*IN PROC. ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS, AND COMPUTERS, PP. 592–596, OCT. 1994.*

# ON THE ROBUSTNESS, CONVERGENCE, AND MINIMAX PERFORMANCE OF INSTANTANEOUS-GRADIENT ADAPTIVE FILTERS

ALI H. SAYED AND MARKUS RUPP

Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106–9560

**Abstract**—The paper establishes several 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 apriori statistical information. It starts with a simple Cauchy-Schwarz inequality for vectors in an Euclidean space and proceeds to derive local and global energy bounds that are shown here to highlight, as well as explain, several relevant aspects of this important class of algorithms.

## I. INTRODUCTION

One of the most widely used algorithms in current practice is the Least-Mean-Squares (LMS) algorithm of Widrow and Hoff (1960). Its simplicity and widespread applicability have led to an enormous interest in the analysis of its performance and convergence properties, and to the introduction of many different variants with the intent of improving several of its characteristics (see, e.g., [1]–[5] and the references therein). But most of the available studies and convergence analysis, however, rely on certain fundamental statistical assumptions that, in many respects, are restrictive and far from the conditions under which the LMS algorithm and its several variants have proven themselves in practical situations. This paper addresses these issues and provides a novel unified analysis of a wide class of instantaneous-gradient adaptive algorithms within a purely deterministic framework. Several new local and global error-energy bounds are established that explain the robustness behaviour of the gradient recursions on a step-by-step basis, as well as over intervals of time. 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 (rather than the Wiener solution). The analysis provided herein leads to a much-needed theoretical validation of this class of adaptive algorithms.

For the special case of the LMS algorithm, it was recently argued in [6], using arguments based on the notion of estimation in indefinite metric spaces, that the LMS algorithm is an optimal $H^\infty$-filter as time progresses to infinity, thus leading to connections with recent work in the fields of robust estimation and control.

The current paper provides a new unified framework that extends this, and other results, in some directions. The derivation given here encompasses a wide range of instantaneous-gradient algorithms with the LMS case being a special example. It also exhibits new local error-energy bounds for the varied gradient recursions. These bounds explain the behaviour of the update-recursions on a local level, i.e., from one time-instant to another. The approach also provides a global optimization criterion that is valid over finite intervals of time, and with no assumptions neither on the noise sequence nor on the data sequence. A convergence analysis is also provided that employs no statistical considerations and establishes the behaviour of the algorithm as time progresses to infinity.

We may also remark that the analysis in this paper extends equally well to the class of IIR gradient-based schemes as well as to filtered error variants. These extensions will be discussed elsewhere (e.g., [10]) and, in fact, can be regarded as special cases of a class of so-called (linear and nonlinear) $H^\infty$ adaptive filters studied in [11].

Finally, we shall use small boldface letters to denote

This material is based on work supported by the National Science Foundation under Award No. MIP-9409319. The work of M. Rupp is also supported by a scholarship from DAAD (German Academic Exchange Service) as well as the scientific division of NATO (Wissenschaftsausschuß der NATO).

vectors and capital boldface letters to denote matrices. Also, the symbol "$*$" will denote Hermitian conjugation (complex conjugation for scalars).

## II. The Stochastic Model

For the sake of illustration and completeness, this section reviews the standard stochastic model that is often used to motivate gradient-descent algorithms. This also serves the purpose of introducing several quantities that are of interest further ahead.

