Study of the Transient Phase of the Forgetting Factor RLS

George V. Moustakides

***Abstract*— We investigate the convergence properties of the forgetting factor RLS algorithm in a stationary data environment. Using the settling time as our performance measure, we show that the algorithm exhibits a variable performance that depends on the particular combination of the initialization and noise level. Specifically when the observation noise level is low (high SNR) RLS, when initialized with a matrix of small norm, it has an exceptionally fast convergence. Convergence speed decreases as we increase the norm of the initialization matrix. In a medium SNR environment, the optimum convergence speed of the algorithm is reduced as compared with the previous case; however, RLS becomes more insensitive to initialization. Finally, in a low SNR environment, we show that it is preferable to initialize the algorithm with a matrix of large norm.**

I. Introduction

THE RECURSIVE least squares (RLS) algorithm is one of the most well-known algorithms used in adaptive filtering and system identification. Its popularity is mainly due to its exceptionally fast convergence speed, which is considered to be optimal in practice and as a measure for comparison (and desired goal) for other algorithms.

Due to its nonlinear nature, the exact theoretical analysis of RLS turns out to be rather complicated. This analytic complexity is particularly apparent in the case of the forgetting factor RLS, which is the most commonly used version of the algorithm. There exists an extensive literature addressing the problem of convergence of RLS under a stationary environment and its performance at steady state [1], [3], [5], [6], [17]. Recent publications, on the other hand, tend to focus mainly on the tracking properties of the algorithm [7], [9], [13]–[15], [20].

Although the performance of RLS, in a stationary environment and during the transient phase, is considered well studied and well understood, there exist certain observations, coming from practice, that cannot be adequately explained with the existing theory. We refer specifically to the variable performance of the algorithm as a function of the initialization of the (exponentially weighted) sample covariance matrix, which is recursively updated in the algorithm. RLS is known to exhibit a significantly faster convergence when initialized with a "small" positive definite matrix (usually of the form of $\delta \boldsymbol{I}$) than when initialized with a "large" one [10, p. 570], [21, p. 476]. The existing analysis cannot distinguish or explain in any sense this variable performance; therefore, there is room for further investigation that concentrates specifically on the initialization. A first effort toward this direction is the statistical analysis of the algorithm for soft and exact initialization [11], [12] but applies only to times $n \leq N$ (where N is the size of the estimation vector).

In this paper, we make a thorough study of the relation between the algorithmic performance and the initialization. Specifically, by analyzing the power of the estimation error vector, we show that the convergence properties of the algorithm not only depend on the initialization but also on the observation noise level. Furthermore, by using the settling time as our performance measure, we prove that the well known (from practice) rule of initialization with a "small" matrix is preferable for cases of high and medium SNR, whereas for low SNR, a "large" matrix must be selected for best results.

The paper is organized as follows. Section II contains the definition of the problem and certain background results. Section III contains our study of the estimation error power and estimates of the settling time. In Section IV, we summarize our results. Simulations are presented in Section V, and finally, Section VI contains the conclusion.

II. Background

Let us consider the linear system

$$y_n = X_n^t W_\star + w_n \qquad n \geq 0 \tag{1}$$

where

$\{y_n\}$ measurable scalar observation sequence;
$\{X_n\}$ measurable vector input data sequence;
$\{w_n\}$ additive observation noise;
$W_\star$ unknown deterministic time invariant vector.

RLS is the well known algorithm that recursively estimates $W_\star$ with the set of equations

$$\begin{aligned} \boldsymbol{R}_n &= (1-\mu)\boldsymbol{R}_{n-1} + X_n X_n^t \\ \epsilon_n &= y_n - X_n^t W_{n-1} \\ W_n &= W_{n-1} + \epsilon_n \boldsymbol{R}_n^{-1} X_n \end{aligned} \tag{2}$$

where

$(1-\mu)$ forgetting factor with $\mu \in [0, 1)$;
W_n estimate of the vector $W_\star$ at time n;
$\boldsymbol{R}_n$ exponentially weighted sample covariance matrix of the input sequence.

Manuscript received April 10, 1996; revised May 27, 1997. This work was supported by the Greek General Secretariat for Research and Technology under Grant ΠΕΝΕΔ96:1584. The associate editor coordinating the review of this paper and approving it for publication was Dr. Nurgun Erdol.

The author is with Department of Computer Engineering and Informatics, University of Patras, Patras, Greece and the Computer Technology Institute (CTI) of Patras, Patras, Greece.

Publisher Item Identifier S 1053-587X(97)07355-8.

In fact, RLS computes recursively $\boldsymbol{P}_n = \boldsymbol{R}_n^{-1}$ [10, p. 569], thus avoiding the direct inversion of the matrix $\boldsymbol{R}_n$ at every time n. However, since we are only concerned with convergence, we are going to assume infinite precision; thus, (2) is equivalent to the usual RLS in the sense that both algorithms yield the same estimates.

Let us now obtain a form of the algorithm that is more suitable for our analysis. By introducing the estimation error vector $\Delta_n = W_n - W_\star$ and using (1), we have

$$\begin{aligned} \boldsymbol{R}_n &= (1-\mu)\boldsymbol{R}_{n-1} + X_n X_n^t \\ \epsilon_n &= w_n - X_n^t \Delta_{n-1} \\ \Delta_n &= \Delta_{n-1} + \epsilon_n \boldsymbol{R}_n^{-1} X_n. \end{aligned} \tag{3}$$

Finally, defining $\mathcal{E}_n = \boldsymbol{R}_n \Delta_n$, we obtain

$$\begin{aligned} \boldsymbol{R}_n &= (1-\mu)\boldsymbol{R}_{n-1} + X_n X_n^t \\ \mathcal{E}_n &= (1-\mu)\mathcal{E}_{n-1} + w_n X_n \\ \Delta_n &= \boldsymbol{R}_n^{-1} \mathcal{E}_n. \end{aligned} \tag{4}$$

Based on this set of equations, we are going to examine the convergence properties of the estimation error vector Δ_n.

A. Initialization of the RLS Algorithm

As was mentioned before, our main objective is to find the relation between convergence speed and initialization. Consequently, let us first identify the points of the algorithm that require initialization. From (2), we can see that W_n and $\boldsymbol{R}_n$ are the only two quantities that must be initialized. The vector W_0 is commonly selected to be zero, whereas the matrix $\boldsymbol{R}_0$ is selected to have the form $\boldsymbol{R}_0 = \delta \boldsymbol{I}$, with $\boldsymbol{I}$ the identity matrix and δ a positive scalar. Regarding the selection of the parameter δ, there exist diverse suggestions in the literature. For example, in [10, p. 570] and [21, p. 476], based on observation from practice, a "small" value is proposed. On the other hand, in theoretical studies [6], [7], [9], the assumption of a "large" value is more common. It turns out that the convergence properties of the algorithm differ significantly, depending on the value of δ being "small" or "large." Furthermore, the same value of δ applied to the same set of data can produce an entirely different performance, depending on the value of μ we use. This suggests that the notion of the size of this quantity ("small" or "large") cannot be defined in absolute terms but must be related, in some sense, to the parameter μ. It is exactly this relation we wish to define next.

