Nonstationary Learning Characteristics of the LMS Algorithm

WILLIAM A. GARDNER, SENIOR MEMBER, IEEE

Abstract — **Upper and lower bounding first-order linear recursions for the mean-squared error realized with the LMS algorithm subjected to a sequence of independent nonstationary training vectors are derived. These bounds coincide to give the exact evolution of mean-squared error for the problem of identification of a nonrecursive time-varying system with white-noise excitation. This leads to an exact formula for time-averaged mean-squared error that is used to study optimization of the step-size parameter for minimum time-average misadjustment. New results on dependence of the minimal step size and the minimum misadjustment on the degree of nonstationarity are obtained.**

I. Introduction

THE TRACKING performance of the LMS algorithm is studied in [1] and [2] for a particular type of time-variant system identification problem. The analysis and conclusions there are based on the *a priori* assumptions that the step size in the algorithm is small and the time variations of the system are slow. The purposes of this paper are i) to present a general approach to studying the tracking performance of the LMS algorithm that is applicable to many types of problems in addition to system identification and ii) to apply this general approach to the particular system identification problem studied in [1] and [2] in order to determine the effects of the *a priori* assumptions of small step size and slow time variations on the results and conclusions presented there.

In brief, it is found that the small step-size assumption alone does not seriously affect the optimization of the step size for minimization of misadjustment for the cases considered, but does lead to the prediction of a misadjustment that for large step sizes can be off by a factor that is as large as 2 to 3. It also removes all dependence of the algorithm's performance on the kurtosis of the data (in which case, for example, the predicted performance is identical for Gaussianly distributed data and uniformly distributed data). In addition, it is found that the slow time-variations assumption coupled with the small step-size assumption has a substantial effect on the results obtained and some of the conclusions drawn. For example, it is shown herein that the optimum step size is not a monotonically increasing function of the degree-of-nonstationarity, and the two components of misadjustment, due to gradient noise and nonstationarity, can be highly unequal when their sum is minimum.

In Section II, the general approach to studying the tracking performance of the LMS algorithm is presented and implicit solutions for bounds on the time-averaged excess mean-squared error are obtained. Then, in Section III, it is shown that for the system identification problem of interest, the implicit solutions reduce to explicit solutions and the upper and lower bounds coincide to produce the exact solution. This solution is then used to study the misadjustment-minimization problem. The section concludes with simulation results that corroborate the theoretical results.

Finally, in Section IV, the results obtained in Section III are compared with results from other studies of the same problem, and discrepancies are explained.

Other work on the tracking performance of the LMS algorithm, which employs various methods of bounding and approximation, is reported in [5]–[11].

II. General Analysis

The LMS algorithm for adaptive adjustment of the N-vector of filter weights W is well known to be given by

$$W(i+1) = W(i) + 2\mu e(i) X(i) \tag{1}$$

where μ is a step-size parameter,[1] $X(i)$ is the filter input vector, $e(i) = d(i) - \hat{d}(i)$ is the error between the desired quantity $d(i)$ and the filter output $\hat{d}(i) = W^T(i)X(i)$, and $2e(i)X(i) = -\nabla e^2(i)$ is the negative gradient of the squared error with respect to $W(i)$. It is assumed that $X(i)$ and $d(i)$ are possibly nonstationary random processes with zero means and finite second and fourth moments, and that the covariance matrices $R(i)$ for the vectors $X(i)$ are positive definite. It is also assumed that $X(i)$ and $d(i)$ are each independent sequences. This significantly simplifying assumption of independent training vectors, which is commonly made in analyses of (1), is rarely true in practice; it is generally true only if the adaptation rate of the algorithm (1) is reduced by incrementing the weight vector only once every K units of time (in which case i is replaced with iK) and K is chosen to be sufficiently large. Nevertheless, theoretical predictions obtained on the basis of this independence assumption have proven useful in

Manuscript received October 27, 1986; revised March 26, 1987. This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR80-0189, 1980–82.

The author is with the Signal and Image Processing Laboratory, Department of Electrical and Computer Engineering, University of California, Davis.

IEEE Log Number 8715935

[1] The step-size parameter μ is the same as that defined in [1] and [2] but differs from that in [3] by a factor of two.

0098-4094/87/1000-1199$01.00 ©1987 IEEE

practice (cf. [3]), and this is confirmed by simulations presented herein. It is also assumed that $e_0(i)$ an $\boldsymbol{X}(i)$ are independent, where $e_0(i)$ is the error that would be realized with the minimum-mean-squared-error weight vector

$$\boldsymbol{W}_0(i) \triangleq \boldsymbol{R}^{-1}(i)\boldsymbol{P}(i) \tag{2}$$

for which $\boldsymbol{P}(i)$ is the covariance vector for $d(i)$ and $\boldsymbol{X}(i)$ (cf. [4]). This simplifying assumption is easily verified for various applications; for example, it is valid for $\boldsymbol{X}(i)$ and $d(i)$ jointly Gaussian (cf. [4]) or for the system identification problem treated in Section III, for which $d(i) = \tilde{\boldsymbol{W}}^T(i)\boldsymbol{X}(i) + n(i)$ for some sequence of system weight vectors $\tilde{\boldsymbol{W}}(i)$ and some measurement noise sequence $n(i)$ that is independent of $\boldsymbol{X}(i)$.

Using only the above assumptions, it is shown in the Appendix that the excess mean-squared error

$$\epsilon_*(i) \triangleq \epsilon(i) - \epsilon_0(i) \tag{3a}$$

in which $\epsilon(i) \triangleq E\{e^2(i)\}$ and (cf. [4])

$$\epsilon_0(i) \triangleq E\{e_0^2(i)\} = \sigma_d^2(i) - \boldsymbol{P}^T(i)\boldsymbol{R}^{-1}(i)\boldsymbol{P}(i) \tag{3b}$$

is bounded above and below by the solutions to first-order linear recursions

$$\epsilon_{\min}(i) \leqslant \epsilon_*(i) \leqslant \epsilon_{\max}(i) \tag{4}$$

where $\epsilon_m(i)$ $(m = \min, \max)$ satisfy

$$\epsilon_m(i+1) = \gamma_m(i)\epsilon_m(i) + \beta(i) \tag{5}$$

with $\epsilon_m(0) = \epsilon_*(0)$. In (5), $\gamma_m(i)$ are the extreme eigenvalues of the matrix

$$\boldsymbol{H}(i) \triangleq \frac{1}{2}\left[\boldsymbol{F}(i) + \boldsymbol{F}^T(i)\right] \tag{6a}$$

where

$$\boldsymbol{F}(i) \triangleq \boldsymbol{R}^{-1/2}(i)E\{\boldsymbol{A}(i)\boldsymbol{R}(i+1)\boldsymbol{A}(i)\}\boldsymbol{R}^{-1/2}(i) \tag{6b}$$

and

$$\boldsymbol{A}(i) \triangleq \boldsymbol{I} - 2\mu\boldsymbol{X}(i)\boldsymbol{X}^T(i). \tag{7}$$

