INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. 2003; **17**:237–258 (DOI: 10.1002/acs.738)

Fixed-point steady-state analysis of adaptive filters

Nabil R. Yousef and Ali H. Sayed*,†,‡

Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA

SUMMARY

The steady-state performance of adaptive filters can vary significantly when they are implemented in finite precision arithmetic, which makes it vital to analyse their performance in a quantized environment. Such analyses can become difficult for adaptive algorithms with non-linear update equations. This paper develops a feedback and energy-conservation approach to the steady-state analysis of quantized adaptive algorithms that bypasses some of the difficulties encountered by traditional approaches. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: adaptive filter; quantized analysis; finite wordlength; fixed point

1. INTRODUCTION

This paper develops an approach to the roundoff error analysis of adaptive filters. The approach is based on showing how a generic quantized adaptive filter can be represented as a cascade of elementary sections, with each section consisting of a lossless mapping in the feedforward path and a feedback interconnection, with roundoff errors acting as disturbances to the system. By studying the energy flow through the cascade, we are able to establish a fundamental error variance relation. Using this relation, we are able to extend results in the infinite precision case to the quantized case with minimal calculations for a large class of adaptive algorithms. We also derive new results.

Thus consider noisy measurements $\{d(i)\}$ that arise from the linear model

$$d(i) = \mathbf{u}_i \mathbf{w}^0 + v(i) \tag{1}$$

where $\mathbf{w}^0$ is a stationary deterministic unknown N × 1 vector that we wish to estimate, $v(i)$ accounts for stationary stochastic measurement noise and modelling errors, and $\mathbf{u}_i$ denotes a *row* input (regressor) vector of stationary stochastic elements. Many adaptive schemes have been developed in the literature for the estimation of $\mathbf{w}^0$ in different contexts (e.g. echo cancellation, channel estimation, channel equalization). In this paper, we focus on the following general class

*Correspondence to: Ali H. Sayed, Department of Electrical Engineering University of California, Los Angeles Los Angeles, CA 90095. Tel.: (3 1 0)267-2142; fax: (3 1 0)206-8495

† E-mail: sayed@ee.ucla.edu

‡ Now with Broadcom Corporation, Irvine, CA.

Contract sponsor grant: National Science Foundation; contract sponsor number: CCR-9732376; ECS-9820765 and CRR-0208573.

Published online 13 March 2003
Received 16 October 2000
Accepted 22 March 2002

Table I. Examples for $f_e(i)$

Algorithm	$fe(i)$
LMS	$e(i)$
NLMS	$e(i)/\|\| \mathbf{u}_i \|\|^2$
LMF	$e^3(i)$
LMMN	$\delta e(i) + (1 - \delta)e^3(i)$
SA	$\text{sign}[e(i)]$
CMA	$y(i)[R2 - \| y(i) \|^2]$

of algorithms:

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \mu \mathbf{u}_i^* \ f_e(i) \tag{2}$$

where $\mathbf{w}_i$ is an estimate for $\mathbf{w}^0$ at iteration i and μ is the step-size. Usually, $f_e(i)$ is a function of the output estimation error, defined by

$$e(i) = d(i) - \mathbf{u}_i \mathbf{w}_i \tag{3}$$

Different choices for $f_e(i)$ result in different adaptive algorithms. For example, Table I defines $f_e(\mathrm{i})$ for several special cases of (2). [§] In Table I, $0 \leqslant \delta \leqslant 1$, R_2 is a positive constant, and $y(i) = \mathbf{u}_i \ \mathbf{w}_i$ is the adaptive filter output.

An important performance measure for an adaptive filter is its steady-state mean-square-error (MSE), which is defined as

$$\text{MSE} = \lim_{i\to\infty} \ \mathrm{E}\left(|e(i)|^2\right) = \lim_{i\to\infty} \ \mathrm{E}\left(|v(i) + \mathbf{u}_i \tilde{\mathbf{w}}_i|^2\right)$$

where $\tilde{\mathbf{w}}_i = \mathbf{w}^0 - \mathbf{w}_i$ denotes the weight error vector.

Under the realistic assumption that (see, e.g. References [1–6]):

A.1. The noise sequence $\{v(i)\}$ is independently and identically distributed (iid) and also statistically independent of the regressor sequence $\{\mathbf{u}_i\}$.

we find that the MSE is equivalently given by[¶]

$$\text{MSE} = \sigma_v^2 + \lim_{i\to\infty} E(|\mathbf{u}_i \tilde{\mathbf{w}}_i|^2) \tag{4}$$

Now the standard way for evaluating (4), and which dominates most derivations in the literature, is the following. First, one assumes, in addition to A.1, that the regression vector $\mathbf{u}_i$ is independent of $\tilde{\mathbf{w}}_i$. Then the above MSE becomes

$$\text{MSE} = \sigma_v^2 + \lim_{i\to\infty} \ \mathrm{Tr}(\mathbf{R}\mathbf{C}_i) \tag{5}$$

where $\mathbf{C}_i = E(\tilde{\mathbf{w}}_i \tilde{\mathbf{w}}_i^*)$ denotes the weight error covariance matrix and $\mathrm{R} = E(\mathbf{u}_i^* \mathbf{u}_i)$ is the input covariance matrix. As is evident from (5), this method of computation requires the determination of the steady-state value of $\mathbf{C}_i$, say $\mathbf{C}_\infty$. In quantized environments, finding $\mathbf{C}_\infty$

[§] The list in the table assumes real-valued data. For complex-valued data, we replace e^3 by $e|e|^2$ and define sign $[a + jb]$ by $\frac{1}{2}$ (sign$[a]$ + $j$sign$[b]$).

