iii) $E[\,|X_n|\,] \rightarrow E[\,|X|\,]$.

Proof: (due to Chung [44])

(i) $\Rightarrow$ (ii): By Theorem B.3, since $X_n \rightarrow X$ in probability, there exists a subsequence $\{X_{n_k}\}$ that converges a.e. to X. Therefore, by part 1 of Theorem B.5 and part 1 of Theorem B.4, $X \in \mathscr{L}^1$ It follows that $[\,X_n - X\,]$ is also uniformly integrable.

Now, for each $\varepsilon > 0$,

$$\begin{aligned}\int_\Omega |X_n - X|\,dP &= \int_{|X_n - X|>\varepsilon} |X_n - X|\,dP + \int_{|X_n - X|\leq\varepsilon} |X_n - X|\,dP \\ &\leq \int_{|X_n - X|>\varepsilon} |X_n - X|\,dP + \varepsilon\end{aligned} \tag{B.18}$$

Since $P\{\,|X_n - X| > \varepsilon\,\} \rightarrow 0$ as $n \rightarrow \infty$ (convergence in probability), then by part 2 of Theorem B.5, the last integral in Equation B.18 also tends to zero. Because $\varepsilon > 0$ is arbitrary,

$$\lim_{n\rightarrow\infty} \int_\Omega |X_n - X|\,dP = 0 \;\Rightarrow\; X_n \rightarrow X \text{ in } \mathscr{L}^1 . \tag{B.19}$$

(ii) $\Rightarrow$ (iii): This is simply part 2 of Theorem B.4, with $p = 1$.

(iii) $\Rightarrow$ (i): Let $X_\infty \overset{\Delta}{=} X$, and define the uniformly bounded sequence

$$X_n^M = \left[\begin{array}{ll} X_n \;, & \text{if}\,|X_n| \leq M \\ M \;, & \text{if}\,|X_n| > M \end{array}\right. \tag{B.20}$$

Clearly, $|X_n^M| \rightarrow |X_\infty^M|$ in probability. Since the sequence is bounded, then by part 2 of Theorem B.1, $|X_n^M| \rightarrow |X_\infty^M|$ in $\mathscr{L}^1$. Consequently, by part 2 of Theorem B.4, $E[\,|X_n^M|\,] \rightarrow E[\,|X_\infty^M|\,]$.

Now, if $\pm M$ are chosen to be points of continuity of the distribution function of X_∞, then the sequence of sub-probability measures $P\{\,|X_n| > M\,\}$ converges vaguely to the sub-probability measure $P\{\,|X_\infty| > M\,\}$. Therefore,

$$\begin{aligned}\int_{|X_n|>M} |X_n| \, dP \\ = E[\,|X_n|\,] - E[\,|X_n^M|\,] \;\to\; E[\,|X_\infty|\,] - E[\,|X_\infty^M|\,] \\ = \int_{|X_\infty|>M} |X_\infty| \, dP.\end{aligned} \tag{B.21}$$

As $M \to \infty$, the last integral goes to zero, independent of n.

Therefore, by definition, (i) is true.

The theorem is proved, because (i) ⇒ (ii) ⇒ (iii) ⇒ (i). ∎

Appendix C. Optional Random Variables (Stopping Times)

The topic of *optional random variables,* also known as *stopping times,* is so important to the study of martingales that it deserves a separate appendix. The concept of a stopping time for a random process evolved from the study of gambling strategies.

Basically, an optional r.v. (stopping time) is a function (rule) of the previous history of a random process (game) that determines when the process is to be stopped (game is to be ended). In a broader sense, the game is not ended; rather, the player only leaves the game at the time determined by the optional r.v. The determination is to be based entirely on past events, so it rules out prescience and other types of foreknowledge as input to the optional r.v.

These ideas are made more rigorous by the following definition.

Definition C.1: A function $v : \Omega \rightarrow \overline{\mathbb{N}} = \{ 1, 2, \ldots \} \bigcup \infty$ is called an **optional random variable** (or a **stopping time**) if the set $\{\omega : v(\omega) = n\}$ is ψ_n-measurable for all $n \in \overline{\mathbb{N}}$, where $\{\psi_n\}$ is an increasing sequence of sub-σ-fields.

The optional r.v. v has associated with it the σ-field ψ_v of subsets of Ω given by

$$\psi_v = \{ A : A \in \psi_\infty, \ A \cap \{\omega : v(\omega) = n\} \in \psi_n \ \forall \ n \in \mathbb{N} \} .$$

ψ_v is sometimes called the ***pre-v-field,*** and the events in ψ_v are prior to v.

The following example is widely used to clarify the preceding definition. It is called a "hitting time."

Example: For every sequence $\{X_n\}$ adapted to the increasing sequence of sub-σ-fields $\{\psi_n\}$ in an arbitrary measurable space (Ω, Ψ), the ***hitting time*** v_A of a subset $A \in \Psi$, defined by

$$v_A(\omega) = \begin{cases} \inf\{ n : X_n(\omega) \in A \} & \text{if } \omega \in \bigcup_{\mathbb{N}} \{ X_n \in A \}, \\ +\infty & \text{otherwise,} \end{cases}$$

is a stopping time. Note that, for each $n \in \mathbb{N}$, the set of ω such that $v_A(\omega) = n$ is defined by

$$\{\omega : v_A(\omega) = n\} = \bigcap_{m<n} \{X_m \in A^c\} \cap \{X_n \in A\}$$

The following lemma characterizes optional random variables and associated σ-fields.

Theorem C.1: A mapping $v : \Omega \to \overline{\mathbb{N}}$ is a stopping time if and only if $\{ v \leq n\} \in \psi_n$ for all $n \in \mathbb{N}$. An event $A \in \psi_\infty$ belongs to the σ-field ψ_v associated with a stopping time v if and only if $A \cap \{ v \leq n\} \in \psi_n$ for all $n \in \mathbb{N}$.

Proof: The lemma follows from

$$\{ v \leq n\} = \bigcup_{m \leq n} \{ v = m \},$$

$$\{ v = n \} = \{ v \leq n\} \cap \{ v \leq n - 1 \}^c,$$

and the fact that ψ_t is an increasing sequence. ■

It follows from Theorem C.1 that $\min_k v_k$ and $\max_k v_k$ of the finite or infinite sequence of stopping times $\{v_k\}$ are also stopping times, because the sets

$$\{\omega : \min_k v_k(\omega) \le n\} = \bigcup_k \{ v_k \le m \}, \quad \text{and} \quad \{\omega : \max_k v_k(\omega) \le n\} = \bigcap_k \{ v_k \le m \},$$

belong to ψ_n for all $n \in \mathbb{N}$. From this fact, it easily follows that, if v is a stopping time, then $\min(v, n)$ and $\max(v, n)$ are also stopping times.

The σ-field associated with an optional r.v. is more fully described by the following lemma and corollary.

Theorem C.2: Let v be a stopping time. A real-valued ψ_∞-measureable function $f : \Omega \rightarrow \overline{\mathbb{R}}$ is ψ_v-measureable if and only if for all $n \in \mathbb{N}$, the restriction of f to $\{\omega : v(\omega) = n\}$ is ψ_n-measureable.

Furthermore, for every positive or integrable real-valued measureable function g defined on the probability space (Ω, Ψ, P),

$$E_{\psi_v \cap \{\omega : v(\omega)=n\}}(g) = E_{\psi_n \cap \{\omega : v(\omega)=n\}}(g).$$

Proof: (from [45]) Suppose $f = I_A$, the indicator function of the set $A \in \psi_\infty$. Then the first part of the lemma follows from the definition of ψ_v. Extension to an arbitrary ψ_∞-measureable function is immediate, using typical methods from real analysis.

Now, suppose g is positive. By the first part of the lemma, $E_{\psi_n \cap \{\omega : v(\omega)=n\}}(g)$ is a positive ψ_v-measureable function. Therefore,

$$\sum_{\overline{\mathbb{N}}} \int_{A \cap \{\omega : v(\omega)=n\}} E_{\psi_n}(g)\, dP = \sum_{\overline{\mathbb{N}}} \int_{A \cap \{\omega : v(\omega)=n\}} g\, dP = \int_A g\, dP = \int_A E_{\psi_v}(g)\, dP$$

because $A \cap \{\omega : v(\omega) = n\} \in \psi_n$ for all $n \in \overline{\mathbb{N}}$. This proves the second part of the lemma for positive functions. The proof can be extended to arbitary real-valued integrable functions using typical methods from real analysis. ■

Corollary C.3: For every sequence X_n adapted to an increasing sequence of sub-σ-fields ψ_n, and for every stopping time v, the random variable X_v , defined by $X_{v(\omega)} = X_n(\omega)$ for all ω such that $v(\omega) = n \in \overline{\mathbb{N}}$, is ψ_v-measureable.

Theorem C.4: For every pair v, v' of optional random variables, the events $\{\omega : v(\omega) < v'(\omega)\}$, $\{\omega : v(\omega) = v'(\omega)\}$, and $\{\omega : v(\omega) \leq v'(\omega)\}$ belong to ψ_v and to $\psi_{v'}$. However, it is also true that $A \in \psi_v$ implies that $A \cap \{\omega : v(\omega) \leq v'(\omega)\} \in \psi_{v'}$. It follows that if $v \leq v'$ for all $\omega \in \Omega$, then $\psi_v \subset \psi_{v'}$.

Proof: [46] Both v and v' are necessarily ψ_∞-measureable, so it is clear that all the sets considered above belong to ψ_∞. But, because

$$\{\omega : v(\omega) < v'(\omega)\} \cap \{\omega : v(\omega) = n\} = \{\omega : v'(\omega) > n\} \cap \{\omega : v(\omega) = n\} \in \psi_n ,$$
$$\{\omega : v(\omega) = v'(\omega)\} \cap \{\omega : v(\omega) = n\} = \{\omega : v'(\omega) = n\} \cap \{\omega : v(\omega) = n\} \in \psi_n ,$$

for all $n \in \mathbb{N}$, the events $\{\omega : v(\omega) < v'(\omega)\}$, and $\{\omega : v(\omega) = v'(\omega)\}$ belong to ψ_v . Therefore, the union $\{\omega : v(\omega) \leq v'(\omega)\}$ also belongs to ψ_v.

Now, by a symmetry argument, $\{\omega : v(\omega) = v'(\omega)\}$ belongs to to $\psi_{v'}$. Similarly, taking the complements of $\{\omega : v(\omega) < v'(\omega)\}$ and $\{\omega : v(\omega) \leq v'(\omega)\}$, and reversing the roles of v and v', it is clear that the same events belong to $\psi_{v'}$.

