# 3 In *Least-Mean-Square Adaptive Filters: New Insights*, S. Haykin and B. Widrow, editors, pp. 79–104, Wiley, NY, 2003.

# Energy Conservation and the Learning Ability of LMS Adaptive Filters

[1]

Ali H. Sayed
Electrical Engineering Department
University of California, Los Angeles

Vitor H. Nascimento
Department of Electronic Systems Engineering
University of São Paulo, Brazil

This chapter provides an overview of interesting phenomena pertaining to the learning capabilities of stochastic-gradient adaptive filters, and in particular those of the least-mean-squares (LMS) algorithm. The phenomena indicate that the learning behavior of adaptive filters is more sophisticated, and also more favorable, than was previously thought, especially for larger step-sizes. The discussion relies on energy conservation arguments and elaborates on both the mean-square convergence and the almost-sure convergence behavior of an adaptive filter.

## 3.1 INTRODUCTION

Adaptive filters are prominent examples of systems that are designed to adjust to variations in their environments in order to meet certain performance criteria. The

[1] This material was based on work supported in part by the National Science Foundation under awards CCR-9732376 and ECS-9820765. The work of V. H. Nascimento was also supported by Grant 2000/09569-6 from FAPESP, Brazil.

learning curve of an adaptive filter is a widely used tool to evaluate how fast and how well an adaptive filter meets (or learns to meet) its objectives. This learning process has been extensively studied in the literature for slowly adapting systems, that is, for systems that employ infinitesimally small step-sizes. This chapter highlights several phenomena that characterize the learning capabilities of adaptive filters when larger step-sizes are used. The phenomena actually occur even for slowly adapting systems but are less pronounced, which explains why they may go unnoticed.

The purpose of the chapter is to provide a straightforward exposition to the topic so that it can provide motivation for further study and analysis. For this reason, the discussion focuses in some detail on a special case that helps illustrate and explain the desired phenomena in their simplest forms. Readers interested in more advanced cases, and in additional details, are referred to the article [1] and the textbook [20].

Among other results, it is argued here that after an initial learning phase, an adaptive filter generally learns at a rate that is higher than that predicted by mean-square theory. It is also argued that even single-tap adaptive filters can exhibit two very distinct rates of convergence; they learn at a slower rate initially and at a faster rate later. Several examples are provided to illustrate these and other effects.

## 3.2 LMS ADAPTIVE FILTERS

As is wellknown, adaptive filters are generally characterized by recursive updates of the form

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) \ + \ \text{a correction term},$$

where $\boldsymbol{w}(n)$ denotes an estimate at time $n$ of a certain unknown weight vector $\boldsymbol{w}_o$ that one wishes to estimate, $\boldsymbol{w}(n+1)$ is the updated weight estimate at time $n+1$, and the correction term determines the direction along which the correction to $\boldsymbol{w}(n)$ is performed. It is assumed that $\boldsymbol{w}_o$ and its estimates are $M$-dimensional column vectors.

Different adaptive schemes differ in the manner in which they evaluate the correction direction. This chapter focuses on adaptive updates of the form

$$\begin{aligned} \boldsymbol{w}(n+1) &= \boldsymbol{w}(n) \ + \ \mu \tfrac{\boldsymbol{u}(n)}{g(\boldsymbol{u}(n))} \, e(n), \quad k \geq 0, \\ e(n) &= d(n) - \boldsymbol{u}(n)^T \boldsymbol{w}(n), \end{aligned} \tag{3.1}$$

where $\boldsymbol{u}(n)$ is the regression column vector at time $n$, $g(\cdot)$ is a positive scalar-valued data nonlinearity, $\mu$ is a positive step-size parameter, and $e(n)$ is the output error at time $n$. The signal $e(n)$ measures the difference between the reference signal $d(n)$ and an estimate for it, which is given by the inner product $\boldsymbol{u}(n)^T\boldsymbol{w}(n)$. The reference $d(n)$ is usually assumed to have risen from a linear model of the form

$$d(n) = \boldsymbol{u}(n)^T \boldsymbol{w}_o + v(n), \tag{3.2}$$

where $v(n)$ denotes possible disturbances or measurement noise. The random variables $\{\boldsymbol{u}(n), d(n), v(n)\}$ are assumed to be zero-mean. Also, the initial condition for recursion (3.1) is taken as $\boldsymbol{w}(0) = 0$, without loss of generality.

The choice $g(\boldsymbol{u}(n)) \equiv 1$ results in the famed LMS algorithm,

$$\begin{aligned} \boldsymbol{w}(n+1) &= \boldsymbol{w}(n) \ + \ \mu\boldsymbol{u}(n)\, e(n), \quad k \geq 0, \\ e(n) &= d(n) - \boldsymbol{u}(n)^T \boldsymbol{w}(n), \end{aligned} \tag{3.3}$$

while the choice $g(\boldsymbol{u}(n)) \equiv 1/(\delta + \|\boldsymbol{u}(n)\|^2)$ results in the normalized least-mean-squares (NLMS) algorithm

$$\begin{aligned} \boldsymbol{w}(n+1) &= \boldsymbol{w}(n) \ + \ \mu\tfrac{\boldsymbol{u}(n)}{\delta+\|\boldsymbol{u}(n)\|^2}\, e(n), \quad k \geq 0, \\ e(n) &= d(n) - \boldsymbol{u}(n)^T \boldsymbol{w}(n), \end{aligned} \tag{3.4}$$

where $\delta$ is a small positive number and $\|\cdot\|$ denotes the Euclidean norm of its vector argument.

## 3.3 THE LEARNING CURVE

The variance of the error signal $e(n)$ is widely accepted as a performance measure; its evolution with time is called the *learning curve* of the adaptive filter:

$$\text{learning curve} \ \stackrel{\Delta}{=} \ \mathrm{E}\, e^2(n), \quad n \geq 0,$$

where E denotes the expectation operator.

By examining the learning curve of an adaptive filter, one can extract useful information about the adaptation process, such as the error variance that remains in steady state (the smaller its value the better), the rate at which the error variance tends to its steady-state value (the faster the better), whether a filter is tending to steady state or not (the latter is indicative of instability), and so on. The learning curve is usually evaluated in one of three manners, which will now be discussed.

A. **Closed-Form Evaluation**. Perhaps the most desirable form of evaluation is a closed-form expression for the learning curve. However, this method of evaluation is possible only in very rare situations, so that it is generally considered to be a formidable task (if at all possible). This is not only because the update equation of an adaptive filter is time-variant and depends nonlinearly on the data, but also because the underlying signal statistics are usually unknown. For these reasons, it is considerably more common in the literature to evaluate the learning curve of an adaptive filter using either of the two methods that follow.

B. **Independence Theory**. This method of evaluation relies on imposing certain simplifying statistical conditions on the data. Among the most widely used conditions are those collectively known as the *independence assumptions*. These assumptions essentially require, whenever necessary, that the underlying random variables be statistically independent of each other in order to facilitate the evaluation of expectations of coupled terms, such as assuming that:

(a) The regression sequence $\{\boldsymbol{u}(n)\}$ is statistically independent and identically distributed (iid).

(b) The reference signal $d(n)$ is statistically independent of the regressors $\{\boldsymbol{u}(m), m \neq n\}$.

(c) The noise sequence $\{v(n)\}$ is also iid and statistically independent of the regression sequence $\{\boldsymbol{u}(n)\}$.

While such assumptions are not valid in most practical cases (e.g., when the regressors $\boldsymbol{u}(n)$ arise from a tapped-delay line implementation, in which case assumption (a) is violated), there is ample evidence in the literature (e.g., [2]–[7]) to support the premise that conclusions obtained under these conditions are sufficiently realistic for slow adaptation scenarios (i.e., for infinitesimally small step-sizes).

C. **Ensemble Averaging**. The third method of evaluation is the most practical and also the most widely used. It relies on controlled simulation or experimentation. In this technique, an adaptive filter is trained repeatedly and the resulting squared error curves are averaged to approximate the variance curve. More specifically, one performs several independent experiments or simulations, say $L$ of them. In each experiment, the adaptive filter is applied for a duration of $N$ iterations, always starting from the same initial condition and under the same statistical conditions for the sequences $\{d(n), \boldsymbol{u}(n), v(n)\}$. From each experiment $i$, a sample error curve is obtained:

$$\text{sample error curve} \;=\; \left\{ e^{(i)}(n),\ 0 \leq n \leq N \right\}.$$

After all $L$ experiments are completed, an approximation for the learning curve is computed by averaging as follows:

$$\text{Ensemble-average curve} \;\triangleq\; \frac{1}{L}\sum_{i=1}^{L}\left[e^{(i)}(n)\right]^2,\ \ 0 \leq n \leq N. \tag{3.5}$$

This method of evaluation is useful for complex filter updates for which closed-form expressions for learning curves are difficult to obtain even under the independence conditions. The method is also useful even for simple filter structures, e.g., when an analysis by the independence theory is not possible or even reliable due, for example, to faster adaptation (a situation that corresponds to non-infinitesimal step-sizes).

As mentioned before, the purpose of this chapter is to highlight several phenomena regarding the learning ability of adaptive filters. This will be pursued, both via examples and analytically, in the next sections.

## 3.4 ERROR MEASURES AND ENERGY RELATION

To begin with, it is helpful to introduce a few error measures and to derive a useful energy relation that will be called upon later in the arguments.

