On the Learning Mechanism of Adaptive Filters

Vítor H. Nascimento and Ali H. Sayed, *Senior Member, IEEE*

***Abstract*—This paper highlights, both analytically and by simulations, some interesting phenomena regarding the behavior of ensemble-average learning curves of adaptive filters that may have gone unnoticed. Among other results, the paper shows that even ensemble-average learning curves of single-tap LMS filters actually exhibit two distinct rates of convergence: one for the initial time instants and another, *faster* one, for later time instants. In addition, such curves tend to converge faster than predicted by mean-square theory and can converge even when a mean-square stability analysis predicts divergence. These effects tend to be magnified by increasing the step size. Two of the conclusions that follow from this work are 1) mean-square stability alone may not be the most appropriate performance measure, especially for larger step sizes. A combination of mean-square stability and almost sure (a.s.) stability seems to be more appropriate. 2) Care is needed while interpreting ensemble-average curves for larger step sizes. The curves can lead to erroneous conclusions unless a large number of experiments are averaged (at times of the order of tens of thousands or higher).**

***Index Terms*—Adaptive filter, almost-sure convergence, Chebyshev's inequality, law of large numbers, learning curve, mean square convergence, rate of convergence.**

I. Introduction

ADAPTIVE filters are inherently nonlinear and time-variant devices that adjust themselves to an ever-changing environment; an adaptive system changes its parameters in such a way that its performance improves through a continuing interaction with its surroundings.

The learning curve of an adaptive filter provides a measure of how fast and how well it reacts to its environment. This learning process has been extensively studied in the literature for slowly adapting systems, that is, for systems that employ infinitesimally small step sizes. In this paper, we will discuss several interesting phenomena that characterize the learning capabilities of adaptive filters when larger step sizes are used. These phenomena actually occur even for slowly adapting systems but are less pronounced, which explains why they may have gone unnoticed.

Manuscript received December 2, 1998; revised November 15, 1999. This work was supported in part by the National Science Foundation under Grants MIP-9796147 and CCR-9732376. In addition, the work of V. H. Nascimento was further supported by a fellowship from Conselho Nacional de Pesquisa e Desenvolvimento—CNPq—Brazil, while on leave from Escola Politécnica da Universidade de São Paulo. The associate editor coordinating the review of this paper and approving it for publication was Dr. Hitoshi Kiya.

V. H. Nascimento was Adaptive Systems Laboratory, Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA. He is now with the Department of Electronic Systems Engineering, Universidade de São Paulo, São Paolo, Brazil (e-mail: vitor@lps.usp.br).

A. H. Sayed is with the Adaptive Systems Laboratory, Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: sayed@ee.ucla.edu).

Publisher Item Identifier S 1053-587X(00)04073-3.

The phenomena however become significantly more pronounced for larger step sizes (faster adaptation) and lead to several observations. In particular, we will show that after an initial phase, an adaptive filter generally learns at a rate that is better than that predicted by mean-square theory, that is, they seem to be "smarter" than we think. We will also show that even simple single-tap adaptive filters actually have two distinct rates of convergence; they learn at a slower rate initially and at a faster rate later. We will also argue that special care is needed in interpreting learning curves. Several examples will be provided.

A. Background and Objectives

As is well known, computable theoretical formulas for learning curves exist only for a few idealized situations. Ensemble-average learning curves (EALC's) are therefore commonly used to analyze and demonstrate the performance of adaptive filters; an EALC is obtained by averaging several error curves over repeated experiments or simulations and by plotting the resulting average curve.

EALC's have been used to extract, among other things, information about the rate of convergence of an adaptive filter, the value of its steady-state error, and choices of step sizes for faster convergence. Under certain independence conditions (see, for example, [1]–[6]), or for infinitesimally small step sizes (e.g., [7]–[19]), it is well understood that data extracted from such EALC's provide reasonably accurate information about the real performance of an adaptive filter.

But what about the performance of an adaptive scheme for larger step sizes and without the independence assumptions?[1] Will EALC's provide satisfactory information in these cases as well? In the process of comparing results obtained from EALC's with results predicted by an exact theoretical analysis for such scenarios (as in [20]–[22]), we noticed considerable differences between simulation and theory. These differences persisted no matter how many experiments we averaged.

A first explanation of the discrepancies was to blame the simulation program and possible numerical errors. After careful study, however, we realized that the differences have an analytical explanation and that they do occur for both small and large step sizes, although they are more pronounced in the latter case. Even more importantly, this led us to observe some other phenomena regarding the behavior of EALC's that may have gone unnoticed in the literature. More specifically, we will establish in this paper the following facts *both* by theory and by simulation.

1) Even ensemble-average learning curves (EALCs) of single-tap adaptive filters actually exhibit two distinct

[1]By larger step sizes, we do not mean step sizes that are necessarily large in value but, rather, step sizes that are not infinitesimally small.

rates of convergence: one for the initial time instants and another, *faster* one, for later time instants.
2) EALC's tend to converge faster than predicted by mean-square theory.
3) EALC's can and do converge even when a mean-square stability analysis predicts divergence.
4) The more experiments we average to construct an EALC, the more time it takes to observe the distinction between theory and simulation. Nevertheless, the difference always exists, and we need only simulate for a longer period (and possibly also reduce the noise level) to observe it.
5) Mean-square analysis alone may not be the most appropriate performance measure for larger step sizes. A *combination* of mean-square (MS) and almost sure (a.s.) stability results seems to be more appropriate.
6) Establishing that an adaptive filter is a.s. convergent (i.e., converges almost surely to zero in the noiseless case) does not necessarily guarantee satisfactory performance.
7) For filters with multiple taps, the behavior of EALC's may be dependent on the initial condition. This dependency does not exist for single tap filters (which are used, for example, in some frequency-domain implementations—see [6] and [23]).

The purpose of this paper is therefore to report on these phenomena and to provide an analytical explanation for their existence. In order to do so, we will resort to both MS and a.s. convergence analyses. In particular, we will show that (for a noiseless filter), MS analysis describes well the initial learning phase of an adaptive filter, whereas a.s. analysis describes well the latter phase of the learning process. We will also support the findings by several simulation results.

We may mention that there are already several works in the literature of adaptive filtering that relate to both kinds of analyses: MS and a.s. analyzses. As examples of MS-based studies, we may cite [7]–[9], [15], [17], and [19]–[21] and for a.s.-based studies [12], [14], [16], [18]. In some of these works (e.g., [14] and [16]), it is actually shown that both methods of analysis provide the same estimates for the rates of convergence of an adaptive filter when the step size is vanishingly small and that (at least for scalar systems) stability in the MS sense implies stability in the a.s. sense [12]. Still, the emphasis so far in the literature has been on the similarity of both approaches for vanishingly small step sizes. This paper focuses on the interesting phenomena and differences that arise when larger step sizes are used and on how to interpret these differences.

II. Simulations and Motivation

The purpose of the examples in this section is to demonstrate that for larger step sizes, there exists a noticeable difference between learning curves derived from MSE analysis and ensemble-average learning curves, even when the latter are constructed by averaging over a large number of repeated experiments. Later in the paper, we will show that this phenomenon has an analytical justification and that (for noiseless filters) it cannot be completely removed by indefinitely increasing the number of repeated experiments. We will also show analytically that this phenomenon disappears for infinitesimally small step sizes.

First, however, let us recall the definitions of learning curves and ensemble-average learning curves.

A. Learning Curves

Two important performance measures for adaptive filters are the mean-square error (MSE) and the mean-square deviation (MSD), which are defined by

$$\begin{aligned}\text{MSE curve} &\triangleq Ee(n)^2 = E\left(y(n) - \mathbf{x}_n^T \mathbf{w}_{n-1}\right)^2 \\ \text{MSD curve} &\triangleq E\|\mathbf{w}_* - \mathbf{w}_n\|^2\end{aligned} \tag{1}$$

where

$\{y(n)\}_{n=1}^{\infty}$ desired sequence;
$\{\mathbf{x}_n\}_{n=1}^{\infty}$ input (regressor) sequence;
$\mathbf{w}_{n-1}$ weight estimate at time $n-1$.

The signal $y(n)$ is further assumed to be generated via the linear model $y(n) = \mathbf{x}_n^T \mathbf{w}_* + v(n)$ for some unknown length-M vector $\mathbf{w}_*$ that is to be estimated. The sequence $\{v(n)\}$ denotes measurement noise. In this paper, we assume that $\mathbf{w}_*$ is constant and that both the input and desired sequences $\{y(n), \mathbf{x}_n\}$ are stationary. We also define the weight error vector $\tilde{\mathbf{w}}_n = \mathbf{w}_* - \mathbf{w}_n$.

The LMS algorithm updates the weight estimates $\{\mathbf{w}_n\}$ by means of the recursion [2], [6]

$$\mathbf{w}_n = \mathbf{w}_{n-1} + \mu \mathbf{x}_n e(n) \tag{2}$$

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

The plot of the MSE as a function of the time instant 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 an analytic expression for the learning curve or for the steady-state MSE, except when the assumptions of *independence theory* [1]–[6] are used. In this framework, we assume the following:

i) The input sequence $\{\mathbf{x}_n\}$ is independent.
ii) The noise sequence $\{v(n)\}$ is independent.
iii) The noise sequence is independent of the input sequence.

Despite the fact that these assumptions are seldom satisfied in practice, it is known that the learning curves that are obtained using the independence theory are good approximations for the true learning curves when the step size μ is vanishingly small (see, e.g., [7], [8], and [15]). However, what about the case of larger step sizes?

There have been studies on the evaluation of the exact learning curve in this case. For example, if the correlation of the sequence $\{\mathbf{x}_n\}$ is zero for large lags (i.e., if there exists a finite N_c such that the correlation of $\mathbf{x}_n$ and $\mathbf{x}_k$ is zero when $|k-n| > N_c$), [20] and [21] discuss a method to analytically evaluate the learning curve of LMS filters that is exact for any step size $\mu > 0$. Unfortunately, this method is computationally feasible only for small filter lengths (at most $M = 6$ or 7, depending on the correlation of the regressor sequence) since its complexity grows extremely fast with the filter length. The method also requires a detailed knowledge of the statistics of

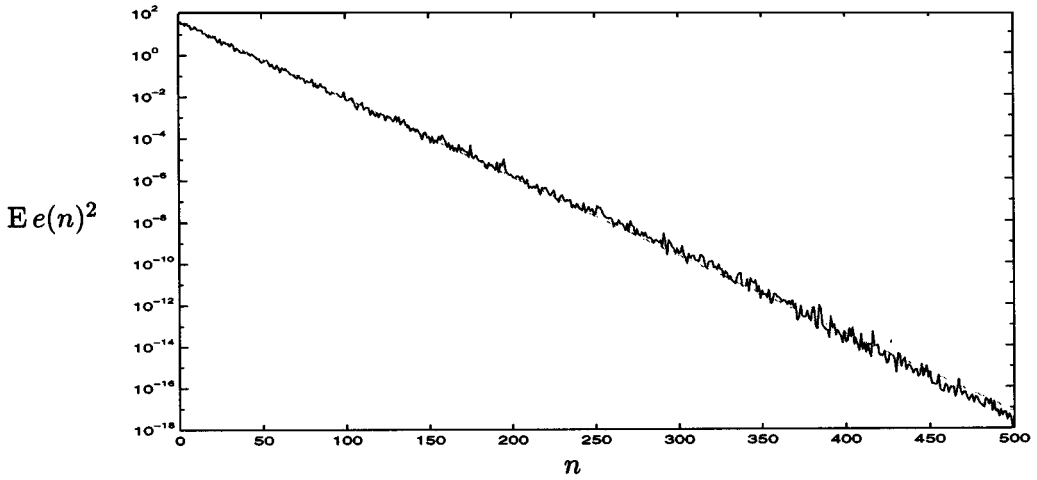

Fig. 1. Learning curves computed by simulation and theoretically with Gaussian iid inputs, $M = 10$, $\mu = 0.08$, and $L = 100$.

the regressor sequence $\{\mathbf{x}_n\}$, which is usually not available. (We may remark that [22] studies the case of large filter lengths and derives a computable lower bound on how large the step size can be for stable mean-square performance.)

B. Ensemble-Average Learning Curves

Therefore, for many situations of relevance, there are no practical ways to evaluate the exact learning curve of an adaptive filter. For this reason, it is common practice to estimate the learning curve by experimentation or by repeated simulations. More specifically, we perform several independent experiments (or simulations), 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 $\{\mathbf{x}_n\}$. From each experiment i, a sample curve $\{e^{(i)}(n), 1 \leq n \leq N\}$ is obtained. After all L experiments are completed, an approximation for the learning curve is computed by averaging as follows:

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

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

Although it is less common, we can also plot the MSD versus time. We will normally refer to this plot also as "the learning curve," or as the MSD *learning curve*, if we need to distinguish between the plots for the MSE and for the MSD. The MSD EALC is

$$E\|\tilde{\mathbf{w}}_n\|^2 \approx \hat{D}(n) = \frac{1}{L}\sum_{i=1}^{L} \left\|\tilde{\mathbf{w}}_n^{(i)}\right\|^2, \quad 1 \leq n \leq N.$$

For vanishingly small step sizes, it is known [7], [8], [15] that an average of a few tens of experiments is enough to obtain experimental learning curves $\hat{E}(n)$ that are good approximations for the actual learning curve (we will also give an analytical justification for this fact later in Theorem 2). It is thus common in the literature to use the average of a few independent repeated experiments to predict or confirm theoretical results by means of simulations (a few relatively recent examples include [3]and [25], which use 10–20 independent experiments, and [24], [27], and [28], which use 100 independent experiments).