So let $\mathbf{w}$ be a *column* vector of $M$ unknown parameters that will be referred to as the *weight vector.* Consider further a zero-mean random signal $d(i)$ and a zero-mean input *row* vector $\mathbf{u}_i$ with $\sigma^2 = E(d^*(i)d(i))$, $\mathbf{R} = E(\mathbf{u}_i^*\mathbf{u}_i)$, and $\mathbf{p} = E(\mathbf{u}_i^* d(i))$. Here the letter $E$ stands for expectation. Let $v(i)$ denote the difference $v(i) = d(i) - \mathbf{u}_i\mathbf{w}$, which thus represents the noise component that explains the mismatch between $d(i)$ and $\mathbf{u}_i\mathbf{w}$. The $v(i)$ is again a zero-mean random variable whose variance will be denoted by $J(\mathbf{w}) = E(v^*(i)v(i))$,

$$J(\mathbf{w}) \;\; = \;\; \sigma^2 - \mathbf{p}^*\mathbf{w} - \mathbf{w}^*\mathbf{p} + \mathbf{w}^*\mathbf{R}\mathbf{w}\,. \qquad (1)$$

The objective is to determine an estimate for the weight vector $\mathbf{w}$, say $\mathbf{w}^o$, so as to minimize the variance $J(\mathbf{w})$ of the noise component $v(i)$. The optimal estimate, $\mathbf{w}^o$, can be easily seen to be the solution of the normal system of equations $\mathbf{p} = \mathbf{R}\mathbf{w}^o$.

## III. Gradient-Descent Algorithms

A major inconvenience of solving the normal equations is that they require apriori knowledge of the autocorrelation and cross-correlation quantities $\mathbf{R}$ and $\mathbf{p}$, respectively. But even if these quantities were available, the $M \times M$ linear system of equations still needs to be solved for the optimal weight $\mathbf{w}^o$. This may require a significant amount of computational effort, especially for large values of $M$.

A way out of this is to employ an approximate gradient-descent solution. In this method, weight estimates are recursively updated along the negative direction of the *instantaneous* gradient of $J(\mathbf{w})$, leading to the so-called LMS recursion:

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu\mathbf{u}_i^*\left[d(i) - \mathbf{u}_i\mathbf{w}_{i-1}\right], \qquad (2)$$

where $\mu$ is a positive step-size parameter and $\mathbf{w}_{-1}$ is an initial guess.

Several other variants (such as $\epsilon$−LMS, $\alpha$−LMS, projection LMS, etc.) 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

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i)\mathbf{u}_i^*\left[d(i) - \mathbf{u}_i\mathbf{w}_{i-1}\right]\;,\;\; \mathbf{w}_{-1}\;, \qquad (3)$$

with many possible choices for $\mu(i)$. Here, and for the sake of generality, a general time-variant step-size $\mu(i)$ will be considered, rather than focus on special choices. Also, it should be added that the framework of this paper is equally applicable to recursions with matrix step-sizes.

## IV. Questions to be Addressed

Several open issues regarding the behaviour of recursion (3) will be addressed in this paper:

(i) The instantaneous-gradient algorithm (3) does not lead to the exact minimization of $J(\mathbf{w})$ in (1). It is thus natural to seek an optimization criterion for which (3) is a natural solution. The answer is provided in Theorem 2 below.

(ii) The second point that is addressed here is the study of the basic properties of recursion (3) on a step-by-step, as well as a global, basis and without any statistical assumptions. This is addressed in Theorem 1 below.

(iii) The last issue to be studied here establishes the convergence, under purely deterministic conditions, of the gradient-based estimates $\mathbf{w}_i$ relative to the true weight vector $\mathbf{w}$, rather than the Wiener solution $\mathbf{w}^o$. This is established in Theorem 3 below.

## V. Local Error-Energy Bounds: Passivity Relations

We now invoke a simple Cauchy-Schwarz argument (see, e.g., last Section in [7]) to establish several local energy bounds that characterize the behaviour of the gradient recursion (3) on a step-by-step basis. For this purpose, it is instructive to ignore the gradient recursion (3) all by itself and to simply note the following: pick *any* positive real number $\mu(i)$ that satisfies $\mu(i)\|\mathbf{u}_i\|_2^2 \leq 1$, and pick *any* vector $\mathbf{q}$ as an estimate for the unknown weight vector $\mathbf{w}$. This is clearly a very crude estimator: it randomly picks a vector $\mathbf{q}$ and uses it as an estimate for $\mathbf{w}$. But still, and because of the condition on $\mu(i)$, this estimator guarantees that the following bound is always satisfied:

$$\frac{|\mathbf{u}_i\mathbf{w} - \mathbf{u}_i\mathbf{q}|^2}{\mu^{-1}(i)\|\mathbf{w} - \mathbf{q}\|_2^2} \leq 1\,. \qquad (4)$$