A. **Error Measures**. It is common to associate with every adaptive scheme of the form (3.1)–(3.2) two estimation errors: the so-called a priori and a posteriori errors,

$$e_a(n) \triangleq \boldsymbol{u}(n)^T \boldsymbol{\epsilon}(n), \qquad e_p(n) \triangleq \boldsymbol{u}(n)^T \boldsymbol{\epsilon}(n+1),$$

where $\boldsymbol{\epsilon}(n)$ denotes the weight error vector that is defined by

$$\boldsymbol{\epsilon}(n) = \boldsymbol{w}_o - \boldsymbol{w}(n).$$

The a priori error, $e_a(n)$, is a measure of how well the estimate $\boldsymbol{u}(n)^T\boldsymbol{w}(n)$ approximates the uncorrupted part $\boldsymbol{u}(n)^T\boldsymbol{w}_o$ of $d(n)$. Likewise, the a posteriori error, $e_p(n)$, is a measure of how well $\boldsymbol{u}(n)^T\boldsymbol{w}(n+1)$ approximates the term $\boldsymbol{u}(n)^T\boldsymbol{w}_o$.

Substituting the data model (3.2) into the defining relation (3.1) for the estimation error $e(n)$, it is easy to see that

$$e(n) \;=\; [\boldsymbol{u}(n)^T\boldsymbol{w}_o + v(n)] \;-\; \boldsymbol{u}(n)^T\boldsymbol{w}(n),$$

so that the errors $\{e(n), e_a(n)\}$ are related via

$$e(n) = e_a(n) + v(n). \tag{3.6}$$

Now under the common and reasonable assumption that:

<u>A.1</u> *The noise sequence $\{v(n)\}$ is iid with variance $\sigma_v^2$, and is statistically independent of the regression sequence $\{\boldsymbol{u}(n)\}$*

we find that

$$\mathrm{E}\, e^2(n) \;=\; \sigma_v^2 \;+\; \mathrm{E}\, e_a^2(n). \tag{3.7}$$

In other words, one finds that both curves

$$\mathrm{E}\, e^2(n) \quad \text{and} \quad \mathrm{E}\, e_a^2(n)$$

can be used to describe the learning behavior of an adaptive filter since they differ only by a constant factor (equal to $\sigma_v^2$). The noise assumption A.1 stated above will be enforced throughout this chapter, and the discussions will therefore focus on studying the behavior of the curve $\mathrm{E}\, e_a^2(n)$.

B. **Energy Relation**. Subtracting $\boldsymbol{w}_o$ from both sides of (3.1) leads to the weight error recursion

$$\boldsymbol{\epsilon}(n+1) = \boldsymbol{\epsilon}(n) - \mu \frac{\boldsymbol{u}(n)}{g(\boldsymbol{u}(n))} e(n). \tag{3.8}$$

Multiplying by $\boldsymbol{u}(n)^T$ from the left, one finds that the errors $\{e_p(n), e_a(n), e(n)\}$ are related via

$$e_p(n) \;=\; e_a(n) \;-\; \mu\frac{\|\boldsymbol{u}(n)\|^2}{g(\boldsymbol{u}(n))}e(n). \tag{3.9}$$

Substituting this relation back into (3.8), one obtains, for nonzero $\boldsymbol{u}(n)$, a recursion that relates all four error quantities $\{\boldsymbol{\epsilon}(n+1), \boldsymbol{\epsilon}(n), e_a(n), e_p(n)\}$:

$$\boldsymbol{\epsilon}(n+1) = \boldsymbol{\epsilon}(n) - \frac{\boldsymbol{u}(n)}{\|\boldsymbol{u}(n)\|^2}[e_a(n) - e_p(n)]. \tag{3.10}$$

Observe that the data nonlinearity function $g(\cdot)$ does not appear explicitly in this relation. Evaluating the energies of both sides of this equation leads to the following energy conservation relation:

$$\|\boldsymbol{\epsilon}(n+1)\|^2 + \frac{1}{\|\boldsymbol{u}(n)\|^2}e_a^2(n) = \|\boldsymbol{\epsilon}(n)\|^2 + \frac{1}{\|\boldsymbol{u}(n)\|^2}e_p^2(n). \tag{3.11}$$

When $\boldsymbol{u}(n) = 0$, it is obviously true that

$$\|\boldsymbol{\epsilon}(n+1)\|^2 = \|\boldsymbol{\epsilon}(n)\|^2\ . \tag{3.12}$$

Both results (3.11) and (3.12) can be grouped together into a single equation by defining

$$\bar{\mu}(n) = \left(\|\boldsymbol{u}(n)\|^2\right)^\dagger = \begin{cases} 0 & \text{if } \boldsymbol{u}(n) = \mathbf{0} \\ \frac{1}{\|\boldsymbol{u}(n)\|^2} & \text{otherwise} \end{cases}$$

in terms of the pseudo-inverse of a scalar, so that one obtains

$$\|\boldsymbol{\epsilon}(n+1)\|^2 + \bar{\mu}(n)e_a^2(n) = \|\boldsymbol{\epsilon}(n)\|^2 + \bar{\mu}(n)e_p^2(n). \tag{3.13}$$

This *energy conservation relation* holds for all adaptive filters whose recursions are of the form (3.1)–(3.2), and was originally developed in [8] in the context of robustness analysis of adaptive filters. No approximations or assumptions are needed to establish (3.13); it is an exact relation that shows how the energies of the weight error vectors at two successive time instants are related to the energies of the a priori and a posteriori estimation errors. Thus the energy relation provides a convenient and powerful framework for carrying out different kinds of performance analysis, both stochastic and deterministic, for a wide range of adaptive filters (see, e.g., [8]–[14] and the textbook [20]). It will be used in the sequel to shed some light on the learning behavior of adaptive filters.

## 3.5 TRANSIENT ANALYSIS

A recursion for the transient behavior of an adaptive filter can be deduced from the energy conservation relation (3.13). For this purpose, one reworks (3.13) so as to express it in an equivalent form that eliminates $e_p(n)$ and keeps only the a priori error $e_a(n)$, whose variance, as indicated earlier, is of interest to the learning behavior of an adaptive filter. Using (3.6) in (3.9) leads to

$$e_p(n) \;=\; \left[1-\mu\frac{\|\boldsymbol{u}(n)\|^2}{g(\boldsymbol{u}(n))}\right]e_a(n) \;-\; \mu\frac{\|\boldsymbol{u}(n)\|^2}{g(\boldsymbol{u}(n))}v(n). \tag{3.14}$$

Substituting this equality into the energy relation (3.13) and expanding terms, one finds, after some straightforward algebra, the equivalent representation:

$$\begin{aligned}\|\boldsymbol{\epsilon}(n+1)\|^2 \;&=\; \|\boldsymbol{\epsilon}(n)\|_A^2 \;+\; \tfrac{\mu^2\|\boldsymbol{u}(n)\|^2}{g^2(\boldsymbol{u}(n))}v^2(n) \\ &\quad -\tfrac{2\mu}{g(\boldsymbol{u}(n))}\left[1-\tfrac{\mu\|\boldsymbol{u}(n)\|^2}{g(\boldsymbol{u}(n))}\right]e_a(n)v(n),\end{aligned} \tag{3.15}$$

where $\boldsymbol{A}$ refers to the following matrix, which is a rank-one modification of the identity matrix,

$$\boldsymbol{A} \;\stackrel{\Delta}{=}\; I-\frac{\mu}{g(\boldsymbol{u}(n))}\left[2\;-\;\frac{\mu\|\boldsymbol{u}(n)\|^2}{g(\boldsymbol{u}(n))}\right]\boldsymbol{u}(n)\boldsymbol{u}(n)^T, \tag{3.16}$$

and the notation $\|x\|_P^2$ refers to the weighted inner product

$$\|x\|_P^2 \stackrel{\Delta}{=} x^T\boldsymbol{P}x.$$

Actually, and in order to be precise, one should write $\boldsymbol{A}(n)$ instead of $\boldsymbol{A}$ to emphasize the fact that $\boldsymbol{A}$ changes with $n$. However, for now, it is sufficient to use the compact notation $\boldsymbol{A}$. The more explicit notation $\boldsymbol{A}(n)$ will be used when it becomes necessary to indicate the time index.

Recursion (3.15) describes the dynamic evolution of the squared norm of the weight-error vector. It will be used in the sequel to characterize the mean-square and almost-sure performances of an adaptive filter.

## 3.6 EXAMPLES OF LEARNING BEHAVIOR

Several examples are presented in this section in order to motivate a handful of phenomena that characterize the learning capability of adaptive filters.

A. Example 1. Let $g(\boldsymbol{u}(n)) \equiv 1$ and thus consider the LMS recursion (3.3). Assume that the regression sequence $\{\boldsymbol{u}(n)\}$ is Gaussian and iid with variance matrix

$$\boldsymbol{R} \;\stackrel{\Delta}{=}\; \mathrm{E}\boldsymbol{u}(n)\boldsymbol{u}(n)^T.$$

Assume further that the reference sequence $d(n)$ is independent of $\{\boldsymbol{u}(m),\ m \neq n\}$. These conditions correspond to a situation in which the independence assumptions are satisfied and, in addition, the learning curve of the adaptive filter can be evaluated in closed form.

