Proc. 36th Annual Allerton Conference on Communication, Control, and Computing,

pp. 242–251, Allerton, IL, Sep. 1998.

Stability of the LMS Adaptive Filter by Means of a State Equation*

Vítor H. Nascimento and Ali H. Sayed

Electrical Engineering Department
University of California
Los Angeles, CA 90095

Abstract

This work studies the mean-square stability of stochastic gradient algorithms without resorting to slow adaptation approximations or to the widely used, yet rarely applicable, independence assumptions. This is achieved by reducing the study of the mean-square convergence of an adaptive filter to the study of the exponential stability of a linear time-invariant state equation. The size of the coefficient matrix of the state equation, however, turns out to grow exponentially fast with the length of the filter so that it becomes computationally infeasible to manipulate the matrix directly. It is instead shown that the coefficient matrix is sparse and has structure. By exploiting these two properties, and by applying a sequence of carefully chosen similarity transformations to the coefficient matrix, an upper bound on the step-size is found that guarantees stability.

1 Introduction

Consider the following estimation problem. Given a zero-mean scalar sequence $\{y(k)\}_{k=0}^{\infty}$ and another sequence of zero-mean length-M column vectors $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$, we seek a length-$M$ column vector $\{\boldsymbol{w}_*\}$ that solves

$$\min_{\boldsymbol{w}} \mathrm{E}\big(y(k) - \boldsymbol{x}_k^T \boldsymbol{w}\big)^2.$$

where E is the expectation operator, k is the time index, and ^{T} is the transposition operator. The sequence $\{y(k)\}$ is called the *desired* sequence and $\{\boldsymbol{x}_k\}$ is the *input* or *regressor* sequence; the sequences $\{y(k), \boldsymbol{x}(k)\}$ are assumed to be jointly stationary. The error sequence is defined by $v(k) = y(k) - \boldsymbol{x}_k^T \boldsymbol{w}_*$ and its variance (also the minimum cost) is denoted by σ_v^2; we also refer to $\{v(k)\}$ as the *noise* sequence.

*This material was based on work supported in part by the National Science Foundation under awards MIP-9796147 and CCR-9732376. The work of V. H. Nascimento was also supported by a fellowship from Conselho Nacional de Pesquisa e Desenvolvimento - CNPq - Brazil, while on leave from LPS-DEE, Escola Politécnica da Universidade de São Paulo, Brazil.

The optimal solution $\boldsymbol{w}_*$ can be fully characterized in terms of the second-order statistics of the random processes $\{y(k), \boldsymbol{x}(k)\}$. In adaptive implementations of the estimator $\boldsymbol{w}_*$, however, these statistics are often replaced by instantaneous approximations. In particular, the famed LMS algorithm (e.g., [1, 2, 3]) computes successive approximations, $\boldsymbol{w}_k$, to $\boldsymbol{w}_*$ through the recurrence relation

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \mu \boldsymbol{x}_k \big(y(k) - \boldsymbol{x}_k^T \boldsymbol{w}_k\big), \quad \text{with initial condition } \boldsymbol{w}_0 \ , \tag{1}$$

where the design parameter μ in (1) is known as the *step-size*. Two important measures of performance for the LMS algorithm are the errors

$$\tilde{\boldsymbol{w}}_k \triangleq \boldsymbol{w}_* - \boldsymbol{w}_k, \qquad\qquad e(k) \triangleq y(k) - \boldsymbol{x}_k^T \boldsymbol{w}_k. \tag{2}$$

Ideally, we would like the adaptive algorithm (1) to reduce $e(k)^2$ from its initial value, and to keep $\mathrm{E}\, e(k)^2$ close to σ_v^2 in steady-state (or, equivalently, $\tilde{\boldsymbol{w}}_k$ close to $\mathbf{0}$).

1.1 Performance with Slow Adaptation

The performance of an adaptive filter is often measured in terms of the variance $\mathrm{E}\, e(k)^2$, also known as the *mean-square error* or MSE for short, and by the trace of the covariance matrix of $\tilde{\boldsymbol{w}}_k$, $\mathrm{Tr}(\mathrm{E}\, \tilde{\boldsymbol{w}}_k \tilde{\boldsymbol{w}}_k^T)$, also called the *mean-square deviation* or MSD. Exact expressions for these measures can be hard to obtain in a general setting, which is in part due to the fact that adaptive systems are naturally time-variant and nonlinear. Nevertheless, there is a vast literature on the analysis of the performance of the LMS algorithm and its variants, which have been motivated by the large range of applications in which these algorithms have been successful. In most of these analyses, simplifying assumptions on the data and the model are imposed in order to make the derivation and the results more tractable.

