# Asymptotic Analysis of Stochastic Gradient-Based Adaptive Filtering Algorithms with General Cost Functions

Rajesh Sharma, *Member, IEEE*, William A. Sethares, *Member, IEEE*, and James A. Bucklew, *Member, IEEE*

***Abstract*— This paper presents an analysis of stochastic gradient-based adaptive algorithms with general cost functions. The analysis holds under mild assumptions on the inputs and the cost function. The method of analysis is based on an asymptotic analysis of fixed stepsize adaptive algorithms and gives almost sure results regarding the behavior of the parameter estimates, whereas previous stochastic analyses typically consider mean and mean square behavior. The parameter estimates are shown to enter a small neighborhood about the optimum value and remain there for a finite length of time. Furthermore, almost sure exponential bounds are given for the rate of convergence of the parameter estimates. The asymptotic distribution of the parameter estimates is shown to be Gaussian with mean equal to the optimum value and covariance matrix that depends on the input statistics. Specific adaptive algorithms that fall under the framework of this paper are signed error least mean squre (LMS), dual sign LMS, quantized state LMS, least mean fourth, dead zone algorithms, momentum algorithms, and leaky LMS.**

## I. Introduction

ADAPTIVE filtering algorithms have become a vital part of many modern communication and control systems. Adaptive filters are used in system identification, adaptive equalization, linear predictive coding, and adaptive differential pulse code modulation, just to name a few [1]. The most popular adaptive filtering algorithm is the least mean squares (LMS) algorithm, which has enjoyed enormous popularity due to its simplicity and robustness. Over the years several variants of LMS have been proposed to overcome certain limitations.

Let $\{X_k\}$ denote a $\mathbb{R}^d$-valued "input process" and $\{D_k\}$ the real valued "desired response process." The goal of LMS is to recursively adjust the weights of a transversal filter $W = [w_1, \cdots, w_d]^T \in \mathbb{R}^d$ so that $E[J(D_k - W^T X_k)]$ is minimized, where $J(e) = e^2/2$ is the mean square "cost" function. The optimum vector $W_k^*$ that solves this problem is given by[1]

$$W_k^* = E[X_k X_k^T]^{-1} E[X_k D_k].$$

Manuscript received May 25, 1995; revised February 13, 1996. The associate editor coordinating the review of this manuscript and appproving it for publication was Dr. Stephen M. McLaughlin.

R. Sharma is with the Signal Processing Department, Environmental Research Institute of Michigan, Ann Arbor, MI 48113-4001 USA (e-mail: sharma@erim.org).

W. A. Sethares and J. A. Bucklew are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706 USA.

Publisher Item Identifier S 1053-587X(96)06675-5.

[1] Assuming $E[X_k X_k^T]$ is invertible.

Assuming stationarity, $E[X_k X_k^T]$ and $E[X_k D_k]$ do not depend on $k$ and hence $W_k^*$ is independent of $k$. The LMS algorithm attempts to solve this minimum mean square error (MSE) problem recursively using the stochastic gradient of the error surface [1], [16]. If $\hat{W}_k$ is an estimate of $W^*$ at the $k$th iteration and $e_k = D_k - \hat{W}_k^T X_k$ is the instantaneous "error" at the $k$th iteration, then

$$\hat{W}_{k+1} = \hat{W}_k - \mu \frac{\partial J(e_k)}{\partial \hat{W}_k} \tag{1}$$

where $\mu > 0$ is called the stepsize. Then (1) can be rewritten as

$$\hat{W}_{k+1} = \hat{W}_k + \mu e_k X_k. \tag{2}$$

A variety of related stochastic gradient algorithms can be specified by varying the cost function $J(\cdot)$. For instance, minimizing the absolute value of the instantaneous error, $J(e_k) = |e_k|$, leads to

$$\hat{W}_{k+1} = \hat{W}_k + \mu \operatorname{sgn}(e_k) X_k$$

which is a variant of LMS known as the signed error LMS, where

$$\operatorname{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases}$$

More generally, for any differentiable cost function $J$ define $f(z) = J'(z)$. The gradient-based recursion for the parameter estimates is then

$$\hat{W}_{k+1} = \hat{W}_k + \mu f(e_k) X_k. \tag{3}$$

One motivation for considering nonmean-square cost functions is the improvement offered in algorithm performance when the interfering noise is non-Gaussian [2]. Another important reason is computational simplicity. For example, signed error LMS is computationally simpler than LMS.

Another common class of cost functions includes both the error $e_k$ and the the parameter estimate $\hat{W}_k$. For instance

$$J(e_k, \hat{W}_k) = \frac{e_k^2 + \lambda \|\hat{W}_k\|^2}{2} \tag{4}$$

leads to the "leaky LMS" algorithms

$$\hat{W}_{k+1} = (1 - \lambda\mu)\hat{W}_k + \mu e_k X_k, \tag{5}$$

where $\lambda > 0$ and $1 - \lambda\mu$ is called the "leakage factor." It is often argued that leaky LMS stabilizes digital implementations of

the adaptive filter and increases its robustness [1, Chapter 19]. Let $\phi: \mathbb{R} \to \mathbb{R}$ be a differentiable function and $f(z) = \phi'(z)$. While the leaky cost function (4) penalizes the magnitude of the parameter estimate, the cost function

$$J(e_k, \hat{W}_k, \hat{W}_{k-1}) = \frac{\phi(e_k) + \frac{\alpha}{\mu}\|\hat{W}_k - \hat{W}_{k-1}\|^2}{2} \quad (6)$$

penalizes (or rewards) changes in the parameter estimates depending on the sign of $\alpha \in (-1, 1)$. Again, using the stochastic gradient approach leads to the algorithm

$$\hat{W}_{k+1} = \hat{W}_k + \mu f(e_k) X_k + \alpha(\hat{W}_k - \hat{W}_{k-1}) \quad (7)$$

which is commonly referred to as a "momentum" algorithm ($\alpha$ is called the momentum factor [11]).

This paper presents a global and local analysis of (3), (7), and (5) for small stepsize $\mu$. The analysis is based on the presentation in Section II of [12] where asymptotic results are presented regarding the behavior of fixed stepsize adaptive algorithms. Analysis could also be based on the presentations in the books by Benveniste *et al.* [18] or Kushner [19]. However, the results presented in [12] enable us to derive almost sure results, whereas the results in [18] and [19] would enable us to derive results that hold only in probability.

