TRANSIENT BEHAVIOR OF FIXED POINT LMS ADAPTATION

Riten Gupta and Alfred O. Hero III

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109-2122

ABSTRACT

We relate the distinguishing features of the fixed point power-of-two step size LMS algorithm's learning curve to the precision of its data and coefficient variables. In particular, we show that the increase in the steady state MSE floor due to finite precision effects is determined primarily by data quantization while the decrease in convergence rate due to finite precision is determined by both data and coefficient quantization. We also derive a condition under which the slowdown phenomenon can be eliminated, given the reference variance and lower bounds on the minimum MSE and optimal weight vector magnitude.

1. INTRODUCTION

The LMS algorithm is the most commonly used adaptive algorithm for such tasks as channel equalization and system identification. The algorithm is used to adapt the coefficients, or weights, of an FIR filter to minimize the mean square error (MSE) between the filter's output and a primary signal. Two important performance measures of the algorithm are its steady state MSE and convergence rate [1]. In practice, the LMS algorithm is implemented in finite precision, and often with fixed point arithmetic [2, 3, 4].

The performance penalty incurred as a result of finite precision implementation has been analyzed in [2] and [3, 4]. In [2], Caraiscos and Liu derived an expression for the increase in steady state MSE due to quantization of data and coefficients using an additive white noise model for the quantizers. Bermudez and Bershad [3, 4] considered an implementation of the LMS algorithm with infinite precision data and finite precision coefficients and, by using a nonlinear model for the coefficient quantizer, derived a recursion that accurately predicts the MSE trajectory of this algorithm. For the algorithm considered in [3, 4] it was shown that the stopping phenomenon is actually a slowdown phenomenon that is always present and renders the analysis of [2] inapplicable.

In this paper we show that the slowdown phenomenon can be eliminated in the general (where both data and coefficients are quantized) fixed point power-of-two step size algorithm by the proper choice of data and coefficient wordlength. Specifically, for most practical cases, more bits should be allocated to coefficients than data to prevent slowdown. We show that the algorithm considered in [4] is a limiting case of the general fixed point LMS algorithm in which slowdown can not be prevented. Finally, we show that both analyses [2] and [3, 4] give useful insights into the behavior of the finite precision LMS algorithm.

2. FIXED POINT LMS ADAPTATION

Given a complex primary signal, y_k, and a complex reference signal, x_k, that is correlated with y_k, the infinite precision complex LMS algorithm adapts the complex coefficients, $\underline{w}_k = [w_{0,k}, \ldots, w_{p-1,k}]^H$ (weight vector), of a p-tap FIR filter such that the mean square error between the filter output, $\hat{y}_k = \underline{w}_k^H \underline{x}_k$, and the primary signal is minimized. Here $\underline{x}_k = [x_k, \ldots, x_{k-p+1}]^T$ is a vector of the p most recent samples of x_k. Both x_k and y_k are assumed to be wide sense stationary. The adaptation of the weight vector is accomplished according to the recursive weight update equation

$$\begin{aligned} \underline{w}_{k+1} &= \underline{w}_k + \mu \underline{x}_k e_k^* \\ e_k &= y_k - \hat{y}_k \end{aligned} \tag{1}$$

where μ is the adaptive gain parameter which determines the transient behavior of the algorithm.

The finite precision LMS algorithm implements the recursive weight update (1) with quantizers in all data and coefficient paths. In this paper we consider a special case of the finite precision algorithm in which all quantizers are uniform scalar quantizers, all arithmetic is done in fixed point, and $\mu = 2^{-q}$ where $q \in \{0, 1, \ldots\}$. We refer to this algorithm as the fixed point power-of-two step size LMS algorithm.

Define the quantization operators $Q_d(\cdot)$ and $Q_c(\cdot)$ as fixed point rounding quantizers with granularities $\Delta_d = 2^{-B_d}$ and $\Delta_c = 2^{-B_c}$, respectively. Thus the data quantizer, $Q_d(\cdot)$, uses B_d bits plus sign and the coefficient quantizer, $Q_c(\cdot)$, uses B_c bits plus sign. Then, with $\underline{w}'_k = Q_c(\underline{w}_k)$, $\underline{x}'_k = Q_d(\underline{x}_k)$, and $y'_k = Q_d(y_k)$, the fixed point algorithm's weight update recursion is

$$\begin{aligned} \underline{w}'_{k+1} &= \underline{w}'_k + Q_c\left(\mu \underline{x}'_k e'^*_k\right) \\ e'_k &= y'_k - \hat{y}'_k \\ &= y'_k - \sum_{i=0}^{p-1} Q_d(w'_{i,k} x'_{k-i}). \end{aligned} \tag{2}$$

This research was supported in part by the Department of Defense Research & Engineering (DDR&E) Multidisciplinary University Research Initiative (MURI) on Low Energy Electronics Design for Mobile Platforms and managed by the Army Research Office (ARO) under grant ARO DAAH04-96-1-0337. Corresponding author: Riten Gupta, robby@eecs.umich.edu.

