# Power vs. Performance Tradeoffs for Reduced Resolution LMS Adaptive Filters

Riten Gupta and Alfred O. Hero
EECS Department
University of Michigan

**Abstract**

Low power implementation of adaptive filters for channel equalization and interference cancelling is important for wireless communications transceivers. In this paper we obtain asymptotic expressions for the increase in mean square error of the widely used LMS adaptive filter when the total power dissipated is decreased by reducing the number of data and filter coefficient bits used in the filter. We also obtain expressions for the power-optimal bit-allocation factor which determines the optimal proportion of bits to allocate to the data, the remainder being allocated to the coefficients. Numerical studies are presented for an exponential memory ISI channel and 4-ary PSK signaling.

**Keywords**

Roundoff and Quantization Analysis, Low-Power Channel Equalization, Optimal Bit Allocation

EDICS numbers: 2.6.4, 2.8.1

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, (734)-764-4110 (voice), (734)-763-8041 (fax). 1301 Beal Ave. #4301, Ann Arbor, MI 48109-2122

# List of Figures

## I. Introduction

In a battery powered receiver, the adaptive equalization function consumes a significant portion of total battery power. For example the SINCGARS combat radio used by the US Army consumes on the average 7 Watts in receive mode of which more than 1 Watt (14 % of total power) is consumed by the channel equalizer [1]. Therefore the channel equalization function is a prime target for power reduction in these handsets. There have been many digital hardware design strategies proposed for power reduction including: reduction of supply voltage, reduction of clock speed and data rate, parallelization and pipelining of operations, using sign-magnitude arithmetic, and differential encoding of data [2], [3]. Another technique, which is the springboard for this paper, is the reduction of the number of bits used to represent the data and control variables in the digital circuit. [4] This bit width reduction strategy is very highly leveraged since it reduces the power dissipation everywhere in the data and control flow paths. This strategy is also very versatile since it can be applied to any hardware architecture and can be easily adjusted in real time by dynamically switching-off bus lines and register bits. However, bit width reduction generally entails a degradation in algorithm performance, as measured by adaptive algorithm convergence rate, steady state mean square error (MSE), and subsequent probability of symbol error. This paper provides an analysis of MSE degradation versus power reduction for a widespread class of adaptive algorithms popularly known as the LMS algorithm.

The LMS algorithm was introduced by Widrow [5] and is one of the most common adaptation algorithms used in practical systems such as adaptive channel equalizers [6], adaptive beamforming antenna arrays [7], and adaptive interference cancellation systems [8]. In channel equalization, the LMS algorithm adapts the filter coefficients of an FIR filter driven by the received data based on an error signal formed by subtracting a training signal from the filtered output. In [9] it was shown that uniform scalar quantization of the error signal is close to optimal for the case of Gaussian inputs. We consider a fully 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 paths, where the quantizers can have different resolutions. To obtain significant power reduction, it may be beneficial to apply different quantizer resolutions during the transient acquisition phase and the steady state tracking phase of the algorithm. In this paper we focus on resolution reduction in the steady state tracking phase.

We first present an asymptotic formula for the increase in steady state mean square error (MSE) due to quantization which generalizes the formulas of Caraiscos and Liu [10] to the case of complex data and coefficients. We then derive a pair of optimal bit-allocation factors which minimize the increase in MSE subject to: 1) a total bit-width constraint (bit budget); and 2) a total power consumption constraint (power budget). Finally we show that QLMS with optimal bit-allocation consumes significantly less power than LMS at little expense in performance. While these results are directly applicable to the generic LMS algorithm in a variety of applications, for concreteness, we concentrate on the case of channel equalization with training in this paper. Numerical examples are given for an IIR (exponential memory) channel.

## II. Register Length and Power

It is well known that the power consumed by the operation of loading successive time samples of a random real-valued sequence into a binary $B$-bit register is proportional to the average number of bit flips per unit time in the register [11]. While for a uniform white sequence the average number of bit flips is $B/2$, in general this average can be much less than $B/2$ for a correlated sequence. This is because for correlated sequences most significant bits (MSB's) have lower probability of transitioning than least significant bits (LSB's). Several models for the power consumption of register loading have been proposed [11]. We propose to use the following simple upper bound (derived in Appendix A) on power consumption for a $B$-bit register, and a zero mean wide-sense stationary Gaussian random sequence:

$$P_B \quad \leq \quad B\eta \cdot \mathrm{erfc}\left(\left[2^B\sqrt{R(0)-R(1)}\right]^{-1}\right) \tag{1}$$

where $\eta$ is the power dissipation-per-bit, which depends on factors such as switching load capacitance and supply voltage, and $R(\tau)$ is the autocorrelation function of the random sequence. Note that it the bound (1) is conservative: if we constrain the right hand side of (1) to some maximum tolerable power dissipation, $P_{max}$ say, then a circuit design which uses register bit width $B = B_{max}$ which solves the equation $P_{max} = B\eta \cdot \mathrm{erfc}\left(\left[2^B\sqrt{R(0)-R(1)}\right]^{-1}\right)$ is guaranteed to consume less power than $P_{max}$.

A plot of $P_{max}$ versus $B$ is given in Figure 1 for an AR(1) sequence with real pole located at $a_1$. Note that as the pole approaches the unit circle the $P_{max}$ curve displays an abrupt threshold occurring at an increasingly high bit-width. Since $\mathrm{erfc}(x) \approx 0$ for $x \geq 1$, the threshold bit-width can be specified by the formula

$$\left[2^{B_{thresh}}\sqrt{R(0)-R(1)}\right]^{-1} = 1$$

which implies

$$B_{thresh} = -\frac{1}{2}\log_2\left(R(0)-R(1)\right).$$

For $|a_1| < 0.8$ the threshold occurs below one bit so in this range of $a_1$ the power dissipation increases almost linearly as a function of $B$. For more highly correlated sequences, the threshold can occur much higher than one bit and thus more bits may be used without significantly increasing power consumption. Conversely, fewer bits are necessary. This increased threshold is due to the fact that only the LSB's flip with high probability in two successive samples of a highly correlated sequence, while for an uncorrelated sequence, all bits have equal probability (0.5) of flipping.

## III. Quantized Adaptive Channel Equalization

Figure 2 shows the block diagram of an adaptive equalization system with two different quantizers, denoted $Q_d$ and $Q_c$, applied to the complex data and to the complex filter coefficients of an adaptive $p$-

tap FIR filter. Here $s_k$ is a complex (baseband) training signal, $y_k$ is a scaled version of $s_k$ $x_k$ is the complex received (baseband) data, and $\hat{y}_k$ is a linear estimate of $y_k$ given the $p$ samples $\underline{x}_k = [x_{k-1}, \ldots, x_{k-p}]^T$ and filter coefficients $\underline{w}_k = [w_{1k}, \ldots, w_{pk}]^H$. The scaling factor $a$ represents a right shift of one or more bits and is used to prevent overflow of the quantized weight vector.

