THEORETICAL ASPECTS OF POWER REDUCTION FOR ADAPTIVE FILTERS

Robby Gupta and Alfred O. Hero

Department of EECS
University of Michigan
Ann Arbor, MI 48109-2122

ABSTRACT

Adaptive filters are used in a number of applications, many of which can benefit from a reduction in power. In this paper we present derivations of the approximate expressions used in [1] for the increase in mean square error of the LMS adaptive algorithm when the total processing power is decreased.

1. INTRODUCTION

The power consumed by the LMS adaptive algorithm can be reduced by a reduction of the number of bits used to represent the data and control variables. Bit width reduction, however, generally entails a degradation in algorithm performance, as measured by steady state mean square error (MSE). This paper provides an analysis of MSE degradation versus power reduction for the LMS algorithm.

The LMS algorithm was introduced by Widrow [2] and is one of the most common adaptation algorithms found in practical systems such as channel equalizers [3]. We consider a quantized version of the LMS algorithm, called QLMS, which is an LMS algorithm implemented with separate uniform scalar quantizers in the data path and the filter coefficient path, where the quantizers can have different resolutions.

We first present a formula for the increase in steady state mean square error (MSE) due to quantization which generalizes the formulas of Caraiscos and Liu [4] to the case of complex data and coefficients. We then derive the optimal bit-allocation factor which minimizes the increase in MSE subject to a total power constraint. We then show that, using this optimal allocation factor, the relation between the LMS algorithm's performance and its power consumption can be derived.

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: Alfred Hero, hero@eecs.umich.edu, (734)-763-0564 (voice), (734)-763-8041 (fax).

2. THE QUANTIZED LMS ADAPTIVE ALGORITHM

The standard LMS algorithm adapts its filter coefficients in an attempt to minimize the quadratic surface specified by the mean squared error: $E[|y_k - \hat{s}_k|^2]$ [2] where y_k is a desired training signal and $\hat{s}_k$ is the filter output. If $\underline{w}_k$ is the filter coefficient (weight) vector and $\underline{x}_k$ is the filter input vector of length p, then the LMS update equation is:

$$\underline{w}_{k+1} = \underline{w}_k + \mu\, \underline{x}_k \left(y_k - \underline{x}_k^H \underline{w}_k\right). \tag{1}$$

where μ is the gain parameter.

Let the operators $Q_d[\cdot]$ and $Q_c[\cdot]$ represent uniform scalar quantization to $B_d + 1$ and $B_c + 1$ bits, respectively. If we quantize all data to $B_d + 1$ bits and all coefficients to $B_c + 1$ bits, then the quantized LMS update equation becomes:

$$\underline{w}_{k+1} = \underline{w}_k + Q_c\left(\mu Q_d(\underline{x}_k) \cdot e'_k\right) \tag{2}$$

where

$$e'_k = Q_d(y_k) - Q_d\left(Q_d(\underline{x}_k^H) \cdot Q_c(\underline{w}_k)\right). \tag{3}$$

is the quantized error signal.

3. POWER CONSUMPTION OF LMS ALGORITHM

In the update formula given by (2) and (3), the calculation of the inner product $Q_d\left(Q_d(\underline{x}_k^H) \cdot Q_c(\underline{w}_k)\right)$ requires p complex multiplications of numbers represented with $B_d + 1$ bits. (Although the weight vector is quantized to $B_c + 1$ bits, the product is stored in $B_d + 1$ bits and therefore, we will assume the multiplication is done in $B_d + 1$ bits.) In addition, this calculation requires $p - 1$ complex additions of $B_d + 1$ bits. Subtracting this inner product from $Q_d(y_k)$ requires an additional complex addition of $B_d + 1$ bits. Next, multiplying this quantity (e'_k) by $Q_d(\underline{x}_k)$ we have p additional complex multiplications of $B_d + 1$ bits. Note that the product will be stored in $B_c + 1$ bits. If we assume the number μ is a power of 2 and hence multiplication by it requires only a bit shift, we are left with the addition by $\underline{w}_k$. This operation requires p complex additions of $B_c + 1$

bits. Therefore, the LMS update formula requires a total of $2p$ complex multiplications of B_d+1 bits, p complex additions of B_d+1 bits, and p complex additions of B_c+1 bits. In terms of total real operations, we have $8p$ multiplications of B_d+1 bits, $4p$ additions of B_d+1 bits, and $4p$ additions of B_c+1 bits.