[¶] The value of $\tilde{\mathbf{w}}_i$ depends only on the past values of $v(i)$ and the current and past values of $\mathbf{u}_i$. Thus, the expected value of the cross terms between $v(i)$ and $\mathbf{u}_i \tilde{\mathbf{w}}_i$ vanishes since $v(i)$ is white and statistically independent of $\mathbf{u}_i$.

is a burden, especially for adaptive schemes with non-linear update equations. The following are the contributions of this work:

1. We develop an energy-conservation approach for evaluating the MSE of a large class of adaptive schemes when implemented in finite precision. This approach bypasses the need for working directly with $\mathbf{C}_i$ or with its limiting value, and it extends the results of [7] to the finite precision case.
2. The approach further establishes the useful conclusion that the finite precision analysis of an adaptive scheme can be obtained almost by inspection from the results in the infinite precision case for a large class of algorithms, such as the LMS, NLMS, LMF, LMMN, and sign algorithms. In contrast, analyses for both cases have usually been carried out separately in the literature.
3. The approach allows us to derive a handful of new results, especially for adaptive filters with non-linear updates for which approaches that require $\mathbf{C}_i$ are not easily applicable, such as the LMF, LMMN, and CM algorithms.

We may remark that quantization errors can affect the performance of an adaptive filter in several ways. For instance, these errors may affect the stability of the adaptive algorithm, i.e. they may cause the algorithm to diverge. They can also degrade the steady-state performance of the adaptive algorithm by causing the filter to attain a higher MSE value than what is expected in the infinite precision case. The instability behaviour is more serious for the class of recursive least squares (RLS) algorithms. In this paper, however, we focus on LMS-like algorithms. Such algorithms are known for their inherent stability and robustness in fixed point environments (see, e.g. References [1–6]). This is because finite precision errors do not tend to affect their transient performance significantly; gradient errors are relatively large in the transient phase, which makes the finite precision effects less significant. However, when LMS-like algorithms approach steady-state, the adaptation error becomes relatively smaller and thus finite precision errors can cause performance degradation in the form of excess mean-square-error. The main goal of this paper is to derive expressions for the MSE of such LMS-like algorithms in a quantized environment.

2. A. MATHEMATICAL MODEL

Figure 1 shows the quantized model used in the paper,[||] and which is widely used in the context of finite precision analyses of adaptive algorithms (see, e.g. References [9–14]). In this figure, $Q[x]$ denotes the fixed point quantization of the value x, and the superscript q distinguishes quantized quantities from infinite precision quantities. Throughout the paper, rounding quantization is considered. It is also assumed that the saturation thresholds of the quantizers are properly chosen such that overflow never occurs and saturation errors are negligible. Thus, only rounding errors are considered. The variance σ^2 of the rounding error, for real-valued

[||] Figure 1 shows only the system model. It does not show the quantizers used to obtain (2). A modified version of this model will be used for the case of the CMA.

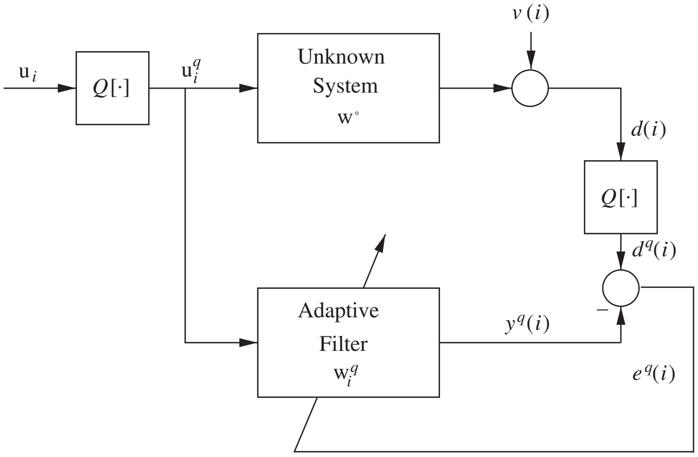

Figure 1. Quantization model.

quantities, is related to the quantizer saturation threshold L according to

$$\sigma^2 = \frac{2^{-2\mathrm{B}}L^2}{12} \tag{6}$$

where it is assumed that the quantizer uses B bits in addition to a sign bit. The values of B and L considered for quantization of the data ($\mathbf{u}_i$, $d(i)$, and $y(i)$) will be denoted by B_d and L_d , and the ones considered for quantization of the filter coefficients will be denoted by B_c and L_c. The corresponding values of σ^2 will be denoted by σ_d^2 and σ_c^2, respectively. For complex-valued quantities, the variance σ^2 is twice the value given in (6). Throughout the paper, we will use σ^2 to denote the rounding error variance for both real and complex-valued quantities. However, the value of σ^2 differs in each case.