0-7803-6293-4/00/$10.00©2000 IEEE

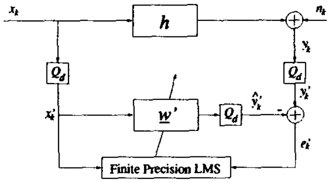

Figure 1: *Finite Precision LMS Algorithm used in system identification configuration.*

Since $\mu = 2^{-q}$, the multiplication by μ in (2) is accomplished by a right shift of q bits.

Figure 1 shows the finite precision LMS algorithm used for identification of an unknown FIR system, h, with impulse response vector $\underline{h} = [h_0, \ldots, h_{L-1}]^T$. Note that with infinite precision and $p = L$, the optimal weight vector is $\underline{w}^o = \underline{h}$ and the minimum MSE is $\xi_{min} = \sigma_n^2 = E[|n_k|^2]$.

In this paper, we consider x_k to be a zero-mean, circular, white Gaussian sequence with variance $\sigma_x^2 \leq 0.1$. With this choice, the probability of quantizer overload is small. In addition, for the simulations, we focus on the case $p = L = 4$.

3. SLOWDOWN PHENOMENON

In [2] it was noted that the fixed point LMS algorithm suffers from a potentially hazardous condition in which the weight update (2) stops prematurely. This condition was called the stopping phenomenon and was believed to occur when the argument of the weight update quantizer, Q_c, in (2) fell into the quantizer's dead zone. Mathematically, this can be expressed as $|\text{Re}\{\mu x'_{k-i} e'^*_k\}| < \frac{\Delta_c}{2}$. for some $i \in \{0, \ldots, p-1\}$. Here $\text{Re}\{\cdot\}$ denotes the real part. We have assumed that the real and imaginary parts of x_{k-i} are uncorrelated and have equal variance. In [3, 4] it was shown that this phenomenon does not stop adaptation as previously believed, but instead severely reduces the convergence rate. Thus a slowdown, not stopping, phenomenon was taking place.

By defining $\xi'_k = E[|e'_k|^2]$, and using the above condition for slowdown as well as the independence assumption, we can determine ξ'_{slow}, the minimum value of ξ'_k before the onset of slowdown. This approach, however, yields an overly conservative estimate as slowdown typically occurs well after the above condition is satisfied. Thus we propose the following condition for slowdown

$$P\left(|\text{Re}\{\mu x'_{k-i} e'^*_k\}| < \frac{\Delta_c}{2}\right) > 1 - \epsilon \qquad (3)$$

for some $i \in \{0, \ldots, p-1\}$ and $0 < \epsilon \ll 1$. To derive an accurate estimate of ξ'_{slow} using (3), we first assume that the transient behavior of the fixed point algorithm is approximately the same as that of the corresponding infinite precision algorithm prior to slowdown. This assumption has been shown to be correct by experimental evidence [3, 4]. Then we can replace $x'_{k-i}e'^*_k$ in (3) with $g_k = x_{k-i}e^*_k$. Next we assume g_k is approximately circular Gaussian with mean zero and variance $\sigma_x^2 \sigma_{e,k}^2$ where $\sigma_{e,k}^2 = E[|e_k|^2]$ is the mean square error trajectory (learning curve) of the infinite precision algorithm. Under the assumption that $\xi'_k = \sigma_{e,k}^2 = E[|y_k - \hat{y}_k|^2]$ prior to slowdown, using (3), we propose the estimate, K, of the iteration at which slowdown commences as the integer, K, for which

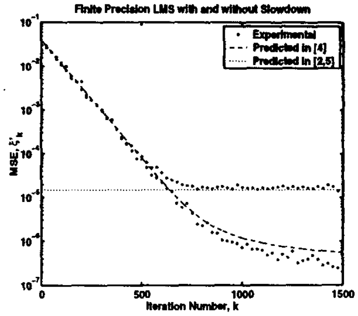

Figure 2: *Experimental and theoretical MSE, ξ'_k, predicted in [4] and [2, 5]. Both curves use $\mu = 1/8$, $\underline{h} = [0.7, -0.2, 0.5, -0.1]^T$, $\sigma_x^2 = 0.05$, $\sigma_n^2 = 10^{-8}$. The slowdown curve uses $B_d = 15$, $B_c = 12$ and real signals while the other uses $B_d = 8$, $B_c = 12$ and complex signals.*

$$\xi'_k < \frac{2^{-2(B_c+1-q)}}{\sigma_x^2[\text{erf}^{-1}(1-\epsilon)]^2} \triangleq \xi'_{slow}, \quad \forall k \geq K. \qquad (4)$$

Experiments have shown that the value of ϵ for which K predicts the actual slowdown time depends only on μ and p. More details follow in Section 6.