In a B_d+1 bit table lookup multiplier, each multiplication requires three table lookups and one B_d bit addition (no sign bit). Therefore, in the LMS update formula, we have $4p$ additions of B_d+1 bits, $4p$ additions of B_c+1 bits, $8p$ additions of B_d bits, and $24p$ table lookups. An addition of B bits requires $B-1$ full adders and 1 half adder. Therefore, we have $16p$ half adders and $4p(3B_d+B_c-2)$ full adders in addition to the $24p$ table lookups. Now, each full adder uses 6 logic gates while each half adder uses 2 gates. Therefore, we have a total of $24p(3B_d+B_c-2)+32p$ logic gates and $24p$ table lookup operations.

Next we define η_g to be the average power consumed per logic gate during an iteration of LMS, and η_t to be the average power per table lookup per bit. Then we have the following expression for total power dissipation per iteration of LMS:

$$P_T = [24p(3B_d + B_c - 2) + 32p]\eta_g + 24p(B_d\eta_t). \quad (4)$$

This expression is linear in the number of bits B_d and B_c and assumes fixed point complex arithmetic and multiplication using table lookup as opposed to adding partial products.

4. PERFORMANCE OF LMS ALGORITHM

Using the same approach as Caraiscos and Liu [4], we obtain an expression for the steady-state increase in MSE of the complex LMS algorithm due to quantization, under a circular Gaussian assumption on $\underline{x}_k$. We assume that the sequence y_k has been properly scaled to prevent overflow in the calculations.

We define:

$$\sigma_d^2 = \frac{1}{12}2^{-2B_d}, \qquad \sigma_c^2 = \frac{1}{12}2^{-2B_c}. \quad (5)$$

These are the variances of the quantization noises added to the data and coefficients, respectively. We also use primed symbols to represent quantized values and define the following quantities:

$$\begin{aligned} \underline{x}'_k &= \underline{x}_k + \underline{\alpha}_k \\ y'_k &= y_k + \beta_k \\ \underline{w}'_k &= \underline{w}_k + \underline{\rho}_k. \end{aligned} \quad (6)$$

The components of the vector $\underline{\alpha}_k$ and β_k are complex numbers whose real and imaginary parts are assumed uncorrelated and have variances σ_d^2. Therefore, the variance of β_k and of each component of $\underline{\alpha}_k$ is $2\sigma_d^2$.

We can now calculate the quantized value of the filter output:

$$\hat{s}'_k = Q_d(\underline{x}'^H_k \underline{w}'_k) = \underline{x}^H_k \underline{w}_k + \underline{x}^H_k \underline{\rho}_k + \underline{\alpha}^H_k \underline{w}_k + \eta_k \quad (7)$$

where η_k is a complex quantization noise with variance $2p\ \sigma_d^2$. This variance arises as a result of quantizing the individual products of the inner product $\underline{x}'^H_k \underline{w}'_k$ and then summing them [4]. We ignore all second-order noise terms.

The total output error is now:

$$y_k - \hat{s}'_k = (y_k - \underline{x}^H_k \underline{w}_k) - (\underline{x}^H_k \underline{\rho}_k + \underline{\alpha}^H_k \underline{w}_k + \eta_k) \quad (8)$$

The first term on the right hand side of of (8) is the error of the unquantized algorithm. The second term is the error due to quantization and will be denoted e_q. Similar to [4], these terms are uncorrelated. Also, $\underline{\alpha}_k$, β_k, $\underline{\rho}_k$, and η_k are assumed uncorrelated. We then have the following expression for the increase in MSE due to quantization:

$$\xi_q = E[|e_q|^2] = E\left[|\underline{x}^H_k \underline{\rho}_k|^2 + |\underline{\alpha}^H_k \underline{w}_k|^2 + |\eta_k|^2\right]. \quad (9)$$