Thus, if $A \in \psi_v$, then

$$A \cap \{\omega : v(\omega) \leq v'(\omega)\} \cap \{\omega : v'(\omega) = n\} = A \cap \{\omega : v(\omega) \leq n\} \cap \{\omega : v'(\omega) = n\} \in \psi_n ,$$

for all $n \in \mathbb{N}$. Consequently, $A \cap \{\omega : v(\omega) \leq v'(\omega)\}$, which obviously belongs to ψ_{∞}, also belongs to $\psi_{v'}$. ▪

Theorem C.5: Let $\{ X_n , \psi_n\}$ be a martingale and α, β be two ψ_n -measureable optional random variables (stopping times) such that $\alpha \leq \beta \leq M < \infty$ a.e.. If $\{ X_n\}$ is uniformly integrable, then the two random variables $\{ X_\alpha , X_\beta \}$ form a martingale relative to $\{ \psi_\alpha , \psi_\beta \}$.

Proof: (due to Chung [47]) From Theorems C.2 and C.3, it follows that $\{ X_\alpha , X_\beta \}$ is adapted to $\{ \psi_\alpha , \psi_\beta \}$. Let $A \in \psi_\alpha$ and $A_i = A \cap \{\omega : \alpha(\omega) = i \}$. Then, for each $k \geq i$, necessarily $A_i \cap \{\omega : \beta(\omega) > k \} \in \psi_k$, because $A_i \in \psi_i$, $\{\omega : \beta(\omega) > k \} \in \psi_k$, and $\psi_i \subset \psi_k$. Now, by virtue of the definitions of these various sets,

$$\begin{aligned} \int_{A_i \cap \{\omega : \beta(\omega) \geq k\}} X_k \, dP &= \int_{A_i \cap \{\omega : \beta(\omega) = k\}} X_k \, dP + \int_{A_i \cap \{\omega : \beta(\omega) > k\}} X_k \, dP \\ &= \int_{A_i \cap \{\omega : \beta(\omega) = k\}} X_\beta \, dP + \int_{A_i \cap \{\omega : \beta(\omega) \geq k+1\}} X_{k+1} \, dP \end{aligned} \tag{C.1}$$

By iterating the relation of Equation C.1, it follows that for each $j \geq i$,

$$\int_{A_i \cap \{\omega : \beta(\omega) \geq i\}} X_i \, dP = \int_{A_i \cap \{\omega : i \leq \beta(\omega) \leq j\}} X_\beta \, dP + \int_{\{\omega : \beta(\omega) \geq j+1\}} X_{j+1} \, dP \tag{C.2}$$

Now, the term on the left side of Equation C.2 is equivalent to $\int_{A_i} X_\alpha \, dP$, because $\beta \geq \alpha = i$ on the set A_i. The same reasoning allows us to change the domain of the integration in the first term on the right to the set $A_i \cap \{\omega : \beta(\omega) \leq j \}$. Now, letting

$j \to \infty$, the last term on the right converges to zero, because α and β are finite. Thus, Equation C.2 can be rewritten as

$$\int_{A_i} X_\alpha \, dP = \int_{A_i} X_\beta \, dP \tag{C.3}$$

Now, by summing over i, the domain of integration becomes the set $A \in \psi_\alpha \subset \psi_\beta$.

It remains to be shown that X_α and X_β are both integrable, but this is almost trivial for any bounded random variable α or β, not necessarily stopping times, because $\alpha \leq M$ implies

$$E(\,|X_\alpha|\,) = \sum_{i=1}^{M} \int_{\{\omega\,:\,\alpha(\omega)=i\,\}} |X_i| \, dP = \sum_{i=1}^{M} E(\,|X_i|\,) < \infty.$$

The same holds for $\beta \leq m$. ■

Theorem C.6: Let $\{X_n\,, \psi_n\}$ be a martingale, and let $\{\alpha_n\}$ be a sequence of stopping times, measureable-ψ_n, such that $\alpha_m \leq \alpha_n < \infty$ a.e. if $m \leq n$.

If for each n, there exists a constant finite M_n such that $\alpha_n \leq M_n$ a.e.,

then $\{\,X_{\alpha_n}\,, \psi_{\alpha_n}\}$ is a martingale.

Also, for each k,

$$E[\,X_1\,] = E[X_k\,] = \lim_{n \to \infty} E[\,X_n\,] \tag{C.4}$$

Proof: That $\{\,X_{\alpha_n}\,, \psi_{\alpha_n}\}$ is a martingale follows from the proof of Theorem C.5, applying it to each pair $\alpha_m\,, M_n$.

In order to prove Equation C.4, note that

$$
\begin{aligned}
E(X_{\alpha_k}) &= \sum_{n=1}^{m} \int_{\{\omega : \alpha_k = n\}} X_n \, dP + \int_{\{\omega : \alpha_k > m\}} X_{\alpha_k} \, dP \\
&= \sum_{n=1}^{m} \int_{\{\omega : \alpha_k = n\}} X_m \, dP + \int_{\{\omega : \alpha_k > m\}} X_{\alpha_k} \, dP \\
&= E(X_m) - \int_{\{\omega : \alpha_k > m\}} X_m \, dP + \int_{\{\omega : \alpha_k > m\}} X_{\alpha_k} \, dP
\end{aligned}
$$

Now, as $m \rightarrow \infty$, the last term tends to zero because $E(|X_{\alpha_k}|) < \infty$, being a martingale; the second to last term also tends to zero because $\alpha_n \leq M_n < \infty$ a.e.. Thus, the second equality in Equation C.4 is proved.

To prove the first equality in Equation C.4, append the stopping time $\alpha_0 = 1$, with $\psi_0 = \psi \subset \psi_1$, to the beginning of the sequence $\{\alpha_n\}$. Note that this alters none of the foregoing proof. But now it is clear that the first equality follows from Theorem C.5 and Corollary D.4. ■

A good example of the kind of process to which the previous theorem is addressed is given by $\alpha_n = \min(\alpha, n)$, where α is a *fixed* stopping time. Thus, the new martingale X_{α_n} is equal to the old martingale X_n until the stopping time α, after which it is a constant.

Appendix D. Conditional Expectation and Martingales

Definition D.1: Given an integrable random variable X and a sub-σ-field ψ, the **conditional expectation** $E[X \mid \psi]$ **of** X **relative to** ψ is any one of the equivalence class of random variables on Ω satisfying the following two properties:

(1) it belongs to ψ;

(2) it has the same integral as X over any set in ψ.

The following lemma provides a characterization of conditional expectation:

Lemma D.1: Let p be a real number in $[1, \infty]$, ψ a sub-σ-field of Ψ in the probability space (Ω, Ψ, P), and E_ψ the expectation conditioned on ψ. Then

1) for all $f \in \mathscr{L}^p$, the conditional expectation $E_\psi(f) \in \mathscr{L}^p$;

2) the operator E_ψ on $\mathscr{L}^p$ is positive, idempotent, and a linear contraction such that $E_\psi(1) = 1$;

3) E_ψ maps $\mathscr{L}^p$ onto $\mathscr{L}^p(\psi)$.

Proof can be found in Neveu [48].

Note that property three indicates conditional expectation is a projection operator. Indeed, for $\mathscr{L}^2$ the orthogonal projection onto the closed vector subspace $\mathscr{L}^2(\psi)$ is called the *conditional expectation with respect to the sub-σ-field ψ of Ψ.*

Theorem D.2: If X is integrable, and $\psi_1 \subset \psi_2$, then

(1) $E[X \mid \psi_1] = E[X \mid \psi_2]$ if and only if $E[X \mid \psi_2] \in \psi_1$, and

(2) $E\{E[X \mid \psi_2] \mid \psi_1\} = E[X \mid \psi_1] = E\{E[X \mid \psi_1] \mid \psi_2\}$.

Proof: Trivially, if and only if $Y \in \psi_1$, then $Y = E[\,Y \mid \psi_1\,]$. Replace Y by $E[\,X \mid \psi_2\,]$, and the first part follows.

Part (1) can now be used to prove the second equation of part (2), because $E[\,X \mid \psi_1\,] \in \psi_1 \subset \psi_2$. To prove the first equation of part (2), consider

$$\begin{aligned} E\{\,E[\,X \mid \psi_2\,] \mid \psi_1\,\}\; dP &= \int_{\psi_1} E[\,X \mid \psi_2\,]\; dP \\ &= \int_{\psi_1 \cap \psi_2} X\; dP = \int_{\psi_1} X\; dP = E[\,X \mid \psi_1\,] \qquad \blacksquare \end{aligned} \tag{D.1}$$

Theorem D.3: Let $\{\,X_n\,\}$ be a submartingale adapted to the increasing sequence $\{\,\psi_n\,\}$ of sub-σ-fields, and let ϕ be an increasing convex function on $\mathbb{R}^1$. If $\phi\,(\,X_n\,)$ is integrable for each n, then $\{\,\phi\,(\,X_n\,)\,\}$ is a submartingale adapted to $\{\,\psi_n\,\}$.

Proof: ϕ is increasing, and

$$X_n \le E[\,X_{n+1} \mid \psi_n\,] \text{ a.e.} \tag{D.2}$$

Therefore,

$$\phi\,(\,X_n\,) \le \phi\,(\,E[\,X_{n+1} \mid \psi_n\,]\,) \text{ a.e.} \tag{D.3}$$

Applying Jensen's inequality (see Theorem E.2) to the right side of Equation D.3,

$$\phi\,(\,E[\,X_{n+1} \mid \psi_n\,]\,) \le E[\,\phi\,(\,X_{n+1}\,) \mid \psi_n\,] \text{ a.e.} \tag{D.4}$$

which, by the definition of a submartingale, proves the theorem. ■

Corollary D.4: If $\{\,X_n\,\}$ is a submartingale adapted to the increasing sequence $\{\,\psi_n\,\}$ of sub-σ-fields, then so is X_n^+, where

$$X_n^+ = \left[\begin{array}{ll} X_n \,, & \text{if } X_n \geq 0 \\ 0 \,, & \text{if } X_n < 0 \end{array} \right.$$

Proof: Theorem D.3 applies, because the function ( )$^+$ is convex and increasing. ∎

Intuitively, it seems that martingales are random processes that ***wander around*** some nominal point. From another perspective, however, these wanderings can be viewed as aperiodic oscillations. Therefore, within the question of convergence of a martingale is imbedded the question of how the martingale oscillates in various regions of its domain.

