A Finite Wordlength Analysis of an LMS-Newton Adaptive Filtering Algorithm

Marcello L. R. de Campos†, Paulo S. R. Diniz‡, and Andreas Antoniou†

†Dept. of Electrical and Computer Engineering, University of Victoria, P.O. Box 3055, Victoria, B.C., Canada, V8W 3P6
‡Prog. de Engenharia Elétrica e Depto. de Eletrônica COPPE/EE/Federal University of Rio de Janeiro, C.P. 68504, Rio de Janeiro, R.J., Brazil 21945

Abstract—The effects of quantization in an LMS-Newton adaptive filtering algorithm are investigated. The algorithm considered uses an optimum convergence factor, that forces the output a posteriori error to become zero in each iteration. The propagation of errors due to quantization in the internal variables of the algorithm is investigated and a closed-form formula for the excess mean square error due to quantization is derived. Fixed-point arithmetic is assumed throughout. Several simulations confirm the accuracy of the formulas presented.

I. INTRODUCTION

An important criterion for the performance of adaptive filters is the level of mean square error (MSE). In the presence of uncorrelated measurement noise, the performance of adaptive filters is degraded either by noisy estimates of the coefficients or by poor ability to track nonstationary systems or environments. These effects tend to increase the MSE. When finite wordlength hardware is used for the implementation, the quantization of signals and coefficients can increase the MSE quite significantly.

In this paper, the effects of quantization on the LMS-Newton (LMSN) algorithm proposed in [1] will be examined. In section 2 the algorithm is presented and its analysis is provided in section 3. Equations describing the quantization errors in the internal variables and the excess MSE due to the quantization effects are derived. In order to verify the accuracy of the formulas presented, some simulations are provided in section 4. Section 5 discusses the results achieved.

The authors are grateful to CAPES/Ministry of Education (Brazil) and Micronet, Networks of Centres of Excellence program (Canada) for supporting this work.

0-7803-1254-6/93 © 1993 IEEE

II. VARIABLE CONVERGENCE FACTOR LMS-NEWTON ALGORITHM

The LMS algorithm in [1] uses a variable convergence factor which is optimized in each iteration to yield zero a posteriori output error. The update equations involved are [2]

\[ e(n) = d(n) - w^H(n-1)u(n) \]  
\[ t(n) = R^{-1}(n-1)u(n) \]  
\[ r(n) = u^H(n)t(n) \]  
\[ w(n) = w(n-1) + 2t(n)e^*(n) / r(n) \]

where \( u(n) \) is the \( M \times 1 \) input vector \( (M \) is the number of coefficients), \( w(n) \) is the \( M \times 1 \) coefficient vector, \( d(n) \) is the desired signal, \( q \) is a reduction factor, and \( R^{-1}(n) \) is an estimate of the inverse of the input autocorrelation matrix \( R \), with \( R = E[u(n)u^H(n)] \). Any good approximation of \( R \) that is positive definite can be used. In [1], Diniz and Biscainho use a variable convergence factor to update the coefficient vector as well as to evaluate \( R^{-1}(n) \). In [2] the memory implications of this procedure are clarified and the possibility of using a constant convergence factor for \( R^{-1}(n) \) which does not affect the optimality characteristics of the algorithm is outlined. The algorithm that uses a variable convergence factor for \( R^{-1}(n) \) will be referred to as algorithm I and the algorithm that uses a constant convergence factor for \( R^{-1}(n) \) will be referred to as algorithm II hereafter. Unfortunately, for both methods of evaluation of the quantized version of \( R^{-1}(n) \), namely, \( R^{-1}(n)Q \), either with variable or constant convergence factor, positive definiteness is not guaranteed, as can be shown by using an analysis similar to that in [3]. This may result in divergence of the coefficients and a stabilization scheme must be incorporated.
III. FINITE PRECISION EFFECTS

In this section, the effect of roundoff errors is examined for the case of fixed-point arithmetic using an approach similar to that in [3]. The following assumptions have been made:

- two's-complement number representation
- additions and subtractions do not introduce quantization errors
- each internal operation is performed with enough bits in the integer part such that no overflow occurs
- quantized versions of all external signals and constants are available
- multiplication and division operations introduce white noise given by

\[ \varepsilon(ab) \triangleq (ab)_Q - ab \]  \hspace{1cm} (5a)

and

\[ \varepsilon(a/b) \triangleq (a/b)_Q - a/b \]  \hspace{1cm} (5b)

respectively.