The driving term $\beta(i)$ in (5) consists of two components

$$\beta(i) = \beta_\nabla(i) + \beta_\Delta(i) \tag{8}$$

where

$$\beta_\nabla(i) \triangleq 4\mu^2\epsilon_0(i)b(i) \tag{9}$$

$$\beta_\Delta(i) \triangleq \left[\Delta^T(i) - 2\bar{\boldsymbol{V}}^T(i)\bar{\boldsymbol{A}}(i)\right]\boldsymbol{R}(i+1)\Delta(i). \tag{10}$$

In (9)

$$b(i) \triangleq \operatorname{tr}\{\boldsymbol{R}(i+1)\boldsymbol{R}(i)\} \tag{11}$$

and, in (10), the overbar denotes expected value and the vector sequence

$$\bar{\boldsymbol{V}}(i) \triangleq \bar{\boldsymbol{W}}(i) - \boldsymbol{W}_0(i) \tag{12}$$

satisfies

$$\bar{\boldsymbol{V}}(i+1) = \bar{\boldsymbol{A}}(i)\bar{\boldsymbol{V}}(i) - \Delta(i) \tag{13}$$

where

$$\Delta(i) \triangleq \boldsymbol{W}_0(i+1) - \boldsymbol{W}_0(i) \tag{14}$$

$$\bar{\boldsymbol{A}}(i) = \boldsymbol{I} - 2\mu\boldsymbol{R}(i). \tag{15}$$

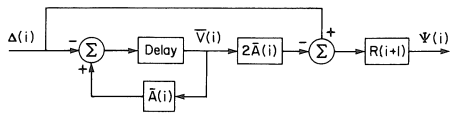


Fig. 1. System model of vector recursion that determines the convergence inhibitor $\beta_\Delta(i)$ due to nonstationarity.

If the N elements of $\boldsymbol{X}(i)$ are i.i.d., then $\boldsymbol{H}(i)$ is diagonal and Toeplitz and therefore $\gamma_{\max} = \gamma_{\min}$. Thus, the upper and lower bounds (4) coincide to give the exact mean-squared error in this case.

If $\boldsymbol{X}(i)$ and $d(i)$ are jointly stationary, $\Delta(i) \equiv \boldsymbol{0}$ and therefore $\beta_\Delta(i) \equiv 0$. Hence, nonstationarity is a necessary source of $\beta_\Delta(i)$. Notice, however, that $\boldsymbol{W}_0$ can be time-invariant even though both $\boldsymbol{X}(i)$ and $d(i)$ are nonstationary. This occurs in the problem of identification of a time-invariant system with nonstationary excitation and measurement noise. In this case, $\Delta(i) \equiv 0$ and therefore $\beta_\Delta(i) \equiv 0$. If the stochastic gradient of the squared error used in (1) is replaced with the nonstochastic gradient, then the socalled gradient noise (Section IV) is identically zero and therefore $\beta_\nabla(i)$ (and also $\beta_\Delta(i)$) vanishes (cf. [3]). Hence, gradient noise is a necessary source of $\beta_\nabla(i)$ (and of $\beta_\Delta(i)$). Notice, however, that ϵ_0 can be zero even though the gradient noise is nonzero. This occurs in the problem of system identification when there is no measurement noise. In this case, $\beta_\nabla(i) \equiv 0$. Since $\epsilon_{\max}(i)$ and $\epsilon_{\min}(i)$ will converge to zero only if $\beta(i)$ is zero or converges to zero, then $\beta_\nabla(i)$ and $\beta_\Delta(i)$ can be interpreted as *convergence inhibitors*.

It follows from (10), (13), and (14) that the convergence inhibitor due to nonstationarity $\beta_\Delta(i)$ can be interpreted as an inner product of the input and output vectors of a time-varying first-order linear vector recursion

$$\beta_\Delta(i) = \Psi^T(i)\Delta(i) \tag{16}$$

where $\Psi(i)$ is the output, corresponding to the input $\Delta(i)$, of the vector recursion represented by the signal-flow diagram shown in Fig. 1. Furthermore, $\beta_\Delta(i)$ admits a particularly simple interpretation in the special case for which only $d(i)$ is nonstationary (as in the application considered in Section III). That is, when $\boldsymbol{X}(i)$ is stationary, the vector recursion shown in Fig. 1 is time-invariant and has the system transfer function matrix (z-transform of unit pulse response matrix) Φ given by

$$\Phi(z) \triangleq \boldsymbol{R}[\boldsymbol{I}z + \bar{\boldsymbol{A}}][\boldsymbol{I}z - \bar{\boldsymbol{A}}]^{-1}. \tag{17}$$

By using a transformation of coordinates to diagonalize Φ, (16) can be reexpressed as

$$\beta_\Delta(i) = [\Psi'(i)]^T\Delta'(i) \tag{18}$$

where

$$\Delta'(i) = \boldsymbol{Q}^{-1}\Delta(i). \tag{19}$$

$\boldsymbol{Q}$ is the matrix of eigenvectors of $\boldsymbol{R}$ (cf. [3]), and $\Psi'(i)$ is the output (corresponding to the input $\Delta'(i)$) of the system with diagonal transfer function matrix $\Phi'(z)$ whose kth

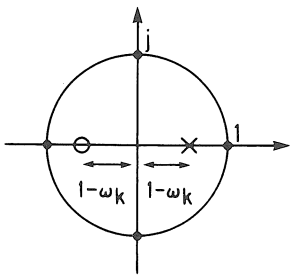


Fig. 2. Pole-zero diagram for the system model shown in Fig. 1 for the special case of stationary $X(i)$.

diagonal term is

$$\Phi'_{kk}(z) = \left[\frac{z+1-\omega_k}{z-1+\omega_k}\right]\lambda_k \tag{20}$$

where λ_k is the kth eigenvalue of $\boldsymbol{R}$ and $\omega_k \triangleq 2\mu\lambda_k$. The pole-zero diagram of this transfer function is shown in Fig. 2. For values of μ satisfying $\mu \ll 1/2\lambda_k$, the pole and zero approach the unit circle and Φ'_{kk} becomes a low-pass filter with high gain ($\gg \lambda_k$). For values of μ satisfying $\mu \simeq 1/2\lambda_k$, the pole and zero approach the origin and Φ'_{kk} becomes an all-pass filter with gain λ_k. Since the size of the output $\Psi'(i)$ determines the size of $\beta_\Delta(i)$, we would like $\Psi'(i)$ to be as small as possible. This involves a tradeoff between small bandwidth and small gain.