However, what about larger step sizes? Will EALC's still provide good approximations for the actual learning curves?

C. Examples

Consider a length $M = 10$ LMS adaptive filter operating with Gaussian inputs with covariance matrix $E\mathbf{x}_n\mathbf{x}_n^T = I$, step size $\mu = 0.08$, and no noise. The learning curve for this case was computed theoretically in [24]. In Fig. 1, we plot the theoretical curve over the first 500 instants of time, in addition to an EALC that is obtained from the average of $L = 100$ simulations. Note how both plots are close to each other.

Consider now the same LMS filter of length $M = 10$ but with the larger step size $\mu = 0.16$ (and noise variance $\sigma_v^2 = 10^{-4}$). Since the independence assumptions are satisfied in this case, it is possible to compute the learning curve $Ee(n)^2$ exactly, as follows [3], [4]. Let the input covariance matrix be $R \triangleq E\mathbf{x}_n\mathbf{x}_n^T$. Under the conditions of the example, $R = I$. Define further the weight error covariance matrix $C_n \triangleq E\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T$. With these definitions, the MSD and MSE are given by $E\|\tilde{\mathbf{w}}_n\|^2 = \mathrm{Tr}\, C_n$ and $Ee(n)^2 = \mathrm{Tr}(RC_n) = \mathrm{Tr}(C_n)$, where $\mathrm{Tr}(\cdot)$ stands for the trace of a matrix. That is, for this example, both the MSD and the MSE depend only on the diagonal entries of C_n. Moreover, it is possible to find recursions for these diagonal entries that do not depend on the off-diagonal entries in the following way. Define the vectors

$$\mathbf{\Gamma}_n \triangleq [(C_n)_{1,1} \quad (C_n)_{2,2} \quad \cdots \quad (C_n)_{M,M}]^T$$
$$\mathbf{L} \triangleq [1 \quad 1 \quad \cdots \quad 1]^T$$

and the matrix $A \triangleq I - 2\mu R + 2\mu^2 R^2 + \mu^2\mathbf{L}\mathbf{L}^T$. Then, the diagonal entries of C_n satisfy the recursion

$$\mathbf{\Gamma}_n = A\mathbf{\Gamma}_{n-1} + \mu^2\sigma_v^2\mathbf{L} \tag{3}$$

with initial condition

$$\mathbf{\Gamma}_0 = [(\tilde{w}_{0,1})^2 \quad \cdots \quad (\tilde{w}_{0,M})^2]^T$$

where $\tilde{w}_{0,i}$ is the ith entry of the error vector $\tilde{\mathbf{w}}_0$, which is assumed deterministic.

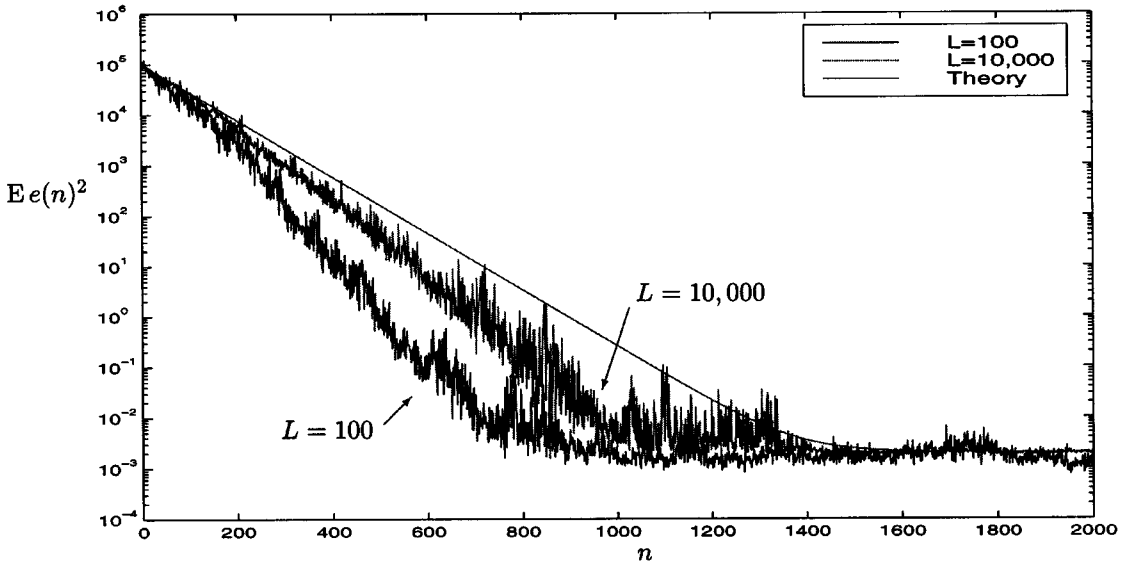

Fig. 2. Learning curves computed by simulation and theoretically with Gaussian independent input vectors, Gaussian noise with $\sigma_v^2 = 10^{-4}$, $M = 10, \mu = 0.16$, $L = 100$, and $L = 10^4$.

In Fig. 2, we plot the resulting theoretical learning curve $Ee(n)^2$ as well as EALC's computed over $L = 100$ and $L = 10\,000$ simulations, all with step size $\mu = 0.16$ (which is twice the value of the step size used to generate Fig. 1).

Note how the simulation curves are now noticeably far (and most of the time) below the (smooth) theoretical curve, although the simulations get closer (but not close enough) to the theoretical curve as L is increased from 100 to 10 000. Note also that the simulation curves converge faster than the theoretical curve. This situation should be compared with Fig. 1, where an almost-perfect agreement was obtained between theory and simulation (in fact, had we computed more iterations in Fig. 1, a distinction between the theoretical curve and the simulations would have been observed as well, just as in Fig. 2).

When the independence assumptions do not hold, these effects still occur, as we show next. Assume now that the input vectors $\{\mathbf{x}_n\}$ are not iid as above but have a delay-line structure of the form $\mathbf{x}_n = \mathrm{col}\{a(n), a(n-1), \ldots, a(n-M+1)\}$. The results of [20] and [21] can be used to obtain, analytically, the learning curve $Ee(n)^2$ for small values of M. In Fig. 3, we plot this theoretical curve, as well as ensemble-average curves for $L = 100$ to $L = 10\,000$, with filter length $M = 2$, step size $\mu = 8.3$, and, for $a(n)$ iid, uniformly distributed between -0.5 and 0.5. With this value of μ, the actual learning curve $Ee(n)^2$ can be shown to diverge (and we observe in the figure that it indeed diverges). However, the simulations show the ensemble-average curves $\hat{E}(n)$ converging [see Fig. 3(a)] for various values of L. Notice further that for increasing L, the ensemble-average curve stays closer to the actual learning curve for a longer period of time toward the beginning of the simulation—although the curves ultimately separate afterwards, with the theoretical curve diverging and the ensemble-average curves converging (no matter how large L is). We will explain this fact analytically (for iid sequences) in Section III-F.

The simulations presented so far show that the behavior of the ensemble-average curves may be significantly different than that of the theoretical learning curves (e.g., convergence can occur even when divergence by MSE analysis is expected; faster convergence can occur even when slower rates are predicted by MSE analysis; the averaged curves and the theoretical curve essentially coincide during the initial training phase but separate thereafter, which indicates the existence of two distinct rates of convergence: a slow one initially and a faster one later). These differences can lead to erroneous conclusions when we attempt to predict performance from simulation results. One interesting fact to stress is that these differences may occur even for very large L.

In the next section, we explain the origin of these effects by focusing first on the scalar LMS case, which serves as a good demonstration and helps highlight the main ideas. In a later section, we extend the analysis to the vector case (see Section IV). The reason why we distinguish between the scalar and the vector cases is that more can be said in the former case, and the analysis methods are also more explicit (and in closed form).

III. 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 $\{\mathbf{x}_n\}$ and zero noise. Thus, assuming $M = 1$, we obtain a single-tap adaptive filter with update equation of the form

$$\begin{aligned} \mathbf{w}_n &= \mathbf{w}_{n-1} + \mu \mathbf{x}_n e(n), \quad y(n) = \mathbf{x}_n \mathbf{w}_* \\ e(n) &= y(n) - \mathbf{x}_n \mathbf{w}_{n-1} \end{aligned} \tag{4}$$

where all variables $\{\mathbf{w}_n, \mathbf{x}_n, e(n)\}$ are now scalar valued. Recall that the weight error vector is denoted by $\tilde{\mathbf{w}}_n = \mathbf{w}_* - \mathbf{w}_n$, which therefore satisfies the recursion $\tilde{\mathbf{w}}_n = (1 - \mu \mathbf{x}_n^2)\tilde{\mathbf{w}}_{n-1}$.

A. Condition for Mean-Square Stability

We first determine conditions on the step size μ for the above one-tap filter to be mean-square stable, i.e., for the variance of

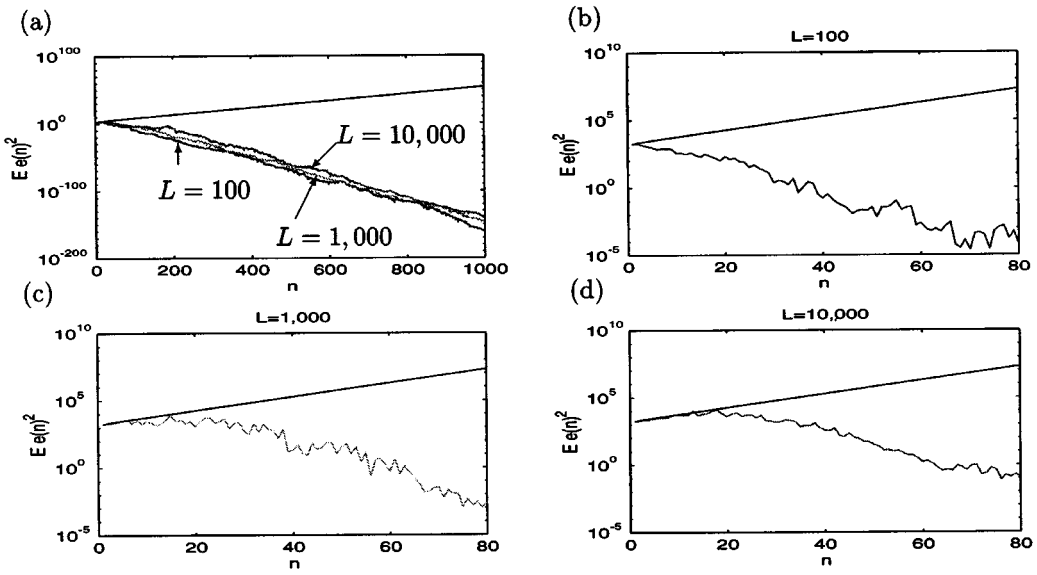

Fig. 3. Learning curves computed by simulation and theoretically with tap-delayed input vectors, $M = 2$, $\mu = 8.3$, and (a) $L = 100$, $L = 1\ 000$, and $L = 10\ 000$. (b) Theoretical curve and $L = 100$ only. (c) Theoretical and $L = 1000$ only. (d) Theoretical and $L = 10\ 000$ only.

$\tilde{\mathbf{w}}_n, E\tilde{\mathbf{w}}_n^2$, to converge to zero. This is a standard step in mean-square stability analysis.

We start by squaring both sides of the above LMS error equation to obtain

$$\tilde{\mathbf{w}}_n^2 = \left(1 - \mu \mathbf{x}_n^2\right)^2 \tilde{\mathbf{w}}_{n-1}^2. \tag{5}$$

This is a stochastic difference equation relating two positive quantities $\tilde{\mathbf{w}}_n^2$ and $\tilde{\mathbf{w}}_{n-1}^2$. The relation between both quantities is a random multiplicative factor, which we denote by $u(n) \triangleq (1-\mu\mathbf{x}_n^2)^2$. Note that from our assumptions on $\{\mathbf{x}_n\}$, it follows that the $u(n)$ are iid. To simplify the notation further, we also denote $Y_n \triangleq \tilde{\mathbf{w}}_n^2$. In our simplified notation, the recursion (5) becomes

$$Y_n = u(n)Y_{n-1} = Y_0 u(1)u(2)\ldots u(n) \tag{6}$$

where the initial condition $Y_0 = \tilde{\mathbf{w}}_0^2$ is assumed deterministic.

As mentioned above, we want to determine conditions under which EY_n converges to zero. For this purpose, we denote the variance and the fourth-order moment of the regressor $\mathbf{x}_n$ by $\sigma_2 \triangleq E\mathbf{x}_n^2$, and $\sigma_4 \triangleq E\mathbf{x}_n^4$. From (6), and using the independence of the $\{u(i)\}$, we then obtain

$$EY_n = (Eu)^n Y_0 = [1 - 2\mu\sigma_2 + \mu^2\sigma_4]^n Y_0 \tag{7}$$

where we are denoting the iid non-negative variables $u(i)$ generically by u (their expected value is equal to $Eu = 1 - 2\mu\sigma_2 + \mu^2\sigma_4$). From the above equation, we conclude that EY_n will converge to 0 if, and only if, μ is such that the mean of u is strictly less than 1. We will refer to the resulting rate of convergence of EY_n, viz., $Eu = 1 - 2\mu\sigma_2 + \mu^2\sigma_4$, as the *MS rate of convergence*. For ease of comparison with a later condition [see (12)], we will rewrite the requirement $Eu < 1$ in the equivalent form (in terms of the natural logarithm)

$$\ln(Eu) < 0. \tag{8}$$

Observe further that the logarithm of the MS rate of convergence is equal to $\ln(Eu)$ [which is a result that we will also invoke later—in the discussion following (12)]. We summarize the above conclusions in the following statement.

