Error-Energy Bounds for Adaptive Gradient Algorithms

Ali H. Sayed, *Member IEEE*, and Markus Rupp

***Abstract*— The paper establishes robustness, optimality, and convergence properties of the widely used class of instantaneous-gradient adaptive algorithms. The analysis is carried out in a purely deterministic framework and assumes no *a priori* statistical information. It employs the Cauchy–Schwarz inequality for vectors in an Euclidean space and derives local and global error-energy bounds that are shown to highlight, as well as explain, relevant aspects of the robust performance of adaptive gradient filters (along the lines of H^∞ theory).**

I. Introduction

ONE of the most widely used adaptive schemes in current practice is the least-mean-squares (LMS) algorithm [1], [2]. Its simplicity and computational efficiency, coupled with its good performance under varied operating conditions, have made the LMS a standard tool in a wide range of applications in signal processing, communications, control, and computations. Its widespread applicability has also led to an enormous interest in the analysis of its performance and convergence properties (e.g., [3]–[14]) and to the introduction of many different variants (e.g., [1], [2], [15]–[17]) with the intent of improving several of its characteristics.

Most of the available analyzes of gradient schemes rely on certain statistical assumptions that are part of the so-called independence theory [1, p. 315]. These assumptions may, in several instances, be restrictive, as pointed out in [1, p. 335] and in earlier references (e.g., [6], [9]). They may also be far from the conditions under which the LMS algorithm and its variants have proven themselves in practical situations. Only a handful of studies have avoided the statistical assumptions, albeit at the expense of either excluding the noise component [5] or requiring additional conditions on the data and the step-size parameter [6], [14].

This paper pursues an analysis of the class of instantaneous-gradient adaptive algorithms (with the LMS being a special case) within a purely deterministic framework. The derivation avoids statistical assumptions and proceeds to establish error-energy bounds that hold independent of stochastic considerations. The error bounds are further shown to explain the robustness behavior of gradient recursions on a step-by-step basis, as well as over intervals of time (cf., H^∞ filtering results). A convergence analysis is also provided that shows, under certain deterministic conditions on the data and noise sequences, that the estimate of the weight vector converges to the true weight vector.

The appealing nature (and, in a sense, the strength) of gradient-type algorithms is primarily due to their simplicity: They are simple to derive, simple to program and implement, and simple to explain. Keeping with this tradition, we try, in this work, to motivate the analysis from very first principles and with minimum background.

A. Notation

We use small boldface letters to denote vectors and capital boldface letters to denote matrices. In addition, the symbol "$*$" denotes Hermitian conjugation (complex conjugation for scalars), and the letter E stands for expectation. The symbol $\boldsymbol{I}$ denotes the identity matrix of appropriate dimensions, and the boldface letter $\mathbf{0}$ denotes either a zero vector or a zero matrix. Finally, the notation $\|\boldsymbol{x}\|$ denotes the Euclidean norm of a vector. All vectors are column vectors except for the input data vector denoted by $\boldsymbol{u}_i$, which is taken to be a row vector.

II. The Steepest Descent and LMS Algorithms

This section reviews the standard stochastic model that is often used to motivate gradient-descent algorithms.

Consider a zero-mean random signal $d(i)$ and a zero-mean M-dimensional input row vector $\boldsymbol{u}_i$ with

$$\sigma^2 = E(d^*(i)d(i)), \quad \boldsymbol{R} = E(\boldsymbol{u}_i^*\boldsymbol{u}_i), \quad \boldsymbol{p} = E(\boldsymbol{u}_i^* d(i)).$$

Consider further a column weight vector $\boldsymbol{w}_i$, and let $e(i)$ denote the estimation error between the desired signal $d(i)$ and the inner product $\boldsymbol{u}_i\boldsymbol{w}_i$,

$$e(i) = d(i) - \boldsymbol{u}_i\boldsymbol{w}_i.$$

The mean-squared error (or cost function) is the variance of $e(i)$ and is given by

$$J(i) = \sigma^2 - \boldsymbol{p}^*\boldsymbol{w}_i - \boldsymbol{w}_i^*\boldsymbol{p} + \boldsymbol{w}_i^*\boldsymbol{R}\boldsymbol{w}_i. \tag{1}$$

This is a quadratic cost function in $\boldsymbol{w}_i$, and the objective is to minimize it. The optimal choice $\boldsymbol{w}^o$ can be easily seen to be the solution of the normal system of equations $\boldsymbol{p} = \boldsymbol{R}\boldsymbol{w}^o$ [1].

Manuscript received June 20, 1994; revised February 12, 1996. This material was based on work supported by the National Science Foundation under Award no. MIP-9409319. The work of M. Rupp was also supported by a scholarship from DAAD (German Academic Exchange Service) as well as the scientific division of NATO. The associate editor coordinating the review of this paper and approving it for publication was Dr. Stephen M. McLaughlin.

A. H. Sayed is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 USA.

M. Rupp was a postdoctoral fellow in the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA. He is now with the Wireless Technology Research Department, Lucent Technologies, 791 Holmdel-Keyport Rd., Holmdel, NJ 07733–0400 USA.

Publisher Item Identifier S 1053-587X(96)05286-5.

A major inconvenience of solving the normal equations is that they require *a priori* knowledge of the autocorrelation and crosscorrelation quantities $\boldsymbol{R}$ and $\boldsymbol{p}$, respectively. However, even if these quantities were available, the $M \times M$ linear system of equations $\boldsymbol{p} = \boldsymbol{R}\boldsymbol{w}^o$ still needs to be solved for the optimal (Wiener) weight $\boldsymbol{w}^o$. This may require a significant amount of computational effort, especially for large values of M.

This problem can be ameliorated by employing an approximate gradient-descent algorithm. In this method, the weight estimates are recursively updated along the negative direction of the *instantaneous* gradient of $J(i)$, leading to the so-called LMS recursion:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu \boldsymbol{u}_i^* [d(i) - \boldsymbol{u}_i \boldsymbol{w}_{i-1}] \tag{2}$$

where μ is a positive constant step-size parameter and $\boldsymbol{w}_{-1}$ is an initial value (or guess).

Several other variants (such as $\epsilon - \text{LMS}, \alpha - \text{LMS}$, and projection LMS—see, e.g., [12]–[16]) have been proposed in the literature with the intent of improving several of the convergence and robustness properties of (2). These employ *time-variant* step-sizes and take the general form (μ in (2) is now replaced by $\mu(i)$ in (3))

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu(i) \boldsymbol{u}_i^* [d(i) - \boldsymbol{u}_i \boldsymbol{w}_{i-1}] \tag{3}$$

with many possible choices for $\mu(i)$, such as

$$\mu(i) = \frac{\alpha}{\|\boldsymbol{u}_i\|^2}, \quad \mu(i) = \frac{\alpha}{\epsilon + \|\boldsymbol{u}_i\|^2},$$
$$\mu(i) = \frac{\alpha}{1 + \alpha \|\boldsymbol{u}_i\|^2}$$

or some other choice, including matrix step-sizes, as briefly indicated in Section VI. Here, α and ϵ are resistive real numbers.