Given the sequence $\{ X_n(\omega) \}$ of r.v.'s, for each point ω in the sample space, the convergence of the sequence depends on the oscillations over finite segments as $n \to \infty$. To be more specific, the sequence will have a limit if and only if the number of oscillations between *any* two rational numbers a and b is finite. The number of oscillations will of course depend upon a, b, and ω. This argument is a standard approach in measure and integration theory.

The astonishing thing about this argument applied to martingales, is that a relatively sharp estimate can be obtained for the expected number of oscillations. The next theorem addresses this topic, but first a definition is required. The definition that follows is rather painstakingly detailed, but this is necessary in the theorem that follows it.

Definition D.2: Let $a < b$. The number γ of **upcrossings of the interval** $[\, a \,, b \,]$ by a sequence of numbers $\{ x_1 \,, \, \dots \,, x_n \}$ is defined as follows. Let

$$\begin{aligned} v_1 &= \min \{ j : 1 \leq j \leq n \,, x_j \leq a \}, \\ v_2 &= \min \{ j : v_1 < j \leq n \,, x_j \geq b \}. \end{aligned}$$

If v_1 or v_2 fails to be defined because such a number j does not exist, then the number of upcrossings γ is set to zero. However, if they *are* defined, then further define

$$\begin{aligned} v_{2k-1} &= \min \{ j : v_{2k-2} < j \leq n \, , x_j \leq a \}, \\ v_{2k} &= \min \{ j : v_{2k-1} < j \leq n \, , x_j \geq b \}. \end{aligned}$$

If any one of this sequence is undefined, then all the subsequent ones are also undefined. Denote the last one defined by v_l, where $l = 0$ in the earlier case where α_1 is undefined. The number of upcrossings γ is defined as $l/2$.

Untangling the definition, it can be seen that γ simply counts the number of times that the sequence progresses from the point at the low end of the interval $[\, a \, , b \,]$ to the point at the high end of the interval; that is, the *upcrossings* of the interval $[\, a \, , b \,]$.

Theorem D.5: Let $\{ X_k \, , \psi_k \}_{k=1}^{n}$ be a martingale, and $-\infty < a < b < \infty$. Let $\gamma^{(n)}_{[a,b]}(\omega)$ denote the number of upcrossings of the interval $[\, a \, , b \,]$ by the sequence $\{ X_k(\omega) \}_{k=1}^{n}$. Then

$$E\left[\gamma^{(n)}_{[a,b]}\right] \leq \frac{E\left[(X_n - a)^{+}\right] - E\left[(X_1 - a)^{+}\right]}{b - a} \leq \frac{E[\, X_n^{+} \,] + \| a \|}{b - a}$$

$$\text{where } \; (X)^{+} = \left[\begin{array}{ll} X \, , & \text{if } X \geq 0 \, ; \\ 0 \, , & \text{otherwise.} \end{array} \right.$$

Proof: (following Chung [49]) Start by considering the case where $X_k(\omega) \geq 0$ a.e. for all $k = 1, \ldots, n$, and $a = 0$. This causes X_{v_j} to be zero whenever j is an odd number, where v_j is defined in Definition D.2, and the x_i in that definition are the $X_i(\omega)$ in the theorem statement.

Also recall from Definition D.2 that, for each ω, the sequence $v_j(\omega)$ is defined only up to $j = l(\omega) \leq n$. But now the definition is modified so that all previously undefined

$v_j(\omega)$ are set to n; thus, $v_j(\omega)$ is defined for all $j = 1, \ldots, n$, for almost all ω. Now, appending $v_0 \equiv 1$ to the start of the sequence, and noting that $v_n = n$ a.e.,

$$\begin{aligned} X_n - X_1 &= X_{v_n} - X_{v_0} = \sum_{j=0}^{n-1} (X_{v_{j+1}} - X_{v_j}) \\ &= \sum_{j \text{ odd}} (X_{v_{j+1}} - X_{v_j}) + \sum_{j \text{ even}} (X_{v_{j+1}} - X_{v_j}). \end{aligned} \tag{D.5}$$

Now consider the sum of the odd-number terms. There are three possibilities, $j < l(\omega)$, $j = l(\omega)$, and $j > l(\omega)$.

Case 1: j odd, and $j + 1 \leq l(\omega)$:

$$X_{v_{j+1}}(\omega) \geq b > 0 = X_{v_j}(\omega)$$

Case 2: j odd, and $j = l(\omega)$:

$$X_{v_{j+1}}(\omega) = X_n(\omega) \geq 0 = X_{v_j}(\omega)$$

Case 2: j odd, and $j > l(\omega)$:

$$X_{v_{j+1}}(\omega) = X_n(\omega) = X_{v_j}(\omega)$$

In all three cases,

$$\sum_{j \text{ odd}} (X_{v_{j+1}} - X_{v_j}) = \sum_{\substack{j \text{ odd} \\ j+1 \leq l(\omega)}} (X_{v_{j+1}} - X_{v_j}) \geq \left[\frac{l(\omega)}{2} \right] b = \gamma_{[0,b]}^{(n)} b \tag{D.6}$$

Note that the general form of the sequence $\{v_j\}$ is

$$1 = v_0 \leq v_1 < v_2 < \dots < v_l = v_{l+1} = \cdots = v_n = n,$$

which is an increasing sequence of optional random variables, because constants are also optional r.v.'s. Therefore, by Lemma C.6, $\{ X_{v_j}, \psi_{v_j} \}$ is a martingale. Thus, for each $j \in \{0, \dots, n-1\}$, $E[X_{j+1} - X_j] = 0$. Therefore the expectation of the sum of the even terms in Equation D.5 is zero.

So, from Equations D.5 and D.6,

$$E[X_n - X_1] \geq E[\gamma^{(n)}_{[0,b]}]\, b, \tag{D.7}$$

which proves the theorem for the case when $a = 0$ and the sequence X_n is non-negative.

To obtain the general case, apply the case already proved to $\{ (X_j - a)^+ \}_{j=1}^{n}$, which is a martingale by Corollary D.4. Then the number of upcrossings of the interval $[a,b]$ by the martingale X_j is equal to the number of upcrossings of the interval $[0, b-a]$ by the modified martingale $X_j - a$. Clearly, then Equation D.7, after the appropriate substitutions, becomes the first inequality in the statement of the theorem. The second inequality follows easily, because $(X_n - a)^+ \leq X_n^+ + |a|$. ■

The upcrossing inequality given by Theorem D.5 is the basis for proving the following basic convergence theorem.

Theorem D.6: If $\{ X_n, \psi_n \}_{n \in \mathbb{N}}$, is a martingale,

then $\{ X_n \}$ converges a.e. to a finite limit.

Proof: By definition, $2E(X_n^+) = E(|X_n|) + E(X_n) \leq 2E(|X_n|) < \infty$, for each $n \in \mathbb{N}$. Therefore, the upcrossings inequality, Theorem D.5, can be applied. That inequality, together with the monotone convergence theorem, shows that the expected number of upcrossings $\gamma_{[a,b]}$ of any interval $[a,b]$ where $a < b$, is almost surely finite.

Therefore, for each pair of rational numbers $a < b$,

$$\{\, \omega : \liminf_n X_n(\omega) < a < b < \limsup_n X_n(\omega) \,\}$$

is a null set (*i.e.*, a set with probability zero). Therefore, the union over all such pairs is also a null set. Therefore, the limit $\lim_{n\to\infty} X_n$ exists, because the union contains the set where $\lim_n \text{sub}\, X_n < \limsup_n X_n$. Applying Fatou's Lemma (Theorem E.8), the limit must be finite almost everywhere.

Theorem D.7: If $\{ X_n , \psi_n \}_{n\in\mathbb{N}}$, is a martingale, then the following propositions are equivalent:

(1) $\{ X_n \}$ is a uniformly integrable sequence;

(2) X_n converges in $\mathscr{L}^1$;

(3) X_n converges a.e. to an integrable X_∞ such that $\{ X_n , \psi_n \}_{n\in\bar{\mathbb{N}}}$, is a martingale, and $E(X_n)$ converges to $E(X_\infty)$.

(4) there exists an integrable random variable Y such that $X_n = E(Y \mid \psi_n)$ for every $n \in \mathbb{N}$.

Proof:

(1) ⇒ (2): Assumption (1) satisfies the premise of Theorem D.6; therefore, $X_n \to X_\infty$ a.e. Now the premises of Theorem B.6 are satisfied, and the condition of uniform integrability is shown there to be equivalent to the condition that $E[\,|X_n|\,] \to E[\,|X|\,]$. Thus X_n converges in $\mathscr{L}^1$.

(2) $\Rightarrow$ (3): By assumption (2), we have $X_n \to X_\infty$ in $\mathscr{L}^1$. Thus, $E(|X_n|) \to E(|X_\infty|) < \infty$, so Theorem D.6 applies, and we have that $X_n \to X_\infty$ a.e. Now, by the defining relation for a martingale, for all $\Lambda \in \psi_n$ and $n > n'$,

$$\int_\Lambda X_n \, dP = \int_\Lambda X_{n'} \, dP \tag{D.8}$$

The right side of the equation converges to $\int_\Lambda X_\infty \, dP$, by the $\mathscr{L}^1$-convergence of the sequence $\{X_n\}$. The resulting equality proves that $\{X_n, \psi_n\}_{n \in \bar{N}}$ is a martingale. Because $\mathscr{L}^1$ convergence implies convergence of expectations, all three conditions in (3) are proved.

(3) $\Rightarrow$ (1): By assumption (3), $\{X_n, \psi_n\}_{n \in \bar{N}}$ is a martingale. Therefore,

$$\int_{X_n^+ > \lambda} X_n^+ \, dP = \int_{X_\infty^+ > \lambda} X_\infty^+ \, dP, \tag{D.9}$$

which proves that $\{X_n^+\}_{n \in N}$ is a uniformly integrable sequence. But we can also write

$$\int_{-X_n^- > \lambda} -X_n^- \, dP = \int_{X_\infty^- > \lambda} X_\infty^- \, dP, \tag{D.10}$$

so $\{-X_n^-\}_{n \in N}$ is also uniformly integrable. Therefore, $X_n = X_n^+ + X_n^-$ is also uniformly integrable.

(3) $\Rightarrow$ (4): This is trivial, because we can substitute the X_∞ in (3) for Y in (4).