Lemma 1 (Mean-Square Stability): Consider the scalar LMS algorithm (4) with stationary iid inputs $\{\mathbf{x}_n\}$. Assume also that the noise is identically zero. Then, $E\tilde{\mathbf{w}}_n^2$ tends to zero if, and only if, the step size μ is such that $\ln(Eu) < 0$. □

Consider now the square error $e(n)^2 = \mathbf{x}_n^2\tilde{\mathbf{w}}_n^2$. Since $\mathbf{x}_n^2$ is stationary and independent of $\tilde{\mathbf{w}}_n^2$, $Ee(n)^2 = \sigma_2 E\tilde{\mathbf{w}}_n^2$. This implies that the behavior of $Ee(n)^2$ is the same as that of $E\tilde{\mathbf{w}}_n^2$, i.e., $Ee(n)^2$ will converge when $E\tilde{\mathbf{w}}_n^2$ does, and the rates of convergence will be the same.

B. Behavior of a Sample Curve

Our experiments in Section II showed that there is a clear distinction between the plots of EY_n and of the ensemble-average curve $(1/L)\sum_{i=1}^{L} Y_n^{(i)}$. We explain this fact in this section.

We focus first on the behavior of a typical (single) curve Y_n and show that for large n, Y_n decays (or increases) at a rate significantly different than that of EY_n. We obtain this result by studying conditions under which a typical curve Y_n converges to zero with probability one (or almost surely). Later, we will show how this effect manifests itself when several such curves are averaged together to yield an ensemble-average curve.

We start by computing the logarithm of Y_n in (6)

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

which shows that the difference $\ln Y_n - \ln Y_0$ is equal to the sum of n iid random variables $\{\ln u(i)\}$. We assume for now that the variance of $\ln u(i)$ is bounded (Theorem 2 gives conditions for this to hold). Therefore, we can use the strong law of large numbers [29] to conclude that

$$\frac{\ln Y_n}{n} \overset{\text{a.s.}}{\to} E(\ln u(i)) \triangleq E(\ln u) \tag{9}$$

where a.s. denotes almost-sure convergence. That is, for large n, $(\ln Y_n)/n$ will almost surely converge to a constant $E\ln u$.

[We will evaluate $E(\ln u)$ for different distributions in Section III-D.] We now need to translate the above result directly in terms of Y_n, instead of its logarithm. To do so, we must find how fast the convergence of $(\ln Y_n)/n$ is to its limit. We use a result from [29, pp. 66 and 437], stating that

$$\limsup_{n\to\infty}\left(\frac{\ln Y_n - \ln Y_0 - nE(\ln u)}{n^{1/2}(\ln\ln n)^{1/2}}\right) = \sqrt{2}\sigma_{\ln u} \quad \text{a.s.} \quad (10)$$

where $\sigma^2_{\ln u}$ denotes the variance of $\ln u$, which we assumed to be finite (see Theorem 2). Relation (10) can be interpreted as follows. Denote by ω the experiment of choosing a regressor sequence $\{\mathbf{x}_n\}_{n=1}^{\infty}$. For each experiment ω, compute the resulting sequence $Y_n(\omega)$ for all $n \geq 1$ (starting always from the same initial condition Y_0). Then, the statement (9) that "$(\ln Y_n)/n$ converges to $E\ln(u)$ a.s." means that the set of experiments

$$\mathcal{Z} = \left\{\omega \text{ such that } \frac{\ln Y_n(\omega)}{n} \to E\ln(u)\right\}$$

has probability 1. Moreover, (10) implies that with probability one, there exists for each experiment a finite positive number $K(\omega)$ (dependent on the experiment) such that for all $n \geq K(\omega)$, the corresponding curve $Y_n(\omega)$ satisfies

$$\ln Y_n(\omega) = nE(\ln u) + \ln Y_0 + \delta(n)$$

where the error $\delta(n)$ satisfies

$$|\delta(n)| \leq \sqrt{2}\sigma_{\ln u} n^{1/2}(\ln\ln n)^{1/2}.$$

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

Therefore, (9) and (10) imply that with probability one, a typical curve $(n, Y_n(\omega))$ will eventually enter and stay inside the set

$$\Theta = \left\{(n, y(n)) :\ y(n) \leq Y_0 e^{nE\ln u} e^{\sqrt{2n\ln(\ln n)}\sigma_{\ln u}}\right\}. \quad (11)$$

In other words, for n large enough, a typical curve $Y_n(\omega)$ will be upper bounded by the curve $Y_0 e^{nE\ln u} e^{\sqrt{2n\ln(\ln n)}\sigma_{\ln u}}$. The convergence of $Y_n(\omega)$ to the above set, however, is not uniform. That is, there is *no* finite K_0 such that for almost all experiments, $(n, Y_n(\omega)) \in \Theta$ for $n \geq K_0$.

Now, since $E\ln u$ does not depend on the time n, the first exponential in (11) dominates the second when n is large, which implies that the upper bound $Y_0 e^{nE\ln u} e^{\sqrt{2n\ln(\ln n)}\sigma_{\ln u}}$ tends to zero if, and only if, $E(\ln u) < 0$. We thus conclude that a typical curve Y_n converges to zero a.s. (or with probability one) if, and only if, the step size μ is such that

$$E\ln u < 0. \quad (12)$$

This leads to a different condition on μ than the one derived for mean-square stability in (8). In addition, note that for large n, when (n, Y_n) is already close to or inside Θ, the rate of convergence of a typical curve Y_n is dictated primarily by the term $e^{nE\ln u}$. This implies that for large n, the logarithm of the rate of convergence of $Y_n = \tilde{\mathbf{w}}_n^2$ is given by $E\ln u$ [which should be contrasted with $\ln Eu$ in the mean-square analysis case right after (8)]. We will refer to the above rate as the *a.s. rate of convergence*. The following theorem has thus been proven.

Theorem 1—A.S. Convergence: Consider the scalar LMS algorithm (4) with stationary iid inputs $\{\mathbf{x}_n\}$. In addition, assume that the noise is identically zero. Then, with probability one, there is a finite constant K (dependent on the realization) such that $(n, \tilde{\mathbf{w}}_n^2)$ stays inside the set Θ defined above for all $n \geq K$. In particular, a typical curve $\tilde{\mathbf{w}}_n^2$ converges to zero with probability one if, and only if, $E\ln u < 0$ (which is equivalent to $E\ln(1-\mu\mathbf{x}_n^2)^2 < 0$). □

We may remark that there are related works in the literature that have studied the a.s. stability of LMS (e.g., [12], [14], and [18]) or even of continuous-time systems (e.g., [10] and [11]). These works, however, do not emphasize the *distinctions* that arise between MS stability and a.s. stability, nor do they note the implications of these differences on the behavior of the EALC's of adaptive filters. Reference [14] obtains condition (12) but compares the a.s. and MS notions of stability only for $\mu \approx 0$ when they in fact agree, thus proving by means of different tools a version of Theorem 2 further ahead for ergodic signals (and for considerably larger values of β). Therefore, while the emphasis so far in the literature has been mainly on the similarity of the a.s. and MS methods of analyses for vanishingly small step sizes, we will instead focus on the phenomena and differences that arise when larger step sizes are used.

C. Comparisons

Comparing the statements of Lemma 1 and Theorem 1, we see that there is a fundamental difference in the conditions required for convergence in both cases. The lemma shows that MS convergence requires the step size μ to be such that $\ln Eu < 0$, whereas the theorem shows that a.s. convergence requires μ to be such that $E\ln u < 0$. The two conditions are not equivalent, and in fact, one implies the other since, for any non-negative random variable u for which Eu and $E\ln u$ both exist, it holds that $E(\ln u) \leq \ln(Eu)$. (This result follows directly from Jensen's inequality since the function $(-\ln x)$ is convex; see, e.g., [29, p. 14]). Therefore, values of μ for which MS convergence occurs always guarantee a.s. convergence while the converse is not true. A value for which $\ln Eu > 0$ (and, thus, MS divergence occurs) can still guarantee a.s. convergence, or $E\ln u < 0$, which explains the phenomenon in Fig. 3.

We will elaborate more on these distinctions in the sequel and explain how they can be used to explain the phenomena that we observed in the simulations in Section II. For now, however, we show that these distinctions disappear for infinitesimally small step sizes (which is a fact that does not depend on a specific input signal distribution). Although, as mentioned above, a version of this result for infinitesimally small step sizes is presented in [14] for ergodic signals and for larger values of β, we provide in Appendix A an independent argument under different assumptions by using a sequence of integration results. (We remark that the requirement on the probability density function in the statement of the theorem below is not restrictive, and it does not rule out most well-known distributions.)

Theorem 2—Rates of Convergence for Small μ: Let $p(x)$ denote the probability density function of the iid regressor sequence $\{\mathbf{x}_n\}$. Assume there exist constants $B < \infty$ and $\beta > 6$

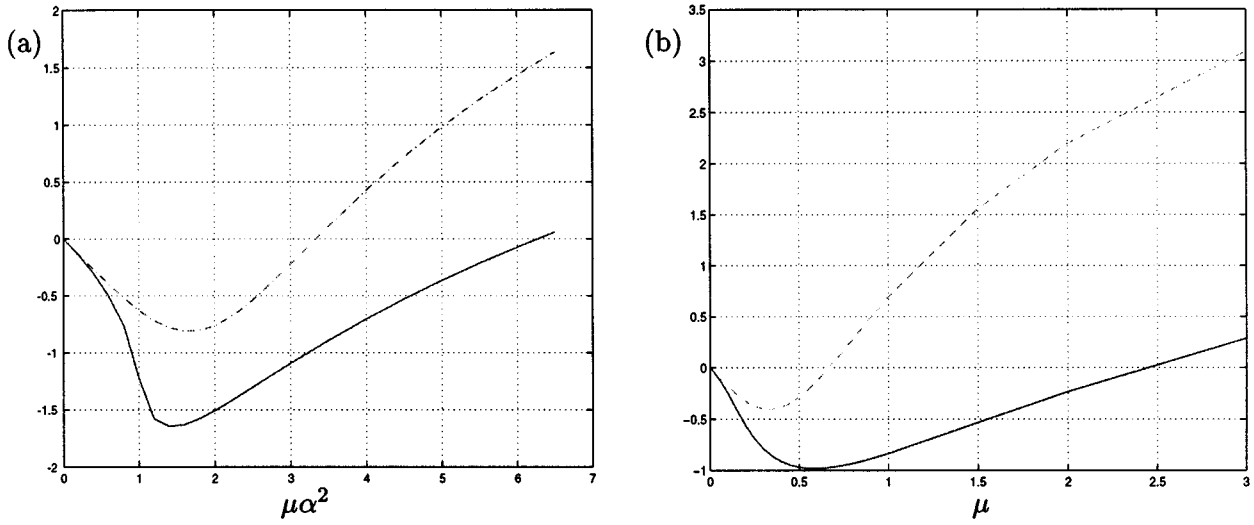

Fig. 4. Graphs of $E\ln(1-\mu\mathbf{x}_n^2)^2$ (continuous line) and $\ln E(1-\mu\mathbf{x}_n^2)^2$ (broken line). (a) Uniformly distributed $\mathbf{x}_n$. (b) Gaussian $\mathbf{x}_n$.

such that $p(x)$ satisfies $p(x) \leq 1/x^\beta$ for $|x| \geq B$. Then, the quantities $E\ln(1-\mu\mathbf{x}_n^2)^2$, var$(\ln(1-\mu\mathbf{x}_n^2)^2)$, and $\ln E(1-\mu\mathbf{x}_n^2)^2$ exist, are finite, and satisfy

$$E\ln\left(1-\mu\mathbf{x}_n^2\right)^2 = -2\mu\sigma_2 + o(\mu)$$
$$\ln E\left(1-\mu\mathbf{x}_n^2\right)^2 = -2\mu\sigma_2 + o(\mu)$$

where $o(\mu)$ is a function satisfying $\lim_{\mu\to 0} o(\mu)/\mu = 0$. □

The theorem therefore shows that $E\ln u$ and $\ln Eu$ are approximately the same when μ is infinitesimally small. This explains why learning curves and EALC's tend to agree reasonably well for such small step sizes. However, we are particularly interested in explaining the discrepancies that arise when larger step sizes are used.

D. Some Examples

We now provide two examples showing that for larger step sizes, the difference between $E\ln u$ and $\ln Eu$ can be considerably large. In particular, the difference can be large around the step size that achieves fastest convergence.

Assume first that $\mathbf{x}_n$ is a uniform random variable with values in the interval $[-\alpha,\alpha]$. For this input distribution, we have $\sigma_2 = \alpha^2/3$ and $\sigma_4 = \alpha^4/5$ so that

$$\ln Eu = \ln E\left(1-\mu\mathbf{x}_n^2\right)^2 = \ln\left(1-2\mu\frac{\alpha^2}{3}+\mu^2\frac{\alpha^4}{5}\right).$$

We can also evaluate $E\ln u$ as a function of $\mu\alpha^2$ explicitly and obtain the equation shown at the bottom of the page.