Indeed, it follows from the independence assumptions that $\boldsymbol{u}(n)$ is independent of $\boldsymbol{\epsilon}(n)$ and that $v(n)$ and $e_a(n)$ are also independent. Taking expectations of both sides of (3.15) with $g(\boldsymbol{u}(n)) \equiv 1$, and using the independence of $v(n)$ and $e_a(n)$, leads to the recursion

$$\mathrm{E}\ \|\boldsymbol{\epsilon}(n+1)\|^2 \ = \ \mathrm{E}\ \left(\|\boldsymbol{\epsilon}(n)\|_A^2\right) \ + \ \mu^2\sigma_v^2\mathrm{Tr}(\boldsymbol{R}), \tag{3.17}$$

with

$$\boldsymbol{A} \ = \ \boldsymbol{I} - \mu\left[2 \ - \ \mu\|\boldsymbol{u}(n)\|^2\right]\boldsymbol{u}(n)\boldsymbol{u}(n)^T.$$

Observe that the weight matrix $\boldsymbol{A}$ is a random variable since it depends on $\boldsymbol{u}(n)$. However, the independence of $\boldsymbol{u}(n)$ and $\boldsymbol{\epsilon}(n)$ permits the replacement of $\boldsymbol{A}$ by a constant matrix (namely, by its mean value). To see this, note that

$$\begin{aligned}
\mathrm{E}\ \left(\|\boldsymbol{\epsilon}(n)\|_A^2\right) &= \mathrm{E}\ (\boldsymbol{\epsilon}(n)^T\boldsymbol{A}\boldsymbol{\epsilon}(n)) \\
&= \mathrm{E}\left[\mathrm{E}\ (\boldsymbol{\epsilon}(n)^T\boldsymbol{A}\boldsymbol{\epsilon}(n))|\boldsymbol{\epsilon}(n)\right] \\
&= \mathrm{E}\boldsymbol{\epsilon}(n)^T\left[\mathrm{E}\ \boldsymbol{A}|\boldsymbol{\epsilon}(n)\right]\boldsymbol{\epsilon}(n) \\
&= \mathrm{E}\ \|\boldsymbol{\epsilon}(n)\|_F^2,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{F} &\stackrel{\Delta}{=} \mathrm{E}(\boldsymbol{A}) \\
&= \boldsymbol{I} - 2\mu\boldsymbol{R} + \mu^2[2\boldsymbol{R}^2 + \mathrm{Tr}(\boldsymbol{R})\boldsymbol{R}]
\end{aligned}$$

and the above value for $\boldsymbol{F}$ follows from the fact that for real-valued Gaussian regressors it holds that

$$\mathrm{E}\|\boldsymbol{u}(n)\|^2\boldsymbol{u}(n)\boldsymbol{u}(n)^T \ = \ 2\boldsymbol{R}^2 + \mathrm{Tr}(\boldsymbol{R})\boldsymbol{R}.$$

In this case, recursion (3.17) is seen to be equivalent to

$$\mathrm{E}\ \|\boldsymbol{\epsilon}(n+1)\|^2 \ = \ \mathrm{E}\ \left(\|\boldsymbol{\epsilon}(n)\|_F^2\right) \ + \ \mu^2\sigma_v^2\mathrm{Tr}(\boldsymbol{R}), \tag{3.18}$$

with $\boldsymbol{A}$ replaced by $\boldsymbol{F}$.

Now consider the choice $\boldsymbol{R} = \sigma_u^2\boldsymbol{I}$. In this case, $\mathrm{Tr}(\boldsymbol{R}) = M\sigma_u^2$ and $\boldsymbol{F}$ becomes a constant multiple of the identity

$$\boldsymbol{F} = \left[1 - 2\mu\sigma_u^2 + \mu^2(M+2)\sigma_u^4\right]\boldsymbol{I},$$

so that the weight-error variance relation (3.18) can be rewritten more directly as

$$\mathrm{E}\|\boldsymbol{\epsilon}(n+1)\|^2 = [1 - 2\mu\sigma_u^2 + \mu^2(M+2)\sigma_u^4] \cdot \mathrm{E}\|\boldsymbol{\epsilon}(n)\|^2 \; + \; \mu^2 M\sigma_u^2\sigma_v^2.$$

Now using (3.7) and the fact that

$$\mathrm{E}e_a^2(n) = \mathrm{E}|\boldsymbol{u}(n)^T\boldsymbol{\epsilon}(n)|^2 = \mathrm{E}\;\boldsymbol{\epsilon}(n)^T\boldsymbol{R}\boldsymbol{\epsilon}(n) = \sigma_u^2\;\mathrm{E}\|\boldsymbol{\epsilon}(n)\|^2,$$

one finds that the learning curve for this example is described in closed form by the recursion

$$\mathrm{E}e^2(n+1) = \left[1 - 2\mu\sigma_u^2 + \mu^2(M+2)\sigma_u^4\right]\mathrm{E}e^2(n) \; + \; 2\mu\sigma_u^2(1-\mu\sigma_u^2)\sigma_v^2,$$

with initial condition $\mathrm{E}e^2(0) = \mathrm{E}d^2(0) \equiv \sigma_d^2$.

It is clear that in this example the learning curve has a single mode and that it will be decaying (i.e., convergent) if, and only if, the step-size $\mu$ is chosen to satisfy

$$\left|1 - 2\mu\sigma_u^2 + \mu^2(M+2)\sigma_u^4\right| < 1. \tag{3.19}$$

Observe that the value of the mode is positive since

$$1 - 2\mu\sigma_u^2 + \mu^2(M+2)\sigma_u^4 = (1-\mu\sigma_u^2)^2 + \mu^2(M+1)\sigma_u^4 > 0.$$

Therefore, condition (3.19) is satisfied for

$$0 < \mu < \frac{2}{\sigma_u^2(M+2)}.$$

When this is the case, the filter is said to be *mean-square stable*. In addition, the fastest convergence rate occurs at the value of $\mu$ that minimizes the magnitude of the corresponding mode, which happens to be

$$\mu^o = \frac{1}{\sigma_u^2(M+2)}.$$

Figure 3.1 shows a plot of the learning curve for the numerical values $\mu = 0.1429$, $M = 5$, $\sigma_u^2 = 1$, and $\sigma_v^2 = 0.01$. For these numerical values, the filter is mean-square stable for step-sizes satisfying

$$0 < \mu < \frac{2}{7} \approx 0.2857.$$

In addition, the fastest convergence occurs at $\mu = 1/7 \approx 0.1429$, with the resulting mode at $0.8571$.

B. Example 2. The learning curve of the previous example is now estimated via ensemble averaging. The adaptive filter is applied several times, starting always

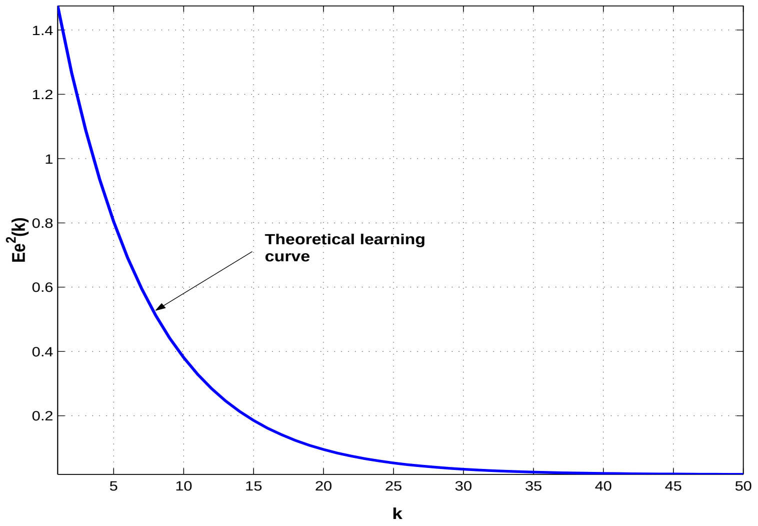

**Fig. 3.1** Theoretical learning curve for the LMS algorithm with Gaussian iid regressors, $M = 5$, $\sigma_u^2 = 1$, $\sigma_v^2 = 0.01$, and $\mu = 0.1429$.

from the same initial condition and maintaining the statistical properties of the data $\{\boldsymbol{u}(n), d(n), v(n)\}$ constant during the experiments. The same numerical values are used for $\mu = 0.1429$, $M = 5$, $\sigma_u^2 = 1$, and $\sigma_v^2 = 0.01$. Figure 3.2 shows the squared-error curves of four such experiments, of duration $N = 50$ samples each.

By averaging several such curves, one obtains the ensemble-average learning curve shown in Fig. 3.3. The figure shows a good fit between the actual learning curve derived in the previous example and the approximate one that is obtained via averaging. In general, the more experiments one averages, the better one expects the fit to be between the actual curve and the approximate curve. This statement seems intuitive but can also be deceptive, as argued later. It is customary in the literature to average around 10 to 300 curves.

C. Example 3. As mentioned earlier, the performance and learning capabilities of adaptive filters have been extensively studied in the literature under slow adaptation approximations (i.e., for vanishingly small step-sizes). However, faster adaptation rates are becoming desirable in order to track fast changing environments more precisely. This mode of operation requires the use of larger (non-infinitesimal) step-sizes. It turns out that some surprising phenomena arise under these conditions that seem to have been overlooked by earlier performance analyses. The phenomena also exist for small step-sizes but are less pronounced.