This follows from the Cauchy-Schwarz inequality $|\mathbf{u}_i\mathbf{w} - \mathbf{u}_i\mathbf{q}|^2 \leq \|\mathbf{u}_i\|_2^2\;\|\mathbf{w} - \mathbf{q}\|_2^2$. We have assumed above that the obvious choice $\mathbf{q} = \mathbf{w}$ is avoided so as to avoid a ratio with zero numerator and denominator. However, here and in later places in the paper, we can avoid this technicality by working all through with differences rather than ratios, say $|\mathbf{u}_i\mathbf{w}-\mathbf{u}_i\mathbf{q}|^2-\mu^{-1}(i)\|\mathbf{w}-\mathbf{q}\|_2^2 \leq 0$. But we shall continue, for now, to express our results in terms of ratios for convenience of exposition.

Continuing with (4), it is certainly true that if its denominator is increased by any positive value, say by the noise term $|v(i)|^2$, then the ratio is still bounded by 1,

$$\frac{|\mathbf{u}_i\mathbf{w}-\mathbf{u}_i\mathbf{q}|^2}{\mu^{-1}(i)\|\mathbf{w}-\mathbf{q}\|_2^2+|v(i)|^2} \leq 1 \, . \tag{5}$$

The inequalities (4) and (5) are valid for *any* data $\mathbf{u}_i$ as long as $\mu(i)\|\mathbf{u}_i\|_2^2 \leq 1$ and they are valid for any choice of $\mathbf{q}$. They are, therefore, certainly valid for a $\mathbf{q}$ that is generated by the gradient recursion (3). So if $\mathbf{q}$ is replaced by $\mathbf{w}_{i-1}$ it follows that

$$\frac{|\mathbf{u}_i\mathbf{w}-\mathbf{u}_i\mathbf{w}_{i-1}|^2}{\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+|v(i)|^2} \leq 1 \, . \tag{6}$$

One might then wonder in what sense does the gradient recursion (3) alter (6)? It can be easily seen that it allows a further tightening of the inequality and to conclude that the following also holds,

$$\frac{\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_i\|_2^2+|e_a(i)|^2}{\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+|v(i)|^2} \leq 1, \tag{7}$$

where we have replaced, for notational convenience, the term $\mathbf{u}_i(\mathbf{w}-\mathbf{w}_{i-1})$ by $e_a(i)$ – also known as the apriori estimation error. This establishes a local error-energy bound: it states that no matter what the value of the noise component $v(i)$ is, and no matter how far the estimate $\mathbf{w}_{i-1}$ is from the true vector $\mathbf{w}$, the sum of the energies of the resulting errors, viz., $\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_i\|_2^2+|e_a(i)|^2$, will always be smaller than or equal to the sum of the energies of the starting errors (or disturbances), $\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+|v(i)|^2$. This can also be interpreted as a passivity relation. In fact, other similar local relations can be established by following similar arguments, and we shall forgo the details here. We instead collect the results into a theorem: let $e_p(i)=\mathbf{u}_i\mathbf{w}-\mathbf{u}_i\mathbf{w}_i$ denote the so-called aposteriori estimation error at time $i$. Also, define the factor $\gamma(i)=[\mu^{-1}(i)-\|\mathbf{u}_i\|_2^2]$.