The most important of these simplifying conditions are known collectively as the *independence assumptions*, which appear in many of the earlier works on LMS and in several recent ones (see, e.g., [1, 4, 5, 6]). Basically, one assumes the following.

I-1. The zero-mean sequences $\big\{y(k),\, \boldsymbol{x}_k\big\}$ are related via a linear model of the form

$$y(k) = \boldsymbol{x}_k^T \boldsymbol{w}_* + v(k) \tag{3}$$

for some unknown $\boldsymbol{w}_*$, and where $v(k)$ is zero-mean with variance σ_v^2. In addition, it is assumed that the sequences $\big\{\boldsymbol{x}_k\big\}$ and $\big\{v(k)\big\}$ are independent and identically distributed, and that they are also independent of each other.

In applications, the assumption of independent regressor vectors is seldom satisfied. For example, in channel equalization, the vectors $\boldsymbol{x}_k$ are formed from a delay line (see (6)), thus $\boldsymbol{x}_n$ shares all but one of the elements of $\boldsymbol{x}_{n+1}$ so that $\{\boldsymbol{x}_n, \boldsymbol{x}_{n+1}\}$ are clearly dependent.

What makes the results obtained with the independence analysis still useful is a result first obtained in [7], showing that when the step-size is infinitesimally small ($\mu \approx 0$), the results obtained using the independence assumptions are good approximations for the actual performance of the LMS algorithm. The conclusions of [7] were later extended to more general settings (but still restricted to the LMS algorithm) in [8, 9, 10]. Similar results for other adaptive algorithms can be obtained using averaging theory and/or the ODE method (e.g., [11, 12, 13, 14]).

1.2 Performance with Faster Adaptation

The above works give a good understanding of the behavior of the LMS algorithm when the step-size μ is sufficiently small (without in fact quantifying how small μ should be). However, from the recursion (1) one can deduce that the rate of convergence of the algorithm is greatly affected by the choice of the step-size. An infinitesimal step-size ($\mu \approx 0$) implies that the weight estimates $\boldsymbol{w}_k$ change very slowly at each iteration, and consequently the convergence rate is small. This can be particularly annoying in nonstationary environments, where very slow convergence may not allow the algorithm to properly track time variations in signal statistics. A designer might then wish to employ larger step-sizes to improve the convergence speed of the algorithm, especially during the initial convergence phase (before steady-state is achieved).

The following questions are therefore relevant and remain largely unanswered.

(i) How small must the step-size be so that the independence-based approximations are still reasonable? Also, for a given value of the step-size, what is the order of magnitude of the error incurred by using these approximations ?

(ii) What is the real performance of the adaptive algorithm when the step-size is not small ?

(iii) How large the step-size can get without compromising filter (mean-square) stability?

(iv) What is the step-size that gives the fastest convergence rate ?

For step-sizes that are not infinitesimally small, there are essentially no results in the literature that predict or confirm the behavior/stability of the LMS algorithm and its variants (see, e.g., the statement in [5] regarding this issue).

The purpose of this paper is to propose a method to study the stability of the LMS algorithm, without relying on the independence assumptions and without assuming beforehand that the step-size is vanishingly small. Since an exact expression for the largest step-size (say, $\mu_{\max}$) that addresses point (iii) above is difficult to obtain, we instead derive an upper bound on how large the step-size μ can be for mean-square stability (say, $\mu < \bar{\mu}$). While the bound $\bar{\mu}$ is not tight (i.e., close to $\mu_{\max}$) at this stage of our analysis, it is, to the authors' knowledge, the first such bound and is also applicable to a generic distribution of the input sequence (and in particular, it even allows for a normally distributed input). The significance of this work is therefore in developing a framework that studies filter stability for step-sizes that are not necessarily infinitesimally small. We are currently pursuing extensions of our analysis in order to obtain tighter upper bounds and to handle larger step-sizes.

