Roundoff Error Analysis of the Recursive Least Square Algorithm

Sasan H. Ardalan
S.T. Alexander

CENTER FOR COMMUNICATIONS AND SIGNAL PROCESSING

North Carolina State University
Raleigh, NC

August 1984

CCSP-TR-84/8

ABSTRACT

A closed form analytical solution is derived for the mean square error due to roundoff in the finite wordlength implementation of the Recursive Least Squares adaptive filter. The result indicates that the algorithm diverges when the predictor coefficients are quantized by rounding in the predictor update recursion. The error term due to rounding in the predicted desired signal is seen to be initially large but relaxes to an additive term as the number of iterations increases. The results closely parallel those obtained in the roundoff error analysis of the Least Mean Squares algorithm where the steady state error power increases inversely to loop gain.

Introduction

The Recursive Least Squares (RLS) algorithm has found wide application in adaptive filtering applications [1,2,3]. However, finite wordlength effects on the digital implementation of this algorithm have not been thoroughly analyzed. Whereas simulation studies exist for the digital implemention of the closely related Kalman filter [7,8], to date no closed form analytical results have been presented.

This paper presents a closed form analytical solution of the steady state error power of the finite wordlength implementation of the RLS algorithm. The result indicates that the algorithm will diverge as the number of iterations approaches infinity. This is in agreement with previous researchers' simulations of the finite wordlength effects on the RLS algorithm which indicate that it diverges as the number of iterations becomes large [6,9]. This result seems to conflict with the fact that the Kalman gain approaches zero as the number of iterations increases. However, it is quite similar to the Least Mean Squares algorithms where the error power due to roundoff is inversely proportional to the loop gain [6,7,10]. The result shows that the divergence phenomena is associated with the rounding in the RLS coefficient update formula. The error due to the estimate rounding consists of two terms. One term is inversely related to the number of iterations and the second term is a constant related to the number of quantization bits.

The result is quite useful in the finite wordlength implementation of the fast transversal RLS algorithm. In this algorithm the forward and backward linear prediction coefficients of the input samples are calculated along with the estimated impulse response in an echo cancellation application for example. Since we show that rounding in the coefficient update formula leads to divergence, the largest possible register size should be allocated to the coefficients. Furthermore, after the initial fast convergence of the RLS algorithm, either the coefficients must be frozen, or as suggested in [8,9], a switch to a different algorithm such as LMS should be done. The register size of the error term or scalar estimate is not as critical.

2. Recursive Least Squares (RLS) Algorithm: Infinite Precision

Consider a linear system with input signal x(n) and output signal y(n). Suppose that the samples x(n) and y(n) can be related by the system impulse response coefficients a_i^*:

$$y(n) = \sum_{i=0}^{N-1} a_i^* x(n-i) = \underline{a}^{*T} \underline{x}(n) \tag{1}$$

where the underscore denotes N-length vectors. Further, we have assumed that the impulse response has insignificant terms beyond N samples. The notation $\underline{x}(n)$ signifies the vector of the last N samples of the input:

$$\underline{x}(n) = [x(n)\ x(n-1)\ \ldots\ x(n-N+1)]^T \tag{2}$$

This paper considers the systems identification problem; that is, we are interested in estimating $\underline{a}^*$ from the sequences x(n) and y(n). The Recursive Least Squares (Kalman) algorithm achieves this by updating a current coefficient estimate $\underline{a}(n-1)$ through the following recursion.

$$\underline{a}(n) = \underline{a}(n-1) + \underline{K}(n)e(n) \tag{3}$$

where

$$e(n) = y(n) - \hat{y}(n) \tag{4}$$

is the prediction error and ,

$$\hat{y}(n) = a^T(n-1)\underline{x}(n) \tag{5}$$

is the prediction of y(n) based on N previous samples of x(n). $\underline{K}(n)$ is sometimes is denoted as the Kalman gain. The Kalman gain can be shown [11] to be

$$\underline{K}(n) = \left[\sum_{i=0}^{n} \underline{x}(i)\underline{x}^T(i) \right]^{-1} \underline{x}(n) \tag{6}$$

