A Variable Step Size LMS Algorithm

Raymond H. Kwong, Member, IEEE, and Edward W. Johnston

Abstract—A new LMS-type adaptive filter with a variable step size is introduced. The step size increases or decreases as the mean-square error increases or decreases, allowing the adaptive filter to track changes in the system as well as produce a small steady state error. The convergence and steady state behavior of the algorithm are analyzed. These results reduce to well-known ones when specialized to the constant step size case.

Simulation results are presented to support the analysis and to compare the performance of the new algorithm with the usual LMS algorithm and another variable step algorithm. They show that the performance of the new algorithm compares favorably with these existing algorithms.

I. INTRODUCTION

One of the most popular algorithms in adaptive signal processing is the least mean square (LMS) algorithm of Widrow and Hoff [1]. It has been extensively analyzed in the literature, and a large number of results on its steady state misadjustment and its tracking performance has been obtained [2]-[8]. The majority of these papers examine the LMS algorithm with a constant step size. The choice of the step size reflects a tradeoff between misadjustment and the speed of adaptation. In [1], approximate expressions were derived which showed that a small step size gives small misadjustment but also a longer convergence time constant. Subsequent works have discussed the issue of optimization of the step size or methods of varying the step size to improve performance [9], [10]. It seems to us, however, that there is as yet no detailed analysis of a variable step size algorithm that is simple to implement and is capable of giving both fast tracking as well as small misadjustment.

In this paper, we propose a variable step size LMS algorithm where the step size adjustment is controlled by the square of the prediction error. The motivation is that a large prediction error will cause the step size to increase to provide faster tracking while a small prediction error will result in a decrease in the step size to yield smaller misadjustment. The adjustment equation is simple to implement, and its form is such that a detailed analysis of the algorithm is feasible under the standard independence assumptions commonly made in the literature [1] to simplify the analysis of LMS algorithms.

The paper is organized as follows. In Section II, we formulate the adaptive system identification problem and describe the new variable step size LMS algorithm. Simplifying assumptions are introduced and their justification discussed. The analysis of the algorithm begins in Section III where the convergence of the mean weight vector is treated. In Section IV, we study the behavior of the mean-square error. Section V contains the steady state results. Conditions for convergence of the mean-square error are given. Expressions for the steady state misadjustment are also derived. In Section VI, simulation results obtained using the new algorithm are described. They are compared to the results obtained for the fixed step size algorithm and the variable step algorithm described in [9]. The improvements in performance over the constant step size algorithm are clearly shown. The simulation results are also shown to correspond closely to the theoretical predictions. Section VII contains the conclusions.

II. A VARIABLE STEP SIZE LMS ALGORITHM

The adaptive filtering or system identification problem being considered is to try to adjust a set of filter weights so that the system output tracks a desired signal. Let the input vector to the system be denoted by \( x_k \) and the desired scalar output be \( d_k \). These processes are assumed to be related by the equation

\[
d_k = x_k^T w_k^* + e_k
\]  

where \( e_k \) is a zero mean Gaussian independent sequence, independent of the input process \( x_k \). Two cases will be considered: \( w_k^* \) equals a constant \( W^* \), and \( w_k^* \) is randomly varying according to the equation

\[
w_{k+1} = a w_k^* + z_k
\]  

where \( a \) is less than but close to 1, and \( z_k \) is an independent zero mean sequence, independent of \( x_k \) and \( e_k \), with covariance \( E\{z_k z_k^T\} = \sigma_z^2 I \delta_{ij} \), \( \delta_{ij} \) being the Kronecker delta function. The first case will be referred to as a stationary system or environment, the second a nonstationary system or environment. They correspond to the models considered in [1]. The input process \( x_k \) is assumed to be a zero mean independent sequence with covariance \( E\{x_k x_k^T\} = R \), a positive definite matrix. This simplifying assumption is often made in the literature [1], [5], [7]. While it is usually not met in practice, analyses based on this assumption give predictions which are often validated in applications and simulations. This will also be the case with our results.
The LMS type adaptive algorithm is a gradient search algorithm which computes a set of weights \( W_k \) that seeks to minimize \( E(d_k - X_k^T W_k)^2 \). The algorithm is of the form

\[
W_{k+1} = W_k + \mu_k X_k \epsilon_k
\]

where

\[
e_k = d_k - X_k^T W_k
\]

and \( \mu_k \) is the step size. In the standard LMS algorithm [1], \( \mu_k \) is a constant. In [9], \( \mu_k \) is time varying with its value determined by the number of sign changes of an error surface gradient estimate. Here, we propose a new algorithm, which we shall refer to as the variable step size (VSS) algorithm, for adjusting the step size \( \mu_k \):

\[
\mu_{k+1} = \alpha \mu_k + \gamma \epsilon_k^2
\]

with

\[
0 < \alpha < 1, \quad \gamma > 0
\]

and

\[
\mu_{k+1} = \begin{cases} 
\mu_{\text{max}} & \text{if } \mu_{k+1} > \mu_{\text{max}} \\
\mu_{\text{min}} & \text{if } \mu_{k+1} < \mu_{\text{min}} \\
\mu_{k+1} & \text{otherwise}
\end{cases}
\]

where \( 0 < \mu_{\text{min}} < \mu_{\text{max}} \). The initial step size \( \mu_0 \) is usually taken to be \( \mu_{\text{max}} \), although the algorithm is not sensitive to the choice. As can be seen from (5), the step size \( \mu_k \) is always positive and is controlled by the size of the prediction error and the parameters \( \alpha \) and \( \gamma \). Intuitively speaking, a large prediction error increases the step size to provide faster tracking. If the prediction error decreases, the step size will be decreased to reduce the misadjustment. The constant \( \mu_{\text{max}} \) is chosen to ensure that the mean-square error (mse) of the algorithm remains bounded. A sufficient condition for \( \mu_{\text{max}} \) to guarantee bounded mse is [7]

\[
\mu_{\text{max}} \leq \frac{2}{3 \text{ tr } (R)}
\]

\( \mu_{\text{min}} \) is chosen to provide a minimum level of tracking ability. Usually, \( \mu_{\text{min}} \) will be near the value of \( \mu \) that would be chosen for the fixed step size (FSS) algorithm. \( \alpha \) must be chosen in the range \((0, 1)\) to provide exponential forgetting. A typical value of \( \alpha \) that was found to work well in simulations is \( \alpha = 0.97 \). The parameter \( \gamma \) is usually small \((4.8 \times 10^{-4} \text{ was used in most of our simulations})\) and may be chosen in conjunction with \( \alpha \) to meet the misadjustment requirements according to formulas presented later. The additional overhead over the FSS algorithm is essentially one more weight update at each time step, so that the increase in complexity is minimal.