**Theorem 1 (Energy Bounds)** *The following local energy bounds always hold at each time instant $i$:*

$$\begin{aligned}
\frac{\|\mathbf{w}-\mathbf{w}_i\|_2^2+\mu(i)|e_a(i)|^2}{\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+\mu(i)|v(i)|^2} &\leq 1 \, , \\
\frac{|e_a(i)|^2+|e_p(i)|^2}{\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+|v(i)|^2} &\leq 1 \, , \\
\frac{\gamma(i)\|\mathbf{w}-\mathbf{w}_i\|_2^2+|e_p(i)|^2}{\gamma(i)\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+|v(i)|^2} &\leq 1 \, , \\
\frac{|e_a(i)|^2+|e_a(i+1)|^2}{\mu^{-1}(i)\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2+|v(i)|^2} &\leq 1 \, ,
\end{aligned}$$

*where it is assumed that $\mu(i)\|\mathbf{u}_i\|_2^2 \leq 1$ for the first three bounds, while $\mu(i) \leq \min\left\{1/\|\mathbf{u}_i\|_2^2, 1/\|\mathbf{u}_{i+1}\|_2^2\right\}$ for the last bound.*

These local bounds show, on a step-by-step basis, how the energies of the apriori and aposteriori residuals compare with the energies of the disturbances due to $v(i)$ and to the weight estimation errors, $(\mathbf{w}-\mathbf{w}_{i-1})$ or $(\mathbf{w}-\mathbf{w}_i)$. They also lead to an important conclusion regarding the optimality and convergence of instantaneous-gradient algorithms as we now verify.

## VI. Global Error-Energy Bounds: Contraction Mapping

Assume we run the gradient recursion (3) from time $i=0$ up to time $N$ and that, at each time instant $i$, the $\mu(i)$ is chosen to satisfy $\mu(i)\ \|\mathbf{u}_i\|_2^2 \leq 1$. It then follows that the first inequality in Theorem 1 holds for each $0 \leq i \leq N$,

$$\mu(i)|e_a(i)|^2 \ \leq\ \|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2-\|\mathbf{w}-\mathbf{w}_i\|_2^2+\mu(i)|v(i)|^2 \, .$$

Summing over $i$ we conclude that we must have (we now use the simplifying notation $\tilde{\mathbf{w}}_i=\mathbf{w}-\mathbf{w}_i$)

$$\frac{\|\tilde{\mathbf{w}}_N\|_2^2\ +\ \sum_{i=0}^N|\bar{e}_a(i)|^2}{\|\tilde{\mathbf{w}}_{-1}\|_2^2+\sum_{i=0}^N|\bar{v}(i)|^2} \ \leq\ 1, \tag{8}$$

where we have also introduced the normalized apriori residuals and the normalized noise signals, $\bar{e}_a(i)=\sqrt{\mu(i)}\ e_a(i)$ and $\bar{v}(i)=\sqrt{\mu(i)}\ v(i)$. The numerator of (8) is the sum of the energies of the normalized apriori residuals $\bar{e}_a(i)$ 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, (8) establishes a global energy bound over the interval of duration $(N+1)$: it states that the (block lower triangular) matrix that maps the normalized noise signals $\{\bar{v}(i)\}_{i=0}^N$ and the initial uncertainty $\tilde{\mathbf{w}}_{-1}$ to the normalized apriori residuals $\{\bar{e}_a(i)\}_{i=0}^N$ and the final weight error $\tilde{\mathbf{w}}_N$ is always a contraction mapping – see Figure 1 further ahead. This means that the 2−induced norm of this mapping, denoted by $\mathcal{T}_N$, is always upper bounded by one ($\|\mathcal{T}_N\|_{2,ind} \leq 1$) – in the language of robust filtering and control (e.g., [8,9]), the 2−induced norm is often referred to as the $H^\infty$−norm (due to connections with a frequency domain interpretation that we forgo here).