The quantization error in each variable is defined as the difference between its quantized and nominal values, i.e.,

\[ \eta_n \triangleq a_Q - a \]  \hspace{1cm} (6)

Algorithms I and II in finite precision are

\[ e(n)_Q = d(n) - [w(n - 1)_Q u(n)]_Q \]  \hspace{1cm} (7)

\[ t(n)_Q = \left[ R^{-1}(n - 1)_Q u(n) \right]_Q \]  \hspace{1cm} (8)

\[ r'(n)_Q = \frac{[r(n)_Q t'(n)_Q]}{N}_Q \]  \hspace{1cm} (9)

\[ g'/(n)_Q = \frac{[r'(n)_Q]}{r'(n)_Q}_Q \]  \hspace{1cm} (10)

\[ r(n)_Q = [t(n)_Q g'(n)]_Q \]  \hspace{1cm} (11)

\[ s(n)_Q = \left[ g(n)_Q \right]_{N/2}_Q \]  \hspace{1cm} (12)

\[ w(n)_Q = w(n - 1)_Q + s(n)_Q \]  \hspace{1cm} (13)

where \( N \) is a power-of-2 normalization factor introduced to avoid overflow, since it can be verified from (3) that the mean value of \( r(n) \) is \( M \).

Matrix \( R^{-1}(n)_Q \) is updated either as

\[
R^{-1}(n)_Q = \left[ \frac{1 + [(2b-1)T(n)_Q]_Q}{[(2b+1)T(n)_Q]_Q} \right]_Q \times \left[ R^{-1}(n - 1)_Q - \left[ \frac{[t(n)_Q t'(n)_Q]}{[2br(n)_Q]_Q} \right]_Q \right]_Q \]  \hspace{1cm} (14)

for algorithm I or as

\[
R^{-1}(n)_Q = \left[ \frac{1}{1 - \alpha} \right]_Q \times \left[ R^{-1}(n - 1)_Q - \left[ \frac{[t(n)_Q t'(n)_Q]}{[1 + \alpha + r(n)_Q]_Q} \right]_Q \right]_Q \]  \hspace{1cm} (15)

for algorithm II [2].

Since \( e(n) = e(n)_Q - \eta_e(n) \), it follows that

\[ e(n)_Q = d(n) - w^H(n - 1)u(n) \]

\[ = d(n) - \left[ w^H(n - 1)Q u(u(n))_Q \right] - \eta_e(n) \]  \hspace{1cm} (16)

Defining \( \varepsilon_1(n) \) as the error introduced when evaluating the inner product \( [w^H(n - 1)_Q u(n)]_Q \) according to (5a), namely,

\[ \varepsilon_1(n)_Q \triangleq \varepsilon \left[ w^H(n - 1)_Q u(n) \right] \]

\[ = [w^H(n - 1)_Q u(n)]_Q - w^H(n - 1)_Q u(n) \]  \hspace{1cm} (17)

we have

\[ d(n) - w^H(n - 1)_Q u(n) \]

\[ = d(n) - w^H(n - 1)_Q u(n) - \eta_e(n) - \varepsilon_1(n) \]  \hspace{1cm} (18)

which yields

\[ \eta_e(n) = -T_e(n - 1)_Q u(n) - \varepsilon_1(n) \]  \hspace{1cm} (19)

An analogous procedure leads to similar equations for the errors in \( t(n) \) and \( r(n) \), i.e.,

\[ \eta_t(n) = \eta_r(n - 1)_Q u(n) + \varepsilon_2(n) \]

\[ \eta_r(n) = \frac{u^H(n)_Q r(n)}{N} + \varepsilon_3(n) \]  \hspace{1cm} (20)

where

\[ \varepsilon_2(n)_Q \triangleq \varepsilon \left[ R^{-1}(n - 1)_Q u(n) \right] \]  \hspace{1cm} (21)

and

\[ \varepsilon_3(n)_Q \triangleq \varepsilon \left[ \frac{u^H(n)_Q t(n)_Q}{N} \right] \]  \hspace{1cm} (22)

