IFAC WORLD CONGRESS, 1984,
BUDAPEST
VOL X

CONVERGENCE RATE DETERMINATION FOR GRADIENT-BASED ADAPTIVE ESTIMATORS*

Robert R. Bitmead, Brian D.O. Anderson

Department of Systems Engineering, Research School of Physical Sciences, Australian National University, Canberra, Australia.

Tung Sang Ng

Department of Electrical Engineering, University of Wollongong, Wollongong, NSW, Australia.

*This work was supported by the Radio Research Board of Australia.

Abstract A convergence rate estimate is derived for the homogeneous gradient-based adaptive linear estimator algorithm. This estimate involves the eigenvalues of the regression vector covariance matrix, yielding a useful measure for the choice of input signals for adaptive parameter estimation. The connection between this criterion and those more familiar from nonadaptive system identification is made and comparisons are drawn between the two areas.

Keywords adaptive systems; identification; learning systems; gradient-based estimators; convergence rate.

1. INTRODUCTION

The standard adaptive linear parameter estimation problem is to take two time series, y_k of scalars and X_k of N-vectors, and to attempt to fit the linear model $\hat{y}_k = X_k'\theta$, where θ is an N-vector parameter, to minimize a criterion such as $E(y_k - \hat{y}_k)^2$. Many recursive algorithms for performing the adaptive estimation of this parameter have the form

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \mu b_k (y_k - X_k'\hat{\theta}_k) \qquad (1.1)$$

where $\hat{\theta}_k$ is the parameter estimate and b_k is a vector function of the X_k data sequence. The simplest and most representative form of these algorithms is the LMS scheme where $b_k = X_k$. We shall examine this algorithm in particular here as the analysis carries over quite directly to many other choices of b_k, one of which will be instanced in the sequel.

One of the analytical problems of examining adaptive linear parameter estimation algorithms is that, in applications involving noises, time-variations or other additive extraneous signals, the parameter estimates do not necessarily converge to a constant in a deterministic or a probabilistic sense. Rather, subject to certain stationarity assumptions, the estimates will converge in a distributional manner (Bitmead (1983)). The development of performance measures for these algorithms often consists of determining or estimating parameters of this limiting distribution such as certain moments. Further, one is also frequently concerned with the quantitative dependence of these measures on the designer-adjustable parameters of the estimation scheme. There appears to be two distinct methodologies for proceeding with this performance analysis. Each approach has its own advantages and, thankfully, produces consistent results with the alternative method.

The first technique, taking a cue from identification theory, examines the adaptive estimation algorithm as a forced linear difference equation for the parameter estimate, e.g.

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \mu X_k (y_k - X_k'\hat{\theta}_k) \qquad (1.2)$$

or, in terms of parameter error $\tilde{\theta}_k$

$$\tilde{\theta}_{k+1} = (I - \mu X_k X_k')\tilde{\theta}_k + \mu X_k n_k \qquad (1.3)$$

where the signal n_k embodies all the additive unmodelled disturbances. This approach has been adopted in (Farden (1981), Farden et al. (1980), Jones (1973), Macchi and Eweda (1983), Abu El Ata (1982), Kim and Davisson (1975)) and, while being technically very sound, *does rely very heavily on* stationarity and mixing assumptions on the underlying processes. The extension of these results to nonstationary applications requires some compromise because of the violation of the stationarity assumption.

The alternative approach is to consider initially only the homogeneous part of (1.3), to prove its exponential asymptotic stability and then to infer performance properties of the forced system. The extension of these results to non-stationary systems is relatively straightforward. However, there is frequently a need for the introduction of further conditions and assumptions to make the inference above. This method has been adopted in (Widrow et al. (1976), Mendel (1973), Bitmead and Anderson (1980), Johnson and Anderson (1981), Weiss and Mitra (1979)) and will be further applied here.