(4) $\Rightarrow$ (3): Assuming (4), Theorem D.2 applies; therefore, for all $n' > n$,

$$E(X_{n'} \mid \psi_n) = E[(Y \mid \psi_{n'}) \mid \psi_n] = E[Y \mid \psi_n] = X_n \tag{D.11}$$

So, by definition, $\{X_n, \psi_n; Y, \psi\}$ is a martingale. It follows that $\{|X_n|, \psi_n; |Y|, \psi\}$ is also a martingale. Therefore, for each $\lambda > 0$,

$$\int_{\{|X_n|>\lambda\}} |X_n| \, dP = \int_{\{|X_n|>\lambda\}} |Y| \, dP \tag{D.12}$$

Therefore,

$$P\{\omega : |X_n| > \lambda\} \leq \frac{1}{\lambda} E(|X_n|) = \frac{1}{\lambda} E(|Y|) \tag{D.13}$$

This proves (3), with Y substituted for X_∞. ■

There is a rather interesting corollary to Theorem D.0.7, known as Lévy's Zero-Or-One Law:

Corollary D.0.8: If $\Lambda \in \psi_\infty$, then $\lim_{n\to\infty} P(\Lambda \mid \psi_n) = 1_\Lambda$ a.e.

Appendix E. Famous (and Useful) Theorems

This appendix presents several well-known theorems and lemmas from measure and probability theory. In most cases, the proofs are omitted because they are readily found in any standard text on the subject.

Theorem E.1 (Chebyshev's Inequality): If ϕ is a strictly positive and increasing function on $(0, \infty)$, with $\phi(u) = \phi(-u)$, and if X is a random variable such that $E[X] < \infty$, then for each $u > 0$,

$$P[\,|X| \geq u\,] \leq \frac{E[\,\phi(X)\,]}{\phi(u)}$$

Proof:

$$E[\,\phi(X)\,] = \int_{\Omega} \phi(X)\, dP \geq \int_{\{\,|X| \geq u\,\}} \phi(X)\, dP \geq \phi(u)\, P[\,|X| \geq u\,] \qquad \blacksquare$$

Definition E.1: A function ϕ on an interval is **convex** if

$$\phi[\,px + (1-p)y\,] \leq p\,\phi[x] + (1-p)\phi[y]$$

for $p \in [\,0,1\,]$, and x and y in the interval of convexity.

Theorem E.2 (Jensen's Inequality): If ϕ is convex on an interval containing the range of X, and X and $\phi(X)$ are integrable random variables, then

3*EQ*0 0 $$\phi(E[X]) \leq E[\phi(X)]$$

Proof is given, for example, by Chung [50].

Lyapunov's inequality can be obtained from Jensen's inequality [50]; it is

$$E[\,|X|^r]^{1/r} \leq E[\,|X|^{r'}]^{1/r'}\ , \quad 1 < r < r' < \infty$$

Theorem E.3 (Hölder's Inequality): Suppose

$$\frac{1}{p} + \frac{1}{q} = 1\,, \quad p > 1\,, \quad q > 1.$$

Then $E[\,|XY|\,] \leq E^{1/p}[\,|X|^p]\,E^{1/q}[\,|X|^q]$.

Proof is given, for example, by Folland [51].

Schwarz's inequality is obtained from Hölder's inequality by substituting $p = q = 2$.

Theorem E.4 (Minkowski's Inequality): If $f, g \in \mathscr{L}^p$, and $1 \leq p < \infty$, then

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p\,.$$

Proof is given, for example, by Folland [52].

Theorem E.5 (The First Borel-Cantelli Lemma): For arbitrary events $\{E_n\}$,

$$\sum_{n=1}^{\infty} P(E_n) < \infty \quad \Rightarrow \quad P(E_n \ i.o.) = 0.$$

Proof is given, for example, by Billingsley [53].

Theorem E.6 (The Second Borel-Cantelli Lemma): If $\{E_n\}$ is an independent sequence

of events, $\sum_{n=1}^{\infty} P(E_n) = \infty \quad \Rightarrow \quad P(E_n \ i.o.) = 1.$

Proof is given, for example, by Billingsley [54].

Theorem E.7 (Monotone Convergence Theorem):

If $0 \leq f_n \uparrow f$ a.e., then $0 \leq \int f_n \, d\mu \uparrow \int f \, d\mu$.

Proof is given, for example, by Billingsley [55].

Theorem E.8 (Fatou's Lemma):

If $0 \leq f_n$ for all n, then $\int \liminf_n f_n \, d\mu \leq \liminf_n \int f_n \, d\mu$.

Proof is given, for example, by Billingsley [56].

Theorem E.9 (Lebesgue's Dominated Convergence Theorem):

If $f_n \leq g$ a.e., where g is integrable, and if $f_n \to f$ a.e.,

then f_n and f are integrable, and $\int f_n \, d\mu \to \int f \, d\mu$.

Proof is given, for example, by Billingsley [57].

The next theorem is not particularly famous; rather it is such a standard result that it is often included in textbooks as a homework problem. It is included here because it is referenced from another proof within this thesis. This theorem considers a sequence of integrals, in which an integrable function is integrated over sets from a sequence that has zero measure in the limit.

Theorem E.10: If $E[\,|X|\,] < \infty$ and $\lim\limits_{n\to\infty} P(\Lambda_n) = 0$,

then $\lim\limits_{n\to\infty} \int_{\Lambda_n} |X|\, dP = 0.$

Proof:

$$\begin{aligned}
\lim_{n\to\infty} \left| \int_{\Lambda_n} X\, dP \right| &\le \lim_{n\to\infty} \int_{\Lambda_n} |X|\, dP \\
&= \lim_{m\to\infty} \lim_{n\to\infty} \left[\int_{\Lambda_n \cap \{X \le m\}} |X|\, dP + \int_{\Lambda_n \cap \{X > m\}} |X|\, dP \right] \\
&\le \lim_{m\to\infty} \lim_{n\to\infty} \left[\int_{\Lambda_n} m \;\; dP + \int_{\{X>m\}} |X|\, dP \right] \\
&= \lim_{m\to\infty} \lim_{n\to\infty} \left[m \;\; P(\Lambda_n) + \sum_{k=m}^{\infty} \int_{A_k} |X|\, dP \right] \qquad \text{(E.1)} \\
&\qquad \text{where } A_k = \{\omega :\; k < |X(\omega)| \le k+1\,\} \\
&= \lim_{m\to\infty} \left[m \times (0) \times + \sum_{k=m}^{\infty} \int_{A_k} (k+1)\, dP \right] \\
&\le 2 \lim_{m\to\infty} \sum_{k=m}^{\infty} \int_{A_k} k\, dP
\end{aligned}$$

But

$$\infty > E[\,|X|\,] = \int_{\Omega} |X| \; dP = \sum_{k=0}^{\infty} \int_{A_k} |X| \; dP > \sum_{k=0}^{\infty} \int_{A_k} k \; dP \tag{E.2}$$

Hence,

$$\lim_{m\to\infty} \sum_{k=m}^{\infty} \int_{A_k} k \; dP = 0, \tag{E.3}$$

which implies

$$\begin{aligned} 0 = \lim_{n\to\infty} \sum_{k=n}^{\infty} \int_{A_k} (k+1) \; dP &\geq \lim_{n\to\infty} \sum_{k=n}^{\infty} \int_{A_k} |X| \; dP \\ \geq \lim_{n\to\infty} \left| \sum_{k=n}^{\infty} \int_{A_k} X \; dP \right| &\geq \lim_{n\to\infty} \left| \int_{\Lambda_n} X \; dP \right| \end{aligned} \tag{E.4}$$

Therefore,

$$\lim_{n\to\infty} \int_{\Lambda_n} X \; dP = 0. \qquad \blacksquare \tag{E.5}$$

Appendix F. Algebraic Expansion of a Term in Equation 5.2.23

This appendix shows the algebraic expansion of the conditional mean term on the right side of Equation 5.2.23, $E_{\psi_k}[G\,\eta_k\,\eta_k^T\,G^T]$. First, note that

$$
\begin{aligned}
G\,\eta_k &= \left[\begin{array}{c|c|c|c|c|c|c|c}
-\theta_A^1 & \cdots & -\theta_A^p & \theta_B^1 & \cdots & \theta_B^m & \mathbf{I}_n & \mathbf{0} \\
\mathbf{0} & & \mathbf{0} & \mathbf{0} & & \mathbf{0} & \mathbf{0} & \mathbf{I}_{(m+p)n}
\end{array}\right]
\begin{bmatrix} w_k \\ v_k \\ \Xi_k^x \\ \mathbf{0} \end{bmatrix} \\
&= \begin{bmatrix} \displaystyle\sum_{i=1}^{p} -\theta_A^i\,w_k^i + \sum_{i=1}^{m}\theta_B^i\,v_k^i + \Xi_k^x \\ \mathbf{0}_{(m+p)n\times 1} \end{bmatrix}.
\end{aligned}
\tag{F.1}
$$

Hence, $E_{\psi_k}\left[(G\,\eta_k)(G\,\eta_k)^T\right] = \left[\begin{array}{c|c} \bigstar_{(n\times n)} & \mathbf{0}_{n\times(m+p)n} \\ \mathbf{0}_{(m+p)n\times n} & \mathbf{0}_{(m+p)n\times(m+p)n} \end{array}\right],$ (F.2)

$$
\begin{aligned}
\text{where } \bigstar_{(n\times n)} = E_{\psi_k}\Bigg[& \sum_{i=1}^{p}\sum_{j=1}^{p} \theta_A^i(\theta_A^j)^T w_k^i\,w_k^j + \sum_{i=1}^{m}\sum_{j=1}^{m} \theta_B^i(\theta_B^j)^T v_k^i\,v_k^j \\
& -\sum_{i=1}^{p}\sum_{j=1}^{m} \theta_A^i(\theta_B^j)^T w_k^i\,v_k^j - \sum_{i=1}^{m}\sum_{j=1}^{p} \theta_B^i(\theta_A^j)^T v_k^i\,w_k^j \\
& -\sum_{i=1}^{p} \theta_A^i(\Xi_k^x)^T w_k^i + \sum_{i=1}^{m} \theta_B^i(\Xi_k^x)^T v_k^i \\
& -\sum_{j=1}^{p} \Xi_k^x(\theta_A^j)^T w_k^j + \sum_{j=1}^{m} \Xi_k^x(\theta_B^j)^T v_k^j + \Xi_k^x(\Xi_k^x)^T \Bigg].
\end{aligned}
\tag{F.3}
$$

Now, recall from Equation 5.1.6 that

$$Q_k^0 \triangleq E\{\eta_k \eta_k^T\} = \left[\begin{array}{ccc|c} R_k & S_{vw}^T(k) & S_{\Xi_{xw}}^T(k) & \\ S_{vw}(k) & Q_k & S_{\Xi_{xv}}^T(k) & \mathbf{0} \\ S_{\Xi_{xw}}(k) & S_{\Xi_{xv}}(k) & \Sigma_k^{xx} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right]. \tag{F.4}$$

Denote the columns $S_{\Xi_k v}(k) \triangleq \left[(S^k_{\Xi_k v})_1 \ldots (S^k_{\Xi_k v})_m\right]$ and $S_{\Xi_k w}(k) \triangleq \left[(S^k_{\Xi_k w})_1 \ldots (S^k_{\Xi_k w})_p\right]$. Also note that $E_{\psi_k}\left[\eta_k \eta_k^T\right] = E\left[\eta_k \eta_k^T\right]$, because the density function $f_{\psi_k}(\eta_k) = f(\eta_k)$.