Example: As an example of the feedback factors $\gamma_m(i)$ in the bounding recursions (5), we consider stationary Gaussian vectors $X(i)$. In this case, (6) reduces to $\boldsymbol{H}(i) = \boldsymbol{G}$, where (cf. [3])

$$\boldsymbol{G} \triangleq [\boldsymbol{I} - 2\mu\boldsymbol{R}]^2 + 4\mu^2\left(\boldsymbol{R}^2 + \operatorname{tr}\{\boldsymbol{R}^2\}\boldsymbol{I}\right) \tag{21}$$

from which it follows that (cf. [3])

$$\gamma_{\substack{\max \\ (\min)}} = \max_{(\min)}\left\{\left[1-2\mu\lambda_j\right]^2 + 4\mu^2\left(\lambda_j^2 + N\lambda_{\text{rms}}^2\right)\right\} \tag{22}$$

where λ_{rms} is the root-mean-square value of the N eigenvalues $\{\lambda_i\}$ of $\boldsymbol{R}$ ($\lambda_{\text{rms}}^2 = \operatorname{tr}\{\boldsymbol{R}^2\}/N$). If $\mu < 1/4\lambda_{\max}$, then

$$\gamma_{\substack{\max \\ (\min)}} = \left(1-2\mu\lambda_{\substack{\min \\ (\max)}}\right)^2 + 4\mu^2\left(\lambda^2_{\substack{\min \\ (\max)}} + N\lambda_{\text{rms}}^2\right). \tag{23}$$

It follows that the separation of the upper and lower bounds (5), which is determined by $\gamma_{\max} - \gamma_{\min}$, can be large when $\lambda_{\max} - \lambda_{\min}$ is large, which occurs when the elements of $X(i)$ are highly correlated. However, as the step size μ is decreased, $\gamma_{\max} - \gamma_{\min}$ decreases for a given $\lambda_{\max} - \lambda_{\min}$.

A performance parameter of particular interest is the average value of the time-variant excess mean-squared error $\epsilon_*(i)$, after initial transients of adaptation have died out. If the time-series of matrices $\boldsymbol{R}(i)$ and $\boldsymbol{P}(i)$ are assumed to possess ergodic properties, then the idealized average

$$\langle\epsilon_*\rangle \triangleq \lim_{Z\to\infty}\frac{1}{Z}\sum_{i=1}^{Z}\epsilon_*(i) \tag{24}$$

in which initial transient effects vanish, can be used. It follows from (4) that

$$\langle\epsilon_{\min}\rangle \leqslant \langle\epsilon_*\rangle \leqslant \langle\epsilon_{\max}\rangle \tag{25}$$

and equality holds in (25) when the elements of $X(i)$ are i.i.d., as discussed in connection with (4). If it is assumed that $X(i)$ is stationary (as it is in the application considered in Section III), then $\boldsymbol{F}(i)$ is independent of time i and, as a result, the eigenvalues $\gamma_m(i)$ are independent of i. Consequently, the time-averaged values of the solutions to the recursions (5) can be obtained simply by equating the time-averaged values of both sides of (5)

$$\langle\epsilon_m\rangle = \gamma_m\langle\epsilon_m\rangle + \langle\beta\rangle \tag{26}$$

which yields

$$\langle\epsilon_m\rangle = \frac{\langle\beta\rangle}{1-\gamma_m}. \tag{27}$$

Moreover, when $\boldsymbol{R}(i)$ is independent of i because $X(i)$ is stationary, then the two components $\langle\beta_\Delta\rangle$ and $\langle\beta_\nabla\rangle$ of $\langle\beta\rangle$ in (27) can be evaluated as follows. Substitution of (11) into (9) yields

$$\langle\beta_\nabla\rangle = 4\mu^2 N\lambda_{\text{rms}}^2\langle\epsilon_0\rangle. \tag{28}$$

To obtain $\langle\beta_\Delta\rangle$, we simply note that it follows from (16) that $\langle\beta_\Delta\rangle$ is the sum of time-average cross-correlations of the elements of the input vector $\Delta(i)$ and the output vector $\Psi(i)$ of the linear time-invariant system depicted in Fig. 1. Thus, we have, as a standard result from the second-order theory of stationary time-series (cf. [4]), that

$$\begin{aligned}\langle\beta_\Delta\rangle &= \sum_{n=-\infty}^{\infty}\operatorname{tr}\left\{\phi^T(n)\hat{\boldsymbol{R}}_\Delta(n)\right\} \\ &= \int_{-1/2}^{1/2}\operatorname{tr}\left\{\Phi^T(e^{j2\pi f})\hat{\boldsymbol{S}}_\Delta(f)\right\}df\end{aligned} \tag{29}$$

where the transfer-function matrix $\Phi(z)$ is given by (17) and $\hat{\boldsymbol{S}}_\Delta(f)$ is the cross-spectral density matrix for $\Delta(i)$

$$\hat{\boldsymbol{S}}_\Delta(f) = \sum_{n=-\infty}^{\infty}\hat{\boldsymbol{R}}_\Delta(n)e^{j2\pi nf} \tag{30}$$

in which $\hat{\boldsymbol{R}}_\Delta(n)$ is the cross-correlation matrix

$$\hat{\boldsymbol{R}}_\Delta(n) = \langle\Delta(i+n)\Delta^T(i)\rangle. \tag{31}$$

Substitution of (14) into (31) yields

$$\hat{\boldsymbol{R}}_\Delta(n) = 2\hat{\boldsymbol{R}}_{\tilde{W}}(n) - \hat{\boldsymbol{R}}_{\tilde{W}}(n-1) - \hat{\boldsymbol{R}}_{\tilde{W}}(n+1). \tag{32}$$

Thus, in this case for which $X(i)$ is stationary, the dependence on the nonstationarity in $d(i)$ of the upper and lower bounds (27) on the performance parameter $\langle\epsilon_*\rangle$ can be completely specified in terms of i) the time-average of the minimum mean-squared error

$$\langle\epsilon_0\rangle = \langle\sigma_d^2\rangle - \langle\boldsymbol{P}^T\boldsymbol{R}^{-1}\boldsymbol{P}\rangle \tag{33}$$

in (28), and ii) the cross-correlation matrix $\hat{\boldsymbol{R}}_{\tilde{W}}$ in (29), (30), and (32) for the sequence of optimum weight-vectors.

In the next section, the formulas (28) and (29) are explicitly evaluated for the system identification problem in which the excitation of the system is stationary white

noise. The resultant explicit formula for the exact value of $\langle \epsilon_* \rangle$, which is given by (27) and (8), is then used to study the step-size optimization problem. However, before proceeding, it should be emphasized that the study of the dependence of the upper and lower bounds (25) on the nonstationarity of the training vectors is considerably more difficult when $X(i)$, as well as $d(i)$, is nonstationary, because then the bounding recursions are time-variant and therefore (26) does not apply.