In studying the properties of non-homogeneous algorithms like (1.3) with the aid of the homogenous algorithm

$$v_{k+1} = (I - \mu X_k X_k')v_k \qquad (1.4)$$

it is desirable primarily to ensure the convergence of v_k to zero and then, secondarily, to determine or approximately quantify the convergence rate. All the available general non-stationary performance measure approximations rely (some implicitly) on knowing the exponential convergence rate of the homogeneous system and utilising bounded input/bounded output properties of linear systems.

Here we shall derive results which demonstrate, in very general cases, that the exponential convergence rate of (1.4) is quantified by the second moment of the X_k process. Conditions are then derived which relate this quantification to the spectrum of X_k. Having determined a convergence rate in terms of the X_k process we then turn to consider these conditions in equation error and output error schemes where X_k is composed of system inputs and outputs. The experiment design problem is considered where one asks the question: given that we may choose our process inputs u_k which then determine the process outputs y_k and these jointly determine X_k, what is a sensible choice of u_k to guarantee a good convergence rate of an adaptive parameter estimator? The response to this enquiry is related to the equivalent problem in nonadaptive system identification.

2 CONVERGENCE RATE QUANTIFICATION

If the homogeneous LMS algorithm (1.4) is iterated over m time steps then we have

$$v_{m+1} = \prod_{i=1}^{m}(I-\mu X_i X_i')v_1 \qquad (2.1)$$

and, writing $A_i = X_i X_i'$ and expanding, this yields

$$\prod_{i=1}^{m}(I-\mu A_i) = I-\mu\sum_{i=1}^{m} A_i + \mu^2\sum_{i=1}^{m}\sum_{j=i+1}^{m} A_j A_i \qquad (2.2)$$

$$-\mu^3\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{k=j+1}^{m} A_k A_j A_i$$

$$+\dots+(-\mu)^m A_m A_{m-1}\cdots A_1$$

$$= I-\mu m\bar{A} + (\mu m)\frac{1}{m}\sum_{i=1}^{m}(\bar{A}-A_i)$$

$$+(\mu m)^2\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=i+1}^{m} A_j A_i$$

$$+\dots+(-\mu m)^m\frac{1}{m^m}A_m A_{m-1}\cdots A_1 \qquad (2.3)$$

where $\bar{A} = E(A_i) = E[X_i X_i']$. We may now rewrite (2.1) using $z_k = v_{mk+1}$ as

$$z_{k+1} = (I-\mu m\bar{A}+Q_k)z_k \qquad (2.4)$$

where Q_k consists of the additional terms in the right hand side of (2.3).

The procedure taken will be to determine the behaviour of (2.4) by invoking the following theorem.

Theorem 1 (Bitmead and Anderson (1981))

The linear difference equation

$$w_{k+1} = \alpha_k w_k$$

with ergodic coefficient matrix α_k will be exponentially asymptotically stable if $\beta = E\|\alpha_k\| < 1$. Further, this exponential rate is at least as fast as β^k.

The application of this result to (2.4) will require the bounding of the effects of the random term Q_k, and this is carried out by considering μ to be small but not vanishing and m to be large but finite. Additional simple arguments are then necessary to show that properties of z_k are reflected in properties of $\{v_k\}$.

We may use the following result to bound some of these terms.

Lemma 1

Suppose that the moments of A_i are geometrically overbounded, i.e. there exist constants $M,\delta < \infty$ such that

$$E\|A_i^k\| \leq M\,\delta^k \qquad (2.5)$$

for all k. Then for all m, the ℓ^{th} term in (2.2) satisfies

$$E\left\|\frac{1}{(m)^\ell}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\dots\sum_{h=k+1}^{m}A_h A_k\dots A_j A_i\right\| \leq \frac{1}{\ell!}M\,\delta^\ell \qquad (2.6)$$

Consequently, we have in (2.4)

$$E\|Q_k\| \leq E\|(\mu m)\frac{1}{m}\sum_{i=1}^{m}(\bar{A}-A_i)\| + M\sum_{i=2}^{m}\frac{(\mu m\delta)^i}{i!} \qquad (2.7)$$