The last term, $E[|\eta_k|^2]$ is equal to $2p\ \sigma_d^2$. For the term $E[|\underline{\alpha}^H_k \underline{w}_k|^2]$ we obtain:

$$E[|\underline{\alpha}^H_k \underline{w}_k|^2] = 2\sigma_d^2\ E[|\underline{w}_k|^2]. \quad (10)$$

Note that this differs from the real case studied in [4] by a factor of 2. In the steady state, and assuming μ is small, this becomes:

$$E[|\underline{\alpha}^H_k \underline{w}_k|^2] = 2\sigma_d^2\ |\underline{w}^o|^2. \quad (11)$$

To calculate $E[|\underline{x}^H_k \underline{\rho}_k|^2]$ we invoke the assumption that $\underline{x}_k$ is a circularly Gaussian random vector and μ is small. In the steady state, this gives:

$$E[|\underline{x}^H_k \underline{\rho}_k|^2] = \frac{p\ \sigma_c^2}{\mu}. \quad (12)$$

Now, using (9), (11), (12), and (5) we obtain the final result:

$$\xi_q = \alpha_c\ 2^{-2B_c} + \alpha_d\ 2^{-2B_d} \quad (13)$$

where

$$\alpha_c = \frac{p}{12\mu}, \qquad \alpha_d = \frac{\|\underline{w}^o\|^2 + p}{6}. \quad (14)$$

The first term in the expression (13) is the MSE due only to quantization of the filter coefficients while the second term represents the MSE due to quantization of the data. Note that ξ_q increases in p at a linear rate, decreases in μ at an inverse square root rate, and decreases in B_d and B_c at an exponential rate. Therefore, the total number of bits allocated gives more leverage over excess MSE than any other of the design parameters.

With these relations the increase in MSE due to quantization, ξ_q, can be plotted as a function of B_d and B_c. A plot for the increase in MSE occurring for $\|\underline{w}^o\| = p = 2$, and $\mu = 0.1$ is given in figure 1.

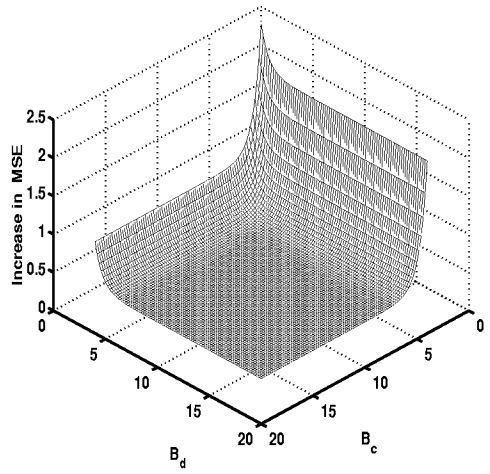

Figure 1: *Excess MSE due to quantization as a function of B_d and B_c*

5. OPTIMAL BIT ALLOCATION FACTOR

We now derive the optimal bit-allocation factor which minimizes the increase in MSE due to quantization of the data and filter coefficients subject to a total power constraint.

Assume that there are a total of B_T bits plus 2 sign bits which are available to allocate between data and coefficients, i.e. $B_T = B_d + B_c$. Further define the data bit allocation factor $\rho = B_d/B_T$. Then we have the obvious relations

$$B_d = \rho B_T, \qquad B_c = (1-\rho)B_T. \tag{15}$$

and we can write (13) as:

$$\xi_q = \alpha_c 2^{-2(1-\rho)B_T} + \alpha_d 2^{-2\rho B_T}. \tag{16}$$

In the following, we show that ξ_q is convex in ρ for $B_T \geq 2$ and thus the optimal ρ can be found by setting the derivative equal to zero.