Denoting $\hat{\theta}(k) \triangleq E_{\psi_k}(\theta)$, from Equation F.3,

$$\begin{aligned} \star_{(n\times n)} = \Bigg[& \sum_{i=1}^{p}\sum_{j=1}^{p} E_{\psi_k}\left[\theta_A^i\,(\theta_A^j)^T\right](R_k)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} E_{\psi_k}\left[\theta_B^i\,(\theta_B^j)^T\right](Q_k)_{i,j} \\ & - \sum_{i=1}^{p}\sum_{j=1}^{m} E_{\psi_k}\left[\theta_A^i\,(\theta_B^j)^T\right][S_{vw}^T(k)]_{i,j} - \sum_{i=1}^{m}\sum_{j=1}^{p} E_{\psi_k}\left[\theta_B^i\,(\theta_A^j)^T\right][S_{vw}(k)]_{i,j} \\ & - \sum_{i=1}^{p} \hat{\theta}_A^i(k)\,(S^k_{\Xi_k w})_i^T + \sum_{i=1}^{m} \hat{\theta}_B^i(k)\,(\Xi_k^x)^T (S^k_{\Xi_k v})_i^T \\ & - \sum_{j=1}^{p} (S^k_{\Xi_k w})_j\,(\hat{\theta}_A^j(k))^T + \sum_{j=1}^{m} (S^k_{\Xi_k v})_j\,(\hat{\theta}_B^j(k))^T + \Sigma_k^{xx} \Bigg]. \end{aligned} \tag{F.5}$$

Now note that

$$\begin{aligned} E_{\psi_k}\left[(\hat{\theta}(k)-\theta)(\hat{\theta}(k)-\theta)^T\right] &= E_{\psi_k}\left[\hat{\theta}(k)\,\hat{\theta}^T(k) - \theta\,\hat{\theta}^T(k) - \hat{\theta}(k)\,\theta^T + \theta\,\theta^T\right] \\ &= \hat{\theta}(k)\,\hat{\theta}^T(k) - E_{\psi_k}(\theta)\,\hat{\theta}^T(k) - \hat{\theta}(k)\,E_{\psi_k}(\theta^T) + E_{\psi_k}(\theta\,\theta^T) \\ &= \hat{\theta}(k)\,\hat{\theta}^T(k) - \hat{\theta}(k)\,\hat{\theta}^T(k) - \hat{\theta}(k)\,\hat{\theta}^T(k) + E_{\psi_k}(\theta\,\theta^T). \end{aligned} \tag{F.6}$$

Hence,

$$E_{\psi_k}(\theta\,\theta^T) \;=\; E_{\psi_k}\Big[(\hat{\theta}(k)-\theta)(\hat{\theta}(k)-\theta)^T\Big] + \hat{\theta}(k)\,\hat{\theta}^T(k) \;\triangleq\; P^{\theta\theta} + \hat{\theta}(k)\,\hat{\theta}^T(k). \tag{F.7}$$

Substituting Equation F.7 into Equation F.5,

$$\begin{aligned}
\bigstar_{(n\times n)} = \Bigg[& \sum_{i=1}^{p}\sum_{j=1}^{p} P^{\theta_A^i\,\theta_A^j}\,(R_k)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} P^{\theta_B^i\,\theta_B^j}\,(Q_k)_{i,j} \\
& - \sum_{i=1}^{p}\sum_{j=1}^{m} P^{\theta_A^i\,\theta_B^j}\,[S_{vw}^T(k)]_{i,j} - \sum_{i=1}^{m}\sum_{j=1}^{p} P^{\theta_B^i\,\theta_A^j}\,[S_{vw}(k)]_{i,j} \\
& + \sum_{i=1}^{p}\sum_{j=1}^{p} \hat{\theta}_A^i(k)\,[\,\hat{\theta}_A^j(k)\,]^T\,(R_k)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{\theta}_B^i(k)\,[\,\hat{\theta}_B^j(k)\,]^T\,(Q_k)_{i,j} \\
& - \sum_{i=1}^{p}\sum_{j=1}^{m} \hat{\theta}_A^i(k)\,[\,\hat{\theta}_B^j(k)\,]^T\,[S_{vw}^T(k)]_{i,j} - \sum_{i=1}^{m}\sum_{j=1}^{p} \hat{\theta}_B^i(k)\,[\,\hat{\theta}_A^j(k)\,]^T\,[S_{vw}(k)]_{i,j} \\
& - \sum_{i=1}^{p} \hat{\theta}_A^i(k)\,(S_{\Xi_k w}^k)_i^T + \sum_{i=1}^{m} \hat{\theta}_B^i(k)\,(\,\Xi_k^x\,)^T\,(S_{\Xi_k v}^k)_i^T \\
& - \sum_{j=1}^{p} (S_{\Xi_k w}^k)_j\,(\,\hat{\theta}_A^j(k)\,)^T + \sum_{j=1}^{m} (S_{\Xi_k v}^k)_j\,(\,\hat{\theta}_B^j(k)\,)^T + \Sigma_k^{xx} \Bigg].
\end{aligned} \tag{F.8}$$

Now, the expansion of $E_{\psi_k}(G)\,Q_k^0\,E_{\psi_k}(G^T)$ will be developed, in order to simplify Equation F.8.

$$E_{\psi_k}(G)\,Q_k^0\,E_{\psi_k}(G^T) = \left[\begin{array}{c|c} E_{\psi_k}(G^x)\,Q_k^0\,E_{\psi_k}[(G^x)^T] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array}\right], \tag{F.9}$$

where $E_{\psi_k}(G^x)\, Q_k^0\, E_{\psi_k}[(G^x)^T] =$

$$\left[\, -\hat{\theta}_A^1(k) \,|\ldots|\, -\hat{\theta}_A^p(k) \,|\, \hat{\theta}_B^1(k) \,|\ldots \hat{\theta}_B^m(k) \,|\, \mathbf{I}_n \,|\, \mathbf{0} \,\right] Q_k^0 \begin{bmatrix} [-\hat{\theta}_A^1(k)]^T \\ \vdots \\ [-\hat{\theta}_A^p(k)]^T \\ [\hat{\theta}_B^1(k)]^T \\ \vdots \\ [\hat{\theta}_B^m(k)]^T \\ \mathbf{I}_n \\ \mathbf{0} \end{bmatrix} \tag{F.10}$$

Now, the j^{th} column of the product $E_{\psi_k}(G^x)\, Q_k^0$, for all $j \in \{1, \ldots, (m+p+1)n\}$, is

$$-\sum_{i=1}^{p} \hat{\theta}_A^i(k)\, [Q_k^0]_{i,j} + \sum_{i=1}^{m} \hat{\theta}_B^i(k)\, [Q_k^0]_{p+i,j} + \sum_{i=1}^{n} \mathbf{e}_i\, [Q_k^0]_{p+m+i,j}\,, \tag{F.11}$$

where $\mathbf{e}_i$ is the i^{th} unit vector in $\mathbb{R}^n$, for all $i \in \{1, \ldots, n\}$.

Substituting the column description of Equation F.11 into Equation F.10,

$$
\begin{aligned}
&E_{\psi_k}(G^x)\,Q_k^0\,E_{\psi_k}[(G^x)^T] \\
&\quad = \sum_{i=1}^{p}\sum_{j=1}^{p} \hat{\theta}_A^i(k)\,[\hat{\theta}_A^j(k)]^T[Q_k^0]_{i,j} + \sum_{i=1}^{p}\sum_{j=1}^{m} \hat{\theta}_A^i(k)\,[\hat{\theta}_B^j(k)]^T[Q_k^0]_{i,p+j} \\
&\qquad\qquad + \sum_{i=1}^{p}\sum_{j=1}^{n} \hat{\theta}_A^i(k)\,\mathbf{e}_j^T[Q_k^0]_{i,p+m+j} \\
&\qquad + \sum_{i=1}^{m}\sum_{j=1}^{p} \hat{\theta}_B^i(k)\,[\hat{\theta}_A^j(k)]^T[Q_k^0]_{p+i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{\theta}_B^i(k)\,[\hat{\theta}_B^j(k)]^T[Q_k^0]_{p+i,p+j} \\
&\qquad\qquad + \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{\theta}_B^i(k)\,\mathbf{e}_j^T[Q_k^0]_{p+i,p+m+j} \\
&\qquad + \sum_{i=1}^{n}\sum_{j=1}^{p} \mathbf{e}_i\,[\hat{\theta}_A^j(k)]^T[Q_k^0]_{p+m+i,j} + \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{e}_i\,[\hat{\theta}_B^j(k)]^T[Q_k^0]_{p+m+i,p+j} \\
&\qquad\qquad + \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{e}_i\,\mathbf{e}_j^T[Q_k^0]_{p+m+i,p+m+j}\,.
\end{aligned}
\tag{F.12}
$$

Substituting the submatrices of Q_k^0 (from Equation F.4) into Equation F.12,

$$
\begin{aligned}
&E_{\psi_k}(G^x)\,Q_k^0\,E_{\psi_k}[(G^x)^T] \\
&\quad = \sum_{i=1}^{p}\sum_{j=1}^{p} \hat{\theta}_A^i(k)\,[\hat{\theta}_A^j(k)]^T[R_k]_{i,j} + \sum_{i=1}^{p}\sum_{j=1}^{m} \hat{\theta}_A^i(k)\,[\hat{\theta}_B^j(k)]^T[S_{vw}^T(k)]_{i,j} \\
&\qquad\qquad + \sum_{i=1}^{p}\sum_{j=1}^{n} \hat{\theta}_A^i(k)\,\mathbf{e}_j^T(S_{\Xi_k w}^k)_i^T \\
&\qquad + \sum_{i=1}^{m}\sum_{j=1}^{p} \hat{\theta}_B^i(k)\,[\hat{\theta}_A^j(k)]^T[S_{vw}(k)]_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{\theta}_B^i(k)\,[\hat{\theta}_B^j(k)]^T[Q_k]_{i,j} \\
&\qquad\qquad + \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{\theta}_B^i(k)\,\mathbf{e}_j^T(S_{\Xi_k v}^k)_i^T \\
&\qquad + \sum_{j=1}^{p} (S_{\Xi_k w}^k)_j[\hat{\theta}_A^j(k)]^T + \sum_{j=1}^{m} (S_{\Xi_k v}^k)_j[\hat{\theta}_B^j(k)]^T \; + \; \Sigma_k^{xx}\,.
\end{aligned}
\tag{F.13}
$$