Let us first introduce some definitions. If $\boldsymbol{F}(z)$ is a matrix function and $f(z)$ a nonnegative scalar function of z with z taking values in some set A_z, we then say that

- $\boldsymbol{F}(z) = \Theta(f)$ when there exist constants c_1, c_2, independent of z, such that $c_1 f(z) \leq \|\boldsymbol{F}(z)\| \leq c_2 f(z)$ for all $z \in A_z$;
- $\boldsymbol{F}(z) = O(f)$ when there exists constant c, independent of z, such that $\|\boldsymbol{F}(z)\| \leq c f(z)$ for all $z \in A_z$;

and where the norm of a matrix $\boldsymbol{F}$ is defined as $\|\boldsymbol{F}\| = (\text{trace}\,\{\boldsymbol{F}^t \boldsymbol{F}\})^{1/2}$.

In our analysis, we will mainly concentrate on cases where $\mu \in [0, \mu_0]$ with $\mu_0 \ll 1$. We can then distinguish a variable as "small" or "large" by comparing it with μ. From the analysis that follows, it turns out that we need to distinguish three sizes for our variables. Specifically, if a variable $a(\mu)$ satisfies $a(\mu) = \Theta(\mu^\alpha)$, then $a(\mu)$ will be characterized as being "small" if $\alpha > 0$, "medium" if $0 \geq \alpha > -1$, and "large" if $-1 \geq \alpha$. Moreover, notice that for small enough μ, we have $\Theta(\mu^\alpha) \ll \Theta(\mu^\beta)$ when $\alpha > \beta$.

Let us now apply the above definition to the initialization of RLS. Consider first the vector W_0. As we said, the most common selection for this quantity is $W_0 = 0$, corresponding to $\Delta_0 = -W_\star$, which is a value of the order of a constant, i.e., a $\Theta(1)$ vector. More generally, we are going to assume $\Delta_0 = \Delta$ with $\Delta = \Theta(1)$ an arbitrary deterministic vector. For the initialization of $\boldsymbol{R}_n$, on the other hand, we will consider $\boldsymbol{R}_0 = \mu^\alpha \boldsymbol{R}$ with $\boldsymbol{R} = \Theta(1)$ an arbitrary deterministic positive definite matrix. According to our definition, $\alpha > 0$ corresponds to a "small" initial value, $0 \geq \alpha > -1$ to a "medium," and $-1 \geq \alpha$ to a "large" one.

Combining the two initialization parts, we have that the complete form of RLS is

$$\begin{aligned} \boldsymbol{R}_n &= (1-\mu)\boldsymbol{R}_{n-1} + X_n X_n^t, \qquad n \geq 1,\ \boldsymbol{R}_0 = \mu^\alpha \boldsymbol{R} \\ \mathcal{E}_n &= (1-\mu)\mathcal{E}_{n-1} + w_n X_n, \qquad n \geq 1,\ \mathcal{E}_0 = \mu^\alpha \mathcal{E} \\ \Delta_n &= \boldsymbol{R}_n^{-1} \mathcal{E}_n \end{aligned} \tag{5}$$

with $\mathcal{E} = \boldsymbol{R}\Delta$ and $\boldsymbol{R}, \mathcal{E}$ are a $\Theta(1)$ matrix and vector, respectively.

Let us now state our assumptions and present some introductory results.

B. Assumptions

A very important point in the study of RLS consists in introducing suitable conditions for the data sequence $\{X_n\}$ that can guarantee some form of boundedness of the inverse matrix $\boldsymbol{R}_n^{-1}$ (persistency of excitation). Next, we are going to present two such conditions—each one guaranteeing persistency of the data—and comment on their specific advantages.

To this end, let us consider the matrix of interest in some detail. From (5), we have that we can write

$$\boldsymbol{R}_n = \mu^\alpha (1-\mu)^n \boldsymbol{R} + \frac{1-(1-\mu)^n}{\mu} \boldsymbol{Q}_n(1-\mu) \tag{6}$$

where we define

$$\boldsymbol{Q}_n(\nu) = \frac{\displaystyle\sum_{j=1}^{n} \nu^{n-j} X_j X_j^t}{\displaystyle\sum_{j=1}^{n} \nu^{n-j}}. \tag{7}$$

Notice that the matrix $\boldsymbol{Q}_n(\nu)$ is the part of $\boldsymbol{R}_n$ influenced by the data sequence, and therefore, it is the part that needs to be controlled. If, for $\boldsymbol{A}, \boldsymbol{B}$, which are two symmetric matrices, we denote with $\boldsymbol{B} \leq \boldsymbol{A}$ the case where the difference $\boldsymbol{A} - \boldsymbol{B}$ is nonnegative definite, we then make the following assumption.

Assumption A1: There exist time n_0 and constant ν_0 satisfying $0 < \nu_0 < 1$ and positive constants c_1, c_2 such that for every $n \geq n_0$ and every $\nu \in [\nu_0\, 1]$, we have $c_1 \boldsymbol{I} \leq \boldsymbol{Q}_n(\nu) \leq c_2 \boldsymbol{I}$. The meaning of this condition is that for any realization of the data and for large enough n, the matrix $\boldsymbol{Q}_n(\nu)$ can be bounded from above and from below uniformly in time n and in ν. This form of persistency is common in the literature for the analysis of RLS [1], [16]. Its main advantage is that it extremely simplifies most proofs; on the other hand, we can see that it is not very realistic since the bounds apply to every realization of the data.

A more reasonable assumption was introduced in [17] and extensively analyzed in [20]. If $\lambda_{\min}\{\boldsymbol{A}\}$ denotes the smallest eigenvalue of a positive definite matrix $\boldsymbol{A}$, $\boldsymbol{Q}_n(1)$ denotes the sample covariance matrix

$$\boldsymbol{Q}_n(1) = \frac{1}{n} \sum_{j=1}^{n} X_j X_j^t \tag{8}$$

and $E\{\cdot\}$ denotes expectation, we then have the following alternative assumption.

Assumption A1′: The data sequence $\{X_n\}$ is stationary, and there exists time n_0 and positive constants c_1, c_2 such that $E\{\lambda_{\min}^{-2}\{\boldsymbol{Q}_{n_0}(1)\}\} \leq c_1$ and $E\{\|X_1\|^4\} \leq c_2$. Assumption A1′, as compared with A1, sets constraints on $\boldsymbol{Q}_n(\nu)$ but only for a specific time instant ($n = n_0$) and a specific value of ν ($\nu = 1$). Furthermore, the constraints involve only moments of certain quantities and not the actual realizations of the random matrix $\boldsymbol{Q}_n(\nu)$, as is the case with A1. In [20], one can find sufficient conditions, set directly on the data sequence $\{X_n\}$, that can guarantee the validity of A1′.

Both assumptions assure validity of our results. Unfortunately, the proof of our main theorem under Assumption A1′ is very lengthy as compared with the simple proof obtained with Assumption A1; on the other hand, it is definitely more elegant and more interesting. A last observation that needs to be made is that in both assumptions, the data sequence $\{X_n\}$ is regarded as being of the order of a constant. It should be noted that this is always possible by proper normalization.

Our second and final assumption refers to the additive noise sequence $\{w_n\}$.