In Section III of [12], the results of Section II of [12] are applied to LMS and its signed variants. For example, it is shown that if $E[X_k X_k^T]$ is positive definite, then the signed error LMS is locally stable about the optimum value. Roughly speaking, this means if the parameter estimate $\hat{W}_k$ is "close" to the optimum value, then the parameter estimates will remain close to the optimum value over a finite time interval. The probability that this happens can be made arbitrarily close to one by choosing $\mu$ sufficiently small. In this paper, it is shown irrespective of where the parameter estimate $\hat{W}_k$ lies, the parameter estimates will enter a "small" ball about the optimum value in finite time. Furthermore, this behavior holds almost surely. These results hold not just for the signed error LMS but for other stochastic gradient algorithms as well. Previous analysis of (3), (7), and (5) has been based on the independence theory approach [1], [2] or a deterministic averaging approach [16]. As examples, the reader is referred to [3] for an independence theory analysis and to [7] for a deterministic averaging analysis of (3). Whereas previous analysis dealt with deterministic periodic inputs and stationary Gaussian inputs, the results of this paper hold for a large class of stationary inputs and cost functions. No independence assumptions are made regarding the input process $\{X_k\}$ or between $X_k$ and $W_k$. Previous stochastic analysis of (3) analyzed mean and mean square behavior of the parameter estimates under such independence assumptions. For example, in [3] it is assumed that $X_k$ is independent of $X_j$ for $j < k$. Clearly, this is violated in many practical situations where $X_k$ contains elements of $X_{k-1}$. Analysis of (3), (7), and (5) for fixed $\mu$ is difficult except for the simplest disturbance-free cases. In general, with disturbances and $\mu$ a fixed small number, $\hat{W}_k$ does not tend to any deterministic limit. In most cases of practical interest the limit does not exist.

Generically, "well-behaved" fixed $\mu$ algorithms start from some arbitrary initial point. The parameter estimates eventually move to a small ball about the optimum value and then fluctuate about that optimum value. Our analysis yields that this behavior holds w.p. 1 (with probability one) for (3), (7), and (5) under mild assumptions. The w.p. 1 results are more appealing from a practical viewpoint than mean, mean square, or probability results, since they hold for any realization of the input to the algorithm. Almost sure bounds are derived for the convergence rate of the parameter estimates $\{\hat{W}_k\}$, and these bounds are compared with the convergence rate results of [3]. Once the parameter estimates have entered a small ball about the optimum value, their fluctuations can be described using the "central limit theorem" of [12]. The parameter estimate $\hat{W}_k$ will be asymptotically Gaussian with mean $W^*$ and a covariance matrix that depends on the input statistics. In [12], the steady state results were derived assuming the components of $X_k$ were independently identically distributed (i.i.d.) for each $k$. In this paper we do not make this assumption. The effect on convergence, stability, and steady-state behavior due to the momentum factor $\alpha$ and the leakage parameter $\lambda$ are discussed.

In [3], the calculus of variations is used to obtain an optimum nonlinearity $f$ that minimizes the misadjustment for a given convergence rate. The results of [3] regarding the mean square behavior of the parameter estimates of (3) are shown to hold under more general conditions (A1–A5 in Section II).

The outline of this paper is as follows: Section II analyzes (3), (7), and (5) and makes comparisons with previous work. Section III applies the results regarding (3) to some popular cost functions. Section IV contains a summary, and the Appendix contains proofs of several results in Section II.

## II. Analysis

This section presents an analysis of (3), (7), and (5). First consider (3). Assume that $D_k = W^{*T} X_k + U_{k+1}$, where $\{U_k\}$ is a sequence of real valued i.i.d random variables and $W^* \in \mathbb{R}^d$. Let $W_k = W^* - \hat{W}_k$ denote the parameter estimate error. In terms of the parameter estimate errors, (3) can be rewritten as

$$W_{k+1} = W_k - \mu f(W_k^T X_k + U_{k+1}) X_k. \quad (8)$$

Assume the following hold.

- A1) $f: \mathbb{R} \to \mathbb{R}$ is nondecreasing, sign preserving, and odd symmetric.
- A2) $\{X_k\}_{k=0}^{\infty}$ is a sequence of zero-mean, stationary ergodic $\mathbb{R}^d$-valued random variables. Assume that the distribution of $X_k$ has a density $p_X$.
- A3) $\{U_k\}$ is a sequence of zero-mean, symmetrically distributed random variables with distribution function $F_u$ and bounded continuous density $h_U$. Also, $\{U_k\}_{k=0}^{\infty}$ is statistically independent of $\{X_k\}_{k=0}^{\infty}$.
- A4) $g(z) = \int_{-\infty}^{+\infty} f(z+u) h_U(u)\, du$ is continuously differentiable with $g'(0) > 0$ and $|g'(z)| \leq L_1 |z|^q + L_2$ for some positive integer $q$ and positive real constants $L_1$ and $L_2$.

- A5) Let $h(z) = \int_{-\infty}^{+\infty}[f(z+u)]^2 h_U(u)\,du$. Assume $E[\|X_k\|^{q+2}] < \infty$ and the following hold for each $K > 0$:

$$E[\sup_{\|w\| \le K} \|X_k\|^2 |f(w^T X_k + U_k)|^2] < +\infty \quad (9)$$

$$E[\sup_{\|w\| \le K} \|X_k\|^2 |h(w^T X_k) - g^2(w^T X_k)|] < +\infty. \quad (10)$$

It follows from A1 that $f$ is continuous except on a countable set of points. A1 allows a great deal of flexibility and certainly in most applications $f$ will have no more than a finite number of discontinuities. The stationary ergodic assumption is sufficiently general to include a large collection of input processes $\{X_k\}$. A1, A3, and A4 imply that $g$ is nondecreasing, odd, sign preserving and $g(x) = 0$ if and only if $x = 0$. The differentiability assumption on $g$ is not stringent, since it is a smoothed version of $f$, and the moment assumptions A5 are not stringent, since the supremum is taken over compact sets in $\mathbb{R}^d$.

Let $R = E[X_k X_k^T]$. For $\alpha \in \mathbb{R}^d$ and $\alpha \neq 0$,

$$\alpha^T R \alpha = \int_{\mathbb{R}^d} |\alpha^T x|^2 p_X(x)\,dx > 0$$

and thus $R$ is positive definite. The parameter estimate error in (8) depends on the stepsize $\mu$. When needed this dependence will be emphasized by writing $W_k^\mu$.

