A NONLINEAR ANALYTICAL MODEL FOR QUANTIZATION EFFECTS IN THE LMS ALGORITHM WITH POWER-OF-TWO STEP SIZE

José Carlos M. Bermudez[†‡] and Neil J. Bershad[‡]

[†]Electronic Instrumentation Lab., Dept of Electrical Engineering, Fed. Univ. of Santa Catarina, Florianopolis, SC 88040-900, Brazil.
[‡]Department of Electrical and Computer Engineering, University of California, Irvine, Irvine, CA 92717, U.S.A..

ABSTRACT

This paper[1] presents a study of the quantization effects in the finite precision LMS algorithm with power-of-two step sizes. Nonlinear recursions are derived for the mean and second moment matrix of the weight vector about the Wiener weight for white gaussian data models and small algorithm step size μ. The solutions of these recursions are shown to agree very closely with the Monte Carlo simulations during all phases of the adaptation process. A design curve is presented to demonstrate the use of the theory to select the number of quantizer bits and the adaptation step size μ to yield desired transient and steady-state behaviors.

1. INTRODUCTION

The least mean squares (LMS) algorithm is one of the most popular algorithms for digital implementation of real-time high-speed adaptive filters. Fixed-point arithmetic is prevalent in such applications [1-3]. Many previous publications have studied the behavior of the finite precision LMS algorithm.

Gitlin et al. [1] were the first to address the so called "stopping phenomenon". They compared the digital and analog LMS implementations for the least attainable residual mean-square errors (MSE). Caraiscos and Liu [2] also presented a steady-state analysis of the quantized LMS algorithm. Their analysis used a linear model for the correlation multiplier. This model approximates the quantization errors by uncorrelated additive white noise sources. Alexander [3] presented a finite precision analysis of the LMS algorithm which included the transient adaptation period. Here, as in [2], a linear model has been used for the quantization operation. The analysis in [3] was based upon an analytical model for the difference between the finite-precision and infinite-precision weight vectors. However, this error vector does not provide direct information regarding the mean-square output error.

Although the linear model is adequate during the early stages of adaptation, its validity lessens as the error decreases and the algorithm converges. Quantization is a nonlinear operation. Thus, its effects on the algorithm behavior can be better predicted using a nonlinear model.

A recent work [8] studied the nonlinear behavior of the quantized LMS algorithm using arbitrary step sizes. For arbitrary μ, the implementation of the LMS updating equation requires two quantizations [9], [12]. A different implementation is possible when the step size is an exact power of two. In this case, the multiplications by the step size μ are usually realized as right shifts. The error and input signals are first multiplied in double precision. The result is then shifted (multiplication by μ) and quantized to single precision. The convergence is controlled by the quantized value of the entire weight update term. This was the problem studied in [1-3] using a linear model and in [4] using a continuous nonlinear function.

This paper analyzes the nonlinear behavior of the quantized LMS algorithm implementation in which products by a power-of-two step size μ are implemented as right shifts.

1.1. Mathematical Model of Quantized LMS

The updating equation for the LMS algorithm is given by [10], [11]

$$W_L(n+1) = W_L(n) + \mu\, \varepsilon_L(n) X(n) \qquad (1)$$

where

$\varepsilon_L(n) = d(n) + z(n) - X^T(n)\, W_L(n)$,

$X(n) = [x_1(n), x_2(n), \cdots, x_N(n)]^T$, $x_i(n) = x(n-i+1)$

:observed data vector with length N

$W_L(n)$:weight vector at time n (length N)

$d(n)$:desired scalar signal

$z(n)$:additive noise

1 This work was supported in part by the Brazilian Research Council (CNPq) under grant No. 201532/93-0.

This paper assumes $\mu = 2^{-d}$, d a positive integer. Also, it is assumed that any multiplication by μ is realized as a right shift. Under these conditions, the most common implementation of (1) performs first the product of the error signal by $x_i(n)$ using double precision [2]. Next, the multiplication by μ is realized shifting the result d positions to the right. Finally, the result is quantized to single precision and added to the current weight vector.