If we assume that quantization is performed by rounding after the additions, then \( \varepsilon_1(n) \), \( \varepsilon_2(n) \), and the elements
The errors in $g'(n)$ and in the variables $r(n)$ and $s(n)$ can be modeled as

\[
\eta_g(n) = t(n)q_{r}(n) + \eta_t(n)g'(n) + \varepsilon_2(n)
\]

and

\[
\eta_r(n) = \frac{q_{r}(n)}{N} + \varepsilon_6(n)
\]

where

\[
\varepsilon_4(n) \triangleq \varepsilon \frac{[e^+](n)q}{\tau'(n)q}
\]

\[
\varepsilon_5(n) \triangleq \varepsilon [t(n)q_{r}g'(n)q]\n\]

and

\[
\varepsilon_6(n) \triangleq \varepsilon \left[ \frac{q_{r}(n)q}{N} \right]
\]

From (13) the quantization error in the coefficient vector can be expressed in terms of the recursive relation

\[
\eta_w(n) = \eta_w(n-1) + \eta_4(n)
\]

The quantization errors in the updating of matrix $R^{-1}(n)$, after neglecting all second-order errors, can be described by

\[
\eta_R(n) \approx \frac{1}{1 - \alpha[n]} \left\{ \eta_R(n-1) - \frac{1}{2\tau(n)} \left[ t(n)q_{r}H(n) + \eta_{r}(n)q_{r}H(n) + \varepsilon_{11}(n) \right] \right. \\
\left. + \frac{t(n)q_{r}H(n)}{2\tau(n)q} \right\} \left[ \eta_{r}(n) + \frac{\varepsilon_{12}(n)}{2b} - \varepsilon_{10}(n) \right] \\
- \left\{ \alpha[n] \frac{\eta_{r}(n)}{\tau(n)} + \frac{\varepsilon_{9}(n)}{2\tau(n)} + \varepsilon_{8}(n) \right\} - \varepsilon_{8}(n)
\]

\[
\times R^{-1}(n) + \varepsilon_{7}(n)
\]

for algorithm I and

\[
\eta_R(n) \approx \frac{1}{1 - \alpha} \left\{ \eta_R(n-1) + \frac{t(n)q_{r}H(n)\eta_{r}(n)}{[\frac{1-\alpha}{\tau(n)} + \tau(n)]^2} - \eta_{r}(n)q_{r}H(n) + t(n)q_{r}H(n) + \varepsilon_{11}(n) \right. \\
\left. - \varepsilon_{10a}(n) \right\} + \varepsilon_{7a}(n)
\]

for algorithm II, where

\[
\varepsilon_{7a}(n) \triangleq \varepsilon \left[ \frac{1 + [(2b - 1)\tau(n)q]q}{[(2b - 1)\tau(n)q]q} \right]
\]

\[
\times \left[ R^{-1}(n-1) - \left\{ \frac{[t(n)q_{r}H(n)]q}{[2b\tau(n)q]q} \right\} \right]
\]

\[
\varepsilon_{7a}(n) \triangleq \varepsilon \left[ \frac{1 + [(2b - 1)\tau(n)q]q}{[(2b - 1)\tau(n)q]q} \right]
\]

\[
\times \left[ R^{-1}(n-1) - \left\{ \frac{[t(n)q_{r}H(n)]q}{[2b\tau(n)q]q} \right\} \right]
\]

\[
\varepsilon_{8}(n) \triangleq \varepsilon \left[ \frac{1 + [(2b - 1)\tau(n)q]q}{[(2b - 1)\tau(n)q]q} \right]
\]

\[
\varepsilon_{10}(n) \triangleq \varepsilon \left[ \frac{[t(n)q_{r}H(n)]q}{[2b\tau(n)q]q} \right]
\]

\[
\varepsilon_{10a}(n) \triangleq \varepsilon \left[ \frac{[t(n)q_{r}H(n)]q}{[2b\tau(n)q]q} \right]
\]

\[
\varepsilon_{11}(n) \triangleq \varepsilon \left[ t(n)q_{r}H(n)q \right]
\]

\[
\varepsilon_{12}(n) \triangleq \varepsilon \left[ 2b\tau(n)q \right]
\]

A. MSE in Finite Precision

By considering the effects of quantization errors on the error signal, the excess MSE due to finite-precision arithmetic can be calculated as

\[
J_{ex}(n) \triangleq E \left[ \eta_{e}(n) \right] = \text{tr} \left[ R K''(n-1) \right] + \sigma_{e}^2
\]

where $\eta_{e}(n)$ is defined in (10) and

\[
K''(n) \triangleq E \left[ \eta_{w}(n-1)\eta_{w}H(n-1) \right]
\]

In (40), it was assumed that $\varepsilon_j(n)$ is a zero-mean white Gaussian noise with variance equal to $\sigma_{e}^2$.