Suppose, for simplicity, that $W_0 \in \mathbb{R}^d$ is a fixed constant (deterministic). As is shown in Appendix I, there exists a continuous $\mathbb{R}^d$-valued deterministic function $W(t)$ defined on $[0, \infty)$, where $\|W(t)\|$ is strictly decreasing and

$$\lim_{t \to \infty} W(t) = 0$$

such that w.p. 1 for any $T > 0$

$$\lim_{\mu \to 0} \max_{0 \le k \le [T/\mu]} \|W_k^\mu - W(k\mu)\| = 0. \quad (11)$$

A direct consequence of (11) is that w.p. 1 the parameter estimate error $\{W_k\}$ will enter a "small" ball about the origin and remain there for a finite length of time. To see this, let $\epsilon > 0$ be fixed and $B_\epsilon = \{w \in \mathbb{R}^d : \|w\| < \epsilon\}$. Since $\lim_{t \to \infty} W(t) = 0$, there exists a $T > 0$ such that for all $t \ge T$, $\|W(t)\| < \epsilon/2$. Then (11) implies there exists a $\mu_0 > 0$ such that for any $\mu \le \mu_0$

$$\max_{0 \le k \le [2T/\mu]} \|W_k^\mu - W(k\mu)\| < \frac{\epsilon}{2}.$$

Therefore, $W_k \in B_\epsilon$ for $[T/\mu] < k \le [2T/\mu]$. In summary, given an initial condition $W_0$, for small stepsizes the parameter estimate errors $\{W_k\}$ will converge to a small ball about the origin and remain in the ball for a finite length of time. In terms of the parameter estimates, this implies convergence to a ball about the optimum value $W^*$. However, (11) does not imply the parameter estimate errors will remain in $B_\epsilon$ forever.

The asymptotic distribution of the parameter estimate errors can be determined by applying Theorem 2 of [12] to (8) and then applying the properties of multidimensional Ornstein–Uhlenbeck processes (the technical details are omitted). For small $\mu$, asymptotically (for large $k$) the parameter estimate error has a Gaussian distribution with mean zero and covariance

$$\frac{\mu E[(f(U_1))^2]}{2g'(0)} I_{d \times d}. \quad (12)$$

The manner in which the parameter estimates fluctuate about the optimum value is described by (12). Note that (12) is similar to the result that was derived in [3] under the assumption that $X_k$ is independent of $X_j$ for $j < k$. Our result (12) relaxes this independence assumption.

Next, we determine the rate at which this convergence to a ball takes place. First, we derive a lower bound on the rate of convergence of the parameter estimate errors in (8). Assume, in addition, that $|g(z)| \le L^*|z|$, where $L^*$ is a finite positive constant. If $g'(0) = \sup_z g'(z)$ (as occurs in the signed error and dual sign algorithms) then $L^*$ can be taken to be $g'(0)$. It is shown in Appendix I that

$$\|W(t)\|^2 = \|W(0)\|^2 - 2\int_0^t G(W(s))\,ds$$

where

$$\begin{aligned} G(w) &= \int_{\mathbb{R}^d} g(w^T x) w^T x p_X(x)\,dx \\ &\le L^* \int_{\mathbb{R}^d} (w^T x)^2 p_X(x)\,dx \\ &= L^* w^T R w \\ &\le L^* \lambda_d \|w\|^2 \end{aligned}$$

and where $\lambda_d$ is the largest eigenvalue of $R$. Therefore

$$\begin{aligned} \frac{d\|W(t)\|^2}{dt} &= -2G(W(t)) \\ &\ge -2L^* \lambda_d \|W(t)\|^2. \end{aligned}$$

Dividing both sides by $\|W(t)\|^2$ and integrating, it follows that

$$\|W(t)\| \ge \|W(0)\| e^{-L^* \lambda_d t}. \quad (13)$$

Similarly, if $|g(z)| \ge L_*|z|$, then $\|W(t)\| \le \|W(0)\| e^{-L_* \lambda_1 t}$ where $\lambda_1$ is the smallest eigenvalue of $R$. However, the condition $|g(z)| \ge L_*|z|$ is rather restrictive, e.g., if $g$ is bounded as in the signed error and dual sign algorithms.

Suppose $X_k$ has compact support. Then there exists a $B_x < \infty$ such that $\|X_k\| \le B_x$ for all $k$. For $z > 0$, let

$$\theta(z) = \sup\{m \in \mathbb{R} : |g(y)| \ge m|y| \text{ for all } |y| \le z\}. \quad (14)$$

Note that in (14) the supremum is taken of a nonempty set and $\theta(z)$ is itself an element of that set. It follows that $\theta(z)$ is a nonincreasing function of $z$. Also, $\lim_{z \to 0} \theta(z) = g'(0)$. Suppose $B > 0$ and $\|W_0\| > B$, then for $w \in \mathbb{R}^d$ such that $\|w\| \le B < \infty$ it follows $G(w) \ge \theta(BB_x)\lambda_1 \|w\|^2$. Let $\|W(t_B)\| = B$. Therefore, for $t \ge t_B$

$$\|W(t)\| \le B e^{-\theta(BB_x)\lambda_1 (t - t_B)}. \quad (15)$$

Since (11) implies that the evolution of the parameter estimate errors in (8) is well approximated by the evolution of $W(k\mu)$ for small $\mu$, (13) and (15) imply that the convergence of the parameter estimate errors to a small ball about the origin can be upper and lower bounded by functions that decay exponentially fast.