The proof of this lemma, presented in the appendix, involves a simple application of the triangle and Hölder inequalities. At this point it is pertinent to remark that the inequality (2.5) holds trivially in the case of almost surely bounded X_k. While this result is not directly applicable to the gaussian case and other distributions of noncompact support because of (2.5) it is possible, using the methods of (Bitmead 1981) to develop equivalent hard bounds analogous to (2.6) subject only to the existence of all moments of A_i.

Now it remains to bound the first term on the right hand side of (2.7) and for this we invoke the following result presented in the appendix as an extension of the work of Ibragimov and Linnik (1971)).

Lemma 2

Let A_i be stationary with mean $\bar{A}$ and let $f^{ij}(\lambda)$ be the spectral density of the i-j component of $A_i-\bar{A}$ which we assume to exist and to be twice differentiable at $\lambda=0$. Let

$$S_m = \frac{1}{m}\sum_{i=1}^{m}(A_i-\bar{A}) \qquad (2.8)$$

then

$$E[|S_m^{ij}|^2] = \frac{2\pi f^{ij}(0)}{m} + \pi m^{-3/2}\left[\frac{d^2}{d\lambda^2}f^{ij}(\lambda)\right]_{\lambda=0} + O(m^{-7/4}) \qquad (2.9)$$

and thus by Holder's inequality,

$$E\|\frac{1}{m}\sum_{i=1}^{m}(\bar{A}-A_i)\| = O(m^{-1/2}) \qquad (2.10)$$

The results of Lemmas 1 and 2 may now be combined to yield the main Theorem.

Theorem 2

Subject to the hypotheses of Lemmas 1 and 2 we may always find a sufficiently small but constant value of the gain μ so that the homogeneous adaptive estimation algorithm (1.4) converges to zero exponentially fast with a rate arbitrarily close to $(1-\mu\alpha)^k$ where α is the minimum eigenvalue of $\bar{A} = E[X_k X_k']$.

Proof

Examining (2.4) in the light of Theorem 1 we see that the convergence rate of the homogeneous algorithm (1.4) is determined by

$$E\|I-\mu m\bar{A} + Q_k\| < \|I-\mu m\bar{A}\| + E\|Q_k\| \qquad (2.11)$$

$E\|Q_k\|$ is overbounded in (2.7) and, from Lemma 2, m may be taken large enough to arbitrarily bound the term in (2.10). Then μ may be chosen to force μm and $\mu m\delta$ in (2.7) to be arbitrarily small thus making $E\|Q_k\|$ arbitrarily small. That the sum in (2.7) tends to zero as $\mu m\delta$ tends to zero may be seen by considering

$$\sum_{i=2}^{m} \frac{(\mu m\delta)^i}{i!} < e^{\mu m\delta} - 1 - \mu m\delta$$

$$< \mu m\delta\,(e-1)$$

provided $0 < \mu m\delta < 1$. The Theorem then follows by relating v_k in (1.4) to z_k in (2.4), to show that the convergence rate is arbitrarily close to $(1-\mu m\alpha)^{k/m}$. Application of the binomial theorem *then establishes the theorem.*

∇∇∇

In the next section we shall discuss the implications of Theorem 2 to the problem of input selection for adaptive estimation, i.e. experiment design. However, our task now will be to relate the convergence conditions of the Theorem statement to the more usual requirements of covariance decay (Farden et al. (1980), Macchi and Eweda (1983), Abu El Ata (1982)) or mixing (Bitmead and Anderson (1980)) of the X_k-process.