A. The Data Model

The analysis in this paper assumes the following model for the given data $\{d(i), \boldsymbol{u}_i\}$

$$d(i) = \boldsymbol{u}_i \boldsymbol{w} + v(i). \tag{4}$$

That is, it assumes that there exists an unknown column vector $\boldsymbol{w}$ that relates $\boldsymbol{u}_i$ and $d(i)$ via a noisy perturbation $v(i)$. The term $v(i)$ may account for measurement noise, modeling errors, or some other uncertainties.

The following issues regarding the behavior of recursion (3) when applied to model (4) are addressed in this paper:

1) A min-max interpretation relative to $\boldsymbol{w}$ is provided in Section IV for the estimates $\boldsymbol{w}_i$ obtained from (3).
2) Deterministic error-energy bounds are established for recursion (3) on step-by-step and global bases when applied to data generated by model (4). This is addressed in Theorem 1.
3) The convergence behavior, under purely deterministic conditions, of the gradient-based estimates $\boldsymbol{w}_i$ to the unknown weight vector $\boldsymbol{w}$ of (4) is discussed in Theorem 2.

III. Energy Bounds or Passivity Relations

We start by invoking a simple Cauchy–Schwarz argument to establish several local energy bounds that characterize the behavior of the gradient recursion (3) on a step-by-step basis.

For this purpose, it is instructive to ignore at this stage the gradient recursion (3) all by itself and to simply note the following general fact. Let $\boldsymbol{w}$ be any unknown weight vector that we wish to estimate, and let $\boldsymbol{u}_i$ be any given input vector at time i. Now, pick *any* positive real number $\mu(i)$ that satisfies

$$\mu(i) \|\boldsymbol{u}_i\|^2 \leq 1, \tag{5}$$

and pick *any* vector $\boldsymbol{q}$ as an estimate for the unknown weight vector $\boldsymbol{w}$. This is clearly a very crude estimator; it randomly picks a vector $\boldsymbol{q}$ and uses it as an estimate for $\boldsymbol{w}$, but still, and because of the condition (5) on $\mu(i)$, this estimator guarantees that the following bound is always satisfied:

$$\frac{|\boldsymbol{u}_i \boldsymbol{w} - \boldsymbol{u}_i \boldsymbol{q}|^2}{\mu^{-1}(i) \|\boldsymbol{w} - \boldsymbol{q}\|^2} \leq 1 \tag{6}$$

since it follows from the Cauchy–Schwarz inequality that

$$|\boldsymbol{u}_i \boldsymbol{w} - \boldsymbol{u}_i \boldsymbol{q}|^2 \leq \|\boldsymbol{u}_i\|^2 \|\boldsymbol{w} - \boldsymbol{q}\|^2.$$

We have also assumed that the obvious choice $\boldsymbol{q} = \boldsymbol{w}$ is excluded in order to avoid a ratio in (6) with zero numerator and denominator. However, here and in later places in the paper, we can avoid this technicality by working through with differences rather than ratios, say

$$|\boldsymbol{u}_i \boldsymbol{w} - \boldsymbol{u}_i \boldsymbol{q}|^2 - \mu^{-1}(i) \|\boldsymbol{w} - \boldsymbol{q}\|^2 \leq 0.$$

However, we shall continue, for now, to express our results in terms of ratios for convenience of exposition.

Now note that the quantity in the numerator of (6) is the square of the error in estimating $\boldsymbol{u}_i \boldsymbol{w}$ by using $\boldsymbol{u}_i \boldsymbol{q}$. Likewise, the quantity in the denominator of (6) is the square of the distance between the true $\boldsymbol{w}$ and its estimate $\boldsymbol{q}$ (weighted by $\mu^{-1}(i)$). Hence, (6) is the ratio of the "energies" of two error quantities: the error in estimating $\boldsymbol{u}_i \boldsymbol{w}$ and the error in estimating $\boldsymbol{w}$.

It is further obvious that if the denominator of (6) is increased by any nonnegative value, say, by the energy of a noise term $|v(i)|^2$, then the ratio will still be bounded by 1

$$\frac{|\boldsymbol{u}_i \boldsymbol{w} - \boldsymbol{u}_i \boldsymbol{q}|^2}{\mu^{-1}(i) \|\boldsymbol{w} - \boldsymbol{q}\|^2 + |v(i)|^2} \leq 1. \tag{7}$$

The denominator is now composed of two energy terms: one relative to the noise signal and the other relative to the error in our guess for $\boldsymbol{w}$.

The inequalities (6) and (7) are valid for *any* data $\boldsymbol{u}_i$ as long as $\mu(i) \|\boldsymbol{u}_i\|^2 \leq 1$ (cf. (5)), and they are valid for any choice of $\boldsymbol{q}$. They are, therefore, certainly valid for a $\boldsymbol{q}$ that has been generated by the gradient recursion (3). Therefore, if instead of $\boldsymbol{q}$ we employ the estimate $\boldsymbol{w}_{i-1}$, it also follows that

$$\frac{|\boldsymbol{u}_i \boldsymbol{w} - \boldsymbol{u}_i \boldsymbol{w}_{i-1}|^2}{\mu^{-1}(i) \|\boldsymbol{w} - \boldsymbol{w}_{i-1}\|^2 + |v(i)|^2} \leq 1. \tag{8}$$

However, how does the fact that the estimate $\boldsymbol{w}_{i-1}$ is generated by recursion (3) alter (8)? It turns out that (3) allows us to further tighten the inequality (8) and to conclude that the following also holds (as we shall promptly verify):

$$\frac{\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_i\|^2+|e_a(i)|^2}{\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+|v(i)|^2}\leq 1 \tag{9}$$

where we have replaced, for notational convenience, the term $\boldsymbol{u}_i(\boldsymbol{w}-\boldsymbol{w}_{i-1})$ by $e_a(i)$, which is also known as the *a priori* estimation error

$$e_a(i)=\boldsymbol{u}_i(\boldsymbol{w}-\boldsymbol{w}_{i-1}).$$

Comparing (9) with (8) we see that the numerator of (9) is larger since the nonnegative term $\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_i\|^2$ has been added to the numerator. However, although the numerator increased in value, the ratio is still guaranteed to be bounded by one. A simple proof of (9) is the following. Starting with the update equation (3), subtracting the true solution $\boldsymbol{w}$ from both sides and squaring, we obtain

$$\begin{aligned}\|\boldsymbol{w}-\boldsymbol{w}_i\|^2&=\|(\boldsymbol{w}-\boldsymbol{w}_{i-1})-\mu(i)\boldsymbol{u}_i^*(d(i)-\boldsymbol{u}_i\boldsymbol{w}_{i-1})\|^2\\&=\|(\boldsymbol{w}-\boldsymbol{w}_{i-1})-\mu(i)u_i^*[e_a(i)+v(i)]\|^2\end{aligned}$$

where we have used $d(i)-\boldsymbol{u}_i\boldsymbol{w}_{i-1}=\boldsymbol{u}_i(\boldsymbol{w}-\boldsymbol{w}_{i-1})+v(i)=e_a(i)+v(i)$. Expanding the right-hand side and rearranging terms leads to the equality

$$\begin{aligned}&\|\boldsymbol{w}-\boldsymbol{w}_i\|^2-\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+\mu(i)|e_a(i)|^2-\mu(i)|v(i)|^2\\&\quad=\mu(i)|e_a(i)+v(i)|^2[\mu(i)\|\boldsymbol{u}_i\|^2-1].\end{aligned}$$