Alternatively, if we denote by $\Delta_N(\mathbf{w}_{-1},v(\cdot))$ the difference between the numerator and the denominator of (8),

$$\Delta_N(\mathbf{w}_{-1},v(\cdot)) = \tag{9}$$

$$\left\{ \|\tilde{\mathbf{w}}_N\|_2^2 \; + \; \sum_{i=0}^{N} |\bar{e}_a(i)|^2 \right\} - \left\{ \|\tilde{\mathbf{w}}_{-1}\|_2^2 \; + \; \sum_{i=0}^{N} |\bar{v}(i)|^2 \right\} ,$$

then we also conclude from the argument pior to (8) that we always have, for any $\mathbf{w}_{-1}$ and $v(\cdot)$,

$$\Delta_N(\mathbf{w}_{-1}, v(\cdot)) \leq 0. \qquad (10)$$

Global bounds similar to (8) and (10) and that are based on aposteriori residuals can also be established by invoking the third inequality in Theorem 1. We shall not pursue these details here for obvious reasons of brevity. Instead, we shall expand on the significance of such global relations. This will be achieved, for instance, by showing how the global relation (8) allows us to provide a statement concerning the minimax nature of gradient algorithms.

## VII. Minimax Optimality of Gradient Recursions

The global property (8) (or (10)) is valid for any initial guess $\mathbf{w}_{-1}$ and for any noise sequence $v(\cdot)$, as long as the $\mu(i)$ are properly bounded. One might then wonder whether the bound in (8) is tight or not. That is, are there disturbances $\{\mathbf{w}_{-1}, v(\cdot)\}$ for which the ratio in (8) can be made arbitrarily close to one (or $\Delta_N$ in (10) arbitrarily close to zero)? The answer is positive. To clarify this, we rewrite the gradient recursion (3) in the alternative form

$$\mathbf{w}_i \;\; = \;\; \mathbf{w}_{i-1} + \mu(i)\mathbf{u}_i^* \left[ e_a(i) + v(i) \right] \; . \qquad (11)$$

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 saying in the values that the $v(\cdot)$ can assume). In this case, the above gradient recursion trivializes to $\mathbf{w}_i = \mathbf{w}_{i-1}$ for all $i$; thus leading to $\mathbf{w}_N = \mathbf{w}_{-1}$ and the ratio in (8) will be one for any $\mathbf{w}_{-1} \neq \mathbf{w}$. Correspondingly, $\Delta_N$ will be zero for any $\mathbf{w}_{-1}$. This means that the maximum value of the ratio in (8), over the unknowns $\{\mathbf{w}_{-1}, v(\cdot)\}$, is equal to one,

$$\max_{\{\mathbf{w}_{-1} \neq \mathbf{w}, v(\cdot)\}} \left\{ \frac{\|\mathbf{w} - \mathbf{w}_N\|_2^2 + \sum_{i=0}^{N} |\bar{e}_a(i)|^2}{\|\mathbf{w} - \mathbf{w}_{-1}\|_2^2 + \sum_{i=0}^{N} |\bar{v}(i)|^2} \right\} = 1. \qquad (12)$$

Also,

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

Another question of interest is the following: how does the gradient recursion (3) compare with other possible recursive algorithms[1] for the update of the weight estimate?

As a motivation, we first consider the important subclass of algorithms that involve update-recursions of the form: $\mathbf{w}_i = \mathbf{w}_{i-1} + \mathbf{f}(\mathbf{u}_i, v(i) + e_a(i)), \;\; \mathbf{f}(\mathbf{u}_i, 0) = \mathbf{0}$, where $\mathbf{f}(\mathbf{x}, z)$ is any (linear or nonlinear) vector function with arguments $(\mathbf{x}, z)$ that satisfies the condition $\mathbf{f}(\mathbf{x}, 0) = \mathbf{0}$. This includes the gradient recursion (3) as well as the RLS recursion [1,7] as special cases. It is immediate to see that for any such algorithm, if we choose $\mathbf{w}_{-1} \neq \mathbf{w}$ and set $v(i) = -e_a(i)$ then the ratio in (12) will be one and, consequently, the maximum over all $(\mathbf{w}_{-1} \neq \mathbf{w}, v(\cdot))$ will necessarily be larger than or equal to 1.