Under the constraint on P_T, we can use (4) and (15) to re-express the total combined number of bits B_T as a function of ρ and P_T

$$B_T = \frac{P_T + 16p\eta_g}{p[\rho(48\eta_g + 24\eta_t) + 24\eta_g]} \tag{17}$$

Now we define the following constants:

$$\begin{aligned} A &\equiv \frac{P_T + 16p\,\eta_g}{p}, \\ B &\equiv 48\,\eta_g + 24\,\eta_t, \qquad C \equiv 24\,\eta_g. \end{aligned} \tag{18}$$

Then from (17) we have:

$$B_T = \frac{A}{B\rho + C} \tag{19}$$

Differentiating (16) with respect to ρ gives:

$$\begin{aligned} \frac{d\xi_q}{d\rho} &= \ln 2\,\alpha_c\, 2^{-2(1-\rho)B_T} \cdot \frac{d}{d\rho}[-2(1-\rho)B_T] \;+ \\ &\quad \ln 2\,\alpha_d\, 2^{-2\rho\, B_T} \cdot \frac{d}{d\rho}[-2\rho\, B_T] \end{aligned} \tag{20}$$

Differentiating again, we have:

$$\begin{aligned} \frac{d^2\xi_q}{d\rho^2} &= \ln 2\,\alpha_c\, 2^{-2(1-\rho)B_T}\left[\frac{d^2}{d\rho^2}[-2(1-\rho)B_T] + \right. \\ &\quad \left. \ln 2\left(\frac{d}{d\rho}[-2(1-\rho)B_T]\right)^2\right] + \\ &\quad \ln 2\,\alpha_d\, 2^{-2\rho B_T}\left[\frac{d^2}{d\rho^2}[-2\rho B_T] + \right. \\ &\quad \left. \ln 2\left(\frac{d}{d\rho}[-2\rho B_T]\right)^2\right]. \end{aligned} \tag{21}$$

Using (19) and (21) we have:

$$\begin{aligned} \frac{d^2\xi_q}{d\rho^2} &= \ln 2\,\alpha_c\, 2^{-2(1-\rho)B_T}\left[-\frac{4AB(B+C)}{(B\rho+C)^3} + \right. \\ &\quad \left. \ln 2 \cdot \frac{4A^2(B+C)^2}{(B\rho+C)^4}\right] + \\ &\quad \ln 2\,\alpha_d\, 2^{-2\rho B_T}\left[\frac{4ABC}{(B\rho+C)^3} + \right. \\ &\quad \left. \ln 2 \cdot \frac{4A^2C^2}{(B\rho+C)^4}\right]. \end{aligned} \tag{22}$$

Now ξ_q is convex if $\frac{d^2\xi_q}{d\rho^2}$ is positive. From (22), it is clear that $\frac{d^2\xi_q}{d\rho^2}$ will be positive if the following condition is met:

$$\ln 2 \cdot \frac{4A^2(B+C)^2}{(B\rho+C)^4} > \frac{4AB(B+C)}{(B\rho+C)^3}. \tag{23}$$

This is equivalent to

$$A > \frac{B(B\rho+C)}{\ln 2(B+C)}. \tag{24}$$

Using (19), this becomes:

$$P_T > \frac{N}{\ln 2} \cdot \frac{B(B\rho+C)}{B+C} - 16p\,\eta_g. \tag{25}$$

Now, using (17), this condition will be satisfied if and only if the following condition is satisfied:

$$B_T > \frac{B}{\ln 2(B+C)} \tag{26}$$

which is equivalent to

$$B_T > \frac{1}{\ln 2}\left[1 - \frac{24\eta_g}{72\eta_g + 24\eta_t}\right]. \tag{27}$$

The term on the right hand side of (27) is clearly less than $1/\ln 2$. This means that $\frac{d^2\xi_q}{d\rho^2}$ will be positive if $B_T > 1/\ln 2$. This condition is clearly true under the assumption $B_T \geq 2$. Therefore, ξ_q is a convex function of ρ.