The difference equation that gives $\eta_{w}(n)$ can be expressed as a function of instantaneous quantization errors, infinite-precision quantities, and $\eta_{R}(n-1)$. Considering that the noise sources in (40) are uncorrelated to each other and that the output error after convergence, $e(n)$, can be modeled as a zero-mean white noise, (40)–(41) give a difference equation for $J_{ex}(n+1)$, which can be solved by neglecting the second-order terms in RR$^{-1}(n-1)$ and assuming independence between $\tau(n)$ and each element of $u(n)u^{H}(n)$ taken separately. If we also assume that $\frac{\eta_{R}}{M^2} \ll 1$ and $E \left[ \frac{1}{\tau(n)} \right] \approx \frac{1}{M}$ then for $\sigma_{e}^2 = \sigma_{e}^2 = \sigma_{e}^2$, we have

\[
J_{ex}(n+1) = \frac{M^2 \sigma_{e}^2}{q} \left[ 1 + \frac{M^2 \sigma_{e}^2 \eta_{e}^2}{M^2 \sigma_{e}^2} \right]
\]

\[
J_{ex}(n+1) = \frac{M^2 \sigma_{e}^2}{q} \left[ 1 + \frac{M^2 \sigma_{e}^2 \eta_{e}^2}{M^2 \sigma_{e}^2} \right]
\]
Unfortunately, for both methods of evaluation of the quantized version of $\mathbf{R}^{-1}(n)$, namely, $\mathbf{R}^{-1}(n)Q$, either with variable or constant convergence factor, positive definiteness is not guaranteed, as can be shown by using an analysis similar to that in [3]. This may result in divergence of the coefficients and a stabilization scheme must be incorporated.

### IV. SIMULATION RESULTS

In order to verify the accuracy of the formula in (42), some simulations were performed with the adaptive filter used in a system identification application. The results obtained are summarized in Tables I and II. Columns 2 and 3 show the excess MSE predicted using (42) and that obtained through simulations on an ensemble of 100 experiments, respectively.

<table>
<thead>
<tr>
<th>Number of Bits</th>
<th>Eqn.(42)</th>
<th>Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>-20.1</td>
<td>-21.9</td>
</tr>
<tr>
<td>7</td>
<td>-26.2</td>
<td>-27.1</td>
</tr>
<tr>
<td>8</td>
<td>-32.2</td>
<td>-32.8</td>
</tr>
<tr>
<td>9</td>
<td>-38.2</td>
<td>-38.6</td>
</tr>
<tr>
<td>12</td>
<td>-56.3</td>
<td>-56.5</td>
</tr>
</tbody>
</table>

Table I shows the results obtained for different values of wordlength and $q = 1$. The unknown system is time-invariant and is of order 20. As can be noted the simulation results agree with the theoretical results, as long as a sufficient wordlength is provided to avoid the coefficients from freezing. Table II shows the results obtained for different values of the parameter $q$, with a constant wordlength of 9 bits. In order to avoid overflow and to simplify the internal scaling, the input signal variance was made 0 dB and 3 additional bits were used to represent the integer part in internal registers in all the finite-precision simulations. The variance of the additional noise used in the simulations shown in Tables I and II is -20 dB. Using this scheme, only the register storing $\tau(n)$ had to be treated separately; its content was shifted 3 bits to the right.

The matrix $\mathbf{R}^{-1}(n)Q$ was evaluated with double-precision floating-point arithmetic, and then quantized. This approach prevents divergence. We are implicitly assuming that some strategy has been used to guarantee the stability of $\mathbf{R}^{-1}(n)Q$, for example the one proposed in [3]. This strategy can be used since, as will be shown in the paper, $\eta_H$ has no significant influence on $J_{ex}(n)$.  

### V. CONCLUSION

An analysis of an LMSN algorithm for adaptive filtering that incorporates a variable convergence factor in finite precision (fixed-point) has been carried out. The analysis includes the propagation of quantization errors in all internal variables, and leads to a closed-form formula for the excess MSE due to finite precision. The experimental results obtained through several simulations agree well with those predicted by using the formula derived.

### REFERENCES