In [3], an optimum nonlinearity $f$ for (8) was computed, which minimized the misadjustment for a given convergence rate, which was fixed by setting $g'(0) = C$. Specifically, the approach of [3] solved the following problem

$$\min_f E[f^2(U_k)]$$

such that

$$g'(0) = C.$$

Note that under appropriate conditions $g'(0) = E[f'(U_k)]$. Why does $g'(0)$ determine the convergence rate of (8)? Roughly speaking, $g'(0)$ determines the convergence rate of the algorithm when the parameter estimate errors are "close" to the origin. The reasoning is as follows: Recall that for small $\mu$ the convergence rate of $W_k$ to a ball about the origin is well approximated by the convergence rate of $W(k\mu)$ to the same ball about the origin [see (11)]. The rate of decay in (13) depends on $L^*\lambda_d$ and the rate of decay in (15) depends on the function $\theta(z)$ and $\lambda_1$. Intuitively, one would expect that the convergence rate of $\{W_k\}$ should depend on more than just the single value $g'(0)$, which only takes into account the behavior of $g$ near the origin. Clearly, the function $\theta(z)$ depends on $g$ in a more global manner. Since $\lim_{z\downarrow 0} \theta(z) = g'(0)$, for parameter estimate errors sufficiently close to the origin,[2] (15) implies that the decay of $W(k\mu)$ can be approximated by $g'(0)\lambda_1$, since $\theta(BB_x)\lambda_1 \approx g'(0)\lambda_1$.

Analysis of the momentum algorithms (7) is similar to the analysis of (3). Under the same assumptions, the parameter estimate errors satisfy

$$W_{k+1} = W_k - \mu f(W_k^T X_k + U_{k+1})X_k + \alpha(W_k - W_{k-1}) \tag{16}$$

where $\alpha \in (-1, 1)$ is the momentum factor. Momentum algorithms can be viewed as linearly filtered gradient algorithms. More specifically, let $\{\Gamma_k\}$ denote the linearly filtered gradients, where $\Gamma_k = \alpha\Gamma_{k-1} + f(e_k)X_k$, then

$$\hat{W}_{k+1} = \hat{W}_k + \mu\Gamma_k.$$

To emphasize the dependence on $\mu$ and $\alpha$, $W_k$ will sometimes be written as $W_k^{\mu,\alpha}$. If A1–A5 hold, then it is shown in Appendix II that there exists a collection of differentiable functions $\{\mathcal{W}_\alpha\}_{\alpha\in(-1,1)}$ mapping $[0, \infty)$ into $\mathbb{R}^d$ such that for each $\alpha$, $\|\mathcal{W}_\alpha\|$ is strictly decreasing

$$\mathcal{W}_\alpha((1-\alpha)t) = \mathcal{W}_0(t) \tag{17}$$

and

$$\lim_{t\to\infty} \mathcal{W}_\alpha(t) = 0.$$

[2] $B$ small in (15).

Furthermore, w.p. 1 for any $T > 0$

$$\lim_{\mu\to 0} \max_{0\le k\le [T/\mu]} \|W_k^{\mu,\alpha} - \mathcal{W}_\alpha(k\mu)\| = 0. \tag{18}$$

Using the central limit results of [14], it follows that asymptotically (as $k \to \infty$) for small $\mu$, $W_k^{\mu,\alpha}$ is distributed according to a Gaussian random vector with mean zero and covariance

$$\frac{\mu E[(f(U_1))^2]}{2(1-\alpha)g'(0)} I_{d\times d}. \tag{19}$$

From (17) it follows that the rate at which $\mathcal{W}_\alpha(t)$ approaches zero as $t \to \infty$ is increased as $\alpha$ is increased. Hence, for small $\mu$, $W_k^{\mu,\alpha}$ will enter a small ball about the origin faster for larger $\alpha$. The results regarding rate of convergence of $\mathcal{W}_\alpha(t)$ to the origin as $t \to \infty$ follow using (17). If $|g(z)| \le L^*|z|$, then

$$\|\mathcal{W}_\alpha(t)\| \ge \|\mathcal{W}_\alpha(0)\| e^{-L^*\lambda_d t/1-\alpha}.$$

Suppose $\|W_0\| > B$. Let $\|\mathcal{W}_\alpha(t_B)\| = B$. Therefore, for $t \ge t_B$

$$\|\mathcal{W}_\alpha(t)\| \le Be^{-[\theta(BB_x)\lambda_1(t-t_B)/1-\alpha]}. \tag{20}$$

It follows using (17) and (19) that $W_k^{\mu(1-\alpha),\alpha}$ and $W_k^{\mu,0}$ will have the same rate of convergence to the origin and will asymptotically have the same covariance matrix. Intuitively, however, $W_k^{\mu(1-\alpha),\alpha}$ will exhibit smoother convergence due to the lowpass filtering.

As stated in the previous section, the cost function (4) gives rise to the leaky LMS update

$$\hat{W}_{k+1} = \hat{W}_k + \mu\{e_k X_k - \lambda\hat{W}_k\}$$

where $\lambda > 0$ is the leakage factor. Assume $D_k = W^{*T}X_k + U_{k+1}, \{X_k\}$ (a sequence of stationary ergodic $\mathbb{R}^d$-valued random variables) is independent of $\{U_k\}$, a sequence of i.i.d. random variables. Also assume $E[\|X_k\|^2]$ and $E[U_k^2]$ are both finite and $R = E[X_k X_k^T]$ is positive definite. For simplicity, assume $\hat{W}_0$ is a fixed deterministic constant. Let $W_\lambda^* = (R + \lambda I)^{-1}RW^*$ and let $W_k = W_\lambda^* - \hat{W}_k$ denote the parameter estimate error. Then

$$W_{k+1} = W_k - \mu[(W_k^T X_k + U_{k+1} + (W^* - W_\lambda^*)^T X_k) \cdot X_k + \lambda W_k - \lambda W_\lambda^*].$$

As shown in Appendix III, w.p. 1 for any $T > 0$

$$\lim_{\mu\to 0} \max_{0\le k\le [T/\mu]} \|W_k^\mu - W(k\mu)\| = 0 \tag{21}$$

where

$$W(t) = \sum_{i=1}^{d} e^{-(\lambda+\lambda_i)t} q_i q_i^T W_0$$

and $W_0 = W_\lambda^* - \hat{W}_0$. Here $q_i$ is the $i$th eigenvector of $R$ corresponding to the eigenvalue $\lambda_i$, that is $R = Q\Lambda Q^T$ and $QQ^T = I$, where $Q = [q_1, q_2, \cdots, q_d]$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$.