Fig. 4(a) compares the plots of $E\ln u$ (the continuous line in the figure) and of $\ln(Eu)$. Note that both plots are close together for small $\mu\alpha^2$ (as predicted by Theorem 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 from the MS and a.s. convergence points of view). In the ranges of $\mu\alpha^2$ for which the curves are significantly different, the rate of convergence of a typical curve Y_n will be significantly different than the rate of convergence of EY_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 $E\tilde{\mathbf{w}}_n^2$ and why these plots are so different for larger step sizes. For sufficiently small step sizes, the rates of convergence of both $E\tilde{\mathbf{w}}_n^2$ and $\tilde{\mathbf{w}}_n^2$ are, with probability one, very close; therefore, we expect that an average of a few simulations will produce a reasonable approximation for $E\tilde{\mathbf{w}}_n^2$. For larger step sizes, and for large n, however, the rate of convergence of $\tilde{\mathbf{w}}_n^2$ is significantly different (and faster) than that predicted by (7). Thus, we should expect to need a larger number of simulations to obtain a good approximation for $E\tilde{\mathbf{w}}_n^2$. This latter point will be better clarified by the variance analysis that we provide in Section III-F.

Another interesting observation is that $E\ln u$ is negative well beyond the point where $\ln(Eu)$ becomes positive. This implies that there is a range of step sizes for which a typical curve Y_n converges to zero with probability one but EY_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

$$E\ln u = \begin{cases} \ln(1-\mu\alpha^2)^2 + \dfrac{4}{\alpha\sqrt{\mu}}\operatorname{arctanh}(\alpha\sqrt{\mu}) - 4, & \text{if } \mu\alpha^2 \leq 1 \\ \ln(1-\mu\alpha^2)^2 + \dfrac{4}{\alpha\sqrt{\mu}}\operatorname{arccoth}(\alpha\sqrt{\mu}) - 4, & \text{if } \mu\alpha^2 > 1. \end{cases}$$

very large values for a long interval of time before converging to zero.

Assume now that $\mathbf{x}_n$ is Gaussian with zero mean and unit variance so that $\sigma_2 = 1$, and $\sigma_4 = 3$. Then, $EY_n = (1 - 2\mu + 3\mu^2)^n Y_0$. We computed $E\ln(1-\mu\mathbf{x}_n^2)^2$ numerically (using the symbolic toolbox in Matlab[2]), obtaining the results shown in Fig. 4(b). Note, again, how $\ln Eu$ and $E\ln u$ are approximately equal for small μ. An interesting fact that appears here is that the value of μ that achieves fastest MS convergence is noticeably smaller than the step size that achieves fastest a.s. convergence. In distributions with heavier tails (such as the exponential), this difference is even more pronounced.

E. Differences Between Theory and Simulation

The above results can thus be used to understand the differences between theoretical and simulated learning curves for large n and for larger step sizes. Indeed, let $\{\omega_l\}_{l=1}^L$ be L independent experiments with the corresponding sample curves $\{Y_n(\omega_l)\}$. We know from Theorem 1 that for each curve, there exists an integer $K(\omega_l)$ such that $Y_n(\omega_l)$ will remain inside the set Θ for all $n \geq K(\omega_l)$. In particular, if the step size is such that $E\ln(1-\mu\mathbf{x}_n^2)^2 < 0$, then this means that with probability one, $Y_n(\omega_l)$ will be converging to zero for all $n \geq K(\omega_l)$.

Now, let $\hat{Y}_n = (1/L)\sum_{l=1}^L Y_n(\omega_l)$ be the EALC. 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 $n \geq \bar{K} = \max_l K(\omega_l)$. This means that eventually (for large enough n), all EALC's will stay far away from the average curve EY_n. This is because the actual average curve EY_n and typical sample curves Y_n will converge at different rates for large n (one rate of convergence is dictated by $\ln Eu$, whereas the other is dictated by $E\ln u$).

Thus, we can say that the a.s. analysis allows us to clarify what happens when we fix L (the number of repeated experiments) and increase n (the time variable). The ensemble-average curve tends to separate from the true average curve for increasing n due to the difference in the convergence rates.

On the other hand, the more simulations we average, the larger we expect $\bar{K}$ to be; therefore, the difference between the ensemble-average curve and the true average will be significant only for increasingly large n. That is, the more we average, the longer it takes for us to see the difference between the ensemble-average curve and the true-average curve. We will explain this fact more clearly in Section III-G by means of a variance analysis. We summarize these conclusions in the following statement for ease of reference.

Lemma 2—A.S. Analysis: Consider the scalar LMS algorithm (4) with stationary iid inputs $\{\mathbf{x}_n\}$. Assume that the noise is identically zero and that the distribution of u is such that $E\ln u < \ln Eu$ (i.e., strict inequality holds). Then, the following conclusions hold.

1) If we fix L, then for large enough n, the ensemble-average curve will be noticeably different from the true-average curve due to different rates of convergence.
2) The more we average (i.e., the larger the value of L), the longer it takes for the difference between the ensemble-average curve and the true-average curve to be noticed.

[2]Matlab is a trademark of The MathWorks, Inc.

F. Variance Analysis

The a.s. convergence analysis of the previous sections establishes that for large enough time n, there always exists a difference between the ensemble-average curve and the true-average curve and that this difference is explained by the fact that the convergence rates of both curves are distinct. In view of this, we will say that the a.s. analysis helps us explain the distinction between both curves for large time instants n.

If we, however, re-examine the curves of Figs. 2 and 3, we see that for small n (that is, close to the beginning of the curves), there is usually a good match between the learning curve and the ensemble-average curve. Put another way, we notice that the rates of convergence of the true learning curve and the ensemble-average curve tend to be identical for these initial time instants. Only for later time instants, do the rates of convergence become different as predicted by the a.s. analysis.

To explain this initial effect, we rely on a different argument that employs Chebyshev's inequality. We start by evaluating the variance of Y_n, $\text{var}\, Y_n$, rather than its mean (as in Section III-A). This is because we will now study the evolution of the following ratio:

$$\rho(n) \triangleq \frac{\sqrt{\text{var}\, Y_n}}{EY_n} \tag{13}$$

which we stress is a function of the time instant n, that is, with each n, we associate a value $\rho(n)$. We claim that for values of n for which $\rho(n) \ll 1$, the average value EY_n will be a good approximation for the values of the sample curve Y_n at these time instants. To see this, assume that $\rho(n) = 0.05$ for a particular value of n. Using Chebyshev's inequality [29, p. 15], we obtain

$$P\left\{|Y_n - EY_n| \geq \frac{1}{2}EY_n\right\} \leq \frac{\rho(n)^2(EY_n)^2}{0.25(EY_n)^2} = 0.01.$$

This means that we have a 99% probability that Y_n will be in the interval $[0.5EY_n, 1.5EY_n]$.

Now, when we form EALC's, we average several sample curves Y_n to obtain

$$\hat{D}(n) = \frac{1}{L}\sum_{i=1}^{L} Y_n^{(i)}.$$

Assuming that we use L independent experiments, then the expected value of $\hat{D}(n)$ is still equal to EY_n. The ratio $\rho(n)$ that is associated with the curves $\{EY_n, \hat{D}(n)\}$ will then be given by

$$\rho'(n) = \frac{\sqrt{\text{var}\, \hat{D}(n)}}{EY_n}$$

where the variance of $\hat{D}(n)$ is equal to $(\text{var}\, Y_n)/L$. This implies that

$$\rho'(n) = \frac{\sqrt{\text{var}\, \hat{D}(n)}}{EY_n} = \frac{\rho(n)}{\sqrt{L}}. \tag{14}$$

That is, the process of constructing EALC's reduces the value of $\rho(n)$. Therefore, if we choose L large enough, the EALC should

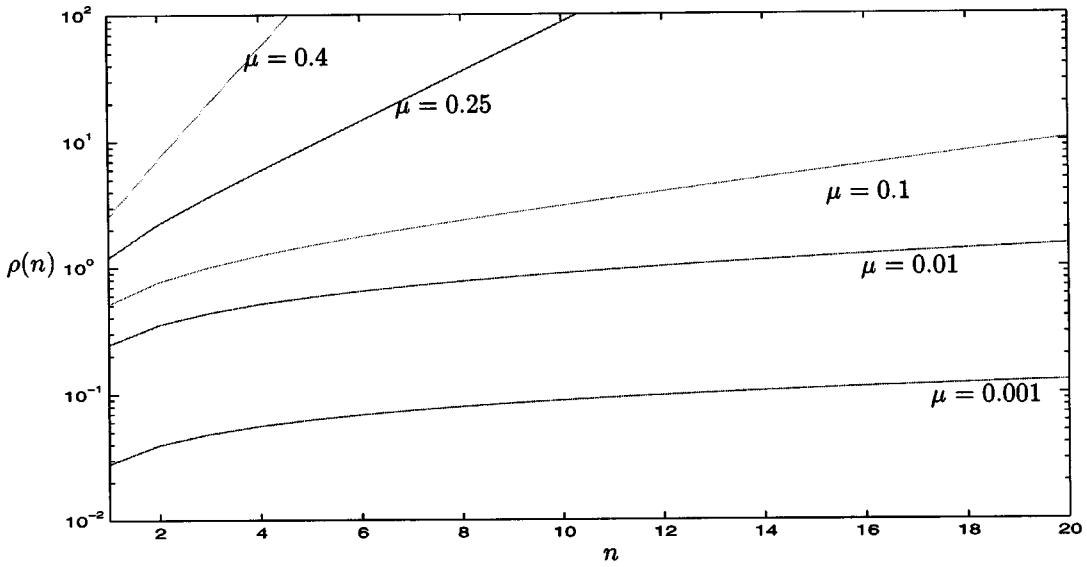

Fig. 5. $\rho(n)$ computed for $\mu = 0.001, 0.01, 0.1, 0.25$, and 0.4.

be a good approximation for EY_n at those time instants where $\rho'(n)$ is sufficiently small.[3]

Thus, a small $\rho(n)$ is desirable to conclude that Y_n or $\hat{D}(n)$ is close to EY_n. However, it turns out that the ratio $\rho(n)$ increases with n (and thus $\hat{D}(n)$ approximates EY_n less effectively for larger n, which is consistent with the results of our a.s. analysis). To see this, we evaluate EY_n^2

$$\begin{aligned} EY_n^2 &= (Eu^2)^n Y_0^2 \\ &= (1 - 4\mu\sigma_2 + 6\mu^2\sigma_4 - 4\mu^3\sigma_6 + \mu^4\sigma_8)^n Y_0^2 \end{aligned}$$

where $\sigma_6 \triangleq E\mathbf{x}_n^6$ and $\sigma_8 \triangleq E\mathbf{x}_n^8$ are assumed finite. Define further

$$\begin{aligned} r_4 &\triangleq Eu^2 = (1 - 4\mu\sigma_2 + 6\mu^2\sigma_4 - 4\mu^3\sigma_6 + \mu^4\sigma_8) \\ r_2 &\triangleq Eu = (1 - 2\mu\sigma_2 + \mu^2\sigma_4). \end{aligned} \tag{15}$$

With these definitions, $\rho(n)$ is given by

$$\rho(n) = \frac{\sqrt{r_4^n - r_2^{2n}}}{r_2^n} = \sqrt{\frac{r_4^n}{r_2^{2n}} - 1}. \tag{16}$$

Now, since

$$\begin{aligned} 0 \le \operatorname{var} Y_1 &= EY_1^2 - (EY_1)^2 \\ &= (Eu^2)Y_0^2 - (Eu)^2 Y_0^2 \\ &= r_4 Y_0^2 - r_2^2 Y_0^2 \end{aligned}$$

we conclude that $r_4 \ge r_2^2$ (with equality only if $\mathbf{x}_n^2$ is a constant with probability one). Therefore, except for this trivial case, $\rho(n)$ is strictly increasing, and thus, $\lim_{n\to\infty} \rho(n) = +\infty$. We have thus proven the following lemma.

Lemma 3—Variance of Y_n: Assume r_2 and r_4 defined above are finite and that the initial condition $\mathbf{w}_0$ to the scalar LMS algorithm (4) is deterministic. Then, the ratio $\rho(n)$ between the standard deviation of Y_n and EY_n is either 0 for all n or is strictly increasing with n and tends to infinity as n tends to infinity. □

Note that from our assumption that Y_0 is deterministic, we obtain $\rho(0) = 0$. In general (for step sizes for which Y_n converges in the MS sense), $\rho(n)$ remains small for some time, which implies (via Chebyshev's inequality) that Y_n is well approximated by EY_n when n is small. We give below examples of the behavior of $\rho(n)$ for two different input distributions.

1) **Binary inputs**. We first give a simple example for which $\rho(n) \equiv 0$. Assume that $\mathbf{x}_n = \pm 1$ with probability 0.5. Then, $u(n) \equiv (1-\mu)^2$ is a constant, and thus, $Y_n \equiv EY_n$. In this trivial case, we have $\sigma_2 = \sigma_4^{1/2} = \sigma_6^{1/3} = \sigma_8^{1/4}$, and $\rho(n) = 0$ for all n.
2) **Gaussian inputs**. Let $\mathbf{x}_n$ be Gaussian with zero mean and unit variance so that $\sigma_2 = 1$, $\sigma_4 = 3$, $\sigma_6 = 15$ and $\sigma_8 = 105$. We plot the value of $\rho(n)$ for several values of μ in the range $0 < \mu < 2/3 = 2\sigma_2/\sigma_4$ in Fig. 5. Note how $\rho(n)$ grows increasingly quickly as μ increases. In addition, note that the rate of increase of $\rho(n)$ is very small for $\mu \approx 0$.