Assumption A2: The additive noise $\{w_n\}$ is stationary, white, zero mean, and independent of the data process $\{X_n\}$ with a variance equal to $\sigma_w^2 = \sigma^2 \mu^\rho$. We assume additive white noise only for simplicity. Similar results can be obtained by considering stationary colored noise independent from the data. Notice that in A2, the noise is not regarded as being of the order of a constant, as was the case for the input data, since we related its power to μ. This is because we intend to analyze the performance of RLS under different SNR levels. According to our definition of size, $\rho > 0$ corresponds to high, $0 \geq \rho > -1$ to medium, and $-1 \geq \rho$ to low SNR.

C. Introductory Results

In the next section, our main goal will be to analyze the convergence properties of the power of the error vector Δ_n. Here, we are going to develop the necessary expressions for the power that will make this analysis possible.

Using (5) and the fact that the additive noise is independent of the data sequence, we can decompose the error power into two parts, namely

$$E\{\Delta_n^t \Delta_n\} = U_n + V_n \tag{9}$$

where

$$U_n = \mu^{2\alpha}(1-\mu)^{2n} \mathcal{E}^t E\{\boldsymbol{R}_n^{-2}\} \mathcal{E} \tag{10}$$

$$V_n = \sigma_w^2 E\left(\text{trace}\left\{\boldsymbol{R}_n^{-1}\left[\sum_{j=1}^{n}(1-\mu)^{2(n-j)} X_j X_j^t\right]\boldsymbol{R}_n^{-1}\right\}\right). \tag{11}$$

Part U_n is due to the fact that our initial estimate W_0 is away from the true value $W_\star$, whereas part V_n is the result of the additive noise. The next theorem introduces suitable estimates for both quantities.

Theorem 1: Let Assumptions A1 (A1′) and A2 be valid, and let n_0 be the time defined in A1 (A1′); then, there exists positive constant μ_0 with $0 < \mu_0 < 1$ such that for any $n \geq n_0$ and any $\mu \in [0, \mu_0]$, we have

$$U_n = \Theta\left(\frac{\mu^{2(\alpha+1)}(1-\mu)^{2n}}{[\mu^{\alpha+1}(1-\mu)^n + 1 - (1-\mu)^n]^2}\right) \tag{12}$$

$$V_n = \sigma_w^2 \Theta\left(\mu \frac{1-(1-\mu)^n}{[\mu^{\alpha+1}(1-\mu)^n + 1 - (1-\mu)^n]^2}\right). \tag{13}$$

Proof: The proof under Assumption A1 is very simple, and it is presented in the Appendix. In the Appendix, we also present the main steps of the proof under Assumption A1′; the complete proof can be found in [19]. ■

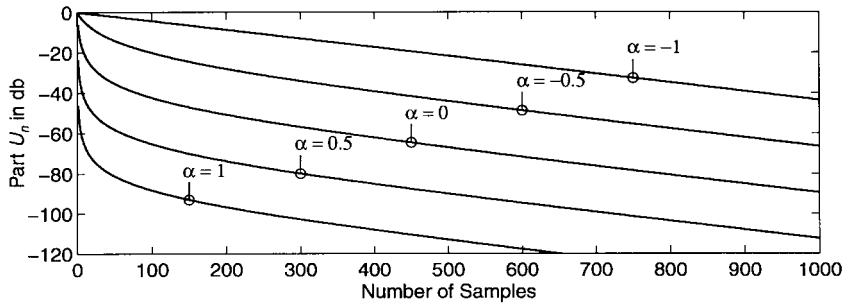
Theorem 1 is the starting point for a detailed study of the two parts U_n, V_n of the estimation error power.

III. Main Results

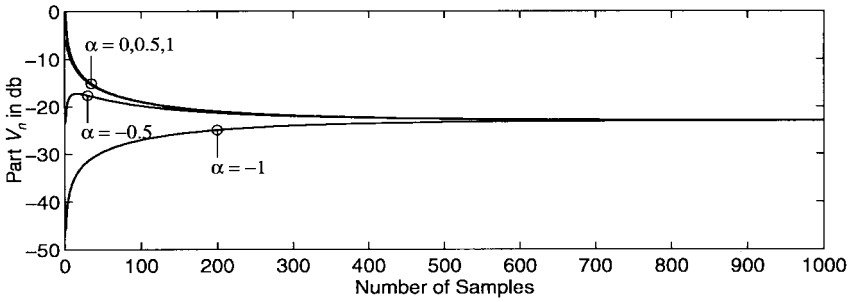
Before proceeding with the analysis of the convergence properties of the error power, let us first introduce our measure of performance. Notice that when any adaptive algorithm is used in practice, it is regarded as having converged when its estimation error power becomes "small." Moreover, the faster the error power becomes "small," the better the algorithm is considered.

According to our definition of size, a quantity is "small" when it is of the order of μ^ϵ with $\epsilon > 0$. Consequently, as measure of speed of convergence, we propose the *settling time* n_Δ required by the power to reach the level μ^ϵ under the constraint, of course, that it will remain below this level for all subsequent time instants. We would like now to stress that we are only interested in estimating the order of magnitude of the settling time n_Δ as a function of μ and not its exact value. The last statement allows for an indirect estimate of n_Δ by first estimating the settling times n_U, n_V of the two parts U_n, V_n of the power and then defining n_Δ as $n_\Delta = \max\{n_U, n_V\}$. This is, of course, possible because, as we said, we are interested only in the order of magnitude of n_Δ.

The next two subsections will be devoted to the development of the necessary estimates for n_U and n_V for all possible combinations of initialization and SNR levels.



(a)



(b)

Fig. 1. Typical form of the two parts of the estimation error power. (a) Part U_n. (b) Part V_n.

A. Study of Part U_n of the Estimation Error Power

In Fig. 1(a), we plot the estimate for U_n obtained in (12) of Theorem 1 for different values of the parameter α. Notice that because of the term $(1-\mu)^{2n}$ in the numerator of this estimate, we have that for $0 < \mu < 1$, it tends exponentially fast to zero [this is also indicated by the asymptotically straight lines in Fig. 1(a)]. Furthermore, by examining the monotonicity properties of the estimate, we conclude that it is strictly decreasing for any value of the parameter α. This suggests that its largest value is achieved for $n = 1$, and since this value is bounded by unity, we have that U_n can at most be of the order of a constant, i.e., $U_n = O(1)$. We now examine the three initialization cases separately.

Case $\alpha > 0$ and $0 < \mu \leq \mu_0 \ll 1$: According to our definition, this case corresponds to a "small" initial value. For $n \geq 1$, we have that $1 - \nu^n = \mu(1 + \cdots + \nu^{n-1}) \geq \mu$ (where for simplicity, from now on, we denote by ν the forgetting factor $1 - \mu$). This means that

$$\frac{\mu^{\alpha+1}\nu^n}{\mu^{\alpha+1}\nu^n + 1 - \nu^n} \leq \frac{\mu^{\alpha+1}\nu^n}{\mu} \leq \mu^\alpha \tag{14}$$

and thus, we conclude that $U_n \leq c\mu^{2\alpha}$ for $n \geq n_0$. In other words, U_n becomes "small" in, at most, n_0 number of steps; therefore, for $\epsilon < 2\alpha$, we have $n_U = \Theta(1)$.