### III. Convergence of the Mean Weight Vector

The VSS algorithm given by (3)-(6) is difficult to analyze exactly. To make the analysis tractable, we introduce the following simplifying assumption.

**Assumption 1:** For the algorithm (3)-(6),

\[
E(\mu_k X_k \epsilon_k) = E(\mu_k)E(X_k \epsilon_k)
\]

This assumption is of course true if \( \mu_k \) is a constant, but cannot really hold for the VSS algorithm. However, we can say that it is approximately true. This is because if \( \gamma \) is small, \( \mu_k \) will vary slowly around its mean value. By writing

\[
E(\mu_k X_k \epsilon_k) = E(\mu_k)E(X_k \epsilon_k) + E\{[\mu_k - E(\mu_k)]X_k \epsilon_k\}
\]

we see that for \( \gamma \) sufficiently small, the second term on the right-hand side of (7) will be small compared to the first. Assumption 1 allows us to derive theoretical results whose predictions are borne out by simulations. Making such simplifying assumptions is not an uncommon practice in the adaptive signal processing literature [1], [5], [7].

We first study the convergence of the mean weight vector. Since the stationary case can be derived from the nonstationary one by setting \( a = 1, \sigma^2 = 0 \) (resulting in \( Z_k = 0 \) with probability one), and \( W^* = W^* \), we shall give the derivation for the nonstationary case only.

By assumption 1,

\[
E(W_{k+1}) = E(W_k) + E(\mu_k)E(X_k \epsilon_k)
\]

\[
= E(W_k) - E(\mu_k)R E(W_k - W_k^*)
\]

Now,

\[
E(W_k^*) = aE(W_k^*)
\]

Thus the error weight vector \( \hat{W}_k = W_k - W_k^* \) satisfies the equation

\[
E(W_{k+1}) = [I - E(\mu_k)R]E(\hat{W}_k) + (1 - a)E(W_k^*)
\]

Equation (8) is stable if and only if

\[
\prod_{k=0}^{n} [I - E(\mu_k)R] \rightarrow 0, \quad \text{as } n \rightarrow \infty.
\]

A sufficient condition for (9) to hold is

\[
E(\mu_k) < \frac{2}{\lambda_{\text{max}}(R)}
\]

where \( \lambda_{\text{max}}(R) \) is the maximum eigenvalue of the matrix \( R \). Furthermore, for any \( |a| < 1 \), \( E(W_k^*) \xrightarrow{k \rightarrow \infty} 0 \). Hence under (9), \( E(\hat{W}_k) \xrightarrow{k \rightarrow \infty} 0 \). The stationary case where \( W_k^* = W^* \) is even simpler in that (8) becomes a homogeneous difference equation. Equation (9) is then a necessary and sufficient condition for \( E(\hat{W}_k) \xrightarrow{k \rightarrow \infty} W^* \).

A stronger but simpler sufficient condition for the convergence of \( E(W_k) \) to \( W^* \) is \( \mu_{\text{max}} < 2/\lambda_{\text{max}}(R) \). This condition is the same as that for the constant step size LMS algorithm.

The convergence of the mean weight vector is of course not sufficient to guarantee convergence of the mean-square error. In the next section, we shall derive equations which describe the behavior of the mean-square error.
IV. MEAN-SQUARE ERROR BEHAVIOR

As is the case for the regular LMS algorithm, the covariance of the weight vector is directly related to the mean-square error. We therefore first analyze the covariance of the weight vector. Let $\hat{W}_k$ be the error weight vector

$$\hat{W}_k = W_k - W_k^*.$$ 

In the nonstationary case, $\hat{W}_k$ satisfies the equation

$$\hat{W}_{k+1} = (I - \mu_k X_k X_k^T)\hat{W}_k + (1 - \alpha)W_k^* - Z_k + \mu_k X_k e_k.$$ 

Since $R$, the covariance matrix of $X_k$, is symmetric, there exist matrices $Q$ and $\Lambda$, with $\Lambda$ diagonal, such that $R = Q^T \Lambda Q$ and $Q^T Q = I$. Let $V_k = Q^T \hat{W}_k$, $X_k = Q^T X_k$, $W_k^* = Q^T W_k^*$, and $Z_k = Q^T Z_k$. Then,

$$E(V_{k+1} V_{k+1}^T) = E[(I - \mu_k X_k^T X_k) V_k V_k^T (I - \mu_k X_k^T X_k^T)] + E[(I - \alpha)W_k^* W_k^T (1 - \alpha)] + E[(1 - \alpha)^2 W_k^* W_k^T] + E(Z_k^T Z_k^T) + E(\mu_k^2 X_k^T X_k^T e_k^2).$$

To proceed further, we shall make the following additional assumptions.

Assumption 1': The step size $\mu_k$ is independent of $X_k$ and $V_k$.