The right-hand side in the above equality is the product of three terms: Two of them are nonnegative ($\mu(i)$ and $|e_a(i)+v(i)|^2$), whereas the third one ($\mu(i)\boldsymbol{u}_i\boldsymbol{u}_i^*-1$) may be positive, negative, or zero, depending on how $\mu(i)$ compares with $\|\boldsymbol{u}_i\|^2$. In particular, for $\mu(i)\|\boldsymbol{u}_i\|^2\leq 1$, the right-hand side is negative or zero, and therefore

$$\|\boldsymbol{w}-\boldsymbol{w}_i\|^2+\mu(i)|e_a(i)|^2\leq\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+\mu(i)|v(i)|^2$$

which is equivalent to the desired inequality (9).

A. Interpretation

In other words, we have established that inequality (9) holds for gradient recursions of the form (3). This can be regarded as a local passivity relation: It states that no matter what the value of the noise component $v(i)$ is and no matter how far the estimate $\boldsymbol{w}_{i-1}$ is from the true vector $\boldsymbol{w}$, the sum of energies $\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_i\|^2+|e_a(i)|^2$ will always be smaller than or equal to the sum of the energies of the starting errors (or disturbances) $\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+|v(i)|^2$.

In fact, other local relations can be established by following similar arguments, and we shall forgo the details here—derivations are provided in Appendix A. We instead collect the results into a theorem. Let

$$e_p(i)=\boldsymbol{u}_i(\boldsymbol{w}-\boldsymbol{w}_i)$$

denote the so-called *a posteriori* estimation error at time i. In addition, define the factor $\gamma(i)=[\mu^{-1}(i)-\|\boldsymbol{u}_i\|^2]$.

Theorem 1: Given the gradient recursion (3) and model (4), the following local energy bounds always hold at each time instant i:

$$\frac{\|\boldsymbol{w}-\boldsymbol{w}_i\|^2+\mu(i)|e_a(i)|^2}{\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+\mu(i)|v(i)|^2}\leq 1, \tag{10}$$

$$\frac{|e_a(i)|^2+|e_p(i)|^2}{\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+|v(i)|^2}\leq 1, \tag{11}$$

$$\frac{\gamma(i)\|\boldsymbol{w}-\boldsymbol{w}_i\|^2+|e_p(i)|^2}{\gamma_{(i)}\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+|v(i)|^2}\leq 1, \tag{12}$$

$$\frac{|e_a(i)|^2+|e_p(i+1)|^2}{\mu^{-1}(i)\|\boldsymbol{w}-\boldsymbol{w}_{i-1}\|^2+|v(i)|^2}\leq 1 \tag{13}$$

where it is assumed that $\mu(i)\|\boldsymbol{u}_i\|^2\leq 1$ for the first three bounds, whereas

$$\mu(i)\leq\min\{1/\|\boldsymbol{u}_i\|^2,1/\|\boldsymbol{u}_{i+1}\|^2\}$$

for the last bound.

The above local bounds show, on a step-by-step basis, how the energies of the *a priori* and *a posteriori* residuals compare with the energies of the disturbances due to $v(i)$ and to the weight estimation errors $(\boldsymbol{w}-\boldsymbol{w}_{i-1})$ or $(\boldsymbol{w}-\boldsymbol{w}_i)$.

Moreover, since the contractivity relation (10) holds for each time instant i, it should also hold globally over an interval of time. Let

$$\tilde{\boldsymbol{w}}_i=\boldsymbol{w}-\boldsymbol{w}_i$$

denote the weight-error vector. Assuming $\mu(i)\|\boldsymbol{u}_i\|^2\leq 1$ over $0\leq i\leq N$, it follows from (10) that

$$\frac{\|\tilde{\boldsymbol{w}}_N\|^2+\sum_{i=0}^{N}|\overline{e}_a(i)|^2}{\|\tilde{\boldsymbol{w}}_{-1}\|^2+\sum_{i=0}^{N}|\overline{v}(i)|^2}\leq 1 \tag{14}$$

where we have introduced the normalized *a priori* residuals and the normalized noise signals

$$\overline{e}_a(i)=\sqrt{\mu(i)}e_a(i),\quad\overline{v}(i)=\sqrt{\mu(i)}v(i).$$

The numerator of (14) is the sum of the energies of the normalized *a priori* residuals $\overline{e}_a(i)$ over $0\leq i\leq N$ and the energy of the final weight-error at time N. Likewise, the sum in the denominator consists of two terms: the energy of the normalized noise signal over the same time interval and the energy of the weight error due to the initial guess. Consequently, (14) states that the (block lower triangular) matrix that maps the normalized noise signals $\{\overline{v}(i)\}_{i=0}^N$ and the initial uncertainty $\tilde{\boldsymbol{w}}_{-1}$ to the normalized *a priori* residuals $\{\overline{e}_a(i)\}_{i=0}^N$ and the final weight error $\tilde{\boldsymbol{w}}_N$ is always a contraction mapping (see Fig. 1). This means that the 2-induced norm of this mapping, which is denoted by $\mathcal{T}_N$, is always upper bounded by one ($\|\mathcal{T}_N\|_{2,\text{ind}}\leq 1$). In the language of robust filtering and control, the 2-induced norm is referred to as the H^∞-norm (due to connections with a frequency domain interpretation that we forgo here (see [18])).

The contractivity of $\mathcal{T}_N$ provides an interesting explanation for the robust behavior of gradient-type algorithms of the form

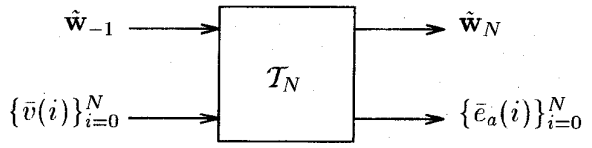


Fig. 1. Causal mapping $calT_N$.

(3): It shows that the energy of the residuals will never exceed the energy of the disturbances. We shall refer to (14) as a global error bound since it is valid over an interval of time.

