Are Ensemble-Average Learning Curves Reliable in Evaluating the Performance of Adaptive Filters?*

VÍTOR H. NASCIMENTO AND ALI H. SAYED

Electrical Engineering Department
University of California
Los Angeles, CA 90024

Abstract

We treat here the computation of the learning curves of the LMS *algorithm by simulation (that is, the computation of the* MSE *as a function of the time instant). Since closed-form analytic expressions for learning curves are quite hard to obtain in most practical situations, one usually approximates learning curves by performing several repeated experiments and by averaging the resulting squared-error curves. We show, both by examples and analytically, that when the step-size is large, this approximation of the* MSE *can be misleading. This is contrary to what one would expect, given the excellent agreement one obtains between simulations and theory for small step-sizes and independent inputs, even using only as few as* 10 *experiments. Our theoretical analysis explains both the good results obtained for small step-sizes, and the discrepancies that arise for large step-sizes.*

1. Introduction

An important performance measure for adaptive filters is the mean-square error (MSE) defined by

$$\mathrm{E}\, e(n)^2 = \mathrm{E}\big(y(n) - \boldsymbol{x}_n^T \boldsymbol{w}_{n-1}\big)^2,$$

where $\{y(n)\}$ is the desired sequence, $\{\boldsymbol{x}_n\}$ is the input (regressor) sequence, and $\boldsymbol{w}_{n-1}$ is the weight estimate at time $n-1$. The signal $y(n)$ is assumed to be generated via

$$y(n) = \boldsymbol{x}_n^T \boldsymbol{w}_* + v(n) \ ,$$

for some unknown vector $\boldsymbol{w}_*$ that should be estimated, and where $\{v(n)\}$ denotes measurement noise.

The famed LMS algorithm updates the weight estimates $\boldsymbol{w}_n$ by means of the recursion [4, 10]

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \mu \boldsymbol{x}_n e(n) \ ,$$

for some initial condition $\boldsymbol{w}_0$ and using a positive step-size parameter μ.

The plot of the MSE as a function of the time n is known as the *learning curve* of the algorithm, and it is dependent on the step-size μ. In general it is not a simple task to find analytical expressions for the learning curve or for the steady-state MSE, except when the assumptions of *independence theory* [4, 10] are used. In practice, the learning curve is estimated by experimentation or repeated simulations. More specifically, several independent simulations are performed, say L of them. In each of the experiments, the LMS algorithm is applied for N iterations, always starting from the same initial condition and under the same statistical conditions for the sequences $\{y(n)\}$ and $\{\boldsymbol{x}_n\}$. From each experiment i, a sample curve $e^{(i)}(n)$, $1 \le n \le N$ is obtained. After all L experiments are completed, an approximation for the learning curve is computed via an averaged curve,

$$\mathrm{E}\, e(n)^2 \approx \hat{E}(n) = \frac{1}{L}\sum_{i=1}^{L} e^{(i)}(n)^2, \quad 1 \le n \le N.$$

$\hat{E}(n)$ is referred to as an *ensemble-average* learning curve.

If the step-size μ is small, an average of few tens of experiments is enough to obtain experimental learning curves $\hat{E}(n)$ that are close to the one predicted by independence theory. This one in turn can be shown to approximate the actual learning curve $\mathrm{E}\, e(n)^2$ (i.e., in the absence of the independence assumptions) to first order in μ [5].

To exemplify this behavior, consider a length $M = 10$ LMS adaptive filter operating with Gaussian inputs with covariance matrix $\mathrm{E}\, \boldsymbol{x}_k \boldsymbol{x}_k^T = I$, step-size $\mu = 0.08$, and no noise. The learning curve for this case was computed theoretically in [6]. In Fig. 1 we plot this theoretical curve, in

*This material was based on work supported in part by the National Science Foundation under Awards MIP-9409319 and CCR-9732376. The work of V. H. Nascimento was also supported by a fellowship from Conselho Nacional de Pesquisa e Desenvolvimento - CNPq - Brazil, while on leave from Escola Politécnica da Universidade de São Paulo.