III. Time-Variant System Identification

For the system identification problem described in [1] and [2], we have

$$d(i) = \tilde{W}^T(i) X(i) + n(i) \tag{34}$$

where $\tilde{W}$ is the N-vector corresponding to the unit-pulse-response of the unknown system which is assumed to have memory length $\leqslant N$, $X(i)$ is the N-vector of samples of the excitation sequence $x(j)$

$$X(i) = [x(i), x(i-1), x(i-2), \ldots, x(i-N+1)]^T \tag{35}$$

and $n(i)$ is measurement noise at the system output. It is assumed that the system excitation $x(i)$ is a stationary process, in which case $X(i)$ is stationary. Consequently, the upper and lower bounds on the excess mean-squared error are given by (27). If it is further assumed that the system excitation is white noise (i.i.d.), then $\lambda_1 = \lambda_2 = \lambda_3 = \cdots = \lambda_N = \sigma_x^2$ and $\gamma_{\max} = \gamma_{\min} \triangleq \gamma$. Therefore, (25) and (27) yield the exact value

$$\langle \epsilon_* \rangle = \frac{\langle \beta \rangle}{1-\gamma} \tag{36}$$

for the performance parameter. Moreover, the unique eigenvalue γ is, from (6), given explicitly by (cf. [3])

$$\begin{aligned} \gamma &= \left(1-2\mu\sigma_x^2\right)^2 + 4\mu^2(\kappa_x + N - 2)\sigma_x^4 \\ &= 1 - 4\mu\sigma_x^2 + 4\mu^2(\kappa_x + N - 1)\sigma_x^4 \end{aligned} \tag{37}$$

where the parameter

$$\kappa_x \triangleq \frac{E\{x^4(i)\}}{\left(E\{x^2(i)\}\right)^2} \tag{38}$$

is the kurtosis of the zero-mean random variable $x(i)$.

The two components of the numerator in (36), which are specified by (28) and (29), can be evaluated explicitly as follows. Since $W_0(i) = \tilde{W}(i)$, then $e_0(i) = n(i)$, and therefore $\langle \epsilon_0 \rangle = \sigma_n^2$, which yields (from (28))

$$\langle \beta_\nabla \rangle = 4\mu^2 N \sigma_x^4 \sigma_n^2. \tag{39}$$

Also, since

$$R = \sigma_x^2 I \tag{40}$$

then (17) and (15) yield

$$\Phi(z) = \sigma_x^2 \left[\frac{z + 1 - 2\mu\sigma_x^2}{z - 1 + 2\mu\sigma_x^2}\right] I. \tag{41}$$

In order to proceed with the evaluation of either the sum or the integral in (29), we must adopt a specific model for the unknown system $\tilde{W}(i)$. In order to facilitate comparison with previous studies of the LMS algorithm for time-variant system identification [1], [2], we adopt the same model used previously. Specifically, it is assumed that the N elements fluctuate independently of each other according to first-order Markov time-series. That is, the time-average cross-correlation matrix for $\tilde{W}(i)$ is given by[2]

$$\hat{R}_{\tilde{W}}(n) = \sigma_w^2 r^{-|n|} I \tag{42}$$

for some r such that $|r| < 1$. Using (32), (41), and (42), it can be shown that (29) reduces to

$$\langle \beta_\Delta \rangle = \frac{4\mu\sigma_x^4\sigma_w^2 N(1-r)}{1 - r\left(1 - 2\mu\sigma_x^2\right)}. \tag{43}$$

It is convenient to normalize the time-averaged excess mean-squared error $\langle \epsilon_* \rangle$ by the time-averaged minimum attainable mean-squared error $\langle \epsilon_0 \rangle$ to obtain a time-averaged counterpart (for nonstationary training vectors) of what is commonly called the *fractional misadjustment* parameter

$$M \triangleq \frac{\langle \epsilon_* \rangle}{\langle \epsilon_0 \rangle}. \tag{44}$$

It follows from (8), (36), (37), (39), and (43) that this misadjustment is given by the sum of two components

$$M = M_\Delta + M_\nabla \tag{45}$$

where

$$M_\nabla \triangleq \frac{\langle \beta_\nabla \rangle / \langle \epsilon_0 \rangle}{1-\gamma} = \frac{\mu N \sigma_x^2}{1 - \mu(\kappa_x + N - 1)\sigma_x^2} \tag{46}$$

$$\begin{aligned} M_\Delta \triangleq \frac{\langle \beta_\Delta \rangle / \langle \epsilon_0 \rangle}{1-\gamma} &= \left[\frac{1}{1 - \mu(\kappa_x + N - 1)\sigma_x^2}\right] \\ &\quad \cdot \left[\frac{\rho(1-r)}{1 - r\left(1 - 2\mu\sigma_x^2\right)}\right] \end{aligned} \tag{47}$$

and the parameter

$$\rho \triangleq N\sigma_x^2\sigma_w^2/\sigma_n^2 \tag{48}$$

is the ratio of time-averaged mean-squared signal to time-averaged mean-squared noise (SNR) at the output $d(i)$ of the unknown system.

The general formulas (45)–(47) can be used to study the dependence of misadjustment on the step-size μ and the degree of nonstationarity, which is determined by r. In order to reduce the number of parameters, we shall focus on Gaussian data in which case $\kappa_x = 3$. The formulas (45)–(47) can be reparameterized as

$$M = \left[c\mu + \frac{\rho}{1 + a\mu}\right] \frac{1}{1 - b\mu} \tag{49}$$

[2] Unlike the model in [1] and [2], for which $\sigma_w^2 = \sigma^2/(1-r^2)$, σ_w^2 is here taken to be independent of r, so that the degree of nonstationarity DNS $= 1 - r$ can be varied independently of the system gain and therefore output SNR (48).

where

$$a \triangleq \frac{2r\sigma_x^2}{1-r} \tag{50a}$$

$$b \triangleq (N+2)\sigma_x^2 \tag{50b}$$

$$c \triangleq N\sigma_x^2. \tag{50c}$$

Using (49), it can be shown that the step-size μ_0 that minimizes the misadjustment is given by

$$\mu_0 = \left\{ \left[1 + \frac{(a-b-c/\rho)c\rho}{(b\rho+c)^2} \right]^{1/2} - 1 \right\} \frac{b\rho+c}{ac} \tag{51}$$

provided that the parameter r is in the range

$$1 > r > r_c. \tag{52}$$

The critical value r_c is specified by

$$r_c \triangleq \left[1 + \frac{2}{2+N(1+1/\rho)} \right]^{-1} > 0. \tag{53}$$

For $r < r_c$, there is no optimum step size in the admissable range, for which $|\gamma| < 1$ so that the algorithm is stable. Since r determines the rate of the fluctuations of the time-variant system, this condition indicates that the *degree of nonstationarity* (DNS) cannot be too high if an optimum step size is to exist. A convenient definition for DNS is

$$\text{DNS} \triangleq 1 - r. \tag{54}$$

Then DNS = 0 for a stationary (time-invariant) system, and the maximum value of D is 1 for over-damped nonstationarity, and 2 for underdamped nonstationarity (cf. (42)). It can be seen from (53) that the critical value of DNS $(1-r_c)$ decreases (the situation worsens) as either N, the number of weights, increases or ρ, the SNR, decreases. If r is outside the range (52) where an optimum exists, then either $r < 0$ or $0 < r < r_c$, and in both cases the misadjustment possesses an infinum at $\mu = 0$, that is, M (and, in fact, both M_∇ and M_Δ) decrease monotonically, and M approaches the minimum $M = \rho$ as $\mu \to 0$. More specifically, $M_\nabla \to 0$ and $M_\Delta \to \rho$ as $\mu \to 0$. In this case, the nonstationarity cannot be tracked and therefore misadjustment is made smallest by minimizing the effects of gradient noise which requires minimizing μ. The adaptive filter converges to the best time-invariant system in this case.