Alternatively, if we denote by $\Delta_N(\boldsymbol{w}_{-1}, v(\cdot))$ the difference between the numerator and the denominator of (14)

$$\begin{aligned}\Delta_N(\boldsymbol{w}_{-1}, v(\cdot)) &= \left\{\|\tilde{\boldsymbol{w}}_N\|^2 + \sum_{i=0}^{N} |\bar{e}_a(i)|^2\right\} \\ &\quad - \left\{\|\tilde{\boldsymbol{w}}_{-1}\|^2 + \sum_{i=0}^{N} |\bar{v}(i)|^2\right\} \end{aligned} \tag{15}$$

then we also conclude from the argument prior to (14) that we always have, for any $\boldsymbol{w}_{-1}$ and $v(\cdot)$

$$\Delta_N(\boldsymbol{w}_{-1}, v(\cdot)) \leq 0. \tag{16}$$

Global bounds that are similar to (14) and (16) and that are based on *a posteriori* rather than *a priori* residuals, can also be established by invoking the third inequality in Theorem 1. For example, assuming $\mu(i)\|\boldsymbol{u}_i\|^2 < 1$, we conclude that

$$\frac{\|\tilde{\boldsymbol{w}}_N\|^2 + \sum_{i=0}^{N} |\bar{\bar{e}}_p(i)|^2}{\|\tilde{\boldsymbol{w}}_{-1}\|^2 + \sum_{i=0}^{N} |\bar{\bar{v}}(i)|^2} \leq 1$$

where we have defined $\bar{\bar{e}}_p(i) = \sqrt{\gamma^{-1}(i)}\, e_p(i)$ and $\bar{\bar{v}}(i) = \sqrt{\gamma^{-1}(i)}\, v(i)$.

In the next section, we expand on the significance of such global relations. This will be achieved, for instance, by showing how the global relation (14) allows us to provide a statement concerning the min-max nature of gradient algorithms (thus complementing the interesting conclusions of [18]).

IV. Minimax Optimality of Gradient Recursions

The global property (14) (or (16)) is valid for any initial guess $\boldsymbol{w}_{-1}$ and for any noise sequence $v(\cdot)$ as long as the $\mu(i)$ are properly bounded as in (5). One might then wonder whether the bound in (14) is tight or not. That is, are there choices $\{\boldsymbol{w}_{-1}, v(\cdot)\}$ for which the ratio in (14) can be made arbitrarily close to one (or Δ_N in (16) arbitrarily close to zero)? The answer is positive. To clarify this point, we rewrite the gradient recursion (3) in the alternative form

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu(i)\boldsymbol{u}_i^*[e_a(i) + v(i)]. \tag{17}$$

We can now envision a noise sequence $v(i)$ that satisfies $v(i) = -e_a(i)$ at each time instant i (after all, we have no say in the values that the $v(\cdot)$ can assume). In this case, the above gradient recursion trivializes to $\boldsymbol{w}_i = \boldsymbol{w}_{i-1}$ for all i, thus leading to $\boldsymbol{w}_N = \boldsymbol{w}_{-1}$, and the ratio in (14) will be one for any $\boldsymbol{w}_{-1} \neq \boldsymbol{w}$. Correspondingly, Δ_N will be zero for any $\boldsymbol{w}_{-1}$. This means that the maximum value of the ratio in (14), over the unknowns $\{\boldsymbol{w}_{-1}, v(\cdot)\}$, is equal to one

$$\max_{\{\boldsymbol{w}_{-1} \neq \boldsymbol{w}, v(\cdot)\}} \left\{ \frac{\|\boldsymbol{w} - \boldsymbol{w}_N\|^2 + \sum_{i=0}^{N} |\bar{e}_a(i)|^2}{\|\boldsymbol{w} - \boldsymbol{w}_{-1}\|^2 + \sum_{i=0}^{N} |\bar{v}(i)|^2} \right\} = 1. \tag{18}$$

In addition

$$\max_{\{\boldsymbol{w}_{-1}, v(\cdot)\}} \{\Delta_N(\boldsymbol{w}_{-1}, v(\cdot))\} = 0.$$

Another question of interest is the following: How does the gradient recursion (3) compare with other possible recursive algorithms for the update of the weight estimate? We assume the algorithms are causal in the sense that the weight estimate at time i is only a function of the data $\{\boldsymbol{u}_j, d(j)\}$ up to and including time i.

Let $\mathcal{A}$ denote any given causal algorithm and assume we perform the following experiment on $\mathcal{A}$: We initialize it with $\boldsymbol{w}_{-1} = \boldsymbol{w}$ and define the noise sequence $v(i)$ in terms of the resulting (successive) *a priori* estimation errors as follows: $v(i) = -e_a(i)$ for $0 \leq i \leq N$. Then, it always holds that

$$\begin{aligned}\sum_{i=0}^{N} |\bar{v}(i)|^2 = \sum_{i=0}^{N} |\bar{e}_a(i)|^2 &\leq \|\boldsymbol{w} - \boldsymbol{w}_N\|^2 \\ &\quad + \sum_{i=0}^{N} |\bar{e}_a(i)|^2\end{aligned}$$

no matter what the resulting value of $\boldsymbol{w}_N$ is. Therefore, this particular choice of initial guess ($\boldsymbol{w}_{-1} = \boldsymbol{w}$) and noise sequence $\{v(\cdot)\}$ will always result in a difference Δ_N that is nonnegative. This implies that for any causal algorithm $\mathcal{A}$, it always holds that

$$\max_{\{\boldsymbol{w}_{-1}, v(\cdot)\}} \{\Delta_N(\boldsymbol{w}_{-1}, v(\cdot))\} \geq 0.$$

For the gradient recursion (3), we were able to show that the maximum has to be exactly zero because the global property (16) provided us with an inequality in the other direction. This may or may not hold for a generic causal algorithm. We can therefore state that among all causal algorithms, the gradient-type recursion (3) is one that solves the following optimization problem:

$$\min_{\text{Algorithm}} \left\{ \max_{\{\boldsymbol{w}_{-1}, v(\cdot)\}} \Delta_N(\boldsymbol{w}_{-1}, v(\cdot)) \right\} \tag{19}$$

and that the optimal value is equal to zero.

As explained before, Δ_N has the following physical interpretation: For any causal algorithm, we define the (block lower) triangular operator $\mathcal{T}_N$ that maps the initial disturbances $\{\boldsymbol{w} - \boldsymbol{w}_{-1}, \bar{v}(\cdot)\}$ to the resulting estimation errors $\{\boldsymbol{w} - \boldsymbol{w}_N, \bar{e}_a(\cdot)\}$. Then, Δ_N measures the difference between the

output energy and the input energy of $\mathcal{T}_N$. The gradient recursion (3) is therefore an algorithm that minimizes the maximum possible difference between these energies over all disturbances. More intuitively, it minimizes the maximum effect of the input disturbances on the resulting estimation-error energy. For the case of a constant step-size μ, this optimality result is in agreement with the H^∞ characterization of the LMS algorithm; this is a conclusion that was first derived in [18] (contrary to the statement on page 70 of the third edition of [1]).

V. Sufficient Convergence Conditions

The contractive relation (14) also has implications on the limiting performance of the errors in the gradient recursion (3) as time progresses to infinity. To clarify this, we proceed with our analysis in a purely deterministic framework and assume no statistical information about the noise sequence and the input data.