Assumption 2: The components of $V_k$ are independent, conditionally Gaussian random variables given $\mu_k$. Assumption 1' is basically a strengthening of assumption 1. Justification for assumption 2 will be discussed at the end of this section. Now, assuming that $\alpha$ is close to 1 so that all terms with $(1 - \alpha)$ can be discarded, we have

$$E(V_{k+1} V_{k+1}^T) = E(V_k V_k^T) - \Delta E\{(\mu_k V_k V_k^T)\} - E(\mu_k V_k V_k^T) \Lambda + E(\mu_k^2 X_k^T X_k V_k V_k^T X_k^T) + \sigma^2 I + E(\mu_k^2 X_k^T X_k^T e_k^2).$$

(10)

Assume that $\mu_{k+1} = \mu_k$. Then

$$E(\mu_{k+1}) = \alpha E(\mu_k) + \gamma E(e_k).$$

(11)

$$E(\mu_{k+1}^2) = \alpha^2 E(\mu_k^2) + 2\alpha \gamma E(\mu_k e_k) + \gamma^2 E(e_k^2).$$

(12)

We can now use the Gaussian moment factoring theorem (see Appendix A) to simplify some of the expressions in the above equations. We have

$$E(\mu_k^2 X_k^T X_k V_k V_k^T X_k^T) = E(\mu_k^2)E(X_k^T X_k^T V_k V_k^T X_k^T) = E(\mu_k^2)[2\Delta E(V_k V_k^T) \Lambda + \Lambda \operatorname{tr}\{(\Lambda E[V_k V_k^T])\}].$$

(13)

Putting (13) into (10) yields

$$E(V_{k+1} V_{k+1}^T) = E(V_k V_k^T) - \Delta E(\mu_k E(V_k V_k^T)) - E(\mu_k)E(V_k V_k^T) \Lambda + E(\mu_k^2)\Lambda \operatorname{tr}\{(\Lambda E[V_k V_k^T])\} + \sigma^2 I + E(\mu_k^2) \xi_{\text{min}}.$$  

(14)

From the definition of $\xi_k$, and the independence of $e_k$,

$$E(e_k^2) = E(e_k^2 - 2e_k X_k V_k + V_k^T X_k^T V_k) = \xi_{\text{min}} + \gamma \operatorname{tr}\{(\Lambda E[V_k V_k^T])\}.  

(15)

In Appendix B, we derive the following approximate expression for $E(e_k^2)$:

$$E(e_k^2) = 3\xi_{\text{min}}^2 + 6\xi_{\text{min}} \gamma \gamma \operatorname{tr}\{(\Lambda E[V_k V_k^T])\} + 3\gamma^2 \xi_{\text{min}}^2 + 6\gamma \gamma \Lambda.$$  

(16)

Now let $G_k$ be a vector whose entries are the diagonal elements of $E(\Lambda V_k V_k^T)$, and let $\mathbf{1}$ be a column vector of $1$'s which is of the same length as $G_k$. Then, using (14)-(16) we obtain the following equation describing $G_k$:

$$G_{k+1} = [I - 2\Delta E(\mu_k) + \Lambda^2 E(\mu_k^2) (2I + 11^T) \mathbf{1}^T] G_k + E(\mu_k^2) \Lambda \mathbf{1} \xi_{\text{min}} + \Lambda \sigma^2.$$  

(17)

$$E(\mu_{k+1}) = \alpha E(\mu_k) + \gamma (\xi_{\text{min}} + \gamma \mathbf{1}^T G_k).$$  

(18)

$$E(\mu_{k+1}^2) = \alpha^2 E(\mu_k^2) + 2\alpha \gamma E(\mu_k \xi_{\text{min}} + \Lambda \mathbf{1}^T G_k) + 3\gamma^2 (\xi_{\text{min}} + \gamma \mathbf{1}^T G_k) + 6\gamma^2 \mathbf{1}^T G_k.$$  

(19)

Note that since under the assumption $E(\mu_k) < 2/\lambda_{\text{max}}(R)$, $E(V_k) \rightarrow 0$, the last term in the right-hand side of (19) will asymptotically be zero. In [5], the excess mse $\xi_{\text{ex}}(k) = \xi(k) - \xi_{\text{min}}$ is shown to be given by

$$\xi_{\text{ex}}(k) = \mathbf{1}^T G_k.$$  

(20)

The mean-square error behavior is now completely described by (17)-(20).

The use of the Gaussian moment factoring theorem is made possible by assumption 2. Assumption 2 is used only in the evaluation of $E(e_k^2)$. Since $E(e_k^2)$ is multiplied by the small quantity $\gamma^2$, we need only to show that assumption 2 holds approximately. We shall show that under conditions required for the convergence of the mean-square error, assumption 2 will indeed hold asymptotically, for small $\mu_k$.

First, the conditionally Gaussian assumption can be justified in the same way as [11] for small $\mu_k$. Now as-
sume that
\[
\lim_{k \to \infty} E(\mu_k) = \bar{\mu}
\]
\[
\lim_{k \to \infty} E(\mu_k^2) = \mu^2.
\]
Then the off-diagonal elements of \(E(V_k V_k^T)\) are determined by a homogeneous equation of the form
\[
E(V_{k+1} V_{k+1}^T) = \rho_j E(V_k V_k^T)\]
where the \(\rho_j\) are given by the following equation:
\[
\rho_j = 1 - \bar{\mu}(\lambda_j + \lambda_j) + 2\mu^2\lambda_j^2
\]
But,
\[
\rho_j - \rho_j^2 = (1 - 2\bar{\mu}\lambda_j + 2\mu^2\lambda_j^2)(1 - 2\bar{\mu}\lambda_j + 2\mu^2\lambda_j^2)
\]
\[
- [1 - \bar{\mu}(\lambda_j + \lambda_j) + 2\mu^2\lambda_j^2]^2
\]
\[
= [(1 - \bar{\mu}\lambda_j)^2 + (2\mu^2 - \bar{\mu}\lambda_j)](1 - \bar{\mu}\lambda_j)^2
\]
\[
+ (2\mu^2 - \bar{\mu}\lambda_j)^2 - [(1 - \bar{\mu}\lambda_j)(1 - \bar{\mu}\lambda_j)]^2
\]
\[
+ (2\mu^2 - \bar{\mu}\lambda_j)^2\lambda_j^2
\]
\[
= (2\mu^2 - \bar{\mu}\lambda_j)(1 - \bar{\mu}\lambda_j) - \lambda_j(1 - \bar{\mu}\lambda_j))^2
\]
\[
\geq 0.
\]
In the next section, we show that one of the conditions for convergence of the mse is that \(\rho_j < 1\), \(\forall j = 1, \ldots, n\). Therefore, under this condition and using (23)
\[
\rho_j < 1. \tag{24}
\]
This means that the off-diagonal components of \(E(V_k V_k^T)\) decrease monotonically to 0, so that the components of \(V_k\) are asymptotically uncorrelated. Finally, it is readily shown that for \(i \neq j\) the conditional expectation \(E(V_k V_k^T | \mu_{i-1})\) satisfies approximately the equation
\[
E(V_{i+1} V_{i+1}^T | \mu_i)\]
\[
= [1 - \mu_i(\lambda_i + \lambda_i) + 2\mu^2\lambda_i \lambda_i] E(V_i V_i^T) \]
and so the components of \(V_k\) are also conditionally asymptotically uncorrelated.

