Convergence of the Delayed Normalized LMS Algorithm with Decreasing Step Size

Sang-Sik Ahn, *Member, IEEE*, and Peter J. Voltz, *Member, IEEE*

Abstract—**In several practical applications of the LMS algorithm, including certain VLSI implementations, the coefficient adaptation can be performed only after some fixed delay. The resulting algorithm is known as the delayed LMS (DLMS) algorithm in the literature. Previous published analyses of this algorithm are based on mean and moment convergence under the independence assumption between successive input vectors. These analyses are interesting and give valuable insights into the convergence properties but, from a practical viewpoint, they do not guarantee the correct performance of the particular realization with which the user must live. In this paper, we consider a normalized version of this algorithm with a decreasing step size $\mu(n)$ and prove the almost sure convergence of the nonhomogeneous algorithm, assuming a mixing input condition and the satisfaction of a certain law of large numbers.**

I. Introduction

Among the many adaptive algorithms, the least mean square (LMS) algorithm, because of its simplicity and ease of computation, is among the most popular in many fields of application including communications [1], [2], system identification [3], [4], adaptive control [5], and signal processing [6], [7]. In most applications, these algorithms are employed as adaptive estimators. For example, in applications such as system identification, they are used to estimate the unknown parameters of a plant within a control system. In signal analysis, they might be used to estimate the statistical parameters of a stochastic process. In all applications, it is essential to understand fully the convergence properties of the algorithm in order to correctly predict the performance and achieve a good system design. Convergence analyzes for LMS-type algorithms have been studied for a long time, and there exist proofs of almost sure convergence for the standard LMS algorithm with constant step size in the homogeneous case [8], [9] and with decreasing step size in the nonhomogeneous case [10].

In some practical situations, when implementing the adaptive algorithm, the error signal can be obtained only after some fixed delay. For example, when employing a decision-directed adaptive equalizer, if we use a decoding procedure such as the Viterbi algorithm, the decoded data, and hence, the error, is not available until after a delay of several symbols because of the inherent decoding delay [11], [12]. A similar problem is also encountered in adaptive reference echo cancellation [13]. Another example occurs in the high-speed signal processing application when the LMS algorithm is implemented using a parallel architecture, such as a pipeline structure or a systolic array. In this case as well, the appropriate form of the algorithm contains an inherent processing delay [14], [15]. In applications such as these, the algorithm realized is a modified version of the LMS algorithm known as the delayed LMS (DLMS) algorithm in the literature.

The convergence properties of the DLMS algorithm with a constant step size have been investigated by Kabal [16], Long *et al.* [17], [18], and Herzberg *et al.* [15], [19]. Kabal derived a stability bound to ensure convergence of the mean of the weight vector under the standard independence assumption, which assumes that the LMS weight vector is statistically independent of the input vector. Long *et al.* considered the special case when the input vector arises from a tapped delay line implementation and derived a necessary condition on the step size for the convergence of the excess mean square error under the independence assumption, among other approximations. In [19], Herzberg *et al.* obtained an upper bound for nonoscillatory convergence of the mean of the weight vector and of the excess mean square error, also under a strict independence assumption. These analyses are interesting and give valuable insights into the convergence properties, but from a practical viewpoint, they do not guarantee convergence of the algorithm for the particular realization of the input with which the user must live. Indeed, in practice, we would prefer probability one convergence, which guarantees convergence for almost all sample functions.

In the homogeneous case the second author has demonstrated almost sure convergence of the normalized version of the DLMS algorithm (DNLMS) for a sufficiently small gain constant under the well-known mixing input assumption [20]. Recently, this result has been extended to ergodic inputs that may not be mixing. In the nonhomogeneous case, however, there exists no rigorous sample convergence analysis for the LMS-type algorithms when they are constrained to operate with a delay in the coefficient adaptation. In this paper, we prove that if the input is mixing and $\mu(n) = a/n$, $a > 0$, then a sufficient condition for almost sure convergence of the nonhomogeneous DNLMS is that the input $X(n)$ and the reference sequence to be modeled $u(n)$ satisfy a certain law of large numbers (see Section III). The purpose of this paper, therefore, is to extend the previous analysis to the more general nonhomogeneous case and to present sample

Manuscript received December 6, 1993; revised June 21, 1996. The associate editor coordinating the review of this paper and approving it for publication was Prof. José M. F. Moura.

S.-S. Ahn is with the Department of Electrical and Information Engineering, Korea University, Jochiwon, Chungnam, Korea.

P. J. Voltz is with the Polytechnic University, Route 110, Farmingdale, NY 11735 USA e-mail: (pvoltz@poly.edu).

Publisher Item Identifier S 1053-587X(96)09047-2.

convergence of the DNLMS algorithm with a decreasing step size.

II. Formulation of the Problem

The LMS algorithm is often used as an adaptive solution to the following problem: Given a reference signal $u(n)$ and a data vector $X(n)$, the linear estimator of $u(n)$ in terms of $X(n)$ is characterized by

$$\hat{u}(n) = X(n)^T W \tag{1}$$

where the weight vector W contains the estimator coefficients. The problem is to find W, which minimizes the mean square error ξ

$$\xi = E\{[u(n) - X(n)^T W]^2\} \tag{2}$$

in which $E\{\cdot\}$ denotes the statistical expectation of the braced quantity, and the superscript T denotes matrix transposition. Since ξ is a quadratic function of W, it has a unique minimum obtained by choosing

$$W_{opt} = R^{-1} r \tag{3}$$

where

$$R = E\{X(n)X(n)^T\} \tag{4}$$

$$r = E\{u(n)X(n)\} \tag{5}$$

assuming the stationarity of $X(n)$ and $u(n)$ and the invertibility of R. In the absence of knowledge about the statistics of $u(n)$ and $X(n)$, we can solve (3) iteratively by the well-known stochastic gradient algorithm:

$$W(n+1) = W(n) + \mu\{u(n) - X(n)^T W(n)\}X(n) \tag{6}$$

where $W(n)$ is the weight vector for the nth iteration cycle, and μ is a constant gain that controls the rate of convergence and the stability of the algorithm. This is the well-known LMS algorithm. Notice here that the standard LMS algorithm minimizes the ordinary mean square error. In some applications, however, it might be more appropriate to minimize the normalized mean square error ξ' given by

$$\xi' = E\left\{\left[\frac{u(n) - X(n)^T W}{\|X(n)\|}\right]^2\right\}. \tag{7}$$