The quantizers $Q_d$ and $Q_c$ are allocated $B_d$ bits plus sign and $B_c$ bits plus sign, respectively, to the real and imaginary parts of their inputs. We assume that the input sequences have been scaled to lie between $-1$ and $+1$. As in [10] we assume that the effect of the operators $Q_d$ and $Q_c$ is to add to the input a complex white noise of variance $2\sigma_d^2 = \frac{1}{6}2^{-2B_d}$ for $Q_d$ and $2\sigma_c^2 = \frac{1}{6}2^{-2B_d}$ for $Q_c$. The standard LMS algorithm adapts the filter coefficients in an attempt to minimize the quadratic surface specified by the mean square error: $E[|y_k - \hat{y}_k|^2]$ [12], [5]. The quantized LMS algorithm, which we call QLMS, implements a recursive filter coefficient update of the form

$$\underline{w}'_{k+1} = \underline{w}'_k + \mu_c Q_c \left(\mu_d \ \underline{x}'_k e'^*_k\right) \tag{2}$$

where

$$e'_k = y'_k - Q_d \left(\underline{w}'^H_k \underline{x}'_k\right). \tag{3}$$

is the quantized error signal. Here $\mu_c$ and $\mu_d$ are gain parameters which control the convergence properties of the algorithm. Note that the standard complex LMS algorithm uses only one gain parameter $\mu$ which is simply the product $\mu = \mu_d \mu_c$. We have used primed symbols to represent quantized values. That is, $\underline{x}'_k = Q_d(\underline{x}_k)$, $y'_k = Q_d(y_k)$, and $\underline{w}'_k = Q_c(\underline{w}_k)$ is the quantized weight vector used in the algorithm. Also, $^*$ denotes complex conjugation.

The total power per iteration of the quantized LMS algorithm is determined by power dissipation of shift, add, multiply, memory load, and memory store operations. This depends on the specific circuit implementation of the FIR filter and control circuitry. In Appendix D, we derive the following formula for power-per-iteration of LMS under the assumption that $\mu = 2^{-q}$ for $q \in \{1, \ldots, B_c + 1\}$:

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

This expression is linear in the number of bits $B_d$ and $B_c$ and assumes fixed point complex arithmetic, cyclic updating of the data stack $[x_{k-1}, \ldots, x_{k-p}]^T$ by overwriting $x_{k-p}$ with $x_k$, multiplication using table lookup as opposed to adding partial products. In (4), we have defined generic power coefficients $\eta_g$ and $\eta_t$ representing power consumption per add and power consumption per table lookup operation, respectively.

### *A. Performance of LMS Algorithm*

The performance of the LMS adaptive algorithm is typically characterized by two quantities: the speed of convergence and the excess MSE [5], [6], [13]. We assume that $x_k$, $y_k$ and $s_k$ are all wide sense stationary random sequences. For a moderate number of filter coefficient quantization bits $B_c$, experiments have shown that the convergence of QLMS in the mean is primarily determined by the quantized data $Q_d(x_k)$. On the other hand, the excess MSE is affected by both the number of data quantization bits $B_d$ and the number of filter coefficient quantization bits $B_c$. Similarly to [10], we make the following high resolution assumptions:

> The quantization error of $Q_d$ and $Q_c$ quantizers is zero mean, white, with variances $2\sigma_d^2$ and $2\sigma_c^2$. Furthermore, these errors are independent of the quantizer inputs. The process $x_k$ is circular Gaussian. The step size $\mu$ is greater than 0, and $B_c$, $B_d \geq 1$. Finally, we assume that the sequence $s_k$ has been properly scaled to lie within the unit circle. (5)

These assumptions are fairly restrictive but enable us to obtain useful closed form expressions.

### A.1 Mean Convergence

Define the $p \times p$ covariance matrix $R_{x'} = E[\underline{x}'_k \underline{x}'^H_k]$ of the quantized data and let this matrix have real non-negative eigenvalues $\{\lambda'_i\}_{i=1}^p$. Further define the cross correlation vector $R_{x'y'} = E[\underline{x}'_k y'^*_k]$. Assume the gain parameter $\mu = \mu_d \mu_c$ satisfies the condition

$$0 < |1 - \mu\lambda'_i| < 1, \qquad i = 1, \ldots, p. \tag{6}$$

In Appendix B we show that the filter coefficients of the QLMS algorithm converge in the mean to a set of optimal weights, $\underline{w}'^o$.

$$\lim_{k \to \infty} E[\underline{w}'_k] = \underline{w}'^o \tag{7}$$

where

$$\underline{w}'^o = R_{x'}^{-1} R_{x'y'}. \tag{8}$$

When the QLMS algorithm converges, the MSE converges as a decaying exponential with the $1/e$ time constant of the slowest mode equal to $\tau_{3dB} = 1/(-\max_i \ln(|1-\mu\lambda'_i|)$, called the adaptation time constant. Note that the speed of convergence generally increases as $\mu$ increases.

### A.2 Steady State Weight Error Covariance

In Appendix C we derive an expression for the steady state quantized weight error covariance for any values of $\mu_d$ and $\mu_c$. Here we assume $\mu = \mu_d \ll 1$ and $\mu_c = 1$, as is the case for the standard complex LMS algorithm. Let $R = E[\underline{x}_k \underline{x}_k^H]$ and let $P_k = E[(\underline{w}'_k - \underline{w}_k)(\underline{w}'_k - \underline{w}_k)^H]$ where $\underline{w}_k$ is the weight vector at time $k$ of the infinite precision LMS algorithm. Then assuming $P_k$ converges as $k \to \infty$, it converges to the steady state covariance matrix

$$P_k = P = p\sigma_c^2(2I - \mu R)^{-1} + \frac{\sigma_c^2}{\mu}R^{-1}. \tag{9}$$

A.3 Excess Mean Square Error

Under the assumptions (5), and under the same assumptions on $\mu_d$, and $\mu_c$ stated in Section III-A.2, the following expression for the steady state mean square error is derived in Appendix C.

$$\xi = E[|s_k - \hat{s}'_k|^2] = \xi_{min} + \xi_{excess} + \xi_q$$

where

$$\xi_{min} = \frac{1}{a^2}\left(\sigma_y^2 - R_{xy}^H R_x^{-1} R_{xy}\right) \tag{10}$$

is the optimal mean-square error of the Weiner filter for the known channel,

$$\xi_{excess} = \frac{1}{2}\mu \mathrm{tr}(R_x)\xi_{min} \tag{11}$$

is the excess MSE due to misadjustment, and

$$\xi_q = \alpha_c \, 2^{-2B_c} + \alpha_d \, 2^{-2B_d} \tag{12}$$

where

$$\alpha_c = \frac{p}{12\mu a^2}, \qquad \alpha_d = \frac{\|\underline{w}^o\|^2 + p}{6a^2}. \tag{13}$$

and $\underline{w}^o = R_x^{-1}R_{xy}$ is the optimal Wiener weight vector for the standard infinite precision complex LMS algorithm.

The term $\xi_q$ is the excess MSE due to quantization of the data and filter coefficients. The first term in the expression (12) is the excess MSE due only to quantization of the filter coefficients while the second term represents the MSE due to quantization of the data. Note that for small $\mu$, the term $\alpha_c$ dominates the excess MSE due to quantization unless $B_c$ is made large. This implies that for small step sizes, a high resolution is required for the filter coefficients. Also worth noting is that $\xi_q$ increases in $p$ at a linear rate, decreases in $\mu$ 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, $\xi_q$, can be plotted as a function of $B_d$ and $B_c$. A plot of the increase in MSE is given in Figs. 3 and 4 for the case of white $s_k$ and for $x_k$ generated by passing $s_k$ through a single pole IIR filter ($p = 2$) with pole at $a_1 = 0.8$ and with LMS gain parameter $\mu = 1/4$.

Observe from the surface plot that the increase in MSE is fairly flat over the range $B_c, B_d > 3$ but undergoes a rapid increase as either $B_c$ or $B_d$ decreases below this range. The objective of the following section is to specify optimal choices of $B_d$ and $B_c$ which minimize this increase under various constraints on $B_d$ and $B_c$.

## IV. Optimal Bit Allocation Strategies

We present expressions for the optimum allocation of bits to data versus filter coefficients under two constraints: total number of bits and total power consumption.