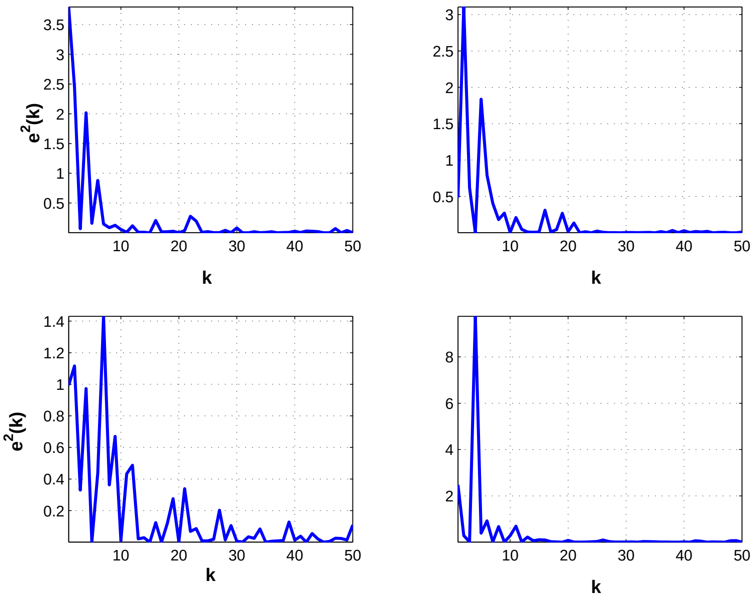

**Fig. 3.2** Four sample squared-error curves for LMS with Gaussian iid regressors, $M = 5$, $\sigma_u^2 = 1$, $\sigma_v^2 = 0.01$, and $\mu = 0.1429$.

In order to illustrate some of these phenomena, consider the same setting of Example 1. Figure 3.4 shows the theoretical learning curve and two ensemble average curves obtained when $\mu = 0.275$ (almost double the value of the previous step-size). The adaptive filter is still mean-square stable for this choice of the step-size (the corresponding mode will be at $0.9794$). One of the ensemble-average curves was obtained by averaging over $L = 500$ experiments, while the other was obtained by averaging over $L = 2500$ experiments. By examining the three curves in the figure, one observes at least five distinctive features:

1. In contrast to the previous case shown in Fig. 3.3 for a smaller step-size, there is not a good match between the theoretical learning curve and the ensemble-average curves.
2. The ensemble-average curves seem to converge faster than the theoretical learning curve.
3. During the initial training phase, all three curves (by theory and by experimentation) coincide reasonably well.
4. The ensemble-average curves seem to exhibit two distinct rates of convergence: an initial rate that agrees with the one predicted by mean-square

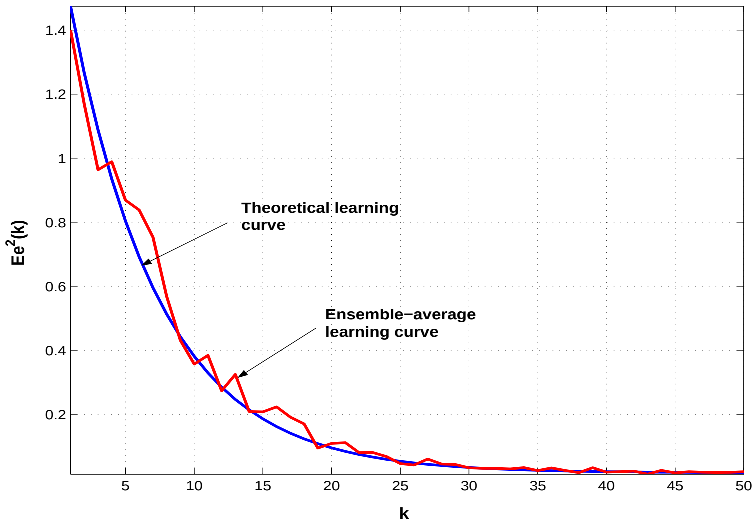

**Fig. 3.3** Theoretical and ensemble-average learning curves for LMS with Gaussian iid regressors, $M = 5$, $\sigma_u^2 = 1$, $\sigma_v^2 = 0.01$, and $\mu = 0.1429$.

theory and a later rate that is faster than that predicted by mean-square theory. Note further that these two distinct rates exist even for this example of an adaptive filter implementation with a single mode.

5. There is even a difference in behavior between the two ensemble-average curves themselves: The higher the number of averaging experiments, the closer the resulting ensemble-average curve is to the theoretical curve. The analysis in a later section will reveal that even if the number of experiments is increased significantly, there will continue to exist a discrepancy between the theoretical curve and the experimental curve.

It should be mentioned that although the earlier discussion was restricted to an example with independence assumptions on the data, these assumptions have actually been *enforced* in the analysis and in the simulations and are therefore valid. Thus the differences in behavior that one sees between the theoretical learning curve and the experimental ones are *not* due to assumptions that are made on the theoretical level and that are not valid on the practical level. In this way, one can conclude that even under these controlled conditions, the differences still exist. Actually, the differences occur even for situations where the independence assumptions are not satisfied (see [1]).

D. Example 4. There is one more phenomenon to highlight before moving on to a justification of the results observed so far. Thus consider again the numerical values used in Example 1, viz., $M = 5$, $\sigma_u^2 = 1$, and $\sigma_v^2 = 0.01$. For these

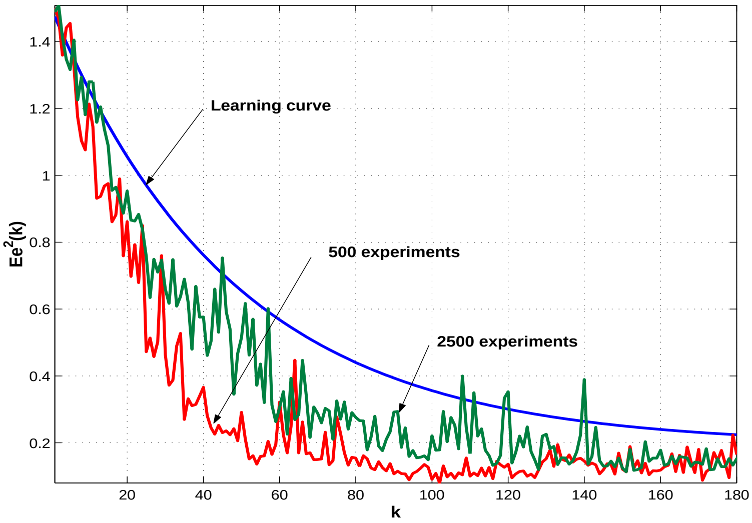

**Fig. 3.4** Theoretical and ensemble-averaged learning curves for LMS with Gaussian iid regressors, $M = 5$, $\sigma_u^2 = 1$, $\sigma_v^2 = 0.01$, and $\mu = 0.275$.

values, the filter was seen to be mean-square stable for step-sizes satisfying $\mu < 0.2857$. The diverging graph in Fig. 3.5 confirms this fact. However, the figure also shows a plot of the ensemble-average curve that is obtained for a larger step-size, $\mu = 0.29$, by averaging over 500 experiments. Mean-square theory predicts instability for this value of $\mu$, while the ensemble-average curve does not seem to diverge. Averaging over a larger number of experiments reveals a similar behavior. An explanation for this behavior is provided later by showing that, for larger step-sizes, there is a noticeable distinction between the mean-square and the almost-sure convergence behaviors of an adaptive filter.

E. Example 5. The earlier examples were concerned with data that satisfy the independence assumptions. Now consider a tapped-delay line implementation with two taps, so that the regression vector at time $n$ has the form

$$\boldsymbol{u}(n)^T = [\begin{array}{cc} u(n) & u(n-1) \end{array}].$$

Observe that due to the shift structure, two successive regressors cannot be independent and that, therefore, this is a situation where the independence assumptions are not valid. Assume further that the entries $\{u(n)\}$ are iid and uniform in the interval $[-0.5, 0.5]$ so that

$$\boldsymbol{R} = \mathrm{E}\boldsymbol{u}(n)\boldsymbol{u}(n)^T = \sigma_u^2 \boldsymbol{I}, \qquad \sigma_u^2 = \frac{1}{12}.$$

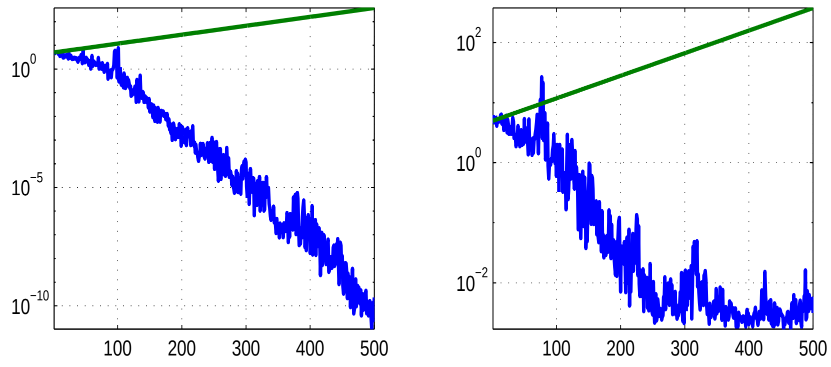

**Fig. 3.5** A comparison of the theoretical learning curve and the ensemble-average learning curve for a case that is mean-square unstable. The theoretical curve is seen to diverge in both cases, while the experimental curves converge. The plot on the left assumes zero noise while the plot on the right uses $\sigma_v^2 = 10^{-4}$. The ensemble-average curves were obtained by averaging over $500$ experiments, with step-size $\mu = 0.29$.

