Performance of the Forgetting Factor RLS during the Transient Phase

George V. Moustakides

Department of Computer Engineering and Informatics, University of Patras, 26500 Patras, GREECE, also Computer Technology Institute (CTI) of Patras, P.O. Box 1122, 21100 Patras, GREECE.

ABSTRACT

We consider the convergence properties of the forgetting factor RLS algorithm in a stationary data environment. We study the dependence of the speed of convergence of RLS with respect to the initialization of the input sample covariance matrix and with respect to the observation noise level. By obtaining estimates of the settling time we show that RLS, in a high SNR environment, when initialized with a matrix of small norm, has a very 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, but RLS becomes more insensitive to initialization. Finally in a low SNR environment it is preferable to start the algorithm with a matrix of large norm.

1. INTRODUCTION

The Recursive Least Squares (RLS) algorithm is one of the most well known algorithms used for adaptive filtering and system identification. Its success is mainly due to its exceptionally fast convergence speed that is considered as optimal in practice and as a measure for comparison (and desired goal) for other algorithms.

Due to its nonlinear nature, the theoretical study of RLS seems to be quite complicated. Complexity increases significantly if we consider the forgetting factor RLS version, which is the most useful version of the algorithm since it is applied in problems where tracking is necessary. Several works in the literature deal with the problem of convergence of RLS in a stationary environment and its corresponding performance at steady state [1],[2],[4],[3],[6].

Although the performance of the algorithm, in a stationary environment, during the transient phase is considered well studied, certain observations coming from practice cannot be explained in a satisfactory manner with the existing theory. Specifically, it is observed in practice that RLS has a much faster convergence rate if the sample covariance matrix (computed in the algorithm) is initialized with a "small" positive definite matrix (usually of the form of $\delta \boldsymbol{I}$) [5, page 484],[8, page 476] suggesting that the initialization with a "large" matrix results in an inferior performance. Unfortunately the existing theory is not capable of distinguishing this performance of the algorithm, meaning that there is a need for further analysis that pays special attention to initialization.

With the present paper we attempt to make a complete study of the performance of RLS under different initialization cases. We will show, by studying the mean and the covariance of the estimation error vector, that the performance of the algorithm depends strongly on the initialization and the observation noise level. To compare the different initialization cases and be able to suggest the most preferable one, we will use the settling time as a measure of speed of convergence. With the help of this measure we will be able to show theoretically that the initialization with a "small" matrix is preferable for cases of high and medium SNR, while in the case of low SNR a "large" matrix is preferable.

2. BACKGROUND MATERIAL

Let us consider the following linear system

$$y_n = X_n^t W_o + w_n \quad n \geq 0 \tag{1}$$

where $\{y_n\}$ is the measurable scalar observation sequence, $\{X_n\}$ the measurable vector input data sequence, $\{w_n\}$ the additive observation noise and W_o an unknown deterministic time invariant vector.

Since we are interested only in convergence speed and not in complexity and computational robustness, we assume infinite accuracy. Then we can show that RLS is equivalent to the following algorithm

$$\begin{aligned} \boldsymbol{R}_n &= (1-\mu)\boldsymbol{R}_{n-1} + X_n X_n^t \\ \mathcal{E}_n &= (1-\mu)\mathcal{E}_{n-1} + w_n X_n \\ \Delta_n &= \boldsymbol{R}_n^{-1}\mathcal{E}_n \end{aligned} \tag{2}$$

where $\Delta_n = W_n - W_o$ is the estimation error vector with W_n the estimate of W_o at time n, μ is the step size and we denote with $\nu = 1-\mu$ the forgetting factor.

0-7803-3629-1/96/$5.00 @ 1996 IEEE

2.1. Initialization of the RLS Algorithm

Initialization in the RLS algorithm is required in two places. That is in W_n (i.e. W_0) and in the matrix $\boldsymbol{R}_n$ (i.e. $\boldsymbol{R}_0$). The vector W_0 is commonly selected to be zero while $\boldsymbol{R}_0 = \delta \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix and δ a constant which is either "small" [5, page 484],[8, page 476] or "large" [4]. The convergence properties of the algorithm are completely different depending on the value of δ being "small" or "large". A fact that needs to be stressed here is that the same value of δ applied to the same set of data can have a completely different performance depending on the value of the step size μ we use. This suggests that the notion of the size ("small" or "large") cannot be defined in absolute terms but only in connection with the step size μ.