V. STEADY STATE MISADJUSTMENT
In this section, we examine the performance of our variable step size algorithm. The figure of merit we shall use is the steady state misadjustment \(M\), which is defined to be
\[
M = \frac{\xi_{cs}}{\xi_{min}}
\]
where \(\xi_{cs}\) is the steady state value of \(\xi_{cs}(k)\). Since \(\xi_{cs}(k)\) is given by \(1^T G_k\), we shall first study the steady state behavior of \(G_k\).

Assume that the first two moments of the step size converge to a steady state value
\[
E(\mu_k) \xrightarrow{k \to \infty} \bar{\mu}
\]
\[
E(\mu_k^2) \xrightarrow{k \to \infty} \mu^2.
\]
The technique of [7] will now be used to derive conditions under which the \(G_k\) equation tends to a steady state solution. Define
\[
\rho_j = 1 - 2\bar{\mu}\lambda_j + 2\mu^2\lambda_j^2
\]
\[
= (1 - \bar{\mu}\lambda_j)^2 + \mu^2\lambda_j^2 + 2(\mu^2 - \bar{\mu}\lambda_j)^2.
\]
Also define
\[
F = I - 2\bar{\mu}\lambda_j + \mu^2\lambda_j^2(2I + 11^T).
\]
The system matrix in (17) for \(G_k\) converges to \(F\) as \(k \to \infty\). Standard results on stability [12] then show that (17) is exponentially stable if the eigenvalues of \(F\) lie strictly inside the unit circle. Now,
\[
det F - \zeta I = \det [\text{diag}(\rho_1 - \zeta, \ldots, \rho_n - \zeta)]
\]
\[
- \mu^2\Lambda^2 11^T
\]
\[
= \left[ \prod_{j=1}^{n} (\rho_j - \zeta) \right] \left[ 1 + \mu^2 1^T \text{diag} \left( \frac{1}{\rho_1 - \zeta}, \ldots, \frac{1}{\rho_n - \zeta} \right) \Lambda^2 1 \right]
\]
\[
= \left[ \prod_{j=1}^{n} (\rho_j - \zeta) \right] \left[ 1 + \mu^2 \sum_{j=1}^{n} \frac{\lambda_j^2}{\rho_j - \zeta} \right].
\]
This expression is of exactly the same form as that appearing in [7]. Using the same arguments, we conclude that necessary and sufficient conditions for the eigenvalues of \(F\) to lie strictly inside the unit circle are
\[
\rho_j < 1, \quad j = 1, 2, \ldots, n
\]
\[
1 + \mu^2 \sum_{j=1}^{n} \frac{\lambda_j^2}{\rho_j - 1} > 0. \tag{27}
\]
These are equivalent to the conditions
\[
\frac{\mu^2}{\bar{\mu}} < \frac{1}{\lambda_j}, \quad j = 1, \ldots, n \tag{27}
\]
and
\[
\sum_{j=1}^{n} \frac{\mu^2 \lambda_j}{\bar{\mu} - \mu^2 \lambda_j} < 2. \tag{28}
\]
Assuming these two conditions to hold then, the solution for \(G_k\) converges to \(G\), where \(G\) is given by the following:
\[
G = (2\bar{\mu}L - \Lambda \bar{\mu}^2 (2I + 11^T))^{-1}(\sigma_1^2 1 + \mu^2 \Lambda^2 1 \xi_{min}).
\]
(29)
Applying the matrix inversion lemma to (29), we obtain

\[ G = \frac{1}{2} (\bar{\mu}I - \mu^2 \Lambda)^{-1} \left[ I + \left( 1 - \frac{1}{4} \bar{\mu} \Lambda^2 \right)^{-1} \mu^2 \Lambda (\bar{\mu}I - \mu^2 \Lambda)^{-1} \right] \cdot [\sigma^2 I + \mu^2 \Lambda \xi_{\text{ref}}]. \]  

(30)

After a little bit of algebra, the following set of equations are obtained:

\[ \xi_{\text{ex}} = Y_{\text{ex}} = \frac{1}{2} \bar{\mu} (\bar{\mu}I - \mu^2 \Lambda)^{-1} \left[ I + \left( 1 - \frac{1}{4} \bar{\mu} \Lambda^2 \right)^{-1} \mu^2 \Lambda (\bar{\mu}I - \mu^2 \Lambda)^{-1} \right] \cdot \mu^2 \Lambda (\bar{\mu}I - \mu^2 \Lambda)^{-1} \mu^2 \Lambda. \]  

(31)

\[ \bar{\mu} = \frac{\gamma (\xi_{\text{min}} + Y_{\text{ex}})}{1 - \alpha} = \frac{\gamma (\xi_{\text{min}} + \xi_{\text{ex}})}{1 - \alpha}. \]  