Case $0 \geq \alpha > -1$ and $0 < \mu \leq \mu_0 \ll 1$: For "medium" initialization values, the situation is different. As we said, U_n is decreasing in n; thus, to find n_U, it is sufficient to find the time when U_n becomes of the order of μ^ϵ (since it will remain under this level afterwards). In other words, for positive constants c_1, c_2 (without loss of generality, we select $0 < c_1 < c_2 < 1$), we like to have

$$c_1\mu^\epsilon \leq \frac{\mu^{2(a+1)}\nu^{2n_U}}{(\mu^{a+1}\nu^{n_U} + 1 - \nu^{n_U})^2} \leq c_2\mu^\epsilon. \tag{15}$$

Solving for ν^{n_U} yields the inequalities

$$\frac{c_1\mu^{\epsilon/2-(\alpha+1)}}{1 - c_1\mu^{\epsilon/2} + c_1\mu^{\epsilon/2-(\alpha+1)}} \leq \nu^{n_U} \leq \frac{c_2\mu^{\epsilon/2-(\alpha+1)}}{1 - c_2\mu^{\epsilon/2} + c_2\mu^{\epsilon/2-(\alpha+1)}}. \tag{16}$$

Notice now that $1-c+c\mu^{\epsilon/2-\alpha-1} \leq 1-c\mu^{\epsilon/2}+c\mu^{\epsilon/2-\alpha-1} \leq 1 + c\mu^{\epsilon/2-\alpha-1}$, and consequently, we can widen the bounds in (16) as

$$\frac{1}{1 + c_1^{-1}\mu^{\alpha+1-\epsilon/2}} \leq \nu^{n_U} \leq \frac{1}{1 + \dfrac{1-c_2}{c_2}\,\mu^{\alpha+1-\epsilon/2}}. \tag{17}$$

For $0 < \epsilon < 2(\alpha + 1)$, if we take logarithms in the above relation and use the approximation $\log(1+x) = \Theta(x)$, which is valid for small x, we obtain $n_U = \Theta(\mu^{\alpha-\epsilon/2})$.

Case $-1 \geq \alpha$ and $0 < \mu \leq \mu_0 \ll 1$: This corresponds to a "large" initialization. For any $\epsilon > 0$, notice that $\mu^{\epsilon/2-(\alpha+1)} \ll \mu^{\epsilon/2}$, and thus, we can write $1-c \leq 1-c\mu^{\epsilon/2}+c\mu^{\epsilon/2-(\alpha+1)} \leq 1$. Applying these inequalities in (16), we can widen the two bounds as

$$c_1\mu^{\epsilon/2-(a+1)} \leq \nu^{n_U} \leq \frac{c_2}{1-c_2}\,\mu^{\epsilon/2-(a+1)} \tag{18}$$

and by taking logarithms and using the same approximation as before, we conclude that $n_U = \Theta(\log(\mu^{-1})/\mu)$.

Case $\alpha = 0$ and $\mu = 0$: We consider this case only for completeness because our estimate is also valid for the unit forgetting factor. Notice that the only possibility for having nontrivial initial values is when we select $\alpha = 0$. From (12), we conclude that

$$U_n = \Theta\left(\frac{1}{n^2}\right). \tag{19}$$

In other words, for $\mu = 0$, the convergence is no longer exponential but of the form of $1/n^2$.

B. Study of Part V_n of the Estimation Error Power

In Fig. 1(b), we have the typical form of the estimate of V_n, given by (13), for different values of the initialization parameter α. We recall that part V_n is mainly due to the additive noise and, as we can see from Fig. 1(b), its limiting value is different from zero. Indeed, from (13), we have that V_n tends exponentially fast (when $1 > \mu > 0$) to a limit that is of the form of $\sigma_w^2\Theta(\mu)$. It is possible to have a better estimate of this limiting value using the results in [2, p. 107], [4], [5], or the analysis in [7] and [9]. Specifically, for small μ, one can show that $\lim_{n\to\infty} V_n$ can be efficiently approximated by $\mu\sigma_w^2 \operatorname{trace}\{\boldsymbol{R}_X^{-1}\}$, where $\boldsymbol{R}_X = E\{X_n X_n^t\}$.

The analysis of V_n is more involved as compared with the previous case for two main reasons. First is the variable behavior of V_n as a function of the parameter α [see Fig. 1(b)]. The second reason is that V_n is related to the observation noise variance σ_w^2, which is parametrized to account for the different SNR levels. We recall that in A2, we defined the noise power to be $\sigma_w^2 = \sigma^2\mu^\rho$ modeling with $\rho > 0$ the high, with $0 \geq \rho > -1$ the medium, and with $-1 \geq \rho$ the low SNR case.

Under the above form of noise power, the limiting value of V_n becomes $\Theta(\mu^{\rho+1})$. If $-1 \geq \rho$ (low SNR), the limiting value of V_n is not "small"; therefore, according to our definition, this case does not converge (in the sense that the error power cannot become and remain "small"). This clearly suggests that we need only estimate the settling time n_V for high and medium SNR, whereas for low SNR, we assign to n_V the value infinity. Let us again consider each initialization case separately.

Case $\alpha > 0$ and $0 < \mu < \mu_0 \ll 1$: From (13), we have that there exist positive constant c_1, c_2 such that

$$c_1\sigma_w^2\mu\frac{1-\nu^n}{(\mu^{\alpha+1}\nu^n+1-\nu^n)^2} \leq V_n \leq c_2\sigma_w^2\mu\frac{1-\nu^n}{(\mu^{\alpha+1}\nu^n+1-\nu^n)^2}. \tag{20}$$

Notice now that for $\alpha > 0$ and $n \geq 1$, we have $0 < \mu^{\alpha+1}\nu^n \leq \mu^{\alpha+1} \leq \mu \leq \mu(1+\cdots+\nu^{n-1}) = 1-\nu^n$. Applying this to (20), we can enlarge the bounds as

$$c_1\sigma_w^2\frac{\mu}{4(1-\nu^n)} \leq V_n \leq c_2\sigma_w^2\frac{\mu}{(1-\nu^n)} \tag{21}$$

or equivalently (since $\sigma_w^2 = \sigma^2\mu^\rho$) that

$$V_n = \sigma_w^2\Theta\left(\frac{\mu}{1-\nu^n}\right) = \Theta\left(\frac{\mu^{\rho+1}}{1-\nu^n}\right). \tag{22}$$

Relation (22) suggests that for "small" initialization, V_n is decreasing in n and practically independent of α. To find the settling time n_V, we must distinguish the different SNR levels.

Consider first $\rho > 0$ (high SNR). Since $1-\nu^n \geq \mu$, from (22), we have that $V_n \leq c\mu^\rho$. In other words, V_n, after at most n_0 number of steps, becomes uniformly "small." Thus, for $0 < \epsilon < \rho$, we have $n_V = \Theta(1)$.

For $0 \geq \rho > -1$ (medium SNR), we need to estimate the first time V_n becomes of the order of μ^ϵ. Following a similar approach as in the case of U_n, we can show that for $0 < \epsilon < (1+\rho)$, we have $n_V = \Theta(\mu^{\rho-\epsilon})$.