4. PERFORMANCE ANALYSIS

The finite precision LMS algorithms analyzed in [3, 4] are of the same type as (2) with the restriction $B_d = \infty$. It was found that under this condition, the slowdown phenomenon was unavoidable. Fortunately, by using a nonlinear coefficient quantizer model, recursions were obtained that accurately predicted the transient behavior, including slowdown.

The analysis in [2] was of the general fixed point algorithm of (2). There, a closed form solution for ξ'_q, the increase in MSE due to finite precision data and coefficients, was derived for the real finite precision LMS algorithm under the assumption that slowdown does not occur. In [5] this derivation was extended to the case of complex data and coefficients giving [1] $\xi'_\infty \approx \xi'_q + \xi_{min}$ for small values of the gain parameter, μ, where

$$\xi'_q \approx \left[\frac{p}{12\mu}\right] 2^{-2B_c} + \left[\frac{\|\underline{w}^o\|^2 + p + 1}{6}\right] 2^{-2B_d}. \qquad (5)$$

Figure 2 shows the learning curves for two implementations of the fixed point LMS algorithm in the system ID configuration. The algorithm exhibiting slowdown is predicted accurately by [4] while the steady state MSE of the algorithm without slowdown is accurately predicted by [2, 5].

[1] Note that ξ'_q differs slightly from ξ_q in [5] as ξ'_q is the asymptotic increase in $E[|y'_k - \hat{y}'_k|^2]$ while ξ_q is the asymptotic increase in $E[|y_k - \hat{y}'_k|^2]$.

μ	ϵ	ν	$B_d = 7$	$B_d = 9$
$\frac{1}{4}$	5×10^{-4}	1.86	0	0
$\frac{1}{8}$	3×10^{-2}	3.54	0	0
$\frac{1}{16}$	1×10^{-1}	4.94	0	0

Table 1: Number of systems of fixed norm for which finite precision algorithm using $B_c = B_d + \lceil \nu \rceil$ exhibits slowdown out of 25 4-tap FIR systems selected at random with the constraint $\|\underline{h}\|^2 = 1$.

μ	ϵ	$B_d = 7$	$B_d = 9$
$\frac{1}{4}$	5×10^{-4}	0	0
$\frac{1}{8}$	3×10^{-2}	0	0
$\frac{1}{16}$	1×10^{-1}	0	0

Table 2: Number of systems of varying norm for which finite precision algorithm using $B_c = B_d + \lceil \nu \rceil$ exhibits slowdown for systems of the type $\underline{h} = \alpha[1/2, 1/2, 1/2, 1/2]^T$ and 10 values of α ranging from 0.1 to 1.0.

The results of [4] are valid since the algorithm with slowdown uses fine data quantization compared to coefficient quantization. Similarly, the prediction of [2, 5] is accurate because the algorithm in question does not exhibit slowdown. In the next section we derive bounds on B_c and B_d that ensure that slowdown does not occur.

5. PREVENTION OF SLOWDOWN

Assuming slowdown does not occur, we have $\xi'_\infty \approx \xi'_q + \xi_{min}$. This implies $\xi'_\infty > \xi'_q|_{B_c=\infty} + \xi_{min}$ and therefore

$$\xi'_\infty > \frac{2^{-2B_d}}{6}(\|\underline{w}^o\|^2 + p + 1) + \xi_{min} \triangleq \xi'_{floor}. \qquad (6)$$

Now, since the infinite precision and finite precision algorithms agree closely before slowdown, slowdown can be prevented by choosing B_c such that $\xi'_{slow} < \xi'_{floor}$. Using (4) and (6), this condition becomes

$$\begin{aligned} 2^{-2(B_c+1-q)} \quad < \quad & \sigma_x^2[\text{erf}^{-1}(1-\epsilon)]^2 \cdot \\ & \left[\frac{2^{-2B_d}}{6}(\|\underline{w}^o\|^2 + p + 1) + \xi_{min}\right]. \quad (7) \end{aligned}$$