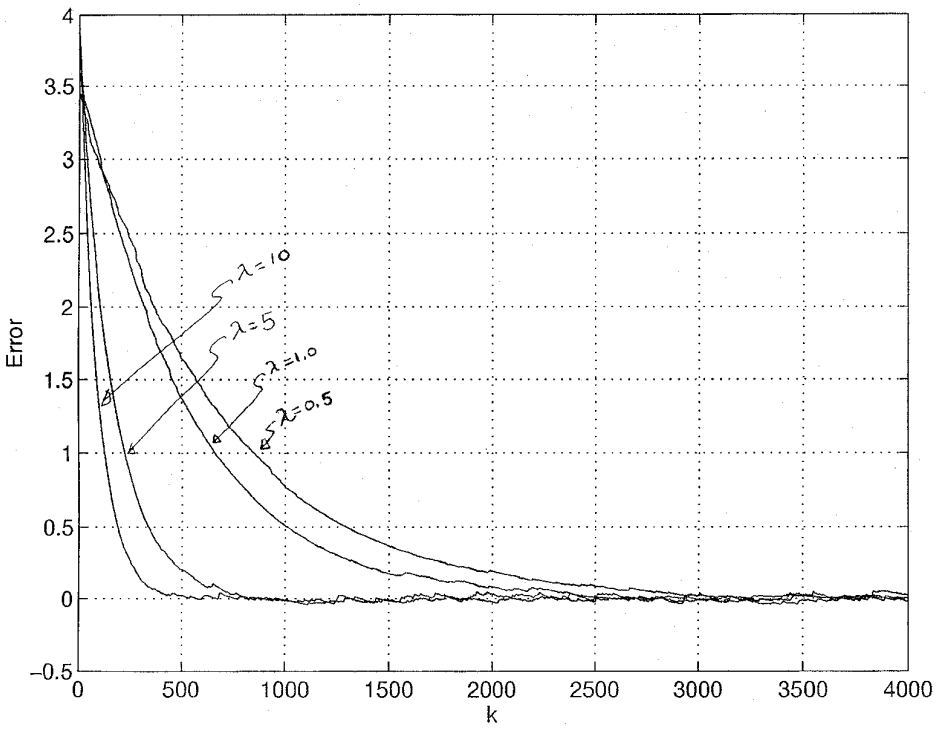



Fig. 1. Error trajectories for various leakage factors $\lambda$.

Note

$$\|W(t)\|^2 = \sum_{i=1}^{d} e^{-2(\lambda+\lambda_i)t}(q_i^T W_0)^2$$

and, thus, by increasing $\lambda > 0$, the rate of decrease of $W(t)$ to the origin as $t \to \infty$ is accelerated. Hence, for small fixed $\mu$ it follows from (21) that the rate at which $\hat{W}_k^{\lambda}$ enters a ball about $W_\lambda^*$ is increased as $\lambda$ is increased. Fig. 1 illustrates this behavior via simulations for the case of a scalar $W_k$ $(d = 1)$. In Fig. 1, *error* trajectories for various leakage factors are plotted. Note that the *error* trajectories correponding to large leakage factors decay faster to the origin.

One consequence of (12) is that asymptotically the parameter estimate error vector for LMS consists of uncorrelated components. However, this need not be the case for the parameter estimate error vector for leaky LMS. It is shown in Appendix III that asymptotically (large $k$) for small $\mu, W_k$ is distributed like a Gaussian random vector with mean zero and covariance $\mu E[U_1^2]\Sigma$, where $\Sigma$ is defined in (33). In general $\Sigma$ is not a diagonal matrix, and thus $W_k$ for large $k$ does not consist of uncorrelated components.

## III. Applications

The results of the previous section regarding (3) are applied to four examples in this section. Assume that A2 and A3 are satisfied in all the examples, since these assumptions do not depend on the particular nonlinearity $f$.

### A. Signed Error Algorithm

Assume $h_U(0) > 0$ and $E[\|X_k\|^2] < \infty$. If $\phi(z) = |z|$, then $f(z) = \operatorname{sgn}(z)$ and (8) corresponds to the signed error variant of the LMS algorithm. A1 is satisfied, since $f$ is discontinuous only at the point $z = 0$. The corresponding $g$ is

$$g(z) = \int_{-\infty}^{+\infty} \operatorname{sgn}(u+z)h_U(u)\,du = 1 - 2F_u(-z).$$

Note that $g'(z) = 2h_U(-z)$ is bounded and continuous by assumption and, hence, A4 holds. Both $f$ and $g'$ are bounded and A5 holds since $E[\|X_k\|^2] < +\infty$. Hence, for small $\mu, W_k$ will enter a small ball about the origin in finite time. For large $k$ and small $\mu, W_k$ is distributed according to a Gaussian random vector with mean zero and covariance

$$\frac{\mu}{4h_U(0)} I_{d\times d}.$$

*Comment:* In [4], the signed error algorithm was analyzed with decreasing stepsizes. It was shown that $\lim_{k\to\infty} W_k = 0$ w.p. 1, where

$$W_{k+1} = W_k - \frac{1}{k}\operatorname{sgn}(W_k^T X_k)X_k. \tag{22}$$

Note in [4] it is assumed that the disturbance $U_k = 0$ for all $k$, which is not the case in many applications. Furthermore, our results hold for small nonzero $\mu$. In practical applications, the stepsize is not allowed to vanish, since this would defeat the "adaptive" purpose of the filter. It should be noted that in [5], the author of [4] relaxed the noise-free assumption in 22.

### B. Dual Sign Algorithm

Let $l_1$ and $l_2$ be positive real numbers with $0 < l_1 < l_2$ and $m > 0$. Assume $h_U(0) > 0$ or $h_U(m) > 0$ and $E[\|X_k\|^2] < \infty$.

If[3]

$$f(z) = l_1 I_{[0,m)}(z) + l_2 I_{[m,\infty)}(z) - l_1 I_{(-m,0)}(z) - l_2 I_{(-\infty,-m]}(z)$$

then (8) corresponds to the dual sign algorithm of [6]. A1 is satisfied, since $f$ is discontinuous at only three points. The corresponding $g$ is

$$g(z) = l_1[F_U(m-z) - F_u(-z)] + l_2[1 - F_U(m-z)] - l_1[F_U(-z) - F_U(-m-z)] - l_2 F_U(-m-z).$$

Note that both $f$ and $g$ are bounded functions and the derivative of $g$ is

$$g'(z) = -l_1 h_U(m-z) + l_1 h_U(-z) + l_2 h_U(m-z) + l_1 h_U(-z) - l_1 h_U(-m-z) + l_2 h_U(-m-z)$$

which is a bounded continuous function. Since $h_U(0)$ or $h_U(m) > 0$ it follows that $g'(0)$ will be strictly positive and, hence, A4 holds. Since $f$ and $g'$ are bounded, A5 is satisfied. Hence, $W_k$ will enter a ball about the origin and for large $k$ and $W_k$ is asymptotically Gaussian with mean zero and covariance

$$\frac{\mu E[f(U_k)^2]}{2[2h_U(m)(l_2 - l_1) + 2h_U(0)l_1]} I_{d\times d}. \tag{23}$$

Similar analyses could by readily accomplished for more general "quantization" functions of the error. Note that when $l_2 = l_1 = 1$, the expression for the covariance matrix reduces to that of the covariance matrix for the signed error algorithm. Hence, using (23) $l_1, l_2, m$ and $\mu$ can be chosen to give a desired steady-state variance.