0-7803-5148-7/98/$10.00©1998 IEEE

addition to an ensemble-average learning curve that is obtained from the average of $L = 100$ simulations. Note how both plots are close to each other.

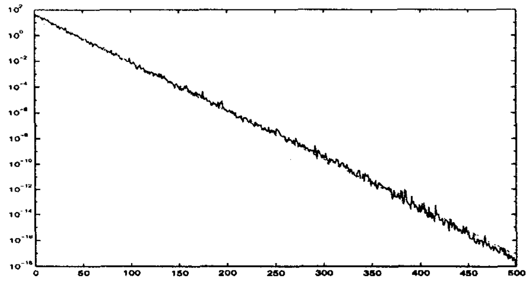

Figure 1. Learning curves computed by simulation and theoretically, with Gaussian iid inputs, $M = 10$, $\mu = 0.08$, and $L = 100$.

Given the good agreement for sufficiently small step-sizes between the ensemble-average learning curve and the actual learning curve, it is now common in the literature to use the average of few independent experiments to predict or confirm theoretical results from simulation results (a few relatively recent examples include [2, 9], which use 10-20 independent experiments, and [7, 8], which use 100 independent experiments).

In this paper, we show by examples and also analytically that for larger step-sizes, it may be necessary to perform a considerably larger number of experiments to correctly approximate the average $\mathrm{E}\, e(n)^2$. In other words, we claim that for large step-sizes more care is needed while interpreting ensemble-average learning curves. These curves can lead to erroneous conclusions unless a large enough number of experiments are averaged (at times of the order of tens of thousands or higher). We study this phenomenon and provide a theoretical justification for its occurrence.

2. Simulations and examples

We start with a few examples. Consider again the adaptive LMS filter of length $M = 10$, with Gaussian input (i.e., the entries of $\boldsymbol{x}_n$ are Gaussian distributed, with zero mean and variance 1), and Gaussian noise $v(n)$ with variance $\sigma_v^2 = 10^{-4}$. In this case, due to independence assumptions, it is possible to compute the learning curve $\mathrm{E}\, e(n)^2$ exactly. In Fig. 2, we plot this theoretical curve, as well as ensemble-average curves computed with $L = 10$, $L = 100$, and $L = 10,000$, now all with step-size $\mu = 0.16$ (which is twice the value of the step-size used to generate Fig. 1). Note how all simulation curves are noticeably far (and, most of the time, below) the (dotted) theoretical curve, although the simulations get closer to the theoretical curve as L is increased. Note also that the simulation curves converge faster than the theoretical curve. This situation should be contrasted with that of Fig. 1, where an almost-perfect agreement was obtained between theory and simulation.

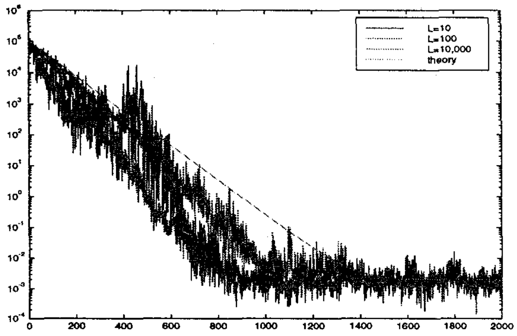

Figure 2. Learning curves computed by simulation and theoretically, with Gaussian independent input vectors, Gaussian noise with $\sigma_v^2 = 10^{-4}$, $M = 10$, $\mu = 0.16$, and $L = 10$, $L = 100$, and $L = 10^4$.

The curve for $L = 10$ in Fig. 2 (the darker line) also shows the mechanism by which the average $\mathrm{E}\, e(n)^2$ is attained — for most of the time, the curve $L = 10$ is below average, but occasionally there are bursts of large error (in this case the largest one is around $n = 500$). These (relatively rare) large bursts, averaged over the ensemble of curves $e^{(i)}(n)$, increase the average to the level predicted by theory.