In our analysis we will be concerned with cases where the step size is small, that is, $\mu \in [0, \mu_0]$ with $\mu_0 \ll 1$. We can then distinguish a variable as "small" or "large" by comparing it to the step size μ. Actually we need to distinguish three different sizes for the variables. Specifically if a variable $a(\mu)$ satisfies $a(\mu) = \Theta(\mu^\alpha)$[1] then $a(\mu)$ can be characterized as "small" if $\alpha > 0$, as "medium" if $0 \geq \alpha > -1$ and as "large" if $-1 \geq \alpha$.

Let us now see how we can initialize the RLS algorithm using the above definition. Consider first the vector W_0. The most common selection for initialization is $W_0 = 0$ corresponding to $\Delta_0 = -W_o$ which is a $\Theta(1)$ vector. For our study we will more generally assume that $\Delta_0 = \Delta$ where $\Delta = \Theta(1)$ is a deterministic vector. For the initialization of $\boldsymbol{R}_n$ we will assume that $\boldsymbol{R}_0 = \mu^\alpha \boldsymbol{R}$ with $\boldsymbol{R}$ a deterministic positive definite matrix satisfying $\boldsymbol{R} = \Theta(1)$. Clearly case $\alpha > 0$ corresponds to a "small" initial value, $0 \geq \alpha > -1$ to a "medium" and $-1 \geq \alpha$ to a "large" one. Referring to the algorithm in (2) this results in $\boldsymbol{R}_0 = \mu^\alpha \boldsymbol{R}$ and $\mathcal{E}_0 = \mu^\alpha \mathcal{E}$ with $\mathcal{E} = \boldsymbol{R}\Delta = \Theta(1)$.

3. ASSUMPTIONS AND MAIN RESULTS

We will assume that the process $\{X_n\}$ is generated by the following linear state space model

$$\begin{aligned} \xi_n &= \boldsymbol{D}\xi_{n-1} + \boldsymbol{E}\zeta_n \\ X_n &= \boldsymbol{F}\xi_n \end{aligned} \tag{3}$$

where the system $\{\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}\}$ is output reachable, that is, the matrix $[\boldsymbol{FE}, \boldsymbol{FDE}, \ldots, \boldsymbol{FD}^{\tau-1}\boldsymbol{E}]$ is of full rank, with τ being the degree of the minimal polynomial of $\boldsymbol{D}$. We also assume that the matrix $\boldsymbol{D}$ has all its eigenvalues strictly inside the unit circle. This model generates a data sequence $\{X_n\}$ with rational spectra. We need to make the following assumptions regarding certain processes involved in our analysis.

[1] $a(f) = \Theta(f)$ means that the norm of $a(f)$ is bounded from above and below by f times some constant while $a(f) = O(f)$ that the norm is bounded only from above.

A1. The input process $\{\zeta_n\}$ in (3) is a stationary zero mean white noise sequence that satisfies the following conditions

- There exist constants $K > 0$, $\gamma > 0$, $x_0 > 0$ such that for all vectors β (of proper length) with $\|\beta\| = 1$ we have for the probability $P(|\beta^t \zeta_1| \leq x) \leq Kx^\gamma$, for all $0 \leq x \leq x_0$.
- $E\{\|\zeta_n\|^8\} < \infty$, where $E\{\cdot\}$ denotes expectation.

A2. The observation noise $\{w_n\}$ in (1) is white, zero mean, with bounded variance σ_w^2. The noise $\{w_n\}$ is also independent of the process $\{\zeta_n\}$ and thus of the data process $\{X_n\}$.