Our discussion builds upon an approach originally suggested in [15], and which will lead naturally to a state-space framework. Basically, the arguments we employ in the future sections can be summarized as follows. We first find a dynamic state-space model for the evolution of the covariance matrix $\mathrm{E}\,\tilde{\boldsymbol{w}}_k\tilde{\boldsymbol{w}}_k^T$; the states of this model will consist of the entries of the covariance matrix in addition to several other relevant quantities. The state equation will be of the form

$$\boldsymbol{\Gamma}_{k+1} = \Phi\boldsymbol{\Gamma}_k + \boldsymbol{b}, \tag{4}$$

where $\boldsymbol{b}$ is a constant vector, $\boldsymbol{\Gamma}_k$ is the state vector, and Φ is a constant matrix. With this model, the largest step-size ($\mu_{\max}$) that guarantees stable performance of the LMS

filter (and therefore answers point (iii) above) will be the largest μ for which Φ is still a stable matrix, i.e.,

$$\mu_{\max} \stackrel{\Delta}{=} \sup\{\mu \ \text{ such that } \rho(\Phi) < 1\} \ , \tag{5}$$

where $\rho(\Phi)$ denotes the *spectral radius* of Φ, i.e., $\rho(\Phi) = \max_i \big|\lambda_i\big(\Phi\big)\big|$.

Unfortunately, determining $\mu_{\max}$ is not a trivial task for two main reasons. First, the eigenvalues of the matrix Φ depend nonlinearly on the step size μ and, secondly, the dimension of Φ grows extremely fast with the filter length (for example, for $M = 6$ the matrix has size $28,181 \times 28,181$). It is therefore computationally infeasible to work directly with Φ; the approach is feasible only for relatively small filter lengths. For this reason, reference [15] considered only the case $M = 2$ (i.e., filter with two taps), while reference [16] used the same method for orders up to $M = 6$ coupled with a numerical procedure (viz., the power method) for the evaluation of the eigenvalues of Φ. For larger filter lengths, we need to develop an alternative procedure for the estimation of $\mu_{\max}$ that does not work directly with the matrix Φ.

The approach we propose in this paper is based on the observation that the matrix Φ, although of large dimensions, is both *sparse* and *structured*. These two properties combined can be used to derive a bound on the step-size for stable performance. [Moreover, the bound will be such that it is not a function of the maximum value that $\|\boldsymbol{x}_k\|$ can attain; the result will depend only on the distribution of the input sequence, and on average quantities.]

2 Structure of the State-Space Model

In the sequel we assume that the regressors $\big\{\boldsymbol{x}_k\big\}$ arise from a tap-delay line, say

$$\boldsymbol{x}_k = \big[a(k-M+1) \quad a(k-M+2) \quad \ldots \quad a(k)\big]^T . \tag{6}$$

where the input sequence $\big\{a(k)\big\}$ is assumed iid, with zero mean, and moments $\sigma_p = \mathrm{E}\, a(k)^p$, for $p \geq 1$. The assumption of iid $\big\{a(k)\big\}$ implies that the variable $a(k)$ is independent of the current weight error vector, $\tilde{\boldsymbol{w}}_k$. We also assume that $\big\{y(k),\, \boldsymbol{x}_k\big\}$ are related via a linear model of the form (3), with a zero-mean iid noise sequence $v(k)$ of variance σ_v^2 and that is independent of the input sequence. Using the LMS update (1) and the model (3), we find that the error equation for LMS is given by

$$\tilde{\boldsymbol{w}}_{k+1} = (I - \mu \boldsymbol{x}_k \boldsymbol{x}_k^T)\tilde{\boldsymbol{w}}_k - \mu \boldsymbol{x}_k v(k) \ . \tag{7}$$

We are interested in conditions under which the MSD is bounded (i.e., conditions under which $\mathrm{Tr}\big(\mathrm{E}\, \tilde{\boldsymbol{w}}_k \tilde{\boldsymbol{w}}_k^T\big)$ forms a bounded sequence). This requires that we study the stability of the matrix Φ in (4) and determine conditions under which its eigenvalues are strictly inside the unit disc. It turns out that, under the above conditions, the noise sequence $\big\{v(k)\big\}$ does not influence the stability of the recursion (4) since it does not enter Φ and only affects the driving term $\boldsymbol{b}$. For this reason, in the remainder of the paper we can assume $v(k) \equiv 0$.

2.1 Obtaining the Linear Model: An Example

We present briefly the main steps of the derivation of the state-space model (4) for $M = 2$. Similar arguments hold for larger values of M.

Recall that our purpose is to determine a recursion that describes the evolution of the entries of the covariance matrix $\mathrm{E}\,\tilde{\boldsymbol{w}}_k\tilde{\boldsymbol{w}}_k^T$ (whose trace determines the MSD). We denote these entries by

$$\mathrm{E}\,\tilde{\boldsymbol{w}}_k\tilde{\boldsymbol{w}}_k^T \triangleq \begin{bmatrix} \gamma_1(k) & \gamma_3(k) \\ \gamma_3(k) & \gamma_2(k) \end{bmatrix}.$$

The variables $\{\gamma_i(k)\}$ will be some of the elements in the state vector $\boldsymbol{\Gamma}_k$. This is because recursions for the $\{\gamma_i(k)\}$ will require that we also propagate other quantities. We do not show how to find all the variables in $\boldsymbol{\Gamma}_k$ here. This example only explains why other variables are necessary, and shows how to obtain a recursion for one of them. Similar derivations can be used for the other variables.