The implementation of (1) described above leads to the finite precision LMS updating equation

$$W_Q(n+1) = W_Q(n) + Q[\mu\, \varepsilon(n) X(n)] \qquad (2)$$

where $Q[\cdot]$ denotes a quantization operation and $\varepsilon(n) = d(n) + z(n) - X^T(n)\, W_Q(n)$, is the error signal of the quantized algorithm. The quantization errors due to the double precision calculations of $\varepsilon(n) x_i(n)$ have been neglected in (2). The input signals are assumed to be properly scaled to avoid overflow errors due to additions.

To render the analysis more tractable, the following typical LMS analysis assumptions are made [5], [7]:

a) The data vector $X(n)$ is statistically independent over time. As a consequence, the present weight and data vectors are statistically independent. Also, $x(n)$ is assumed a stationary zero-mean independent Gaussian sequence. Thus the data covariance matrix $R_{XX} = E[X(n)X^T(n)] = \sigma_x^2 I$ (I = identity matrix);

b) The desired data $d(n)$ is a stationary zero-mean Gaussian sequence, correlated with $X(n)$;

c) The noise sequence $z(n)$ is zero-mean, Gaussian and statistically independent of any other signal.

Using these assumptions, the analysis leads to results that are representative of several practical applications [3], [5].

2. LMS ALGORITM WITH QUANTIZED UPDATE

A two's complement rounding quantizer with step-size Δ is assumed in the analysis (Fig. 1).

2.1. Mean Behavior

It is mathematically more convenient to investigate the statistics of (2) about the optimum Wiener weight vector $W_0 = R_{XX}^{-1} R_{dX}$, where $R_{dX} = E[d(n)X(n)]$ and $R_{XX} = E[X(n)X^T(n)]$. Here, $E[\cdot]$ denotes statistical expectation. Letting $V(n) = W_Q(n) - W_0$ and inserting into (2) yields

$$V(n+1) = V(n) + Q[\mu\, \varepsilon(n) X(n)] \qquad (3)$$

Averaging both sides of (3) yields

$$E[V(n+1)] = E[V(n)] + E\left[Q\left[\mu \left\{d(n) + z(n) - V^T(n)X(n) - W_0^T X(n)\right\} X(n)\right]\right] \qquad (4)$$

For simplicity, the expectations in (4) are taken in two steps, first on the data and then on $V(n)$. Conditioning (4) on $V(n)$ yields

$$E[V(n+1)|V(n)] = V(n) + E[Q[\mu\, \varepsilon(n) X(n)]|V(n)] \qquad (5)$$

The quantization function can be expressed as

$$Q[y] = y - g(y) = y - \sum_{k=-\infty}^{\infty} b_k\, e^{jk\frac{2\pi}{\Delta} y} \qquad (6)$$

where $g(y)$ is periodic with Fourier series coefficients

$$b_k = \begin{cases} j(-1)^k \left(\dfrac{\Delta}{2k\pi}\right) & , k \neq 0 \\ 0 & , k = 0 \end{cases} \qquad (7)$$

Using (6) and (7) into (5), with $y = \mu\, \varepsilon(n) x_i(n)$ and $V(n) = [v_1(n) \cdots v_N(n)]^T$, yields, for the ith row

$$E[v_i(n+1)|V(n)] = v_i(n) + E[\mu\, \varepsilon(n) x_i(n)|V(n)] - \sum_{k=-\infty}^{\infty} b_k\, E\left[e^{jk\frac{2\pi}{\Delta}\mu\, \varepsilon(n) x_i(n)} | V(n)\right] \qquad (8)$$

Each expectation within the summation is the characteristic function of the product $\varepsilon(n) x_i(n)$, conditioned on $V(n)$. Conditioned on $V(n)$, $\varepsilon(n)$ and $x_i(n)$ are zero-mean, Gaussian and correlated. Thus, the characteristic function of their product is given by [10]

$$E\left[e^{j\left(k\frac{2\pi}{\Delta}\mu\right)\varepsilon(n) x_i(n)} | V(n)\right] = \left\{1 - 2j\left(k\frac{2\pi}{\Delta}\mu\right)\rho\, \sigma_{\varepsilon|V}\, \sigma_x + \left(k\frac{2\pi}{\Delta}\mu\right)^2 \sigma_{\varepsilon|V}^2\, \sigma_x^2 \left(1 - \rho^2\right)\right\}^{-1/2} \qquad (9)$$