Finally, substituting Equation F.13 into Equation F.8,

$$\begin{aligned}\star_{(n\times n)} = &\sum_{i=1}^{p}\sum_{j=1}^{p} P^{\theta_A^i\,\theta_A^j}(R_k)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} P^{\theta_B^i\,\theta_B^j}(Q_k)_{i,j} \\ &-\sum_{i=1}^{p}\sum_{j=1}^{m} P^{\theta_A^i\,\theta_B^j}[S_{vw}^T(k)]_{i,j} - \sum_{i=1}^{m}\sum_{j=1}^{p} P^{\theta_B^i\,\theta_A^j}[S_{vw}(k)]_{i,j} \\ &+ E_{\psi_k}(G^x)\,Q_k^0\,E_{\psi_k}[(G^x)^T].\end{aligned} \tag{F.14}$$

Using the partitioning of the matrix $P_{k+1|k}^{\theta\theta}$ given in Equation 5.3.10, and the definition of Q_k^0 given in Equation F.4, Equation F.14 can be rewritten as

$$\begin{aligned}\star_{(n\times n)} = &\sum_{i=1}^{p}\sum_{j=1}^{p} P_{i,j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{i,j} + \sum_{i=1}^{m}\sum_{j=1}^{m} P_{p+i,p+j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{p+i,p+j} \\ &-\sum_{i=1}^{p}\sum_{j=1}^{m} P_{i,p+j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{i,p+j} - \sum_{i=1}^{m}\sum_{j=1}^{p} P_{p+i,j}^{\theta\theta}(k+1|k)\,(Q_k^0)_{p+i,j} \\ &+ E_{\psi_k}(G^x)\,Q_k^0\,E_{\psi_k}[(G^x)^T].\end{aligned} \tag{F.15}$$

References

[1] H. W. Sorenson, "Least-Squares Estimation: from Gauss to Kalman," *IEEE Spectrum*, Vol. 7, No. 7, July 1970, pp. 63-68.

[2] Karl Friedrich Gauss, *Theory of Motion of the Heavenly Bodies*, New York: Dover Publications, 1963, p. 249.

[3] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, March 1960, pp. 35-45.

[4] W. Jakoby and M. Pandit, "Prediction-error-method for Recursive Identification of Nonlinear Systems," *Automatica*, Vol. 23, No. 4, July 1987, pp. 491-496.

[5] G. Salut, J. Aguilar-Martin, and S. Lefebvre, "New Results on Optimal Joint Parameter and State Estimation of Linear Stochastic Systems," *ASME J. of Dynamic Systems, Measurement and Control*, Vol. 102, March 1980, pp. 28-34.

[6] Gerard Salut, "Identifiabilité d'un système dynamique linéaire invariant," *Comptes Rendus Hebdomadaires des Seances de L'Academie des Sciences, Serie A*, t.278, pp. 181-184, 1974.

[7] Y. C. Chen, J. D. Aplevich, and W. J. Wilson, "Simultaneous Estimation of State Variables and Parameters for Multivariable Linear Systems with Singular Pencil Models," *IEE Proceedings*, Vol. 133, Part D, No. 2, March 1986, pp. 65-72.

[8] H. F. VanLandingham and M. A. Hopkins, "Deadbeat Parameter Identification of DARMA Processes," *Proceedings of the 24th Annual Allerton Conference on Communication, Control, and Computing*, University of Illinois at Urbana-Champaign, October 1986.

[9] Lennart Ljung and Torsten Söderström, *Theory and Practice of Recursive Identification*, Cambridge MA: The MIT Press, 1983, Chapter 2.

[10] N. K. Sinha and B. Kuszta, *Modeling and Identification of Dynamic Systems*, New York: Van Nostrand Reinhold, 1983, Chapter 4.

[11] H. Robbins and S. Monro, "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, Vol. 22, pp. 400-407, 1951.

[12] E. G. Gladysev, "On Stochastic Approximation," *Theory of Probability and Its Applications*, Vol. 10, pp. 275-278, 1965.

[13] Lennart Ljung and Torsten Söderström, *op.cit.*, pp. 33-36.

[14] *ibid.*, pp. 51-54.

[15] Lawrence W. Nelson and Edwin Stear, "The Simultaneous On-Line Estimation of Parameters and States in Linear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-21, February 1976, pp. 94-98.

[16] R. A. Padilla, C. S. Padilla, and S. P. Bingulac, "Comments on 'The Simultaneous On-Line Estimation of Parameters and States in Linear Systems'," *IEEE Transactions on Automatic Control*, Vol. AC-23, February 1978, pp. 96-97.

[17] Robert J. Fitzgerald, "Divergence of the Kalman Filter," *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 6, pp. 736-747, December 1971.

[18] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*, Englewood Cliffs NJ: Prentice-Hall, Inc., 1984, pp. 362-366.

[19] C. T. Chen, *Introduction to Linear System Theory*, New York: Holt, Rinehart and Winston, Inc., 1970, p. 294.

[20] Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes, Second Edition*, New York: McGraw-Hill Book Company, 1984, p. 144.

[21] Avner Friedman, *Stochastic Differential Equations and Applications, Volume 1*, New York: Academic Press, 1975, p.53.

[22] B. D. O. Anderson, and J. B. Moore, *Optimal Filtering*, Englewood Cliffs NJ: Prentice-Hall, Inc., 1979, p. 94.

[23] *ibid.*, pp. 26-27.

[24] D. G. Luenberger, "An Introduction to Observers," *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 6, pp. 596-602, 1971.

[25] Robert R. Bitmead and Brian D. O. Anderson, "Exponentially Convergent Behavior of Simple Stochastic Adaptive Estimation Algorithms," *Proceedings of the 17th IEEE Conference on Decision and Control*, pp. 580-585, December 1978.

[26] Richard M. Johnstone, C. Richard Johnson, Jr., Robert R. Bitmead, and Brian D. O. Anderson, "Exponential Convergence of Recursive Least Squares with Exponential Forgetting Factor," *Proceedings of the 21st IEEE Conference on Decision and Control*, pp. 994-997, December 1982.

[27] Harold J. Kushner, "Convergence of Recursive Adaptive and Identification Procedures Via Weak Convergence Theory," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 6, pp 921-930, December 1977.

[28] Lennart Ljung, "Analysis of Recursive Stochastic Algorithms," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 4, pp. 551-575, August 1977.

[29] Kai Lai Chung, *A Course in Probability Theory*, New York: Harcourt, Brace & World, 1968, pp. 312-313.

[30] Jan Sternby, "On Consistency for the Method of Least Squares Using Martingale Theory," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 3, June 1977, pp. 346-352.

[31] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, New York: Academic Press,Inc., 1977, pp. 124-129.

[32] *ibid.*, pp. 44-47.

[33] Karl Johan Åström, "Theory and Applications of Adaptive Control -- A Survey," *Automatica*, Vol. 19, No. 5, pp 471-486, 1983.

[34] Ya. Z. Tsypkin, "Adaptation, training and self-organization in automatic systems," *Automation and Remote Control*, Vol. 27, pp. 16-51, 1966.

[35] R. E. Kalman, "Design of a Self-Optimizing Control System," *Transactions of the ASME*, Vol. 80, pp. 468-478, 1958.

[36] Bjorn Wittenmark, "Stochastic adaptive control methods: a survey," *International Journal of Control*, Vol. 21, No. 5, pp. 705-730, 1975.

[37] Graham C. Goodwin and Kwai Sang Sin, *op.cit.*, p. 180.

[38] Brian D. O. Anderson and John B. Moore, *op.cit.*, pp. 131-132.

[39] T. Nishimura, "On the *a Priori* Information in Sequential Estimation Problems," *IEEE Transactions on Automatic Control*, Vol. AC-11, No. 2, April 1966, pp. 197-204, and Vol. AC-12, Vo. 1, February 1967, p. 123.

[40] H. Heffes, "The Effect of Erroneous Models on the Kalman Filter Responses," *IEEE Transactions on Automatic Control*, Vol. AC-11, No. 3, July 1966, pp. 541-543.

[41] C. J. Harris and S. A. Billings, eds., *Self-Tuning and Adaptive Control: Theory and Applications*, New York: Peter Peregrinus Ltd., 1981.

[42] Hugh F. VanLandingham, *Introduction to Digital Control Systems*, New York: Macmillan Publishing Company, 1985, pp. 359-363.

[43] Russell S. Kemp, *Pseudo-Linear Identification and Its Application to Adaptive Control*, M. S. Thesis, Virginia Polytechnic Institute and State University, March 1987.

[44] Kai Lai Chung, *op.cit.*, pp. 90-91.

[45] J. Neveu, author, and T. P. Speed, translator, *Discrete-Parameter Martingales*, New York: North-Holland Publishing Co., 1975, pp. 11-12, p.21.

[46] *ibid.*, p.22.

[47] Kai Lai Chung, *op.cit.*, pp. 298-300.

[48] J. Neveu, author, and T. P. Speed, translator, *op.cit.*, pp. 11-12.

[49] *ibid.*, pp. 304-305.

[50] *ibid.*, pp. 45-46.

[51] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, New York: John Wiley and Sons, Inc., 1984, pp. 174-175.

[52] *ibid.*, p. 175.

[53] Patrick Billingsley, *Probability and Measure*, New York: John Wiley and Sons, Inc., 1986, p.53.

[54] *ibid.*, p. 55.

[55] *ibid.*, p. 211.

[56] *ibid.*, p. 212.

[57] *ibid.*, p. 213.

Bibliography

Albert, Arthur E., and Leland A. Gardner, Jr., *Stochastic Approximation and Nonlinear Regression*, Cambridge MA: The MIT Press, 1967.

Alengrin, G., and G. Favier, "Algorithmes de Réalisation Stochastique pour l'Estimation du Gain Stationnaire du Filtre de Kalman dans le Cas de Systèmes Multivariables," *RAIRO Automatique Systems Analysis and Control*, Vol. 15, No. 1, pp. 19-30, 1981.

Anderson, Brian D. O., "Exponential Data Weighting in the Kalman-Bucy Filter," *Information Sciences*, Vol. 5, pp. 217-230, 1973.