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

$$B_d = \rho B_T, \qquad\qquad B_c = (1-\rho)B_T. \tag{14}$$

### *A. Total Bit Budget Constraint*

Under a constraint on $B_T$ the objective becomes to minimize the increase $\xi_q$ in MSE with respect to $\rho$. Graphically, this is the same as minimizing $\xi_q$ along the diagonal line $B_T = B_d + B_c$ of slope $-1$ in the $B_c$, $B_d$ plane shown on Fig. 4. It is shown in the appendix that $\xi_q$ is convex as a function of $\rho$ with a single minimum occurring at the point $\rho = \rho^*$

$$\begin{aligned}\rho^* &= \frac{1}{4B_T}\log_2\left(\frac{\alpha_d}{\alpha_c}\right) + \frac{1}{2}\\ &= \frac{1}{4B_T}\log_2\left(2\mu\frac{\|\underline{w}^o\|^2 + p}{p}\right) + \frac{1}{2},\end{aligned} \tag{15}$$

and the corresponding minimum value of $\xi_q$ is

$$\begin{aligned}\min_\rho \xi_q &= 2^{-B_T+1}\sqrt{\alpha_d\alpha_c}\\ &= 2^{-B_T-1}\frac{1}{3}\sqrt{\frac{2p}{\mu}(\|\underline{w}^o\|^2 + p)}.\end{aligned} \tag{16}$$

Equation (15) only holds when $\rho \in [0,1]$, i.e.

$$2^{-2B_T} \le \frac{\alpha_d}{\alpha_c} \le 2^{2B_T}.$$

Observe that the optimal bit allocation factor $\rho^*$ converges to $1/2$ as the combined register length $B_T$ goes to infinity. This is the regime where the standard textbook 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. Note that if $\alpha_d = \alpha_c$ then $\rho^* = 1/2$, implying again that an equal number of bits should be allocated to both data and filter coefficients ($B_d = B_c$). In this case, equation (12) indicates that the data quantization and coefficient quantization contribute equally to the increase in MSE.

In typical implementations, e.g. where AGC is implemented to prescale the data to unit variance, the gain parameter is chosen such that $\mu \ll 1$ to ensure convergence. In particular, if $\mu < 1/4$ and $\|\underline{w}^o\|^2 \leq p$ then $\rho^*$ is less than $1/2$ and more bits should be allocated to the filter coefficients than to the data.

### *B. Total Power Budget Constraint*

Under a constraint on total power budget, $P_T$, we can use (4) and (14) to re-express the total combined number of bits $B_T$ as a function of $\rho$ and $P_T$

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

In the appendix we show that for $B_T \geq 2$, $\xi_q$ is once again a convex function of $\rho$ with unique minimum at $\rho = \rho^{**}$

$$\rho^{**} = \frac{\log_2\left[\frac{24\eta_g\alpha_d}{(72\eta_g+24\eta_t)\alpha_c}\right]\frac{24\eta_g p}{P_T+16\eta_g p} + 2}{-\log_2\left[\frac{24\eta_g\alpha_d}{(72\eta_g+24\eta_t)\alpha_c}\right]\frac{(48\eta_g+24\eta_t)p}{P_T+16\eta_g p} + 4}, \tag{18}$$

which gives the corresponding minimum MSE

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

where $B_T$ is given in (17).

Observe that the optimal bit allocation factor $\rho^{**}$ converges to the standard allocation $1/2$ as the total power constraint $P_T$ is relaxed to infinity. As $P_T$ decreases, the standard allocation becomes suboptimal.

## V. Numerical Example: 4-PSK

Here we briefly consider the case of a Gaussian noise IIR channel with a single pole at $a_1 = 0.8$, a 4-PSK signal $s_k$ of unit variance, noiseless training sequence $y_k = as_k$, and a 2-tap LMS filter with gain coefficient $\mu = 1/4$ and scaling factor $a = 1/4$. This corresponds to a rather severe exponential memory channel with intersymbol interference (ISI) extending over 5 to 10 data samples.

Figure 5 shows the $B_T$-constrained optimal data bit allocation factor $\rho^*$ as a function of $B_T$ superimposed on a plot of the resultant optimal MSE, $\xi$. Note that for large $p$, the formula (12) for $\xi_q$ is independent of any channel effects (i.e. independent of $R_x$ and $R_{xy}$). Also note that $\rho^* \approx 1/2$.

Figure 6 shows the $P_T$-constrained optimal data bit allocation factor $\rho^{**}$ as a function of $P_T$ superimposed on a plot of MSE. The power coefficients used are $\eta_g = 1$ milliwatt and $\eta_t = 100$ milliwatts. Relation (17) with $\rho = 1/2, \rho^*, \rho^{**}$,is plotted in Fig. 7. This plot indicates that with the power-optimal bit allocation factor, $\rho^{**}$, for a design constraint of $P_T \leq 20$, $B_T = 11$ bits (13 total bits) can be used while only $B_T = 8$ bits (10 bits total) can be used with the standard bit allocation of $\rho = 1/2$ or with $\rho = \rho^*$. Figure 6 indicates that under this $P_T$ constraint ($P_T \leq 20$), with $\rho = \rho^{**}$, a steady-state MSE of 0.035 can be achieved. However, Fig. 5 indicates that under the same constraint, with $\rho = \rho^* \approx 1/2$, the steady-state MSE is 0.076. Thus the steady-state MSE has been reduced by a factor of more than two by using the $P_T$-optimal allocation factor under this power constraint.

Finally, in Fig. 8 we plot the MSE as a function of $P_T$ for both the $P_T$-optimal and the $B_T$-optimal allocation factors $\rho^{**}$ and $\rho^*$ as well as for the standard allocation 1/2. Note that, as expected, for a given power constraint the $P_T$-optimal bit allocation yields lower MSE than the $B_T$-optimal bit allocation and the standard allocation, $\rho = 1/2$. We also see that operating with power consumption $P_T \approx 20$, the MSE can be reduced by more than a factor of two by using the power-optimal bit allocation.

## VI. Conclusion

We have derived expressions for optimal bit allocation for adaptive LMS algorithms under combined register length constraints and total power constraints. These expressions 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 or register length constraints get very large. For typical LMS implementations, it is optimal to allocate more bits to the coefficients than to the data.

## Appendix A
## Derivation of (1)

Consider a zero-mean wide-sense stationary Gaussian random sequence $x_k$, whose values are scaled to within $-1$ and $+1$, quantized to $B$ bits and loaded into a $B$-bit register. Let $n_k$ denote the number of bits that flip when loading the value $x_{k+1}$ into the register. Then

$$E[n_k] = \sum_{i=1}^{B} i \; P(n=i) = \sum_{i=0}^{B} P(n>i).$$

Next we note that

$$\sum_{i=1}^{B} P(n=i) \leq \sum_{i=0}^{B} P(n>i) \leq \sum_{i=1}^{B} B \cdot P(n=i).$$

Therefore,

$$P(n>0) \leq E[n_k] \leq B \cdot P(n>0). \tag{A-1}$$

Now $P(n>0)$ is equal to the probability that at least one bit flips when loading $x_{k+1}$ into the register. This is equal to one minus the probability that no bits flip. Therefore

$$P(n>0) = 1 - P(Q(x_k) = Q(x_{k+1})) \tag{A-2}$$

where $Q(x)$ is $x$ quantized to $B$ bits. Now, the granularity of the quantizer is $2^{-B+1}$. Therefore, (A-2) becomes

$$P(n>0) = 1 - P(|x_k - x_{k+1}| < 2^{-B+1}).$$

As, $x_k - x_{k+1}$ is Gaussian with mean zero and variance $2R(0) - 2R(1)$ where $R(\tau) = E[x_k x_{k+\tau}]$, we have

$$P(n > 0) = 1 - \sqrt{\frac{2}{\pi}} \int_0^{\frac{2^{-B+1}}{\sqrt{2R(0)-2R(1)}}} e^{-\frac{t^2}{2}} dt,$$

or equivalently

$$P(n > 0) = \text{erfc}\left( \left[ 2^B \sqrt{R(0) - R(1)} \right]^{-1} \right).$$