In this case, the optimum vector W'_{opt} is obtained as the solution of the following Wiener–Hopf equation:

$$E\left\{\left[\frac{u(n) - X(n)^T W'_{opt}}{\|X(n)\|}\right]\frac{X(n)}{\|X(n)\|}\right\} = 0. \tag{8}$$

Thus,

$$W'_{opt} = R'^{-1} r' \tag{9}$$

where

$$R' = E\left\{\frac{X(n)X(n)^T}{\|X(n)\|^2}\right\} \tag{10}$$

$$r' = E\left\{\frac{u(n)X(n)}{\|X(n)\|^2}\right\}. \tag{11}$$

R' and r' now involve a normalization factor, and the stationarity of $X(n)$ and $u(n)$ and the invertibility of R' are assumed. Therefore, we have a different gradient estimate $\hat{\nabla}(n)$ for ξ' given by

$$\hat{\nabla}(n) = -\frac{2\{u(n) - X(n)^T W(n)\}X(n)}{\|X(n)\|^2}. \tag{12}$$

In this case, the stochastic gradient algorithm, with a time-varying step size $\mu(n)$ should be changed to

$$W(n+1) = W(n) + \mu(n)\{u(n) - X(n)^T W(n)\} \cdot \frac{X(n)}{\|X(n)\|^2}. \tag{13}$$

This is the normalized LMS (NLMS) algorithm [6], [22]–[25] in which the LMS correction term is scaled by the squared norm of the input vector. The choice of the step size reflects a tradeoff between the amount of weight misadjustment and the speed of adaptation. It is well known in the LMS case that a small step size gives small misadjustment in the steady state but a slow convergence rate, and vice versa.

Now, as mentioned previously, in some practical situations, the error signal can not be observed until after some fixed delay. In such applications, the implemented algorithm, by analogy with (13), becomes

$$\begin{aligned} W(n+1) = W(n) &+ \mu(n-d) \\ &\cdot \{u(n-d) - X(n-d)^T W(n-d)\} \\ &\cdot \frac{X(n-d)}{\|X(n-d)\|^2}. \end{aligned} \tag{14}$$

This is a modified version of the NLMS known as the delayed NLMS (DNLMS) in the literature [20]. In the present work, we analyze this algorithm and present a set of sufficient conditions for the almost sure convergence of $W(n)$ to W'_{opt}.

Recall that W'_{opt} is defined as the weight vector that minimizes the normalized mean squared error in (7). It is clear from (7) that the mean squared error can be zero if and only if $u(n) \equiv X(n)^T W'_{opt}$, i.e., if and only if $u(n)$ can be modeled exactly by a linear combination of the input vector components, and in such cases, W'_{opt} becomes the appropriate linear weight vector. When such modeling is possible, the algorithm is said to homogeneous because the driver term in the equation governing the weight error vector ((22), which will be developed later) goes to zero, and we end up with a homogeneous difference equation. In the homogeneous case, it is possible to find W'_{opt} exactly by using the adaptive algorithm in (14) with a *constant* step size μ. To see that this well-known fact is at least plausible, note that if the update $W(n-d)$ ever takes the value W'_{opt}, the correction term becomes zero due to perfect modeling, and the adaptation stops. If, on the other hand, $u(n)$ cannot be perfectly modeled by such a linear parameterization, then the error term $u(n) - X(n)^T W(n)$ can never be *identically* zero even if $W(n) \equiv W'_{opt}$. In this case, with a fixed step size, it is easy to see from (14) that the adaptation never stops, and $W(n)$ cannot converge to *anything*, let alone to W'_{opt}. As is well known, the best that can be hoped for in this case is that with small enough μ, the weight vector update $W(n)$ remains close to W'_{opt}, and the steady state error variance is not too large. Since it is known that this weight error variance shrinks as μ is decreased, it is possible to identify W'_{opt} even in the nonhomogeneous case by successively decreasing a time varying gain parameter $\mu(n)$, provided the gain is not reduced too fast. If the gain is reduced too quickly, the algorithm can stall, as the example to be presented later will show.

There have been a few studies of LMS algorithms with update delay. The convergence properties of the DLMS algorithm (nonnormalized) with a *constant* step size μ have been investigated by Kabal [16], Long *et al.* [17], [18], and Herzberg *et al.* [15], [19]. Kabal derived a stability bound to ensure convergence of the mean of the weight vector under the standard independence assumption, which assumes that the LMS weight vector is statistically independent of the input vector. A stability region for μ was obtained that, when interpreted in the context of the normalized algorithm, becomes

$$0 < \mu < \frac{2}{\lambda_{\max}} \sin \frac{\pi}{2(2d+1)}$$

where $\lambda_{\max}$ is the largest eigenvalue of the autocorrelation matrix R', which is given in (10). It is important to realize that although this bound does guarantee stability of the mean of the weight vector, it does not guarantee stability of the weight vector itself. A simple example suffices to demonstrate this. Suppose that the input $X(n)$ to the NLMS algorithm is a sequence of independent vectors, each of which is composed of N independent Gaussian random variables with identical variance σ^2. This input certainly satisfies the independence assumption, and it can be shown that $R' = (1/N)I$ in this case. Note that with $d = 0$ (which results in the normalized LMS algorithm without delay), the bound above becomes $0 < \mu < 2/\lambda_{\max}$, or, for this example, $0 < \mu < 2N$. Thus, for this example, the *mean* of the weight vector will certainly remain stable for $0 < \mu < 2N$, but it is well known and can easily be verified by simulation that for the standard NLMS algorithm, the weight vector itself will become unstable if $\mu > 2$. This example illustrates the weakness of a mean convergence analysis and reinforces the importance of an investigation of almost sure convergence.

Long *et al.* considered the special case when $X(n)$ arises from a tapped delay line implementation and developed a convergence region that, when applied to DNLMS, is

$$0 < \mu < \frac{2}{1 + 2\left(\frac{d}{N}\right)}.$$

Notice that this result reduces to the correct one when $d = 0$, and it yields the intuitively satisfying conclusion that if d is small compared with the dimension of the input vector, then the delay should have minimal impact on the stability problem. From a practical point of view, this provides a useful stability bound under the conditions imposed by the analysis in [17] and [18]. That analysis, while more encompassing than a mean convergence analysis and very useful in practice, still relies on independence assumptions, among others, and does not address the issue of sample convergence.