The two extreme cases of high degree of nonstationarity $(r \to r_c)$ and low degree of nonstationarity $(r \to 1)$ are of particular interest. It follows from (49)–(51) that as $r \to 1$

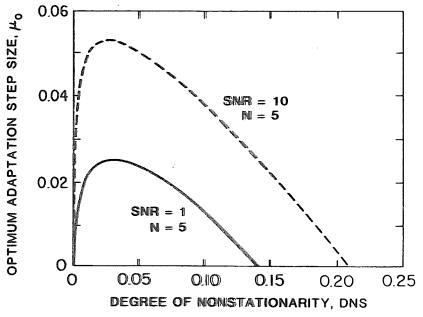
$$\mu_0 \to \frac{\left[\rho(1-r^2)/N\right]^{1/2}}{2\sigma_x^2} \to 0 \tag{55}$$

$$M(\mu_0) \to \left[N\rho(1-r^2)\right]^{1/2} \to 0 \tag{56}$$

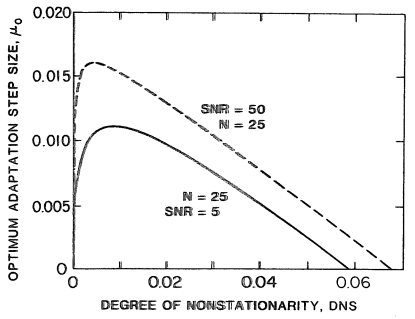
and as $r \to r_c$ (assuming $r_c \nrightarrow 1$)

$$\mu_0 \to \frac{\left[2+N(1+1/\rho)\right]^{-1} - (1-r)/2r}{2\sigma_x^2} \to 0 \tag{57}$$

$$M(\mu_0) \to \rho. \tag{58}$$

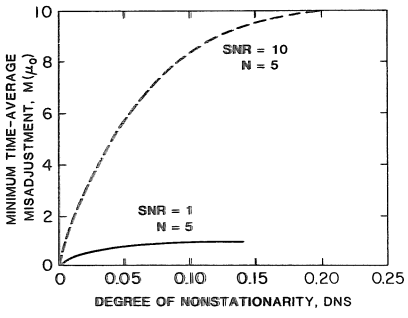


(a)

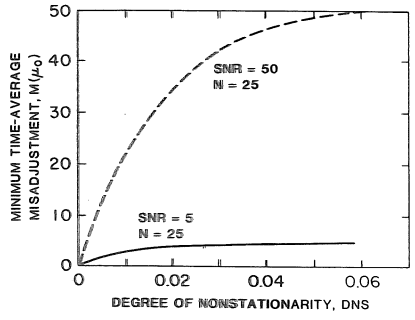


(b)

Fig. 3. Optimum adaptation step-size parameter μ_0 as a function of degree of nonstationarity DNS. (a) $N = 5$, SNR = 1, 10. (b) $N = 25$, SNR = 5, 50.



(a)



(b)

Fig. 4. Minimum time-average misadjustment $M(\mu_0)$ as a function of degree of nonstationarity DNS. (a) $N = 5$, SNR = 1, 10. (b) $N = 25$, SNR = 5, 50.

It follows that μ_0 has a maximum within the range (52) of DNS.

To illustrate the dependence of the optimal step-size μ_0 and the minimum misadjustment $M(\mu_0)$ on the filter-length parameter N, the degree-of-nonstationarity parameter r, and the signal-to-noise-ratio parameter ρ, Figs. 3–6

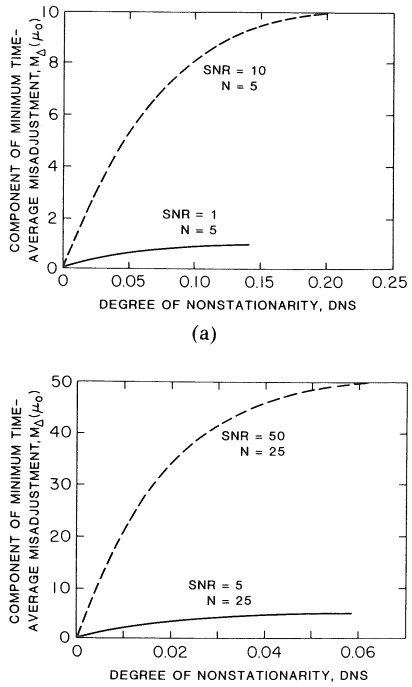


Fig. 5. Component of minimum time-average misadjustment $M_\Delta(\mu_0)$, due to nonstationarity, as a function of degree of nonstationarity DNS. (a) $N = 5$, SNR $= 1$, 10. (b) $N = 25$, SNR $= 5$, 50.

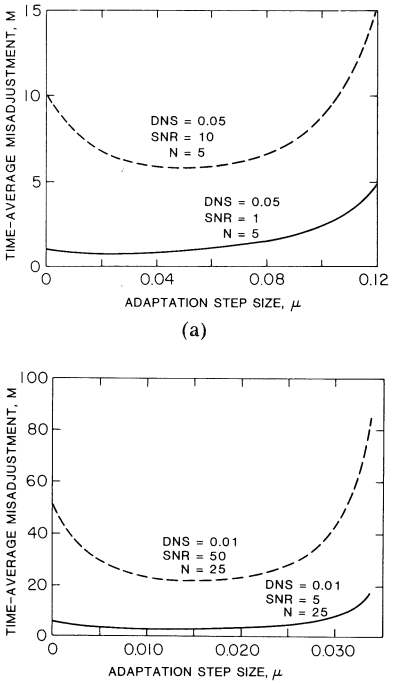


Fig. 7. Time-average misadjustment $M(\mu)$ as a function of adaptation step-size μ. (a) $N = 5$, SNR $= 1$, 10, DNS $= 0.05$. (b) $N = 25$, SNR $= 5$, 50, DNS $= 0.01$.