3. RLS Roundoff Error Analysis

We are interested in analyzing the degradation in the finite wordlength implementation of (6). As a first step, we are specifically interested in the errors introduced by rounding the products of the Kalman gain and the prediction error in (3). We will use the following model for the computation of these products:

$$\beta_i(n) = K_i(n)\; e'(n) + \mu_i(n), = 1, \ldots, N \;. \tag{7}$$

where $\mu_i(n)$ is the ith component of the rounding error. In (7) we will initially investigate the case of infinite precision in the calculation of the Kalman gain vector $\underline{K}(n)$. Define $\varepsilon(n)$ as the roundoff error introduced by quantizing the finite wordlength inner product computation,

$$\hat{y}'(n) = \underline{x}^T(n)\underline{a}'(n-1) + \varepsilon(n) \tag{8}$$

For rounding, $\mu_i(n)$ and $\varepsilon(n)$ can be modelled as random zero mean, uncorrelated noise processes. Sripad [5] has shown this to be an accurate model for register lengths greater or equal to 8 bits. This model is quite common and and very often used in the analysis of finite register effects [6,10]. Using this model, we have:

$$E[\mu_i(n)] = 0 \tag{9}$$

$$E[\varepsilon(n)] = 0 \tag{10}$$

$$E[\mu_i(m)\mu_j(n)] = \sigma_\mu^2\; \delta_{mn} \tag{11}$$

$$E[\varepsilon(m)\varepsilon(n)] = \sigma_\varepsilon^2\; \delta_{mn} \tag{12}$$

where δ_{mn} is the Kronecker delta: $\delta_{mn} = 1$ if m=n, D otherwise.

In general σ_μ and σ_ε are functions of the register lengths of $\underline{a}(n)$ and $\hat{y}(n)$. Thus, using (3) and (7) the finite wordlength recursive update equation becomes:

$$\underline{a}'(n) = \underline{a}'(n-1) + \underline{K}(n)e'(n) + \underline{\mu}(n) \tag{13}$$

where

$$\underline{\mu}(n) = [\mu_1(n)\ \mu_2(n)\ \ldots\ \mu_N(n)]^T \tag{14}$$

The prime denotes the finite wordlength estimate. Now

$$e'(n) = y(n) - \hat{y}'(n) \tag{15}$$

From (1), (5), and (8) we obtain:

$$e'(n) = \underline{x}^T(n)\ [\underline{a}^* - \underline{a}'(n-1)] - \varepsilon(n) \tag{16}$$

Substituting (16) into (13) we have,

$$\underline{a}'(n) = \underline{a}'(n-1) - \underline{K}(n)\underline{x}^T(n)\ [\underline{a}'(n-1) - \underline{a}^*] - \underline{K}(n)\varepsilon(n) + \underline{\mu}(n) \tag{17}$$

If we subtract $\underline{a}^*$ from both sides of (17) and substitute

$$\underline{\theta}(n) = \underline{a}'(n) - \underline{a}^* \tag{18}$$

we obtain,

$$\underline{\theta}(n) = \underline{\theta}(n-1) - \underline{K}(n)\underline{x}^T(n)\underline{\theta}(n-1) - \underline{K}(n)\varepsilon(n) + \underline{\mu}(n) \tag{19}$$

Rewrite (19) as

$$\underline{\theta}(n) = [I-\underline{K}(n)\underline{x}^T(n)]\underline{\theta}(n-1) - \underline{K}(n)\varepsilon(n) + \underline{\mu}(n) \tag{20}$$

Let

$$\underline{\xi}(n) = -\ \underline{K}(n)\varepsilon(n) + \underline{\mu}(n) \tag{21}$$

Then

$$\underline{\theta}(n) = [I-\underline{K}(n)\underline{x}^T(n)]\underline{\theta}(n-1) + \underline{\xi}(n)$$