Case $0 \geq \alpha > -1$ and $0 < \mu < \mu_0 \ll 1$: By setting $x = \nu^n$, we can study the monotonicity properties of V_n and since $0 < \mu^{\alpha+1} < 1$, we can show that V_n, increasing at first, attains a maximum value $\Theta(\mu^{\rho-\alpha})$ and then decreases monotonically to its steady state [see Fig. 1(b), $\alpha = -0.5$]. We must now distinguish two cases. If the maximum value $\Theta(\mu^{\rho-\alpha})$ is "small," i.e., $\rho > \alpha$, then V_n is uniformly "small," and we have convergence in n_0 number of steps. Thus, for $0 < \epsilon < \rho-\alpha$, we have $n_V = \Theta(1)$. If the maximum is not "small," i.e., $\rho \leq \alpha$, then we have to identify the time instant (after the occurrence of the maximum), where V_n reaches the level μ^ϵ, and this will be the settling time n_V. To estimate n_V, if we define $x = 1-\nu^n$, we obtain inequalities involving second-order polynomials in x. By solving these inequalities, we can show that for $0 < \epsilon < 1+\rho$, we have $n_V = \Theta(\mu^{\rho-\epsilon})$.

Case $-1 \geq \alpha$ and $0 < \mu < \mu_0 \ll 1$: For this case, we can show that V_n is monotonically increasing [see Fig. 1(b)]. Since for high and medium SNR we have that the steady state value of V_n is "small," this suggests that V_n will also be "small" for all time instants. In other words, for $0 < \epsilon < 1+\rho$, we have $n_V = \Theta(1)$.

Case $\alpha = 0$ and $\mu = 0$: As we did for U_n, we consider here the case of unit forgetting factor for V_n. From (13) in Theorem 1, we obtain

$$V_n = \Theta\left(\frac{1}{n}\right). \tag{23}$$

In other words, the part of the estimation error power due to the additive noise tends to zero as $1/n$. Comparing this value with the corresponding obtained for U_n, we conclude that for the unit forgetting factor, the estimation error power tends to zero as $1/n$ and is mainly due to the additive noise and not the initial conditions. This is in agreement with [10, pp. 576–578].

IV. Discussion of the Results

In this section, we are going to summarize our results; furthermore, based on our analysis, we will be able to explain several characteristics of the algorithm known from practice. Finally, for every noise level, we are going to propose the initialization that yields the best possible convergence speed.

Table I contains the estimates of the settling times n_U, n_V of the two parts of the estimation error power. For each settling time, we also present the range of values of ϵ for which the estimate is valid. The last column contains the total settling time $n_\Delta = \max\{n_U, n_V\}$ required by the error power to reach the level μ^ϵ. Notice that we consider only high and medium SNR since for low SNR, the settling time n_V is infinite, resulting also in an infinite total settling time n_Δ. To be able to compare the settling times for the different initialization cases, we need to introduce the following definition.

Definition: An initialization α_1 will be preferable to an initialization α_2 if there exists $\epsilon_0 > 0$ such that the first initialization has a smaller settling time for all $\epsilon \in (0, \epsilon_0)$.

In other words, we are interested in values of ϵ that are close to zero corresponding to the largest possible "small" values for the level μ^ϵ.

TABLE I
SETTLING TIMES FOR U_n, V_n AND THE TOTAL ESTIMATION ERROR POWER FOR DIFFERENT COMBINATIONS OF SNR AND INITIALIZATION VALUES

α	n_U		n_V		$n_\Delta = \max\{n_U, n_V\}$	
High SNR ($\rho > 0$)						
$\alpha > 0$	$\Theta(1)$	$0 < \epsilon < 2\alpha$	$\Theta(1)$	$0 < \epsilon < \rho$	$\Theta(1)$	$0 < \epsilon < \min\{2\alpha, \rho\}$
$0 \geq \alpha > -1$	$\Theta(\mu^{\alpha-\epsilon/2})$	$0 < \epsilon < 2(1+\alpha)$	$\Theta(1)$	$0 < \epsilon < \rho - \alpha$	$\Theta(\mu^{\alpha-\epsilon/2})$	$0 < \epsilon < \min\{2(1+\alpha), \rho - \alpha\}$
$-1 \geq \alpha$	$\Theta(\frac{\log(\mu^{-1})}{\mu})$	$0 < \epsilon$	$\Theta(1)$	$0 < \epsilon < 1 + \rho$	$\Theta(\frac{\log(\mu^{-1})}{\mu})$	$0 < \epsilon < 1 + \rho$
Medium SNR ($0 \geq \rho > -1$)						
$\alpha > 0$	$\Theta(1)$	$0 < \epsilon < 2\alpha$	$\Theta(\mu^{\rho-\epsilon})$	$0 < \epsilon < 1 + \rho$	$\Theta(\mu^{\rho-\epsilon})$	$0 < \epsilon < \min\{2\alpha, 1 + \rho\}$
$0 \geq \alpha \geq \rho$	$\Theta(\mu^{\alpha-\epsilon/2})$	$0 < \epsilon < 2(1+\alpha)$	$\Theta(\mu^{\rho-\epsilon})$	$0 < \epsilon < 1 + \rho$	$\Theta(\mu^{\rho-\epsilon})$	$0 < \epsilon < \min\{2(1+\alpha), 1 + \rho\}$
$\rho > \alpha > -1$	$\Theta(\mu^{\alpha-\epsilon/2})$	$0 < \epsilon < 2(1+\alpha)$	$\Theta(1)$	$0 < \epsilon < \rho - \alpha$	$\Theta(\mu^{\alpha-\epsilon/2})$	$0 < \epsilon < \min\{2(1+\alpha), \rho - \alpha\}$
$-1 \geq \alpha$	$\Theta(\frac{\log(\mu^{-1})}{\mu})$	$0 < \epsilon$	$\Theta(1)$	$0 < \epsilon < 1 + \rho$	$\Theta(\frac{\log(\mu^{-1})}{\mu})$	$0 < \epsilon < 1 + \rho$

Using the above definition and focusing on the last column of Table I, we can summarize our results for the three SNR levels as follows.

High SNR ($\rho > 0$): From Table I, we have that with "small" initialization ($\alpha > 0$), RLS converges almost instantly and is basically insensitive to the exact initialization value. For "medium" initialization, the settling time increases with increasing initialization. Finally, with "large" initialization, RLS exhibits the worst possible settling time because we can show that $\Theta(\log(\mu^{-1})/\mu) \gg \Theta(\mu^{-\beta})$ for $0 \leq \beta < 1$. Consequently, in a high SNR environment, initializing with a "small" value is definitely the most preferable initialization since it results in an extremely fast convergence.

Medium SNR ($0 \geq \rho > -1$: For this noise level, the optimum speed of the algorithm is significantly reduced as compared with the previous case. We notice from Table I that we no longer have any convergence in $\Theta(1)$ number of steps. On the other hand, RLS becomes more insensitive to the initialization. Notice that for all $\alpha \geq \rho$, corresponding to the "small" and part of the "medium" initialization values, the performance of RLS is almost indistinguishable. The settling time starts to increase significantly only when the initialization becomes large enough ($\rho > \alpha$). That this is in fact the case can be seen by comparing the settling times for values of ϵ in the intersection of the corresponding intervals. For these values of ϵ, we have $\alpha - \epsilon/2 < \rho - \epsilon$, which yields $\Theta(\mu^{\alpha-\epsilon/2}) \gg \Theta(\mu^{\rho-\epsilon})$. Finally, for medium SNR, we can again show that the largest settling time is obtained with "large" initialization ($-1 \geq \alpha$).