G. Two Rates of Convergence

Let us consider again the differences between theory and simulation. Assume that we fix the time instant n and compare the values of EY_n and $\hat{D}(n)$ at that particular time instant for different values of L. We know from the expression for $\rho'(n)$ that the larger the value of L is, the smaller the value of $\rho'(n)$ will be. Hence, the more we average, the closer will the value of

[3]Although a small $\rho(k)$ implies that $Y_k \approx EY_k$ and $\hat{D}(k) \approx EY_k$, a large $\rho(k)$ *does not imply* that the difference $|Y_k - EY_k|$ or $|\hat{D}(k) - EY_k|$ should be large with a significant probability (but it does hint that this may be the case). For example, take a random variable z satisfying

$$z = \begin{cases} 10, & \text{with probability } 10^{-4} \\ 10^{-2}, & \text{with probability } 0.9999. \end{cases}$$

In this case $Ez^2 = 0.0101$ and $Ez = 0.0110$, and thus, $\rho = 9.1$. Despite the large value of ρ, $|z - Ez| \le 10^{-3} = 0.1Ez$ in 99.99% of the realizations.

$\hat{D}(n)$ be to that of EY_n. That is, the closer will both curves be at that time instant n. This again confirms an earlier conclusion in Lemma 2, viz., that the more we average, the longer it takes for us to see the differences between both curves.

Another major conclusion that follows from the a.s. and variance analyzes is that the LMS recursion exhibits *two different rates of convergence* (even for single-tap adaptive filters) At first, for small n, a sample curve Y_n is close to EY_n and, therefore, converges at a rate that is determined by $E \ln u$. For larger n, the sample curve Y_n will converge at a rate that is determined by $\ln Eu$.

We can now justify our sixth claim in the introduction, viz., that the knowledge that an adaptive filter is a.s. convergent does not necessarily guarantee satisfactory performance. Thus, assume that a filter is a.s. stable but that MS is unstable. It follows from our analysis that a learning curve will tend to diverge in the first iterations (by following the divergent mean-square learning curve), and only after an *unknown* interval of time will the learning curve start to converge. In simulations, we noticed that the estimation error may reach quite large values before starting to decrease again. Therefore, the performance of an a.s. stable filter need not always be satisfactory in applications.

IV. Theoretical Analysis in the Vector Case

In this section, we extend the ideas presented above to larger filter lengths. It turns out that the behavior of the LMS algorithm for filter lengths $M > 1$ is richer than what we saw in the scalar case and is (except when the step size is vanishingly small) very dependent on the actual input distribution. Therefore, the examples shown in this section cannot be exhaustive, i.e., the examples do not show all possible kinds of behavior—but they do illustrate the phenomena in which we are interested. As before, we will provide MS, a.s., and variance analyses. We start with the latter and explain how to compute the variance of $\|\tilde{\mathbf{w}}_n\|^2$ in the vector case (by generalizing the results of Section III-F).

A. Variance Analysis

We continue to assume that the input sequence $\{\mathbf{x}_n \in \mathbb{R}^M\}$ is iid and that the noise is identically zero $(v(n) \equiv 0)$. The individual entries of each regressor vector $\mathbf{x}_n$, however, are not assumed to be independent. The ratio $\rho(n)$ is defined in the vector case as

$$\rho(n) = \frac{\sqrt{\text{var}(\|\tilde{\mathbf{w}}_n\|^2)}}{E\|\tilde{\mathbf{w}}_n\|^2} \tag{17}$$

whose computation requires that we evaluate both the mean and the variance of $\|\tilde{\mathbf{w}}_n\|^2$. We consider first the evaluation of the mean.

1) Evaluating the Mean of $\|\tilde{\mathbf{w}}_n\|^2$: Recall that in Section II, we computed $E\|\tilde{\mathbf{w}}_n\|^2$ by finding a recursion for the diagonal entries of $C_n = E\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T$ and by using the fact that $E\|\tilde{\mathbf{w}}_n\|^2 = \text{Tr}\, C_n$. In that section, a recursion for the diagonal entries of C_n was all we needed since we assumed that the individual entries of $\mathbf{x}_n$ were independent. When the entries of $\mathbf{x}_n$ are not independent, which is the general case we are treating here, the off-diagonal elements of C_n should enter into the recursion. Therefore, let us first show how such a general recursion can be obtained.

Subtracting $\mathbf{w}_*$ from both sides of the LMS recursion (2), we obtain the error equation

$$\tilde{\mathbf{w}}_n = \left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right)\tilde{\mathbf{w}}_{n-1}. \tag{18}$$

Then, we have

$$\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T = \left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right)\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T\left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right). \tag{19}$$

Taking expectations and using the independence of $\mathbf{x}_n$ from $\mathbf{w}_{n-1}$, we obtain

$$\begin{aligned} C_n = C_{n-1} - \mu RC_{n-1} - \mu C_{n-1}R \\ + \mu^2 E\left(\mathbf{x}_n\mathbf{x}_n^T C_{n-1}\mathbf{x}_n\mathbf{x}_n^T\right) \end{aligned} \tag{20}$$

where the last expectation is in general difficult to evaluate in closed form, except when the entries of $\mathbf{x}_n$ are mutually independent.[4] To address the above general case, it is necessary to know all the fourth-order moments and cross correlations between the entries of $\mathbf{x}_n$. Assuming that these fourth-order moments are known, we can simplify (20) using Kronecker products, as we now show.

The Kronecker product of two matrices $A \in \mathbb{R}^{m_a \times n_a}$ and $B \in \mathbb{R}^{m_b \times n_b}$ is defined as [30]

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,n_a}B \\ \vdots & & \vdots \\ a_{m_a,1}B & \cdots & a_{m_a,n_a}B \end{bmatrix}. \tag{21}$$

This operation has several useful properties, but the one that interests us is the following. Define the symbol vec(A) to represent an $m_a n_a$ column vector formed by stacking the columns of the matrix A one above the other. Let $C = AXB$, where A, B, and X are matrices of compatible dimensions. Then, the following equality holds [30, p. 254]:

$$\text{vec}(C) = (B^T \otimes A)\,\text{vec}(X). \tag{22}$$

Applying this property to (20) and using the independence of $\{\mathbf{x}_n\}$, we obtain

$$\text{vec}(C_n) = A_n\text{vec}(C_{n-1}) \tag{23}$$

where

$$\begin{aligned} A_n = I_{M^2} - \mu(R \otimes I_M) - \mu(I_M \otimes R) \\ + \mu^2 E\left(\mathbf{x}_n\mathbf{x}_n^T \otimes \mathbf{x}_n\mathbf{x}_n^T\right) \end{aligned}$$

and I_r represents the identity matrix of dimension r. We thus have a recursion for vec(C_n), which can be used to evaluate Tr(C_n) and, consequently, the mean $E\|\tilde{\mathbf{w}}_n\|^2$.

In the following, we will often use repeated Kronecker products, as in $A \otimes A \otimes A$. We will denote such "Kronecker powers" as $A^{\otimes 2} = A \otimes A$ and similarly for $k > 2$.

2) Evaluating the Variance of $\|\tilde{\mathbf{w}}_n\|^2$: We still need to evaluate the numerator of $\rho(n)$ in (17), which requires that we evaluate var$(\|\tilde{\mathbf{w}}_n\|^2)$. We start by noting that

$$\text{var}(\|\tilde{\mathbf{w}}_n\|^2) = E\|\tilde{\mathbf{w}}_n\|^4 - (E\|\tilde{\mathbf{w}}_n\|^2)^2 \tag{24}$$

and that, as shown in Appendix B, we can rewrite $E\|\tilde{\mathbf{w}}_n\|^4$ as

$$E\|\tilde{\mathbf{w}}_n\|^4 = E\,\text{Tr}\left(\left(\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T\right)^{\otimes 2}\right).$$

[4]When $\mathbf{x}_n$ is Gaussian, this problem can also be simplified and be reduced to the case of independent entries in $\mathbf{x}_n$.

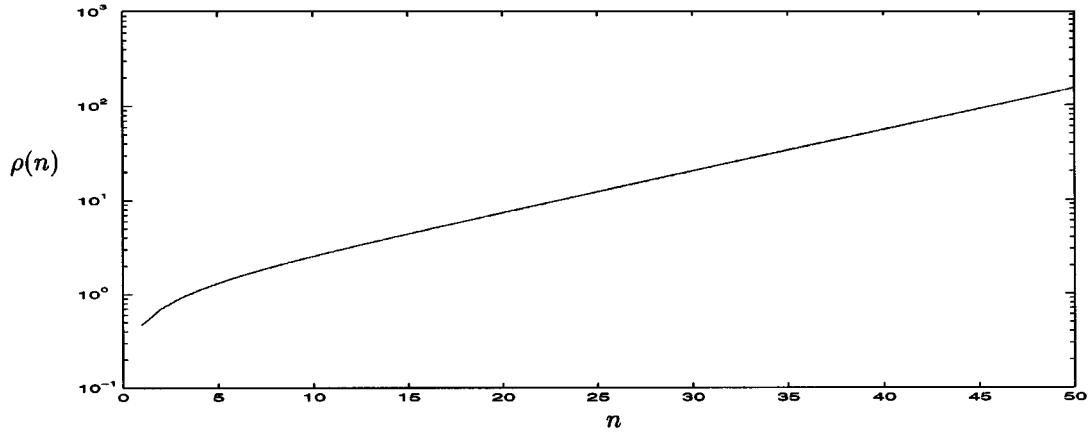

Fig. 6. $\rho(n)$, computed for $M = 2, \mu = 0.25$, and Gaussian regressor sequence with $R = I$.

Using this result, we can establish the following recursion for $\text{vec}((\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T)^{\otimes 2})$.

Lemma 4—Recursion for Variance Calculation: The expected value $E(\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T)^{\otimes 2}$ can be computed from the recursion

$$\begin{aligned} &E\,\text{vec}\left(\left(\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T\right)^{\otimes 2}\right) \\ &\quad = E\left(\left(I-\mu\mathbf{x}_n\mathbf{x}_n^T\right)^{\otimes 4}\right) E\,\text{vec}\left(\left(\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T\right)^{\otimes 2}\right) \\ &\quad \triangleq \Psi E\,\text{vec}\left(\left(\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T\right)^{\otimes 2}\right). \end{aligned} \tag{25}$$

The above recursion allows us to evaluate $E\|\tilde{\mathbf{w}}_n\|^4$, which in turn can be used in (24) to evaluate $\text{var}(\|\tilde{\mathbf{w}}_n\|^2)$. Thus, in principle, we know how to evaluate the ratio $\rho(n)$ in the vector case. A drawback of this method is that the matrix Ψ lies in $\mathbb{R}^{M^4\times M^4}$, and it becomes difficult to solve the recursion of Lemma 4 explicitly for large filter lengths. If the entries of $\mathbf{x}_n$ are mutually independent, several elements of Ψ vanish, and sparse matrix techniques can be used to simplify the problem.

In any case, our recursions allow us to evaluate $\rho(n)$ [as defined in (17)]. An example with Gaussian inputs and $M = 2$ is shown in Fig. 6 with the curve for $\rho(n)$ for $R = I$ and $\mu = 0.25$. The value of μ is chosen to be close to the value that achieves fastest convergence of $E\|\tilde{\mathbf{w}}_n\|^2$ in this case. Notice that, as in the scalar case, the simulation shows $\rho(n)$ growing with n. It also shows that $\rho(n)$ assumes relatively small values at the beginning of the simulation so that there will be good agreement between the actual learning curve and the EALC for small n. Fig. 8 further ahead confirms this effect for filters of length $M = 100$.

B. A.S. Convergence: Solution for a Simplified Model

As we mentioned in the scalar case, the variance analysis explains reasonably well the initial behavior of the EALC, but it cannot predict the behavior for large n. For that, we need an a.s. convergence analysis similar to what we did in Section III-B. We start by considering a *simplified* model here that will show that the effects we observed in the scalar case still exist in the vector case. It will also show that some new effects arise, especially the sensitivity of the behavior of the EALC to the direction of the initial condition. In Section IV-C, we will present a method of analysis that applies to more general models and input distributions.

Therefore, let $\mathbf{e}^{(i)}$ represent the ith basis vector, i.e., $e_j^{(i)} = 1$ if $i = j$ and zero otherwise, and assume that the input sequence $\{\mathbf{x}_n\}$ is of the form

$$\mathbf{x}_n = r(n)\mathbf{s}_n \tag{26}$$

where $r(n)$ is a random variable with zero mean. The vector $\mathbf{s}_n$ is independent of $r(n)$ and satisfies $\mathbf{s}_n = \mathbf{e}^{(i)}$ with probability p_i. In other words, $\mathbf{x}_n$ may assume only one out of M orthogonal directions (a similar model was used in a different context in [27]). Note that the entries of $\mathbf{x}_n$ are *dependent* in this case. As we did before, we assume that the noise is identically zero.

With these definitions, the weight vector $\tilde{\mathbf{w}}_n$ is given by

$$\begin{aligned} \tilde{\mathbf{w}}_n &= \prod_{i=1}^{n}\left(I-\mu\mathbf{x}_i\mathbf{x}_i^T\right)\tilde{\mathbf{w}}_0 \\ &= \text{diag}\left\{\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(1)}\right],\ldots\right. \\ &\qquad \left.\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(M)}\right]\right\}\tilde{\mathbf{w}}_0. \end{aligned}$$

Using this relation, we can compute $\|\tilde{\mathbf{w}}_n\|^2$ and $e(n)^2$ as follows:

$$\|\tilde{\mathbf{w}}_n\|^2 = \sum_{l=1}^{M}\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2\tilde{w}_{0,l}^2 \tag{27}$$

$$e(n)^2 = r(n+1)^2\tilde{w}_{0,j}^2\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(j)}\right]^2 \tag{28}$$

with probability p_j.