We can write

$$d^q(i) = d(i) + \tilde{d}(i), \quad \mathbf{u}_i^q = \mathbf{u}_i + \tilde{\mathbf{u}}_i, \quad y^q(i) = \mathbf{u}_i^q\mathbf{w}_i^q + \tilde{y}(i) \tag{7}$$

where $\tilde{d}(i)$ is the system output quantization error with variance σ_d^2, $\tilde{\mathbf{u}}_i$ is a vector of input data quantization errors with σ_d^2 being the variance of each of its entries, and $\tilde{y}(i)$ is the quantization error that occurs in computing the term $\mathbf{u}_i^q\mathbf{w}_i^q$. The variance of $\tilde{y}(i)$, $\sigma_{\tilde{y}}^2$, depends on the procedure by which $y^q(i)$ is computed. If all N products involved in $\mathbf{u}_i^q\mathbf{w}_i^q$ are computed with high precision, summed, and the final result is quantized to B_d bits, then $\sigma_{\tilde{y}}^2$ is approximately equal to σ_d^2. If each one of the N products is quantized to B_d bits first, $\sigma_{\tilde{y}}^2$ is equal to $N\,\sigma_d^2$. The quantized estimation error $e^q\,(i)$ is given from (7) by

$$e^q(i) = d^q(i) - y^q(i) = e(i) + \tilde{e}(i) \tag{8}$$

where $\tilde{e}(i) = \tilde{d}(i) - \tilde{y}(i) - \tilde{\mathbf{u}}_i\mathbf{w}^0 + \tilde{\mathbf{u}}_i\tilde{\mathbf{w}}_i$. Obviously, $\tilde{e}(i)$ is a zero-mean sequence. Here, we note that, in steady-state, the variance of the term $\tilde{\mathbf{u}}_i\tilde{\mathbf{w}}_i$ is equal to $\sigma_d^2\ E\ (\|\tilde{\mathbf{w}}_i\|^2)$. If we assume that $E\ (\|\tilde{\mathbf{w}}_i\|^2) \ll 1$ in steady-state, which is usually valid in practical applications, then we can approximate $\tilde{e}(i)$ by $\tilde{e}(i) \approx \tilde{d}(i) - \tilde{y}(i) - \tilde{\mathbf{u}}_i\mathbf{w}^0$, with variance $\sigma_{\tilde{e}}^2 = \sigma_d^2 + \sigma_{\tilde{y}}^2 + \sigma_d^2\|\mathbf{w}^0\|^2$. We denote the quantized error function by $f_e^q(i)$.

Taking the above quantizations into consideration, the adaptive recursion (2) becomes

$$\begin{aligned} w_{i+1}^q &= w_i^q + Q[\mu\, \mathbf{u}_i^{q*}\ \ f_e^q(i)] \\ &= w_i^q + \mu\, \mathbf{u}_i^{q*}\ \ f_e^q(i) - \mathbf{m}_i \end{aligned} \tag{9}$$

where $\mathbf{m}_i$ is a vector of multiplication quantization errors in the update term $\mu\mathbf{u}_i^{q*}\ \ f_e^q(i)$, which is defined for convenience of notation by

$$\mathbf{m}_i \overset{\triangle}{=} \mu\, \mathbf{u}_i^{q*}\ \ f_e^q(i) - Q[\mu\, \mathbf{u}_i^{q*}\ \ f_e^q(i)]$$

each entry of $\mathbf{m}_i$ has variance σ_c^2. The weight error vector is now defined as

$$\tilde{\mathbf{w}}_i = \mathbf{w}^0 - \mathbf{w}_i^q \tag{10}$$

3. QUANTIZED ENERGY RELATION

We start by defining the a-priori and a-posteriori estimation errors,

$$e_a(i) = \mathbf{u}_i^q \tilde{\mathbf{w}}_i, \quad e_p(i) = \mathbf{u}_i^q(\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i)$$

Using (3) and (8), it is easy to see that the errors $\{e^q(i), e_a(i)\}$ are related via

$$e^q(i) = e_a(i) + v(i) + \tilde{e}(i)$$

If we subtract $\mathbf{w}^0$ from both sides of (9) and multiply by $\mathbf{u}_i^q$ from the left, we also find that the errors $\{e_p(i), e_a(i), e^q(i)\}$ are related via

$$e_p(i) = e_a(i) - \mu\|\mathbf{u}_i^q\|^2 f_e^q(i) \tag{11}$$

Substituting (11) into (9), we obtain the update relation

$$\tilde{\mathbf{w}}_{i+1} = \tilde{\mathbf{w}} - \frac{\mathbf{u}_i^{q*}}{\|\mathbf{u}_i^q\|^2}[e_a(i) - e_p(i)] + \mathbf{m}_i$$

By evaluating the energies of both sides of this equation and using a similar procedure to that used in References [7, 8], we obtain

$$\|\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i\|^2 + \frac{1}{\|\mathbf{u}_i^q\|2}|e_a(i)|^2 = \|\tilde{\mathbf{w}}_i\|^2 + \frac{1}{\|\mathbf{u}_i^q\|^2}|e_p(i)|^2 \tag{12}$$

When $\mathbf{u}_i^q = 0$, it is obviously true that

$$\|\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i\|^2 = \|\tilde{\mathbf{w}}_i\|^2 \tag{13}$$

Both results (12) and (13) can be grouped together into a single equation by defining

$$\bar{\mu}(i) = (\|\mathbf{u}_i^q\|^2)^\dagger$$

in terms of the pseudo-inve rse of a scalar,[**] so that we obtain

$$\|\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i\|^2 + \bar{\mu}(i)|e_a(i)|^2 = \|\tilde{\mathbf{w}}_i\|^2 + \bar{\mu}(i)|e_p(i)|^2 \tag{14}$$

This energy conservation relation, first established in References [15–17], holds for all adaptive algorithms whose recursions are of the form given by (2). *No approximations or assumptions are*

[**] For a scalar x, the pseudo-inverse $x^\dagger$ is equal to $1/x$ for $x \neq 0$ and equal to zero for $x = 0$.