where ρ is the correlation coefficient of $\varepsilon(n)$ and $x_i(n)$, conditioned on $V(n)$. The second moment $\sigma_{\varepsilon|V}^2$ is

$$\sigma_{\varepsilon|V}^2 = E[\varepsilon^2(n)|V(n)] = \xi_0 + \sigma_x^2\, V^T(n) V(n) \qquad (10)$$

where ξ_0 is the MSE using the Wiener filter. For small μ, the weight fluctuations are small and $\sigma_{\varepsilon|V}^2(V(n))$ is

concentrated near its mean. Thus, an accurate approximation for the expectation over $V(n)$ in (9) can be obtained replacing $\sigma^2_{\varepsilon|V}$ by its mean. Taking the expected value of (10) yields

$$E\left[\sigma^2_{\varepsilon|V}\right] = E\left[\varepsilon^2(n)\right] = \xi_0 + \sigma_x^2 \,\mathrm{tr}\left[K_{VV}(n)\right] \tag{11}$$

Using (6)- (10), it can be shown [9] that

$$E[V(n+1)] = \left\{1 - \mu\sigma_x^2\left[1 + 2\sum_{k=1}^{\infty}\frac{(-1)^k}{[C(k)]^{3/2}}\right]\right\} E[V(n)] \tag{12}$$

$$\text{with } C(k) = 1 + \left(\frac{2k\pi}{\Delta}\mu\sigma_x^2\right)^2\left(\frac{\xi_0}{\sigma_x^2} + \mathrm{tr}\left[K_{VV}(n)\right]\right)$$

and where $K_{VV}(n) = E\left[V(n)V^T(n)\right]$ is the correlation matrix of the weight error vector $V(n)$ and $\mathrm{tr}\left[K_{VV}(n)\right]$ is the trace of $K_{VV}(n)$.This recursion describes the mean behavior of the weight error vector.

2.2. Second Moment Behavior

Postmultiplying (3) by its transpose, taking the expected value and determining the trace yields

$$\mathrm{tr}\left[K_{VV}(n+1)\right] = \mathrm{tr}\left[K_{VV}(n)\right] + 2E\left[V^T(n)Q\left[\mu\,\varepsilon(n)X(n)\right]\right] + E\left[Q\left[\mu\,\varepsilon(n)X^T(n)\right]\,Q\left[\mu\,\varepsilon(n)X(n)\right]\right] \tag{13}$$

Evaluating the expectations in (13), it can be shown that [9]

$$\mathrm{tr}\left[K_{VV}(n+1)\right] = \left\{1 - 2\mu\sigma_x^2\left\{1 + 2\sum_{k=1}^{\infty}\frac{(-1)^k}{[C(k)]^{3/2}}\right\} + N\left(\mu\sigma_x^2\right)^2\left\{1 + 4\sum_{k=1}^{\infty}\frac{(-1)^k}{[C(k)]^{3/2}}\right\}\right\}\mathrm{tr}\left[K_{VV}(n)\right]$$

$$+ N\left(\mu\sigma_x^2\right)^2\left(\frac{\xi_0}{\sigma_x^2}\right)\left\{1 + 4\sum_{k=1}^{\infty}\frac{(-1)^k}{[C(k)]^{3/2}}\right\}$$

$$- 2N\left(\frac{\Delta}{2\pi}\right)^2\sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty}\frac{(-1)^{k+\ell}}{k\ell}\left\{\frac{1}{[C(k+\ell)]^{1/2}} - \frac{1}{[C(k-\ell)]^{1/2}}\right\} \tag{14}$$

Equation (14) describes the time evolution of the trace of the weight-error correlation matrix. Using (11) and (14), the MSE performance can be recursively determined.