(32)

\[ \mu^2 = \frac{2 \alpha \gamma \bar{\mu} (\xi_{\text{min}} + Y_{\text{ex}})^2 + 3 \gamma^2 (\xi_{\text{min}} + Y_{\text{ex}})^2 + 6 \gamma^2 \bar{\mu} Y_{\text{ex}}}{1 - \alpha^2}. \]  

(33)

For small values of misadjustment, \( 2 \bar{\mu} Y_{\text{ex}} \ll (\xi_{\text{min}} + Y_{\text{ex}})^2 \), so that

\[ \bar{\mu} \approx \frac{2 \alpha \gamma \bar{\mu} (\xi_{\text{min}} + Y_{\text{ex}}) + 3 \gamma^2 (\xi_{\text{min}} + Y_{\text{ex}})^2}{1 - \alpha^2}. \]  

(34)

The choice of \( \alpha \) and \( \gamma \) are clearly important for the convergence of \( G \). Here we give a simple sufficient condition on \( \alpha \) and \( \gamma \) to guarantee convergence of \( G_1 \) when \( \xi_{\text{ex}} \leq \xi_{\text{min}} \), which is the usual situation.

Using the results of [7], we see that the following condition is sufficient to guarantee that (27) and (28) are satisfied:

\[ \frac{\mu^2}{2 \bar{\mu}} \leq \frac{1}{3 \text{ tr} (R)}. \]  

(35)

Combining (32) and (34), we get

\[ \frac{\mu^2}{2 \bar{\mu}} = \frac{\alpha \gamma (\xi_{\text{min}} + \xi_{\text{ex}}) + 3 \gamma^2 (\xi_{\text{min}} + \xi_{\text{ex}})^2}{1 - \alpha^2} \]
\[ = \frac{\gamma (\xi_{\text{min}} + \xi_{\text{ex}})(3 - \alpha)}{2(1 - \alpha^2)} \]
\[ < \frac{3 \gamma \xi_{\text{min}}}{1 - \alpha^2}. \]  

(36)

since \( 0 < \alpha < 1 \). Thus, a sufficient condition on \( \alpha \) and \( \gamma \) for the convergence of \( G_1 \) is

\[ 0 < \frac{3 \gamma \xi_{\text{min}}}{1 - \alpha^2} \leq \frac{1}{3 \text{ tr} (R)}. \]  

(37)

Using the above equations, we can now derive the expressions for the misadjustment.

**Stationary Misadjustment:** For stationary systems (i.e., \( \sigma^2 = 0 \)), we can write \( \xi_{\text{ex}} \) as

\[ \xi_{\text{ex}} = \frac{1}{Y_{\text{ex}}} \xi_{\text{min}} Y_{\text{ex}} \]  

(38)

where

\[ Y = \frac{1}{2} (\bar{\mu}I - \mu^2 \Lambda)^{-1} \mu^2 \Lambda. \]  

Let \( y = Y_{\text{ex}} \). Then

\[ y = \frac{1}{2} \sum_{i=1}^{n} \frac{\mu^2 \lambda_i}{\mu - \mu^2 \lambda_i}. \]  

(39)

The misadjustment \( M \) can then be written as

\[ M = \frac{y}{1 - y}. \]  

(40)

Equation (40) does not give an explicit expression for the misadjustment, since \( y \) depends on \( M \) through \( \bar{\mu} \) and \( \alpha \).

We shall discuss the solution of the nonlinear equation for \( \xi_{\text{ex}} \) later in connection with the nonstationary case. However, we note that if \( \mu_k \) is fixed to be a constant, say \( 2 \mu' \), then \( y \) is given by

\[ y = \sum_{i=1}^{n} \frac{\mu' \lambda_i}{1 - 2 \mu' \lambda_i}. \]  

(34)

which is the same result as that obtained in [5]. Also, we can get rather simple expressions for \( M \), based on approximations valid for small \( M \), as we shall now show.

First, we observe that from the definition of \( Y \), \( Y \) satisfies

\[ 2 \bar{\mu} Y = \mu^2 \Lambda (2Y + 1). \]

For small values of misadjustment, the components of \( Y \) are \( \ll 1 \) so that

\[ 2 \bar{\mu} Y \approx \mu^2 \Lambda. \]  

We have therefore the following approximate expression for \( y \):

\[ 2 \bar{\mu} Y = \mu^2 \Lambda. \]  

(41)

Substituting the expression (34) for \( \bar{\mu} \), (32) for \( \mu \), and the fact that

\[ \bar{\mu} = \mu^2 \Lambda = \frac{\xi_{\text{min}}}{1 - y} \]

into (41), we find, after a little algebra, that \( y \) satisfies the following quadratic equation:

\[ y(1 - y) = \frac{(3 - \alpha) \gamma \xi_{\text{min}} \text{ tr } (R)}{2(1 - \alpha^2)}. \]

Since \( y \ll 1 \), the correct root to take is given by

\[ y = \frac{1 - 2 \left( \frac{3 - \alpha) \gamma \xi_{\text{min}} \text{ tr } (R)}{1 - \alpha^2} \right)^{1/2}}{2}. \]  

(42)

Finally, we arrive at the following approximate expression for the misadjustment \( M \), valid for small misadjustments in the stationary case:

\[ M = \frac{1 - 2 \left( \frac{3 - \alpha) \gamma \xi_{\text{min}} \text{ tr } (R)}{1 - \alpha^2} \right)^{1/2}}{1 + \left( \frac{1 - 2 \left( \frac{3 - \alpha) \gamma \xi_{\text{min}} \text{ tr } (R)}{1 - \alpha^2} \right)^{1/2}}{2}. \]  

(43)
This expression can be used to determine, for a fixed value of \( \alpha \), the value of \( \psi \) to achieve a desired level of misadjustment.