When the independence assumptions do not hold, these effects still occur. In the next example, the input vectors $\boldsymbol{x}_n$ are not iid, but have a delay-line structure. Let the sequence from which the elements of $\boldsymbol{x}_n$ are taken be $\{a_n\}$. Assuming that this sequence is iid uniformly distributed around -0.5 and 0.5, for $M = 2$, the results of [3] can be used to obtain, analytically, the learning curve $\mathrm{E}\, e(n)^2$. In Fig. 3 we plot this theoretical curve, as well as ensemble-average curves for $L = 100$ to $L = 10,000$, with step-size $\mu = 8.3$.

With this value of μ, $\mathrm{E}\, e(n)^2$ diverges (and the analytical learning curve indeed increases with n), but the simulations show $\frac{1}{L}\sum_{i=1}^{L} e^{(i)}(n)^2$ converging (see Fig. 3 (a)). For $L = 100$, there is no hint of divergence (b). Only for $L = 1,000$ and $10,000$ we can see that there is something wrong; the curve in (c) is increasing from $n = 1$ to $n = 5$, and the curve in (d) is increasing from $n = 1$ to $n = 10$.

These simulations show that the behavior of the ensemble-average curves may be significantly different than that of the theoretical learning curves, if the step-size is large. This may lead to wrong conclusions when one attempts to predict performance from simulation results.

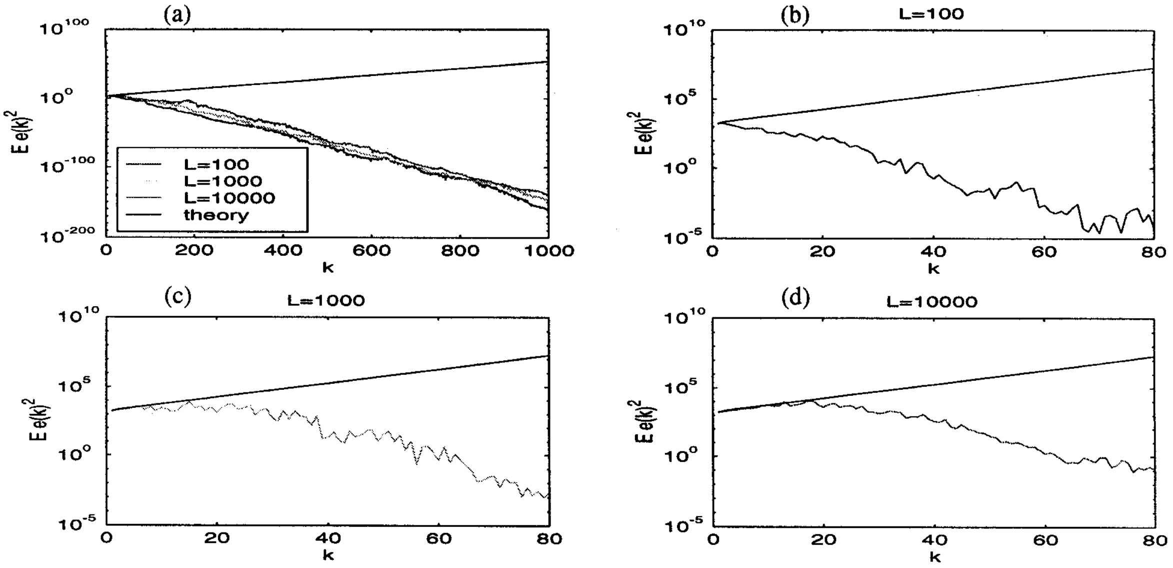

Figure 3. Learning curves computed by simulation and theoretically, with tap-delayed input vectors, $M = 2$, $\mu = 8.3$, and $L = 100$, $L = 1,000$, and $L = 10,000$ (a); theoretical curve and $L = 100$ only (b); theoretical and $L = 1000$ only (c); theoretical and $L = 10,000$ only (d).