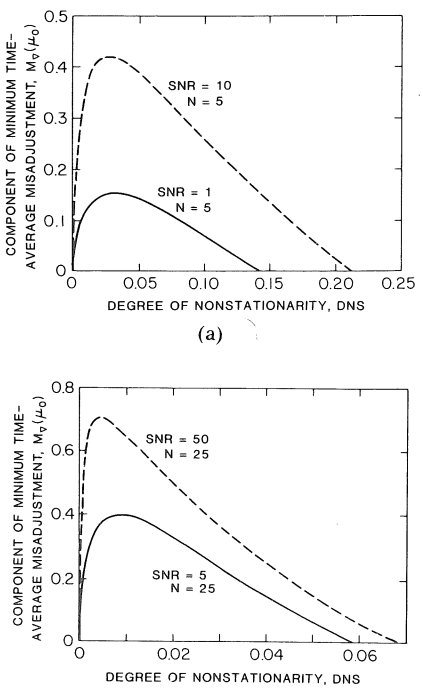


Fig. 6. Component of minimum time-average misadjustment $M_\nabla(\mu_0)$, due only to gradient noise, as a function of degree of nonstationarity DNS. (a) $N = 5$, SNR $= 1$, 10. (b) $N = 25$, SNR $= 5$, 50.

show graphs of μ_0, $M(\mu_0)$, $M_\nabla(\mu_0)$, and $M_\Delta(\mu_0)$ as functions of DNS for several values of N and $\rho = \text{SNR}$. Since the only effect of σ_x^2 is to scale μ, σ_x^2 was set equal to unity. As shown in Fig. 3, μ_0 increases from 0 as the degree of nonstationarity increases from 0 (r decreases from 1) and reaches a peak (between $r = 0.95$ and 1 in all cases considered) and then μ_0 decreases (at a slower rate) to 0 as r approaches the critical value r_c. As shown in Fig. 4, $M(\mu_0)$ increases from 0 as the degree of nonstationarity increases from 0, and approaches a maximum (ρ) as the degree of nonstationarity approaches the critical value ($r \to r_c$). For higher degrees of nonstationarity ($r < r_c$), M cannot be made less than ρ but can be made arbitrarily close to ρ by choosing a sufficiently small value of μ, say μ_*. Specifically, it follows from (45)–(48) that

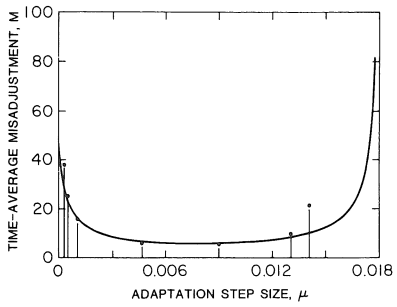
$$\mu_* = a \min\left\{\rho/N\sigma_x^2, 1/(\kappa_x + N - 1)\sigma_x^2\right\} \qquad (59)$$

where $a \ll 1$, say $a = 1/10$.

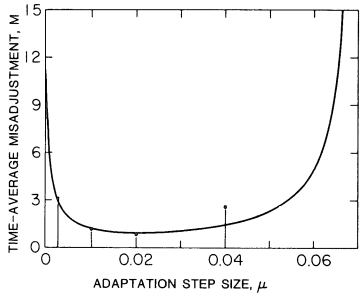
It should be mentioned that since the minimum M in the worst case equals the SNR ρ, it would appear that the worst case is better when the SNR is smaller. But recall that M is *fractional* misadjustment. Consider as an alternative the time-averaged mean-squared error, normalized by the time-averaged mean-squared value of the quantity being estimated, $d(i)$. When $M = \rho$, this quantity is equal to $1 + 1/\rho$. Therefore, decreasing the SNR does indeed increase this measure of error, even though M decreases.

No optimization study is complete without some analysis of sensitivity. In order to show how sensitive performance (i.e., misadjustment M) is to deviations of the step-size μ from its optimum value of μ_0, graphs of M as a function of μ for several values of N, SNR and DNS are displayed in Fig. 7(a) and (b). It can be seen that M can, for example, be increased by a factor of from 2 to 5 by increasing μ from μ_0 by a factor of from 2 to 4.

Simulations: In order to corroborate the theoretical results presented here, simulations were conducted for two sets of parameters: i) $N = 25$, $\rho = 50$, DNS $= 0.002$, and ii)



(a)



(b)

Fig. 8. Time-average misadjustment M as a function of adaptation step-size μ: Theoretical (smooth curve) and simulated (data points). (a) $N = 25$, SNR = 50, DNS = 0.002. (b) $N = 5$, SNR = 10, DNS = 0.004.

$N = 5$, $\rho = 10$, DNS = 0.004, and the resultant time-averaged misadjustment was measured for various values of the step-size parameter μ. In order to measure the mean-squared error $\epsilon(i)$, an ensemble of 300 statistically independent sample paths of $x(i)$ and $n(i)$, both from unit-variance white Gaussian noise generators (software), was averaged over. In order to measure the time-average of $\epsilon(i)$, the 3583 time samples from $i = 512$ to $i = 4095$ were averaged (initial transients had died away by time $i = 512$). Each of the N weight sequences in the unknown system weight vector was an independent sample path of a Gauss-Markov process. These N sample paths were fixed throughout the ensemble of excitation sequences $x(i)$ and measurement noise sequences $n(i)$. The results are shown in Fig. 8(a) and (b). The theoretical graph of $M(\mu)$ is shown as a solid curve, and the data points from the simulations are shown superimposed. It can be seen that agreement between theory and simulation is quite good.

IV. Comparison with Other Work

As explained in [3], many studies of the LMS algorithm use the assumption of a small step-size μ to make various simplifying assumptions (cf. [12]–[14]). But for values of μ that are not sufficiently small, the theoretical results obtained do not always agree with less approximate results obtained without the assumption of small μ. Although the discrepancies described in [3] are all for the case of stationary training vectors, discrepancies can also arise for the case of nonstationary training vectors studied in this paper. In fact, the problem is *potentially* more critical for nonstationary training vectors because a major objective in this case is to solve for and study the optimum step size, which need not be sufficiently small to justify certain

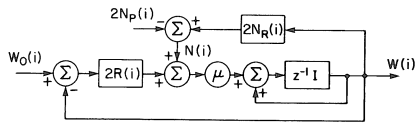


Fig. 9. Exact model of the LMS algorithm (1).

approximations. To investigate this, a comparison was made between the results presented here and those obtained in the previous studies [1], [2] of the same time-variant system identification problem.