**Nonstationary Misadjustment:** We focus on the situation where (34) for \( \psi \) is valid. Substituting (32) into (34) gives

\[
\psi = \frac{(3 - \alpha)^2 (\xi_{\text{min}} + \xi_{\text{ex}})^2}{(1 - \alpha)(1 - \alpha^2)}.
\]

(44)

From (31), we have

\[
2\xi_{\text{ex}} = \sum_{i=1}^{n} \frac{\sigma_i^2}{\mu - \mu_i^2} + \sum_{i=1}^{n} \frac{\mu_i \lambda_i}{\mu - \mu_i^2} (\xi_{\text{min}} + \xi_{\text{ex}}).
\]

(45)

Equations (32), (44), and (45) can be solved by iteration to find \( \xi_{\text{ex}} \). Note that condition (28) required for the stability of the \( G_i \) equation also guarantees convergence of the iteration, starting say at \( \xi_{\text{ex}} = 0 \).

We can once again note that for constant step size \( \mu_k = \mu \) and equal eigenvalues, (45) simplifies to

\[
\xi_{\text{ex}} = \frac{n \sigma_i^2}{2\mu} + \frac{\mu \lambda \xi_{\text{min}}}{2(1 + 2\mu\lambda)}
\]

which is the same result as that in [13]. Also, if we assume that \( \mu \lambda \ll \mu \), and that \( \mu \lambda \text{tr}(R) \ll 2\mu \), (31) reduces to

\[
\xi_{\text{ex}} = \frac{n \sigma_i^2}{2\mu} + \mu \text{tr}(R) \xi_{\text{min}}.
\]

If we now let \( \mu_k \) be a constant, say \( 2\mu \), the above becomes

\[
\xi_{\text{ex}} = \frac{n \sigma_i^2}{4\mu} + \mu \text{tr}(R) \xi_{\text{min}}
\]

(46)

which is the result given in [1].

**VI. Simulation Results**

In this section, we describe simulations performed to verify the theory developed in the previous sections, and to compare experimentally the performance of the new variable step size (VSS) algorithm to that of the fixed step size (FSS) algorithm and Harris's VS (variable step) algorithm [9]. These simulations are system modeling experiments. In each case, an adaptive filter is placed in parallel with the system to be modelled. This setup is the same as that used in [1], and is shown in Fig. 1. The input \( \epsilon \) is assumed to be an independent, zero mean, Gaussian random sequence with correlation matrix \( R = I \) in all the simulations except the one presented in Fig. 8. In that simulation, \( \epsilon \) is a correlated sequence, and we shall supply the details for that case separately later. In addition, white noise with variance \( \xi_{\text{min}} \) is added to the output to prevent exact modeling. In order to simulate a nonstationary environment, the weights of the transversal filters are determined by the output of a bank of low-pass filters driven by white noise with variance \( \sigma_w^2 \). In the simulations presented here, the number of weights \( n = 4 \). Since time averages as a measure of statistical behavior is not useful in nonstationary situations, an ensemble of filters is simulated and averages are taken over the ensemble. Here, there are 100 filters in the ensemble, each with an independent input sequence.

For completeness, we summarize Harris's algorithm below. Each adaptive weight \( \mu_i(k) \) is adjusted according to the equation

\[
\mu_i(k + 1) = \mu_i(k) + \lambda_i(k)x_i(k)e(k), \quad i = 0, \ldots, n.
\]

Each of the step sizes \( \mu_i \) is updated as follows: i) If \( m_0 \) consecutive sign changes in \( x_i(\ell) \) have occurred, then \( \mu_i(k) = \mu_i(k - 1) / |\alpha| \). ii) If \( m_1 \) consecutive identical signs have occurred, \( \mu_i(k) = \alpha \mu_i(k - 1) \) where \( |\alpha| \ll 1 \). In addition, \( \mu_i \) is restricted to the range \( [\mu_{\text{min}}, \mu_{\text{max}}] \) to guarantee stability of the algorithm. It should be noted that although the FSS algorithm is the simplest to implement, neither the new VSS algorithm nor Harris's algorithm is significantly more complex to implement.

Fig. 2 shows the behavior of the FSS algorithm and the new VSS algorithm in a stationary environment. In order to show clearly the different mse characteristics in Fig. 2, the data has been smoothed using a first-order low-pass filter, and plotted on a linear scale focused on the steady state region. (The mse data to be presented in later graphs are plotted on a semilog scale without smoothing.) The parameters used in the VSS algorithm are \( \alpha = 0.97 \), \( \gamma = 4.8 \times 10^{-5} \), \( \mu_{\text{max}} = 0.1 \), and \( \mu_{\text{min}} = 10^{-5} \). The same values for \( \mu_{\text{max}} \) and \( \mu_{\text{min}} \) are used in the simulations shown in Figs. 2-7, except in connection with Fig. 3 where \( \mu_{\text{min}} = 8.25 \times 10^{-5} \). The value for \( \alpha \) appears to be a good choice for all the experiments, while the value for \( \gamma \) is chosen arbitrarily. The step sizes for the two FSS simulations have been chosen to give comparable misadjustment or convergence rate to the VSS algorithm. Notice that the misadjustment level of the FSS algorithm with a small step size is achieved, but the convergence rate of the FSS algorithm with a large step size is also achieved.
by the VSS algorithm. The VSS algorithm has reduced the tradeoff between misadjustment and convergence rate.

Fig. 3 compares the VSS algorithm with Harris's algorithm and the FSS algorithm. Notice that both the VSS and Harris's algorithm provide much faster convergence than the FSS algorithm for the same level of misadjustment. The parameters used in Harris's algorithm are $\alpha = 2$, $m_0 = 4$, $m_1 = 5$. The parameters for the other two algorithms are the same as those in Fig. 2.

Fig. 4 shows the behavior of the three algorithms in a nonstationary environment. The fixed step size $\mu$ and the $\alpha$ and $\gamma$ parameters for the VSS algorithm have been chosen to give minimum misadjustment. The parameters used for the VSS algorithm are $\alpha = 0.97$ and $\gamma = 7.65 \times 10^{-4}$. Since there is no theoretical analysis of the misadjustment in Harris's algorithm, there is no guidance available to choose the design parameters. For this simulation, they are chosen to be $\alpha = 2$, $m_0 = 3$, and $m_1 = 3$. Here $\sigma^2_f = 10^{-3}$. As may be seen, the VSS algorithm has better or at least as good a performance as the other two algorithms.