Application of this formula to choosing B_c requires knowledge of both the minimum MSE, ξ_{min}, and the optimal weight vector magnitude, $\|\underline{w}^o\|$. Although these quantities are usually unknown a priori by the designer, a sufficient condition for preventing slowdown can be obtained by lower bounding these values. In the particular case where $2^{-2B_d} \gg \xi_{min} \geq 0$ and $\|\underline{w}^o\|^2 \geq \psi$, the sufficient condition in B_c, B_d for no slowdown is

$$B_c > B_d + \nu \qquad (8)$$

where

$$\nu = q - 1 - \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{6}(\psi + p + 1)[\text{erf}^{-1}(1-\epsilon)]^2\right). \qquad (9)$$

6. SYSTEM ID DESIGN EXAMPLES

In this section, we consider fixed point LMS system identification of a 4-tap FIR system using step size, μ, chosen from $\{1/16, 1/4, 1/2\}$, and $p = 4$. To determine ϵ for each value of μ, the fixed point algorithm was simulated with a baseline set of parameters for which slowdown is exhibited. In this case, these parameters are $B_c = B_d = 10$, $\sigma_x^2 = 0.05$, $\sigma_n^2 = 10^{-8}$, and $\underline{h} = [1/2, 1/2, 1/2, 1/2]$. Then, ϵ was computed by determining the actual slowdown point and inverting (4). Selection of ϵ by this procedure has been shown experimentally to provide accurate predictions of ξ'_{slow} and, consequently, valid wordlength constraints. Furthermore, these results remain valid as the FIR system, $\underline{h}$, the reference variance, σ_x^2, and the wordlengths, B_c and B_d, are varied. Of course, as μ and p vary, ϵ must be recomputed.

With ϵ known for each value of μ, ν in (8) can be calculated for various systems, $\underline{h}$, and various signal powers, σ_x^2. To demonstrate the validity of the ϵ-selection procedure, ν was calculated using (9) with $\psi = 1$, $p = 4$, and $\sigma_x^2 = 0.05$. The algorithm was then simulated with $\sigma_x^2 = 0.05$, $\sigma_n^2 = 10^{-8}$, $B_c = B_d + \lceil \nu \rceil$ and $B_d \in \{7, 9\}$ with 25 randomly chosen FIR systems, $\underline{h}$, satisfying the constraint $\|\underline{h}\|^2 = 1$. The resulting experimental learning curves were inspected for slowdown. Table 1 shows that for all cases, slowdown was not exhibited. The robustness of the procedure to changing $\|\underline{h}\|^2$ can be seen from Table 2. Here, the systems, $\underline{h}$, are of the form $\underline{h} = \alpha[1/2, 1/2, 1/2, 1/2]^T$. Ten values of α were chosen between 0.1 and 1.0. The value ν was calculated using $\psi = \alpha^2$ and the same ϵ determined from the baseline experiment. Again, for all systems considered, slowdown was not exhibited. Similar results were obtained when the system, $\underline{h}$, was fixed and σ_x^2 was varied, although they are not shown here. Note that there is a possibility for the bound (8) to be too conservative. It is therefore possible that for these examples, the value ν prevents slowdown but is larger than necessary.

Figure 3 shows three fixed point LMS learning curves with $\mu = 1/4$ and $\underline{h} = [0.7, -0.2, 0.5, -0.1]^T$. The wordlengths, B_c and B_d, are chosen to satisfy (8). Clearly, slowdown is not evident on these curves. Note also that the prediction of [2, 5] is accurate. Finally, note that the fixed point algorithm behavior is the same as that of the infinite precision algorithm in the transient region. This validates the assumption made earlier.

7. SLOWDOWN VS. MSE TRADEOFF

Based on the inequality (8), which gives the condition under which slowdown is avoided, we conclude that the convergence rate of the general fixed point power-of-two step size LMS algorithm is dependent on both data and coefficient wordlengths. Specifically, if $B_c > B_d + \nu$, the convergence rate is approximately equal to that of the infinite precision algorithm. As B_c falls below the threshold wordlength, $B_d + \nu$, the convergence rate should decrease.