We have already remarked following Lemma 1 that the Theorem requirements for boundedness of $O(\mu^2)$ terms in $E\|Q_k\|$ will be satisfied if the X_k-process is bounded or at least has finite moments of all orders. These restrictions only relate to the distribution of X_k. The other requirement that the spectral density $f(\lambda)$ be twice differentiable at $\lambda=0$ is a statement concerning the timewise correlation of X_k. To see this consider for simplicity the corresponding continuous time case

$$\left.\frac{d^2}{d\lambda^2} f(\lambda)\right|_{\lambda=0} = \left.\frac{d^2}{d\lambda^2} \int_{-\infty}^{\infty} e^{-j\lambda t} R(t)dt\right|_{\lambda=0}$$

$$= \int_{-\infty}^{\infty} - t^2 R(t)dt \qquad (2.12)$$

when $R(t)$ is the autocorrelation function of A_i. The existence of the second derivative of $f(\lambda)$ is thus a requirement that the integral in (2.12) be well defined. For this to occur it is sufficient that

$$\int_{-\infty}^{\infty} t^2 \, |R(t)| dt < \infty$$

so that the differentiability restriction may be interpreted as a covariance decay condition on A_i. Since $A_i = X_i X_i'$ we in turn have a covariance decay requirement on X_i. This is entirely consistent with the known convergence requirements for these algorithms.

3 EXPERIMENT DESIGN FOR ADAPTIVE ESTIMATION

In practice, the algorithm gain μ is typically chosen to be very small (usually < .001 divided by the input signal power). The reasons for this choice of small μ are twofold. Firstly, it may be demonstrated (Farden et al. (1980), Macchi and *Eweda* (1983), Abu El Ata (1982)) that, for stationary systems, the asymptotic parameter error variance is linear in μ so that time-invariant steady state performance requirements dictate that μ should be small in order that the parameter estimates be smooth. Secondly, in order that the adaptive estimation problem of slowly time-varying parameters be well-defined one requires that the estimator time constants should be sufficiently longer than the system time constants. To see this latter point recall that the conditions of the previous section involved μm being small, where m reflected the convergence rate of the Cesaro sums in (2.8) and hence reflected the system time constants. The choice of m is thus related to the covariance of the input process, and the condition number of $\bar{A}$ plays an important role in determining the maximum allowable value of μ. Convergence requirements in the independent case demand $\mu\lambda_{max}(\bar{A}) < 1$ (Widrow et al. (1976)) while the convergence rate analysis of the previous section would favour $\mu\lambda_{min}(\bar{A})$ close to one. It is well known that poor condition number severely affects the performance of gradient-based minimization methods (Luenberger, (1973)) and this is clearly reflected here.

Given that μ is chosen small on the grounds above and that Theorem 2 applies, we are justified in attempting to choose an input process to maximize the convergence rate. The question of adjusting μ is not addressed in this section but rather the questions of input selection for a particular choice of μ. In order to maximise this convergence rate it is necessary to maximise the minimum eigenvalue of $\bar{A}$ subject to input power constraints and we shall next address certain problems in achieving this.

For finite impulse response or moving average systems one typically has

$$X_k = (u_k \; u_{k-1} \cdots u_{k-N})'$$

where u_k is the scalar system input process. This is the well-known adaptive transversal filter, and the best choice of input u_k to maximize convergence rate is u_k an independent and identically distributed sequence. Performance in this case has been closely studied (Widrow et al. (1976)). When dealing with infinite impulse response or autoregressive-moving average systems

$$X_k = (y_{k-1} \;\; y_{k-2} \cdots y_{k-N} \; u_k \cdots u_{k-N})'$$

so that the actual system parameters have an implicit effect in determining the best input sequence u_k for adaptive parameter estimation.

The area of system identification has given rise to many efforts to confront the problem of experiment design (Goodwin and Payne (1977)). There is a natural performance benchmark here, the Cramer-Rao lower bound on the covariance of the estimates, and optimal experiment design in this context is directed towards ensuring that, since the Cramer-Rao bound is given by the inverse of the Fisher information matrix M, the input process extremizes a suitable scalar function of M in an attempt to minimize the parameter error covariance. (This function above typically involves an expectation taken over the a priori parameter distribution.)