3. SIMULATION EXAMPLES

Fig. 2 depicts a simple system identification problem. W^* is the weight vector to be identified. The components of W^* are the values of 31 equally spaced samples of a time-delayed raised-cosine function. Fig. 3 displays Monte Carlo simulations (100 runs) of $\mathrm{tr}\left[K_{VV}(n)\right]$ for $N = 31$, $\sigma_x^2 = 1/9$, $\mu = 2^{-5}$, $\xi_0 = E\left[z^2(n)\right] = 10^{-12}$ and several values of $\Delta = 2^{-b}$. The theoretical curves were determined using (14). The theoretical predictions and the simulation results are in excellent agreement. Fig. 4 presents the time evolution of the MSE for the same parameters. Clearly, the analytical results can be used to predict the behavior of the quantized LMS algorithm (2). Fig. 5 displays the MSE after 5000 iterations for $\sigma_x^2 = 1$, $\xi_0 = \sigma_z^2 = 10^{-8}$, $N = 31$ and several values of $\Delta = 2^{-b}$ and $\mu = 2^{-d}$. The dots (•) were obtained from the theoretical recursions. The lines (—) were drawn by cubic spline interpolation for easier visualization.

4. CONCLUSIONS

This paper presented a study of the quantization effects in the finite precision LMS algorithm with power-of-two step sizes. Deterministic nonlinear recursions were derived for the mean and second moment matrix of the weight vector about the Wiener weight for white gaussian data models and small algorithm step size μ. The numerical solutions of these recursions were shown to agree very closely with the Monte Carlo simulations. A design curve has been presented to demonstrate how the theory can be used to select the number of bits and the step size μ to yield a desired algorithm performance.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Rui Seara of Federal University of Santa Catarina for the helpful discussions regarding the digital implementation of the LMS algorithm.

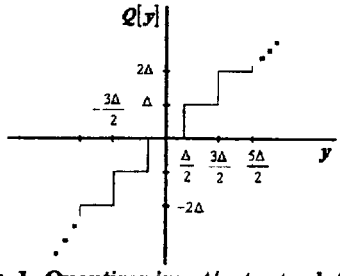

Fig. 1- Quantizer input/output relation

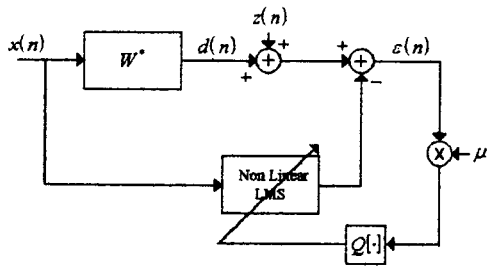

Fig. 2- System identification model

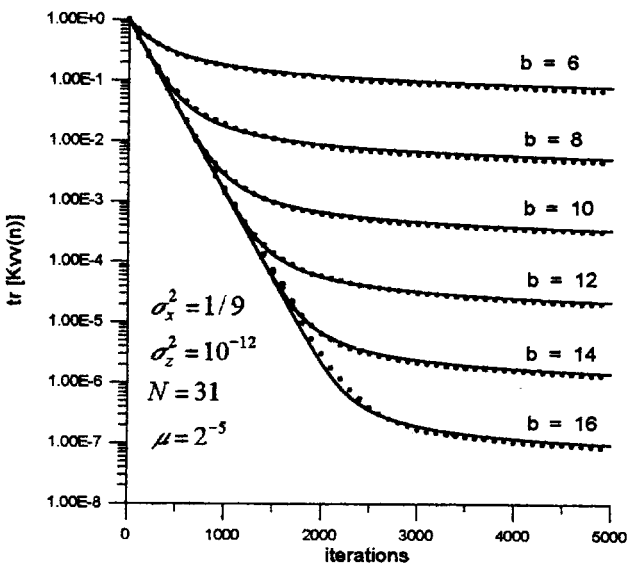

Fig. 3. Simulations (•) versus theory (–) for the time evolution of $\mathrm{tr}[K_{VV}(n)]$. Simulations using updating equation (2) and $\Delta = 2^{-b}$ in Fig. 1.