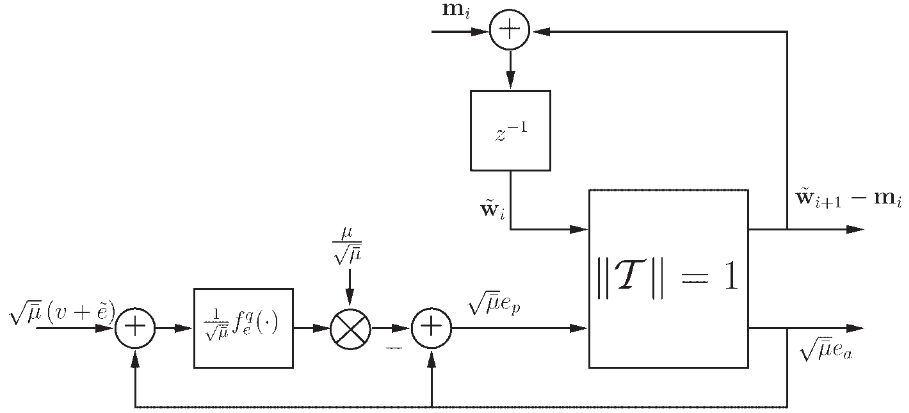
$\mathbf{m}_i$
z^{-1}
$\tilde{\mathbf{w}}_i$
$\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i$
$\|\mathcal{T}\| = 1$
$\frac{\mu}{\sqrt{\bar{\mu}}}$
$\sqrt{\bar{\mu}}(v + \tilde{e})$
$\frac{1}{\sqrt{\bar{\mu}}} f_e^q(\cdot)$
$-$
$\sqrt{\bar{\mu}}e_p$
$\sqrt{\bar{\mu}}e_a$

Figure 2. Lossless mapping and a feedback loop.

needed to establish (14); it is an exact relation that shows how the energies of the weight error vectors at two successive time instants are related to the energies of the a priori *and* a posteriori estimation errors. The relation also has an interesting system-theoretic interpretation. It establishes that the mapping from $\{\tilde{\mathbf{w}}_i, \sqrt{\bar{\mu}(i)e_p(i)}\}$ to $\{\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i, \sqrt{\bar{\mu}(i)e_a(i)}\}$ is energy preserving (or lossless). Furthermore, combining (14) with (11), we see that both relations establish the existence of the feedback configuration shown in Figure 2, where T denotes a lossless map and z^{-1} denotes the unit delay operator. Here, we can see that the quantization error vector $\mathbf{m}_i$ acts as a disturbance input to the system. Such a disturbance plays the same role as that of system non-stationarity [18]. On the other hand, the data quantization error $\tilde{e}(i)$ is added to the plant noise $v(i)$.

3.1. Relevance to fixed point analysis

Relation (14) has several ramifications. It was used in References [15–17] to study the robustness and l_2-stability of adaptive filters and in References [7, 18] to study the steady-state and tracking performances of various adaptive algorithms. It was also used in References [19, 20] to study the transient performance of adaptive filters. Here, we show its significance to finite precision analyses of adaptive algorithms.

First, we impose the following modelling assumption.

A.2. Quantization errors are zero-mean, mutually independent, and independent of all other signals.

This assumption is typical in the context of finite precision analysis of adaptive algorithms (see e.g. References [9–14]), and it enables the derivation of closed-form expressions for the steady-state MSE. A more sophisticated non-linear model for treating quantization errors, which takes into account quantizer underflow effects, has been used in Reference [21] for the LMS algorithm; though it does not lead to closed-form expressions.

Imposing the equality $E\,(\|\tilde{\mathbf{w}}_{i+1}\|^2) = E\,(\|\tilde{\mathbf{w}}_i\|^2)$ in steady-state, and using (11) and A.2 it is straightforward to verify that the energy relation (14) leads to

$$E(\bar{\mu}(i)|e_a(i)|^2) = \mathrm{Tr}(\mathrm{M}) + E\left(\bar{\mu}(i)\left|e_a(i) - \frac{\bar{\mu}}{\bar{\mu}(i)} f_e^q(i)\right|^2\right) \tag{15}$$

where $\mathbf{M} = E\left(\mathbf{m}_i\mathbf{m}_i^*\right)$. For iid multiplication errors, Tr(**M**) $= N\,\sigma_c^2$. This equation can now be solved for the steady-state excess mean-square-error (EMSE):

$$\zeta \overset{\triangle}{=} \lim_{i\to\infty}\ \text{E}\ \left(|e_a(i)|^2\right)$$

Observe from (4) that the desired MSE is given by MSE $= \sigma_v^2 + \zeta$, so that finding ζ is equivalent to finding the MSE.

3.2. *A class of error functions*

We now focus on the class of algorithms whose error functions satisfy the following condition:

$$f_e^q(i) = f_e^q(\mathbf{u}_e^q, e^q(i)) = f_e(\mathbf{u}_e^q, e^q(i)) + \eta(i) \tag{16}$$

where $\eta\,(i)$ is a zero-mean random variable, which is statistically independent of all other algorithm quantities. Note that $\eta\,(i)$ is the error in calculating $f_e^q(\mathbf{u}_e^q, e^q(i))$ from $\mathbf{u}_i^q$ and $e^q(i)$. The variance of $\eta\,(i)$ also depends on the adaptive algorithm used.

