[3] A. K. Mandal, "Generation of Lyapunov functions and its application in system design," Ph.D. dissertation, Univ. Calcutta, Calcutta, India, 1971.
[4] A. K. Mandal, T. Mukherjee, and A. K. Choudhury, "Generation of Lyapunov functions for linear, nonlinear and time-varying systems," private communication.
[5] R. E. Kalman and J. E. Bertram, "Control system analysis and design via the second method of Lyapunov," *Trans. ASME (J. Basic Engr.), Ser. D*, vol. 82, pp. 371–393, 1960.
[6] A. K. Mandal, "Lyapunov functions for linear time-varying systems and their applications," private communication.
[7] A. K. Mandal and A. K. Choudhury, "Design of a class of nonlinear time-varying systems" private communication.

A Note on Stochastic Approximation Algorithms in System Identification

H. G. KWATNY

***Abstract*—This correspondence considers the application of stochastic approximation algorithms to a broad class of system identification problems. Both asymptotic and initial convergence properties of the algorithms are discussed. A suboptimal procedure for parameter selection and a means of convergence acceleration are suggested.**

Introduction

A variety of interesting problems of system identification, both linear and nonlinear, can be formulated in terms of the discrete-time linear model

$$y_n = \boldsymbol{a}_n' \boldsymbol{Y}_n + \eta_n \tag{1}$$

relating the output y_n (scalar, for simplicity), input $\boldsymbol{Y}_n$ (random s-vector), unknown parameter $\boldsymbol{a}_n$ (s-vector), and random noise η_n [2]–[5], [7]. Numerous papers have been concerned with the on-line estimation of $\boldsymbol{a}_n$ using stochastic approximation algorithms (for example, [3]–[7]).

It appears to be well known, although not explicitly stated, that if (1) is an exact representation, then unbiased estimates of $\boldsymbol{a}_n$ can be obtained by these methods without prior knowledge of the statistical parameters of $\boldsymbol{Y}_n$ and η_n, provided $\boldsymbol{Y}_n$ and η_n are statistically orthogonal (and, of course, $\boldsymbol{Y}_n$ repeatedly spans s-space). If $\boldsymbol{Y}_n$ and η_n are not orthogonal, then the estimators are generally biased and some statistical parameters of $\boldsymbol{Y}_n$ and/or η_n must be known in order to remove the bias [4]–[7]. However, the requirements for a nonparametric formulation can frequently be met (but not frequently enough, unfortunately). This is, perhaps, the strongest attraction for stochastic approximation methods.

It should also be noted that it is generally assumed that $\boldsymbol{Y}_n$ and η_n are temporally independent sequences (as well as being mutually independent). However, this assumption is not necessary and greatly restricts the class of problems which can be cast into the form of (1). In [3] it was shown to be sufficient that $\boldsymbol{Y}_n$ (and η_n) become independent at a geometric rate and actually this can be relaxed to a harmonic rate which correlates with conditions originally presented by Sakrison [1].

Even under such fortuitous circumstances, most proposed (nonparametric) stochastic approximation algorithms have the disadvantage that, although asymptotic convergence is assured, very little can be done to control the initial convergence properties. However, an algorithm proposed in [3] and which also appeared in [6] has some very favorable characteristics in this regard. These will be explored in the following.

Manuscript received July 12, 1971; revised February 4, 1972.
The author is with the College of Engineering, Drexel University, Philadelphia, Pa.

Properties of Stochastic Approximation Algorithms

Discussion of the major points is facilitated by comparing the behavior of the more-or-less standard algorithm

$$\hat{\boldsymbol{a}}_n = \hat{\boldsymbol{a}}_n - \frac{k}{n+1}(\hat{\boldsymbol{a}}_n' \boldsymbol{Y}_n - y_n)\boldsymbol{Y}_n \tag{2}$$

with the proposed algorithm

$$\hat{\boldsymbol{a}}_{n+1} = \hat{\boldsymbol{a}}_n - \frac{k}{n+1}\frac{(\hat{\boldsymbol{a}}_n' \boldsymbol{Y}_n - y_n)\boldsymbol{Y}_n}{\|\boldsymbol{Y}_n\|^2}. \tag{3}$$

It can be shown (as in [3]) that if the joint density functions of $\boldsymbol{Y}_n$ and η_n satisfy $P(\boldsymbol{Y}_p|\boldsymbol{Y}_n) = P(\boldsymbol{Y}_p) + o((p-n)^{-1})$, $P(\eta_p|\eta_n) = P(\eta_p) + o((p-n)^{-1})$ for $p - n \to \infty$ and that variations in $\boldsymbol{a}_n$ vanish asymptotically at the rate n^{-w} where $w > 1$, then, in both cases, the mean-square estimation error $\langle\|\boldsymbol{\rho}_n\|^2\rangle$, where $\boldsymbol{\rho}_n = \hat{\boldsymbol{a}}_n - \boldsymbol{a}_n$, is bounded by a quantity x_n satisfying the difference equation

$$x_{n+1} = \xi_n x_n + \frac{B}{(n+1)^\nu}; \qquad x_0 = \langle\|\boldsymbol{\rho}_0\|^2\rangle \tag{4}$$

where B depends mainly on the observation noise and is proportional to k^2 and $\nu = \min(2w - 1, 2)$. The sequence ξ_n will be discussed below.