Figure 3.6 shows the ensemble-average curves that are obtained by averaging over $L = 100$ and $L = 1000$ experiments for $\mu = 7.9$. It is seen that in both cases, the averaged curves converge, as opposed to the theoretical curve, which is divergent for this value of the step-size (see [1]). Observe in addition that the larger the value of $L$, the longer the averaged curve stays closer to the theoretical curve before ultimately converging away from it.

## 3.7 MEAN-SQUARE CONVERGENCE

The examples in the previous section indicate that the behavior of the ensemble-average curves may show significant differences in relation to the behavior of the theoretical learning curve. An explanation for the origin of these differences is pursued in the following sections, which focus in some detail on the case of a single-tap adaptive filter.

Thus assume that $M = 1$, in which case $\boldsymbol{w}(n)$ and $\boldsymbol{u}(n)$ become scalars. Assume further that the noise signal $v(\cdot)$ is negligible enough so that its effect can be ignored. In this case, the energy recursion (3.15) collapses to

$$\boldsymbol{\epsilon}^2(n+1) \quad = \quad \boldsymbol{A}(n)\boldsymbol{\epsilon}^2(n), \tag{3.20}$$

where $\boldsymbol{A}(n)$ is the random scalar variable

$$\boldsymbol{A}(n) \;\stackrel{\Delta}{=}\; 1 - \frac{2\mu\boldsymbol{u}^2(n)}{g(\boldsymbol{u}(n))} + \frac{\mu^2\boldsymbol{u}^4(n)}{g^2(\boldsymbol{u}(n))} \;=\; \left(1 - \frac{\mu\boldsymbol{u}^2(n)}{g(\boldsymbol{u}(n))}\right)^2 .$$

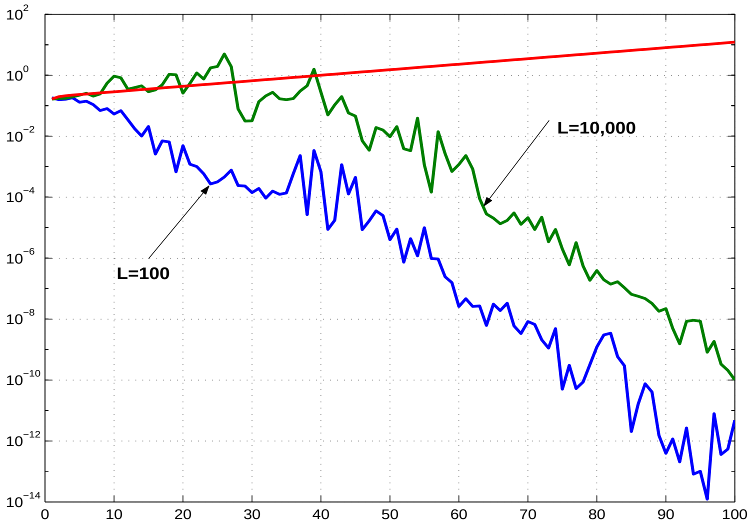

**Fig. 3.6** A comparison of the theoretical learning curve and the ensemble-average learning curves for an unstable tapped-delay line implementation with uniform input. The ensemble-average curves were obtained by averaging over $100$ and $10000$ experiments with step-size $\mu = 7.9$.

Observe that the dependency of $\boldsymbol{A}$ on $n$ is now indicated explicitly by means of a time index $n$.

Assuming that the $\{\boldsymbol{u}(n)\}$ are iid, it follows that the evolution of the variance of $\boldsymbol{\epsilon}(n)$ is described by difference equation

$$\mathrm{E}\boldsymbol{\epsilon}^2(n+1) = (\mathrm{E}\boldsymbol{A}(n)) \, \left(\mathrm{E}\boldsymbol{\epsilon}^2(n)\right). \tag{3.21}$$

In other words, the dynamics of the mean-square behavior of the filter is determined by $\mathrm{E}[\boldsymbol{A}(n)]$, which denotes the model of the above first-order recursion. Moreover, since the output error is given by

$$e(n) = d(n) - \boldsymbol{u}(n)^T \boldsymbol{w}(n) = \boldsymbol{u}(n)^T \boldsymbol{\epsilon}(n),$$

we find that

$$\mathrm{E}e^2(n) = \sigma_u^2 \, \mathrm{E}\boldsymbol{\epsilon}^2(n),$$

so that studying the evolution of $\mathrm{E}\boldsymbol{\epsilon}^2(n)$ is equivalent to studying the learning curve of the filter. Hence, the analysis in the sequel focuses on the behavior of $\boldsymbol{\epsilon}^2(n)$.

Now it is clear from (3.21) that the filter will be mean-square stable if, and only if, the step-size $\mu$ is chosen such that

$$\mathrm{E}(\boldsymbol{A}) < 1 \quad \Longleftrightarrow \quad \mathrm{E}\left(1 - \frac{\mu \boldsymbol{u}^2}{g(\boldsymbol{u})}\right)^2 < 1.$$

Observe that since all variables are stationary, by assumption, the time index $n$ is being dropped for compactness of notation, with $\{\boldsymbol{A}, \boldsymbol{u}\}$ written instead of $\{\boldsymbol{A}(n), \boldsymbol{u}(n)\}$.

The expectation of $\boldsymbol{A}$ is fully characterized in terms of the second and fourth moments of the normalized random variable

$$\bar{\boldsymbol{u}} \triangleq \frac{\boldsymbol{u}}{\sqrt{g(\boldsymbol{u})}}.$$

Indeed, let

$$\sigma_{\bar{u}}^2 \triangleq \mathrm{E}\bar{\boldsymbol{u}}^2, \quad \rho_{\bar{u}}^4 \triangleq \mathrm{E}\bar{\boldsymbol{u}}^4.$$

With these definitions, one gets

$$\mathrm{E}(\boldsymbol{A}) = 1 - 2\mu\sigma_{\bar{u}}^2 + \mu^2\rho_{\bar{u}}^4,$$

which describes a second-order equation in $\mu$. The condition $\mathrm{E}(\boldsymbol{A}) < 1$ then leads to the interval

$$0 < \mu < \frac{2\sigma_{\bar{u}}^2}{\rho_{\bar{u}}^4}.$$

In the LMS case, when $g(\boldsymbol{u}) \equiv 1$, the definitions of the moments collapse to

$$\sigma_{\bar{u}}^2 = \sigma_u^2, \quad \rho_{\bar{u}}^4 = \rho_u^4.$$

For ease of comparison with a later condition (see (3.27)), it is convenient to rewrite the requirement $\mathrm{E}(\boldsymbol{A}) < 1$ in the equivalent form (in terms of the natural logarithm):

$$\ln \mathrm{E}(\boldsymbol{A}) < 0 \quad \text{where} \quad \boldsymbol{A} \triangleq \left(1 - \frac{\mu\boldsymbol{u}^2}{g(\boldsymbol{u})}\right)^2, \tag{3.22}$$

where $u$ is an iid random variable.

## 3.8 ALMOST-SURE CONVERGENCE

In order to account for the differences between the theoretical and the experimental learning curves, this section now examines the behavior of a single (or typical) squared-error curve.

Starting from (3.20) and iterating it from time $0$ up to time $n$, one arrives at the expression

$$\boldsymbol{\epsilon}^2(n) = (\boldsymbol{A}(n-1)\boldsymbol{A}(n-2)\ldots\boldsymbol{A}(0))\,\boldsymbol{\epsilon}^2(0), \tag{3.23}$$

where the initial condition $\boldsymbol{\epsilon}(0)$ is deterministic (and equal to $\boldsymbol{w}_o$ since, by assumption, $\boldsymbol{w}_o = 0$). In addition, as explained in the previous section, the $\boldsymbol{A}(m)$ are realizations of an iid random variable $\boldsymbol{A}$ defined by

$$\boldsymbol{A} \triangleq \left(1 - \frac{\mu\boldsymbol{u}^2}{g(\boldsymbol{u})}\right)^2$$

Taking the natural logarithm of both sides of (3.23), and dividing by $n$ leads to

$$\frac{1}{n}\ln \boldsymbol{\epsilon}^2(n) = \frac{1}{n}\ln \boldsymbol{\epsilon}^2(0) \; + \; \frac{1}{n}\sum_{m=0}^{n-1} \ln \boldsymbol{A}(m). \tag{3.24}$$

Assuming that the variance of the random variable $\ln \boldsymbol{A}$ is bounded, then one can invoke the strong law of large numbers [18] to conclude that, as $n \to \infty$,

$$\frac{\ln \boldsymbol{\epsilon}^2(n)}{n} \quad \xrightarrow{a.s.} \quad \mathrm{E}\ln \boldsymbol{A}, \tag{3.25}$$

where "a.s." denotes almost-sure convergence. In other words, for large enough $n$, the curve $(\ln \boldsymbol{\epsilon}^2(n))/n$ converges almost surely to the constant value $\mathrm{E}\ln \boldsymbol{A}$. But what about the sample curve $\boldsymbol{\epsilon}^2(n)$ itself? The answer also follows from the strong law of larger numbers, which guarantees that, with probability 1, for each experiment $\Omega$, there exists a finite integer $K(\Omega)$ (dependent on the experiment) such that for all $n \geq K(\Omega)$, the sample curve $\boldsymbol{\epsilon}^2(n)$ will be upper bounded by the curve

$$\boldsymbol{\epsilon}^2(n) \; \leq \; \boldsymbol{\epsilon}^2(0) \; \exp^{n\mathrm{E}\ln \boldsymbol{A}} \; \exp^{\sqrt{2n\ln(\ln n)}\sigma_{\ln A}} \tag{3.26}$$

where $\sigma^2_{\ln A}$ denotes the variance of $\ln \boldsymbol{A}$.