To solve for ρ^{**}, the optimal bit allocation factor, we set $\frac{d\xi_q}{d\rho}$ equal to zero. After some algebraic manipulation, this gives the optimal allocation factor:

$$\rho^{**} = \frac{\log_2\left[\frac{24\eta_g\alpha_d}{(72\eta_g+24\eta_t)\alpha_c}\right]\frac{24\eta_g p}{P_T+16\eta_g p} + 2}{-\log_2\left[\frac{24\eta_g\alpha_d}{(72\eta_g+24\eta_t)\alpha_c}\right]\frac{(48\eta_g+24\eta_t)p}{P_T+16\eta_g p} + 4}, \qquad (28)$$

which gives the corresponding minimum MSE increase:

$$\min_{\rho} \xi_q \quad = \quad \alpha_c\, 2^{-2(1-\rho^{**})B_T} + \alpha_d\, 2^{-2\rho^{**}B_T}$$

where B_T is given in (17).

Observe that the optimal bit allocation factor ρ^{**} converges to the standard allocation $1/2$ as the total power constraint P_T is relaxed to infinity. This is the regime where the standard allocation is optimal: allocate an equal number of bits to data as to filter coefficients. As register length decreases or convergence speed increases, the standard allocation becomes suboptimal.

6. POWER VS. MEAN SQUARE ERROR

Figure 2 shows the P_T-constrained optimal data bit allocation factor ρ^{**} as a function of P_T superimposed on a plot of the increase in MSE due to quantization. The power coefficients used are $\eta_g = 1$ milliwatt and $\eta_t = 10$ milliwatts. The vector length is $p = 2$. Note that MSE does not degrade significantly until P_T falls below approximately 1.2 Watts (normalized). At this breakdown point the optimal data bit allocation factor is approximately $\rho^* = 0.25$. We can use relation (17) with $\rho = \rho^{**}$ to find the corresponding B_T as a function of P_T. We find that $P_T = 1.2$ corresponds to $B_T \approx 6$, but the optimal ρ^{**} tells us to allocate only 1 bit plus sign to the data and 5 bits plus sign to the filter coefficients. This reduction is because the coefficient operations α_c dominate the data factor α_d in the relation (13).

7. CONCLUSION

We have derived expressions for the optimal bit allocation for adaptive LMS algorithms under a total power constraint. This expression can easily be specialized to a specific hardware implementation for computation of the number of bits to allocate to data and filter coefficients. A general conclusion is that the standard design strategy of allocating an equal number of bits to the data and filter coefficients is optimal only as the power gets very large. For typical LMS implementations, it is optimal to allocate more bits to the coefficients than to the data. In particular, we have found that for an LMS algorithm implemented with fixed point arithmetic and using table-lookup multiplication, it is possible to reduce the total number of bits to 2 data bits and 6 coefficient bits without significant increase in steady-state MSE.

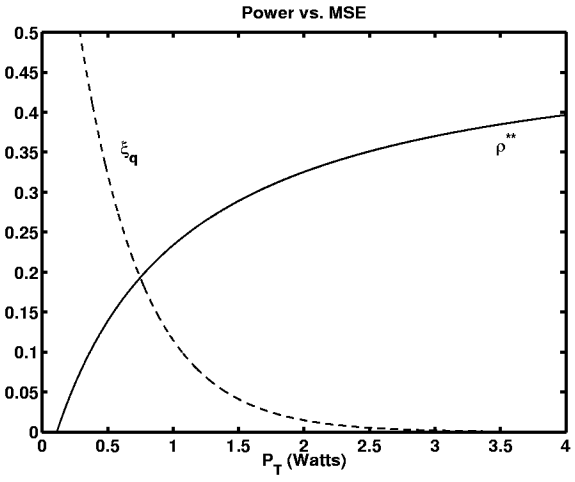

Figure 2: *Optimal data bit allocation factor under P_T constraint and total MSE as function of P_T.*

8. REFERENCES

[1] A. O. Hero and R. Gupta, "Optimal bit allocation for the quantized lms adaptive algorithm," in *Proc. of the IEEE Workshop on Statistical Signal and Array Processing*, Portland, OR, Sept. 1998.

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

[3] J. R. Treichler, C. R. Johnson, and M. G. Larimore, *Theory and Design of Adaptive Filters*, Wiley, New York, 1987.

[4] 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.