Low SNR ($-1 \geq \rho$: Although, for this case, the settling time is infinite, we can still make some important remarks concerning the performance of RLS. From the analysis of the previous section, we have seen that part U_n of the error power can at most be of the order of a constant. On the other hand, part V_n has a steady-state value that is $\Theta(\mu^{1+\rho})$. According to our definition (when $-1 > \rho$), this corresponds to a "large" value, and therefore, it is significantly larger than U_n. This suggests that the leading part of the error power is V_n. In Section IV-B, we have seen that V_n is decreasing in n for "small" initialization, unimodal for "medium," and increasing for "large" initialization [see Fig. 1(b)]. Since all cases converge to the same steady-state value, we conclude that for low SNR, initialization with a "large" matrix is clearly preferable.

As a general remark we have that for the most practically interesting SNR levels (high and medium), RLS achieves its best performance with "small" initialization. Moreover, the optimal performance is insensitive to the exact "small" value used. This characteristics was also observed in practice [21, p. 476].

Comments: We have seen that for high SNR, if RLS is initialized with a small matrix, it converges almost instantly. Although this property might seem "intuitively obvious," we must stress that this is not at all the case. Consider for instance the LMS algorithm and assume that there is no additive noise (infinite SNR). Even under this ideal condition, the convergence speed of LMS is exponential and of the form of $(1 - \mu c)^n$. It is easy to verify that with this form of convergence, the settling time of LMS is of the order of $\Theta(\log(\mu^{-1})/\mu)$; in other words, it is comparable to the worst possible settling time of RLS. Exponential convergence is common to several known families of adaptive algorithms [8], suggesting that for all these cases, the corresponding settling time is again of the order of $\Theta(\log(\mu^{-1})/\mu)$.

With this last remark in mind, we can definitely say that the convergence speed of RLS, for high SNR and "small" initialization, is exceptional. In addition, however, its convergence speed for "small" or "medium" initialization and high or medium SNR does not follow the common practice of other algorithms. This is because the corresponding settling time is of the form of $\Theta(\mu^{-\beta})$, where $1 > \beta \geq 0$, which is, in order of magnitude, significantly smaller than the settling time $\Theta(\log(\mu^{-1})/\mu)$ required by other known adaptive algorithms.

In our opinion, this very important property stems from the fact that RLS is one of the few algorithms that can estimate exactly $W_\star$ in a constant number of steps when there is no noise. We can see that this is true by considering the estimate W_n for $n \geq N$. If we make the initialization matrix tend to zero ($\alpha \to \infty$), then for any nonzero forgetting factor, we obtain $W_n = W_\star$ [provided the matrix $\boldsymbol{Q}_n(1 - \mu)$ is nonsingular]. To our knowledge, the only other algorithm that has this property is the instrumental variables adaptive

algorithm [22] that satisfies a recursion similar to (2) but uses in the regression vector, instead of X_n, the instruments Z_n. It is thus expected that the instrumental variables algorithm will have convergence properties similar to RLS.

Based on our analysis, we can also make a remark about the tracking capability of the algorithm. RLS, once in steady state, cannot track abrupt changes in $W_\star$ as efficiently as during the initial transient phase. Indeed, if RLS is in steady state and there is a sudden change in the vector $W_\star$, then the continued application of the algorithm corresponds to an initialization with a "large" initial value (with time 0 being the instant of change). This is so because, during steady state, the matrix $\boldsymbol{R}_n$ is of the order of $\Theta(\mu^{-1})$ corresponding to a "large" value. From our analysis, we know that this form of initialization produces the worst settling time (for medium and high SNR). Consequently, when we have an abrupt change in the vector $W_\star$, it is preferable to restart the algorithm, initializing $\boldsymbol{R}_0$ with a "small" value, than continuing to apply RLS.

The initialization we used in our paper corresponds to the soft constraint initialization scheme [10], [12], which consists of adding in the normal least squares criterion the term $(1-\mu)^n\mu^\alpha W_n^t\boldsymbol{R}W_n$. The resulting problem always has a unique solution given by the recursion in (2), whereas the normal least squares has an infinite number of solutions for $n < N$. Our analysis also applies to the case of exact initialization [11]. This scheme consists in finding the minimum norm solution of the least squares problem for time $n < N$. One can show that the minimum norm solution is a limiting case of the soft initialization corresponding to $\boldsymbol{R} = \boldsymbol{I}$ and $\alpha \to \infty$. In other words, exact initialization corresponds to soft initialization with a (very) "small" initial value and, therefore, has the properties of this initialization case.

A last comment we must make is that in [7] and [9], we can find more efficient estimates of the error power than the ones introduced here. The key point in deriving these estimates is the property that two matrices $(P_k, \overline{P}_k)$, corresponding to our $\mu^{-1}\boldsymbol{R}_n^{-1}$, $\mu^{-1}(E\{\boldsymbol{R}_n\})^{-1}$, have norms of the order of a constant. It can be shown that this requirement is met only when $-1 \geq \alpha$ (actually the analysis in [7] and [9] corresponds to $\alpha = -1$), which, as we have seen, is the least important from a practical point of view, as far as transient phase is concerned.

V. Simulations

In this section, we perform several simulations to verify the validity of our theoretical analysis. We consider an FIR system of length $N = 10$, where the vector $W_\star$ is composed of ten random numbers in the interval $[-1, 1]$. The data process $\{X_n\}$ satisfies $X_n = [x_n x_{n-1} \cdots x_{n-9}]^t$, where $\{x_n\}$ is a random ARMA sequence generated by passing white Gaussian noise through an IIR system with transfer function

$$H(z) = \frac{1 + 2z^{-1} + 3z^{-2}}{(1 - 1.1314z^{-1} + 0.64z^{-2})(1 + 0.9z^{-1})}. \quad (24)$$

To the output process $W_\star^t X_n$, we add a zero mean white Gaussian noise $\{w_n\}$ to generate the sequence $\{y_n\}$.

For the initialization of RLS, we consider four values for the parameter α, namely, $\alpha = 1, 0, -0.5, -1$. The initialization matrix $\boldsymbol{R}$ is selected to be $\sigma_x^2\boldsymbol{I}$ with σ_x^2 the variance of x_n; moreover, we select $W_0 = 0$ and a forgetting factor equal to 0.995. We apply RLS to 100 independent sets of data and for each time step n, we average the resulting squared norm of the estimation error vector to form an estimate of the error power at time n.

Fig. 2 depicts the performance of RLS for SNR values 40, 10, and -20 dB (corresponding to high, medium, and low SNR). We notice the exact agreement between simulations and our theoretical analysis. In particular, in Fig. 2(a) (high SNR), we observe that for $\alpha = 1$, we also have a very fast convergence that the settling time increases with decreasing α. In Fig. 2(b) (medium SNR), the performances for $\alpha = 1, 0, -0.5$ are almost indistinguishable, whereas $\alpha = -1$ has a significantly larger settling time. In Fig. 2(c) (low SNR), we can see that $\alpha = -1$ has an overall better performance, as was predicted by our analysis.