Recall the definition of the weight-error vector $\tilde{\boldsymbol{w}}_i = \boldsymbol{w} - \boldsymbol{w}_i$, which therefore measures the error of the weight-estimate relative to the *true* weight vector $\boldsymbol{w}$ of model (4), rather than the Wiener solution $\boldsymbol{w}^o$. It follows from the gradient recursion (3) that $\tilde{\boldsymbol{w}}_i$ satisfies the update relation

$$\tilde{\boldsymbol{w}}_i = \tilde{\boldsymbol{w}}_{i-1} - \overline{\boldsymbol{u}}_i^* [\overline{e}_a(i) + \overline{v}(i)] \tag{20}$$

where $\overline{e}_a(i) = \sqrt{\mu(i)} e_a(i)$, $\overline{v}(i) = \sqrt{\mu(i)} v(i)$, and $\overline{u}_i = \sqrt{\mu(i)} u_i$.

Our purpose is to provide sufficient conditions that would guarantee the weight-error vector $\tilde{\boldsymbol{w}}_i$ in (20) to tend to zero as time progresses.

The analysis in this section is based on the following two deterministic assumptions (see the example and simulation in next section):

1) *Finite normalized-noise energy:* Our first condition requires the normalized noise measurement $\{\overline{v}(\cdot) = \sqrt{\mu(\cdot)} v(\cdot)\}$ to have finite energy, i.e.

$$\sum_{i=0}^{\infty} \mu(i) |v(i)|^2 < \infty. \tag{21}$$

2) *Persistent excitation:* Our second condition requires the normalized input rows $\{\overline{\boldsymbol{u}}_i = \sqrt{\mu(i)} \boldsymbol{u}_i\}$ to be persistently exciting. By this, we mean that there exists a finite integer K such that the smallest singular value of

$$\begin{bmatrix} \overline{\boldsymbol{u}}_i \\ \vdots \\ \overline{\boldsymbol{u}}_{i+K} \end{bmatrix} \tag{22}$$

is uniformly bounded from below by a positive quantity, say, δ, for sufficiently large i.

The following result is now immediate, a proof of which follows from the contractivity relation (14) and can be found in the companion paper [21] (where the condition on $\mu(i)\|\boldsymbol{u}_i\|^2$ is further relaxed and allowed to be bounded by 2).

Theorem 2: Assume that $\mu(i)\|\boldsymbol{u}_i\|^2 \leq 1$ and $\|\tilde{\boldsymbol{w}}_{-1}\| < \infty$. If $\{\overline{v}(\cdot)\}$ has finite energy, then $\overline{e}_a(i) \to 0$. If $\{\overline{\boldsymbol{u}}_i\}$ is further persistently exciting, then $\boldsymbol{w}_i \to \boldsymbol{w}$.

Note that if $\overline{v}(\cdot)\}$ is instead a finite-power sequence (rather than finite-energy), i.e., if $\overline{v}(\cdot) = \sqrt{\mu(\cdot)} v(\cdot)$ satisfies

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} \mu(i) |v(i)|^2 = P_v < \infty$$

then, in this case, we conclude from (16) that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} \mu(i) |e_a(i)|^2 \leq P_v.$$

In other words, a bounded noise power leads to a bounded estimation error power.

A. Discussion and a Numerical Example

There is an abundance of (almost-sure) convergence results for stochastic gradient algorithms in the literature, (e.g., [6], [14], [19]). They all exhibit a set of sufficient conditions that include, among others, requirements on the step-size parameter $\mu(i)$ and on the distribution of the input data. For example, it is usually required for almost-sure convergence of $\boldsymbol{w}_i$ (to some fixed point) that the step-size parameter $\mu(i)$ be chosen such that [6]

$$\sum_{i=0}^{\infty} \mu(i) = \infty. \tag{23}$$

If one adopts this requirement for the choice of $\mu(i)$, then the finite-energy condition (21) would require the noise sequence $v(i)$ to vanish, i.e., $v(i) \to 0$.

We include an example to show that convergence may still occur when (23) is violated, and the noise does not vanish. For this purpose, we construct an example with a nonvanishing noise sequence that satisfies conditions (21) and (22) and, therefore, guarantees the convergence of $\boldsymbol{w}_i$ to $\boldsymbol{w}$. The example however is such that the noise-to-signal energy decays to zero. That is, although the noise signal never vanishes, the input data vector $\boldsymbol{u}_i$ becomes more powerful (energy-wise) as time progresses.

Assume the noise sequence is never vanishing and that it is bounded by a certain constant, say, $|v(i)| \leq \alpha$ for all i and for some finite $\alpha > 0$. There is no restriction on how big or how small α can be.

Assume further that the time-variant step-size $\mu(i)$ is taken to be

$$\mu(i) = \frac{1}{(i+1)^2} \quad \text{for} \quad i \geq 0. \tag{24}$$

It is easy to check that this choice violates (23) since

$$\sum_{i=0}^{\infty} \mu(i) = \sum_{i=0}^{\infty} \frac{1}{(i+1)^2} < \infty.$$

However, the finite-energy condition (21) is not violated since

$$\sum_{i=0}^{\infty} \mu(i) |v(i)|^2 \leq |\alpha|^2 \sum_{i=0}^{\infty} \mu(i) < \infty.$$

Assume further that the data $\{d(i)\}$ is generated via

$$d(i) = \boldsymbol{u}_i \boldsymbol{w} + v(i)$$

for a noise sequence that satisfies the above boundedness requirement and where the $\boldsymbol{u}_i$ are 1×3 row vectors that are constructed as follows. Let $\{\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2\}$ denote the basis vectors

$$\boldsymbol{e}_0 = [1 \quad 0 \quad 0], \quad \boldsymbol{e}_1 = [0 \quad 1 \quad 0],$$
$$\boldsymbol{e}_2 = [0 \quad 0 \quad 1]$$

and choose for $i = 0, 3, 6, 9, \cdots$ (multiples of 3)

$$\boldsymbol{u}_i = 0.1 \cdot (i+1) \cdot \boldsymbol{e}_0, \boldsymbol{u}_{i+1} = 0.1 \cdot (i+2) \cdot \boldsymbol{e}_1,$$
$$\boldsymbol{u}_{i+2} = 0.1 \cdot (i+3) \cdot \boldsymbol{e}_2. \tag{25}$$

That is, for the first six time instants

$$\boldsymbol{u}_0 = [0.1 \quad 0 \quad 0], \quad \boldsymbol{u}_3 = [0.4 \quad 0 \quad 0],$$
$$\boldsymbol{u}_1 = [0 \quad 0.2 \quad 0], \quad \boldsymbol{u}_4 = [0 \quad 0.5 \quad 0],$$
$$\boldsymbol{u}_2 = [0 \quad 0 \quad 0.3], \quad \boldsymbol{u}_5 = [0 \quad 0 \quad 0.6].$$

In other words, the $\boldsymbol{u}_i$ are multiples of the basis vectors, with the coefficient changing from one time instant to another.

If we pick any i, say, w.l.o.g. a multiple of 3, then it can be easily seen, using the choice (24) for $\mu(i)$, that

$$\begin{bmatrix} \overline{\boldsymbol{u}}_i \\ \overline{\boldsymbol{u}}_{i+1} \\ \overline{\boldsymbol{u}}_{i+2} \end{bmatrix} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \tag{26}$$

which is always a full rank matrix. This means that the sequence $\{\sqrt{\mu(i)}\boldsymbol{u}_i\}$ is persistently exciting, and we can choose $K = 3$ in the definition of persistence of excitation after (22).