The recursive equation (20) can be written solved for $\underline{\theta}(n)$ by expanding for n= 0,1,2,...and recognizing the pattern that emerges. This procedure produces:

$$\underline{\theta}(n) = \left\{\prod_{i=1}^{n}\ [I-\underline{K}(i)\underline{x}^T(i)]\right\}\underline{\theta}(0) + \sum_{i=1}^{n} \underline{\xi}^T(i)\left\{\prod_{j=i+1}^{n}[I-\underline{K}(j)\underline{x}^T(j)]\right\} \tag{22}$$

Note that we define

$$\prod_{j=n+1}^{n} [\cdot] = 1$$

The term

$$\underline{\theta}_0(n) = \prod_{i=1}^{n} [I-\underline{K}(i)\underline{x}^T(i)]\underline{\theta}(0) \tag{23}$$

is exactly the parameter error vector which would be obtained by an infinite precision RLS algorithm. That is,

$$\theta_0(n) = \underline{a}(n) - \underline{a}^* \quad . \tag{24}$$

The term

$$\underline{\psi}(n) = \sum_{i=1}^{n} \prod_{j=i+1}^{n} [I-\underline{K}(j)\underline{x}^T(j)]\underline{\xi}(i) \tag{25}$$

represents the additional parameter error due to roundoff. Thus,

$$\underline{\theta}(n) = \underline{\theta}_0(n) + \underline{\psi}(n) \tag{26}$$

We also have using (18),

$$\underline{a}'(n) = \underline{a}(n) + \underline{\psi}(n) \tag{27}$$

Define the excess prediction error due to roundoff:

$$\zeta(n) = e(n) - e'(n) \tag{28}$$

Substituting (16) into (28) and using (1), (4), and (5) produces

$$\zeta(n) = \underline{\psi}^T(n-1)\underline{x}(n) - \varepsilon(n) \tag{29}$$

A common assumption [6] is that the quantization noises $\underline{\psi}(n)$ and $\varepsilon(n)$ are uncorrelated with each other, as well as with the data $\underline{x}(n)$. Using this property produces:

$$\sigma_\zeta^2 = E\{\zeta^2(n)\} = E\{\underline{\psi}^T(n-1)\underline{x}(n)\underline{x}^T(n)\underline{\psi}(n-1)\} + \sigma_\varepsilon^2 \tag{30}$$

$$= \text{Trace } [R_x E\{\psi(n-1)\psi^T(n-1)] + \sigma_\varepsilon^2 \tag{31}$$

where

$$R_x = E\{\underline{x}(n)\underline{x}^T(n)\} \tag{32}$$

is the input signal autocorrelation matrix. Hence, to derive the excess mean square error caused by roundoff quantization we need to derive

$$R_\psi = E\{\underline{\psi}(n-1)\underline{\psi}^T(n-1)\} \tag{33}$$

From (25):

$$\underline{\psi}^T(n-1) = \sum_{i=1}^{n-1} \underline{\xi}^T(i) \prod_{j=i+1}^{n-1} [I-\underline{x}(j)\underline{K}^T(j)] \tag{34}$$

therefore, R_ψ becomes:

$$E\{\underline{\psi}(n-1)\underline{\psi}^T(n-1)\} = E\{ \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \prod_{k=i+1}^{n-1} [I-\underline{K}(k)\underline{x}^T(k)]\underline{\xi}(i)\underline{\xi}^T(j) \prod_{m=j+1}^{n-1} [I-\underline{x}(m)\underline{K}^T(m)]\} \tag{35}$$

Since k=i+1 and m=j+1, $\underline{\xi}(i)$ occurs in time before $\underline{K}(k)$, x(k). Therefore, the $\underline{\xi}(\cdot)$ in (35) are uncorrelated with $\underline{K}(\cdot)$ and $\underline{x}(\cdot)$, which allows a substantial simplification in (35). This can be seen by examining the structure of $E\{\underline{\xi}(i)\underline{\xi}^T(j)\}$. From (21) and using the property that $\mu(i)$ and $\varepsilon(j)$ are uncorrelated and zero mean, we have

$$E\{\underline{\xi}(i)\underline{\xi}^T(j)\} = E\{\underline{\mu}(i)\underline{\mu}^T(j) + \varepsilon(i)\varepsilon(j)\underline{K}(i)\underline{K}^T(j)\} \tag{36}$$