The Cramer-Rao lower bound is met if $E(\tilde{\theta}\tilde{\theta}^T) = M$ and, for moving average systems, M consists of the covariance matrix of the input process scaled by the measurement noise variance, while for autoregressive moving average systems it comprises the scaled covariance of a vector composed of input and output processes. In our notation here

$$M = E(X_k X_k') = \hat{A} \qquad (3.1)$$

and the correspondence between the adaptive and nonadaptive schemes is clear although it should be stated here that the estimate of convergence rate in the adaptive case is not necessarily tight and may be conservative. In spite of this disclaiming rider it is apparent that, for a given adaptive estimation problem with a particular system and small μ, the estimate convergence rate will be enhanced by maximizing the minimum eigenvalue of $\bar{A}$.

There are several common scalar performance measures of M and each appears to have some advantageous properties and some negative ones. Maximizing the trace of M has been used but this can lead to parameter nonidentifiability (Grewal and Glover (1975)). Maximizing det M (or more commonly minimizing -log det M) has better features as well as independence from parameter scaling and an interesting frequency domain signal to noise ratio interpretation (Qureshi and Ng (1982)). However, because det M=0 if and only if M is singular, this criterion can be optimized by identifying perfectly the projection of θ in any one particular direction, regardless of estimator performance in other directions. To this extent, this criterion does permit the trading off of identifier performance in one direction to the detriment of other directions.

The performance measure being advanced here is to maximize the minimum eigenvalue of M, which is equivalent to minimizing the maximum eigenvalue of the parameter error covariance. Thus the criterion attempts to control mean squared parameter error in every direction. It does suffer from being scaling dependent and analytically difficult to work with, however.

What is important in this context is that the adaptive experiment design problem has a very similar formulation to that of system identification thus allowing the application of well-studied input selection procedures from this area.

It certainly is not true that adaptive estimators of the form of (1.1) are efficient in the statistical sense but rather that convergence rate improvement for these adaptive schemes is mathematically similar to nonadaptive experiment design. Thus the parameter error covariance is not equal to the Cramer-Rao lower bound of M^{-1}. In the case of independent X_k it is readily shown that the asymptotic parameter error covariance is given to first order in μ by $\frac{1}{2}\mu\sigma^2$, where σ^2 is the variance of the Wiener residuals $y_k - X_k'\theta$, which is not the same as M^{-1}.

An alternative algorithm may be constructed, if M is known or approximated, which is closer in spirit to the asymptotically efficient methods used in nonadaptive systems identification based on least-squares estimation:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \mu M^{-1} X_k (y_k - X_k'\hat{\theta}_k) \qquad (3.2)$$

or

$$\hat{\theta}_{k+1} = (I - \mu M^{-1} X_k X_k')\tilde{\theta}_k + \mu M^{-1} X_k n_k \qquad (3.3)$$

where n_k is the Wiener residual. This algorithm may be written

$$M^{1/2}\tilde{\theta}_{k+1} = (I - \mu M^{-1/2} X_k X'_k M^{-1/2})M^{1/2}\tilde{\theta}_k + \mu M^{-1/2} X_k n_k \qquad (3.4)$$

from which one sees that the convergence rate will be linear in μ and independent of M, while the asymptotic error covariance may be approximated (assuming independent X_k and to first order in μ) by the solution P to

$$P = (I-\mu I)P(I-\mu I) + \mu^2 M^{-1}\sigma^2$$

or $P = \mu \frac{\sigma^2}{2} M^{-1}$. This modified algorithm (3.2) demonstrates the connection between the experiment design problems as well as the trade-off between adaptive estimator convergence rate and parameter error covariance.