### C. Dead Zone

Assume $E[\|X_k\|^4] < \infty$. Let

$$f(z) = (z+c)I_{(-\infty,-c]}(z) + (z-c)I_{[c,\infty)}(z)$$

where $c$ is a finite positive constant that determines the width of the dead zone. A1 is satisfied, since $f$ is continuous. Also

$$g(z) = (z-c)(1 - F_U(c-z)) + (z+c)F_U(-c-z) + \int_{-c-z}^{c-z} -u h_U(u)\, du.$$

Furthermore

$$g'(z) = 1 + F_U(-c-z) - F_U(c-z)$$

which is a bounded continuous function. Note that $g'(0) = 1 - (F_U(c) - F_U(-c))$. Using arguments similar to those found in [12], it follows that (8) will be locally stable if $g'(0)R$ is positive definite. Note $g'(0) \geq 0$ and $g'(0) = 0$ if and only if $F_U(c) - F_U(-c) = 1$. Thus, we assume $F_U(c) - F_U(-c) < 1$, to ensure local stability. Under these conditions, A4 and A5 are satisfied. Therefore, the parameter estimates will enter a small ball about the origin for small $\mu$ and will be asymptotically Gaussian with mean zero and covariance

$$\frac{\mu E[f(U_k)^2]}{2(1 - (F_U(c) - F_U(-c)))} I_{d\times d}. \tag{24}$$

[3] $I_A(z) = 1$ if $z \in A$ and zero otherwise.

By decreasing $c \downarrow 0$ the variance

$$E[\|W_k\|^2] = d\mu \frac{E[f(U_k)^2]}{2(1 - (F_U(c) - F_U(-c)))}$$

decreases and by increasing $c \uparrow \infty$ the variance increases. Note that $c = 0$ is the case for LMS.

### D. Cubic Nonlinearity

If $\phi(z) = z^4/4$ then $f(z) = z^3$ and (8) corresponds to the least mean fourth (LMF) adaptive algorithm [8]. The corresponding $g$ is $g(z) = z^3 + 3\sigma_U^2 z$, where $\sigma_U^2 = E[U_k^2]$. A1–A5 hold, and the conclusions of the previous examples hold. In particular, the asymptotic distribution of $W_k$ is Gaussian with mean zero and covariance

$$\frac{\mu E[U_k^6]}{6\sigma_U^2} I_{d\times d}.$$

Analogous results hold for cost functions of the form $\phi(z) = z^\rho \rho$, where $\rho$ is an even integer.

## IV. Summary

This paper analyzed stochastic gradient adaptive algorithms of the general form

$$\hat{W}_{k+1} = \hat{W}_k - \mu \frac{\partial J}{\partial \hat{W}_k}$$

where $\hat{W}_k$ is the parameter estimate at the $k$th iteration, $\mu > 0$ and $J$ is a cost function. Several examples were given, including leaky LMS, momentum algorithms, cubic nonlinearities, dead zone algorithms and various sign algorithms. The results of Section II and the examples of Section III can be mixed and matched to consider a large variety of algorithms such as cubic nonlinearity with momentum.

It was shown that for small $\mu$, the parameter estimates $\{\hat{W}_k\}$ will enter a small ball about the optimum weight vector $W^*$ and remain inside the ball for finite time. For large $k$, the parameter estimates fluctuate about $W^*$ according to a Gaussian distribution. These results hold under mild assumptions on the cost function, the input process, and for any initial starting point. Furthermore, the results hold w.p. 1, whereas previous works considered convergence in the mean, convergence in the mean square, and convergence in probability. Also, almost sure exponential bounds are derived on the rate of convergence.

The effects on algorithm performance due to the incorporation of a leakage factor $\lambda$ in (4) are discussed. It is shown that the rate of convergence is increased as $\lambda$ is increased. Under the assumption that $D_k = X_k^T W^* + U_{k+1}$, it is shown for $\lambda > 0, \mu$ small and $k$ large, the parameter estimate error vector $W_k$ does not necessarily consist of asymptotically correlated components, in contrast to the parameter estimate error vector of LMS, which does consist of uncorrelated components. For the momentum algorithms, the effects on convergence and steady-state behavior due to the momentum factor $\alpha$ are discussed. It is shown that by increasing $\alpha$, the convergence speed and the steady-state variance of the parameter estimates is increased.

As with all asymptotic results ($\mu \to 0$) it is extremely difficult to say precisely how small $\mu$ needs to be for the behavior of the algorithm to be within $\epsilon$ of the asymptotic behavior. This is true for all but the simplest cases (noise free). For stepsizes $\mu$ used in practice, it has been observed that the results accurately predict algorithmic behavior [12]. The steady-state results along with the convergence rate results presented in this paper allow one to obtain a good idea regarding algorithm behavior and how this behavior depends on various parameters. For example, the expression for the steady-state covariance allows one to select a stepsize $\mu$ to achieve a prescribed steady-state variance or predict the dependence of the steady-state variance on other algorithm parameters (For example, for dead zone algorithms the dependence of the steady state variance on the dead zone parameter $d$ is made precise by the expression for the steady state covariance; see Section III). Also, the results presented in this paper allow one to compare various algorithms with regards to convergence and steady state. The main benefit of the general results presented in this paper is that they unify the treatment of a large class of algorithms that were treated separately in numerous papers under the classical independence assumptions.

## APPENDICES

In the Appendix, the results of Section II are derived.