From (9) through (12) we obtain,

$$E\{\underline{\xi}(i)\underline{\xi}^T(j)\} = \begin{bmatrix} \sigma_\mu^2 \; I + \sigma_\varepsilon^2 \; E\{\underline{K}(i)\underline{K}^T(i)\}, & i=j \\ 0, & i\neq j \end{bmatrix} \tag{37}$$

Now, define as in [11,12] the sample autocorrelation matrix

$$R(i) = \frac{1}{i} \sum_{j=0}^{i} \underline{x}(j)\underline{x}^T(j) \tag{38}$$

or,

$$\left[\sum_{j=0}^{i} \underline{x}(j)\underline{x}^T(j)\right]^{-1} = \frac{1}{i} R^{-1}(i) \tag{39}$$

Hence from (6) we can write the Kalman gain in the form

$$\underline{K}(i) = \frac{1}{i} R^{-1}(i)\underline{x}(i)$$

We have therefore,

$$E\{\underline{K}(i)\underline{K}^T(i)\} = E\{\frac{1}{i^2} R^{-1}(i)\ \underline{x}(i)\underline{x}^T(i)R^{-1}(i)\} \tag{40}$$

As $i\to\infty$, $R(i) \to R_x$ and (40) becomes,

$$E\{\underline{K}(i)\underline{K}^T(i)\} = \frac{1}{i^2} R_x^{-1} \tag{41}$$

where we are assuming x(n) is ergodic. Therefore,

$$R_\xi = E\{\underline{\xi}(i)\underline{\xi}^T(i)\} \cong \sigma_\mu^2\ I + \sigma_\varepsilon^2\ \frac{1}{i^2}\ R_x^{-1} \tag{42}$$

At this point, one must consider the structure of the R_x matrix. Since we are examining a systems identification application, we may assume that the excitation signal is a white, gaussian, uncorrelated sequence. Therefore,

$$R_x = \sigma_x^2 I \tag{43}$$

and

$$R_\xi = \sigma_\xi^2 I \tag{44}$$

where

$$\sigma_\xi^2 = \sigma_\mu^2 + \frac{\sigma_\varepsilon^2}{i^2\ \sigma_x^2} \tag{45}$$

Now, since $E\{\underline{\xi}(i)\underline{\xi}^T(j)\} = 0$, $i\neq j$, the only non-zero terms of (35) occur when $i=j$. Using this fact and (44) in (35) we obtain

$$R_\psi = E\{\underline{\psi}(n)\underline{\psi}^T(n)\} = \sigma_\xi^2 \sum_{i=1}^{n} \prod_{k=i+1}^{n} E\{[I-\underline{K}(k)\underline{x}^T(k)]\ [I-\underline{x}(k)\underline{K}^T(k)]\} \tag{46}$$

An initial simplification will help:

$$E\{\underline{K}(k)\underline{x}^T(k)\} = E\{\frac{1}{k} R^{-1}(k)\underline{x}(k)\underline{x}^T(k)\} = \frac{1}{k} I \quad , \tag{47}$$

This is for the "steady state" solution.

Calculating the expectation in (46) we have

$$E\{\cdot\} = E\{I - \underline{x}(k)\underline{K}^T(k) - \underline{K}(k)\underline{x}^T(k) + \underline{K}(k)\underline{x}^T(k)\underline{x}(k)\underline{K}^T(k)\}$$

$$= I - \frac{2}{k}I + E\{\frac{1}{k^2} R^{-1}(k)\underline{x}(k)\underline{x}^T(k)\underline{x}(k)\underline{x}^T(k)R^{-1}(k)\}$$

$$= I - \frac{2}{k} I + \frac{1}{k^2} \quad R^{-1}(k)E\{\underline{x}(k)\underline{x}^T(k)\underline{x}(k)\underline{x}^T(k)\}R^{-1}(k)$$