Now the first exponential in (3.26) dominates the second when $n$ is large, which implies that the upper bound tends to zero if, and only if, $\mathrm{E}(\ln \boldsymbol{A}) < 0$. It thus follows that a typical curve $\boldsymbol{\epsilon}^2(n)$ converges to zero almost surely (or with probability 1) if, and only if, the step-size $\mu$ is chosen such that

$$\mathrm{E}\ln \boldsymbol{A} < 0, \quad \text{where} \quad \boldsymbol{A} \triangleq \left(1 - \frac{\mu \boldsymbol{u}^2}{g(\boldsymbol{u})}\right)^2, \tag{3.27}$$

where $\boldsymbol{u}$ again is an iid random variable. This leads to a different condition on $\mu$ than the one derived in (3.22) for mean-square stability.

## 3.9 COMMENTS AND DISCUSSION

Comparing the conditions (3.22) and (3.27) for mean-square and almost-sure convergence behaviors, one sees that there is a clear distinction between them. The two conditions are not equivalent and, in fact, one always implies the other since, for any nonnegative random variable $\boldsymbol{A}$ for which $\mathrm{E}\boldsymbol{A}$ and $\mathrm{E}\ln \boldsymbol{A}$ both exist, it holds that

$$\mathrm{E}\ln \boldsymbol{A} \leq \ln \mathrm{E}\boldsymbol{A}.$$

Therefore, values of the step-size $\mu$ for which mean-square convergence occurs always guarantee almost-sure convergence while the converse is not true: a value for which $\ln \mathrm{E}\boldsymbol{A} > 0$ (and thus for which mean-square divergence occurs) can still guarantee almost-sure convergence, or $\mathrm{E}\ln \boldsymbol{A} < 0$ — which explains the phenomenon in Fig. 3.5.

However, these distinctions disappear for infinitesimally small step-sizes, which explains why the phenomena described before can pass unnoticed at this level of adaptation. This is a consequence of the fact that under some reasonable assumptions about the probability density function of the random variable $\{\boldsymbol{u}\}$, it holds that (see, e.g., [1])

$$\mathrm{E}\ln \boldsymbol{A} = \ln \mathrm{E}\boldsymbol{A} + o(\mu), \tag{3.28}$$

where $o(\mu)$ is a function satisfying $\lim_{\mu\to 0} o(\mu)/\mu = 0$. That is, both conditions, which are functions of $\mu$, coincide for vanishingly small $\mu$. This explains why learning curves and ensemble-average learning curves tend to agree reasonably well for such small step-sizes.

It should be mentioned that connections between mean-square convergence and almost-sure convergence have been studied before in the literature (see, e.g., [15]–[17]). However, these earlier studies are concerned with the case of vanishingly small step-sizes for which both notions of performance tend to agree. The development in this chapter elaborates on the discrepancies that arise when larger step-sizes are employed.

To illustrate the above points further, and to highlight that for larger step-sizes, the difference between $\mathrm{E}\ln \boldsymbol{A}$ and $\ln \mathrm{E}\boldsymbol{A}$ can be significant, Fig. 3.7 shows the plots of these terms as functions of $\mu$ for the case of Gaussian $\{\bar{\boldsymbol{u}}\}$ with unit variance (this case arises, for example, when $g(\boldsymbol{u}) \equiv 1$ and $\boldsymbol{u}$ itself is Gaussian with unit variance). In this case,

$$\sigma_{\bar{u}}^2 = 1, \quad \rho_{\bar{u}}^4 = 3.$$

Note that both plots are close together for small $\mu$ but that they become significantly different as $\mu$ increases. Observe also that $\mathrm{E}\ln \boldsymbol{A}$ is negative well beyond the point where $\ln\big(\mathrm{E}\boldsymbol{A}\big)$ becomes positive. This implies that there is a range of step-sizes for which a typical curve $\boldsymbol{\epsilon}^2(n)$ converges to zero with probability 1, but $\mathrm{E}\boldsymbol{\epsilon}^2(n)$ diverges. This explains the simulations in Fig. 3.5 and 3.6. This is not a paradox. Since the convergence is not uniform, there is a small (but nonzero) probability that a sample curve $\boldsymbol{\epsilon}^2(n)$ will exist such that it assumes very large values for a long interval of time before converging to zero. Finally, note that the value of $\mu$ that achieves the fastest mean-square convergence is noticeably smaller than the step-size that achieves the fastest almost-sure convergence.

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. In other words, the almost-sure analysis condition allows one to clarify what happens when $L$ is fixed (the number of experiments) and $n$ is increased (the time dimension); the ensemble-average curve tends to separate from the true average curve for increasing $n$ due to the difference in the convergence rates.

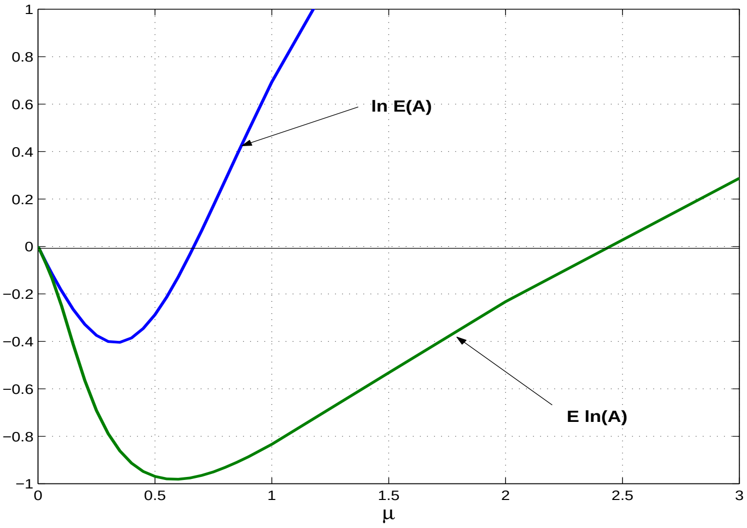

**Fig. 3.7** Graphs of $\mathrm{E}\ln \boldsymbol{A}$ and $\ln \mathrm{E}\boldsymbol{A}$ for Gaussian $\bar{\boldsymbol{u}}(n)$.

## 3.10 VARIANCE ANALYSIS

While the almost-sure analysis provides an explanation for the behavior of the ensemble-average curves for large $n$, one observes from the curves of Fig. 3.4 that for small $n$, i.e., close to the beginning of the curves, there is usually good agreement between the learning curve and the ensemble-average curves. This initial behavior can be explained by resorting to a variance analysis, which focuses on evaluating the variance of $\boldsymbol{\epsilon}^2(n)$, as opposed to its mean,

$$\mathrm{var}(\boldsymbol{\epsilon}^2(n)) = \mathrm{E}\boldsymbol{\epsilon}^4(n) \;-\; \left(\mathrm{E}\boldsymbol{\epsilon}^2(n)\right)^2 .$$

This approach is motivated by Chebyshev's inequality, which asserts that for any random variable $z$ it holds that

$$\mathrm{Prob}\left\{\left|z - \mathrm{E}z\right| \geq \kappa\right\} \leq \frac{\sigma_z^2}{\kappa^2}.$$

In other words, Chebyshev's inequality provides a bound on the probability that a random variable will lie outside an interval around its mean. The bound is seen to depend on the variance of the random variable. Hence, the smaller the variance of $z$, the smaller the bound. This is rather intuitive since the smaller the variance of a random variable, the more likely it is for the variable to assume values close to its mean.

Thus, define the ratio

$$\gamma(n) \triangleq \frac{\sqrt{\text{var}(\boldsymbol{\epsilon}^2(n))}}{\text{E}(\boldsymbol{\epsilon}^2(n))} \tag{3.29}$$

which is time dependent in general. It then follows from the above Chebyshev's inequality that

$$\text{Prob}\left\{\left|\boldsymbol{\epsilon}^2(n) - \text{E}\boldsymbol{\epsilon}^2(n)\right| \geq \frac{1}{2}\text{E}\boldsymbol{\epsilon}^2(n)\right\} \leq 4\gamma^2(n).$$

For example, the bound evaluates to $0.01$ for $\gamma(n) = 0.05$. This means that there is a $99$ percent probability that $\boldsymbol{\epsilon}^2(n)$ will be close to its mean (and, more specifically, lie within the interval $\left[0.5\text{E}\boldsymbol{\epsilon}^2(n),\ 1.5\text{E}\boldsymbol{\epsilon}^2(n)\right]$). Therefore, the smaller the value of $\gamma(n)$, the closer one expects the sample curve $\boldsymbol{\epsilon}^2(n)$ to be to the theoretical learning curve at that time instant.