In addition, the requirement (5) is satisfied here since for any i

$$\mu(i)\|\boldsymbol{u}_i\|^2 = \frac{0.1^2(i+1)^2}{(i+1)^2} = 0.01 \leq 1. \tag{27}$$

In fact, note that we do not need to restrict the coefficient in (25) to be the same and equal to 0.1 for all i. It can be taken to be any nonzero number whose absolute value is smaller than 1 (say, sufficiently bounded away from zero). It can also vary from one time instant to another. In this case, the resulting matrix in (26) will still be full rank, and the condition (27) will still be satisfied.

Therefore, all the requirements of Theorem 2 are met and we conclude that if the gradient recursion (3) is applied to the data $\{d(i), \boldsymbol{u}_i\}$ of this example, then we must obtain convergence to $\boldsymbol{w}$, i.e., $\boldsymbol{w}_i \to \boldsymbol{w}$.

Fig. 2 confirms the above discussion. It is the result of a MATLAB[1] simulation. The data was generated for $\boldsymbol{w}^* = [1 \quad 3 \quad 0.5]$ with zero initial guess $\boldsymbol{w}_{-1} = 0$, and the noise level was allowed to take random values in the range $[-0.5, 0.5]$ (i.e., $|v(i)| \leq 0.5$ for all i). In addition, the coefficient 0.1 in (25) was replaced by a nonzero random number always less than 1 and sufficiently bounded away from zero. The figure shows the convergence of both $\tilde{\boldsymbol{w}}_i$ and $e_a(i)$ to zero in about 200 iterations.

[1] Matlab is a copyright of The MathWorks Inc.

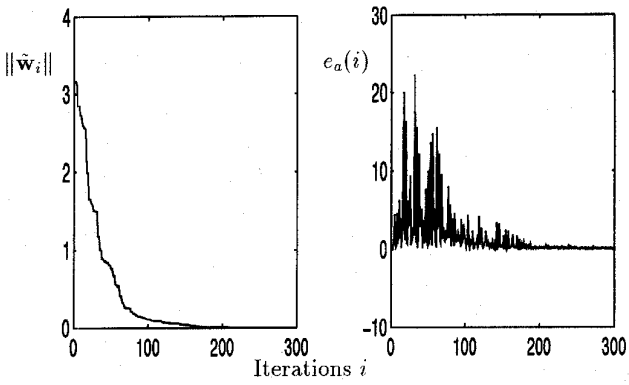


Fig. 2. Convergence of $\tilde{\boldsymbol{w}}_i$ and $e_a(i)$ to zero.

VI. Algorithms with Matrix "Step-Sizes"

The derivation in the earlier sections allows us to extend the results to cases where the step-size parameter is assumed to be a matrix quantity, diagonal or not, thus extending earlier results in the literature (see, e.g., [16] where, motivated by statistical considerations, a constant diagonal step-size matrix with exponential entries was used). We thus consider a recursive update of the form

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \boldsymbol{D}_i \boldsymbol{u}_i^* [d(i) - \boldsymbol{u}_i \boldsymbol{w}_{i-1}] \tag{28}$$

where a general (not necessarily diagonal) time-variant positive-definite step-size matrix $\boldsymbol{D}_i$ is allowed.

Following similar arguments to what we have done before, we obtain the following result.

Theorem 3: Given (28), the following three relations hold for $0 < \boldsymbol{u}_i \boldsymbol{D}_i \boldsymbol{u}_i^* \leq 1$:

$$\frac{|e_a(i)|^2}{\tilde{\boldsymbol{w}}_{i-1}^* \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{w}}_{i-1} - \tilde{\boldsymbol{w}}_i^* \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{w}}_i + |v(i)|^2} \leq 1$$

$$\frac{|e_p(i)|^2}{[1 - \boldsymbol{u}_i \boldsymbol{D}_i \boldsymbol{u}_i^*][\tilde{\boldsymbol{w}}_{i-1}^* \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{w}}_{i-1} - \tilde{\boldsymbol{w}}_i^* \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{w}}_i] + |v(i)|^2} \leq 1$$

$$\frac{|e_a(i)|^2 + |e_p(i)|^2}{\tilde{\boldsymbol{w}}_{i-1}^* \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{w}}_{i-1} + |v(i)|^2} \leq 1$$

while

$$\frac{|e_a(i)|^2 + |e_a(i+1)|^2}{\tilde{\boldsymbol{w}}_{i-1}^* \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{w}}_{i-1} + |v(i)|^2} \leq 1$$

holds for $0 < \max\{\boldsymbol{u}_i \boldsymbol{D}_i \boldsymbol{u}_i^*, \boldsymbol{u}_{i+1} \boldsymbol{D}_i \boldsymbol{u}_{i+1}^*\} \leq 1$.

If $\boldsymbol{D}_i$ is a scalar multiple of a constant positive-definite matrix $\boldsymbol{D}$, say

$$\boldsymbol{D}_i = \gamma(i)\boldsymbol{D}, \quad \gamma(i) > 0, \quad \boldsymbol{D} > \boldsymbol{0},$$

then we can extend the earlier global relations to this case and verify that

$$\frac{(\boldsymbol{w} - \boldsymbol{w}_N)^* \boldsymbol{D}^{-1} (\boldsymbol{w} - \boldsymbol{w}_N) + \sum_{i=0}^{N} |\overline{e}_a(i)|^2}{(\boldsymbol{w} - \boldsymbol{w}_{-1})^* \boldsymbol{D}^{-1} (\boldsymbol{w} - \boldsymbol{w}_{-1}) + \sum_{i=0}^{N} |\overline{v}(i)|^2} \leq 1$$

for $0 < \boldsymbol{u}_i \boldsymbol{D}_i \boldsymbol{u}_i^* \leq 1$, whereas

$$\frac{(\boldsymbol{w}-\boldsymbol{w}_N)^* \boldsymbol{D}^{-1}(\boldsymbol{w}-\boldsymbol{w}_N) + \sum_{i=0}^{N} |\bar{\bar{e}}_a(i)|^2}{(\boldsymbol{w}-\boldsymbol{w}_{-1})^* \boldsymbol{D}^{-1}(\boldsymbol{w}-\boldsymbol{w}_{-1}) + \sum_{i=0}^{N} |\bar{\bar{v}}(i)|^2} \leq 1$$

for $0 < \boldsymbol{u}_i \boldsymbol{D}_i \boldsymbol{u}_i^* < 1$. Here

$$\begin{aligned}
\bar{e}_a(i) &= \sqrt{\gamma(i)}\, e_a(i), \bar{v}_i = \sqrt{\gamma(i)}\, v(i), \\
\bar{\bar{e}}_a(i) &= \sqrt{(\gamma^{-1}(i) - \boldsymbol{u}_i \boldsymbol{D} \boldsymbol{u}_i^*)^{-1}}\, e_a(i), \\
\bar{\bar{v}}(i) &= \sqrt{(\gamma^{-1}(i) - \boldsymbol{u}_i \boldsymbol{D} \boldsymbol{u}_i^*)^{-1}}\, v(i).
\end{aligned}$$

VII. Concluding Remarks

We have provided a time-domain analysis of gradient-based adaptive schemes with emphasis on robustness, optimality, and convergence issues. This was achieved by highlighting a fundamental contraction mapping property of the algorithm (3), as summarized in (14). The result states that a gradient recursion always results in a contraction mapping from the initial disturbances (noise and initial uncertainty) to the final estimation errors (viz., *a priori* estimation errors and final weight estimate). This conclusion was motivated by first establishing several error bounds on a local level, as summarized by Theorem 1. The bounds involve not only the *a priori* estimation error but the *a posteriori* estimation error and combinations of both as well.