Next with $\eta$ the power consumption per bit flip, we have: $P_B = \eta E[n_k]$, where $P_B$ is the average power consumed during each register load. Using this and (A-1), we obtain the following two bounds on $P_B$:

$$\begin{aligned} P_B &\leq B\eta \cdot \text{erfc}\left( \left[ 2^B \sqrt{R(0) - R(1)} \right]^{-1} \right) \\ P_B &\geq \eta \cdot \text{erfc}\left( \left[ 2^B \sqrt{R(0) - R(1)} \right]^{-1} \right). \end{aligned}$$

## Appendix B
## Derivation of (7)

Consider the following mean square error cost function:

$$\xi' = E[|e'_k|^2]. \tag{A-3}$$

where $e'_k$ is given in (3). Using the definition of $Q_d$ and $e'_k$

$$e'_k = y'_k - \underline{w}'^H_k \underline{x}'_k + n^*_k$$

where $n^*_k = \underline{w}'^H_k \underline{x}'_k - Q_d(\underline{w}'^H_k \underline{x}'_k)$. As in [10] we approximate $n_k$ as zero-mean, white, with variance $2\sigma_d^2$ and uncorrelated with $\underline{w}'^H_k \underline{x}'_k$ and $y'_k$. Therefore, we have

$$\xi' = E[|e'_k|^2] = E[|y'_k - \underline{w}'^H_k \underline{x}'_k|^2] + 2\sigma_d^2.$$

Taking the gradient of this cost function with respect to $\underline{w}'_k$ we get

$$\frac{d\xi'}{d\underline{w}'_k} = 2R_{x'}\underline{w}'_k - 2R_{x'y'}.$$

Setting this to zero and solving for $\underline{w}'_k$ we obtain the optimal weight vector (8).

Now, from (2) and (3), we see that the QLMS update is of the form

$$\underline{w}'_{k+1} = \underline{w}'_k + \mu \underline{x}'_k (y'^*_k - \underline{x}'^H_k \underline{w}'_k + n_k) + \underline{\sigma}_k$$

where $n_k$ is defined above and $\underline{\sigma}_k$ is a zero-mean complex white noise vector whose components have variance $2\sigma_c^2$. Using the method of [13] we obtain (7) and (6). First define

$$\delta \underline{w}'_{k+1} = \underline{w}'_{k+1} - \underline{w}'_k = \mu \underline{x}'_k e'^*_k + \underline{\sigma}_k,$$

and

$$\epsilon'_k = y'_k - \underline{w}'^{oH}\underline{x}'_k, \qquad \underline{\tilde{w}}'_k = \underline{w}'_k - \underline{w}'^{o}.$$

Then we have

$$\begin{aligned} e'_k &= \epsilon'_k + \underline{w}'^{oH}\underline{x}'_k - \underline{w}'^{H}_k\underline{x}'_k + n^*_k \\ &= \epsilon'_k - \underline{\tilde{w}}'^{H}_k\underline{x}'_k + n^*_k. \end{aligned}$$

Next use, $\underline{\tilde{w}}'_{k+1} - \underline{\tilde{w}}'_k = \delta\underline{w}'_{k+1}$ to obtain

$$\underline{\tilde{w}}'_{k+1} = (I - \mu\underline{x}'_k\underline{x}'^{H}_k)\underline{\tilde{w}}'_k + \mu\underline{x}'_k\epsilon'^*_k + \mu\underline{x}'_k n_k + \underline{\sigma}_k.$$

Similarly to [13], we assume that $\underline{x}'_k$ and $\epsilon'_k$ are independent. Using this and the fact that $n_k$ and $\underline{\sigma}_k$ are zero-mean, we see that

$$E[\underline{\tilde{w}}'_{k+1}|\underline{x}'_1,\ldots,\underline{x}'_k,\epsilon'_1,\ldots,\epsilon'_k] = \underline{\tilde{w}}'_k - \mu R_{x'}\underline{\tilde{w}}'_k$$

and taking expectations,

$$E[\underline{\tilde{w}}'_{k+1}] = (I - \mu R_{x'})E[\underline{\tilde{w}}'_k]$$

which is completely analogous to the standard unquantized LMS algorithm. Using the same analysis as in [13] (Sec. 5.2), we conclude that $E[\underline{w}'_k] \to \underline{w}'^{o}$ as long as condition (6) is met.

## Appendix C
## Derivation of (12) and (9)

The weight update formula for the unquantized LMS algorithm is

$$\underline{w}_{k+1} = \underline{w}_k + \mu\ \underline{x}_k\left(y^*_k - \underline{x}^H_k\underline{w}_k\right) \tag{A-4}$$

where $\mu = \mu_d\mu_c$. We define

$$\sigma^2_d = \frac{1}{12}2^{-2B_d}, \qquad \sigma^2_c = \frac{1}{12}2^{-2B_c}. \tag{A-5}$$

These are the variances of the quantization noises associated with quantizers having $B_d$ and $B_c$ bits (plus sign), respectively. We use primed symbols to represent quantized values and define the following quantities

$$\begin{aligned} \underline{x}'_k &= \underline{x}_k + \underline{\alpha}_k \\ y'_k &= y_k + \beta_k \\ \underline{w}'_k &= \underline{w}_k + \underline{\rho}_k. \end{aligned} \tag{A-6}$$

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

Thus we have the representation for the quantized value of the filter output

$$\hat{y}'_k = Q_d(\underline{w}'^H_k \underline{x}'_k) = \underline{w}_k^H \underline{x}_k + \underline{\rho}_k^H \underline{x}_k + \underline{w}_k^H \underline{\alpha}_k + \eta_k \tag{A-7}$$

where $\eta_k$ 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 $\underline{w}'^H_k \underline{x}'_k$ and then summing them [10]. As in [10], we have ignored the noise product term $\underline{\rho}_k^H \underline{\alpha}_k$ in (A-7), as its power is of order $\sigma_d^2 \sigma_c^2$.

The total error is now

$$\begin{aligned} s_k - \hat{s}'_k &= \frac{1}{a}(y_k - \underline{w}_k^H \underline{x}_k) - \\ &\quad \frac{1}{a}(\underline{\rho}_k^H \underline{x}_k + \underline{w}_k^H \underline{\alpha}_k + \eta_k) \end{aligned} \tag{A-8}$$

The first term on the right hand side of (A-8) is the error of the unquantized algorithm. The second term is the error due to quantization and will be denoted $e_q$. Under the hypothesis (5), these terms are uncorrelated and the MSE for the quantized LMS algorithm is

$$\begin{aligned} \xi &= \frac{1}{a^2} E[|y_k - \underline{w}_k^H \underline{x}_k|^2] + \\ &\quad \frac{1}{a^2} E[|\underline{\rho}_k^H \underline{x}_k + \underline{w}_k^H \underline{\alpha}_k + \eta_k|^2]. \end{aligned} \tag{A-9}$$