Anderson, Brian D. O., "Exponential Stability of Linear Equations Arising in Adaptive Identification," *IEEE Transactions on Automatic Control*, Vol. AC-22, No.1, pp. 83-88, February 1977.

Anderson, Brian D. O., "Exponential Convergence and Persistent Excitation," *Proceedings of the 21st IEEE Conference on Decision and Control*, pp. 12-17, December 1982.

Anderson, Brian D. O., and C. Richard Johnson, Jr., "Exponential Convergence of Adaptive Identification and Control Algorithms," *Automatica*, Vol. 18, No. 1, pp. 1-13, 1982.

Anderson, Brian D.O., and John B. Moore, *Optimal Filtering*, Englewood Cliffs NJ: Prentice-Hall, Inc., 1979.

Anderson, Brian D.O., and John B. Moore, "The Kalman-Bucy Filter as a True Time-Varying Wiener Filter," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-1, No. 2, pp. 119-128, April, 1981.

Aplevich, J. D., "A System Representation for Control, Digital Signal Processing, and Estimation," *Proceedings of the IEEE*, Vol. 67, No. 11, pp. 1557-1559, November 1979.

Aplevich, J. D., "Time-Domain Input-Output Representations of Linear Systems," *Automatica*, Vol. 17, pp. 509-522, 1981.

Åström, Karl Johan, "Maximum Likelihood and Prediction Error Methods," *Automatica*, Vol. 16, pp. 551-574, 1980.

Åström, Karl Johan, and P. Eykhoff, "System Identification -- A Survey," *Automatica*, Vol. 7, pp. 123-162, 1971.

Balakrishnan, A. V., "A Martingale Approach to Linear Recursive State Estimation," *SIAM Journal of Control*, Vol. 10, No. 4, pp. 754-766, November 1972.

Balakrishnan, A. V., *Kalman Filtering Theory*, New York: Optimization Software, Inc., 1984.

Balakrishnan, A. V., and V. Peterka, "Identification in Automatic Control," *Automatica*, Vol. 5, pp. 817-829, 1969.

Bartle, Robert G., *The Elements of Real Analysis, Second Edition*, New York: John Wiley & Sons, 1964.

Bazaraa, Mokhtar S., and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, New York: John Wiley and Sons, 1979.

Bierman, Gerald J., *Factorization Methods for Discrete Sequential Estimation*, New York: Academic Press, Inc., 1977.

Billingsley, Patrick, *Probability and Measure, Second Edition*, New York: John Wiley & Sons, 1986.

Bitmead, Robert R., and Brian D. O. Anderson, "Exponentially Convergent Behavior of Simple Stochastic Adaptive Estimation Algorithms," *Proceedings of the 17th IEEE Conference on Decision and Control*, pp. 580-585, December 1978.

Bitmead, Robert R., and Brian D. O. Anderson, "Performance of Adaptive Estimation Algorithms in Dependent Random Environments," *IEEE Transactions on Automatic Control*, Vol. AC-25, No. 4, pp. 788-794, August 1980.

Bitmead, Robert R., and Brian D. O. Anderson, "Lyapunov Techniques for the Exponential Stability of Linear Difference Equations with Random Coefficients," *IEEE Transactions on Automatic Control*, Vol. AC-25, No. 4, pp. 788-794, August 1980.

Chen, Y. C., J. D. Aplevich, and W. J. Wilson, "Simultaneous Estimation of State Variables and Parameters for Multivariable Linear Systems with Singular Pencil Models," *IEE Proceedings*, Vol. 133, Part D, No. 2, pp. 65-72, March 1986.

Chung, Kai Lai, *A Course in Probability Theory*, New York: Harcourt, Brace & World, Inc., 1968.

de Larminat, Phillipe, and Christian Doncarli, "Une méthode d'identification récursive des systèmes stochastiques linéaires multivariables," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences, Serie A*, t. 284, pp. 1405-1408, June 1977.

Denham, Michael J., "Canonical Forms for the Identification of Multivariable Linear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, pp 646-656, December 1974.

Dickenson, B. W., T. Kailath, and M. Morf, "Canonical Matrix Fraction and State-Space Descriptions for Deterministic and Stochastic Linear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, pp. 656-667, December 1974.

Doob, J. L., *Stochastic Processes*, New York: John Wiley & Sons, Inc., 1953.

Doob, J. L., "What is a Martingale?" *American Mathematical Monthly*, Vol. 78, pp. 451-463, 1971.

Eweda, Eweda, and Odile Macchi, "Convergence of an Adaptive Linear Estimation Algorithm," *IEEE Transactions on Automatic Control*, Vol. AC-29, No. 2, pp. 119-127, February 1984.

Eykhoff, Pieter, "Process Parameter and State Identification," *Automatica*, Vol. 4, pp. 205-233, 1968.

Eykhoff, Pieter, *System Identification: Parameter and State Estimation*, New York: John Wiley and Sons, 1974.

Fasol, K. H., and H. P. Jorgl, "Principles of Model Building and Identification," *Automatica*, Vol. 16, pp. 505-518, 1980.

Fitzgerald, Robert J., "Divergence of the Kalman Filter," *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 6, pp. 736-747, December 1971.

Folland, Gerald B., *Real Analysis: Modern Techniques and their Applications*, New York: John Wiley & Sons, 1984.

Friedlander, Benjamin, "System Identification Techniques for Adaptive Signal Processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 2, pp. 240-246, April 1982.

Friedman, Avner, *Stochastic Differential Equations and Applications, Volume 1*, New York: Academic Press, 1975.

Gauss, Karl Friedrich, *Theory of Motion of the Heavenly Bodies*, New York: Dover Publications, 1963.

Gelb, Arthur, ed., *Applied Optimal Estimation*, Cambridge MA: The MIT Press, 1974.

Gladysev, E. G., "On Stochastic Approximation," *Theory of Probability and Its Applications*, Vol. 10, pp. 275-278, 1965.

Glover, Keith, and Jan C. Williams, "Parametrizations of Linear Dynamical Systems: Canonical Forms and Identifiability," *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, pp. 640-645, December 1974.

Godfrey, K. R., "Correlation Methods," *Automatica*, Vol. 16, pp. 527-534, 1980.

Golub, Gene H., and Charles F. Van Loan, *Matrix Computations*, Baltimore MD: The Johns Hopkins University Press, 1983.

Goodwin, Graham C., and Robert L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*, New York: Academic Press, 1977.

Goodwin, Graham C., Kwai Sang Sin, *Adaptive Filtering, Prediction and Control*, Englewood Cliffs NJ: Prentice-Hall, Inc., 1984.

Goodwin, Graham C., and Eam Khwang Teoh, "Persistency of Excitation in the Presence of Possibly Unbounded Signals," *IEEE Transactions on Automatic Control*, Vol. AC-30, No. 6, pp. 595-597, June 1985.

Graupe, Daniel, and Eli Fogel, "Convergence of Sequential Algorithms for Identifying Stable and Unstable Processes," *Proceedings of the 6th Symposium on Nonlinear Estimation*, San Francisco CA, September, 1975.

Graupe, Daniel, and Eli Fogel, "A Unified Sequential Identification Structure Based on Convergence Considerations," *Automatica*, Vol. 12, pp. 53-59, 1976.

Graupe, Daniel, *Identification of Systems*, Huntington NY: Robert E. Krieger Publishing Company, 1976.

Hawkes, Richard M., and John B. Moore, "Performance Bounds for Adaptive Estimation," *Proceedings of the IEEE*, Vol. 64, No. 8, pp. 1143-1150, August 1976.

Heffes, H., "The Effect of Erroneous Models on the Kalman Filter Response," *IEEE Transactions on Automatic Control*, Vol. AC-11, No. 3, pp. 541-543, July 1966.

Hopkins, Mark A., and Hugh F. VanLandingham, "Optimal Joint Parameter and State Estimation of Linear Stochastic MIMO Systems Using Pseudo-Linear Identification (PLID)," *Proceedings of the 25th Annual Allerton Conference on Communication, Computing, and Control*, University of Illinois at Urbana-Champaign, October 1987.

Ioannou, Petros A., and Petar V. Kokotovic, "An Asymptotic Error Analysis of Identifiers and Adaptive Observers in the Presence of Parasitics," *IEEE Transactions on Automatic Control*, Vol. AC-27, No. 4, pp. 921-927, August 1982.

Isermann, Rolf, "Practical Aspects of Process Identification," *Automatica*, Vol. 16, pp. 575-587, 1980.

Jakoby, W., and M. Pandit, "A Prediction-error-method for Recursive Identification of Nonlinear Systems," *Automatica*, Vol. 23, No. 4, pp. 491-496, July 1987.

Jazwinski, Andrew H., *Stochastic Processes and Filtering Theory*, New York: Academic Press, 1970.

Johnstone, Richard M., C. Richard Johnson, Jr., Robert R. Bitmead, and Brian D. O. Anderson, "Exponential Convergence of Recursive Least Squares with Exponential Forgetting Factor," *Proceedings of the 21st IEEE Conference on Decision and Control*, pp. 994-997, December 1982.

Kailath, Thomas, editor, *Linear Least-Squares Estimation*, Stroudsburg PA: Dowden, Hutchinson & Ross, Inc., 1977.

Kalman, R. E., "Design of a Self-Optimizing Control System," *Transactions of the ASME*, Vol. 80, pp. 468-478, 1958.

Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, pp. 35-45, March 1960.

Kaminsky, Paul G., Arthur E. Bryson, Jr., and Stanley F. Schmidt, "Discrete Square Root Filtering: A Survey of Current Techniques," *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 6, pp. 727-735, December 1971.

Karlin, Samuel, and Howard M. Taylor, *A First Course in Stochastic Processes, Second Edition*, New York: Academic Press, 1975.

Kaufman, Howard, and Daniel Beaulier, "Adaptive Parameter Identification," *IEEE Transactions on Automatic Control*, Vol. AC-17, pp. 729-731, October 1972.

Kazakos, Dimitri, "New Convergence Bounds for Bayes Estimators," *IEEE Transactions on Information Theory*, Vol. IT-27, No. 1, pp. 97-104, January 1981.

Kemp, Russell S., *Pseudo-Linear Identification and Its Application to Adaptive Control*, M. S. Thesis, Virginia Polytechnic Institute and State University, March 1987.

Kopp, P. E., *Martingales and Stochastic Integrals*, New York: Cambridge University Press, 1984.

Krishnan, Venkatarama, *Nonlinear Filtering and Smoothing: An Introduction to Martingales, Stochastic Integrals and Estimation*, New York: John Wiley & Sons, 1984.