For the LMS and sign algorithms, we can see that this condition is satisfied with $\sigma_\eta^2 = 0$. For the LMF and LMMN algorithms, if the quantity $(e^q(i))^3$ is calculated via a look-up table or if the term is first calculated in a high precision, then both algorithms satisfy the condition in (16), with σ_η^2 equal to σ_d^2 and $2\sigma_d^2 + (1-\delta)^2\sigma_d^2$ for the LMF and LMMN algorithms, respectively. To obtain the error function of the NLMS algorithm, the norm $||\mathbf{u}_i^q||^2$ is usually calculated in $2B_d$ bits precision, then $1/||\mathbf{u}_i^q||^2$ is obtained using a look-up table, multiplied by $e^q(i)$, and quantized to B_d bits [12]. In this case, we have

$$\eta^{\text{NLMS}}(i) = e_1(i)e^q(i) + e_2(i)$$

where $e_1(i)$ and $e_2(i)$ are two zero-mean random variables of variance σ_d^2. In this case, the NLMS algorithm does not generally satisfy condition (16). However, note that, in steady-state and due to A.2, the variance of the term $e_1(i)e^q(i)$ is equal to σ_d^2 (MSE $+\ \sigma_d^2$), which can be neglected with respect to σ_d^2, since (MSE $+\ \sigma_d^2) \ll 1$, which is reasonable in practical applications. In this case, the NLMS satisfies condition (16) with $\sigma_\eta^2\ =\ \sigma_d^2$. On the other hand, the CMA does not satisfy (16).

Using (8), (16), and A.2, we can rewrite the error variance relation (15), in terms of $e_a(i)$ and $\bar{v}(i) \overset{\triangle}{=} v(i) + \bar{e}(i)$ as

$$E(\bar{\mu}(i)|e_a(i)|^2) = |\text{Tr}(\mathbf{M}) + \mu^2\sigma_\eta^2\text{Tr}(\text{R}^q) + E\left(\bar{\mu}(i)\left|e_a(i) - \frac{\mu}{\bar{\mu}(i)}f_e(\mathbf{u}_i^q,\ \ e_a(i) + \bar{v}(i))\right|^2\right) \tag{17}$$

where $\text{Tr}(\mathbf{R}^q) = E(\mathbf{u}_i^{q*}\mathbf{u}_i^q) = \text{Tr}(\mathbf{R}) + N\sigma_d^2$.

Moreover, for the infinite precision case, Equation (17) is given by [7]:

$$E(\bar{\mu}(i)|e_a(i)|^2) = E\left(\bar{\mu}(i)\left|e_a(i) - \frac{\mu}{\bar{\mu}(i)}f_e(\mathbf{u}_i,\ \ e_a(i) + v(i))\right|^2\right) \tag{18}$$

where all the quantities are now defined in terms of $\{\mathbf{u}_i,\ \ v(i)\}$ instead of $\left\{\mathbf{u}_i^q, \bar{v}(i)\right\}$. Comparing (17) with (18), we can observe the following. If we replace $\{\mathbf{u}_i,\ \ v(i)\}$ by $\left\{\mathbf{u}_i^q, \bar{v}(i)\right\}$, and add the two terms Tr($\mathbf{M}$) and $\mu^2\sigma_\eta^2\text{Tr}(\mathbf{R}^q)$ to the right-hand side of the infinite precision variance relation (18), we obtain the finite precision variance relation (17).

Thus, instead of directly solving (17) for ζ, we can do the following. First, we evaluate both sides of the infinite precision variance relation (18). Second, we replace terms in the result that are related to $\{\mathbf{u}_i,\ \ v(i)\}$ by the corresponding terms that are related to $\{\mathbf{u}_i^q, \bar{v}(i)\}$ and add the two terms Tr($\mathbf{M}$) and $\mu^2\sigma_\eta^2\text{Tr}(\mathbf{R}^q)$ to the right-hand side. Finally, we solve the resulting equation for $\zeta = E(|e_a(i)|^2)$. Fortunately, we do not need to perform the first step as this is already done in the infinite precision analysis [7]. This is a useful observation in the context of finite precision analysis of adaptive algorithms, as it shows how to extend the results of the infinite precision case to those of the quantized case with minimal effort if (16) is satisfied. In the literature, both cases have generally been studied separately.

Here, we want to stress that the approach presented in this paper is not restricted by the validity of (16). In fact, it can be used regardless of (16). If, for a specific algorithm, (16) is valid, then the infinite precision results can be extended to the finite precision case. If not, then Equation (15) can be solved to get the EMSE in the general case. In this paper, we will find the EMSE for CMA, for which (16) does not hold.

4. QUANTIZED ANALYSIS

We now apply the above general procedure to various adaptive algorithms from Table I. Due to space limitations, we omit some trivial details and only highlight the main steps in the arguments. The reader will soon realize the convenience of working with (18).

4.1. The LMS algorithm

First, we solve both the infinite and finite precision energy equations (18) and (17) for the LMS algorithm, and show how to extract the results of the quantized case from those of the infinite precision case. Later we directly apply our procedure to other algorithms.