The first term on the right hand side of (A-9) is the MSE of the standard unquantized LMS algorithm and is equal to $\xi_{min} + \xi_{excess}$ given in (10) and (11) [5], [13]. The second term is the excess MSE due to quantization,

$$\xi_q = E[|e_q|^2].$$

Under the assumptions (5), $\underline{\alpha}_k$, $\beta_k$, $\underline{\rho}_k$, and $\eta_k$ are all uncorrelated and

$$\xi_q = \frac{1}{a^2}\left( E\left[|\underline{\rho}_k^H \underline{x}_k|^2\right] + E\left[|\underline{w}_k^H \underline{\alpha}_k|^2\right] + E\left[|\eta_k|^2\right]\right). \tag{A-10}$$

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

$$E[|\underline{w}_k^H \underline{\alpha}_k|^2] = 2\sigma_d^2\ E[||\underline{w}_k||^2].$$

Note that this differs from the real case studied in [10] by a factor of 2. In the steady state, this becomes

$$E[|\underline{w}_k^H \underline{\alpha}_k|^2] = 2\sigma_d^2\ \left( ||\underline{w}^o||^2 + \frac{1}{2} p\ \mu\ \xi_{min} \right). \tag{A-11}$$

Finally, for the first term in (A-10)

$$E[|\underline{\rho}_k^H \underline{x}_k|^2] = \mathrm{tr}\Big( E\left[\underline{\rho}_k \underline{\rho}_k^H\right] R_x \Big). \tag{A-12}$$

It remains to find an expression for $E[\underline{\rho}_k \underline{\rho}_k^H]$. Let $e_k = y_k - \underline{w}_k^H \underline{x}_k$ and define $\zeta_k = \beta_k - \eta_k$. Note that the variance of $\zeta_k$ is $2(1+p)\sigma_d^2$. Then we have

$$e'_k = y'_k - \hat{y}'_k = e_k - \underline{\rho}_k^H \underline{x}_k - \underline{w}_k^H \underline{\alpha}_k + \zeta_k$$

The quantized LMS update equations (2), (3) can now be written

$$\begin{aligned}
\underline{w}'_{k+1} &= \underline{w}'_k + \mu_c Q_c \left[\mu_d \ \underline{x}'_k \left(y'_k - Q_d[\underline{w}'^H_k \underline{x}'_k]\right)^*\right] \\
&= \underline{w}'_k + \mu_c Q_c[\mu_d \ \underline{x}'_k e'^*_k] \\
&= \underline{w}'_k + \mu(\underline{x}_k e^*_k - \underline{x}_k \underline{x}_k^H \underline{\rho}_k - \\
&\quad \underline{x}_k \underline{\alpha}_k^H \underline{w}_k + \underline{x}_k \zeta^*_k + \underline{\alpha}_k e^*_k) + \mu_c \underline{\sigma}_k
\end{aligned}$$

where $\underline{\sigma}_k$ is a quantization noise of variance $2\sigma_c^2$ and all second-order noise terms, e.g. $\underline{\rho}_k^H \underline{\alpha}_h$ have been ignored.

Using (A-4) and (A-6), $\underline{\rho}_k$ can be seen to obey the recursion

$$\underline{\rho}_{k+1} = F_k \underline{\rho}_k + \underline{b}_k \tag{A-13}$$

where

$$F_k = I - \mu \underline{x}_k \underline{x}_k^H$$

and

$$\underline{b}_k = \mu(-\underline{x}_k \underline{\alpha}_k^H \underline{w}_k + \underline{x}_k \zeta^*_k + \underline{\alpha}_k e^*_k) + \mu_c \underline{\sigma}_k \tag{A-14}$$

Next define $P_k = E[\underline{\rho}_k \underline{\rho}_k^H]$ and $Q_k = E[\underline{b}_k \underline{b}_k^H]$. Then from (A-13) and the assumptions (5)

$$P_{k+1} = E[F_k P_k F_k] + Q_k. \tag{A-15}$$

Using (A-14), we get

$$\begin{aligned}
Q_k &= 2\mu^2 \left(\sigma_d^2 \ E[\|\underline{w}_k\|^2] R_x + (1+p)\sigma_d^2 \ R_x + \right. \\
&\quad \left. E[|e_k|^2]\sigma_d^2 I\right) + 2\mu_c^2 \sigma_c^2 I
\end{aligned}$$

In the steady state, this becomes

$$\begin{aligned}
Q_k &= 2\mu^2 \left(\sigma_d^2 \ R_x \|\underline{w}^o\|^2 + (1+p)\sigma_d^2 \ R_x + \right. \\
&\quad \left. \xi_{min} \sigma_d^2 I\right) + 2\mu_c^2 \sigma_c^2 I
\end{aligned} \tag{A-16}$$

Dropping the time subscripts and writing $R_x$ as $R$, the $i,j$th element of $E[FPF]$ is

$$\begin{aligned}
E[FPF]_{ij} &= E\left[\sum_k \sum_l (\delta_{ik} - \mu x_i x^*_k)(\delta_{lj} - \mu x_l x^*_j) p_{kl}\right] \\
&= p_{ij} - \mu(PR)_{ij} - \mu(RP)_{ij} + \\
&\quad \mu^2 \sum_k \sum_l E[x_i x^*_j x^*_k x_l] p_{kl}
\end{aligned}$$

Now under the assumption that $\underline{x}_k$ is a circular Gaussian random vector [14],

$$E[x_i x_j^* x_k^* x_l] = R_{ij}R_{lk} + R_{ik}R_{lj}.$$

Therefore,

$$\begin{aligned} E[FPF]_{ij} &= P_{ij} - \mu(PR)_{ij} - \mu(RP)_{ij} + \\ &\quad \mu^2 R_{ij}\mathrm{tr}(RP) + \mu^2 (RPR)_{ij}. \end{aligned} \tag{A-17}$$

While in [10], the last term is neglected, here we take it into account. From (A-15) and (A-17) we obtain the recursion

$$\begin{aligned} P_{k+1} &= P_k - \mu(P_k R + RP_k) + \mu^2 R\ \mathrm{tr}(RP_k) \\ &\quad + \mu^2 RP_k R + Q_k \end{aligned}$$

In the steady state, $P_{k+1} = P_k$ and this becomes

$$\mu(P_k R + RP_k) = \mu^2 R\ \mathrm{tr}(RP_k) + \mu^2 RP_k R + Q_k. \tag{A-18}$$

As a candidate solution, assume that the matrix $P = P_k$ has the same eigenvectors as $R$. We will verify that this is a valid assumption below. Then $RP = PR$ and

$$P = (2\mu I - \mu^2 R)^{-1}(\mu^2 \mathrm{tr}(RP)I + 2\mu^2 qI + rR^{-1}) \tag{A-19}$$

where $q = \sigma_d^2(\|\underline{w}^o\|^2 + 1 + p)$ and $r = 2\mu^2\xi_{min}\sigma_d^2 + 2\mu_c^2\sigma_c^2$. This gives

$$PR = (2\mu I - \mu^2 R)^{-1}(\mu^2 \mathrm{tr}(RP)R + 2\mu^2 qR + rI).$$

Taking the trace and solving for $\mathrm{tr}(RP)$ we get

$$\mathrm{tr}(RP) = \frac{\mathrm{tr}[(2\mu I - \mu^2 R)^{-1}(2\mu^2 qR + rI)]}{1 - \mathrm{tr}[(2\mu I - \mu^2 R)^{-1}(\mu^2 R)]}. \tag{A-20}$$

Then, if $\{\lambda_i\}_{i=1}^p$ are the eigenvalues of $R$, we have, from (A-12), and (A-20)

$$\mathrm{tr}(RP) = E[|\underline{\rho}_k^H \underline{x}_k|^2] = \frac{\sum_{i=1}^p \frac{2\mu^2 q\lambda_i + r}{2\mu - \mu^2\lambda_i}}{1 - \sum_{i=1}^p \frac{\mu^2\lambda_i}{2\mu - \mu^2\lambda_i}}. \tag{A-21}$$

It follows from (A-19) that the eigenvalues $\{\delta_k\}_{k=1}^p$ of $P$ are

$$\delta_k = \frac{\mu \mathrm{tr}(RP) + 2\mu q}{2 - \mu\lambda_k} + \frac{r}{\lambda_k(2\mu - \mu^2\lambda_k)}$$

and therefore

$$\begin{aligned} P &= (\mu \mathrm{tr}(RP) + 2\mu q)(2I - \mu R)^{-1} + \\ &\quad r(2\mu R - \mu^2 R^2)^{-1}. \end{aligned} \tag{A-22}$$

Substitution of (A-21) into (A-22) yields an expression for $P$ which satisfies (A-18). This verifies that the candidate solution is indeed the steady state solution.