Let $\{\tilde{w}_{k,1}, \tilde{w}_{k,2}\}$ denote the entries of the error $\tilde{\boldsymbol{w}}_k$, i.e., $\tilde{\boldsymbol{w}}_k = \begin{bmatrix} \tilde{w}_{k,1} & \tilde{w}_{k,2} \end{bmatrix}^T$. Now using the error equation (7), and the the assumptions on $a(k)$, we find that

$$\gamma_1(k+1) = \gamma_1(k) - 2\mu\,\mathrm{E}\big(a(k-1)^2\tilde{w}_{k,1}^2\big) + \mu^2\,\mathrm{E}\big(a(k-1)^4\tilde{w}_{k,1}^2\big) + \mu^2\sigma_2\,\mathrm{E}\big(a(k-1)^2\tilde{w}_{k,2}^2\big).$$

The right-hand side of this recursion depends on the additional terms

$$\mathrm{E}\big(a(k-1)^2\tilde{w}_{k,1}^2\big), \quad \mathrm{E}\big(a(k-1)^4\tilde{w}_{k,1}^2\big), \quad \mathrm{E}\big(a(k-1)^2\tilde{w}_{k,2}^2\big), \quad \text{and} \quad \mathrm{E}\big(a(k-1)^2\tilde{w}_{k,1}\tilde{w}_{k,2}\big).$$

Hence, a complete recursion for $\gamma_1(k)$ requires that we also determine recursions for these quantities. If we denote the above terms in succession by $\{\gamma_4(k), \gamma_5(k), \gamma_6(k), \gamma_7(k)\}$, then, for example, the recursion that results for $\gamma_4(k+1)$ will be

$$\gamma_4(k+1) = \sigma_2\gamma_1(k) - 2\mu\sigma_2\gamma_4(k) + \sigma_2\gamma_5(k) + \mu^2\sigma_4\gamma_6(k).$$

The state vector in this case will therefore be of dimension 7, $\boldsymbol{\Gamma}_k = \begin{bmatrix} \gamma_1(k) & \cdots & \gamma_7(k) \end{bmatrix}^T$. This procedure can in principle be repeated for any filter order M (and, in fact, as shown in [16], similar state-space models can be obtained even in situations where the $a(k)$ form a correlated sequence). However, the number of state variables will grow exponentially fast with the filter order, which makes it infeasible to pursue this line of reasoning for larger filter lengths (larger than 7 for example). Instead we focus on the structure of the coefficient matrix Φ.

2.2 Structure of Φ

As we mentioned above, a major drawback of the state-space model (4) is that its order (the size of Φ) grows exponentially fast with the filter length M. In this section, we highlight several structural properties of Φ and use them in the next section to study the stability of the LMS recursion without having to form any large matrices.

Indeed, it can be shown that Φ is highly sparse and has considerable structure. The sparseness of Φ was used in [16] to obtain approximations for its largest eigenvalue; however, the exponential growth of Φ limits the use of this technique to filters of smaller orders (say up to 6 or 7). To work with larger filter lengths, it is necessary to study the structure of Φ with more detail. It can be shown that Φ has the following properties (some are evident from the case $M = 2$ above, others only arise for larger values of M):

1. The number of nonzero elements in each row of Φ does not exceed $(M+1)^2$.

2. The number of rows of Φ is finite for finite M.

3. It is possible to select a small number of rows of Φ (no more than $16M-12$ different rows) such that *all other rows* of Φ are *permutations* of these rows. Consider the following *contrived* example,

$$\Phi = \begin{bmatrix} 1 & -2\mu & 0 & 0 & \mu^2\sigma_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu^2\sigma_4 & 0 & 0 & 0 & -2\mu \\ \sigma_2 & 0 & -2\mu\sigma_2 & 0 & \mu^2\sigma_6 & 0 & 0 & 0 & 0 \\ 0 & -2\mu\sigma_2 & 0 & 0 & 0 & 0 & \mu^2\sigma_6 & \sigma_2 & 0 \\ 0 & 1 & 0 & -2\mu & 0 & \mu^2\sigma_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2\mu & \mu^2\sigma_4 & 0 \end{bmatrix}$$