1) Mean-Square Analysis: Let $Er(i)^2 = \sigma_2$, and $Er(i)^4 = \sigma_4$. Since all $\mathbf{s}_i$ and $r(i)$ are independent and $E\mathbf{s}_i^T\mathbf{e}^{(l)} = p_l$, the MSD and MSE are given by

$$\begin{aligned} E\|\tilde{\mathbf{w}}_n\|^2 &= \sum_{l=1}^{M}(1-2\mu\sigma_2 p_l+\mu\sigma_4 p_l)^n\tilde{w}_{0,l}^2 \\ Ee(n)^2 &= \sigma_2\sum_{l=1}^{M}p_l(1-2\mu\sigma_2 p_l+\mu\sigma_4 p_l)^n\tilde{w}_{0,l}^2. \end{aligned} \tag{29}$$

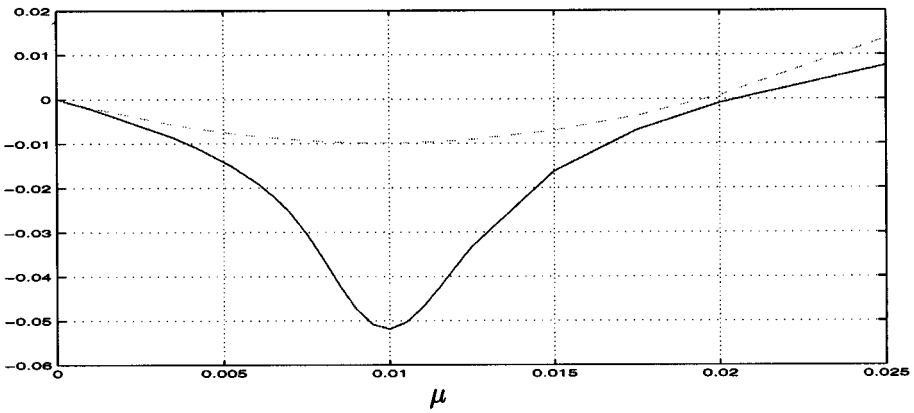

Fig. 7. Graphs of $E\ln(1-\mu r^2)^2$ (continuous line) and $\ln E(1-\mu r^2)^2$ (broken line), for $\mathcal{X}^2$ distribution with 100 degrees of freedom.

These relations express the MSD and the MSE in terms of exponential terms that depend on the factors $(1-2\mu\sigma_2 p_l+\mu\sigma_4 p_l)$, which are equal to

$$(1-2\mu\sigma_2 p_l+\mu\sigma_4 p_l)=E\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2.$$

Therefore, the MS convergence of all the modes will require that μ be such that

$$\ln E\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2<0 \tag{30}$$

or, equivalently, $\ln E[1-2p_l\mu\sigma_2+p_l\mu^2\sigma_4]<0$ for all $1\le l\le M$.

2) A.S. Analysis: Consider now one of the products in (27), i.e.,

$$P_l\triangleq\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2\tilde{w}_{0,l}^2.$$

Since this product has the same form as (5) in the scalar case, we can use our results of Section III-B to analyze its behavior. Evaluating the logarithm of P_l, we can verify that

$$\begin{aligned}\frac{\ln P_l}{n}&\overset{\text{a.s.}}{\to}E\ln\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2\\&=p_l E\ln[1-\mu r(i)^2]^2.\end{aligned}$$

We thus conclude that a.s. convergence requires [in contrast with (30)]

$$E\ln\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2<0 \tag{31}$$

or, equivalently, $E\ln(1-\mu r(i)^2)^2<0$. The distinction between conditions (30) and (31) highlights again the same phenomenon that occurred in the scalar case, viz., for large n, the rates of convergence of the true learning curve and the EALC will be distinct with the latter decaying faster.

3) Sensitivity to the Initial Condition: A new feature of the vector case is that the behavior of $\|\tilde{\mathbf{w}}_n\|^2$ is now dependent on the direction of the initial condition $\tilde{\mathbf{w}}_0$. Indeed, assume for example that all the probabilities p_l are equal, i.e., $p_l=1/M$, and that all the entries of the vector $\tilde{\mathbf{w}}_0$ are also equal. To further simplify the discussion, we normalize $\tilde{\mathbf{w}}_0$ so that $\|\tilde{\mathbf{w}}_0\|=1$, that is, we choose $\tilde{\mathbf{w}}_{0,l}=\pm1/\sqrt{M}$. In this situation, the norm of the weight error vector becomes

$$\|\tilde{\mathbf{w}}_n\|^2=\frac{1}{M}\sum_{l=1}^{M}\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2 \tag{32}$$

where the distribution of each of the terms in the sum is exactly the same. This means that $\|\tilde{\mathbf{w}}_n\|^2$ is, in fact, an average of M (*not* independent) scalar learning curves, each described by a term of the form

$$\tilde{q}_{n,l}=\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(l)}\right]^2\tilde{q}_{0,l},\quad \tilde{q}_{0,l}=1.$$

That is

$$\|\tilde{\mathbf{w}}_n\|^2=\frac{1}{M}\sum_{l=1}^{M}\tilde{q}_{n,l}.$$

Therefore, we should expect the variance of $\|\tilde{\mathbf{w}}_n\|^2$ to be smaller than that of each term in the sum [as we saw in (14) and in Section III-E].

On the other hand, if (for example) $\tilde{\mathbf{w}}_{0,1}=1$ and $\tilde{\mathbf{w}}_{0,2}=\cdots=\tilde{\mathbf{w}}_{0,M}=0$, then

$$\|\tilde{\mathbf{w}}_n\|^2=\prod_{i=1}^{n}\left[1-\mu r(i)^2\mathbf{s}_i^T\mathbf{e}^{(1)}\right]^2. \tag{33}$$

Since there is no averaging effect in the computation of the norm anymore, we should see exactly the same kind of behavior as for the scalar LMS algorithm. For other values of the initial condition $\tilde{\mathbf{w}}_0$, we have an averaging effect between the extremes of (32) and (33).

In Fig. 7, we plot the curves $E\ln(1-\mu r^2)^2$ and $\ln E(1-\mu r^2)^2$ for a variable r^2 that is distributed as a $\mathcal{X}^2$ variable with 100 degrees of freedom (this is exactly the distribution of $\|\mathbf{y}\|^2$ if the entries of the random vector $\mathbf{y}\in\mathbb{R}^{100}$ are Gaussian independent variables with unit variance). For a variable r with this distribution, we have $\sigma_2=Er^2=M$ and $\sigma_4=Er^4=M^2+2M$. Assuming that $p_1=p_2=\cdots=p_M=1/M$, we conclude from (30) that

$$\ln E(1-\mu r^2)^2=\ln(1-2\mu+(M+2)\mu^2). \tag{34}$$

Summarizing, the above discussion shows that an LMS filter with length $M>1$ and with input satisfying (26) will behave

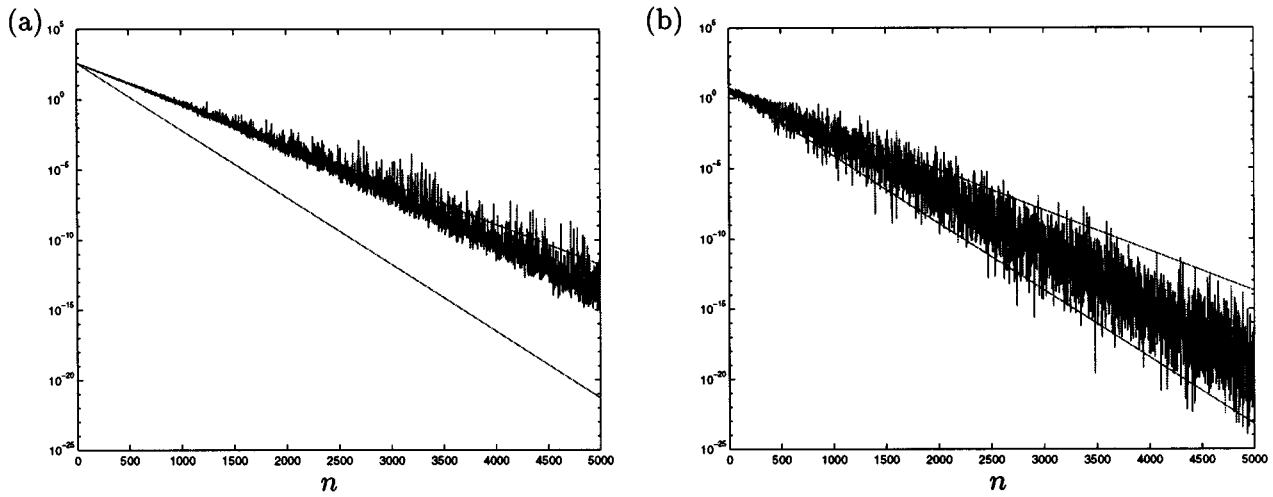

Fig. 8. $\hat{E}(n)^2$ computed with $L = 1000$, $M = 100$ and $\mu = 0.0042$. The input sequence satisfies (26), and r^2 is a $\mathcal{X}^2$. (a) All entries of the initial condition $\bar{\mathbf{w}}_0$ are equal. (b) Only the first entry of the initial condition $\bar{\mathbf{w}}_0$ is nonzero. The upper smooth curves are $Ee(n)^2$ computed theoretically, and the lower curves are the rate of convergence predicted by a.s. convergence analysis.

in a manner similar to that of a scalar LMS filter for which $\mathbf{x}_n$ has the same probability distribution as $r(n)$ but with three main differences.

1) The rate of convergence will now be smaller (depending on the values of the probabilities p_l).
2) A single realization of the error $\|\tilde{\mathbf{w}}_n\|^2$ will tend to be close to its mean $E\|\tilde{\mathbf{w}}_n\|^2$ for a longer time because of the averaging performed when computing the norm (32).
3) The behavior of an EALC is sensitive to the initial condition.

Fig. 8 illustrates the above results for a filter with 100 taps and such that r^2 is a $\mathcal{X}^2$ with 100 degrees of freedom and mean 100. In Fig. 8(a), all entries of the initial condition $\tilde{\mathbf{w}}_0$ are equal, whereas only the first entry $\tilde{w}_{0,1}$ in the initial condition for Fig. 8(b) was nonzero. Both plots show EALC's computed with $L = 1\,000$. Note how the first simulation stays close to $Ee(n)^2$ for a longer time, as we predicted above. Note, however, that in both simulations, the EALC's eventually tend to decrease with the (fastest) rate predicted by a.s. analysis (which, in this case, is equal to 0.9646, whereas the (slowest) rate predicted by MS analysis is 0.9905).

C. A.S. Convergence: A Solution for General Models

The analysis in the previous section assumed a special regression sequence $\{\mathbf{x}_n\}$ [see (26)].

Although restrictive, the resulting simplified model showed that the effects we observed in the scalar case still occur in the vector case. We now provide an analysis that applies to general regression vectors $\mathbf{x}_n$.

Thus, using (18), we obtain

$$\tilde{\mathbf{w}}_n = \left[\prod_{i=1}^{n}\left(I - \mu\mathbf{x}_i\mathbf{x}_i^T\right)\right]\tilde{\mathbf{w}}_0 \triangleq \Phi_n\tilde{\mathbf{w}}_0$$

where we defined the state-transition matrix Φ_n. In the simplified model prior to (27), the matrix Φ_n was assumed diagonal, which led to (27). Now, we get $\|\tilde{\mathbf{w}}_n\|^2 = \tilde{\mathbf{w}}_0^T\Phi_n^T\Phi_n\tilde{\mathbf{w}}_0$.

The rate of convergence of $\|\tilde{\mathbf{w}}_n\|^2$ will be dependent on the modes (eigenvalues) of $\Phi_n^T\Phi_n$. For the simple model (26) of the previous section, we were able to determine the properties of each individual eigenvalue of $\Phi_n^T\Phi_n$. In order to extend the analysis to more general input distributions, we study in this section the evolution of the determinant of $(\Phi_n^T\Phi_n)$, i.e., we now study the product of the eigenvalues of $\Phi_n^T\Phi_n$ and compare this product with $\det E(\Phi_n^T\Phi_n)$ since

$$E\|\tilde{\mathbf{w}}_n\|^2 = \tilde{\mathbf{w}}_0^T E\left(\Phi_n^T\Phi_n\right)\tilde{\mathbf{w}}_0.$$

1) Mean-Square Determinant Analysis: The computation of $\det\left(E\Phi_n^T\Phi_n\right)$ can be performed in the case of iid input regressors $\{\mathbf{x}_n\}$ by using our recursion for $\mathrm{vec}(C_n)$ in (23). Indeed

$$\begin{aligned}\det\left(E\Phi_n^T\Phi_n\right) = \det E\big[&\left(I - \mu\mathbf{x}_1\mathbf{x}_1^T\right) \quad \cdots \\ &\left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right)\left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right) \quad \cdots \\ &\left(I - \mu\mathbf{x}_1\mathbf{x}_1^T\right)\big].\end{aligned}$$

On the other hand, from (20), we obtain

$$\begin{aligned}C_n = E\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T = E\big[&\left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right) \quad \cdots \\ &\left(I - \mu\mathbf{x}_1\mathbf{x}_1^T\right)C_0\left(I - \mu\mathbf{x}_1\mathbf{x}_1^T\right) \quad \cdots \\ &\left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right)\big].\end{aligned} \tag{35}$$

The covariance C_n can be evaluated using Kronecker products, as we showed in (23). We can use the same method to compute $\det(E\Phi_n\Phi_n^T)$ as follows. Let F_n be obtained from (35) but with C_0 replaced by the identity matrix. Since $\{\mathbf{x}_n\}$ is stationary and iid, the order of the matrices in the product is irrelevant, and $F_n = E\Phi_n\Phi_n^T$. Therefore, we have

$$\det\left(E\Phi_n\Phi_n^T\right) = \det F_n \tag{36}$$

where $\mathrm{vec}(F_n)$ satisfies (with initial condition $F_0 = I$)

$$\mathrm{vec}(F_n) = A_n\,\mathrm{vec}(F_{n-1})$$

with A_n defined according to the equation following (23).