In addition, the simulations are consistently below the theoretical curves, and in fact after some time they tend to converge *faster* than predicted by theory (around $n = 0$, the convergence rate predicted by the theory is a good approximation). In the next sections we study the reasons for this behavior, assuming the filter length is 1 ($M = 1$) and that the input is iid.

3. Theoretical analysis in the scalar case

A simple model is used in this section to explain the differences observed between the simulations and theoretical results. More specifically, we study the scalar LMS recursion with independent and identically-distributed stationary inputs $\{x_n\}$. We also assume that x_n is uniformly distributed between $-\alpha$ and α, and that the noise is zero.

3.1 Condition for mean-square stability

With the above assumptions in mind, we square both sides of the LMS error equation to obtain

$$\tilde{w}_n^2 = \left(1 - \mu x_n^2\right)^2 \tilde{w}_{n-1}^2 \, , \qquad (1)$$

where $\tilde{w}_n = w_n - w_*$. This is a stochastic difference equation relating two positive quantities, $\tilde{w}_n^2$ and $\tilde{w}_{n-1}^2$. The relation between both quantities is a random multiplicative factor, which we denote by $u(n) = \left(1 - \mu x_n^2\right)^2$. To simplify the notation, we also denote $Y_n = \tilde{w}_n^2$. Our simplified model is then

$$Y_n = u(n) Y_{n-1} = Y_0 u(1) u(2) \ldots u(n). \qquad (2)$$

Note that from our assumptions on $\{x_n\}$, it follows that the $u(n)$ are iid. Our experiments in the last section showed that there is a distinction between the plots of $\mathrm{E}\,Y_n$ and of $\frac{1}{L}\sum_{l=1}^{L} Y_n^{(l)}$ over several experiments.

In this section we study this model, assuming that the initial condition, $Y_0 = \tilde{w}_0^2$, is deterministic. The first task is to find under which conditions $\mathrm{E}\,Y_n$ converges. Note first that the variance of x_n is $\sigma_x^2 = \alpha^2/3$ and its fourth-order moment is $\sigma_4 = \alpha^4/5$. From (2), and using the independence of the $u(i)$, we obtain

$$\mathrm{E}\,Y_n = \left[1 - \mu\frac{2\alpha^2}{3} + \mu^2\frac{\alpha^4}{5}\right]^n Y_0. \qquad (3)$$

From this equation, one concludes that $\mathrm{E}\,Y_n$ converges to 0 if, and only if, the coefficient on the right-hand side is strictly less than 1, i.e., $0 < \mu\alpha^2 < 10/3$. Observe also from $\mathrm{E}\,Y_n = Eu(n)\,\mathrm{E}\,Y_{n-1}$ that the logarithm of the rate of convergence of $\mathrm{E}\,Y_n$ is equal to $\ln Eu(n)$ (a result that we shall invoke later).

What we just did was the standard mean-square stability analysis using independence theory. The theory thus tells us that the recursion (1) will be mean-square stable if, and only if, $0 < \mu\alpha^2 < 10/3$.

3.2 Behavior of a sample curve

We now study the behavior of a typical curve Y_n. We show in the remainder of this section that, for large n and large step-size μ, Y_n decays (or increases) at a rate significantly different than that of $\mathrm{E}\,Y_n$, predicted by (3). We obtain this result by studying conditions under which Y_n converges to zero with probability one (or almost surely).