The matrix Φ above has two different classes of rows. Rows numbers 1, 2, 5, and 6 form one class, and rows 3 and 4 form the other. In each class, the entries appearing in all rows are the same, although the order is different. For example, the nonzero entries in row 1 are $\mathcal{A}_1 = \{1, -2\mu, \mu^2\sigma_4\}$. The nonzero entries of rows 2, 5, and 6 also belong to $\mathcal{A}_1$. Similarly, the elements of rows 3 and 4 belong to the set $\mathcal{A}_2 = \{\sigma_2, -2\mu\sigma_2, \mu^2\sigma_6\}$. We name the sets $\mathcal{A}_1$ and $\mathcal{A}_2$ the *coefficient sets* of Φ, and we say that any two rows belonging to the same class are *p-equivalent* (i.e., they are permutations of each other).

The real matrix Φ has a similar structure; all its rows take their nonzero entries from one of (at most) $16M-12$ different coefficient sets. The coefficient sets of Φ are listed below for $M > 2$ (the integer p lies in the interval $1 \le p \le 4(M-1)$ – we can also specify the number of times each term will appear on a given row):

$$\begin{aligned}
\mathcal{A}_1 &= \Big\{1,\ \ -2\mu,\ \ \mu^2\sigma_2,\ \ 2\mu^2,\ \ \mu^2,\ \ -2\mu\sigma_1,\ \ 2\mu^2\sigma_1\Big\}, \\
\mathcal{A}_2 &= \Big\{1-2\mu\sigma_2+\mu^2\sigma_4,\ \ -2\mu\big(\sigma_1-\mu\sigma_3\big),\ \ \mu^2\sigma_2\Big\} \\
\mathcal{A}_3 &= \Big\{1,\ \ -\mu,\ \ \mu^2,\ \ -\mu\sigma_1,\ \ 2\mu^2\sigma_1,\ \ \mu^2\sigma_2\Big\} \\
\mathcal{A}_4 &= \Big\{1-\mu\sigma_2,\ \ -\mu(1-2\mu\sigma_2),\ \ -\mu\sigma_1,\ \ \mu^2\sigma_1,\ \ \mu^2\sigma_3\Big\} \\
\mathcal{A}_{1,p} &= \Big\{\sigma_p,\ \ -2\mu\sigma_p,\ \ \mu^2\sigma_{p+2},\ \ 2\mu^2\sigma_p,\ \ \mu^2\sigma_p,\ \ -2\mu\sigma_{p+1},\ \ 2\mu^2\sigma_{p+1}\Big\} \\
\mathcal{A}_{2,p} &= \Big\{\sigma_p-2\mu\sigma_{p+2}+\mu^2\sigma_{p+4},\ \ -2\mu(\sigma_{p+1}-\sigma_{p+3}),\ \ \mu^2\sigma_{p+2}\Big\} \\
\mathcal{A}_{3,p} &= \Big\{\sigma_p,\ \ -\mu\sigma_p,\ \ \mu^2\sigma_p,\ \ -\mu\sigma_{p+1},\ \ 2\mu^2\sigma_{p+1},\ \ \mu^2\sigma_{p+2}\Big\} \\
\mathcal{A}_{4,p} &= \Big\{\sigma_p-\mu\sigma_{p+2},\ \ -\mu(\sigma_p-2\mu\sigma_{p+2}),\ \ -\mu\sigma_{p+1},\ \ \mu^2\sigma_{p+1},\ \ \mu^2\sigma_{p+3}\Big\}\ .
\end{aligned}$$

4. Those rows of Φ that have unity entries can in fact have only a single unity entry. Moreover, some of the rows of Φ have ones on the main diagonal and these are generated only by the coefficient sets $\{\mathcal{A}_1, \mathcal{A}_3\}$. For each such row with a unity entry on the main diagonal, there exists a single row from either sets $\{\mathcal{A}_{1,2}, \mathcal{A}_{3,2}\}$ (i.e., with $p = 2$) with entries in the same column positions and which can be obtained from it by the following substitutions (compare the elements in the sets $\mathcal{A}_1$ and $\mathcal{A}_{1,p}$, as well as $\mathcal{A}_3$ and $\mathcal{A}_{3,p}$):

$$1 \leftarrow \sigma_2\ ,\ \ \sigma_1 \leftarrow \sigma_3\ ,\ \ \sigma_2 \leftarrow\ \sigma_4\ .$$

5. Pick a row with unity on its main diagonal and the corresponding row with the σ_2 entry in the same column position as the one on the diagonal. We can always

associate with these two rows, by means of suitable permutations, a square block whose size can vary from 2 up to M and which has the following generic form (shown here for sizes 4 and 3, and denoted by B_4 and B_3, respectively):

$$B_4 \triangleq \begin{bmatrix} 1 & -2\mu & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \sigma_2 & -2\mu\sigma_2 & 0 & 0 \end{bmatrix}, \quad B_3 \triangleq \begin{bmatrix} 1 & -2\mu & 0 \\ 0 & 0 & 1 \\ \sigma_2 & -2\mu\sigma_2 & 0 \end{bmatrix}. \tag{8}$$

The relevant characteristic of each block is the position of the 1's: one is on the main diagonal while the others are on the upper diagonal.