In [15], Herzberg *et al.* proposed a systolic architecture for implementing the LMS algorithm, and they showed that the form of the algorithm that naturally arises from the systolic architecture is the DLMS algorithm with a delay equal to one less than the dimension of $X(n)$. In [19], the same authors obtained an upper bound for nonoscillatory convergence of the mean of the weight vector and of the excess mean square error, also under a strict independence assumption.

The main contribution of this paper is to provide a rigorous convergence analysis for the decreasing step size case that requires no independence assumption and guarantees almost sure convergence of the weight error vector rather than focusing on its mean value only. The results show that for any delay d, the algorithm will converge almost surely under the stated conditions.

III. Assumptions and Convergence Proof

In this section, we summarize the assumptions and explain the convergence proof, referring to the Appendix where appropriate. We first present and discuss the mixing condition [21] and the law of large numbers, which are assumed in the proof.

Our first assumption is that the input vector sequence is mixing in the sense that there exist a finite integer T and an $\alpha > 0$ such that for any constant nonzero N vector h, the following holds for all n:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left\{ \frac{h^T X(n+i)}{\|X(n+i)\|} \right\}^2 \geq \alpha \|h\|^2. \tag{15}$$

Basically, this condition is that over any time interval of length T, the components of $X(n)$ have an average length of at least α in any direction. To examine this property further, let us rewrite (15) as

$$h^T \left\{ \frac{1}{T} \sum_{i=0}^{T-1} \frac{X(n+i)X(n+i)^T}{\|X(n+i)\|^2} \right\} h \geq \alpha \|h\|^2 \tag{16}$$

and note that $X(n+i)X(n+i)^T / \|X(n+i)\|^2$ is a projection matrix. Therefore, if a sequence of N vectors is restricted to any proper subspace of R^N, then it cannot be mixing since there exists an N vector h that is orthogonal to the subspace. It is also clear that T must be greater than or equal to dimension N. An equivalent rephrasing of the mixing condition is

$$\lambda_{\min} \left\{ \frac{1}{T} \sum_{i=0}^{T-1} \frac{X(n+i)X(n+i)^T}{\|X(n+i)\|^T} \right\} \geq \alpha. \tag{17}$$

Our second assumption is that the sequences

$$\left\{ \frac{X(n)X(n)^T}{\|X(n)\|^2} \right\}$$

and

$$\left\{ \frac{u(n)X(n)}{\|X(n)\|^2} \right\}$$

satisfy the law of large numbers in the sense that

A1)
$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{X(i)X(i)^T}{\|X(i)\|^2}=E\left\{\frac{X(i)X(i)^T}{\|X(i)\|^2}\right\}=R' \quad \text{a.s.}$$

A2)
$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{u(i)X(i)}{\|X(i)\|^2}=E\left\{\frac{u(i)X(i)}{\|X(i)\|^2}\right\}=r' \quad \text{a.s.}$$

Notice, under A1) and A2), that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{u(i)X(i)-X(i)X(i)^TW'_{opt}}{\|X(i)\|^2}=r'-R'W'_{opt}=0 \quad \text{a.s.} \tag{18}$$

since $W'_{opt}=R'^{-1}r'$. We will use this property to prove the convergence of the nonhomogeneous DNLMS algorithm. A1) and A2) are analogous to the assumptions

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X(i)X(i)^T=E\{X(i)X(i)^T\}=R \quad \text{a.s.} \tag{19}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}u(i)X(i)=E\{u(i)X(i)\}=r \quad \text{a.s.} \tag{20}$$

which are sometimes made for the standard LMS algorithm. In this paper, we prove that if the input is mixing, then A1) and A2) are sufficient conditions for the almost sure convergence of the NLMS and DNLMS algorithms with $\mu(n)=a/n$.

Although the proof is given for $\mu(n)=a/n$, the following looser conditions are all that are required on the $\mu(n)$ sequence in much of the development, and they are called out where appropriate.

B1)
$$\sum_{n=1}^{\infty}\mu(n)=\infty$$

B2)
$$\frac{\mu(n)^2}{\mu(n+d+T)}\to 0, \qquad n\to\infty.$$

Now, the convergence properties of the DNLMS algorithm are best discussed in terms of the weight error vector $Y(n)$, which is defined as

$$Y(n)=W(n)-W'_{opt}. \tag{21}$$

Subtracting W'_{opt} from both sides of (14) and using (21) gives

$$\begin{aligned}Y(n+1)=&Y(n)-\mu(n-d)\frac{X(n-d)X(n-d)^T}{\|X(n-d)\|^2}\\&\cdot Y(n-d)+\mu(n-d)\\&\cdot\{u(n-d)-X(n-d)^TW'_{opt}\}\\&\cdot\frac{X(n-d)}{\|X(n-d)\|^2}.\end{aligned} \tag{22}$$

Equation (22) is a $(d+1)$th-order vector stochastic recursion describing the trajectory of the weight error vector $Y(n)$ in terms of the reference scalar sequence $u(n)$ and the input vector sequence $X(n)$. For analysis, by back substitution, we decompose the DNLMS algorithm into a term corresponding to the NLMS algorithm and residual terms that depend on delayed weight error vectors and input vectors. For notational simplicity, let us define

$$P(n)=\frac{X(n)X(n)^T}{\|X(n)\|^2} \tag{23}$$

and

$$Z(n)=\frac{u(n)X(n)-X(n)X(n)^TW'_{opt}}{\|X(n)\|^2}. \tag{24}$$

Then, (22) can be written in the simpler form

$$\begin{aligned}Y(n+1)=&Y(n)-\mu(n-d)P(n-d)Y(n-d)\\&+\mu(n-d)Z(n-d)\end{aligned} \tag{25}$$

or

$$\begin{aligned}Y(n+1)-Y(n)=&-\mu(n-d)P(n-d)Y(n-d)\\&+\mu(n-d)Z(n-d).\end{aligned} \tag{26}$$