The min-max optimality of the gradient recursion (3) was also addressed in Section IV. Recursion (3) is usually derived as an approximate solution for the minimization of the quadratic cost function $J(i)$ in (1): The approximation is due to the use of instantaneous estimates for the second-order statistics of the data. The significance of the discussion in Section IV is that it provides an optimality criterion for gradient recursions. In simple terms, it states that the gradient approximations are in fact min-max filters. This result was first established for the the LMS algorithm, and in the infinite-horizon case, $i \to \infty$ in [18] by exploiting connections with estimation in indefinite metric spaces and H^∞-theory. We have considered here general time-variant steps sizes as in (3) and have a stronger finite-horizon bound in (14) as well.

We may finally remark that the approach of this paper can be extended to more general scenarios that arise, for instance, in adaptive schemes for IIR identification, in filtered-error variants, and in Gauss–Newton methods. The relevant details can be found in [20]–[22].

Appendix A
Proof of the Local Properties of Theorem 1

Let $\hat{v}(i) = d(i) - \boldsymbol{u}_i \boldsymbol{w}_{i-1}$, and note that $e_a(i) = \boldsymbol{u}_i \tilde{\boldsymbol{w}}_{i-1}$ and

$$\begin{aligned}
e_p(i) &= \boldsymbol{u}_i \tilde{\boldsymbol{w}}_{i-1} - \mu(i) \|\boldsymbol{u}_i\|^2 \hat{v}(i). \\
e_a(i+1) &= \boldsymbol{u}_{i+1} \tilde{\boldsymbol{w}}_{i-1} - \mu(i) \boldsymbol{u}_{i+1} \boldsymbol{u}_i^* \hat{v}(i). \\
v(i) &= \hat{v}(i) - \boldsymbol{u}_i \tilde{\boldsymbol{w}}_{i-1}.
\end{aligned}$$

1) To prove the third property in Theorem 1, we note that

$$\begin{aligned}
\|\tilde{\boldsymbol{w}}_i\|^2 &= \|\tilde{\boldsymbol{w}}_{i-1} - \mu(i) \boldsymbol{u}_i^* \hat{v}(i)\|^2, \\
|v(i)|^2 &= |\hat{v}(i) - \boldsymbol{u}_i \tilde{\boldsymbol{w}}_{i-1}|^2, \\
|e_p(i)|^2 &= |\boldsymbol{u}_i \tilde{\boldsymbol{w}}_{i-1} - \mu(i) \|\boldsymbol{u}_i\|^2 \hat{v}(i)|^2.
\end{aligned}$$

Using these expressions for $\|\tilde{\boldsymbol{w}}_i\|^2$, $|v(i)|^2$, and $|e_p(i)|^2$, we get

$$\begin{aligned}
&[\mu^{-1}(i) - \|\boldsymbol{u}_i\|^2]\{\|\tilde{\boldsymbol{w}}_{i-1}\|^2 - \|\tilde{\boldsymbol{w}}_i\|^2\} \\
&\quad + |v(i)|^2 - |e_p(i)|^2 \\
&\quad = (1 - \mu(i)\|\boldsymbol{u}_i\|^2)|\hat{v}(i)|^2 \geq 0.
\end{aligned}$$

2) To prove the second property in Theorem 1, we note that it requires

$$\begin{aligned}
&\mu^{-1}(i)\|\boldsymbol{w} - \boldsymbol{w}_{i-1}\|^2 + |v(i)|^2 - |e_a(i)|^2 \\
&\quad - |e_p(i)|^2 \geq 0.
\end{aligned}$$

Replacing the quantities $v(i), e_a(i)$, and $e_p(i)$ in terms of $\hat{v}(i)$ and $\tilde{\boldsymbol{w}}_{i-1}$, the above inequality then collapses to the following quadratic inequality in $\hat{v}(i)$ and $\tilde{\boldsymbol{w}}_{i-1}$:

$$[\tilde{\boldsymbol{w}}_{i-1}^* \hat{v}^*(i)] \boldsymbol{P} \begin{bmatrix} \tilde{\boldsymbol{w}}_{i-1} \\ \hat{v}(i) \end{bmatrix} \geq 0$$

where

$$\boldsymbol{P} = \begin{bmatrix} \mu^{-1}(i)\boldsymbol{I}_M - \boldsymbol{u}_i^* \boldsymbol{u}_i & -(\boldsymbol{I}_M - \mu(i)\boldsymbol{u}_i^* \boldsymbol{u}_i)\boldsymbol{u}_i^* \\ -\boldsymbol{u}_i(\boldsymbol{I}_M - \mu(i)\boldsymbol{u}_i^* \boldsymbol{u}_i) & 1 - \mu^2(i)\boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{u}_i \boldsymbol{u}_i^* \end{bmatrix}.$$

The leading block of $\boldsymbol{P}$, viz., $\mu^{-1}(i)\boldsymbol{I}_M - \boldsymbol{u}_i^* \boldsymbol{u}_i$ is a positive-semidefinite matrix due to the condition on $\mu(i)$. The Schur complement of $\boldsymbol{P}$ with respect to this block is equal to $(1 - \mu(i)\|\boldsymbol{u}_i\|^2)$, which is again nonnegative. We therefore conclude that $\boldsymbol{P}$ is a positive-semidefinite matrix, and the required inequality is thus guaranteed.