3 Stability of the LMS Filter

Now that several properties of Φ have been highlighted, we return to the problem of deriving a bound for the step-size that guarantees a uniformly bounded covariance matrix $\mathrm{E}\,\tilde{\boldsymbol{w}}_k\tilde{\boldsymbol{w}}_k^T$ (*cf.* (5)).

As mentioned before, the eigenvalues of Φ depend nonlinearly on μ and, hence, it is mathematically intractable to derive an expression for the spectral radius of Φ in terms of μ. For this reason, we shall proceed via an alternative route. For any $M \times M$ square matrix A, it holds that its spectral radius is upper bounded by its ∞-norm [18], i.e., $\rho(A) \leq \|A\|_\infty$, where $\|A\|_\infty$ is the maximum absolute row sum of A. Therefore, if we can find a μ that guarantees $\|\Phi(\mu)\|_\infty < 1$, then it also guarantees $\rho\big(\Phi(\mu)\big) < 1$.

We mentioned in the previous section that the rows of Φ belong to $16M - 12$ classes of p-equivalent rows, and that each row has only $O(M^2)$ nonzero entries. Hence, in principle, $\|\Phi\|_\infty$ can be evaluated with little effort. Unfortunately, however, several rows of Φ contain the element $\{1\}$, which makes $\|\Phi\|_\infty$ always larger than 1 regardless of μ.

To overcome this difficulty, we propose to work with a modified matrix Φ that has the same spectral radius as the original Φ but smaller ∞-norm. We achieve this by showing how to construct a similarity transformation T such that there exists a $\bar{\mu} > 0$ satisfying

$$\|T^{-1}\Phi(\mu)T\|_\infty < 1 \quad \text{for all} \quad \mu < \bar{\mu}. \tag{9}$$

This is useful because, in view of the fact that similarity transformations preserve eigenvalues, the above $\bar{\mu}$ will also guarantee $\rho\big(\Phi(\mu)\big) < 1$ for all $\mu < \bar{\mu}$.

Special care must be taken while constructing the similarity transformation in order to preserve as much structure as possible and in order to keep the ∞-norm of $T^{-1}\Phi_M(\mu)T$ easily computable. We cannot reproduce here all the details involved in the construction procedure. We only highlight the main steps.

We construct T as a sequence of elementary similarity transformations T_k. Each one is defined once the special blocks B_k that we mentioned in (8) above have been identified. So consider a generic block B_k, which we regard as the leading block in a submatrix with k rows and multiple columns, say (for a block of size 4)

$$\left[\begin{array}{cccc|cccccc} \boxed{1} & -2\mu & 0 & 0 & -2\mu & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & \boxed{1} & 0 & 0 & -2\mu & -2\mu & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \boxed{1} & 0 & 0 & 0 & 0 & -2\mu & -2\mu & \dots \\ \sigma_2 & -2\mu\sigma_2 & 0 & 0 & -2\mu\sigma_2 & 0 & 0 & 0 & 0 & 0 & \dots \end{array}\right]$$

Recall that each block B_k has a row with a 1 on the main diagonal and $k-2$ rows with a 1 in the upper diagonal. Our task is to replace all the 1's with smaller terms. This can be achieved by a similarity transformation T_k that is defined as follows:

1) Rows that have a 1 on the main diagonal are replaced by rows that have $1-2\mu\sigma_2$ on the diagonal. All other elements are $O(\mu^2)$.

2) Rows that have a 1 in off-diagonal positions are replaced by rows that have $O(\mu^{1/3})$ on these positions. All other elements will be $O(\mu^{2/3})$.