## APPENDIX I

Assume A1–A5 hold. Recall that the parameter estimate errors satisfy (8). For $x \in \mathbb{R}^d, w \in \mathbb{R}^d$ and $U \in \mathbb{R}$ define $H: \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ as

$$H(w, x, U) = f(w^T x + U)x.$$

Note that (8) can be rewritten as

$$W_{k+1} = W_k - \mu H(W_k, X_k, U_{k+1}). \tag{25}$$

Let $\nu_X$ and $\nu_U$ denote the measure induced on $\mathbb{R}^d$ and $\mathbb{R}$ by $X_k$ and $U_k$, respectively. Then for each fixed $w \in \mathbb{R}^d, H(w, \cdot, \cdot)$ is continuous almost surely on $\mathbb{R}^d \times \mathbb{R}$ with respect to the measure $\nu_X \times \nu_U$.

*Proof:* $C_f = \{z \in \mathbb{R}: f \text{ is discontinuous at } z\}$ and let

$$S_w = \{(x, U) \in \mathbb{R}^{d+1}: H(w, \cdot, \cdot) \text{ is discontinuous at } (x, U)\}.$$

Then

$$S_w \subset \bigcup_{z \in C_f} \{(x, U): w^T x + U = z\}.$$

It follows $\{(x, U): w^T x + U = z\}$ is a linear variety in $\mathbb{R}^{d+1}$ where the subspace has dimension $d$. Since $\nu_X \times \nu_U$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{d+1}$ it follows $\nu_X \times \nu_U[\{(x, U): w^T x + U = z\}] = 0$ for any $z \in \mathbb{R}$. Since $C_f$ is countable, the result follows.

Let $W_\mu(t) = W_{[t/\mu]}$ and let $K > 0$ be such that $K > \|W_0\|$. Define the stopped process $W_\mu^{\tau_\mu^K}(t) = W_\mu(t \wedge \tau_\mu^K)$ where

$$\tau_\mu^K = \inf\{t \geq 0: \|W_\mu(t)\| \geq K\}.$$

Then Corollary 1 of [12] applies to (25) and implies the existence of a continuous $\mathbb{R}^d$-valued function $W(t)$ defined on $[0, \infty)$ such that w.p. 1 for any $T > 0$

$$\lim_{\mu \to 0} \sup_{t \leq T \wedge \tau^K} \|W_\mu^{\tau_\mu^K}(t) - W(t)\| = 0$$

and

$$W(t) = W_0 - \int_0^t \hat{H}(W(s))\, ds$$

for all $t < \tau^K$, where

$$\tau^K = \inf\{t \geq 0: \|W(t)\| \geq K\}$$

and $\hat{H}(w) = \int_{\mathbb{R}^d} g(w^T x) x p_X(x)\, dx$ is continuously differentiable. It follows, using Corollary 23 on page 38 of [20], that $\tau^K = \infty$, since $K > \|W_0\|$. Hence, $\lim_{\mu \to 0} \tau_\mu^K = \infty$ and thus we do not need to work with a stopped process. Therefore

$$\|W(t)\|^2 = \|W_0\|^2 - 2 \int_0^t G(W(s))\, ds$$

where

$$G(w) = \int_{\mathbb{R}^d} g(w^T x) w^T x p_X(x)\, dx \tag{26}$$

and w.p. 1, for any $T > 0$

$$\lim_{\mu \to 0} \sup_{t \leq T} \|W_\mu(t) - W(t)\| = 0.$$

Next, it follows that $G(\cdot)$ is a continuous function and $G(w) \geq 0$. Furthermore, $G(w) = 0$ if and only if $w = 0$. Therefore, for any $r > 0$

$$\inf_{\|w\| \geq r} |G(w)| > 0. \tag{27}$$

Hence, (27) implies that $\lim_{t \to \infty} W(t) = 0$ and, thus, (11) holds.

## APPENDIX II

Assume A1–A5 are satisfied. Then Theorem 1 of [14] implies that there exists a collection $\{\mathcal{W}_\alpha\}_{\alpha \in [-\alpha^*, \alpha^*]}$ of continuous $\mathbb{R}^d$-valued functions defined on $[0, \infty)$ such that

$$\|\mathcal{W}_\alpha(t)\|^2 = \|W_0\|^2 - \frac{2}{1-\alpha} \int_{\mathbb{R}^d} \cdot \mathcal{W}_\alpha(t)^T x g(\mathcal{W}_\alpha(t)^T x) p_X(x)\, dx \tag{28}$$

and for each $\alpha, \mathcal{W}_\alpha(t)$ is the unique solution of the ODE

$$\dot{w} = \frac{1}{1-\alpha} \hat{H}(w) \tag{29}$$

with initial condition $w(0) = W_0$. Furthermore, w.p. 1 for

any $T > 0$

$$\lim_{\mu \to 0} \max_{0 \le k \le [T/\mu]} \|W_k^{\mu,\alpha} - \mathcal{W}_\alpha(k\mu)\| = 0. \tag{30}$$

Using (29) and (30), (17) and (18) both follow. Also, (19) follows as a direct application of Theorem 2 of [14].

## APPENDIX III

In this Appendix, results for leaky LMS are derived. Note $W_\lambda^* = (R + \lambda I)^{-1} R W^* = W^* - (R + \lambda I)^{-1} \lambda W^*$ and $W^* = W_\lambda^* + R^{-1} \lambda W_\lambda^*$. Then the parameter estimate error $W_k = W_\lambda^* - \hat{W}_k$ satisfies

$$W_{k+1} = W_k - \mu[(W_k^T X_k + U_{k+1} + (W^* - W_\lambda^*)^T X_k) \cdot X_k + \lambda W_k - \lambda W_\lambda^*]. \tag{31}$$