Now, adding up d terms of the difference in (26) and solving for $Y(n-d)$ yields

$$\begin{aligned}Y(n-d)=&Y(n)+\sum_{i=0}^{d-1}\mu(n-2d+i)P(n-2d+i)\\&\cdot Y(n-2d+i)-\sum_{i=0}^{d-1}\mu(n-2d+i)\\&\cdot Z(n-2d+i)\end{aligned} \tag{27}$$

and substituting (27) into (22), we have

$$\begin{aligned}Y(n+1)=&[I-\mu(n-d)P(n-d)]\\&\cdot Y(n)-\mu(n-d)P(n-d)\\&\cdot\sum_{i=0}^{d-1}\mu(n-2d+i)P(n-2d+i)Y(n-2d+i)\\&+\mu(n-d)P(n-d)\sum_{i=0}^{d-1}\mu(n-2d+i)\\&\cdot Z(n-2d+i)+\mu(n-d)Z(n-d)\end{aligned} \tag{28}$$

where I denotes the identity matrix of dimension N. Again, for simplicity, let us define $Q(n)$, $\Gamma(n)$, and $L(n)$ as follows:

$$Q(n)=I-\mu(n-d)P(n-d), \tag{29}$$

$$\begin{aligned}\Gamma(n)=&-\mu(n-d)P(n-d)\sum_{i=0}^{d-1}\mu(n-2d+i)\\&\cdot P(n-2d+i)Y(n-2d+i)\end{aligned} \tag{30}$$

and

$$\begin{aligned}L(n)=&\Gamma(n)+\mu(n-d)P(n-d)\sum_{i=0}^{d-1}\mu(n-2d+i)\\&\cdot Z(n-2d+i)+\mu(n-d)Z(n-d).\end{aligned} \tag{31}$$

Then, we have the following form of the DNLMS algorithm:

$$Y(n+1)=Q(n)Y(n)+L(n) \tag{32}$$

whose solution is given by

$$Y(n+1) = \Psi(n+1, 1)Y(1) + \sum_{j=1}^{n} \Psi(n+1, j+1)L(j) \tag{33}$$

$$\begin{aligned} &= \Psi(n+1, 1)Y(1) + \sum_{j=1}^{n} \Psi(n+1, j+1)\Gamma(j) \\ &\quad + \sum_{j=1}^{n} \mu(j-d)\Psi(n+1, j+1)P(j-d) \\ &\quad \cdot \sum_{i=0}^{d-1} \mu(j-2d+i)Z(j-2d+i) \\ &\quad + \sum_{j=1}^{n} \mu(j-d)\Psi(n+1, j+1)Z(j-d) \end{aligned} \tag{34}$$

where the transition matrix is defined as

$$\Psi(n, k) = Q(n-1)Q(n-2) \cdots Q(k+1)Q(k); \quad n > k \tag{35}$$

and $\Psi(n, n) = I$. Now, taking the norm of (34) with $\mu(n) = a/n$, $a > 0$, and using (24), we have

$$\begin{aligned} \|Y(n+1)\| &= \left\| \Psi(n+1,1)Y(1) + \sum_{j=1}^{n} \Psi(n+1,j+1)\Gamma(j) \right\| \\ &\quad + a \left\| \sum_{j=1}^{n} \frac{\Psi(n+1, j+1)}{j-d} Z(j-d) \right\| \\ &\quad + a^2 \left\| \sum_{j=1}^{n} \frac{\Psi(n+1, j+1)}{j-d} \cdot P(j-d) \sum_{i=0}^{d-1} \frac{Z(j-2d+i)}{j-2d+i} \right\|. \end{aligned} \tag{36}$$

Now, if we can show that each term on the RHS of (36) converges to zero as n tends to infinity, then our proof will be done. To begin, note that the first term in (36) represents the homogeneous DNLMS algorithm defined by $Z(n) = 0$ in (25). The convergence properties of that algorithm are derived in the Appendix, and we note here that the first term in (36) is the identical to (A-4), and as shown in the Appendix:

$$\lim_{n\to\infty} \left\| \Psi(n+1, 1)Y(1) + \sum_{j=1}^{n} \Psi(n+1, j+1)\Gamma(j) \right\| = 0. \tag{37}$$

Furthermore, the second term on the RHS of (36) is simpler than the third and can be shown to converge in the same manner. Therefore, we only need to show the convergence of the third term in (36) to zero as n tends to infinity to finish the proof. We start with the following inequalities:

$$\begin{aligned} &\left\| \sum_{j=1}^{n} \frac{\Psi(n+1, j+1)}{j-d} P(j-d) \sum_{i=0}^{d-1} \frac{Z(j-2d+i)}{j-2d+i} \right\| \\ &\le \sum_{i=0}^{d-1} \left\| \sum_{j=1}^{n} \frac{\Psi(n+1, j+1)}{j-d} P(j-d) \cdot \frac{Z(j-2d+i)}{j-2d+i} \right\| \\ &\le \sum_{i=0}^{d-1} \left\| \sum_{j=1}^{n_0-1} \frac{\Psi(n+1, j+1)}{j-d} P(j-d) \cdot \frac{Z(j-2d+i)}{j-2d+i} \right\| + \sum_{i=0}^{d-1} \left\| \sum_{j=n_0}^{n} \frac{\Psi(n+1, j+1)}{j-d} \cdot P(j-d) \frac{Z(j-2d+i)}{j-2d+i} \right\|. \end{aligned} \tag{38}$$

In addition, note that from (A-12) in the Appendix

$$\lim_{n\to\infty} \|\Psi(n, n_0)\| = 0 \tag{39}$$

for *any* fixed n_0 and that the first term in (38) contains a finite number of terms, each of which converges to zero by (39). Consequently, to complete the proof that $Y(n) \to 0$, i.e., $W(n) \to W'_{opt}$ as $n \to \infty$, it only remains to be shown that the second term in (38) can be upper bounded by an arbitrarily small number by choosing n_0 sufficiently large. To this end, for any $i = 1, 2, \cdots, d-1$, we introduce $T(n)$ in (40), which appears at the bottom of the page. Then, from (18)

$$\lim_{n\to\infty} \frac{\|T(n)\|}{n} = 0 \quad \text{a.s.}$$

Hence, for any $\delta > 0$, there exists $N'_1(\delta, \omega)$ such that for $n > N'_1$

$$\frac{\|T(n)\|}{n} < \delta \quad \text{a.s.} \tag{41}$$

Now, choose $n_0 > \max(N'_0, N'_1)$, where N'_0, as defined in (A-12), is the integer lower bound on n_0 required for the exponential bound on $\|\Psi(n, n_0)\|$. To proceed further, for

$$\begin{aligned} T(n) &= \sum_{k=d+1}^{n} \frac{u(k-d+i)X(k-d+i) - X(k-d+i)X(k-d+i)^T W'_{opt}}{\|X(k-d+i)\|^2} \\ &= \sum_{k=d+1}^{n} Z(k-d+i). \end{aligned} \tag{40}$$

notational simplicity, let us define

$$\text{LHS} \equiv \left\| \sum_{j=n_0}^{n} \frac{\Psi(n+1, j+1)}{j-d} \frac{P(j-d)}{j-2d+i} Z(j-2d+i) \right\|. \tag{42}$$