On the basis of (4), both estimators can be shown to converge in the mean; in fact, the asymptotic behavior is expressed by the relations

$$\langle\|\boldsymbol{\rho}_n\|^2\rangle \sim \frac{2B}{2k\gamma - (\nu-1)} n^{-(\nu-1)}, \qquad \text{for } \nu - 1 < 2k\gamma$$
$$= 0(n^{-2k\gamma}), \qquad \text{for } \nu - 1 \geq 2k\gamma \tag{5}$$

for estimator (2), and

$$\langle\|\boldsymbol{\rho}_n\|^2\rangle \sim \frac{2B}{2k\beta - (\nu-1)} n^{-(\nu-1)}, \qquad \text{for } \nu - 1 < 2k\beta$$
$$= 0(n^{-2k\beta}), \qquad \text{for } \nu - 1 \geq 2k\beta \tag{6}$$

for estimator (3), where $\gamma = \inf_n \gamma_n$, $\beta = \inf_n \beta_n$, where γ_n is the smallest eigenvalue of the positive-definite matrix $\langle \boldsymbol{Y}_n \boldsymbol{Y}_n' \rangle$, and β_n is the smallest eigenvalue of the positive-definite matrix $\langle \boldsymbol{Y}_n \boldsymbol{Y}_n' / \|\boldsymbol{Y}_n\|^2 \rangle$. Since the maximum value of $\nu - 1$ is 1, it is evident that the asymptotic convergence of either estimator (2) or (3) will never be faster than $1/n$ regardless of how large $2k\gamma$ or $2k\beta$ may be.

It is interesting to focus on the transient response of (4). In [3] it is shown that $|\xi_n| < 1$ for all $n \geq N_0$ where

$$N_0 = \text{integral part}\left[\frac{k}{2} \times \frac{\mu}{\gamma}\right] \tag{7}$$

for estimator (2), where $\mu = \sup \mu_n$, μ_n being the largest eigenvalue of the positive-definite matrix $\langle \|\boldsymbol{Y}_n\|^2 \boldsymbol{Y}_n \boldsymbol{Y}_n' \rangle$, and γ is defined above. For estimator (3)

$$N_0 = \text{integral part}\left[\frac{k}{2}\right]. \tag{8}$$

This means that the transient part of (4) will converge monotonically for $n \geq N_0$. Alternatively, it is possible for the estimate to diverge from the desired vector for $n < N_0$. Since the quantity μ/γ is not known beforehand, it is impossible to determine N_0 for estimator (2). This is typical of most stochastic approximation procedures. Furthermore, μ/γ may be quite large, especially if the dimension of $\boldsymbol{a}$ is large. This means that there is likely to be an initial period of divergence although asymptotic convergence is assured. The novel feature of estimator (3) is that N_0 as given by

(8) can be controlled by judicious choice of k. By choosing $k = 1$, $N_0 = 0$ so that the transient term is monotonically decreasing from the inception of the iteration process.

A Suboptimal Procedure

From (5) or (6), it is seen that it is always possible to choose k sufficiently large so that the maximum rate of asymptotic convergence is obtained, i.e., since $\nu - 1$ is at most 1, select $k > \frac{1}{2}\gamma$ or $k > \frac{1}{2}\beta$. Unfortunately, γ and β are not known *a priori*. On the other hand, the disturbance coefficient B increases with k^2, and hence it is undesirable to choose k larger than necessary. To further complicate the problem, the number N_0 increases with k. In the case of estimator (2), nothing can be done in view of the limited *a priori* information in the way of selecting a near-optimum value of k. An advantage of estimator (3) is that something more can be done. To begin with, suppose that Y_n is s-dimensional and let $\lambda_1, \lambda_2, \cdots, \lambda_s$ be the eigenvalues of the positive-definite matrix $\langle Y_n Y_n'/\|Y_n\|^2\rangle$. Then

$$\operatorname{tr}\left\langle \frac{Y_n Y_n'}{\|Y_n\|^2} \right\rangle = \lambda_1 + \lambda_2 + \cdots + \lambda_s. \tag{9}$$

However,

$$\operatorname{tr}\left\langle \frac{Y_n Y_n'}{\|Y_n\|^2} \right\rangle = \left\langle \frac{Y_n' Y_n}{\|Y_n\|^2} \right\rangle = 1. \tag{10}$$

Now, a rather conservative upper bound for the minimum eigenvalue λ_1 can be obtained by assuming all of the eigenvalues to be equal. In this case, (9) and (10) yield $\lambda_1 = 1/s$. Thus, a conservative estimate of β is $1/s$, and hence we should choose k accordingly, say, $k = s$. Note that s is the number of unknown parameters and may be quite large. Once k is specified, N_0 is given by (8). In order to retain the property that the transient response of x_n decreases from the start of the iteration procedure, the process is started with $n = N_0$.