Next our main goal is to study the behavior of the power of the estimation error vector Δ_n for the RLS algorithm defined in (2). For our study we are going to assume that Assumptions *A1, A2* are valid. Although not explicitly stated we assumed (and will continue to assume) that the moments of the process $\{\zeta_n\}$ and thus of $\{X_n\}$ are $\Theta(1)$ quantities. This will not be the case though with the variance σ_w^2 of the observation noise because our intention is to study RLS in a high, medium and low SNR environment. Consequently, later in this section, we will relate σ_w^2 to the step size μ.

Let us now try to analyze the power $E\{\|\Delta_n\|^2\}$. The power satisfies

$$E\{\|\Delta_n\|^2\} = \|E\{\Delta_n\}\|^2 + \text{trace}\{\text{Cov}\{\Delta_n\}\} \tag{4}$$

where $\text{Cov}\{T\} = E\{(T - E\{T\})(T - E\{T\})^t\}$ denotes the covariance of the random vector T. Because of the above decomposition we will study separately the mean and the covariance.

By assumption the additive observation noise $\{w_n\}$ is a zero mean white process independent of $\{\zeta_n\}$, thus it will also be independent of the data process $\{X_n\}$ and consequently of the matrix $\boldsymbol{R}_n$. This allows us to write

$$E\{\Delta_n\} = E\{\boldsymbol{R}_n^{-1}\mathcal{E}_n\} = \mu^\alpha \nu^n E\{\boldsymbol{R}_n^{-1}\}\mathcal{E} \tag{5}$$

and

$$\text{Cov}\{\Delta_n\} = U_n + V_n \tag{6}$$

where

$$U_n = \mu^{2\alpha}\nu^{2n}\text{Cov}\{\boldsymbol{R}_n^{-1}\mathcal{E}\} \tag{7}$$

$$V_n = \sigma_w^2 E\Big\{\boldsymbol{R}_n^{-1}\Big(\sum_{j=1}^{n} \nu^{2(n-j)} X_j X_j^t\Big)\boldsymbol{R}_n^{-1}\Big\} \tag{8}$$

We have now the following theorem where we estimate the mean and the two parts of the covariance matrix.

Theorem 1: If $\mu \in [0, \mu_0]$ with $\mu_0 < 1$ then there exists integer n_0 independent of μ such that

the following relations hold for $n \geq n_0$

$$E\{\Delta_n\} = \Theta\left(\frac{\mu^{\alpha+1}\nu^n}{\mu^{\alpha+1}\nu^n + 1 - \nu^n}\right) \quad (9)$$

$$U_n = O\left(\mu \frac{\mu^{2(\alpha+1)}\nu^{2n}(1-\nu^n)}{\left(\mu^{\alpha+1}\nu^n + 1 - \nu^n\right)^4}\right) \quad (10)$$

$$V_n = \sigma_w^2 \Theta\left(\mu \frac{1-\nu^n}{\left(\mu^{\alpha+1}\nu^n + 1 - \nu^n\right)^2}\right) \quad (11)$$

Proof: The proof can be found in [7]. ■

The above theorem constitutes the main tool for studying the mean and the covariance of the estimation error vector.

Next we are confronted with the problem of estimating the speed of convergence of the RLS algorithm for the different initialization cases. We would like to define a quantity that measures the speed in a way that is consistent with the practical feeling. Notice that when we apply RLS we usually pay attention to the power $E\{\|\Delta_n\|^2\}$ and expect this quantity to become "small". Since the notion of "small" is now well defined we will use it to estimate the time n_s (settling time) required by the power to achieve "small" values. More specifically we will try to estimate the smallest possible time n_s for which we have

$$E\{\|\Delta_n\|^2\} = O(\mu^{2\epsilon}), \quad \text{for all } n \geq n_s \quad (12)$$

where $\epsilon > 0$. If such a condition cannot be satisfied for any $\epsilon > 0$ and any n then $n_s = \infty$. As we will see this can happen when the SNR is low.