Now recall that the ensemble-average learning curve is constructed by averaging together several sample curves $\boldsymbol{\epsilon}^2(n)$ to obtain, say,

$$\hat{D}(n) = \frac{1}{L}\sum_{i=1}^{L}\left[\boldsymbol{\epsilon}^2(n)\right]^{(i)}.$$

Assuming that the $L$ experiments are independent, then the expected value of the averaged curve $\hat{D}(n)$ is still equal to $\text{E}\boldsymbol{\epsilon}^2(n)$. However, the ratio $\gamma(n)$ that is associated with $\hat{D}(n)$ will be smaller and given by

$$\gamma'(n) = \frac{\sqrt{\text{var}(\hat{D}(n))}}{\text{E}\,\boldsymbol{\epsilon}^2(n)} = \frac{\gamma(n)}{\sqrt{L}}.$$

That is, the process of constructing ensemble-average curves reduces the value of $\gamma(n)$ by a factor of $\sqrt{L}$.

Although a small $\gamma(n)$ is desirable to conclude that $\boldsymbol{\epsilon}^2(n)$ or $\hat{D}(n)$ is close to $\text{E}\boldsymbol{\epsilon}^2(n)$, it turns out that $\gamma(n)$ increases with $n$ (and thus $\hat{D}(n)$ approximates $\text{E}\,\boldsymbol{\epsilon}^2(n)$ less effectively for larger $k$, which is consistent with the results of the almost-sure analysis). To see this, define again the moments (assumed finite):

$$\begin{aligned}\sigma_{\bar{u}}^2 &\triangleq \text{E}\bar{\boldsymbol{u}}^2, \quad \rho_{\bar{u}}^4 \triangleq \text{E}\bar{\boldsymbol{u}}^4\\ \xi_{\bar{u}}^6 &\triangleq \text{E}\bar{\boldsymbol{u}}^6, \quad \eta_{\bar{u}}^8 \triangleq \text{E}\bar{\boldsymbol{u}}^8,\end{aligned}$$

where $\bar{\boldsymbol{u}}$ denotes the normalized variable $\boldsymbol{u}/\sqrt{g(\boldsymbol{u})}$. Then

$$\begin{aligned}\text{E}\boldsymbol{\epsilon}^4(n) &= \left(\text{E}\boldsymbol{A}^2\right)^n \boldsymbol{\epsilon}^4(0)\\ &= \left(1 - 4\mu\sigma_{\bar{u}}^2 + 6\mu^2\rho_{\bar{u}}^4 - 4\mu^3\xi_{\bar{u}}^6 + \mu^4\eta_{\bar{u}}^8\right)^n \boldsymbol{\epsilon}^4(0).\end{aligned}$$

Define further the coefficients

$$\begin{aligned} r_4 &\triangleq \mathrm{E}\boldsymbol{A}^2 = \big(1 - 4\mu\sigma_{\bar{u}}^2 + 6\mu^2\rho_{\bar{u}}^4 - 4\mu^3\xi_{\bar{u}}^6 + \mu^4\eta_{\bar{u}}^8\big) \\ r_2 &\triangleq \mathrm{E}\boldsymbol{A} = \big(1 - 2\mu\sigma_{\bar{u}}^2 + \mu^2\rho_{\bar{u}}^4\big). \end{aligned} \tag{3.30}$$

It holds that $r_4 \geq r_2^2$ (with equality only if $\boldsymbol{u}(n)^2$ is a constant with probability 1). With these definitions, $\gamma(n)$ is given by

$$\gamma(n) = \frac{\sqrt{r_4^n - r_2^{2n}}}{r_2^n} = \sqrt{\frac{r_4^n}{r_2^{2n}} - 1}. \tag{3.31}$$

Therefore, except for the trivial case of a constant $\boldsymbol{u}(n)$, $\gamma(n)$ is strictly increasing, and thus

$$\lim_{n\to\infty} \gamma(n) = +\infty,$$

while $\gamma(0) = 0$ (due to the assumption of a deterministic $\boldsymbol{\epsilon}^2(0)$).

## 3.11 ASYMMETRY OF THE PROBABILITY DISTRIBUTION

The variance analysis in the previous section shows that that in the initial adaptation steps, $\boldsymbol{\epsilon}^2(n)$ tends to stay close to its mean, $\mathrm{E}\,\boldsymbol{\epsilon}^2(n)$, since the variance of $\boldsymbol{\epsilon}^2(n)$ is small. As time progresses, the variance grows, and one expects that $\boldsymbol{\epsilon}^2(n)$ will wander farther and farther away from its mean. In principle, this could mean that $\boldsymbol{\epsilon}^2(n)$ will assume equally likely large and small values. However, this is usually not the case. As time increases, $\boldsymbol{\epsilon}^2(n)$ assumes small values more often than large values, and its probability density function becomes more and more asymmetric!

To explain this behavior, return to (3.24) and rewrite it as follows

$$\ln\left(\frac{\boldsymbol{\epsilon}^2(n)}{\boldsymbol{\epsilon}^2(0)}\right) = \sum_{m=0}^{n-1} \ln \boldsymbol{A}(m). \tag{3.32}$$

Define also

$$\omega = \mathrm{E}\big(\ln \boldsymbol{A}(m)\big), \qquad \sigma^2 = \mathrm{E}\big(\ln \boldsymbol{A}(m) - \omega\big)^2,$$

which are constants since $\big\{\boldsymbol{u}(n)\big\}$ is assumed stationary. Assuming that both $\omega$ and $\sigma^2$ are finite, one can use the Central Limit Theorem [19] to conclude that, for $n \to \infty$,

$$\frac{1}{\sigma\sqrt{n}}\left[\ln\left(\frac{\boldsymbol{\epsilon}^2(n)}{\boldsymbol{\epsilon}^2(0)}\right) - n\omega\right] \sim N(0,1),$$

that is, the quantity on the left-hand side tends to a normal distribution with zero mean and unit variance. It then follows that, as $n$ increases, the distribution of $\boldsymbol{\epsilon}^2(n)$

can be well approximated by the following probability density function:

$$p_{\epsilon^2}(x) = \frac{1}{x\sigma\sqrt{2\pi n}} e^{-\frac{1}{2n\sigma^2}\left(\ln(x/\epsilon^2(0)) - n\omega\right)^2}, \quad x > 0.$$

Figure 3.8 shows $\omega$ and $\sigma^2$ for $\bar{\boldsymbol{u}}(n)$ uniformly distributed between $-0.5$ and $0.5$. Note the behavior similar to that seen in Fig. 3.7, where $\bar{\boldsymbol{u}}(n)$ is Gaussian. The next figures show $p_{\epsilon^2}(x)$ for several situations (in all cases, the vertical bar indicates the position of $\mathrm{E}\ln(\boldsymbol{\epsilon}^2(n)/\boldsymbol{\epsilon}^2(0))$).

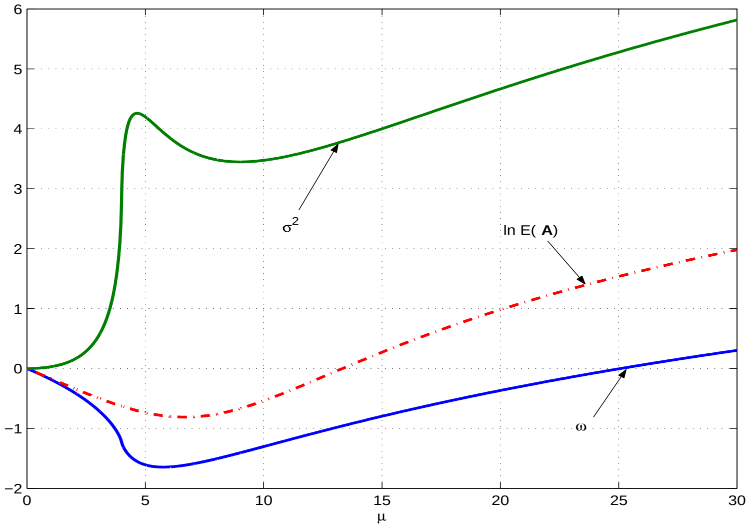

**Fig. 3.8** Graphs of $\omega = \mathrm{E}\ln\boldsymbol{A}$, $\sigma^2 = \mathrm{E}\big(\ln\boldsymbol{A} - \omega\big)^2$ and $\ln\mathrm{E}\boldsymbol{A}$ for $\bar{\boldsymbol{u}}(n)$ uniformly distributed between $-0.5$ and $0.5$.

The plots in Fig. 3.9 show the probability density function (pdf) for $\mu = 0.1$, $n = 10$ (left plot) and $n = 500$ (right plot). In this case, one has from Fig. 3.8 that $\omega = -1.679\ 10^{-2} \approx \ln\mathrm{E}\,\boldsymbol{A} = 1.668\ 10^{-2}$, and $\sigma^2 = 2.271\ 10^{-4}$. Since $\omega \approx \ln\mathrm{E}\,\boldsymbol{A}$ and $\sigma^2$ is small, one expects the learning curve to approximate well the behavior of a single run of the filter. This expectation is confirmed by the pdfs of $\boldsymbol{\epsilon}^2(10)$ and $\boldsymbol{\epsilon}^2(500)$, which show that $\boldsymbol{\epsilon}^2(n)$ tends to stay close to its mean.