Table I shows that the new algorithm is also less sensitive to variations in the level of nonstationarity than the FSS algorithm. Optimal parameters for both the FSS and VSS algorithms are calculated for a given level of nonsta-
Fig. 6. (a) Comparison of MSE from simulation and theoretical prediction in nonstationary environment. (b) Comparison of step size from simulation and theoretical prediction in nonstationary environment.

Fig. 7. Ratio of MSE of VSS algorithm over FSS algorithm.

Table I

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$\xi_{\min}$ Fixed Step Size</th>
<th>$\xi_{\min}$ Variable Step Size</th>
<th>Predicted Misadjustment Fixed Step Size</th>
<th>Predicted Misadjustment Variable Step Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0001</td>
<td>1.0</td>
<td>7.0</td>
<td>6.8</td>
<td>7.1</td>
</tr>
<tr>
<td>0.001</td>
<td>1.0</td>
<td>14.2</td>
<td>14.4</td>
<td>13.9</td>
</tr>
<tr>
<td>0.01</td>
<td>1.0</td>
<td>88.3</td>
<td>68.9</td>
<td>81.9</td>
</tr>
<tr>
<td>0.0</td>
<td>0.5</td>
<td>4.41</td>
<td>2.65</td>
<td>3.47</td>
</tr>
<tr>
<td>0.0</td>
<td>2.0</td>
<td>4.41</td>
<td>8.80</td>
<td>3.47</td>
</tr>
</tbody>
</table>

The parameters used forclassical the FSS algorithm have $\mu = 0.029$.

However, the VSS algorithm is more sensitive to increases in the level of $\xi_{\min}$. The FSS algorithm maintains a constant level of misadjustment independent of the value of $\xi_{\min}$, while the misadjustment of the VSS algorithm increases and decreases as $\xi_{\min}$. This is of course to be expected since $\mu$ for the VSS algorithm increases with $\xi_{\min}$, resulting therefore is a larger value of the misadjustment. If desired, the parameter $\gamma$ can be decreased to reduce the level of misadjustment. These results are also shown in Table I, where the parameters used are $\alpha = 0.97$, $\gamma = 7.65 \times 10^{-4}$, and $\mu = 0.0165$.

Figs. 5(a), (b) and 6(a), (b) compare the theoretical predictions of the VSS algorithm for the mean-square error and the step size, described in the previous sections, to the results of simulations in both stationary and nonstationary environments. It can be seen that our analysis agrees well with simulation results.

One of the main features of the new algorithm is the ability to increase the step size for improved tracking when changes in the system to be modeled occur. Fig. 7 shows the ratio of the VSS mse to that of the FSS algorithm. Where changes occur in the system, we see that the ratio decreases to less than one. For example, when mse ratio is 0.8, the VSS mse is only 80% of the FSS mse. The results displayed in Fig. 7 clearly show the responsiveness of the VSS algorithm.

In Fig. 8, we show simulation results with a nonwhite input sequence in a stationary environment. The $x_i$ input sequence is generated by

$$x_i = 0.9x_{i-1} + b_i$$

where $b_i$ is a zero mean independent Gaussian sequence with $\sigma^2 = 1$. The parameters used for Harris’s algorithm are $\alpha = 2$, $m_0 = 2$, $m_1 = 3$, $\mu_{\min} = 10^{-4}$, $\mu_{\max} = 0.01$, $\mu_0 = 0.01$. The parameters used in the VSS algorithm are $\alpha = 0.97$, $\gamma = 10^{-4}$, $\mu_{\min} = 10^{-4}$, $\mu_{\max} = 0.01$, $\mu_0 = 0.01$. The plot has been smoothed for better contrast. Fig. 8 shows results similar to the previous simulations: the VSS algorithm has faster convergence and lower steady state mse than the FSS algorithm and Harris’s algorithm. The difference in convergence speed is
not as dramatic in this example because \( \mu_{\text{max}} \) is only twice the value of the fixed step size. While Harris’s algorithm performed well initially, it was difficult to find values of \( m_0 \) and \( m_1 \) that would give good steady state results.

VII. CONCLUSIONS

A new LMS type algorithm has been introduced which uses a variable step size to reduce the tradeoff between misadjustment and tracking ability of the fixed step size LMS algorithm. The variable step size algorithm also reduces sensitivity of the misadjustment to the level of nonstationarity.

A significant feature of the new algorithm is that approximate formulas can be derived to predict the misadjustment in both stationary and nonstationary environments. These theoretical predictions agree well with simulations of the algorithm.

Comparison of the new algorithm with the fixed step size algorithm and another variable step algorithm due to Harris et al. shows that it has superior performance to the fixed step size algorithm, and performs at least as well as Harris’s algorithm.

APPENDIX A

GAUSSIAN MOMENT FACTORING THEOREM

For zero mean Gaussian random variables \( x_i \), \( i = 1, \ldots, 4 \), the following result holds:

\[
E(x_1 x_2 x_3 x_4) = E(x_1 x_2)E(x_3 x_4) + E(x_1 x_3)E(x_2 x_4) + E(x_1 x_4)E(x_2 x_3).
\]

Applying the above result to a zero mean Gaussian random vector \( X \) with \( E(XX^T) = \Lambda \), \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \), we obtain the following:

\[
E(XX^TAXX^T) = A \Lambda A + A \Lambda^T A + \Delta \text{tr}(\Lambda A).
\]

Details can be found in [5]. If \( x_1 \) has mean \( \bar{x}_1 \), then on letting \( \bar{x}_i = x_i - \bar{x}_i \), we have

\[
E(x_i x_j x_k x_l) = E(\bar{x}_i \bar{x}_j)E(\bar{x}_k \bar{x}_l) + E(\bar{x}_i \bar{x}_k)E(\bar{x}_j \bar{x}_l) + E(\bar{x}_i \bar{x}_l)E(\bar{x}_j \bar{x}_k).
\]