One criticism frequently levelled at experiment design is that, as the optimal design depends on unknown system parameters, there is a circularity in its justification. This criticism is frequently answered by using a Bayesian approach and averaging over a prior distribution, however in the general areas of adaptive estimation (in contrast to nonadaptive system identification) where tracking of slowly varying parameters is often desired, one frequently will have a reasonable knowledge of the system parameter values thus allowing effective 'causal' experiment design.

4 CONCLUSIONS

We have presented a method for deriving a lower bound on the exponential convergence rate of the homogeneous part of gradient-based adaptive parameter estimators. The convergence rate is dependent on the minimum eigenvalue of the covariance matrix of the X_k process and, as a result, if the input signal is to be chosen to enhance the convergence rate, a useful criterion of performance is the magnitude of this smallest eigenvalue.

The comparison was drawn between this performance measure for adaptive estimators and the experiment design criteria for nonadaptive system identification. The important point here was to show that while these two problems have seemingly disparate origins there are considerable similarities and the input selection rules from identification may be carried over essentially intact to adaptive estimation.

We have concentrated specifically on the LMS gradient-based adaptive estimation algorithm because it provides performance criteria which are usually representative of the class of such schemes as was evidenced in section 3. Equally, the exact analysis of the nonhomogeneous equations was not presented with dependent inputs, since the effect of dependence appears most critically in the convergence rate of the homogeneous equation, and the nonhomogeneous error covariances can be *related to those in the independent* X_k *theory.*

The importance of the results presented here is that they allow the approximate quantification of convergence rate and, more importantly, they allow the development of useful input selection procedures to maximize this rate. This represents a significant extension from the usual 'persistence of excitation' (Bitmead and Anderson (1980)) requirements and would be readily interfaced with other restricted input designs for adaptive systems (Ioannou and Kokotivic (1983)).

Appendix

Proof of Lemma 1

The difficult component here is, the demonstration of (2.6). Using the Holder inequality repeatedly we have

$$E\|A_h A_k \ldots A_i\| < E^{1/\ell}\|A_h\|^\ell \, E^{1/\ell}\|A_k\|^\ell \ldots E^{1/\ell}\|A_i\|^\ell$$
$$< M\delta^\ell$$

from (2.5).

Thus the term on the left hand side of (2.6) is bounded above by

$$\frac{1}{m^\ell}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{h=k+1}^{m} M\delta^\ell < \frac{M\delta^\ell}{m^\ell}\int_0^m\int_i^m \ldots \int_k^m dh\,dk..di$$
$$= \frac{M\delta^\ell}{m^\ell}\int_0^m\int_0^i \ldots \int_0^k dh\,dk..di$$
$$= \frac{M\delta^\ell}{\ell!}$$

Proof of Lemma 2

The proof of Theorem 18.2.1 of (Ibragimov and Linnik (1971)) shows that if a stationary scalar sequence X_i is given, and *one defines*

$$T_m = \sum_{i=1}^{m} X_i$$

and if $f(\lambda)$, the spectral density function of X, is continuous at $\lambda=0$, then

$$\left|\mathrm{var}\, T_m - 2\pi f(0)m\right| < 2\pi m \max_{|\lambda|<m^{-1/4}} \left|f(\lambda)-f(0)\right| + O(m^{-1/2}) \qquad \text{(A.1)}$$

Now if $f(\lambda)$ has first and second derivatives at $\lambda=0$ then, since $f'(0) = 0$, by Taylor's Theorem

$$\left|f(\lambda) - f(0)\right| = \frac{\lambda^2}{2} f''(0) + O(\lambda^3)$$

and (A.1) yields

$$\frac{\mathrm{var}\, T_m}{m^2} = \frac{2\pi f(0)}{m} + O[\pi m^{-3/2} f''(0)] + O(m^{-7/4}) + O(m^{-5/2})$$

and the Lemma follows by identifying T_m/m with S_m^{ij}, X_i with the i-j component of $A_i - \bar{A}$, and $f(\lambda)$ with $f^{ij}(\lambda)$.