Compute the logarithm of Y_n,

$$\ln Y_n = \ln Y_0 + \sum_{i=1}^{n} \ln u(i) \ ,$$

which is therefore a sum of independent and identically distributed random variables with bounded variance (this last fact follows from the distribution of the $u(i)$). Therefore, we can use the law of large numbers [1] to conclude that

$$\frac{\ln Y_n}{n} \xrightarrow{a.s.} \mathrm{E}\big(\ln u(i)\big) \triangleq \mathrm{E}(\ln u). \tag{4}$$

That is, for large n, $(\ln Y_n)/n$ will almost surely converge to a constant, $\mathrm{E}\ln u$. We evaluated this expectation as a function of $\mu\alpha^2$, $\mathrm{E}\ln u$=

$$\begin{cases} \ln\left(1-\mu\alpha^2\right)^2 + \frac{4}{\alpha\sqrt{\mu}}\mathrm{arctanh}\left(\alpha\sqrt{\mu}\right) - 4, & \mu\alpha^2 \le 1, \\ \ln\left(1-\mu\alpha^2\right)^2 + \frac{4}{\alpha\sqrt{\mu}}\mathrm{arccoth}\left(\alpha\sqrt{\mu}\right) - 4, & \mu\alpha^2 > 1, \end{cases}$$

and plotted the result in Fig. 4 further ahead.

We now need to translate the above result directly in terms of Y_n, instead of its logarithm. To do so, we need to find how fast is the convergence of $\left(\ln Y_n\right)/n$ to its limit. We use a result from [1, pp. 66 and 437] stating that

$$\limsup_{n\to\infty} \frac{\ln Y_n - \ln Y_0 - n\,\mathrm{E}\big(\ln u\big)}{n^{1/2}\left(\ln\ln n\right)^{1/2}} = \sqrt{2}\,\sigma_{\ln u} \quad \text{a.s.} \tag{5}$$

where $\sigma^2_{\ln u}$ is the variance of $\ln u$. This relation can be interpreted as follows. Denote by ω the experiment of choosing a sequence $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$. For each experiment ω, compute the sequence $Y_n(\omega)$ for all $n \ge 1$ (starting always from Y_0). Equation (5) implies that, with probability one (over the experiments ω), there exists a positive number $K(\omega)$ such that for all $n \ge K(\omega)$, $Y_n(\omega)$ satisfies

$$\ln Y_n(\omega) = n\,\mathrm{E}\big(\ln u\big) + \ln Y_0 + \delta(n),$$

where the error $\delta(n)$ satisfies

$$|\delta(n)| \le \sqrt{2}\,\sigma_{\ln u} n^{1/2}\left(\ln\ln n\right)^{1/2}.$$

We stress that $K(\omega)$ depends on the experiment ω.

Therefore, (4) implies that, with probability one, the curve $(n, Y_n(\omega))$ will eventually enter (and stay in) the set

$$\Theta = \left\{ (n,\, y(n)) :\ y(n) \le Y_0 e^{n\,\mathrm{E}\ln u} e^{\sqrt{2n\ln(\ln n)}\sigma_{\ln u}} \right\}. \tag{6}$$

Unfortunately, the convergence is not uniform, that is, there is *no* finite K_0 such that for almost all experiments, $\left(n,\, Y_n(\omega)\right) \in \Theta$ for $n \ge K_0$.

Since $\mathrm{E}\ln u$ does not depend on the time n, the first exponential in (6) dominates the second when n is large, which implies that

$$f(n) \triangleq e^{n\,\mathrm{E}\ln u} e^{\sigma_{\ln u}\sqrt{2n\ln(\ln n)}} \to 0$$

if, and only if, $\mathrm{E}\ln u < 0$.

Note also that for large n, when $\left(n, Y_n\right)$ is already close to or inside Θ, the rate of convergence of Y_n is dictated primarily by the term $e^{n\,\mathrm{E}\ln u}$. This implies that, for large n, the rate of convergence of $Y_n = \tilde{\boldsymbol{w}}_n^2$ is given primarily by $e^{\mathrm{E}\ln\left(1-\mu \mathrm{x}_n^2\right)^2}$. We conclude that Y_n converges to zero a.s. if, and only if, $\mathrm{E}\ln u < 0$. This leads to a different condition on $\mu\alpha^2$ than the one derived for mean-square stability. Moreover, the logarithm of the rate of convergence of Y_n is equal to $E\ln u$. This should be compared with $\ln Eu$, the logarithm of the rate of convergence of EY_n. One of the implications of this result is that Y_n converges to zero with probability one for $10/3 < \mu\alpha^2 < 6.1$, even though $\mathrm{E}\,Y_n$ diverges for $\mu\alpha^2$ in that range.