APPENDIX B

In this Appendix, we derive (16) which gives an expression for \( E(\epsilon^2) \). We begin with the definition of \( \epsilon_i \) to obtain:

\[
E(\epsilon_i^2) = E(\epsilon_i^4) - 4E(\epsilon_i^2)X_i^2V_i + 6E(\epsilon_i^2)X_i^2V_i - 4E(X_i^4)X_i^2V_i + V_i^2X_i^2V_i^2X_i^2V_i.
\]

Since \( \epsilon_i \) is assumed to be a zero mean Gaussian independent sequence, (B.1) simplifies to

\[
E(\epsilon_i^4) = 3\epsilon_{\text{max}}^2 + 6\epsilon_{\text{min}}^2 \text{tr} [AE(V_i^2)] + E(V_i^2)X_i^2V_i^2X_i^2V_i.
\]

To simplify the notation, when dealing with the process \( V_i \), we shall use \( E_i \) to denote conditional expectation given \( \mu_i \). For example,

\[
E_i(V_i^2) = E(V_i^2|\mu_i-1).
\]

and

\[
E_i(V_{i+1}^2) = E(V_{i+1}^2|\mu_i).
\]

Let \( V_i = E(V_i) \) and \( \bar{V}_i = V_i - E_i \), and let \( \tilde{v}_i \), \( \bar{v}_i \), and \( \bar{x}_i \) denote the ith components of \( \tilde{V}_i \), \( \bar{V}_i \), and \( X_i \), respectively. Since \( V_i \) is assumed to be conditionally Gaussian given \( \mu_i \), and its components and those of \( X_i \) are assumed to be mutually independent, we can apply the Gaussian moment factoring theorem (Appendix A) to the last term of (B.2) to get

\[
E_i(V_i^2)X_i^2V_i^2X_i^2V_i = \sum_{i,j,l,m} E_i(v_j v_l v_m)E(x_i x_j x_l x_m) + \sum_{i,j,l,m} [E_i(\tilde{v}_j \tilde{v}_l \tilde{v}_m)E_i(\tilde{v}_i) + E_i(\tilde{v}_i \tilde{v}_j \tilde{v}_l)E_i(\tilde{v}_m)] + E_i(\tilde{v}_i \tilde{v}_j \tilde{v}_l \tilde{v}_m)E_i(\tilde{v}_i)E_i(\tilde{v}_j)E_i(\tilde{v}_l)E_i(\tilde{v}_m) + E_i(\tilde{v}_i \tilde{v}_j \tilde{v}_l \tilde{v}_m)E_i(\tilde{v}_i)E_i(\tilde{v}_j)E_i(\tilde{v}_l)E_i(\tilde{v}_m).
\]
Noting that
\[ [E(x_i x_j) E(x_i x_m) + E(x_i x_j) E(x_i x_m) + E(x_i x_m) E(x_i x_j)] \]
\[ = [\lambda_i \lambda_j \delta_{ij} \delta_{im} + \lambda_i \lambda_j \delta_{ij} \delta_{jm} + \lambda_i \lambda_j \delta_{im} \delta_{jm}] \]
we find, after some tedious algebra,
\[ E_x (V_i^T X_i X_i^T V_i V_i^T X_i X_i^T V_i) \]
\[ = 3 \sum_{i,j} E_x (u_i^2) E_x (u_i^2) (\lambda_i \lambda_j + 2 \lambda_i \delta_{ij}) \]
\[ - 6 \sum_i [E_x (u_i)]^2 \lambda_i^2. \]  \hspace{1cm} (B.4)

Discarding terms which contain \((1 - a)\), it is readily seen that
\[ [E_x (u_i (k + 1))]^2 = (1 - \mu_k \lambda_i)^4 [E_x (u_i (k))]^2 \]
so that
\[ \sum_i [E_x (u_i)]^2 \lambda_i^2 = \frac{E \left( \sum_i [E_x (u_i)]^2 \lambda_i^2 \right)}{\sum i [E(x_i (k) - 1))]^2 \lambda_i^2.} \]  \hspace{1cm} (B.5)

Next, we will show that
\[ E[x_i (v_i + 1)] = E[(v_i + 1)^2] E[(v_i)^2]. \]  \hspace{1cm} (B.6)

First observe that
\[ E_x (v_i (k + 1)) = (1 - 2 \lambda_i \mu_k + 2 \mu_k \lambda_i^2) E_x (v_i (k)) \]
\[ + \mu_k^2 \lambda_i \sum_j \lambda_j E_x (u_j (k)) + \sigma^2 \lambda_i^2 \lambda_{\min}. \]

Since \( \mu_k \) is small, terms involving cubic or higher powers in \( \mu_k \) are small compared to \( \mu_k^2 \) and will be discarded. Thus
\[ E[x_i (v_i (k + 1))] \]
\[ - E [v_i (k + 1)] E [v_i (k + 1)] \]
\[ = 4 [E (u_i) - E (u_i) \lambda_i \lambda_j E_x (v_i (k)) E_x (v_i (k))] \]
\[ = 4 \mu_k E_x (E_x (u_i)) [E_x (u_i)] \lambda_i \lambda_j E_x (v_i (k)) E_x (v_i (k)). \]  \hspace{1cm} (B.7)

As explained in remarks about assumption 1, Section III, \( \mu_k = E (\mu_k) \) is small when \( \gamma \) is small. Since ultimately these expressions are used in the evaluation of \( E (\varepsilon_i^2) \), which in turn is multiplied by \( \gamma^2 \), we are justified in concluding (B.6).

Finally, since
\[ E[V_i^T X_i^T V_i^T X_i^T V_i V_i^T X_i^T V_i] \]
\[ = E[E_x(V_i^T X_i^T V_i^T X_i^T V_i)] \]  \hspace{1cm} (B.8)
combining (B.2), (B.4)-(B.6), and (B.8), we obtain (16).