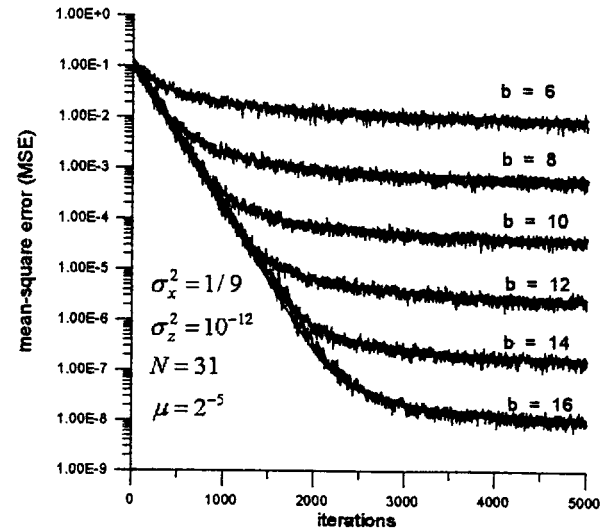

Fig. 4. Simulations (—) versus theory (–) for the MSE. Simulations using updating equation (2) and $\Delta = 2^{-b}$ in Fig. 1.

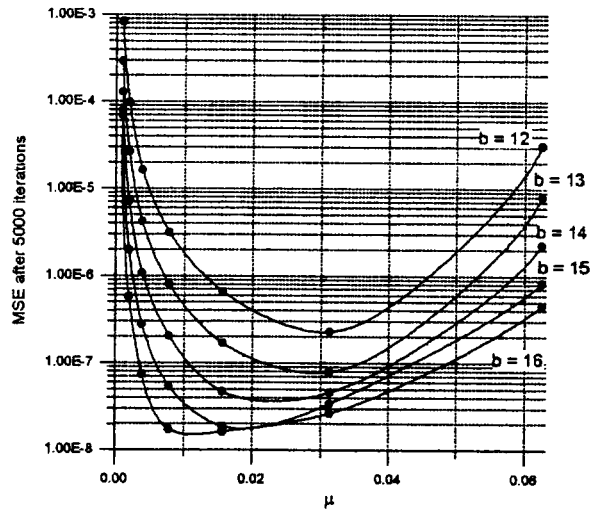

Fig. 5. MSE after 5000 iterations for $\sigma_x^2 = 1$, $\xi_0 = \sigma_z^2 = 10^{-8}$, $N = 31$ and several values of $\Delta = 2^{-b}$ and $\mu = 2^{-d}$. Lines (—) are cubic spline interpolations for easier visualization.

REFERENCES

[1] R. D. Gitlin, J. E. Mazo and M. G. Taylor, "On the design of gradient algorithms for digitally implemented adaptive filters," *IEEE Trans.CT*, pp. 125-136, Mar 1973.

[2] C. Caraiscos and B. Liu, "A roundoff error analysis of the LMS adaptive algorithm," *IEEE Trans. ASSP*, pp. 34-41, Feb 1984.

[3] S. T. Alexander, "Transient weight misadjustment properties for the finite precision LMS algorithm," *IEEE Trans. ASSP*, pp. 1250-1258, Sept 1987.

[4] N. J. Bershad, "On weight update saturation nonlinearities in LMS adaptation," *IEEE Trans. ASSP*, pp. 623-630, Apr 1990.

[5] D. L. Duttweiller, "Adaptive filter performance with nonlinearities in the correlation multiplier," *IEEE Trans. ASSP*, pp. 578-586, Aug 1982.

[6] N. J. Bershad, "On the optimum data nonlinearity in LMS adaptation," *IEEE Trans. ASSP*, pp. 69-76, Feb 1986.

[7] N. J. Bershad, "On error-saturation nonlinearities in LMS adaptation," *IEEE Trans. ASSP*, pp. 440-452, Apr 1988.

[8] J.C.M. Bermudez and N.J. Bershad, "Nonlinear quantization effects in the LMS algorithm - analytical models for the MSE transient and convergence behavior," Proc. 28th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Oct 1994.

[9] N.J. Bershad and J.C.M. Bermudez, "A nonlinear analytical model for the quantized LMS algorithm - the power-of-two step size case," submitted to the *IEEE Trans. on Signal Processing*, Aug 1994..

[10] J. Omura and T. Kailath, *Some Useful Probability Distributions*, Technical Report No. 7050-6, Systems Theory Lab., Stanford University, Stanford, CA, pp.88, September 1965.