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

$$\sum_{i=0}^{N} |\bar{v}(i)|^2 \leq \|\mathbf{w} - \mathbf{w}_N\|_2^2 + \sum_{i=0}^{N} |\bar{e}_a(i)|^2 \; ,$$

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

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

For the gradient recursion (3) we were able to show that the maximum has to be exactly zero because the global property (10) already provided us with an inequality in the other direction. This may or may not hold for any other causal algorithm. We can therefore state that among all causal algorithms, the gradient-type recursion (3) is one that solves the following optimization problem:

$$\min_{Algorithm} \left\{ \max_{\{\mathbf{w}_{-1}, v(\cdot)\}} \;\; \Delta_N(\mathbf{w}_{-1}, v(\cdot)) \right\} , \qquad (13)$$

and that the optimal value is equal to zero. As explained before, $\Delta_N$ has the following physical interpretation: for any causal algorithm we define the (block lower) triangular operator $\mathcal{T}_N$ that maps the initial disturbances $\{\tilde{\mathbf{w}}_{-1}, \bar{v}(\cdot)\}$ to the resulting estimation errors $\{\tilde{\mathbf{w}}_N, \bar{e}_a(\cdot)\}$. Then $\Delta_N$ measures the difference between the output energy and the input energy of $\mathcal{T}_N$. The gradient recursion (3) is thus 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 over the resulting estimation-error energy.

[1] We assume the algorithms are causal in the sense that the weight estimate at time $i$ is only a function of the data $\{\mathbf{u}_j, d(j)\}$ up to and including time $i$.

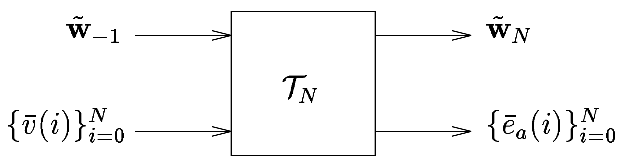
$\tilde{\mathbf{w}}_{-1}$
$\{\bar{v}(i)\}_{i=0}^{N}$
$\mathcal{T}_N$
$\tilde{\mathbf{w}}_N$
$\{\bar{e}_a(i)\}_{i=0}^{N}$

Figure 1: *Causal mapping $\mathcal{T}_N$.*

**Theorem 2 (Minimax Optimality)** *Among all causal estimators that recursively estimate the unknown weight vector $\mathbf{w}$, starting with an initial guess $\mathbf{w}_{-1}$ and producing successive estimates $\{\mathbf{w}_0, \ldots, \mathbf{w}_N\}$ and successive residuals $\{e_a(0), \ldots, e_a(N)\}$, the gradient recursion (3) is one solution that solves the min-max problem (13), where $\Delta_N$ is as defined in (9). Moreover, the optimal (i.e., the minimum) value is equal to zero.*

## VIII. Convergence of Gradient Algorithms

We now study the convergence of the gradient recursion (3) from a deterministic point of view, and without assuming any statistical information. We instead require the following:

(i) *Finite noise energy:* $\sum_{i=0}^{\infty} \mu(i)|v(i)|^2 < \infty$.

(ii) *Persistent excitation.* That is, the input vectors $\{\sqrt{\mu(i)}\mathbf{u}_i\}$ are such $\lim_{i\to\infty} \mu(i)|\mathbf{u}_i\mathbf{x}|^2 = 0$ implies $\mathbf{x} = \mathbf{0}$.