Infinite precision case: For LMS, in the infinite precision case, we have $f_e(i) = e(i) = e_a(i) + v(i)$. Substituting into (18) and using A.1, it follows immediately that

$$2\mu\zeta^{\text{LMS}} = \mu^2 E(\|\mathbf{u}_i\|^2 |e_a(i)^2|) + \mu^2\sigma_v^2\text{Tr}(\mathbf{R}) \tag{19}$$

To solve for ζ LMS we consider three cases:

1. For sufficiently small μ, we can assume that the term $\mu^2 E(\|\mathbf{u}_i\|^2 |e_a(i)^2)$ is negligible relative to the second term on the right-hand side of (19), so that

$$\zeta^{\text{LMS}} = \frac{\mu}{2}\sigma_v^2\text{Tr}(\mathbf{R}) \qquad (\text{small}\ \ \mu) \tag{20}$$

2. For larger values of μ, Equation (19) can be solved by imposing the following assumption:[††]

A.3. At steady state, $\|\mathbf{u}_i\|^2$ is statistically independent of $|e_a(i)^2|$.

This assumption in fact becomes realistic for long filter lengths. Furthermore, it becomes exact for the case of constant modulus data that arises in some adaptive filtering applications

[††] By larger values of μ we do not mean a large μ, but rather step-size values that are not infinitesimally small and still guarantee filter stability.

(see, e.g. Reference [22]). Using A.3, and (19), we directly obtain [1, 2]

$$\zeta^{\mathrm{LMS}} = \frac{\mu\sigma_v^2 \mathrm{Tr}(\mathbf{R})}{2 - \mu \mathrm{Tr}(\mathbf{R})} \quad (\text{large } \ \mu) \tag{21}$$

3. For Gaussian white-input signals $\left(\mathbf{R} = \sigma_u^2 \mathbf{I}\right)$, Equation (19) can be more accurately solved by imposing the widely used independence assumption [2]:

A.4. At steady state, $\tilde{\mathbf{w}}_i$ is statistically independent of $\mathbf{u}_i$

to yield the well-known result

$$\zeta^{\mathrm{LMS}} = \frac{\mu\sigma_v^2 \mathrm{Tr}(\mathbf{R})}{2 - \mu(N + \lambda)\sigma_u^2} \quad (\text{Gaussian}) \tag{22}$$

where N is the filter length, $\lambda = 1$ if the $\{\mathbf{u}_i\}$ are complex-valued and $\lambda = 2$ if the $\{\mathbf{u}_i\}$ are real-valued.

Finite precision case: Now for the quantized case, substituting in (17) and using **A.1** and **A.2**, we obtain

$$2\mu\zeta^{\mathrm{LMS}} = \mathrm{Tr}(\mathbf{M}) + \mu^2 E\left(\|\mathbf{u}_i^q\|^2 |e_a(i)^2|\right) + \mu^2\sigma_v^2 \mathrm{Tr}(\mathbf{R}^q)$$

For small enough values of μ, we have

$$\zeta^{\mathrm{LMS}} = \tfrac{1}{2}(\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\sigma_{\bar{v}}^2 \mathrm{Tr}(\mathbf{R}^q)) \tag{23}$$

where $\sigma_{\bar{v}}^2 = \sigma_v^2 + \sigma_{\tilde{e}}^2$. For larger values of μ, using A.3, we obtain

$$\zeta^{\mathrm{LMS}} = \frac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\sigma_{\bar{v}}^2 \mathrm{Tr}(\mathbf{R}^q)}{2 - \mu \mathrm{Tr}(\mathbf{R}^q)} \tag{24}$$

For Gaussian white-input signals, the quantized error variance relation can be solved using **A.4.** to yield

$$\zeta^{\mathrm{LMS}} = \frac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\sigma_{\bar{v}}^2 \mathrm{Tr}(\mathbf{R}^q)}{2 - \mu(N + \lambda)\sigma_u^2} \quad (\text{Gaussian}) \tag{25}$$

where $\sigma_{u^q}^2 = \sigma_u^2 + \sigma_d^2$.

The same results were obtained in the literature by analysing the finite precision LMS recursion (see, e.g. References [9, 13]). This can require some tedious algebra. Here, we see that instead of solving the finite precision energy relation (17), we used the solution of the infinite precision energy relation (18), replaced $\{\mathbf{u}_i, \ \ v(i)\}$ by $\left\{\mathbf{u}_i^q, \bar{v}(i)\right\}$, and added the two terms Tr($\mathbf{M}$) and $\mu^2\sigma_\eta^2\mathrm{Tr}(\mathbf{R}^q)$ to the RHS. Solving for $\zeta = E\left(|e_a(i)^2|\right)$, we obtained the quantized results. We may add that most of these steps can be done just by inspection.

Moreover, we can see that, unlike the infinite precision case, the EMSE is not a monotonically increasing function of μ. In the finite precision case there exists an optimum step-size (μ_0), which minimizes the EMSE. This value is given by

$$\mu_0^{\mathrm{LMS}} = \frac{1}{\sigma_{\bar{v}}}\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\mathrm{Tr}(\mathbf{R}^q)}}$$

The corresponding minimum achievable EMSE ζ_0^{LMS} is given, from (20), by

$$\left(\zeta_0^{\mathrm{LMS}}\right) = \sigma_{\bar{v}}\sqrt{\mathrm{Tr}(\mathbf{M})\mathrm{Tr}(\mathbf{R}^q)}$$

The same expressions were obtained in Reference [13].