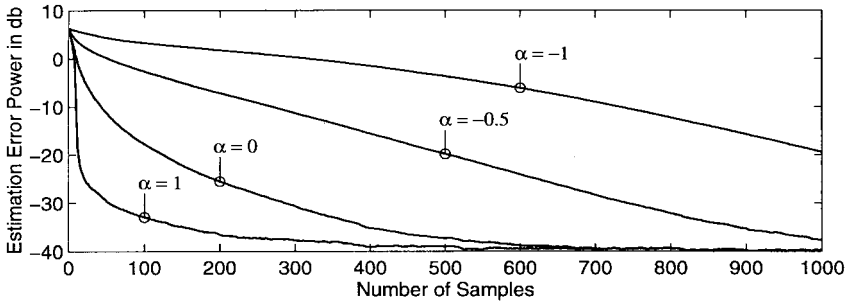
Finally, in Fig. 3, we can see the performance of RLS before and after an abrupt change in the vector $W_\star$. As was explained in Section IV, the convergence speed of the algorithm is significantly reduced if the algorithm is in steady state as compared with the corresponding speed during the initial transient phase. We observe that the simulations support our remark.

VI. Conclusion

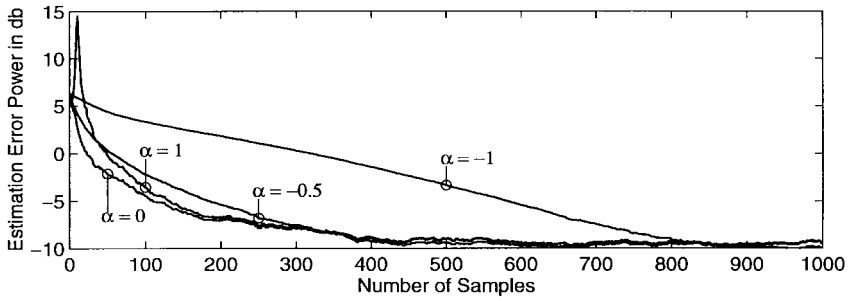
We presented a theoretical analysis of the convergence properties of the RLS algorithm. Specifically, we examined the dependence of the convergence speed to the initialization of the sample covariance matrix and the observation noise level. We proved that RLS, in a high SNR environment, converges in a finite number of steps if its sample covariance matrix is initialized with a matrix of "small" norm. The speed of convergence is decreased as the norm of the initialization matrix is increased. In a medium SNR environment, the optimal speed of the algorithm is reduced significantly as compared with the optimal speed of the previous case, but the algorithm becomes more insensitive to initialization. Finally, in a low SNR environment, it is preferable to initialize the algorithm with a matrix of "large" norm since this yields the best overall performance. Our analysis has also indicated that for high and medium SNR levels, the convergence properties of RLS for "small" or "medium" initialization are exceptional as compared with the corresponding properties of other commonly used adaptive algorithms. In fact, RLS can have an order of magnitude better convergence speed than most such algorithms.

Appendix

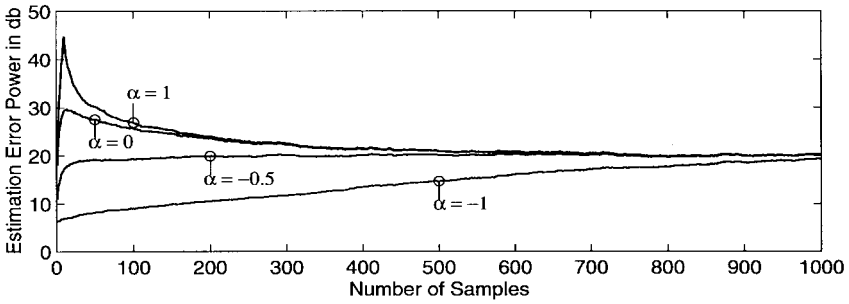
Proof of Theorem 1 under Assumption A1: We will only show the expression for U_n. Following similar steps, we can show the corresponding expression for V_n. For the positive matrix $\boldsymbol{R}$, we have $\lambda_{\min}(\boldsymbol{R})\boldsymbol{I} \leq \boldsymbol{R} \leq \lambda_{\max}(\boldsymbol{R})\boldsymbol{I}$. Combining (6) and Assumption A1, we conclude that, for small enough μ, if we select $c' = \min\{c_1, \lambda_{\min}(\boldsymbol{R})\}$ and



(a)



(b)



(c)

Fig. 2. Performance of RLS for different initializations. (a) SNR $= 40$ dB. (b) SNR $= 10$ dB. (c) SNR $= -20$ dB.

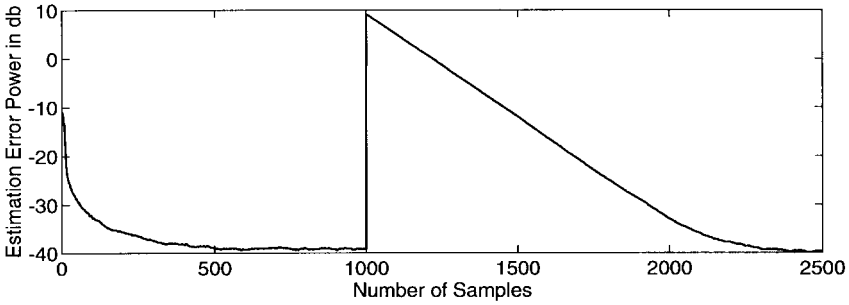


Fig. 3. Difference in performance of RLS during initial transient phase and after an abrupt change in the vector $W_\star$.

$c'' = \max\{c_2, \lambda_{\max}(\boldsymbol{R})\}$, then

$$c'\left(\mu^\alpha \nu^n + \frac{1-\nu^n}{\mu}\right)\boldsymbol{I} \le \boldsymbol{R}_n \le c''\left(\mu^\alpha \nu^n + \frac{1-\nu^n}{\mu}\right)\boldsymbol{I}. \tag{25}$$

Applying this expression in (10) yields the desired bounds. In fact, it is easy to see that we have a much stronger result. The proposed bounds bound any realization of the squared norm of the error and not only its expectation (which is the error power).

Proof of Theorem 1 under Assumption $\mathcal{A}1'$: We will only highlight the main steps of the proof.

Step 1: We first show that

$$\sup_{n\ge k} E\{\lambda_{\min}^{-s}[\boldsymbol{Q}_n(\nu)]\} \le \frac{2^s}{\nu^{2ks}} E\{\lambda_{\min}^{-s}[\boldsymbol{Q}_k(1)]\}\boldsymbol{I}. \tag{26}$$

In other words, if the sample covariance matrix $\boldsymbol{Q}_k(1)$ has a finite sth-order moment for the inverse of its smallest eigenvalue, then the same property will be true for the matrix $\boldsymbol{Q}_n(\nu)$ uniformly in time and for any forgetting factor away from zero.

The proof of this statement follows the same lines of the corresponding proof in [20] with only some minor modifi-

cation to include the case of unit forgetting (something that was not possible with the proof in [20]); for more details, see [19]. With this step, we have boundedness of the second-order moments of the inverse of $\lambda_{\min}[\boldsymbol{Q}_n(\nu)]$ for $n \geq n_0$ and any forgetting factor away from zero.