Then, using (40), this can be expressed as

$$\text{LHS} = \left\| \sum_{j=n_0}^{n} \frac{\Psi(n+1, j+1)}{j-d} \frac{P(j-d)}{j-2d+i} \cdot \{T(j-d) - T(j-d-1)\} \right\|. \tag{43}$$

At this point, applying Abel's formula for partial summation of a sequence to (43) yields

$$\begin{aligned}\text{LHS} = \Bigg\| &\sum_{j=n_0-1}^{n-1} \left\{ \frac{\Psi(n+1, j+1)}{j-d} \frac{P(j-d)}{j-2d+i} - \frac{\Psi(n+1, j+2)}{j+1-d} \frac{P(j+1-d)}{j+1-2d+i} \right\} T(j-d) \\ &+ \frac{\Psi(n+1, n+1)}{n-d} \frac{P(n-d)}{n-2d+i} T(n-d) \\ &- \frac{\Psi(n+1, n_0)}{n_0-1-d} \frac{P(n_0-1-d)}{n_0-1-2d+i} \cdot T(n_0-1-d) \Bigg\|.\end{aligned}$$

Now, recalling

$$\begin{aligned}\Psi(n+1, j+1) &= \Psi(n+1, j+2)\Psi(j+2, j+1), \\ \Psi(j+2, j+1) &= Q(j+1) \\ &= I - \frac{a}{j+1-d} \cdot \frac{X(j+1-d)X(j+1-d)^T}{\|X(j+1-d)\|^2}\end{aligned}$$

and using (41), we have

$$\begin{aligned}\text{LHS} \le &\Bigg\| \sum_{j=n_0-1}^{n-1} \left\{ \frac{1}{j-d} \left[I - \frac{a}{j+1-d} \cdot \frac{X(j+1-d)X(j+1-d)^T}{\|X(j+1-d)\|^2} \right] \frac{P(j-d)}{j-2d+i} - \frac{1}{j+1-d} \frac{P(j+1-d)}{j+1-2d+i} \right\} \\ &\cdot \Psi(n+1, j+2)T(j-d) \Bigg\| + \delta + \delta \\ \le &\Bigg\| \sum_{j=n_0-1}^{n-1} \left\{ \frac{1}{j-d} \frac{P(j-d)}{j-2d+i} - \frac{1}{j+1-d} \frac{P(j+1-d)}{j+1-2d+i} \right\} \cdot \Psi(n+1, j+2)T(j-d) \Bigg\| \\ &+ a \Bigg\| \sum_{j=n_0-1}^{n-1} \frac{1}{(j-d)(j+1-d)} \cdot \frac{X(j+1-d)X(j+1-d)^T}{\|X(j+1-d)\|^2} \frac{P(j-d)}{j-2d+i} \cdot \Psi(n+1, j+2)T(j-d) \Bigg\| + 2\delta \\ \le &\sum_{j=n_0-1}^{n-1} \left\| \frac{1}{j-d} \frac{P(j-d)}{j-2d+i} - \frac{1}{j+1-d} \frac{P(j+1-d)}{j+1-2d+i} \right\| \cdot \|\Psi(n+1, j+2)\| \, \|T(j-d)\| \\ &+ a \sum_{j=n_0-1}^{n-1} \frac{\|\Psi(n+1, j+2)\|}{j+1-d} \frac{\|T(j-d)\|}{j-d} + 2\delta \\ \le &\, 2 \sum_{j=n_0-1}^{n-1} \frac{\|\Psi(n+1, j+2)\|}{j-2d+i} \frac{\|T(j-d)\|}{j-d} + \left[a \sum_{j=n_0-1}^{n-1} \frac{\|\Psi(n+1, j+2)\|}{j+1-d} + 2 \right] \delta \\ \le &\left\{ (2+a) \sum_{j=n_0-1}^{n-1} \frac{\|\Psi(n+1, j+2)\|}{j-2d} + 2 \right\} \delta. \\ &[\text{by (41)}].\end{aligned} \tag{44}$$

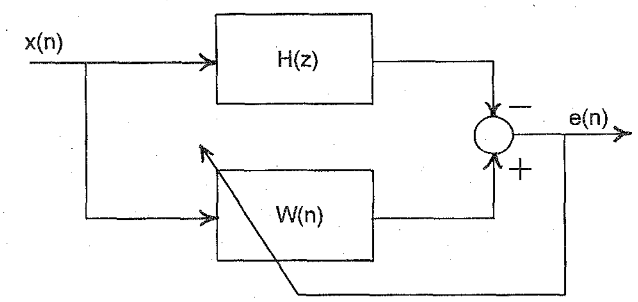
Finally, the proof reduces to showing that the sum

$$\sum_{j=n_0-1}^{n-1} \frac{\|\Psi(n+1, j+2)\|}{j-d} \tag{45}$$

can be upper bounded by a constant independent of n and n_o. This can indeed be done by appealing to the exponentially decaying character of Ψ, but the details are tedious, and the interested reader is referred to [27].

IV. A SIMULATION EXAMPLE

Consider the simplified system identification/modeling scenario depicted in Fig. 1. The IIR system function of the system under scrutiny is $H(z) = 1/(1 - az^{-1})$, and we attempt to model it as an FIR system with $N = 5$ impulse response coefficients. Both the IIR system and the FIR model are excited by white Gaussian noise as shown, and the output error is used to update the model. If $a = 0.6$, the impulse response of the IIR system will contain significant energy beyond the $n = 5$ delay, and the five-tap FIR model cannot exactly reproduce the system. Nevertheless, we apply the DNLMS algorithm (with $d = 5$) with decreasing step size in an attempt to model the system as best we can in the minimum mean squared error sense given in (7). It can easily be shown that the optimum FIR approximation in this case will be $W_{opt} =$



$$W(n+1) = W(n) + \mu(n-d)\{u(n-d) - X(n-d)^T W(n-d)\} \frac{X(n-d)}{\|X(n-d)\|^2}$$

$$X(n) = [x(n), x(n-1), \ldots\ldots, x(n-N+1)]^T$$

Fig. 1. System identification setup.