It follows from the global property (8) that (with $0 < \mu(i)\ \|\mathbf{u}_i\|_2^2 \leq 1$)

$$\sum_{i=0}^{N} \mu(i)|e_a(i)|^2 \ \leq \ \|\tilde{\mathbf{w}}_{-1}\|_2^2 + \sum_{i=0}^{N} \mu(i)|v(i)|^2 .$$

We then conclude from the finite energy assumption on $\bar{v}(\cdot)$ and from the boundedness of $\|\tilde{\mathbf{w}}_{-1}\|_2^2$ that

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

This means that the infinite series $\sum_{i=0}^{\infty} \mu(i)|e_a(i)|^2$ is convergent, which implies, by a classical result in mathematical analysis, that the sequence $\{\mu(i)|e_a(i)|^2\}$ converges to zero, or $\lim_{i\to\infty} \sqrt{\mu(i)}e_a(i) = 0$. If we now further replace $e_a(i)$ by its definition, we conclude that $\lim_{i\to\infty} \mu(i)|\mathbf{u}_i\tilde{\mathbf{w}}_{i-1}|^2 = 0$. It then follows from the persistent excitation assumption (e.g., [5]) that we must have $\lim_{i\to\infty} \tilde{\mathbf{w}}_{i-1} = \mathbf{0}$.

**Theorem 3 (Convergence)** *Assume $\mu(i)\ \|\mathbf{u}_i\|_2^2 \leq 1$ for all $i$ and that the normalized noise sequence has finite energy, $\sum_{i=0}^{\infty} \mu(i)|v(i)|^2 < \infty$. It then follows that the normalized apriori residuals obtained via the gradient recursion (3) tend to zero, $\lim_{i\to\infty} \sqrt{\mu(i)}e_a(i) = 0$. If the input vectors $\{\sqrt{\mu(i)}\ \mathbf{u}_i\}$ are further persistently exciting then $\lim_{i\to\infty} \mathbf{w}_i = \mathbf{w}$.*

## IX. Concluding Remarks

We finally remark that the point of view taken in this work can be extended to deal with gradient-type recursions that often arise in IIR modeling, as well as with variants that employ filtered error quantities. In these cases some nonlinearities arise that can still be properly handled within the framework of the current paper. These extensions will be treated elsewhere (see, e.g., [10,11] and the references therein). In particular, the gradient recursion (3), as well as the IIR extensions, can be viewed as special cases of so-called (linear and nonlinear) $\mathrm{H}^{\infty}-$adaptive filters studied in [11].

### Acknowledgment

The authors are grateful to T. Kailath, B. Hassibi, and B. Halder for their helpful comments.

## References

[1] S. Haykin, *Adaptive Filter Theory*, NJ: Prentice Hall, second edition, 1991.

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

[3] N. J. Bershad, "Behaviour of the $\epsilon-$normalized LMS algorithm with Gaussian inputs," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. ASSP-35, no. 5, pp. 636–644, May 1987.

[4] W. A. Gardner, "Nonstationary learning characteristics of the LMS algorithm," *IEEE Transactions on Circuits and Systems*, vol. CAS-34, pp. 1199–1207, 1987.

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

[6] B. Hassibi, A. H. Sayed, and T. Kailath, "LMS is $\mathrm{H}^{\infty}$ optimal," *Proc. Conference on Decision and Control*, vol. 1, pp. 74–79, San Antonio, Texas, December 1993.

[7] A. H. Sayed, and T. Kailath, "A state-space approach to adaptive RLS filtering," *IEEE Signal Processing Magazine*, vol. 11, no. 3, pp. 18–60, July 1994.

[8] J. C. Doyle, K. Glover, P. Khargonekar, and B. Francis, State-space solutions to standard $\mathrm{H}_2$ and $\mathrm{H}_{\infty}$ control problems, *IEEE Transactions on Automatic Control*, vol. AC-34, no. 8, pp. 831–847, August 1989.

[9] P.P. Khargonekar and K. M. Nagpal, Filtering and smoothing in an $\mathrm{H}^{\infty}-$ setting, *IEEE Trans. on Automatic Control*, vol. AC-36, pp. 151–166, 1991.

[10] M. Rupp and A. H. Sayed, "On the stability and convergence of Feintuch's algorithm for adaptive IIR filtering," to appear in *Proc. ICASSP*, 1995.

[11] A. H. Sayed and M. Rupp, "A class of nonlinear adaptive $\mathrm{H}^{\infty}$-filters with guaranteed $l_2$-stability," *submitted for publication.*