Refer now to Fig. 4 and compare the plots of $\mathrm{E}\ln u$ as a function of $\mu\alpha^2$ (the continuous line in the figure) and of $\ln\left(\mathrm{E}\,u\right)$ (which corresponds to the logarithm of the rate of convergence of $\mathrm{E}\,Y_n$ from (3)). Note that both plots are close together for small $\mu\alpha^2$, but they become significantly different as $\mu\alpha^2$ increases. In particular, they are quite different at the minima of each plot (which correspond to the fastest rates of convergence). In the ranges of $\mu\alpha^2$ for which the curves are significantly different, the rate of convergence of Y_n will be significantly different than the rate of convergence of $\mathrm{E}\,Y_n$ for large n.

With this result we can explain why the ensemble-average curves computed for small step-sizes are close to the "theoretical" predictions using $\mathrm{E}\,\tilde{\boldsymbol{w}}_n^2$, and why these plots are so different for large step-sizes.

For small step-sizes, the rates of convergence of both $\mathrm{E}\,\tilde{\boldsymbol{w}}_n^2$ and of $\tilde{\boldsymbol{w}}_n^2$ are, with probability one, very close, so we expect that an average of a few simulations will produce a reasonable approximation for $\mathrm{E}\,\tilde{\boldsymbol{w}}_n^2$.

For large $\mu\alpha^2$ and large n, with probability one the rate of convergence of $\tilde{\boldsymbol{w}}_n^2$ is significantly different (and faster) than (3), so we should expect to need a larger number of simulations to obtain a good approximation to $\mathrm{E}\,\tilde{\boldsymbol{w}}_n^2$.

Another interesting observation is that $\mathrm{E}\ln u$ is negative well beyond the point where $\ln\left(\mathrm{E}\,u\right)$ becomes positive. This implies that there is a range of step-sizes for which Y_n converges to zero with probability one, but $\mathrm{E}\,Y_n$ diverges. This explains the simulations in Fig. 3.

This is not a paradox. Since the convergence is not uniform, there is a small (but nonzero) probability that a sample curve Y_n will exist such that it assumes very large values

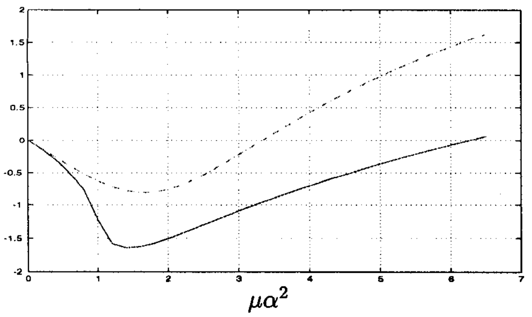

Figure 4. Graphs of $\mathrm{E}\ln(1-\mu x_{\mathrm{n}}^2)^2$ (continuous line) and $\ln\mathrm{E}(1-\mu x_{\mathrm{n}}^2)^2$ (broken line).

for a long interval of time before converging to zero. The following theorem has been proved.

Theorem 1. *Consider the scalar* LMS *algorithm with iid inputs* $\{x_n\}$*. Assume that the noise is identically zero and that* x_n *is a stationary random variable uniformly distributed between* $-\alpha$ *and* α*. Then, with probability one, there is a finite constant* K *(dependent on the realization) such that* $(n,\, \tilde{w}_n^2)$ *stays inside the set* Θ *defined above for all* $n \geq K$*. In particular,* $\tilde{w}_n^2$ *converges to zero with probability one if and only if* $\mathrm{E}\ln(1-\mu x_n^2)^2 < 0$.