Next we specialize to the two cases $\mu = \mu_d \ll 1$, $\mu_c = 1$ and $\mu = \mu_c \ll 1$, $\mu_d = 1$. In the first case, (A-21) gives

$$\begin{aligned} \mathrm{tr}(RP) &= \frac{\mu \mathrm{tr}[(2qR + 2\xi_{min}\sigma_d^2 I)(2I - \mu R)^{-1}]}{1 - \mu \mathrm{tr}[R(2I - \mu R)^{-1}]} + \frac{p\sigma_c^2}{\mu} \\ &\approx \frac{p\sigma_c^2}{\mu} \end{aligned} \tag{A-23}$$

and (A-22) along with (A-23) gives

$$\begin{aligned} P &= p\sigma_c^2(2I - \mu R)^{-1} + 2\mu q(2I - \mu R)^{-1} + \\ &\quad 2\mu\xi_{min}\sigma_d^2(2R - \mu R^2)^{-1} + \\ &\quad 2\sigma_c^2(2\mu R - \mu^2 R^2)^{-1} \\ &\approx p\sigma_c^2(2I - \mu R)^{-1} + \frac{\sigma_c^2}{\mu} R^{-1}. \end{aligned} \tag{A-24}$$

In the second case, (A-21) gives

$$\begin{aligned} \mathrm{tr}(RP) &= \frac{\mu \mathrm{tr}[(2qR + 2\xi_{min}\sigma_d^2 I + 2\sigma_c^2 I)(2I - \mu R)^{-1}]}{1 - \mu \mathrm{tr}[R(2I - \mu R)^{-1}]} \\ &\approx 0 \end{aligned} \tag{A-25}$$

and (A-22) along with (A-25) gives

$$\begin{aligned} P &= 2\mu q(2I - \mu R)^{-1} + 2\mu\xi_{min}\sigma_d^2(2R - \mu R^2)^{-1} \\ &\approx 0. \end{aligned} \tag{A-26}$$

Now, in the first case, ($\mu = \mu_d \ll 1$, $\mu_c = 1$) using (A-10), (A-11), and (A-23) we get

$$\xi_q = \frac{1}{a^2}\left[\frac{p\ \sigma_c^2}{\mu} + 2\sigma_d^2\ (\|\underline{w}^o\|^2 + p)\right] \tag{A-27}$$

from which, by using (A-5), we obtain (12).

## Appendix D
## Derivation of (4)

In the LMS update formula given by (2) and (3), the calculation of the inner product $\underline{w}_k'^H \underline{x}_k'$ 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 $y_k'$ requires an additional complex addition of $B_d+1$ bits. Next, multiplying the conjugate of this quantity ($e_k'$) by $\underline{x}_k'$ 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 $\mu = 2^{-q}$ for some $q \in \{1,\ldots,B_c+1\}$ 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 $\eta_g$ to be the average power consumed per logic gate during an iteration of LMS. Then we have

$$P_g = [24p(3B_d+B_c-2)+32p]\eta_g \tag{A-28}$$

where $P_g$ is the power consumption per iteration of LMS attributed to logic gates. We assume that the power consumed per table lookup operation is linear in the number of bits being multiplied and we define $\eta_t$ to be the average power per table lookup per bit. We then have

$$P_{tl} = 24p(B_d\eta_t) \tag{A-29}$$

where $P_{tl}$ is the power consumption per iteration of LMS attributed to table lookup operations. The total power consumption per iteration of LMS is then

$$P_T = P_g + P_{tl}. \tag{A-30}$$

Equations (A-28), (A-29), and (A-30) lead to (4).

We can derive a similar formula for multiplication using partial product accumulation. In such a multiplier, each $B$-bit multiplication requires $B^2$ full adders and $B^2$ AND gates [15]. Therefore, the $8p$ multiplications in each iteration of LMS require $8p(B_d+1)^2$ full adders and $8p(B_d+1)^2$ AND gates for a total of $56p(B_d+1)^2$ gates. The $4p$ additions of $B_d+1$ bits and the $4p$ additions of $B_c+1$ bits require $4p(6B_d+6B_c+4)$ gates. This gives the following relation for power per iteration of LMS using partial product accumulation multiplication:

$$P_T = [56pB_d^2 + 138pB_d + 24pB_c + 72p]\eta_g. \tag{A-31}$$

Using equations (4) and (A-31) along with (14), we can plot $P_T$ as a function of $B_T$ for both power relations. These plots are shown in figure 9 for $\rho = 1/2$, $\eta_g = 1$ milliwatt, $\eta_t = 10$ milliwatts, and $p = 2$.

Note that the formula derived using partial product accumulation is quadratic in $B_T$ while the formula derived using log table multiplication is linear in $B_T$.

## Appendix E
## Derivation of (15) and (16)

Using equations (12) and (14) we can write the excess MSE due to quantization as

$$\xi_q = \alpha_c 2^{-2(1-\rho)B_T} + \alpha_d 2^{-2\rho B_T}. \tag{A-32}$$

Differentiating this equation with respect to $\rho$ we get

$$\begin{aligned}\frac{d\xi_q}{d\rho} &= 2\ln 2\ \alpha_c\ B_T \cdot 2^{-2(1-\rho)B_T} - 2\ln 2\ \alpha_d\ B_T \cdot 2^{-2\rho B_T} \\ &= 2\ln 2\ B_T\ \left(\alpha_c 2^{-2(1-\rho)B_T} - \alpha_d 2^{-2\rho B_T}\right). \end{aligned} \tag{A-33}$$

Differentiating again we get

$$\begin{aligned}\frac{d^2\xi_q}{d\rho^2} &= (2\ln 2\ B_T)^2 \left(\alpha_c 2^{-2(1-\rho)B_T} + \alpha_d 2^{-2\rho B_T}\right) \\ &= (2\ln 2\ B_T)^2\ \xi_q > 0. \end{aligned} \tag{A-34}$$

Equation (A-34) shows that for $\rho \in [0,1]$, $\frac{d^2\xi_q}{d\rho^2} > 0$ and therefore, $\xi_q$ is a convex function of $\rho$. Now, setting $\frac{d\xi_q}{d\rho}$ in (A-33) equal to zero we have

$$\begin{aligned}\left.\frac{d\xi_q}{d\rho}\right|_{\rho=\rho^*} &= 0 \\ &= 2\ln 2\ B_T\ \left(\alpha_c 2^{-2(1-\rho^*)B_T} - \alpha_d 2^{-2\rho^* B_T}\right),\end{aligned}$$

from which we get

$$\alpha_c 2^{-2(1-\rho^*)B_T} = \alpha_d 2^{-2\rho^* B_T}.$$

This implies

$$\frac{\alpha_d}{\alpha_c} = 2^{4\rho^* B_T - 2B_T}. \tag{A-35}$$

Equation (A-35) leads directly to (15).