In order to gain some insight into how the overall model of the LMS algorithm is affected by the commonly made small-μ approximations, we compare the exact model with the approximate model used in [1]. The exact model of interest can be obtained directly from (1) simply by substitution of the following identity for the negative stochastic gradient

$$-\nabla e^2(i) = 2e(i)\boldsymbol{X}(i) = 2\boldsymbol{R}(i)\left[\boldsymbol{W}_0(i) - \boldsymbol{W}(i)\right] + 2\left[\boldsymbol{R}(i) - \boldsymbol{X}(i)\boldsymbol{X}^T(i)\right]\boldsymbol{W}(i) - 2\left[\boldsymbol{P}(i) - d(i)\boldsymbol{X}(i)\right] \tag{60}$$

where $\boldsymbol{W}_0(i) = \boldsymbol{R}^{-1}(i)\boldsymbol{P}(i)$, $e(i) = d(i) - \hat{d}(i)$, and $\hat{d}(i) = \boldsymbol{X}^T(i)\boldsymbol{W}(i)$. This substitution yields

$$\boldsymbol{W}(i+1) = \boldsymbol{W}(i) + \mu\left\{2\boldsymbol{R}(i)\left[\boldsymbol{W}_0(i) - \boldsymbol{W}(i)\right] + \boldsymbol{N}(i)\right\} \tag{61}$$

where

$$2\boldsymbol{R}(i)\left[\boldsymbol{W}_0(i) - \boldsymbol{W}(i)\right] = -\nabla\epsilon(i) \tag{62}$$

is the negative nonstochastic gradient and

$$\boldsymbol{N}(i) \triangleq \nabla\epsilon(i) - \nabla e^2(i) = 2\boldsymbol{N}_R(i)\boldsymbol{W}(i) - 2\boldsymbol{N}_P(i) \tag{63}$$

is the socalled *gradient noise* in which

$$\boldsymbol{N}_R(i) \triangleq \boldsymbol{R}(i) - \boldsymbol{X}(i)\boldsymbol{X}^T(i) \tag{64}$$

$$\boldsymbol{N}_P(i) \triangleq \boldsymbol{P}(i) - \boldsymbol{X}(i)d(i). \tag{65}$$

A signal-flow diagram of the recursion in (61) and (63) is shown in Fig. 9. The components (64) and (65) of $\boldsymbol{N}(i)$ in (63) are called the *autocorrelation matrix noise* and the *cross-correlation vector noise*, respectively.

The model (61) of the LMS algorithm, as shown in Fig. 9, depicts $\boldsymbol{W}(i)$ as the response of a deterministic first-order linear recursion driven by a deterministic input $\boldsymbol{W}_0(i)$ plus a random noise input $\boldsymbol{N}(i)$. Unfortunately, part of the random noise input is obtained through a random feedback matrix from the system output $\boldsymbol{W}(i)$. Thus, the overall system is neither deterministic nor linear. This considerable complication is not discussed in [1] and [2], but ignoring it can be justified by arguing that if μ is sufficiently small and i is sufficiently large, then $\boldsymbol{W}(i)$ in (63) can be approximated by $\boldsymbol{W}_0(i)$, thereby decoupling the input noise $\boldsymbol{N}(i)$ from the output.

Now let us consider the accuracy of this and the corresponding small-μ approximations. The formulas given in [1, eqs. (51) and (74)] (but without the additional ap-

proximations [1, eqs. (75)–(77)]) for misadjustment are

$$M_{\nabla} \cong \mu N \sigma_x^2 \tag{66}$$

$$M_{\Delta} \cong \left[\frac{1}{1-\mu\sigma_x^2}\right]\left[\frac{\rho(1-r)}{1-r(1-2\mu\sigma_x^2)}\right]. \tag{67}$$

The discrepancy between (46) and (66) is the result of a small-μ approximation in [1] and [2], and we can see from (46) that it is negligible if (cf. [3])

$$\mu \ll \frac{1}{(\kappa_x + N - 1)\sigma_x^2}. \tag{68}$$

The discrepancy between (47) and (67) is the result of the assumption in [1] and [2] that M_{Δ} can be obtained simply by setting the gradient noise equal to zero. Although it is true that this makes $M_{\nabla} = 0$, it also changes the feedback factor γ. Specifically, (67) results from using $\gamma = (1 - 2\mu\sigma_x^2)^2$ (which is valid only for the nonstochastic gradient descent algorithm, cf. [3]) instead of (37). It follows from (37) that this will be a close approximation if (68) holds. By inspecting Figs. 7 and 8, it can be seen that the optimum step size does indeed satisfy (68) for the parameter values considered. However, a consequence of the discrepancy between (47) and (67) is that (67) predicts too small a misadjustment for large step sizes. For example, for large N, (67) is a factor of 2 smaller than (47) for $\mu = 1/2N\sigma_x^2$ and (67) is a factor of 3 smaller than (47) for $\mu = 2/3N\sigma_x^2$. This is particularly important in view of the simulations in Fig. 8, which reveal that even (47) predicts too small a misadjustment for large step sizes.

Another consequence of the discrepancy between (47) and (67) is that, unlike (47), (67) shows no dependence of performance on the kurtosis κ_x of the data. Experiment shows that the LMS algorithm does indeed behave differently for, say, uniformly distributed data ($\kappa_x = 1$) than for Gaussianly distributed data ($\kappa_x = 3$). An additional consequence is that when the first factor in (67) is ignored because of small μ, as done in [1] and [2], then M_{Δ} decreases monotonically as μ increases. However, M_{Δ} given by (47) is not necessarily monotonic in μ when r is not very close to unity.

Another comparison with the results in [1] and [2] can be made in order to determine the effects of the assumption of very slow nonstationarity used there. By making such an assumption before optimization of μ, the resultant formula for the optimum μ turns out to be the asymptotic result (55). Thus, (55) corroborates the result in [1] and [2]. However, the substantial difference between (55) and the more general formula (51) reveals that some of the general conclusions drawn in [1] and [2] do not apply when the nonstationarity is not very slow. Specifically, it is assumed in [1] and [2] that μ is sufficiently small and r is sufficiently close to unity to approximate (67) by [1, eq. (80)]

$$M_{\Delta} \cong \frac{\rho(1-r)}{2\mu\sigma_x^2}. \tag{69}$$

Using (66) and (69), it is shown that $M_{\Delta} = M_{\nabla}$ when their sum is minimized, and the minimal value of μ is [1, eq. (86)], [2, eq. (82)]

$$\mu_0 = \frac{1}{\sigma_x^2}\left[\frac{\rho(1-r)}{2N}\right]^{1/2}. \tag{70}$$

However, the exact formulas (46) and (47) yield highly unequal M_{Δ} and M_{∇} when their sum is minimized for values of r not very close to unity, but still close. For example, Figs. 5 and 6 shows that for DNS = 0.04 (r = 0.96), $M_{\Delta} \cong 10 M_{\nabla}$ when $M = M_{\nabla} + M_{\Delta}$ is minimum. Also, the optimum step size (51) is not monotonically decreasing in r (except for sufficiently small DNS) as it is in (70).

Because of this limited applicability of the results in [1] and [2], the significant conclusion drawn in [2], regarding the favorable comparison of the statistical efficiency of the LMS algorithm with that of the least-squares algorithm, cannot automatically be applied to nonstationarity that is not very slow.

V. Conclusions

A general approach to studying the tracking behavior of the LMS algorithm, which is based on upper and lower bounding linear recursions for excess mean-squared error, has been presented. The bounds are close together when the data $x(i)$ being adaptively filtered is not too highly correlated, and they are identical when there is no data correlation (white noise). The recursions are time-invariant when the data being filtered is stationary, but the time-varying driving terms reflect the nonstationarity of the desired signal $d(i)$. For the case of stationary $x(i)$, evaluation of the time-averaged value of the bounds on excess mean-squared error is quite tractable.