For example, the above matrix will become after the transformation T_k,

$$\left[\begin{array}{cccc|cccccc} \boxed{1-2\mu\sigma_2} & 0 & 0 & 0 & -4\mu^2\sigma_2 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & \boxed{\mu^{1/3}} & 0 & 0 & -2\mu^{2/3} & -2\mu^{2/3} & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & \boxed{\mu^{1/3}} & 0 & 0 & 0 & 0 & -2\mu & -2\mu & \ldots \\ \mu^{-1/3}\sigma_2 & 0 & 2\mu^{2/3}\sigma_2 & 0 & & 0 & 0 & 0 & 0 & 0 & \ldots \end{array}\right]$$

Once all such transformations are carried out, the absolute row sum of each row will be strictly less than one. It then becomes possible to compute an upper bound $\bar{\mu} \leq \mu_{\max}$ by seeking the largest value of μ that results in an $\infty-$norm for the transformed matrix $\|T^{-1}\Phi(\bar{\mu})T\|_\infty$ that is less than one. This step leads to the following optimization problem (with $1 \leq p \leq 4(M-1)$),

$$\bar{\mu} \;=\; \max\left(\mu_a \mu_f\right) \tag{10}$$

subject to the scalars $\{\mu_f, \mu_a\}$ satisfying the inequalities (11) to (16) below:

$$\begin{aligned} \left|1-2\mu_f\mu_a\sigma_2\right| + (M-1)^2\mu_f^2\left|1-2\mu_f\mu_a\sigma_2\right| + \mu_f^2\left|\mu_a\sigma_2 - 2\mu_f\mu_a^2\sigma_4\right| + \\ +2(M-2)(M-1)^2\mu_f^2 + 2(M-2)(M+1)\mu_f^2\mu_a\sigma_2 + (M-2)(M-1)^3\mu_f^3 + \\ +5(M-2)(M-1)\mu_f^3\mu_a\sigma_2 + 2(M-2)\mu_f^3\mu_a^{3/2}\sigma_3 < 1, \end{aligned} \tag{11}$$

$$\begin{aligned} \left|1-2\mu_f\mu_a\sigma_2\right| + 2(M-1)(M^2-3M+1)\mu_f^2 + \\ +2(M^2-M-3)\mu_f^2\mu_a\sigma_2 + (M-1)^2(M^2-3M+1)\mu_f^3 + \\ +(5M^2-15M+9)\mu_f^3\mu_a\sigma_2 + 2(M-2)\mu_f^3\mu_a^{3/2}\sigma_3 < 1. \end{aligned} \tag{12}$$

$$\mu_f^{\frac{1}{3(M-2)}} + 2(M-1)\mu_f^{\frac{2}{3}} + (M-1)^2\mu_f^{\frac{5}{6}} + \mu_f^{\frac{5}{6}}\mu_a\sigma_2 < 1, \tag{13}$$

$$\left|1-2\mu_f\mu_a\sigma_2 + \mu_f^2\mu_a^2\sigma_4\right| + (M-1)^2\mu_f^2\mu_a\sigma_2 + 2(M-1)\mu_f^2\mu_a^{3/2}\sigma_3 < 1, \tag{14}$$

$$\begin{aligned} \mu_a^{p/2}\sigma_p + (M^2-3M+2)\mu_f\mu_a^{p/2}\sigma_p + (M-1)^2\mu_f^2\mu_a^{p/2}\sigma_p + \\ +2\mu_f\mu_a^{(p+1)/2}\sigma_{p+1} + 2(M-1)\mu_f^2\mu_a^{(p+1)/2}\sigma_{p+1} + \mu_f^2\mu_a^{(p+2)/2}\sigma_{p+2} < \mu_f^{\frac{1}{3}}, \end{aligned} \tag{15}$$

$$\begin{aligned} \left|\mu_a^{p/2}\sigma_p - 2\mu_f\mu_a^{(p+2)/2}\sigma_{p+2} + \mu_f^2\mu_a^{(p+4)/2}\sigma_{p+4}\right| + (M-1)^2\mu_f^2\mu_a^{(p+2)/2}\sigma_{p+2} + \\ +2(M-1)\mu_f^2\mu_a^{(p+3)/2}\sigma_{p+3} < 1 \end{aligned} \tag{16}$$

The solid curve in Figure 1 shows a plot of the solution $\bar{\mu}$ of (10) for a sequence $\{a(k)\}$ that is normally distributed with $\sigma_2 = 0.01$. The rate of decay of $\bar{\mu}$ is dependent on the signal distribution and is proportional to $1/M^5$ in this example, as indicated by the broken line.

While a tighter bound for $\bar{\mu}$ that decays slower with M can be obtained by careful selection of more specialized similarity transformations T_k, the purpose of this work has been to suggest a framework for the derivation of such bounds and for the stability analysis of adaptive schemes without prior assumption of vanishingly small step-sizes.