Step 2: We show that

$$E\{\lambda^s_{\min}[\boldsymbol{Q}_k(\nu)]\} \leq E\{\lambda^s_{\max}[\boldsymbol{Q}_k(\nu)]\} \leq E\{\|X_1\|^{2s}\}. \quad (27)$$

This can be easily shown using Minkowski's inequality. Because of Assumption A1′, we have that the second-order moment of the smallest and largest eigenvalue of $\boldsymbol{Q}_n(\nu)$ is bounded uniformly in time and in ν.

Step 3: Since for any symmetric positive definite matrix $\boldsymbol{A}$ we have $\lambda_{\min}(\boldsymbol{A})\boldsymbol{I} \leq \boldsymbol{A} \leq \lambda_{\max}(\boldsymbol{A})\boldsymbol{I}$, we can conclude from (6) that

$$\left\{\mu^\alpha \nu^n c' + \frac{1-\nu^n}{\mu}\lambda_{\min}[\boldsymbol{Q}_n(\nu)]\right\}\boldsymbol{I} \leq \boldsymbol{R}_n \leq \left\{\mu^\alpha \nu^n c'' + \frac{1-\nu^n}{\mu}\lambda_{\max}[\boldsymbol{Q}_n(\nu)]\right\}\boldsymbol{I} \quad (28)$$

where c', c'' are, respectively, the smallest and the largest eigenvalue of $\boldsymbol{R}$. Applying these inequalities in (10), we find the following bounds for U_n.

$$E\left(\frac{\mu^{2(\alpha+1)}\nu^{2n}\|\mathcal{E}\|^2}{\{\mu^{\alpha+1}\nu^n c'' + (1-\nu^n)\lambda_{\max}[\boldsymbol{Q}_n(\nu)]\}^2}\right) \leq U_n \leq E\left(\frac{\mu^{2(\alpha+1)}\nu^{2n}\|\mathcal{E}\|^2}{\{\mu^{\alpha+1}\nu^n c' + (1-\nu^n)\lambda_{\min}[\boldsymbol{Q}_n(\nu)]\}^2}\right). \quad (29)$$

Step 4: Using Jensen's inequality, we can show that the expectation in the lower bound of U_n can be bounded from below by the expression required by Theorem 1.

Step 5: Using the property that for nonnegative quantities a_1, a_2, b_1, b_2, we have

$$\frac{a_1+a_2}{b_1+b_2} \leq \max\left\{\frac{a_1}{b_1}, \frac{a_2}{b_2}\right\} \quad (30)$$

and thus

$$\left(\frac{a_1+a_2}{b_1+b_2}\right)^2 \leq \max\left\{\left(\frac{a_1}{b_1}\right)^2, \left(\frac{a_2}{b_2}\right)^2\right\} \leq \left(\frac{a_1}{b_1}\right)^2 + \left(\frac{a_2}{b_2}\right)^2. \quad (31)$$

We can show that the expectation in the upper bound of U_n divided by the desired expression can be bounded from above by a constant. This concludes the proof. ■

REFERENCES

[1] B. D. O. Anderson and C. R. Johnson, Jr., "Exponential convergence of adaptive identification and control algorithms," *Automatica*, vol. 18, no. 1, pp. 1–13, 1982.

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations.* New York: Springer-Verlag, 1990.

[3] S. Bittanti and M. Campi, "Adaptive RLS algorithms under stochastic excitation-L^2 convergence analysis," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 963–967, Aug. 1991.

[4] J. A. Bucklew, T. G. Kurtz, and W. A. Sethares, "Weak convergence and local stability properties of fixed step size recursive algorithms," *IEEE Trans. Inform. Theory*, vol. 39, pp. 966–978, May 1993.

[5] E. Eleftheriou and D. D. Falconer, "Tracking properties and steady state performance of RLS adaptive filter algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1097–1110, 1986.

[6] E. Eweda and O. Macchi, "Convergence of the RLS and LMS adaptive filters," *IEEE Trans. Circuits Syst.*, vol. 34, pp. 799–803, July 1987.

[7] L. Guo, L. Ljung, and P. Priouret, "Performance analysis of forgetting factor RLS," *Int. J. Adapt. Contr. Signal Process.*, vol. 7, pp. 525–537, 1993.

[8] L. Guo and L. Ljung, "Exponential stability of general tracking algorithms," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 1376–1387, Aug. 1995.

[9] ———, "Performance analysis of general tracking algorithms," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 1388–1402, Aug. 1995.

[10] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[11] N. E. Hubing and S. T. Alexander, "Statistical analysis of the fast exact initialization of RLS algorithms," in *Proc. 23rd Asilomar Conf. Signals, Syst. Comput.*, Oct. 1989.

[12] ———, "Statistical analysis of the soft constraint initialization of RLS algorithms," in *Proc. ICASSP*, Albuquerque, NM, 1990, pp. 1277–1280.

[13] L. Ljung and S. Gunnarsson, "Adaptive tracking in system identification, A survey," *Automatica*, vol. 26, no. 1, pp. 7–22, 1990.

[14] L. Ljung and P. Priouret, "A result on the mean square tracking error," *Int. J. Adapt. Contr. Signal Process.*, vol. 5, no. 4, pp. 231–250, 1991.

[15] ———, "Remarks on the mean square tracking error," *Int. J. Adapt. Contr. Signal Process.*, vol. 5, no. 6, pp. 395–403, 1991.

[16] R. L. Lozano, "Convergence analysis of recursive identification algorithms with forgetting factor," *Automatica*, vol. 19, pp. 95–97, 1983.

[17] O. Macci and E. Eweda, "Compared speed and accuracy of RLS and LMS algorithms with constant forgetting factors," *Traitement du Signal*, vol. 22, pp. 255–267, 1988.

[18] G. V. Moustakides and S. Theodoridis, "Fast Newton transversal filters, a new class of adaptive estimation algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 2184–2193, Oct. 1991.

[19] G. V. Moustakides, "Performance analysis of constant forgetting factor RLS," *Comput. Technol. Inst. Rep.*, Nov. 1996.

[20] M. Niedźwiecki and L. Guo, "Nonasymptotic results for finite-memory WLS filters," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 198–206, Feb. 1991.

[21] S. Orfanidis, *Optimum Signal Processing, an Introduction*, 2nd ed. New York: McGraw-Hill, 1990.

[22] T. Söderström and P. G. Stoica, *Instrumental Variable Methods for System Identification.* Berlin: Springer-Verlag, 1983.

George V. Moustakides was born in Drama, Greece, in 1955. He received the diploma in electrical engineering from the National Technical University of Athens, Greece, in 1979, the M.Sc. in systems engineering from the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, in 1980, and the Ph.D. in electrical engineering from Princeton University, Princeton NJ, in 1983.

From 1983 to 1986, he held a research position at the Institut de Resherche en Informatique et Systemes Aleatoires (IRISA-INRIA), Rennes, France, and, from 1987 to 1990, he held a research position at the Computer Technology Institute (CTI) of Patras, Patras, Greece. From 1991 to 1996, he was an Associate Professor with the Department of Computer Engineering and Informatics, University of Patras, and since 1996, he has been a Professor in the same department. His interests include adaptive estimation algorithms, design of classical filters, theory of optimal stopping times, and biomedical signal processing.