Application of this general approach to the time-variant system identification problem studied in [1] and [2] yields exact results that corroborate most of the approximate results in [1] and [2] but show that several results and conclusions require substantial modification when the assumption of very slow time variation made there is not satisfied. Specifically, the optimum step size is not, in general, a monotonically increasing function of degree of nonstationarity, and the two components of misadjustment, due to gradient noise and nonstationarity, are not equal in general when misadjustment is minimum. Also, there is some question concerning the favorable comparison of the LMS algorithm with the least-squares algorithm when the degree of nonstationarity is not very low.

It should be kept in mind that the analysis presented here assumes that the algorithm is implemented with infinite precision multiplication and addition. In practice, there is a tradeoff in the step size between two opposing effects of finite precision [15], [16], and it is conceivable that such a tradeoff could overshadow the tradeoff for optimum tracking with infinite precision that is studied here.

Appendix
Derivation of Bounding Recursions

Under the first of the two independence assumptions made in the text between (1) and (2), it can be shown (cf.

[3]) that[3]

$$\epsilon_*(i) = E\{V^T(i)R(i)V(i)\} \tag{A1}$$

where $V(i)$ is defined by analogy with (12). Also, the LMS algorithm (1) is easily manipulated into the form

$$V(i+1) = A(i)V(i) + B(i) - \Delta(i) \tag{A2}$$

where $A(i)$ and $\Delta(i)$ are defined by (7) and (14), and

$$B(i) \triangleq 2\mu e_0(i)X(i). \tag{A3}$$

Substitution of (A2) and (A3) into (A1), and use of the two independence assumptions, yields (cf. [3])

$$\epsilon_*(i+1) = E\{V^T(i)C(i)V(i)\} + \beta(i) \tag{A4}$$

where $\beta(i)$ is defined by (8)–(10), and

$$C(i) \triangleq E\{A(i)R(i+1)A(i)\}. \tag{A5}$$

Now, consider the quantity

$$\begin{aligned} c &\triangleq E\{V^T(i)C(i)V(i)\} \\ &= E\{V^T(i)R^{1/2}(i)F(i)R^{1/2}(i)V(i)\} \end{aligned} \tag{A6}$$

where

$$F(i) \triangleq R^{-1/2}(i)C(i)R^{-1/2}(i). \tag{A7}$$

This quantity can be upper and lower bounded by

$$c_{\min} \leqslant c \leqslant c_{\max} \tag{A8}$$

where

$$c_m \triangleq \gamma_m E\{V^T(i)R(i)V(i)\} \tag{A9}$$

for $m = \min$, max, and $\{\gamma_m\}$ are the extreme eigenvalues of the matrix $H(i)$ defined by (6). Substitution of (A1) into (A9), and (A9) and (A6) into (A8), yields the bounds

$$\epsilon_*(i+1) \underset{m=\min}{\overset{m=\max}{\lessgtr}} \gamma_m \epsilon_*(i) + \beta(i) \tag{A10}$$

on (A4). It follows from (A10) that $\epsilon_*(i)$ is bounded above and below by (4) and (5), which is the desired result.

Acknowledgment

The author expresses his gratitude to Dr. M. Hajivandi for his assistance with optimization of the step-size parameter μ in the system identification example, and to Mr. W. A. Brown for his assistance with the simulations.

[3]Although formula (A1) is commonly treated as being valid for the LMS algorithm (cf. [12]), it cannot be valid unless $W(i)$ is independent of $X(i)$, as explained in [3]).

References

[1] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, pp. 1151–1162, Aug. 1976.

[2] B. Widrow and E. Walach, "On the statistical efficiency of the LMS algorithm with nonstationary inputs," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 211–221, Mar. 1984.

[3] W. A. Gardner, "Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique," *Signal Processing*, vol. 6, pp. 113–133, Apr. 1984. (Errata: *Signal Processing*, vol. 12, p. 211, Mar. 1987.

[4] W. A. Gardner, *Introduction to Random Processes with Applications to Signals and Systems*. New York: Macmillan, 1985.

[5] T. P. Daniell and J. E. Brown III, "Adaptation in nonstationary applications," in *Proc. 9th IEEE Symp. Adaptive Processes* (Austin, TX), Dec. 1970.

[6] P. Monsen, "Linear estimation in an unknown quasi-stationary environment," *IEEE Trans. Syst., Man. Cyber.*, vol. 1, pp. 216–222, Jan. 1971.

[7] R. R. Bitmead and B. D. O. Anderson, "Exponentially convergent behavior of simple stochastic adaptive estimation algorithms," in *Proc. 17th IEEE Conf. on Decision and Control* (San Diego, CA), Jan. 1979, pp. 580–585.

[8] J. T. Rickard, M. J. Dentino, and J. R. Zeidler, "Detection performance of an adaptive processor in nonstationary noise," in *Proc. IEEE Conf. on Acoust., Speech, Signal Process.* (Washington, DC), Apr. 1979, pp. 136–139.

[9] D. C. Farden, "Tracking properties of adaptive signal processing algorithms," *IEEE Trans. Acoust., Speech, Signal. Process.*, vol. ASSP-29, pp. 439–446, 1981.

[10] N. J. Bershad, "Tracking characteristics of the LMS adaptive line enhancer-response to a linear chirp signal to noise," *IEEE Trans. Acoust., Speech, Signal. Process.*, vol. ASSP-28, pp. 504–516, Oct. 1980.

[11] J. R. Treichler, "Response of the adaptive line enhancer to chirped and doppler-shifted sinusoids," *IEEE Trans. Acoust. Speech, Signal. Process.*, vol. ASSP-28, pp. 343–348, June 1980.

[12] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

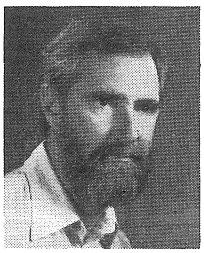
[13] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[14] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*. New York: Wiley, 1980.

[15] C. Caraiscos and B. Liu, "A round-off analysis of the LMS adaptive algorithm," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-32, pp. 34–41, Feb. 1984.

[16] J. M. Cioffi, "Limited precision effects in adaptive filtering," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 821–833, 1987.

✠



William A. Gardner (S'64–M'67–SM'84) was born in Palo Alto, CA, on November 4, 1942. He received the M.S. degree from Stanford University, Stanford, CA, in 1967, and the Ph.D. degree from the University of Massachusetts, Amherst, in 1972, both in electrical engineering.

He was a Member of the Technical Staff at Bell Laboratories in Massachusetts from 1967 to 1969. He has been a faculty member at the University of California, Davis, since 1972, where he is a Professor of Electrical and Computer Engineering. His research interests are in the general area of statistical signal processing, with primary emphasis on the theories of time-series analysis, stochastic processes, and signal detection and estimation. Professor Gardner is the author of *Introduction to Random Processes with Applications to Signals and Systems* (Macmillan, 1985) and *Statistical Spectral Analysis: A Nonprobabilistic Theory*, Prentice-Hall (in press).