To evaluate the fourth moments

$$E\{\underline{x}(k)\underline{x}^T(k)\underline{x}(k)\underline{x}^T(k)\}$$

we note that, in general, if z_1, z_2, z_3, z_4 are real zero mean Gaussian random variables, then [13]

$$E\{z_1 z_2 z_3 z_4\} = E[z_1 z_2]E[z_3 z_4] + E[z_1 z_3]\ E[z_2 z_4] + E[z_1 z_4]E[z_2 z_3]$$

In [14] this result is extended from scalars to vectors and the fourth moments are found to be

$$E\{\underline{x}(k)\underline{x}^T(k)A\underline{x}(k)\underline{x}^T(k)\} = 2R_x A R_x + R_x Tr\{R_x A\}$$

where A is a positive definite matrix and $R_x = E\{x(k)x^T(k)\}$.

Substituting $A=I$ and $R_x=\sigma^2{}_x I$ we obtain

$$E\{\underline{x}(k)\underline{x}^T(k)\underline{x}(k)\underline{x}^T(k)\} = 2\sigma^4{}_x I + \sigma^4{}_x NI$$

Thus,

$$E\{\cdot\} = (1 - \frac{2}{k})I + \frac{1}{k^2 \sigma_x^4} \quad \sigma_x^4 \ (N+2)\ I$$

$$E\{\cdot\} = (1 - \frac{2}{k} + \frac{N+2}{k^2})I \tag{48}$$

Substituting (48) into (46),

$$R_\psi = \sigma_\xi^2 \sum_{i=1}^{n} \prod_{k=i+1}^{n} (1 - \frac{2}{k} + \frac{N+2}{k^2})I \tag{49}$$

We shall examine the term

$$\sum_{i=1}^{n} \prod_{k=i+1}^{n} \left[1 - \frac{2}{k} + \frac{N+2}{k^2}\right] = \sum_{i=1}^{n} \prod_{k=i+1}^{n} \left[(\frac{k-1}{k})^2 + \frac{N+1}{k^2} \right] \tag{50}$$

When i is small and n large, the initial terms in the bracket will be large and the term $(N+1)/k^2$ will dominate. However, for large n, these initial terms are multiplied by terms with magnitudes less than one. Hence, their effect is reduced. Based on this argument, terms where k is large (by large we mean k>>N) contribute the most to (5) because less multiplication terms of magnitude less than one are involved and the terms rapidly approach unity. Hence, we may neglect the second term in (50). By assuming steady state we have that n is large. Therefore,

$$\prod_{k=i+1}^{n} (\frac{k-1}{k})^2 = \left\{ \frac{(i\cdot(i+1)(i+2)\ldots(n-1)}{(i+1)(i+2)\ldots n} \right\}^2 = \frac{i^2}{n^2} \tag{51}$$

Substituting into (49) we obtain

$$R_\psi = \frac{1}{n^2} \sum_{i=1}^{n} i^2 \sigma_\xi^2 \tag{52}$$

Substituting for $\sigma_\xi{}^2$ from (45),

$$R_\psi = \frac{1}{n^2} \sum_{i=1}^{n} [i^2 \sigma_\mu^2 + \frac{\sigma_\varepsilon^2}{\sigma_x^2}]I \tag{52}$$

$$R_\psi = \sigma_\mu^2 \frac{1}{n^2} \frac{n(n+1)(2n+1)}{6} I + \frac{\sigma_\varepsilon^2}{\sigma_x^2} \frac{1}{n} I \tag{53}$$

$$R_{\psi} = \left(\frac{n}{3}\sigma_{\mu}^{2} + \frac{\sigma_{\varepsilon}^{2}}{\sigma_{x}^{2}}\frac{1}{n}\right)I \tag{54}$$

We can now obtain the mean square error due to roundoff noise by substituting (54) into (32)

$$\sigma_{\zeta}^{2} = \text{Trace}\left[\sigma_{x}^{2}\frac{n}{3}\sigma_{\mu}^{2} I + \frac{1}{n}\sigma_{\varepsilon}^{2} I\right] + \sigma_{\varepsilon}^{2} \tag{55}$$