To find the minimum value of $\xi_q$ we use (15) in (A-32).

$$\min_{\rho} \xi_q = \alpha_c 2^{-2(1-\rho^*)B_T} + \alpha_d 2^{-2\rho^* B_T}. \tag{A-36}$$

Equation (A-36) leads directly to (16).

## Appendix F
## Derivation of (18)

Define the following constants

$$\begin{aligned} A &\equiv \frac{P_T + 16p\ \eta_g}{p}, \\ B &\equiv 48\ \eta_g + 24\ \eta_t, \\ C &\equiv 24\ \eta_g. \end{aligned} \tag{A-37}$$

Then from (17) we have

$$B_T = \frac{A}{B\rho + C} \tag{A-38}$$

Differentiating (A-32) with respect to $\rho$ gives

$$\begin{aligned} \frac{d\xi_q}{d\rho} &= \ln 2\ \alpha_c\ 2^{-2(1-\rho)B_T} \cdot \frac{d}{d\rho}[-2(1-\rho)B_T] \quad + \\ & \ln 2\ \alpha_d\ 2^{-2\rho\ B_T} \cdot \frac{d}{d\rho}[-2\rho\ B_T] \end{aligned} \tag{A-39}$$

Differentiating again, we have

$$\begin{aligned} \frac{d^2\xi_q}{d\rho^2} &= \ln 2\ \alpha_c\ 2^{-2(1-\rho)B_T} \left[\frac{d^2}{d\rho^2}[-2(1-\rho)B_T] + \right. \\ & \left. \ln 2 \left(\frac{d}{d\rho}[-2(1-\rho)B_T]\right)^2\right] + \\ & \ln 2\ \alpha_d\ 2^{-2\rho B_T} \left[\frac{d^2}{d\rho^2}[-2\rho B_T] + \right. \\ & \left. \ln 2 \left(\frac{d}{d\rho}[-2\rho B_T]\right)^2\right]. \end{aligned} \tag{A-40}$$

Now, from (A-38) we have

$$\begin{aligned} \frac{d}{d\rho}[-2\rho B_T] &= -\frac{2AC}{(B\rho + C)^2}, \\ \frac{d}{d\rho}[-2(1-\rho)B_T] &= \frac{2A(B+C)}{(B\rho + C)^2}, \\ \frac{d^2}{d\rho^2}[-2\rho B_T] &= \frac{4ABC}{(B\rho + C)^3}, \\ \frac{d^2}{d\rho^2}[-2(1-\rho)B_T] &= -\frac{4AB(B+C)}{(B\rho + C)^3}. \end{aligned} \tag{A-41}$$

Using (A-41) in (A-40) gives

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

$$\ln 2 \; \alpha_d \; 2^{-2\rho B_T} \left[ \frac{4ABC}{(B\rho + C)^3} + \right.$$
$$\left. \ln 2 \cdot \frac{4A^2C^2}{(B\rho + C)^4} \right]. \tag{A-42}$$

Now, $\xi_q$ is convex if $\frac{d^2\xi_q}{d\rho^2}$ is positive. From (A-42), it is clear that $\frac{d^2\xi_q}{d\rho^2}$ will be positive if the following condition is met:

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

This is equivalent to

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

Using (A-37), this becomes

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

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

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

which is equivalent to

$$B_T > \frac{1}{\ln 2} \left[ 1 - \frac{24\eta_g}{72\eta_g + 24\eta_t} \right]. \tag{A-43}$$

The term on the right hand side of (A-43) 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, $\xi_q$ is once again a convex function of $\rho$. To solve for $\rho^{**}$, we set $\frac{d\xi_q}{d\rho}$ equal to zero. Using (A-39) and (A-41) we have

$$\alpha_c \; 2^{-2(1-\rho^{**}) \cdot \frac{A}{B\rho^{**}+C}} \cdot \frac{2A(B+C)}{(B\rho^{**} + C)^2} \quad =$$
$$\alpha_d \; 2^{-2\rho^{**} \cdot \frac{A}{B\rho^{**}+C}} \cdot \frac{2AC)}{(B\rho^{**} + C)^2}.$$

This implies

$$4A\rho^{**} - 2A = \log_2 \left( \frac{\alpha_d}{\alpha_c} \cdot \frac{C}{B+C} \right) \cdot (B\rho^{**} + C)$$

which is equivalent to

$$\rho^{**} = \frac{2A + C \log_2 \left( \frac{\alpha_d}{\alpha_c} \cdot \frac{C}{B+C} \right)}{4A - B \log_2 \left( \frac{\alpha_d}{\alpha_c} \cdot \frac{C}{B+C} \right)} \tag{A-44}$$

Equations (A-37) and (A-44) lead to (18).

# References


[1] J. McChesney, *Notes for MURI Seminar on Low Energy Electronics Design for Mobile Platforms, The University of Michigan*, Oct. 1997.

[2] K. K. Parhi, "Algorithm transformation techniques for concurrent processors," *IEEE Proceedings*, vol. 77, pp. 1879–1895, Dec. 1989.

[3] A. P. Chandrasekan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *IEEE Proceedings*, vol. 83, pp. 498–523, April 1995.

[4] A. Peled and B. Liu, *Digital Signal Processing: Theory, Design, and Implementation*, Wiley, New York, 1976.

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

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

[7] B. Widrow and *et al*, "Adaptive antenna systems," *IEEE Proceedings*, vol. 55, pp. 2143–2159, 1967.

[8] W. F. Gabriel, "Adaptive arrays – an introduction," *IEEE Proceedings*, vol. 64, pp. 239–272, Feb. 1976.

[9] S. Douglas and T.-Y. Meng, "Optimum error quantization for lms adaptation," in *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, May 1991.

[10] 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.

[11] D. Singh, J. M. Rabaey, M. Pedram, F. Catthoor, S. Rajgopal, N. Sehgal, and T. J. Mozden, "Power conscious CAD tools and methodologies: a perspective," *IEEE Proceedings*, vol. 83, pp. 5701–594, April 1995.

[12] B. Widrow, J. McCool, and M. Ball, "Complex LMS algorithm," *IEEE Proceedings*, vol. 63, pp. 719–720, 1975.

[13] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[14] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, Upper Saddle River, NJ, 1997.

[15] R. H. Katz, *Contemporary Logic Design*, Benjamin-Cummings, 1994.