Convergence Acceleration

In some instances, when β is quite small, the above estimate can be far too conservative and poor convergence is obtained. This situation is accompanied by a large spread in the eigenvalues of $\langle Y_n Y_n'/\|Y_n\|^2\rangle$ (the matrix is ill conditioned), and arbitrarily increasing the value of k is generally unsatisfactory. In such cases it has been found advantageous to use a modified algorithm as follows. At the time of computation of $\hat{a}_{n+1}$, in addition to the estimate correction provided in (3), a correction orthogonal to Y_n and lying in the plane of Y_n and Y_{n-1} is added. The algorithm is

$$\hat{a}_{n+1} = \hat{a}_n - \frac{k}{n+1}\left[\frac{Y_n}{\|Y_n\|^2}(\hat{a}_n' Y_n - Y_n) + \frac{Z_n}{\|Z_n\|^2}(\hat{a}_n Z_n - \|Y_n\|^2 y_{n-1} - (Y_n' Y_{n-1}) y_n)\right] \tag{11}$$

where

$$Z_n = \|Y_n\|^2 Y_{n-1} - (Y_n' Y_{n-1}) Y_n. \tag{12}$$

This algorithm converges to the true value of the parameter and its convergence properties may be characterized in terms of parameters similar to those used above. The proof parallels the proof of convergence of (3) as outlined in [3]. In this case, the property $N_0 =$ integral part $[k/2]$ is retained. A conservative choice of k can be shown to be $s/2$. An example of the application of the algorithm can be found in [3].

References

[1] D. J. Sakrison, "Application of stochastic approximation methods to system optimization," Res. Lab. Electron., Mass. Inst. Technol., Cambridge, Tech. Rep. 391, 1962.
[2] Y. C. Ho and R. C. K. Lee, "Identification of linear dynamic systems," *Inform. Contr.*, vol. 8, pp. 93–110, Feb. 1965.
[3] H. G. Kwatny and D. W. C. Shen, "Identification of nonlinear systems using a method of stochastic approximation," in *Proc. 8th Joint Automatic Control Conf.*, 1967, pp. 814–826.
[4] G. N. Saridis and G. Stein, "Stochastic approximation algorithms for linear discrete-time system identification," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 515–523, Oct. 1968.
[5] ——, "A new algorithm for linear system identification," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-13, pp. 592–594, Oct. 1968.
[6] Y. Bar-Shalom and S. C. Schwartz, "Application of stochastic approximation to on-line system identification," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-15, pp. 606–607, Oct. 1970.
[7] A. N. Netravali and R. J. P. De Figueiredo, "On the identification of nonlinear dynamical systems," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 28–36, Feb. 1971.

Optimal Smoothing for Continuous-Time Systems with Multiple Time Delays

K. K. BISWAS AND A. K. MAHALANABIS

***Abstract*—Equations for the smoothed state estimate and for the error covariances of a continuous-time system with multiple time delays, based on observations involving time delays, are derived through a combination of discretization, state augmentation, and subsequent dediscretization procedures.**

Introduction

Although the existing literature contains smoothing results for non-time-delayed systems, no attempt seems to have been made to extend these to the case of time-delayed continuous-time systems. Priemer and Vacroux have reported some results for the discrete-time version of the problem, which were obtained through projection arguments [1], [2]. The aim of this correspondence is to report smoothing solutions for a continuous-time system having both transportation and observation lags. These are obtained by first discretizing the continuous-time problem and then employing a state augmentation technique [3] that converts the given problem into a non-time-delayed higher dimensional filtering problem. The desired smoothing solutions are then obtained from the components of the higher dimensional filtering equations. Finally, a formal limiting procedure is utilized to derive the continuous-time smoothing solutions.

Problem Statement

Consider a time-delayed continuous-time system modeled by the following equations:

$$\dot{x}(t) = \sum_{i=0}^{L} F_i(t) x(t - \alpha_i) + w(t) \tag{1}$$

$$y(t) = v(t) + \sum_{i=0}^{M} H_i(t) x(t - \beta_i) \tag{2}$$

where x is the n-vector system state, y is the m-vector observation, α_i and β_i represent, respectively, the ith delay in the system and observation equations with $\alpha_i > \alpha_{i-1}$; $\beta_i > \beta_{i-1}$; $\alpha_0 = \beta_0 = 0$. L and M denote the total number of delays in the system and observation.

The noise processes $w(t)$ and $v(t)$ are assumed to be independent, zero-mean, white, Gaussian processes with covariances $Q(t)$ and $R(t)$, respectively, with $R(t)$ positive definite.

If it is assumed that $t = t_k$ (the kth sampling instant), $\alpha_i = d_i T$, and $\beta_i = h_i T$ where T is the sampling interval, the discrete-time equivalents of (1) and (2) can be obtained in the form [4]

$$x(t_k + T) = \sum_{i=0}^{L} A_i(t_k) x(t_k - d_i T) + w(t_k) T^{1/2} \tag{3}$$

Manuscript received January 18, 1972.
The authors are with the Department of Electrical Engineering, Indian Institute of Technology, New Delhi-29, India.