Finally

$$\boxed{\sigma_{\zeta}^{2} = \frac{n}{3} N\sigma_{\mu}^{2}\sigma_{x}^{2} + \frac{N}{n}\sigma_{\varepsilon}^{2} + \sigma_{\varepsilon}^{2}} \tag{56}$$

4. Discussion

We have derived an expression for the mean square error introduced by roundoff quantization in the RLS algorithm (52). This expression shows that however small the error introduced by rounding the Kalman gain multiplied by the error term, the algorithm will diverge as $n \to \infty$. However, since during the initial convergence of RLS algorithms n is small and $\|\underline{K}(n)\|$ is large, the errors are small. Thus, expression (52) leads to the conclusion that in the finite wordlength implementation of the RLS algorithm, after on the order of N iterations, the estimates must be frozen or the adaptation modified. Otherwise, divergence will occur. Thus, the algorithm will initially converge rapidly to a small error in FWL implementation, but will begin to diverge. This is an analytical agreement with the simulation results observed in [8,9].

5. References

[1] D.D. Falconer and L. Ljung, "Application of Fast Kalman Estimation to Adaptive Equalization," IEEE Trans. on Comm., Vol. COM-26, No. 10, October 1978, pp. 1439-1446.

[2] B. Friedlander, "System Identification Techniques for Adaptive Noise Cancelling," IEEE Trans. on Acous., Speech, and Sig. Processing, Vol. ASSP-30, October 1982, pp. 699-709.

[3] E.H. Satorius and J.D. Pack, "Application of Least Squares Lattice Algorithms to Adaptive Equalization," IEEE Trans. on Comm., Vol. COM-29, No. 2, February 1981, pp. 136-142.

[4] P. Moroney, Issues in the Implementation of Digital Feedback Compensators, MIT Press, Cambridge, MA,1983.

[5] A. Sripad, "Models for Finite Precision Arithmetic, with Application to the Digital Implementation of Kalman Filters," Ph.D. dissertation, Washington University, St. Louis, MO,1978.

[6] C. Caraiscos, and B. Liu, "A Roundoff Error Analysis of the LMS Adaptive Algorithm," IEEE Trans. on Acous., Speech, and Sig. Processing, Vol. ASSP-32, No. 1, February 1984, pp. 34-41.

[7] M. Bellanger and C. Cengiz, "On Computational Complexity in Adaptive Digital Filters," Proc., 1983 IEEE Internation Conference on Acous., Speech, and Sig. Proc., Boston, MA, Apr. 1983.

[8] F. Ling and J.G. Proakis, "Numerical Accuracy and Stability: Two Problems of Adaptive Estimation Algorithms Caused by Round-off Error," Proc., 1983 IEEE Internation Conference on Acous., Speech, and Sig. Proc., San Diego, CA, March 1984.

[9] J.M. Cioffi and T. Kailath, "Fast, Recursive Least Squares Transversal Filters for Adaptive Filtering," IEEE Trans. on Acous., Speech, and Sig. Proc., Vol. ASSP-32, April 1984, pp. 304-337.

[10] R.D. Gitlin, et. al., "On the Design of Gradient Algorithms for Digitally Implemented Adaptive Filters," IEEE Trans. Circuit Th., Vol. CT-20, pp. 125-136, March 1973.

[11] L. Ljung, M. Morf, and D.D. Falconer, "Fast Calculation of Gain Matrices for Recursive Estimation Schemes," Int. Jour. of Control, Vol. 27, No. 1, 1978, pp. 1-19.

[12] D. Godard, "Channel Equalization Using a Kalman filter for Fast Data Transmission," IBM Jour. of Res. and Dev., May 1974, pp. 267-273.

[13] W.B. Davenport, Probability and Random Processes, McGraw-Hill, New York, 1970.

[14] L.L. Horowitz and K.P. Senne, "Performance Advantage of Complex LMS for Controlling Narrow-Band Adaptive Arrays," IEEE Trans. on Acous., Speech, and Sig. Proc., Vol. ASSP-29, No. 3, June 1981.