We will shortly present an example where the computation of $\det\left(E\Phi_n\Phi_n^T\right)$ simplifies, and a simple formula for its rate of convergence can be obtained.

2) A.S. Determinant Analysis: The determinant of Φ_n satisfies

$$\det \Phi_n = \prod_{i=1}^{n} \left[\det\left(I - \mu \mathbf{x}_i \mathbf{x}_i^T\right)\right] = \prod_{i=1}^{n} (1 - \mu \|\mathbf{x}_i\|^2)$$

where we used the fact that the matrix $I - \mu \mathbf{x}_i \mathbf{x}_i^T$ has $M-1$ eigenvalues at 1 and one eigenvalue at $1 - \mu\|\mathbf{x}_i\|^2$. We then obtain

$$\det\left(\Phi_n \Phi_n^T\right) = \prod_{i=1}^{n} (1 - \mu \|\mathbf{x}_i\|^2)^2$$

which has the same form as a scalar LMS algorithm with input sequence $\{\|\mathbf{x}_n\|\}$ so that

$$\frac{1}{n} \ln \det\left(\Phi_n \Phi_n^T\right) \stackrel{\text{a.s.}}{\rightarrow} E \ln (1 - \mu \|\mathbf{x}_i\|^2)^2.$$

Therefore, all of our previous results can be directly applied to this case. In particular, the rate of convergence (or divergence) of $\det \Phi_n \Phi_n^T$ for large n is a.s. given by $e^{E \ln (1-\mu\|\mathbf{x}_i\|^2)^2}$, which in general will be different than the rate obtained from (36).

To explicitly find the a.s. rate of convergence, it is necessary to know the distribution of $\|\mathbf{x}_n\|^2$, which depends on the distribution of $\mathbf{x}_n$ itself. We consider a few special cases below.

For example, let $\{\mathbf{x}_n\}$ be iid and such that the entries of each vector are mutually independent and Gaussian with unit variance. We saw in the previous section that in this case, $\|\mathbf{x}_n\|^2$ is distributed as a $\mathcal{X}^2$ with M degrees of freedom. In this case, the computation of $\det E\Phi_n^T \Phi_n$ simplifies considerably, as follows:

$$E\left(I - \mu \mathbf{x}_i \mathbf{x}_i^T\right)^2 = (1 - 2\mu + (M+2)\mu^2) I.$$

Since this is a multiple of the identity, $\det E\left(\Phi_n^T \Phi_n\right)$ reduces to

$$\det\left(E\Phi_n^T \Phi_n\right) = (1 - 2\mu + (M+2)\mu^2)^{Mn}.$$

This is similar to the expression that we obtained for the simplified model of Section IV-B2 [see (34) and Fig. 7], except that the factor $(1-2\mu+(M+2)\mu^2)$ is now raised to the power M. This means that for $M = 100$, the plots of Fig. 7 (with the vertical scale multiplied by 100) also apply to this example. Note that this example and that of Section IV-B2 are in fact very different—in this section, $\mathbf{x}_n$ may take *any* direction in $\mathbb{R}^M$, unlike what happened in the previous example. It only happens that the determinants have the same properties in both situations.

As another example for the computation of $E(1-\mu\|\mathbf{x}_n\|^2)^2$, assume that the entries of $\mathbf{x}_n$ have the same (non-Gaussian) distribution and are independent. In this situation, we can use the central limit theorem [29, p. 112] to conclude that for large M, the distribution of $\|\mathbf{x}_n\|^2$ will be approximately Gaussian with mean $M\sigma_2$ and variance $M(\sigma_{x,4} - \sigma_{x,2}^2) + M^2 \sigma_{x,2}^2$, where $\sigma_{x,2}$ and $\sigma_{x,4}$ are, respectively, the variance and the fourth moment of each entry of $\mathbf{x}_n$. This is true as long as both $\sigma_{x,2}$ and $\sigma_{x,4}$ are finite.

V. Concluding Remarks

In this paper, we have shown that there are situations in which the actual behavior of the LMS errors is significantly different than that of their averages. These situations arise when one uses larger step sizes (i.e., noninfinitesimal) to obtain faster convergence. Our simulations and analyses 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, we must be careful when analyzing ensemble-average learning curves.

Moreover, it follows from Section III-G that the performance of an a.s. convergent adaptive filter may be poor if the filter is not also MS stable. Looking at these results from another perspective, we might conclude that with larger step sizes, we should take into account both average and a.s. points of view for design purposes in order to get a clearer perspective.

We have proven our claims analytically and studied the behavior of the scalar LMS algorithm in detail. We also extended the conclusions to the vector case and showed that additional effects arise here.

Although our analysis was performed only for the LMS algorithm, a similar behavior can be expected by some other stochastic gradient algorithms.

Appendix A

In this Appendix, we prove the statement of Theorem 2. We do so by showing that both $E \ln(1-\mu\mathbf{x}_n^2)^2$ and $\ln E(1-\mu\mathbf{x}_n^2)^2$ are differentiable with respect to μ at $\mu = 0$ and that both derivatives are equal at that point. Now, the derivative of the second function evaluates to

$$\begin{aligned} \frac{d}{d\mu} \ln(1 - \mu\sigma_2 + \mu^2\sigma_4)\Big|_{\mu=0} &= -\frac{\sigma_2}{1 - \mu\sigma_2 + \mu^2\sigma_4}\Big|_{\mu=0} \\ &= -2\sigma_2. \end{aligned}$$

The evaluation of the other derivative is more involved and will be obtained in several steps in the lemmas below. The first lemma proves that $E \ln(1-\mu x^2)^2$ is well defined for all $\mu \geq 0$.

Lemma 5: Under the conditions of Theorem 2, the expected value

$$E \ln (1 - \mu x^2)^2 = \int_{-\infty}^{\infty} \ln (1 - \mu x^2)^2 p(x)\, dx$$

exists and is finite for all $\mu \geq 0$.

Proof: Let δ be a positive constant such that $1/\sqrt{\mu} + \delta > B$, and split the above integral as

$$\begin{aligned} E \ln (1 - \mu x^2)^2 &= \int_{-\infty}^{-\frac{1}{\sqrt{\mu}} - \delta} \ln (1 - \mu x^2)^2 p(x)\, dx \\ &\quad + \int_{-\frac{1}{\sqrt{\mu}} - \delta}^{\frac{1}{\sqrt{\mu}} + \delta} \ln (1 - \mu x^2)^2 p(x)\, dx \\ &\quad + \int_{\frac{1}{\sqrt{\mu}} + \delta}^{\infty} \ln (1 - \mu x^2)^2 p(x)\, dx \\ &\stackrel{\triangle}{=} I_1 + I_2 + I_3. \end{aligned}$$

Using the assumptions of Theorem 2, all three terms $\{I_1, I_2, I_3\}$ can be bounded. Indeed, using the assumption that $\sup_x p(x) < \infty$, we first have that

$$I_2 \leq 2 \left(\sup_x p(x)\right) \int_0^{\frac{1}{\sqrt{\mu}} + \delta} \left|\ln (1 - \mu x^2)^2\right| dx.$$

Now, by noting that $\ln(1-\mu x^2)^2$ is nonpositive if $0 \le x \le \sqrt{2/\mu}$ and positive otherwise and by evaluating the integral

$$\int \ln(1-\mu x^2)^2\,dx$$

we can conclude that I_2 is bounded. Consider next the term I_3. Using the fact that $1/\sqrt{\mu}+\delta > B$ and the assumptions of Theorem 2, we have

$$I_3 \le \int_{\frac{1}{\sqrt{\mu}}+\delta}^{\infty} \left|\frac{\ln(1-\mu x^2)^2}{x^\beta}\right| dx.$$

Since $\ln(1-\mu x^2)^2 < x$ for all positive x and $\beta > 6$, the above integral is finite. The first term I_1 can be bounded in a similar manner. □

With a small modification, the same arguments can be used to prove that $\mathrm{var}(E\ln(1-\mu x^2)^2)$ is finite.

Having proved that $E\ln(1-\mu x^2)^2$ exists, we now show that this function is differentiable at $\mu = 0$. Unfortunately, we cannot simply apply the formula

$$\frac{d}{d\mu}\int_{-\infty}^{\infty} \ln(1-\mu x^2)^2 p(x)\,dx = \int_{-\infty}^{\infty} \frac{\partial \ln(1-\mu x^2)^2}{\partial \mu} p(x)\,dx$$

because $\ln(1-\mu \mathbf{x}_n^2)^2$ is not a bounded function, and its derivative is not integrable, except at $\mu = 0$ [31, pp. 236–239]. We need to compute the derivative of $E\ln(1-\mu \mathbf{x}_n^2)^2$ directly from the definition, that is, we will show that

$$\lim_{\mu\to 0} \frac{\int_{-\infty}^{\infty} \ln(1-\mu x^2)^2 p(x)\,dx}{\mu} - \int_{-\infty}^{\infty} (-2x^2) p(x)\,dx = 0.$$

The computation of the above limit is carried out in the three lemmas below. The first two results show that we can avoid the singular points at $x = \pm(1/\sqrt{\mu})$ by restricting the integration limits to $-\mu^{-1/4+\gamma}$ and $\mu^{-1/4+\gamma}$, where $\gamma < 1/4$ is a small positive parameter.

Lemma 6: Assume that the conditions of Theorem 2 hold and that μ is small enough such that $\mu^{-1/4+\gamma} \ge B$. Then, there exists a finite constant C_1 such that

$$\left| \int_{-\infty}^{\infty} \ln(1-\mu x^2)^2 p(x)\,dx - \int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}} \ln(1-\mu x^2)^2 p(x)\,dx \right| < C_1 \mu^{(1/4-\gamma)(\beta-1)}. \quad (37)$$

Proof: Let Δ denote the expression on the left-hand side of the above inequality. Then, we can bound it by

$$\Delta \le 2\int_{\mu^{-1/4+\gamma}}^{\infty} \left|\frac{\ln(1-\mu x^2)^2}{x^\beta}\right| dx.$$

Performing the change of variables $y = \sqrt{\mu}x$, we further obtain

$$\Delta \le 2\mu^{\frac{\beta-1}{2}} \int_{\mu^{1/4+\gamma}}^{\infty} \left|\frac{\ln(1-y^2)^2}{y^\beta}\right| dy < 2\mu^{\frac{\beta-1}{2}} \left[\int_{\mu^{1/4+\gamma}}^{0.5} \frac{|\ln 0.75|}{y^\beta}\,dy + \int_{0.5}^{\infty} \left|\frac{\ln(1-y^2)^2}{y^\beta}\right| dy\right].$$

By evaluating the above integrals, we can easily verify that

$$\int_{0.5}^{\infty} \left|\frac{\ln(1-y^2)^2}{y^\beta}\right| dy = C_{11} < \infty$$

and

$$\int_{\mu^{1/4+\gamma}}^{0.5} \frac{|\ln 0.75|}{y^\beta}\,dy = C_{12}\mu^{-(1/4+\gamma)(\beta-1)} + C_{13}$$

for some finite constants $\{C_{11}, C_{12}, C_{13}\}$. Inequality (37) follows from these results. ◇

Lemma 7: The inequality below is satisfied under the conditions of the previous lemma.

$$\left| -2\int_{-\infty}^{\infty} x^2 p(x)\,dx + \int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}} 2x^2 p(x)\,dx \right| < C_2 \mu^{(1/4-\gamma)(\beta-3)}$$

where C_2 is a finite constant.

Proof: Since $\beta > 6$, we have

$$2\left| -\int_{-\infty}^{\infty} x^2 p(x)\,dx + \int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}} x^2 p(x)\,dx \right| < 4\left|\int_{\mu^{-1/4+\gamma}}^{\infty} \frac{x^2}{x^\beta}\,dx\right| = \frac{4}{\beta-3}\mu^{(1/4-\gamma)(\beta-3)}.$$

◇

Up to now, we have shown that for sufficiently small μ

$$\left|\int_{-\infty}^{\infty} \left[\frac{\ln(1-\mu x^2)^2}{\mu} + 2x^2\right] p(x)\,dx\right| < \left|\int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}} \left[\frac{\ln(1-\mu x^2)^2}{\mu} + 2x^2\right] p(x)\,dx\right| + C_1\mu^{[(1/4-\gamma)\beta-(5/4-\gamma)]} + C_2\mu^{(1/4-\gamma)(\beta-3)} \quad (38)$$

for some finite constants $\{C_1, C_2\}$, and $0 < \gamma < 1/8$ (note that the exponents of μ in the above inequality can be made positive if $\beta > 6$ by choosing a sufficiently small γ).