4.2. The NLMS algorithm

For the normalized NLMS algorithm,

$$f_e(i) = e(i)/\|\mathbf{u}_i\|^2$$

In this case, relation (18) and assumption A.1 lead to the equality

$$(2\mu - \mu^2)E\left(\frac{|e_a(i)|^2}{\|\mathbf{u}_i\|^2}\right) = \mu^2\sigma_v^2 E\left(\frac{1}{\|\mathbf{u}_i\|^2}\right) \tag{26}$$

Again this is an exact equality. We consider two cases.

(1). Under assumption A.3, we have

$$E\left(\frac{|e_a(i)|^2}{\mu^2\|\mathbf{u}_i\|^2}\right) = E\left(|e_a(i)|^2\right)\ E\left(\frac{1}{\mu^2\|\mathbf{u}_i\|^2}\right)$$

Thus, the solution of (26) becomes

$$(2\mu - \mu^2)E|e_a(i)|^2 E\left(\frac{1}{\|\mathbf{u}_i\|^2}\right) = \mu^2\sigma_v^2 E\left(\frac{1}{\|\mathbf{u}_i\|^2}\right) \tag{27}$$

This result is in fact *exact* for constant modulus data. Now, replacing $\{\mathbf{u}_i,\ \ v(i)\}$ by $\{\mathbf{u}_i^q, \bar{v}(i)\}$, and adding $\mathrm{Tr}(\mathbf{M}) + \mu^2\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)$ to the RHS, the quantized energy relation is given by

$$(2\mu - \mu^2)E|e_a(i)|^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right) = \mathrm{Tr}(\mathbf{M}) + \mu^2\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu^2\sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right) \tag{28}$$

Solving for ζ, we obtain

$$\zeta^{\mathrm{NLMS}} = \frac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu\sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right)}{(2-\mu)E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right)} \tag{29}$$

The values of μ_0^{NLMS} and ζ_0^{LMS} are thus approximately given by

$$\mu_0^{\mathrm{NLMS}} = \sqrt{\mathrm{Tr}(\mathbf{M})/\left[\sigma_d^2 \mathrm{Tr}(\mathbf{R}^q) + \sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right)\right]} \tag{30}$$

$$\zeta_0^{\mathrm{NLMS}} = \sqrt{\mathrm{Tr}(\mathbf{M})/\left[\sigma_d^2 \mathrm{Tr}(\mathbf{R}^q) + \sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right)\right]}\Bigg/E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right) \tag{31}$$

(2). In some works (see, e.g. Reference [3] p. 443), the following approximation is instead used:

$$E\left(\frac{|e_a(i)^2|}{\|\mathbf{u}_i\|^2}\right) \approx \frac{E\left(|e_a(i)^2|\right)}{E\left(\|\mathbf{u}_i\|^2\right)}$$

in which case the solution of (26) becomes

$$\frac{(2\mu - \mu^2)}{\mathrm{Tr}(\mathbf{R})} E\left(|e_a(i)^2|\right) = \mu^2 \sigma_v^2 E\left(\frac{1}{\|\mathbf{u}_i\|^2}\right) \tag{32}$$

Again, replacing $\{\mathbf{u}_i,\ v(i)\}$ by $\{\mathbf{u}_i^q, \bar{v}(i)\}$, and adding $\mathrm{Tr}(\mathbf{M}) + \mu^2\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)$ to the RHS, the quantized energy relation is given by

$$\frac{(2\mu - \mu^2)}{\mathrm{Tr}(\mathbf{R}^q)} E\left(|e_a(i)^2|\right) = \mathrm{Tr}(\mathbf{M}) + \mu^2\sigma_d^2 \mathrm{Tr}(\mathbf{R}^q) + \mu^2\sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right) \tag{33}$$

Solving for ζ, we obtain

$$\zeta^{\mathrm{NLMS}} = \frac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu\sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right)}{(2-\mu)} \mathrm{Tr}(\mathbf{R}^q) \tag{34}$$

Here ζ_0^{NLMS} is still given by (30), while ζ_0^{NLMS} is now given by

$$\zeta_0^{\mathrm{NLMS}} = \mathrm{Tr}(\mathbf{R}^q)\sqrt{\mathrm{Tr}(\mathbf{M})\left[\sigma_d^2 \mathrm{Tr}(\mathbf{R}^q) + \sigma_{\bar{v}}^2 E\left(\frac{1}{\|\mathbf{u}_i^q\|^2}\right)\right]} \tag{35}$$

The finite precision analysis of the NLMS was previously done in Reference [12] in a less direct way and under stronger assumptions.