On the other hand, one can see from Fig. 3.10 that for $\mu = 2.0$ the behavior is quite different (now one has $\omega = -0.4005$, $\ln\mathrm{E}\,\boldsymbol{A} = 0.3331$, and $\sigma^2 = 0.1521$. That is, $\omega$ is significantly larger than $\ln\mathrm{E}\,\boldsymbol{A}$, and the variance is large). Even for $n = 10$ the pdf of $\boldsymbol{\epsilon}^2(n)$ is already quite asymmetric, a characteristic that becomes more pronounced as $n$ increases. In this situation, $\boldsymbol{\epsilon}^2(n)$ is much more likely to be smaller, rather than larger, than its average.

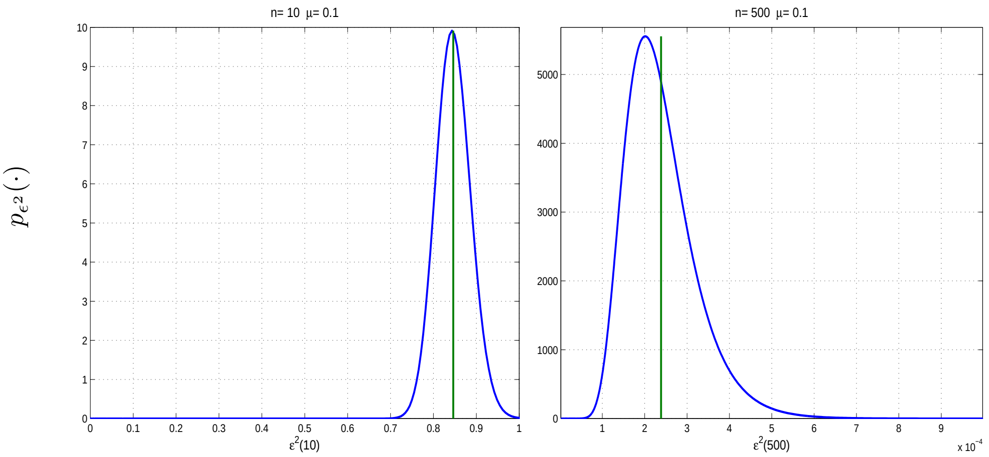

**Fig. 3.9** Left: Graph of $p_{\epsilon^2}(x)$ for $\bar{\boldsymbol{u}}(n)$ uniformly distributed between $-0.5$ and $0.5$, $\mu = 0.1$, $n = 10$. Right: Graph of $p_{\epsilon^2}(x)$ for $\bar{\boldsymbol{u}}(n)$ uniformly distributed between $-0.5$ and $0.5$, $\mu = 0.1$, $n = 500$.

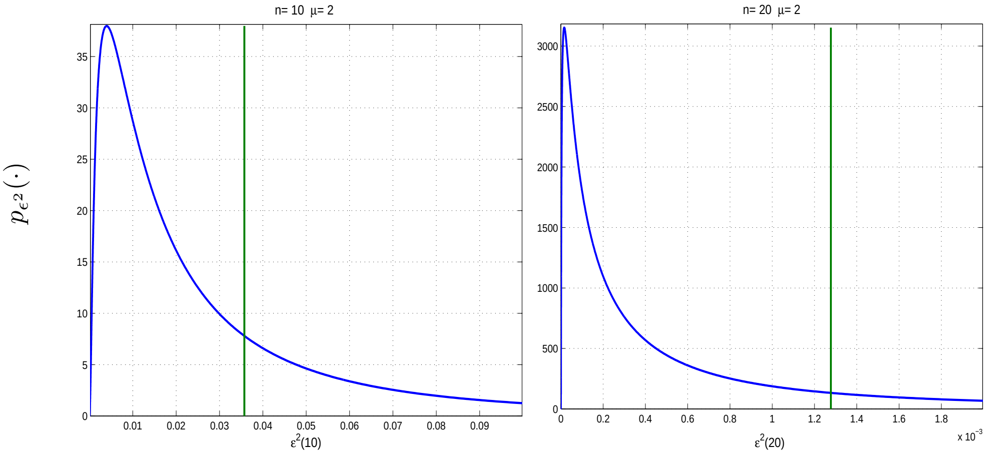

**Fig. 3.10** Left: Graph of $p_{\epsilon^2}(x)$ for $\bar{\boldsymbol{u}}(n)$ uniformly distributed between $-0.5$ and $0.5$, $\mu = 2.0$, $n = 10$. Right: Graph of $p_{\epsilon^2}(x)$ for $\bar{\boldsymbol{u}}(n)$ uniformly distributed between $-0.5$ and $0.5$, $\mu = 2.0$, $n = 20$.

## 3.12 TWO RATES OF CONVERGENCE

The above discussion can be used to compare the values of $\mathrm{E}\boldsymbol{\epsilon}^2(n)$ and $\hat{D}(n)$ when $n$ is fixed but $L$ is allowed to vary. Indeed, it follows from the expression for $\gamma'(n)$ that the larger the value of $L$ is, the smaller the value of $\gamma'(n)$ will be. Hence, the more experiments one averages, the closer the value of $\hat{D}(n)$ will be to that of $\mathrm{E}\,\boldsymbol{\epsilon}^2(n)$.

Another conclusion that follows from the almost-sure and variance analyses is that an adaptive filter recursion exhibits two different rates of convergence (even for single-tap adaptive filters). At first, for small $n$, a sample curve $\boldsymbol{\epsilon}^2(n)$ is close to $\mathrm{E}\boldsymbol{\epsilon}^2(n)$ and therefore converges at a rate that is determined by $\mathrm{E}\ln\boldsymbol{A}$. For larger $n$, the sample curve $\boldsymbol{\epsilon}^2(n)$ will converge at a rate that is determined by $\ln\mathrm{E}\boldsymbol{A}$.

A final remark: The knowledge that an adaptive filter is almost-sure convergent does *not* necessarily guarantee satisfactory performance! Thus assume that a filter is almost-sure stable but mean-square unstable. It follows from the earlier analysis that a sample error 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.

## 3.13 CONCLUDING REMARKS

This chapter provided an overview of two recent developments in the understanding of adaptive filter performance. One result pertains to an energy conservation relation that turns out to provide a convenient framework for the analysis of the performance of a wide range of adaptive filters (cf. [8]–[14] and [20]). A second result pertains to the learning abilities of adaptive filters, especially for larger step-sizes, where several interesting phenomena arise that seem to indicate that the learning behavior of adaptive filters is more sophisticated than was originally thought. More details can be found in [1], such as extensions of the arguments to the vector case.

## REFERENCES

1. V. H. Nascimento and A. H. Sayed, “On the learning mechanism of adaptive filters,” *IEEE Trans. Signal Process.*, **48**(6), p. 1609 (2000).

2. J. E. Mazo, “On the independence theory of equalizer convergence,” *The Bell System Technical Journal*, **58**, p. 963 (1979).

3. O. Macchi and E. Eweda, “Second-order convergence analysis of stochastic adaptive linear filtering,” *IEEE Trans. Automatic Control*, **28**(1), p. 76 (1983).

4. A. Feuer and E. Weinstein, “Convergence analysis of LMS filters with uncorrelated Gaussian data,” *IEEE Trans. Acoust. Speech Signal Process.*, **33**(1), p. 222 (1985).

5. V. Solo and X. Kong, *Adaptive Signal Processing Algorithms,* Prentice Hall, NJ, 1995.

6. H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer, 1997.

7. H. J. Butterweck, “A wave theory of long adaptive filters,” *IEEE Trans. Circuits and Systems–I*, **48**, p. 739 (2001).

8. A. H. Sayed and M. Rupp, “A time-domain feedback analysis of adaptive algorithms via the small gain theorem,” *Proc. SPIE*, **2563**, p. 458, San Diego, CA (1995).

9. A. H. Sayed and M. Rupp, “Robustness issues in adaptive filtering,” in *DSP Handbook*, Chapter **20**, CRC Press, 1998.

10. J. Mai and A. H. Sayed, “A feedback approach to the steady-state performance of fractionally-spaced blind adaptive equalizers,” *IEEE Trans. Signal Process.*, **48**(1), p. 80 (2000).

11. N. R. Yousef and A. H. Sayed, “A unified approach to the steady-state and tracking analyses of adaptive filters,” *IEEE Trans. Signal Process.*, **49**(2), p. 314 (2001).

12. M. Rupp and A. H. Sayed, “A time-domain feedback analysis of filtered-error adaptive gradient algorithms,” *IEEE Trans. Signal Process.*, **44**(6), p. 1428 (1996).

13. A. H. Sayed and T. Y. Al-Naffouri, “Mean-square analysis of normalized leaky adaptive filters,” *Proc. ICASSP*, **6**, p. 3873, Salt Lake City, Utah, 2001.

14. T. Y. Al-Naffouri and A. H. Sayed, “Transient analysis of data-normalized adaptive filters,” *IEEE Trans. Signal Process.*, **51**(3), pp. 639–652, March 2003.

15. R. R. Bitmead and B. D. O. Anderson, “Adaptive frequency sampling filters,” *IEEE Trans. Circuits and Systems*, **28**(6), p. 524 (1981).

16. R. R. Bitmead, B. D. O. Anderson, and T. S. Ng, “Convergence rate determination for gradient-based adaptive estimators,” *Automatica*, **22**, p. 185 (1986).

17. H. J. Kushner and F. J. Vázquez-Abad, “Stochastic approximation methods for systems over an infinite horizon,” *SIAM Journal of Control and Optimization*, **34**(2), p. 712 (1996).

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

19. D. Williams, *Probability with Martingales*, Cambridge University Press, 2000.

20. A. H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley, NY, 2003.