Because of the decomposition we made to the power with (4),(6) we can see that the settling time n_s is the largest among the three settling times corresponding to the three parts that constitute the power. Specifically $n_s = \max\{n_m, n_u, n_v\}$ where n_m, n_u, n_v are the smallest time instants for which respectively we have

$$\begin{aligned} E\{\Delta_n\} &= O(\mu^{\epsilon}) \quad \text{for all } n \geq n_m \\ U_n &= O(\mu^{2\epsilon}) \quad \text{for all } n \geq n_u \\ V_n &= O(\mu^{2\epsilon}) \quad \text{for all } n \geq n_v \end{aligned} \quad (13)$$

There is a slight ambiguity in our definition for the settling time coming from the fact that the parameter ϵ is not explicitly defined. We resolve this problem with the following definition:

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

With the above definition we efface the ambiguity by considering values of ϵ corresponding to the largest possible "small" values for the expression $\mu^{2\epsilon}$.

α	n_s
High SNR ($\rho > 0$)	
$\alpha > 0$	$\Theta(1)$
$0 \geq \alpha > -1$	$\Theta(\mu^{\alpha-\epsilon})$
$-1 \geq \alpha$	$\Theta(\frac{\log(\mu^{-1})}{\mu})$
Medium SNR ($0 \geq \rho > -1$)	
$\alpha > 0$	$\Theta(\mu^{\rho-2\epsilon})$
$0 \geq \alpha \geq \rho$	$\Theta(\mu^{\rho-2\epsilon})$
$\rho > \alpha > -1$	$\Theta(\mu^{\alpha-\epsilon})$
$-1 \geq \alpha$	$\Theta(\frac{\log(\mu^{-1})}{\mu})$

Table 1. Estimates of the settling for different combinations of the parameters α and ρ.

To proceed with the estimation of the settling time n_s we must distinguish different SNR environments. We will thus assume that $\sigma_w^2 = \Theta(\mu^{\rho})$ where ρ is a real parameter. According to our definition, $\rho > 0$ corresponds to high SNR, $0 \geq \rho > -1$ to medium and $-1 \geq \rho$ to low SNR. Notice that under the above form of noise power the limiting value of V_n can be shown to be of the form $\Theta(\mu^{\rho+1})$. Clearly if $-1 \geq \rho$ (low SNR) the limiting value is no longer "small" and the algorithm has a bad steady state performance. According to our definition, such a case has infinite settling time.

We can estimate the three different settling times n_m, n_u, n_v using Theorem 1 and form Table 1 for the settling time $n_s = \min\{n_m, n_u, n_v\}$ as a function of α and ρ. We must stress that for each case listed in the table there exists an interval $(0\ \epsilon_0)$ where the estimate of the settling time is valid.

4. DISCUSSION OF THE RESULTS

Let us consult Table 1 and try to draw conclusions for the performance of RLS.

Case of High SNR ($\rho > 0$)

From Table 1, by comparing the different expressions for n_s, we have that the settling time is increasing with decreasing α. For "small" initial values ($\alpha > 0$) RLS converges almost instantly and is basically insensitive to "small" initialization. For "medium" initialization values the settling time increases with increasing initial value. Finally for "large" initial values we have the worst possible settling time.

Case of Medium SNR ($0 \geq \rho > -1$)

In this noise environment the optimum speed of the algorithm is significantly reduced as compared to the previous case. On the other hand RLS seems to be rather insensitive to the initialization value. For all $\alpha \geq \rho$, corresponding to "small" and part of "medium" initialization values, the

performance of RLS is almost indistinguishable. The settling time starts to increase significantly only when the initial value becomes large enough ($\rho > \alpha$) and continues to have the worst performance for "large" values.

Case of Low SNR ($-1 \geq \rho$)

Even though the settling time is infinite for this case we can still draw conclusions regarding the most preferable initialization. It is possible to show that the leading part of the power for this case is part V_n of the covariance matrix. This part, can be shown, to have smaller values when the algorithm is initialized with "large" initial value. This suggests that for this SNR case, initialization with a "large" value is preferable [7].