REFERENCES

Abu, El Ata, S. (1982). Asymptotic Behaviour of an Adaptive Estimation Algorithm with Application to M-Dependent Data. IEEE Trans. Automat. Contr., AC-27, 1255-1257.

Bitmead, R.R. and B.D.O. Anderson (1980). Performance of Adaptive Estimation Algorithms in Dependent Random Environments. IEEE Trans. Automat. Contr., AC-25, 788-794.

Bitmead, R.R. and B.D.O. Anderson (1981). Adaptive Frequency Sampling Filters. IEEE Trans. Circuits Syst., CAS-28, 524-534.

Bitmead, R.R. (1981). Convergence Properties of LMS Adaptive Estimators with Unbounded Dependent Inputs. Proc. 20th IEEE Conf. Decision and Control, San Diego, CA., 607-612.

Bitmead, R.R. (1983). Convergence in distribution of LMS-Type Adaptive Parameter Estimates. IEEE Trans. Automat. Contr., AC-28, 54-60.

Farden, D.C., J.C. Goding Jr.,and K. Sayood (1980). On the 'desired behaviour' of adaptive signal processing algorithms. *Proc. 1979 IEEE Int. Conf. Acoust., Speech, Signal* Processing, 466-469.

Farden, D.C. Farden (1981). Tracking Properties of Adaptive Signal Processing Algorithms. IEEE Trans. Acoust., Speech, Signal *Processing.*, ASSP-29, 439-446.

Goodwin, G.C. and R.L. Payne (1977). Dynamic System Identification: Experiment Design and Data Analysis, Academic Press, New York.

Grewal, M.S. and K. Glover (1975). Relationships between Identifiability and Input Selection. Proc. IEEE Conf. Decision and Control, Phoenix AZ., 526-528.

Ibragimov, I.A. and Yu. V. Linnik (1971). Independent and Stationary Sequences of Random Variables. Wolters-Noordhoff, Groningen.

Ioannou, P.A. and P.V. Kokotovic (1983). Adaptive Systems with Reduced Models. Springer-Verlag, Berlin.

Johnson Jr., C.R., and B.D.O. Anderson (1981). Sufficient Excitation and Stable Reduced-order Adaptive IIR Filtering. IEEE Trans. Acoust., Speech, Signal Processing, ASSP-29, 1212-1215.

Jones, S.K. (1973). Adaptive Linear Estimation for Stationary M-Dependent Processes. Ph.D. dissertation, Dept. Elec. Eng., Southern *Methodist University*, Dallas, TX.

Kim, J.R., and L.D. Davisson (1975). Adaptive Linear Estimation for Stationary M-Dependent Processes. IEEE Trans. Infor. Theory, IT-21, 23-31.

Luenberger, D.G. (1973). Introduction to Linear and Nonlinear Programming. Addison-Wesley, Reading, MA.

Macchi, O., and E. Eweda (1983). Second Order Convergence Analysis of Stochastic Adaptive Linear Filtering. IEEE Trans. Automat. Contr., AC-28, 76-85.

Mendel, J. (1973). Discrete Techniques of Parameter Estimation: The Equation Error Formulation. Marcel-Dekker, New York.

Qureshi, Z.H., and T.S. Ng (1982). Parameter Estimation: The D_s-Optimality Case. *SIAM J.* Control Optimiz., 20, 713-721.

Sondhi, M.M. and D. Mitra (1976). New Results on the Performance of a Well-known Class of Adaptive Filters. Proc. IEEE, 64, 1583-1597.

Weiss, A. and D. Mitra (1979). *Digital Adaptive* Filters: Conditions for Convergence, Rates of Convergence, Effects of Noise and Errors Arising from the Implementation. IEEE Trans Info Theory, IT-25, 637-652.

Widrow, B., J.M. McCool, M.G. Larimore and C.R. Johnson Jr. (1976). Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter. Proc. IEEE, 64, 1151-1162.