In order to bound the remaining integral above, we need to find out the dependence on x of the convergence of $(\ln(1-\mu x^2)^2/\mu) + 2x^2$ to zero as $\mu \to 0$. This is done in the next lemma. ◇

Lemma 8: The following inequality holds for all x and $\mu x^2 \le 1/5$:

$$0 > \frac{\ln(1-\mu x^2)^2}{\mu} + 2x^2 > -\frac{2x^2}{\frac{1}{\mu x^2}-1}. \quad (39)$$

Proof: First, note that

$$\frac{\ln(1-\mu x^2)^2}{\mu}+2x^2=\ln\left[\frac{(1-\mu x^2)^{2/\mu}}{e^{-2x^2}}\right]. \tag{40}$$

We find a bound for this function by studying the convergence of $(1-a/n)^n$ to e^{-a} as $n\to\infty$.

We begin our analysis by noting that the sequence $\{(1+(1/n))^n\}_{n=0}^{\infty}$ is strictly increasing and upper bounded by $1+1+2^{-1}+2^{-2}+\cdots+2^{-(n-1)}$. This implies that the inequality below holds for $m>n$

$$0<\left(1+\frac{1}{m}\right)^m-\left(1+\frac{1}{n}\right)^n<\frac{1}{2^n}+\cdots+\frac{1}{2^{m-1}}.$$

Taking this inequality to the limit as $m\to\infty$ and dividing the result by e, we have

$$1>\frac{\left(1+\frac{1}{n}\right)^n}{e}>1-\frac{1}{2^{n-1}e}. \tag{41}$$

Next, we translate this inequality to the case $(1-1/n)^n$ by considering the change of variables $t=-m-1$. Thus, note that we can write

$$\left(1+\frac{1}{m}\right)^m=\left(1-\frac{1}{t+1}\right)^{-t-1}=\left(1+\frac{1}{t}\right)^t\left(1+\frac{1}{t}\right)$$

so that applying (41) to this relation, we obtain, for $m<-1$

$$\begin{aligned}1-\frac{1}{m+1}&>\frac{\left(1+\frac{1}{m}\right)^m}{e}\\&>\left(1-\frac{2^{m+2}}{e}\right)\left(1-\frac{1}{m+1}\right)>1\end{aligned}$$

where the last inequality is true for $|m|\geq 5$. Performing the change of variables $m=-n/a$ (for some $a>0$), we further obtain

$$1+\frac{1}{\frac{n}{a}-1}>\frac{\left(1-\frac{a}{n}\right)^{-\frac{n}{a}}}{e}>1.$$

Finally, raising these inequalities to the power $-a$ and taking the logarithm, we find that

$$0>\ln\left[\frac{\left(1-\frac{a}{n}\right)^n}{e^{-a}}\right]>-a\ln\left[1+\frac{1}{\frac{n}{a}-1}\right].$$

Now, since for a small positive number ϵ, it holds that

$$\ln(1+\epsilon)=\epsilon-\frac{\epsilon^2}{2}+\frac{\epsilon^3}{3}-\cdots<\epsilon$$

we get for n large enough

$$0>\ln\left[\frac{\left(1-\frac{a}{n}\right)^n}{e^a}\right]>-\frac{a}{\frac{n}{a}-1}.$$

Applying this inequality to (40) with $a=2x^2$ and $n=2/\mu$, we obtain the desired result (39). □

With the above lemma, we can bound the remaining integral in (38), as below. Assume that $\mu^{1/2+2\gamma}<1/5$. It then follows from Lemma 8 that

$$\begin{aligned}&\left|\int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}}\left(\frac{\ln(1-\mu x^2)^2}{\mu}+2x^2\right)p(x)\,dx\right|\\&\quad<\left(\sup_x p(x)\right)\int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}}\frac{2x^2}{\frac{1}{\mu x^2}-1}\,dx.\end{aligned}$$

Now note that

$$\frac{2x^2}{\frac{1}{\mu x^2}-1}=\frac{2\mu x^4}{1-\mu x^2}$$

so that in the interval $|x|\leq\mu^{-1/4+\gamma}$, and for small enough μ, the above expression is smaller than

$$\frac{2\mu\mu^{-1+4\gamma}}{1-\mu\mu^{-1/2+2\gamma}}=\frac{2\mu^{4\gamma}}{1-\mu^{1/2+2\gamma}}.$$

Integrating, we get

$$\begin{aligned}&\left(\sup_x p(x)\right)\int_{-\mu^{-1/4+\gamma}}^{\mu^{-1/4+\gamma}}\frac{2x^2}{\frac{1}{\mu x^2}-1}\,dx\\&\quad<\frac{2\mu^{-1/4+5\gamma}}{1-\mu^{1/2+2\gamma}}\left(\sup_x p(x)\right).\end{aligned}$$

Substituting this result into (38), we conclude that all inequalities are satisfied with positive powers of μ if $\beta>6$ (and $\gamma>1/20$) so that

$$\begin{aligned}&\frac{d}{d\mu}\int_{-\infty}^{\infty}\ln(1-\mu x^2)^2p(x)\,dx\\&\quad=-2\int_{-\infty}^{\infty}x^2p(x)\,dx=-2\sigma_2\end{aligned}$$

which is our desired result.

Appendix B

Here, we prove the statement of Lemma 4. From (19), we have

$$\begin{aligned}&\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T\otimes\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T\\&\quad=\left[\left(I-\mu\mathbf{x}_n\mathbf{x}_n^T\right)\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T\left(I-\mu\mathbf{x}_n\mathbf{x}_n^T\right)\right]^{\otimes 2}.\end{aligned} \tag{42}$$

This expression can be simplified using another property of Kronecker products. For any matrices A,B,C, and D, it holds that [30, p. 244]

$$(A\otimes B)(C\otimes D)=(AC)\otimes(BD). \tag{43}$$

Now, apply this property to (42) with $A=B=(I-\mu\mathbf{x}_n\mathbf{x}_n^T)$ and $C=D=\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T(I-\mu\mathbf{x}_n\mathbf{x}_n^T)$ to obtain

$$\begin{aligned}\left(\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T\right)^{\otimes 2}&=\left(I-\mu\mathbf{x}_n\mathbf{x}_n^T\right)^{\otimes 2}\\&\quad\times\left[\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T\left(I-\mu\mathbf{x}_n\mathbf{x}_n^T\right)\right]^{\otimes 2}.\end{aligned}$$

Applying (43) again, now with $A = B = \tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T$ and $C = D = (I - \mu\mathbf{x}_n\mathbf{x}_n^T)$, we obtain

$$\left(\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^T\right)^{\otimes 2} = \left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right)^{\otimes 2}\left(\tilde{\mathbf{w}}_{n-1}\tilde{\mathbf{w}}_{n-1}^T\right)^{\otimes 2} \times \left(I - \mu\mathbf{x}_n\mathbf{x}_n^T\right)^{\otimes 2}.$$

We can now apply (22) and take expected values to obtain the desired result.

ACKNOWLEDGMENT

The authors wish to thank Dr. P. Williams (Litton Data Systems, Agoura Hills, CA) and N. J. de Moraes Nascimento (Cyclades Brasil, São Paulo) for helpful discussions and feedback.

REFERENCES

[1] B. Widrow *et al.*, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, pp. 1151–1162, 1976.
[2] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.
[3] A. Feuer and E. Weinstein, "Convergence analysis of LMS filters with uncorrelated Gaussian data," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-33, pp. 222–229, Feb. 1985.
[4] J. B. Foley and F. M. Boland, "A note on the convergence analysis of LMS adaptive filters with Gaussian data," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1087–1089, July 1988.
[5] O. Macchi, *Adaptive Processing: The LMS Approach with Applications in Transmission*. New York: Wiley, 1995.
[6] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[7] J. E. Mazo, "On the independence theory of equalizer convergence," *Bell Syst. Tech. J.*, vol. 58, pp. 963–993, May–June 1979.
[8] S. K. Jones, R. K. Cavin, and W. M. Reed, "Analysis of error-gradient adaptive linear estimators for a class of stationary dependent processes," *IEEE Trans. Inform. Theory*, vol. 28, pp. 318–329, Mar. 1982.
[9] O. Macchi and E. Eweda, "Second-order convergence analysis of stochastic adaptive linear filtering," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 76–85, Jan. 1983.
[10] F. Koziny, "A survey of stability of stochastic systems," *Automatica*, vol. 5, pp. 95–112, 1969.
[11] A. Parthasarathy and R. M. Evan-Iwanowski, "On the almost sure stability of linear stochastic systems," *SIAM J. Appl. Math.*, vol. 34, no. 4, pp. 643–656, June 1978.
[12] R. R. Bitmead and B. D. O. Anderson, "Adaptive frequency sampling filters," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 524–533, June 1981.
[13] R. R. Bitmead, "Convergence properties of LMS adaptive estimators with unbounded dependent inputs," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 477–479, May 1984.
[14] R. R. Bitmead, B. D. O. Anderson, and T. S. Ng, "Convergence rate determination for gradient-based adaptive estimators," *Automatica*, vol. 22, pp. 185–191, 1986.
[15] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[16] H. J. Kushner and F. J. Vázquez-Abad, "Stochastic approximation methods for systems over an infinite horizon," *SIAM J. Contr. Optim.*, vol. 34, no. 2, pp. 712–756, Mar. 1996.
[17] L. Guo, L. Ljung, and G.-J. Wang, "Necessary and sufficient conditions for stability of LMS," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 761–770, June 1997.
[18] V. Solo, "The stability of LMS," *IEEE Trans. Signal Processing*, vol. 45, pp. 3017–3026, Dec. 1997.
[19] G. V. Moustakides, "Exponential convergence of products of random matrices: Application to adaptive algorithms," *Int. J. Adapt. Contr. Signal Process.*, vol. 12, no. 7, pp. 579–597, Nov. 1998.
[20] S. Florian and A. Feuer, "Performance analysis of the LMS algorithm with a tapped delay line (two-dimensional case)," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1542–1549, Dec. 1986.
[21] S. C. Douglas and W. Pan, "Exact expectation analysis of the LMS adaptive filter," *IEEE Trans. Signal Processing*, vol. 43, pp. 2863–2871, Dec. 1995.
[22] V. H. Nascimento and A. H. Sayed, "Stability of the LMS adaptive filter by means of a state equation," in *Proc. 36th Annu. Allerton Conf. Commun., Contr., Comput.*, Allerton, IL, Sept. 1998, pp. 242–251.
[23] J. Shynk, "Frequency-domain and multirate adaptive filtering," *Signal Process. Mag.*, vol. 1, pp. 15–37, 1992.
[24] M. Rupp, "The behavior of LMS and NLMS algorithms in the presence of spherically invariant processes," *IEEE Trans. Signal Processing*, vol. 41, pp. 1149–1160, Mar. 1993.
[25] S. Vembu, S. Verdú, R. A. Kennedy, and W. Sethares, "Convex cost functions in blind equalization," *IEEE Trans. Signal Processing*, vol. 42, pp. 1952–1960, Aug. 1994.
[26] N. J. Bershad, "Analysis of the normalized LMS algorithm with Gaussian inputs," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 793–806, 1986.
[27] D. T. M. Slock, "On the convergence behavior of the LMS and the normalized LMS algorithms," *IEEE Trans. Signal Processing*, vol. 41, pp. 2811–2825, Sep. 1993.
[28] M. Tarrab and A. Feuer, "Convergence and performance analysis of the normalized LMS algorithm with uncorrelated Gaussian data," *IEEE Trans. Inform. Theory*, vol. 34, pp. 680–691, July 1988.
[29] R. Durrett, *Probability: Theory and Examples*, 2nd ed. Duxbury, U.K.: Duxbury, 1996.
[30] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, MA: Cambridge Univ. Press, 1994.
[31] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

Vítor H. Nascimento was born in São Paulo, Brazil. He received the B.S. and M.S. degrees (with distinction) in electrical engineering from Escola Politécnica, University of São Paulo, in 1989 and 1992, respectively, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 1999.

He is currently an Assistant Professor with the Department of Electrical Engineering, University of São Paulo, where he also taught between 1990 and 1994. His research interests include digital signal processing, estimation, adaptive filtering, and control.

Ali H. Sayed (SM'99) received the Ph.D. degree in electrical engineering in 1992 from Stanford University, Stanford, CA.

He is Associate Professor of Electrical Engineering at the University of California, Los Angeles (UCLA). He has over 130 journal and conference publications, is co-author of the research monograh *Indefinite Quadratic Estimation and Control* (Philadelphia, PA: SIAM, 1999) and of the graduate-level textbook *Linear Estimation* (Englewood Cliffs, NJ: Prentice-Hall, 2000). He is also co-editor of the volume *Fast Reliable Algorithms for Matrices with Structure* (Philadelphia, PA: SIAM, 1999). He has contributed several articles to engineering and mathematical encyclopedias and handbooks and has served on the program committees of several international meetings. His research interests span several areas including adaptive and statistical signal processing, linear and nonlinear filtering and estimation, interplays between signal processing and control methodologies, and reliable and efficient algorithms for large-scale structured computations. He is a member of the editorial boards of the *SIAM Journal on Matrix Analysis and Its Applications* and the *International Journal of Adaptive Control and Signal Processing*, has served as co-editor of special issues of the journal *Linear Algebra and Its Applications*. (To learn more about his work, see the website of the UCLA Adaptive Systems Laboratory at www.ee.ucla.edu/asl.)

Dr. Sayed received the 1996 IEEE Donald G. Fink Award. He is past associate editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.