3) To prove the fourth property in Theorem 1, we note that it requires

$$\begin{aligned}
&\mu^{-1}(i)\|\boldsymbol{w} - \boldsymbol{w}_{i-1}\|^2 + |v(i)|^2 - |e_a(i)|^2 \\
&\quad - |e_a(i+1)|^2 \geq 0.
\end{aligned}$$

Replacing the quantities $v(i), e_a(i)$, and $e_a(i+1)$ in terms of $\hat{v}(i)$ and $\tilde{\boldsymbol{w}}_{i-1}$, the above inequality then collapses to the following quadratic inequality in $\hat{v}(i)$ and $\tilde{\boldsymbol{w}}_{i-1}$,

$$[\tilde{\boldsymbol{w}}_{i-1}^* \hat{v}^*(i)] \boldsymbol{P} \begin{bmatrix} \tilde{\boldsymbol{w}}_{i-1} \\ \hat{v}(i) \end{bmatrix} \geq 0$$

where now

$$\boldsymbol{P} = \begin{bmatrix} \mu^{-1}(i)\boldsymbol{I}_M - \boldsymbol{u}_{i+1}^* \boldsymbol{u}_{i+1} & -(\boldsymbol{I}_M - \mu(i)\boldsymbol{u}_{i+1}^* \boldsymbol{u}_{i+1})\boldsymbol{u}_i^* \\ -\boldsymbol{u}_i(\boldsymbol{I}_M - \mu(i)\boldsymbol{u}_{i+1}^* \boldsymbol{u}_{i+1}) & 1 - \mu^2(i)\boldsymbol{u}_{i+1} \boldsymbol{u}_i^* \boldsymbol{u}_i \boldsymbol{u}_{i+1}^* \end{bmatrix}.$$

The leading block of $\boldsymbol{P}$, viz., $\mu^{-1}(i)\boldsymbol{I}_M - \boldsymbol{u}_{i+1}^* \boldsymbol{u}_{i+1}$ is a positive-semidefinite matrix due to the condition on $\mu(i)$. The Schur complement of $\boldsymbol{P}$ with respect to this block is equal to $(1 - \mu(i)\|\boldsymbol{u}_i\|^2)$, which is again

nonnegative. We therefore conclude that $\boldsymbol{P}$ is a positive-semidefinite matrix, and the required inequality is thus guaranteed.

ACKNOWLEDGMENT

The authors would like to thank B. Halder, B. Hassibi, T. Kailath, and an anonymous reviewer for helpful feedback on earlier drafts of this work.

REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991, 2nd ed.

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

[3] B. Widrow *et al.*, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, pp. 1151–1162, 1976.

[4] J. E. Mazo, "On the independence theory of equalizer convergence," *Bell Syst. Tech. J.*, vol. 58, pp. 963–993, 1979.

[5] R. R. Bitmead and B. D. O. Anderson, "Performance of adaptive estimation algorithms in dependent random environments," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 788–794, 1980.

[6] D. C. Farden, "Stochastic approximation with correlated data," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 105–113, 1981.

[7] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, "Comparison of the convergence of two algorithms for adaptive FIR digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, no. 3, pp. 670–678, June 1981.

[8] O. Macchi and E. Eweda, "Convergence analysis of self-adaptive equalizers," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 161–176, 1984.

[9] W. A. Gardner, "Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis and critique," *Signal Processing*, vol. 6, pp. 113–133, 1984.

[10] A. Feuer and E. Weinstein, "Convergence analysis of LMS filters with uncorrelated Gaussian data," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–33, no. 1, pp. 222–230, Feb. 1985.

[11] K. H. Shi and F. Kozin, "On almost sure convergence of adaptive algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 471–474, 1986.

[12] N. J. Bershad, "Analysis of the normalized LMS algorithm with Gaussian inputs," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 5, pp. 793–806, 1986.

[13] ——, "Behavior of the ϵ-normalized LMS algorithm with Gaussian inputs," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 5, pp. 636–644, May 1987.

[14] V. Solo, "The limiting behavior of LMS," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1909–1922, 1989.

[15] R. W. Harris, D. M. Chabries, and F. A. Bishop, "A variable step (VS) adaptive filter algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 2, pp. 309–316, Apr. 1986.

[16] S. Makino, Y. Kaneda, and N. Koizumi, "Exponentially weighted step-size NLMS adaptive filter based on the statistics of a room impulse response," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 101–108, Jan. 1993.

[17] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction, and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

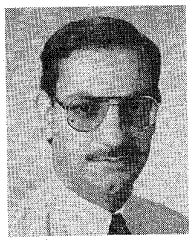
[18] B. Hassibi, A. H. Sayed, and T. Kailath, "LMS is H^∞ optimal," in *Proc. Conf. Decision Contr.*, vol. 1, San Antonio, TX, Dec. 1993, pp. 74–79; also in *IEEE Trans. Signal Processing*, vol. 44, no. 2, pp. 267–280, Feb. 1996.

[19] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

[20] A. H. Sayed and M. Rupp, "A class of adaptive nonlinear H^∞-filters with guaranteed l_2-stability," in *Proc. IFAC Symp. Nonlinear Contr. Syst. Design*, Tahoe City, CA, vol. 1, June 1995, pp. 455–460; also to appear in *Automatica*.

[21] M. Rupp and A. H. Sayed, "A time-domain feedback analysis of filtered-error adaptive gradient algorithms," *IEEE Trans. Signal Processing*, vol. 44, no. 6, pp. 1428–1439, June 1996.

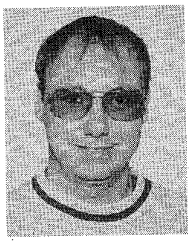
[22] ——, Robustness of Gauss–Newton recursive methods: A deterministic feedback analysis," to appear in *Signal Processing*, vol. 50, no. 1, 1996.



Ali H. Sayed (M'92) was born in São Paulo, Brazil. In 1981, he graduated in first place in the National Lebanese Baccalaureat, with the highest score in the history of the examination. In 1987, he received the degree of Engineer in Electrical Engineering from the University of São Paulo, Brazil, and was the first-place graduate of the School of Engineering. In 1989, he received, with distinction, the M.S. degree in electrical engineering from the University of São Paulo, Brazil, and was awarded a FAPESP fellowship for overseas studies. In 1992, he received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

From September 1992 to August 1993, he was a Research Associate with the Information Systems Laboratory at Stanford University, after which he joined, as an Assistant Professor, the Department of Electrical and Computer Engineering at the University of California, Santa Barbara. His research interests are in the areas of adaptive and statistical signal processing, linear and nonlinear filtering and estimation, interplays between signal processing and control methodologies, system theory, interpolation theory, and reliable and efficient computations for structured problems.

Dr. Sayed is a member of SIAM, and ILAS. He was awarded the Institute of Engineering Prize and the Conde Armando Alvares Penteado Prize, both in Brazil in 1987. His Ph.D. thesis on structured algorithms received a special mention for the Householder Prize in Numerical Linear Algebra in 1993. He is a recipient of a 1994 NSF Research Initiation Award and of the 1996 IEEE Donald G. Fink Prize.



Markus Rupp received the Diploma in electrical engineering from FHS Saarbruecken, Germany, and Universitaet of Saarbruecken in 1984 and 1988, respectively. He received the Doctoral degree in 1993 summa cum laude from the TH Darmstadt in the field of acoustical echo cancellation.

During this time, he was also a lecturer in digital signal processing and high frequency techniques at FHS. He was awarded a DAAD postdoctoral fellowship and spent the time from November 1993 until September 1995 in the Department of Electrical and Computer Engineering at the University of California, Santa Barbara. Since October 1995, he has been with Lucent Technologies (previously AT&T), Wireless Technology Research Group, Holmdel, NJ, where he is currently working on new adaptive equalization schemes for wireless telephones. His interests involve algorithms for speech applications, speech enhancement and recognition, echo compensation, adaptive filter theory, H_∞-filtering, neural nets, classification algorithms, and signal detection.