Kudva, Prabhakar, and Kumpati S. Narendra, "The Discrete Adaptive Observer," *Proceedings of the 13th IEEE Conference on Decision and Control*, Phoenix AZ, pp. 307-312, December 1974.

Kumar, Rajendra, and John B. Moore, "Convergence of Adaptive Minimum Variance Algorithms via Weighting Coefficient Selection," *IEEE Transactions on Automatic Control*, Vol. AC-27, No. 1, pp. 146-153, February 1982.

Kushner, Harold J., *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, Cambridge MA: The MIT Press, 1984.

Kushner, Harold J., "Convergence of Recursive Adaptive and Identification Procedures Via Weak Convergence Theory," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 6, pp 921-930, December 1977.

Lainiotis, Demetrios G., "Partitioning: A Unifying Framework for Adaptive Systems, I: Estimation," *Proceedings of the IEEE*, Vol. 64, No. 8, pp. 1126-1143, August 1976.

Landau, Ioan D., "Martingale Convergence Analysis of Adaptive Schemes -- A Feedback Approach," *IEEE Transactions on Automatic Control*, Vol. AC-27, No. 3, pp. 716-719, June 1982.

Lawrence, Dale A., and C. Richard Johnson, Jr., "Recursive Parameter Identification Algorithm Stability Analysis Via Pi-Sharing," *IEEE Transactions on Automatic Control*, Vol. AC-31, No. 1, pp. 16-24, January 1986.

Liptser, R. S., and A. N. Shiryayev, *Statistics of Random Processes II: Applications*, New York: Springer-Verlag, 1974.

Ljung, Lennart, "Consistency of the Least-Squares Identification Method," *IEEE Transactions on Automatic Control*, Vol. AC-21, No. 5, pp. 779-781, October 1976.

Ljung, Lennart, "Analysis of Recursive Stochastic Algorithms," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 4, pp. 551-575, August 1977.

Ljung, Lennart, "On Positive Real Transfer Functions and the Convergence of Some Recursive Schemes," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 4, pp. 539-551, August 1977.

Ljung, Lennart, "Strong Convergence of a Stochastic Approximation Algorithm," *The Annals of Statistics*, Vol. 6, No. 3, pp. 680-696, 1978.

Ljung, Lennart, "Convergence Analysis of Parametric Identification Methods," *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 5, pp. 770-783, October 1978.

Ljung, Lennart, "Recursive Identification Techniques," *Mathematical Learning Models -- Theory and Algorithms: Proceedings of a Conference*, pp. 126-137, New York: Springer-Verlag, 1983.

Ljung, Lennart, "On the Estimation of Transfer Functions," *Automatica*, Vol. 21, No. 6, pp. 677-696, 1985.

Ljung, Lennart, "Asymptotic Variance Expressions for Identified Black-Box Transfer Function Models," *IEEE Transactions on Automatic Control*, Vol. AC-30, No. 9, pp. 834-844, September 1985.

Ljung, Lennart, and Torsten Söderström, *Theory and Practice of Recursive Identification*, Cambridge MA: The MIT Press, 1983.

Ljung, Lennart, and Zhen-Dong Yuan, "Asymptotic Properties of Black-Box Identification of Transfer Functions," *IEEE Transactions on Automatic Control*, Vol. AC-30, No. 6, pp. 514-530, June 1985.

Lozano-Leal, R., "Convergence Analysis of Recursive Identification Algorithms with Forgetting Factor," *Automatica*, Vol. 19, No. 1, pp. 95-97, January 1983.

Lüders, Gerd, and Kumpati S. Narendra, "A New Canonical Form for an Adaptive Observer," *IEEE Transactions on Automatic Control*, Vol. AC-19, pp. 117-119, April 1974.

Luenberger, David G., "An Introduction to Observers," *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 6, December 1971.

Maybeck, Peter S., *Stochastic Models, Estimation, and Control, Volume 2*, New York: Academic Press, 1982.

Mayne, D. Q., "A Canonical Model for Identification of Multivariable Linear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-17, pp. 728-729, October 1972.

Mehra, Raman K., "Optimal Input Signals for Parameter Estimation in Dynamic Systems -- Survey and New Results," *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, pp. 753-768, December 1974.

Mendel, Jerry M., *Discrete Techniques of Parameter Estimation: The Equation Error Formulation*, New York: Marcel Dekker, Inc., 1973.

Moore, John B., "On Strong Consistency of Least Squares Identification Algorithms," *Automatica*, Vol. 14, pp. 505-509, 1978.

Nelson, Lawrence W., and Edwin Stear, "The Simultaneous On-Line Estimation of Parameters and States in Linear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-21, No. 1, pp. 94-98, February, 1976. (Also see Padilla, R. A.)

Nevel'son, N. B., and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*, Providence RI: American Mathematical Society, 1973.

Neveu, J., and T. P. Speed (translator), *Discrete-Parameter Martingales*, New York: North-Holland Publishing Co., 1975.

Nishimura, T., "On the a priori Information in Sequential Estimation Problems," *IEEE Transactions on Automatic Control*, Vol. AC-11, No. 2, pp. 197-204, April 1966. (Also see: correction/extension from Vol. AC-12, No. 1, p. 123, February 1967.)

Padilla, R. A., C. S. Padilla, and S. P. Bingulac, "Comments on 'The Simultaneous On-Line Estimation of Parameters and States in Linear Systems'," *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 1, pp. 96-97, February 1978. (Attached to Nelson paper.)

Papoulis, Athanasios, *Probability, Random Variables, and Stochastic Processes, Second Edition*, New York: McGraw-Hill Book Company, 1984.

Robbins, Herbert, and Sutton Monro, "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, Vol. 22, pp. 400-407, 1951.

Ruckebusch, Guy, "Théorie géométrique de la Représentation Markovienne," *Annales de L'Institut Henri Poincaré, Section B: Calcul des Probabilités et Statistique*, Vol. XVI, No. 3, pp.225-297, 1980.

Sage, A. P., and C. D. Wakefield, "Maximum likelihood identification of time varying and random system parameters," *International Journal of Control*, Vol. 16, No. 1, pp. 81-100, 1972.

Salut, Gerard, "Identifiabilité d'un système dynamique linéaire invariant," *Comptes Rendus Hebdomadaires des Séances de L'Academie des Sciences, Serie A*, t. 278, pp. 181-184, January 14, 1974.

Salut, Gerard, J. Aguilar-Martin, and S. Lefebvre, "New Results on Optimal Joint Parameter and State Estimation of Linear Stochastic Systems," *Transactions of the ASME*, Vol. 102, pp. 28-34, March 1980.

Schlee, F. H., C. J. Standish, and N. F. Toda, "Divergence in the Kalman Filter," *AIAA Journal*, Vol. 5, No. 6, pp. 1114-1120, June 1967.

Sinha, N. K., and B. Kuszta, *Modelling and Identification of Dynamic Systems*, New York: Van Nostrand Reinhold Company, 1983.

Söderström, Torsten, "Identification of Stochastic Linear Systems in the Presence of Input Noise," *Automatica*, Vol. 17, No. 5, pp. 713-725, 1981.

Söderström, Torsten, Lennart Ljung, and I. Gustavsson, "A Theoretical Analysis of Recursive Identification Methods," *Automatica*, Vol. 14, pp. 231-244, 1978.

Solo, V., "The Convergence of AML," *IEEE Transactions on Automatic Control*, Vol. AC-24, No. 6, pp. 958-962, December 1979.

Solo, V., "Some Aspects of Recursive Parameter Estimation," *International Journal of Control*, Vol. 32, No. 3, pp. 395-410, 1980.

Solo, V., "The Second-Order Properties of a Time Series Recursion," *The Annals of Statistics*, Vol. 9, No. 2, pp. 307-317, 1981.

Sorenson, Harold W., "Least-squares estimation: from Gauss to Kalman," *IEEE Spectrum*, Vol. 7, No. 7, pp. 63-68, July 1970.

Sorenson, Harold W., and J. E. Sacks, "Recursive Fading Memory Filtering," *Information Sciences*, Vol. 3, pp. 101-119, 1971.

Steinway, W. J., and J. L. Melsa, "Discrete Linear Estimation for Previous Stage Noise Correlation," *Automatica*, Vol. 7, pp. 389-391, 1971.

Sternby, Jan, "On Consistency for the Method of Least Squares Using Martingale Theory," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 3, pp. 346-352, June 1977.

Stewart, G. W., *Introduction to Matrix Computations*, New York: Academic Press, Inc. 1973.

Stigum, Bernt P., "Asymptotic Properties of Dynamic Stochastic Parameter Estimation," *Journal of Multivariate Analysis*, Vol. 4, pp. 351-381, 1974.

Stoica, Petre, Jan Holst, and Torsten Söderström, "Eigenvalue Location of Certain Matrices Arising in Convergence Analysis Problems," *Automatica*, Vol. 18, No. 4, pp. 487-489, 1982.

Strejc, V., "Least Squares Parameter Estimation," *Automatica*, Vol. 16, pp. 535-550, 1980.

Toyooka, Yasuyuki, "Second-order Expansion of Mean Squared Error Matrix of Generalized Least Squares Estimator with Estimated Parameters," *Biometrika*, Vol. 69, No. 1, pp. 269-273, 1982.

Tse, Edison, and Michael Athans, "Optimal Minimal-Order Observer-Estimators for Discrete Linear Time-Varying Systems," *IEEE Transactions on Automatic Control*, Vol. AC-15, No. 4, pp. 416-426, August 1970.

VanLandingham, Hugh F., *Introduction to Digital Control Systems*, New York: Macmillan Publishing Company, 1985.

VanLandingham, Hugh F., and Mark A. Hopkins, "Deadbeat Parameter Identification of DARMA Processes," *Proceedings of the 24th Annual Allerton Conference on Communication, Control, and Computing*, University of Illinois at Urbana-Champaign, October 1986.

Walter, E., G. Le Cardinal, and P. Bertrand, "On the Identifiability of Linear State Systems," *Mathematical Biosciences*, Vol. 31, pp. 131-141, 1976.

Weiss, Alan, and Debasis Mitra, "Digital Adaptive Filters: Conditions for Convergence, Rates of Convergence, Effects of Noise and Errors Arising from the Implementations," *IEEE Transactions on Information Theory*, Vol. IT-25, No. 6, pp. 637-652, November 1979.

Zhdanov, A. I., and O. A. Katsyuba, "On Consistent Estimates of the Solutions of Ill-Posed Stochastic Algebraic Equations in the Identification of the Parameters of Linear Difference Operators," *Engineering Cybernetics*, Vol. 19, No. 5, pp. 121-127, September/October, 1981.