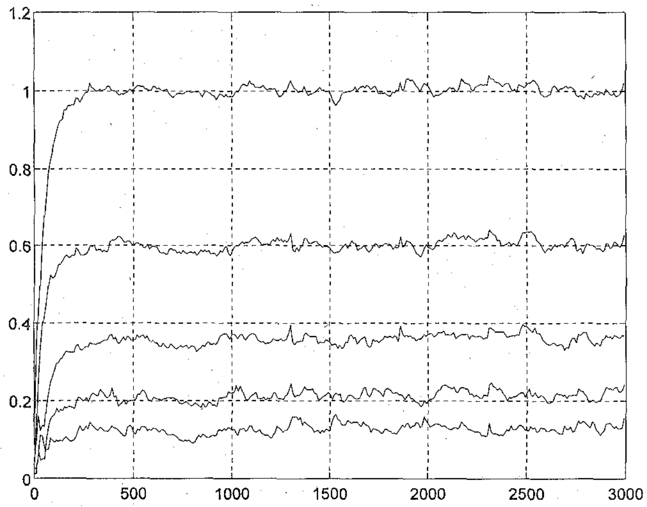


Fig. 2. Convergence of weight vector with $\mu = 0.1$.

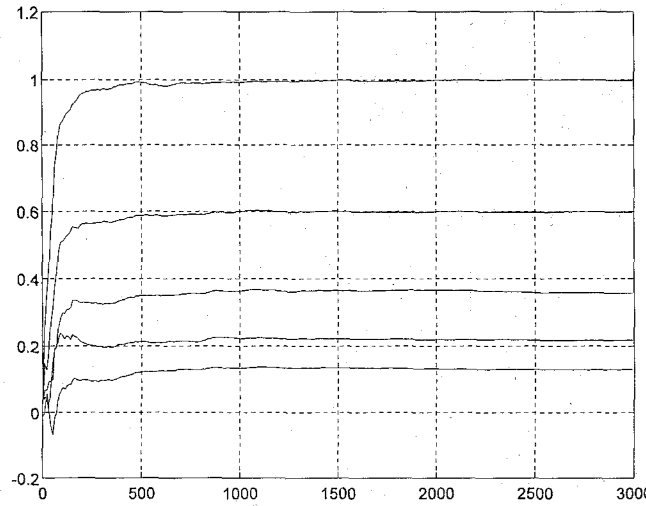


Fig. 3. Convergence of weight vector with $\mu \sim 1/n$.

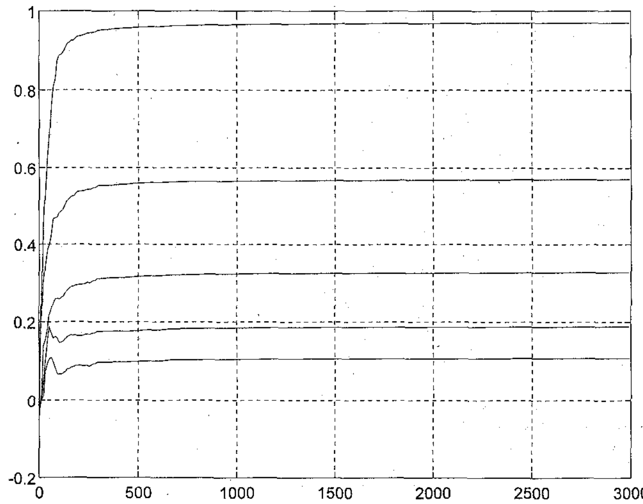


Fig. 4. Convergence of weight vector with $\mu \sim 1/n^2$.

[1, 0.6, 0.36, 0.216, 0.1296]. Fig. 2 shows the convergence of the five weight vector components when the gain μ is constant at $\mu = 0.1$. While the initial convergence seems reasonable, the steady-state error never decreases to zero because the imperfect modeling gives rise to a nonhomogeneous algorithm, as discussed earlier.

Fig. 3 shows the convergence of the five weight vector components when μ decreases as $1/n$ with a starting value of 0.1. Here, the initial convergence is similar to Fig. 2, but the weight vector ultimately converges correctly to W_{opt}. Finally, Fig. 4 illustrates the situation when the gain is decreased as $1/n^2$. In this case, the decrease in the gain sequence occurs too rapidly, μ_n does not satisfy assumption B1), and the adaptation stalls before the correct weight vector is attained.

V. Conclusion

The main result of this paper has been to provide a rigorous proof of sample convergence for the nonhomogeneous normalized DLMS algorithm with decreasing step size. The results show that if the input is mixing and satisfies a law of large numbers (see A1) and A2)), then the weight vector converges to W_{opt} for *any* d when $\mu(n) = 1/n$. The results can be viewed as extending previous work in several ways. On one hand, previous work on the delayed LMS algorithm is given more rigor and extended to the normalized version of the algorithm. On the other hand, previous work on the LMS type algorithms with decreasing step size is extended to the DNLMS algorithm. Incidentally, the example given in Section II clearly shows the shortcomings of any analysis that focuses solely on mean convergence by demonstrating that such analyses can lead to very misleading stability results when the actual convergence of $W(n)$ to a fixed value is desired.