3.3 Differences between theory and simulation

The above result can be used to understand the differences between theoretical and simulated learning curves for large step-sizes, as we now explain. Let $\{\omega_l\}_{l=1}^{L}$ be L independent experiments, with the corresponding $Y_n(\omega_l)$ and $K(\omega_l)$, and let $\hat{Y}_n = \frac{1}{L}\sum_{l=1}^{L} Y_n(\omega_l)$ be the ensemble-averaged learning curve. Since $(n, Y_n(\omega_l))$ stays inside Θ for $n \geq K(\omega_l)$, $(n, \hat{Y}_n)$ will also stay inside Θ for $n \geq \bar{K} = \sup K(\omega_l)$. This means that eventually (for large enough n), all ensemble-averaged learning curves will stay far from the average $\mathrm{E}\,Y_n$ (if $\mu\alpha^2$ is large). Nevertheless, the more simulations we average, the larger we expect $\bar{K}$ to be, so the difference between the ensemble-averaged curves and the true average will be significant only for increasingly large n.

Consider now the square error $e(n)^2 = x_n^2\tilde{w}_n^2$. Since x_n^2 is stationary and independent of $\tilde{w}_n^2$, the qualitative behavior of $e(n)^2$ is the same as that of $\tilde{w}_n^2$, that is, $e(n)^2$ converges when $\tilde{w}_n^2$ does, and the rates of convergence are the same.

4. Concluding remarks

We have shown that there are situations in which the behavior of the LMS errors are significantly different than that of their averages. These situations arise when one uses large step-sizes (to obtain faster convergence). Our simulations and analysis show that in some cases, it may be necessary to average a significantly large number of simulations to obtain a good approximation to the mean-square behavior of an adaptive filter. In particular, one must be careful when analyzing ensemble-average learning curves when the step-size is large.

Looking at the same results from another perspective, we might conclude that, for large step-size, the average performance may not be a good design parameter.

Although our analysis was performed only for the scalar LMS algorithm, one should expect to observe similar behavior in several other gradient-based algorithms and their variants, such as signed-LMS, leaky-LMS, CMA, etc. We are currently working on the extension of the analysis to the vector case and to other signal distributions.

Acknowledgment. The authors wish to thank Dr. P. Williams (Litton Data Systems) and Mrs. N. J. de Moraes Nascimento (Xontech) for helpful discussions.

References

[1] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, 2nd edition, 1996.

[2] A. Feuer and E. Weinstein. Convergence analysis of LMS filters with uncorrelated Gaussian data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(1):222–229, February 1985.

[3] S. Florian and A. Feuer. Performance analysis of the LMS algorithm with a tapped delay line (two-dimensional case). *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(6):1542–1549, December 1986.

[4] S. Haykin. Adaptive Filter Theory, 3rd edition, Prentice Hall, NJ, 1996.

[5] S. Jones, R. Cavin, and W. Reed. Analysis of error-gradient adaptive linear estimators for a class of stationary dependent processes. *IEEE Transactions on Information Theory*, IT-28:318–329, March 1982.

[6] M. Rupp. The behavior of LMS and NLMS algorithms in the presence of spherically invariant processes. *IEEE Transactions on Signal Processing*, SP-41:1149–1160, March 1993.

[7] D. Slock. On the convergence behavior of the LMS and the normalized LMS algorithms. *IEEE Transactions on Signal Processing*, 41(9):2811–2825, September 1993.

[8] M. Tarrab and A. Feuer. Convergence and performance analysis of the normalized LMS algorithm with uncorrelated Gaussian data. *IEEE Transactions on Information Theory*, 34(4):680–691, July 1988.

[9] S. Vembu, S. Verdú, R. Kennedy, and W. Sethares. Convex cost functions in blind equalization. *IEEE Transactions on Signal Processing*, 42(8):1952–1960, August 1994.

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