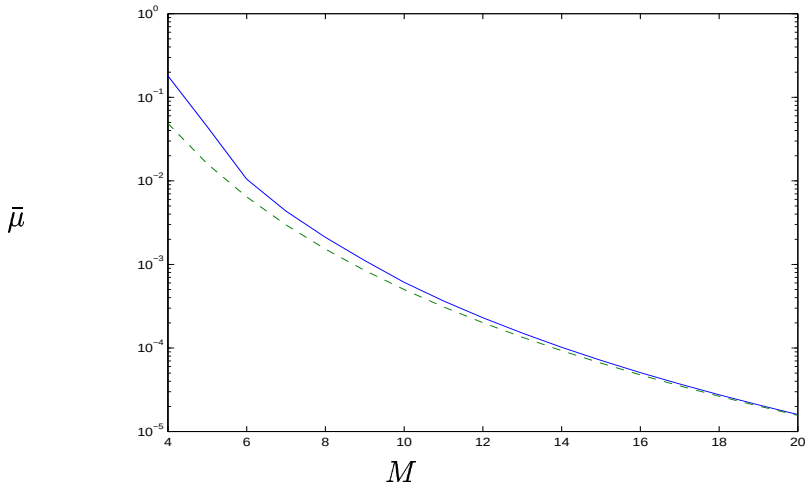

Figure 1: A plot of $\bar{\mu}$ versus filter size M for a Gaussian distribution.

4 Some Concluding Remarks

There are still several issues that remain to be addressed. These include weakening the assumptions on $\{a(k)\}$ (such as allowing for correlated input sequence), obtaining a more relaxed bound on the step-size (one that decreases slower with M), and weakening the condition on the noise $\{v(k)\}$. These issues are currently under investigation.

Another issue of interest and which is discussed in the companion article [17] is the following. Once an analysis is performed on the expected behavior of an adaptive scheme (as we did above), it is common in practice to confirm the results by means of simulations by generating so-called ensemble-average learning curves. These curves are usually obtained by averaging over no more than 100-200 repeated experiments and they provide approximations for the evolution of the error measures $\mathrm{E}\, e(k)^2$ or $\mathrm{E}\, \|\tilde{\boldsymbol{w}}_k\|^2$ as a function of time. It has been observed in practice that these averaged curves tend to match theoretical results reasonably well for sufficiently small step-sizes and that, therefore, they tend to provide a good approximation for the expected performance of an adaptive filter.

For larger step-sizes, however, we observed that care must be taken in trying to validate or predict the performance of an adaptive scheme by means of ensemble-average learning curves. It can be shown that, for larger μ, the number of simulations that must be performed in order to obtain a good approximation to $\mathrm{E}\, e(k)^2$ or $\mathrm{E}\, \|\tilde{\boldsymbol{w}}_k\|^2$ can be of the order of tens or hundreds of thousands, depending on the distribution of the input sequence. A smaller number of simulations can lead to erroneous or deceptive conclusions. An analytical explanation and more details on this effect are given in [17].

References

[1] B. Widrow et al. Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proceedings of the IEEE*, 64:1151–1162, 1976.

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

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

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

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

[6] K. Mayyas and T. Aboulnasr. The leaky LMS algorithm: MSE analysis for gaussian data. *IEEE Transactions on Signal Processing*, 45(4):927–934, April 1997.

[7] J. E. Mazo. On the independence theory of equalizer convergence. *The Bell System Technical Journal*, 58:963–993, 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 Transactions on Information Theory*, IT-28:318–329, March 1982.

[9] O. Macchi and E. Eweda. Second-order convergence analysis of stochastic adaptive linear filtering. *IEEE Transactions on Automatic Control*, AC-28(1):76–85, January 1983.

[10] O. Macchi. Adaptive Processing: The LMS Approach with Applications in Transmission. Wiley, NY, 1995.

[11] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22:551–575, August 1977.

[12] W. A. Sethares, B. D. O. Anderson, and C. R. Johnson. Adaptive algorithms with filtered regressor and filtered error. *Mathematics of Control, Signals, and Systems*, 2:381–403, 1989.

[13] V. Solo and X. Kong. *Adaptive Signal Processing Algorithms*. Prentice-Hall, Englewood Cliffs, NJ, 1995.

[14] 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):712–756, March 1996.

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

[16] S. C. Douglas and W. Pan. Exact expectation analysis of the LMS adaptive filter. *IEEE Transactions on Signal Processing*, 43(12):2863–2871, December 1995.

[17] V. H. Nascimento and A. H. Sayed. Are ensemble-average learning curves reliable in evaluating the performance of adaptive filters? In *Proc. Asilomar Conference on Signals, Systems, and Computers*, CA, 1998.

[18] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1987.