Appendix
Homogeneous DNLMS Algorithm Analysis

If

$$u(n) = X(n)^T W'_{opt} \tag{A-1}$$

so that $Z(n) = 0$ in (25), then (22) reduces to the following homogeneous algorithm:

$$\begin{aligned} Y(n+1) = &Y(n) - \mu(n-d) \frac{X(n-d)X(n-d)^T}{\|X(n-d)\|^2} \\ &\cdot Y(n-d). \end{aligned} \tag{A-2}$$

Now, following the derivation from (23) to (31) in Section III with $Z = 0$ yields the simple form of the homogeneous DNLMS algorithm:

$$Y(n+1) = Q(n)Y(n) + \Gamma(n). \tag{A-3}$$

where $Q(n)$ and $\Gamma(n)$ are defined in (29) and (30), respectively. The solution of the above recursive equation from the beginning is given by

$$Y(n+1) = \Psi(n+1, 1)Y(1) + \sum_{i=1}^{n} \Psi(n+1, i+1)\Gamma(i) \tag{A-4}$$

or, iterating (A-3) T times from any n yields

$$Y(n+T) = \Psi(n+T, n)Y(n) + \sum_{i=0}^{T-1} \cdot \Psi(n+T, n+i+1)\Gamma(n+i) \tag{A-5}$$

where Ψ is the transition matrix defined in (35). Notice, from (A-5) and (30) that we need $Y(n-2d)$, $Y(n-2d+1), \cdots, Y(n+T-1)$ to calculate $Y(n+T)$. Now, taking the norm of (A-5) gives

$$\begin{aligned} \|Y(n+T)\| &\le \|\Psi(n+T, n)\| \, \|Y(n)\| + \sum_{i=0}^{T-1} \cdot \|\Psi(n+T, n+i+1)\| \, \|\Gamma(n+i)\| \\ &\le \|\Psi(n+T, n)\| \, \|Y(n)\| + \sum_{i=0}^{T-1} \|\Gamma(n+i)\|, \\ &\quad [\text{since } \|\Psi(\cdot, \cdot)\| \le 1]. \end{aligned} \tag{A-6}$$

Now, to derive the bound on $\|Y(n+T, n)\|$, we bound $\Gamma(n)$ first. Suppose that

$$\|Y(i)\| \le M, \qquad n-2d \le i \le n+T-1 \tag{A-7}$$

for some n; then, from (30)

$$\|\Gamma(n)\| = \left\| -\mu(n-d)P(n-d) \sum_{i=0}^{d-1} \mu(n-2d+i) \cdot P(n-2d+i)Y(n-2d+i) \right\| \tag{A-8}$$

$$\begin{aligned} &\le \mu(n-2d)^2 \sum_{i=0}^{d-1} \|Y(n-2d+i)\| \\ &\quad [\text{since } \|P(\cdot)\| = 1] \end{aligned} \tag{A-9}$$

$$\le \mu(n-2d)^2 \, dM. \tag{A-10}$$

At this point, to obtain a bound on $\|\Psi(n+T, n\|$, we note that $\Psi(\,)$ is the transition matrix of a decreasing gain homogeneous normalized LMS algorithm with no update delay, as the reader can easily recognize. In [21], Weiss and Mitra derived and exponential bound on this transition matrix for a fixed gain parameter and mixing inputs, and we note here that a similar technique can be used to generalize their result to the case of a decreasing gain parameter. For details, we refer the interested reader to [27]. The result of this derivation is that there exists N_0' such that

$$\|\Psi(n+T, n)\| \le 1 - \alpha\beta T\mu(n-d+T), \quad n_0 > N_0' \tag{A-11}$$

$$\|\Psi(n, n_0)\| \le \exp\left[-\alpha\beta T \sum_{i=1}^{m} \mu(n_0 - d + iT) \right], \quad n_0 > N_0' \tag{A-12}$$

where $m = [(n-n_0)/T]$ and $[x]$ denotes the greatest integer function of x. Finally, substituting (A-10) and (A-11) into (A-6), we will have

$$\|Y(n+T)\| \le \delta M$$

where

$$\delta(n) = 1 - \alpha\beta T\mu(n-d+T) + dT\mu(n-2d)^2. \tag{A-13}$$

Moreover, to see the behavior of $\delta(n)$ as $n \to \infty$, we rewrite (A-13) as

$$\delta(n) = 1 - T\mu(n-d+T)\left\{ \alpha\beta - \frac{dT\mu(n-2d)^2}{\mu(n-d+T)} \right\}. \tag{A-14}$$

Then, by (B2), there exists $\eta > 0$ and $N_2 > N_0'$ such that for $n > N_2$

$$\alpha\beta - \frac{dT\mu(n-2d)^2}{\mu(n-d+T)} \ge \eta > 0. \tag{A-15}$$

Thus,

$$\delta(n) \le 1 - \eta T\mu(n-d+T) \tag{A-16}$$

so that

$$\|Y(n+T)\| \le M \exp\left[-\eta T\mu(n-d+T)\right]. \tag{A-17}$$

Therefore, condition (A-7) can be shifted up one unit in time to yield

$$\|Y(i)\| \le M, \qquad n-2d+1 \le i \le n+T. \tag{A-18}$$

With condition (A-18), we may repeat the whole development from (A-8)–(A-17) to get

$$\|Y(n+T+1)\| \le M \exp\left[-\eta T\mu(n-d+T+1)\right]. \tag{A-19}$$

Furthermore, repeating from (A-8)–(A-17) $(T+2d-2)$ more times with the appropriately shifted conditions, we have for $0 \le i \le T+2d-1$

$$\begin{aligned} &\|Y(n+T+i)\| \\ &\quad \le M \exp\left[-\eta T\mu(n-d+T+T+2d-1)\right] \\ &\quad \le M \exp\left[-\eta T\mu(n-d+T+T+2d)\right] \end{aligned} \tag{A-20}$$

since $\mu(n)$ is decreasing. For the next interval of length $(T+2d)$, using (A-20) as the starting bound and repeating the steps from (A-8)–(A-20) yields

$$\begin{aligned} &\|Y(n+T+T+2d+i)\| \\ &\quad \le M \exp\left(-\eta T\{\mu(n-d+T+T+2d) \right. \\ &\quad\quad \left. + \mu[n-d+T+2(T+2d)]\}\right) \\ &\quad\quad 0 \le i \le T+2d-1. \end{aligned}$$

Therefore, it follows that $\|Y(n)\|$ of the jth interval of length $(T+2d)$ from N_2 is bounded by

$$\begin{aligned} &\|Y[n+T+i+(j-1)(T+2d)]\| \\ &\quad \le M \exp\left\{-\eta T \sum_{k=1}^{j} \mu[n-d+T+k(T+2d)]\right\} \\ &\quad 0 \le i \le T+2d-1. \end{aligned} \tag{A-21}$$

Finally, if we let $j \to \infty$ with (A-21), we obtain the required convergence. That is, for $0 \le i \le T+2d-1$

$$\begin{aligned} &\lim_{j\to\infty} \|Y[n+T+i+(j-1)(T+2d)]\| \\ &\quad \le M \exp\left\{-\eta T \sum_{k=1}^{\infty} \mu[n-d+T+k(T+2d)]\right\}. \end{aligned}$$

Therefore, by (B1)

$$\lim_{n\to\infty} Y(n) = 0 \quad \Rightarrow \quad \lim_{n\to\infty} W(n) = W'_{opt}. \qquad \text{Q.E.D}$$

REFERENCES

[1] A. Gersho, "Adaptive equalization of highly dispersive channels for data transmission," *Bell Syst. Tech. J.*, vol. 48, pp. 55–70, 1969.