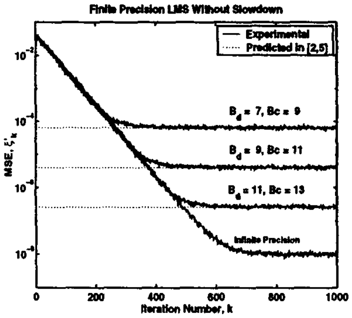

Figure 3: *Experimental and theoretical MSE, ξ'_k, predicted in [2, 5]. All curves use $\mu = 1/4$, $\underline{h} = [0.7, -0.2, 0.5, -0.1]^T$, $\sigma_x^2 = 0.05$, $\sigma_n^2 = 10^{-8}$.*

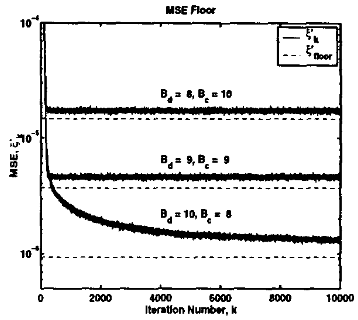

Figure 4: *Experimental MSE, ξ'_k, and theoretical MSE floor, ξ'_{floor} with $\mu = 1/2$, $\underline{h} = [0.7, -0.2, 0.5, -0.1]^T$, $\sigma_x^2 = 0.05$, $\sigma_n^2 = 10^{-8}$, $B_T = 18$.*

Although the bound (6) uses the results of [2], which are applicable only in the absence of slowdown, (6) is still valid for implementations in which slowdown occurs. In such cases, the formula (5) will indeed be overly conservative. However, the steady state will eventually be reached and the finite precision algorithm will obey the bound (6). It is in the steady state region that (6) becomes useful. Note that when $B_d = \infty$, (6) becomes $\xi'_\infty > \xi_{min}$. This is precisely the behavior observed and analyzed in [3, 4]. Figure 4 shows the learning curves for three fixed point systems with $\mu = 1/2$, along with the value ξ'_{floor} for each system. This figure shows that (6) is a tight lower bound even when slowdown occurs. Thus we conclude that the MSE floor is determined by the data wordlength, although the steady-state region may take many iterations to reach.

These results suggest a tradeoff between slowdown and steady state MSE. To see this, assume the fixed point algorithm must use a total of $B_T + 2$ bits with $B_T = B_d + B_c$. If the constraint (8) is satisfied, the slowdown phenomenon will be eliminated completely and the convergence rate will be almost identical to that of the infinite precision algorithm. On the other hand, if (8) is not met and B_d is increased, slowdown will occur, but if given enough iterations to converge completely, the steady state MSE will be reduced. The three systems in Figure 4 all use $B_T = 18$. Observe that as expected, the data wordlength determines the MSE floor while the convergence rate is determined by the quantity $B_c - B_d$. Note, however, from this figure that choosing B_c too large can result in increased MSE without much improvement in convergence rate. In this case, it is beneficial to use more data bits than coefficient bits as the decrease in convergence rate is not drastic compared to the decrease in MSE. For smaller values of μ, however, it is beneficial to use more coefficient bits as slowdown can have a more significant impact.

8. CONCLUSION

We have analyzed the fixed point power-of-two step size LMS algorithm and derived the threshold MSE which determines the onset of the slowdown phenomenon under the independence assumption on the LMS update term, $\underline{x}_k e_k^*$. We have also derived a constraint on the wordlengths of the data and coefficients under which the slowdown phenomenon can be avoided.

We conclude that the steady state MSE floor of the fixed point power-of-two step size LMS algorithm is determined primarily by the data wordlength, while the convergence rate depends on both data and coefficient wordlengths.

We also conclude that the analyses in [2] and [3, 4] are both valid and give correct performance predictions for special cases of the general fixed point LMS algorithm. Specifically, [2] provides a closed-form solution for the increase in steady state MSE when the constraint (8) is met and [3, 4] gives a deterministic recursion for the learning curve when the data resolution is high. It remains to combine the efforts of [2] and [3, 4] to determine a recursion for the learning curve in the general case of coarse data and coefficient quantization. It also remains to relax the independence assumption on the LMS update term. Finally, an analytic process for selection of ϵ using μ and p is yet unavailable.

9. REFERENCES

[1] B. Widrow and S. D. Stearns, *Adaptive signal processing*, Prentice Hall, Englewood Cliffs NJ, 1985.

[2] C. Caraiscos and B. Liu, "A roundoff error analysis of the LMS adaptive algorithm," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. ASSP-32, no. 1, pp. 34–41, Feb. 1984.

[3] J. C. M. Bermudez and N. J. Bershad, "A nonlinear analytical model for the quantized LMS algorithm - the arbitrary step size case," *IEEE Transactions on Signal Processing*, vol. SP-44, no. 5, pp. 1175–1183, May 1996.

[4] N. J. Bershad and J. C. M. Bermudez, "A nonlinear analytical model for the quantized LMS algorithm - the power-of-two step size case," *IEEE Transactions on Signal Processing*, vol. SP-44, no. 11, pp. 2895–2900, Nov. 1996.

[5] R. Gupta and A. O. Hero, "Theoretical aspects of power reduction for adaptive filters," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Phoenix, AZ, March 1999.