Comments: For high SNR the optimum settling time is $\Theta(1)$ while for medium SNR it becomes $\Theta(\mu^{\rho-2\epsilon})$. In other words the optimum speed of convergence for RLS depends on the SNR value and increases with increasing SNR. Also for the most practically interesting SNR cases (high and medium SNR) the performance of RLS seems to have a limited sensitivity to initialization provided the initial matrix R_0 is small enough. This characteristic was also observed in practice [8, page 476].

The RLS algorithm, once in steady state, has a reduced ability to track abrupt changes in the regression vector W_o as compared to its convergence speed during the initial transient phase. Indeed if RLS is in steady state then R_n is of the order of $\Theta(\mu^{-1})$, that is, a "large" value and we have seen that this yields the worst possible settling time (for medium and high SNR).

5. SIMULATIONS

We consider an FIR system where the vector W_o 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 an ARMA sequence generated by passing white noise through an IIR system with transfer function

$$H(z) = \frac{1 + 2z^{-1} + 3z^{-2}}{(1 - 1.1314z^{-1} + 0.64z^{-2})(1 + 0.9z^{-1})} \tag{14}$$

To the output process $W_o^t X_n$ we add a zero mean white noise $\{w_n\}$.

We apply the RLS algorithm with forgetting factor $\nu = 0.995$. For the initialization we use $\alpha = 1, 0, -0.5, -1$. The initialization matrix R is selected to be $\sigma_x^2 I$, with σ_x^2 the variance of x_n and $W_0 = 0$. We apply the algorithm on 100 independent sets of data and for each time step n, we compute the corresponding sample mean estimation error power. Figs. 1(a), (b) and (c) depict the performance of RLS for SNR values 40 db, 10 db and -20 db (corresponding to high, medium and low SNR). We can see that there is exact agreement between the simulations and our conclusions in Section 4.

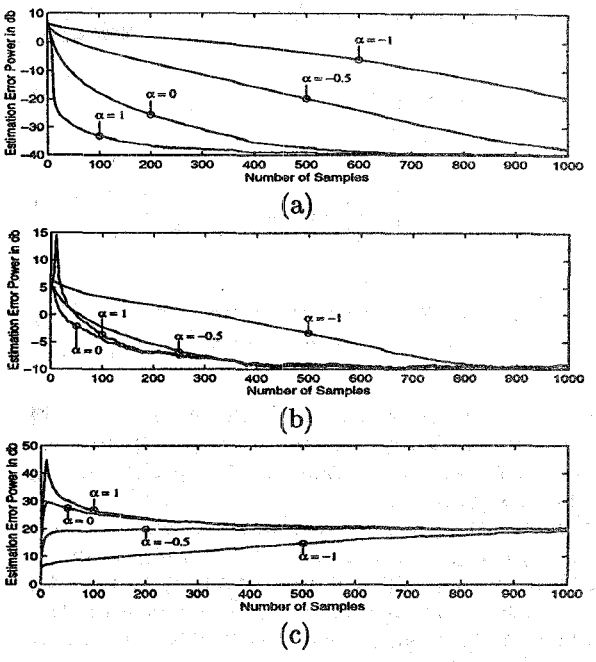

Figure 1. Performance of RLS for different initializations, (a) SNR=40 db, (b) SNR=10 db, (c) SNR=-20 db.

REFERENCES

[1] B.D.O. Anderson and C.R. Johnson Jr., *Automatica*, vol. 18, no. 1, pp. 1-13, 1982.

[2] S. Bittanti and M. Campi, "Adaptive RLS algorithms under stochastic excitation-L^2 convergence analysis," *IEEE Transactions on Autom. Contr.*, vol. 36, no. 8, pp. 963-967, Aug. 1991.

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

[4] E. Eweda and O. Macchi, "Convergence of the RLS and LMS adaptive filters," *IEEE Trans. on Circ. and Syst.*, vol. 34, no. 7, pp. 799-803, Jul. 1987.

[5] S. Haykin, *Adaptive Filter Theory*, 2nd edition. Englewood Cliffs, NJ: Prentice Hall, 1991.

[6] 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.

[7] G.V. Moustakides, "Study of the transient phase of the forgetting factor RLS," submitted to the *IEEE Trans. on Signal Proc.*, under review.

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