[2] E. Mazo, "Analysis of decision directed equalizer convergence," *Bell Syst. Tech. J.*, vol. 59, pp. 1857–1876, 1980.

[3] G. Goodwin and K. Sin, *Adaptive Filtering, Prediction, and Control.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[4] I. D. Landau, "Unbiased recursive identification using model reference adaptive techniques," *IEEE Trans. Automat. Contr.*, vol. AC-21, no. 2, pp. 194–202, Apr. 1976.

[5] S. Sastry and M. Bodson, *Adaptive Control—Stability, Convergence, and Robustness.* Englewood Cliffs, NJ: Prentice-Hall, 1989.

[6] S. Haykin, *Adaptive Filter Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1991.

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

[8] V. Solo, "The limiting behavior of LMS," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1909–1922, Dec. 1989.

[9] O. Macchi and E. Eweda, "Second-order convergence analysis of stochastic adaptive linear filtering," *IEEE Trans. Automat. Contr.*, vol. AC-28, no. 1, pp. 76–85, Jan. 1983.

[10] E. Eweda and O. Macchi, "Convergence of an adaptive linear estimation algorithm," *IEEE Trans. Automat. Contr.*, vol. AC-29, no. 2, pp. 119–127, Feb. 1984.

[11] T. N. Morrissey, Jr., "Analysis of decoders for convolutional codes by stochastic sequential machine methods," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 460–469, July 1970.

[12] G. D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 363–378, July 1970.

[13] D. D. Falconer, "Adaptive reference echo cancellation," *IEEE Trans. Commun.*, vol. COM-30, pp. 2083–2094, Sept. 1982.

[14] K. Konstantinides and N. Kanopoulos, "Design alternatives for adaptive digital lattice filters and a new bit-serial architecture," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 737–742, July 1987.

[15] H. Herzberg, R. Haimi-Cohen, and Y. Be'ery, "A systolic array realization of an LMS adaptive filter and the effects of delayed adaptation," *IEEE Trans. Signal Processing*, vol. 40, no. 11, pp. 2799–2803, Nov. 1992.

[16] P. Kabal, "The stability of adaptive minimum mean square error equalizers using delayed adjustment," *IEEE Trans. Commun.*, vol. COM-31, no. 3, pp. 430–432, Mar. 1983.

[17] G. Long, F. Ling, and J. Proakis, "The LMS algorithm with delayed coefficient adaptation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 9, pp. 1397–1405, Sept. 1989.

[18] ——, "Correction to 'The LMS algorithm with delayed coefficient adaptation,'" *IEEE Trans. Signal Processing*, vol. 40, pp. 230–232, Jan. 1992.

[19] R. Haimi-Cohen, H. Herzberg, and Y. Be'ery, "Delayed adaptive LMS filtering: Current results," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 1273–1276.

[20] P. Voltz, "Sample convergence of the normalized LMS algorithm with feedback delay," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1991, pp. 2129–2132.

[21] A. Weiss and D. Mitra, "Digital adaptive filters: Conditions for convergence, effects of noise and errors arising from the implementation," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 6, pp. 637–652, Nov. 1979.

[22] R. R. Bitmead and B. D. O Anderson, "Performance of adaptive estimation algorithms in dependent random environments," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 4, pp. 788–794, Aug. 1980.

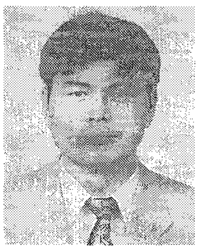
[23] ——, "Lyapunov techniques for the exponential stability of linear difference equations with random coefficients," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 4, pp. 782–787, Aug. 1980.

[24] P. Voltz and F. Kozin, "Almost-sure convergence of adaptive algorithm by projections," *IEEE Trans. Automat. Contr.*, vol. 34, no. 3, pp. 325–327, 1989.

[25] D. Shi and F. Kozin, "On almost sure convergence of adaptive algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-31, no. 5, pp. 471–474, May 1986.

[26] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, no. 8, pp. 1151–1162, Aug. 1976.

[27] S. Ahn, "Convergence properties of LMS type algorithms with update delay," Ph.D. dissertation, Polytechnic Univ., Farmingdale, NY, Jan. 1994.



Sang-Sik Ahn (M'95) received the B.S. degree in electronic engineering and the M.S. degree in electrical engineering from Korea University, Seoul, in 1983 and 1985, respectively, and the Ph.D. degree in electrical engineering from Polytechnic University, Farmingdale, NY, in 1994.

Since 1995, he has been an Assistant Professor of Electrical and Information Engineering at Korea University, Seochang, Jochiwon, Korea. Prior to this, he had been with the GoldStar Central Research Laboratory as a Senior Research Engineer from 1984 to 1987 and as a Principal Research Engineer from 1994 to 1995. From 1991 to 1993, he was a Teaching Fellow at Polytechnic University. His current research interests include adaptive signal processing, communication systems, and VLSI implementation of DSP algorithms.

Peter J. Voltz (M'81) received the Bachelors, Masters, and Ph.D. degrees in electrical engineering from Polytechnic University, Farmingdale, NY, in 1980, 1981, and 1987, respectively.

From 1981 to 1985, he was employed full time at Hazeltine Corporation, Greenlawn, NY, where he was involved in the design, analysis, and simulation of bandwidth efficient and spread spectrum radio systems. He has been a faculty member at Polytechnic University since the fall of 1987. His research interests have included the study of adaptive algorithms and their convergence behavior, parameter identification techniques in controlled Markov chains, and source localization in the underwater environment. More recently, he has been moving toward the general area of wireless communication networks, with specific interests including optimization of signal acquisition and tracking techniques in dynamic multipath environments and adaptive power control and hand-off issues in cellular CDMA networks.