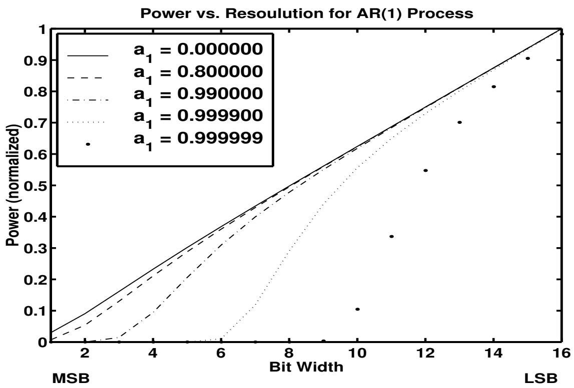


Fig. 1. *Power versus bit width $B$ as function of AR parameter $a_1$ for loading a $B$-bit register with successive samples of an AR(1) process. Curves are normalized relative to power $P_{16}$ consumed for a white sequence in a 16 bit register. In the plot the bit width is reduced from right to left by eliminating LSB's*

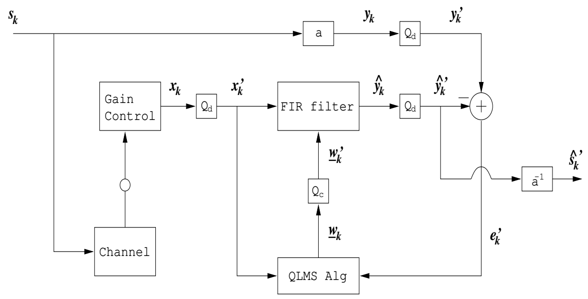


Fig. 2. *Quantized LMS adaptive channel equalizer with training sequence $s_k$.*

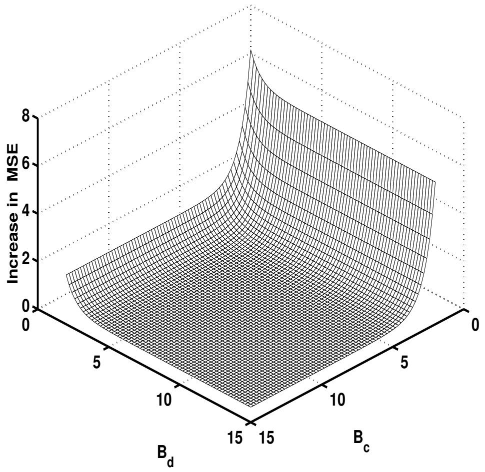


Fig. 3. *Excess MSE due to quantization as a function of $B_d$ and $B_c$ for single pole IIR channel.*

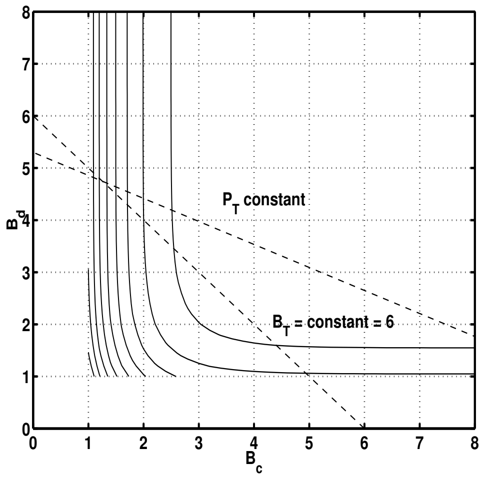


Fig. 4. *Excess MSE due to quantization as a function of $B_d$ and $B_c$ for single pole IIR channel.*

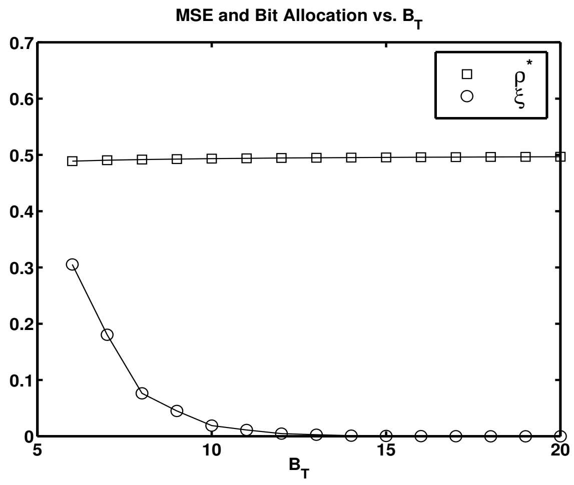


Fig. 5. *IIR channel with 4-PSK input. Optimal data bit allocation factor under $B_T$ constraint and total MSE as function of $B_T$.*

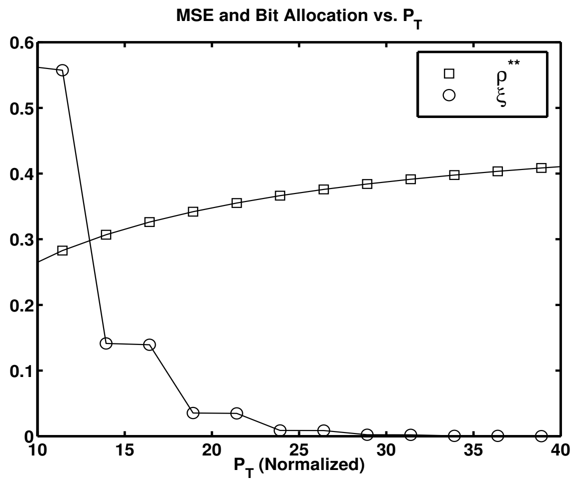


Fig. 6. *IIR channel with 4-PSK input. Optimal data bit allocation factor under $P_T$ constraint and total MSE as function of $P_T$.*

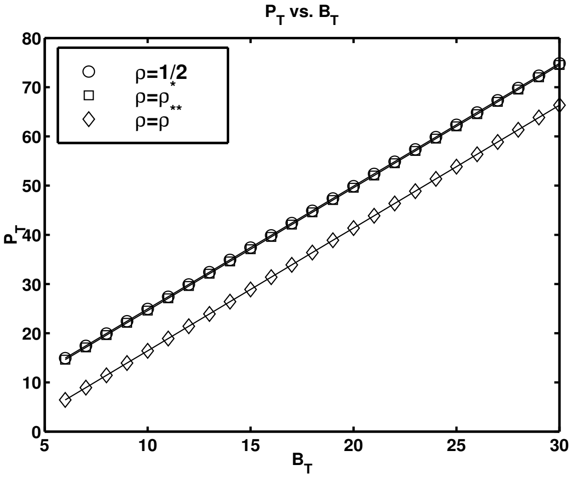


Fig. 7. *Power $P_T$ as function of $B_T$ for $\rho = \frac{1}{2}, \rho^*, \rho^{**}$*

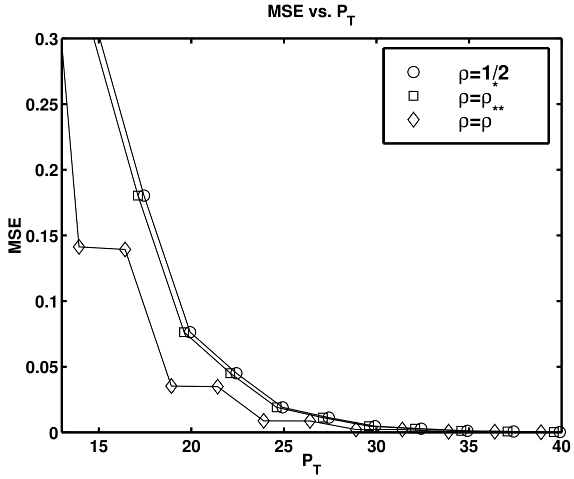


Fig. 8. *IIR channel with 4-PSK input. MSE as function of $P_T$ for $\rho = \frac{1}{2}, \rho^*, \rho^{**}$*

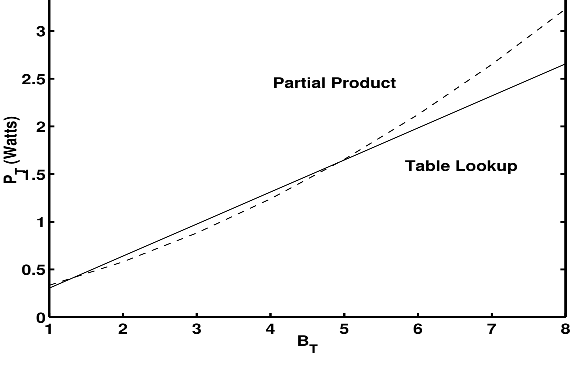


Fig. 9. *$P_T$ vs. $B_T$ for different power relations*