Equation (31) can be rewritten as

$$W_{k+1} = W_k + \mu H(W_k, X_k, U_{k+1}) \tag{32}$$

where $H: \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is defined in an obvious way. Corollary 1 of [12] applies to (32). Hence, w.p. 1 for any $T > 0$

$$\lim_{\mu \to 0} \max_{0 \le k \le [T/\mu]} \|W_k^\mu - W(k\mu)\| = 0$$

where $W(t)$ is the solution of the ODE $\dot{w} = \hat{H}(w)$ with initial condition $w(0) = W_0$, where $\hat{H}(w) = -(R + \lambda I)w$. Therefore

$$W(t) = \sum_{i=1}^{d} e^{-(\lambda + \lambda_i)t} q_i q_i^T W_0.$$

Let $V_\mu(t) = W_\mu(t) - W(t)/\sqrt{\mu}$. Let $Z_k = X_k X_k^T (W^* - W_\lambda^*) - \lambda W_\lambda^*$. There are a variety of mixing conditions (see [17, ch. 4]) on $\{Z_k\}$ which imply[4]

$$\sqrt{\mu} \sum_{k=0}^{[t/\mu]-1} Z_k \Rightarrow Z(t)$$

where $Z(t)$ is $\mathbb{R}^d$-dimensional Brownian motion with covariance matrix

$$R_Z = E[Z_0 Z_0^T] + \sum_{k=1}^{\infty} (E[Z_0 Z_k^T] + E[Z_k Z_0^T]).$$

Theorem 2 of [12] implies $V_\mu$ converges weakly to $V$ where

$$V(t) = B(t) - (R + \lambda I) \int_0^t V(s)\, ds$$

where $B(t)$ is mean zero $\mathbb{R}^d$-dimensional Brownian motion with covariance matrix $R_B = E[U_1^2]R + R_Z$. Let $R_B = MJM^T$ and $R = Q\Lambda Q^T$, where $J$ and $\Lambda$ are diagonal matrices containing the eigenvalues of of $R_B$ and $R$, respectively and $M$ and $Q$ are orthogonal matrices. That is $J = \mathrm{diag}(\gamma_1, \gamma_2, \cdots, \gamma_d)$ and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ and $QQ^T = I$ and $MM^T = I$. Then [14] it follows that for large $k$ and small $\mu$, $W_k$ is asymptotically Gaussian with mean zero and covariance $\mu E[U_1^2]\Sigma$ with

$$\Sigma = \sum_{k=1}^{d} \sum_{l=1}^{d} \sum_{m=1}^{d} q_k q_k^T m_l m_l^T q_m q_m^T \frac{\gamma_l}{2\lambda + \lambda_k + \lambda_m} \tag{33}$$

where $q_k$ is the $k$th column of $Q$ and $m_k$ is the $k$th column of $M$.

[4] See [17] for the definition of $\Rightarrow$.

## REFERENCES


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

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

[3] S. C. Douglas, T. H.-Y. Meng, "Stochastic gradient adaptation under general error criteria," *IEEE Trans. Signal Processing,* vol. 42, pp. 1335–1351, June 1994.

[4] E. Eweda, "Almost sure convergence of a decreasing gain sign algorithm for adaptive filtering," *IEEE Trans. Acoust., Speech, Signal Processing,* vol 36, pp. 1669–1671, Oct. 1988.

[5] ——, "Convergence of the sign algorithm for adaptive filtering with correlated data," *IEEE Trans. Inform. Theory,* vol. 37, pp. 1450–1457, Sept. 1991.

[6] C. P. Kwong, "Dual sign-algorithm for adaptive filtering," *IEEE Trans. Commun.*, vol. COM-34, no. 12, pp. 1272–1275, Dec. 1986.

[7] W. A. Sethares, "Adaptive algorithms with nonlinear data and error," *IEEE Trans. Signal Processing,* vol. 40, pp. 2199–2206, Sept. 1992.

[8] E. Walach and B. Widrow, "The least mean fourth (LMF) adaptive algorithm and its family," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 275–283, Mar. 1984.

[9] N. J. Bershad and L. Z. Qu, "On the probability density function of the complex scalar LMS adaptive weights," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 43–56, Jan. 1989.

[10] J. G. Proakis, "Channel identification for high speed digital communications," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 916–922, Dec. 1974.

[11] S. Roy and J. J. Shynk, "Analysis of the momentum LMS algorithm," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 38, no. 12, Dec. 1990.

[12] J. A. Bucklew, T. G. Kurtz, and W. A. Sethares, "Weak convergence and local stability properties of fixed step size recursive algorithms," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 966–978, May, 1993.

[13] J. A. Bucklew and W. A. Sethares, "The covering problem and $\mu$-dependent adaptive algorithms," *IEEE Trans. Signal Processing,* vol. 42, no. 10, pp. 2616–2627, Oct., 1994.

[14] R. Sharma, W. A. Sethares, and J. A. Bucklew, "Analysis of momentum adaptive filtering algorithms," *IEEE Trans. Signal Processing*, submitted for publication.

[15] R. Durrett, *Probability: Theory and Examples.* CA: Wadsworth Brooks/Cole, 1991.

[16] C. R. Johnson, Jr., *Lectures on Adaptive Parameter Estimation.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[17] P. Billingsley, *Convergence of Probability Measures.* New York: Wiley, 1968.

[18] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations.* New York: Springer-Verlag, 1990.

[19] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, MIT Press Series in Signal Processing, Optimization, and Control.* Cambridge, MA: MIT Press, 1984.

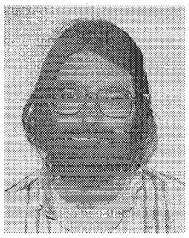
[20] M. Vidyasagar, *Nonlinear System Analysis.* Englewood Cliffs, NJ: Prentice-Hall, 1993.



**Rajesh Sharma** (S'92–M'95) was born in Srinagar, India, on September 20, 1970. He received the B.S. degree (with distinction) and the M.S. and Ph.D. degrees all from the University of Wisconsin, Madison, in 1991, 1992, and 1995, respectively.

Since 1995, he has been with the Signal Processing Department of the Environmental Research Institute of Michigan, Ann Arbor. His research interests include image segmentation, computer vision, adaptive algorithms, and applied probability.

Dr. Sharma is a member of Eta Kappa Nu.



**William A. Sethares** (S'84–M'86–S'86–M'87) received the B.A. degree in mathematics from Brandeis University, Waltham, MA, and the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY.

He has served with Raytheon Company, Wayland, MA, as a Systems Engineer and is currently on the faculty of the Department of Electrical and Computer Engineering at the University of Wisconsin, Madison. His research interests include adaptive systems in signal processing, communications, control, and electronic music.

**James A. Bucklew** (S'75–M'79) received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1979.

He is currently a Professor with the Department of Electrical and Computer Engineering and the Department of Mathematics of the University of Wisconsin, Madison. His research interests are in the applications of probability to signal processing and communication problems.

Dr. Bucklew is the recipient of a Presidential Young Investigator Award (1984). He has served as the Associate Editor at Large (1989–1992) and the Associate Editor for Detection (1992) for the IEEE TRANSACTIONS ON INFORMATION THEORY.