4.3. The LMF and LMMN algorithms

For the least-mean mixed-norm (LMMN) algorithm with real-valued data (the case of complex-valued data is considered further ahead towards the end of this section), we have [23]:

$$f_e(i) = \delta e(i) + (1-\delta)e^3(i)$$

The least-mean fourth (LMF) algorithm corresponds to the special case $\delta = 0$ [24]. Introduce, for compactness of notation,

$$\bar{\delta} = 1 - \delta, \quad E\left(|v(i)|^4\right) = \xi_4^v, \quad E\left(|v(i)^6|\right) = \xi_6^v$$

By ignoring higher-order terms and by using A.1, the energy equation (18) implies that

$$2\mu b\zeta^{\mathrm{LMMN}} = \mu^2 a\mathrm{Tr}(\mathbf{R}) + \mu^2 cE\left(\|\mathbf{u}_i\|^2|e_a(i)|^2\right) \tag{36}$$

where we introduced the constants

$$a = \delta^2\sigma_v^2 + 2\delta\bar{\delta}\xi_4^v + \bar{\delta}^2\xi_6^v \tag{37}$$

$$b = \delta + 3\bar{\delta}\sigma_v^2 \tag{38}$$

$$c = \delta^2 + 12\delta\bar{\delta}\sigma_v^2 + 15\bar{\delta}\xi_4^v \tag{39}$$

We again consider three cases:

(1). For values of μ that are small enough so that the term $\mu^2 cE\left(\|\mathbf{u}_i\|^2|e_a(i)|^2\right)$ could be ignored, Equation (36) becomes

$$2\mu b\zeta^{\mathrm{LMMN}} = \mu^2 a\mathrm{Tr}(\mathbf{R})$$

Replacing $\{\mathbf{u}_i,\ v(i)\}$ by $\{\mathbf{u}_i^q, \bar{v}(i)\}$, and adding $\mathrm{Tr}(\mathbf{M}) + \mu^2\left(2+\bar{\delta}^2\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)$ to the RHS, the quantized energy relation is given by

$$2\mu b\zeta^{\mathrm{LMMN}} = \mathrm{Tr}(\mathbf{M}) + \mu^2\left(2+\bar{\delta}^2\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu^2 a\ \ \mathrm{Tr}(\mathbf{R}^q)$$

where the constants a, b, and c are now defined by

$$a = \delta^2\sigma_{\bar{v}}^2 + 2\delta\bar{\delta}\xi_4^{\bar{v}} + \bar{\delta}^2\xi_6^{\bar{v}} \tag{40}$$

$$b = \delta + 3\bar{\delta}\sigma_{\bar{v}}^2 \tag{41}$$

$$c = \delta^2 + 12\delta\bar{\delta}\sigma_{\bar{v}}^2 + 15\bar{\delta}\xi_4^{\bar{v}} \tag{42}$$

$\xi_{\bar{v}}^4 = E\left(|\bar{v}(i)|^4\right)$, and $\xi_{\bar{v}}^6 = E\left(|\bar{v}(i)|^6\right)$. Solving for ζ, we get

$$\zeta^{\mathrm{LMMN}} = \frac{1}{2b}\left[\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\left(2+\bar{\delta}^2\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu a\ \ \mathrm{Tr}(\mathbf{R}^q)\right] \quad (\text{small}\ \ \mu) \tag{43}$$

The values of μ_0^{LMMN} and ζ_0^{LMMN} are thus given by

$$\mu_0^{\mathrm{LMMN}} = \sqrt{\mathrm{Tr}(\mathbf{M})/\left[\left(2+\bar{\delta}^2\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + a\ \ \mathrm{Tr}(\mathbf{R}^q)\right]} \tag{44}$$

$$\zeta_0^{\mathrm{LMMN}} = \frac{1}{b}\sqrt{\mathrm{Tr}(\mathbf{M})\left[\left(2+\bar{\delta}^2\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + a\ \ \mathrm{Tr}(\mathbf{R}^q)\right]} \tag{45}$$

For $\delta = 0$, the above expression collapses to

$$\zeta^{\mathrm{LMF}} = \frac{1}{6\sigma_v^2}\left[\mu^{-1}\mathrm{Tr}(\mathbf{M}) + 3\mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu\xi_6^{\bar{v}}\mathrm{Tr}(\mathbf{R}^q)\right] \quad (\text{small}\ \ \mu) \tag{46}$$

Corresponding values of μ_0^{LMF} and ζ_0^{LMF} are given by

$$\mu_0^{\mathrm{LMF}} = \sqrt{\mathrm{Tr}(\mathbf{M})/\left[3\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \xi_6^{\bar{v}}\mathrm{Tr}(\mathbf{R}^q)\right]} \tag{47}$$

$$\zeta_0^{\mathrm{LMF}} = \frac{1}{3\sigma_v^2}\sqrt{\mathrm{Tr}(\mathbf{M})/\left[3\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \xi_6^{\bar{v}}\mathrm{Tr}(\mathbf{R}^q)\right]} \tag{48}$$

Note that we did not need the independence assumption (A.4) to obtain these results.

(2). For larger values of μ, using A.3 again, and following the same procedure, we get the following new expressions for the EMSE:

$$\zeta^{\text{LMMN}} = \frac{\mu^{-1}\text{Tr}(\mathbf{M}) + \mu\left(2+\bar{\delta}^2\right)\sigma_d^2\text{Tr}(\mathbf{R}^q) + \mu a \;\; \text{Tr}(\mathbf{R}^q)}{2b - \mu c\text{Tr}(\mathbf{R}^q)} \quad (\text{large} \;\; \mu) \tag{49}$$

$$\zeta^{\text{LMF}} = \frac{\mu^{-1}\text{Tr}(\mathbf{M}) + 3\mu\sigma_d^2\text{Tr}(\mathbf{R}^q) + \mu\xi_6^{\tilde{v}} \;\; \text{Tr}(\mathbf{R}^q)}{6\sigma_{\tilde{v}}^2 - 15\mu\xi_4^{\tilde{v}}\text{Tr}(\mathbf{R}^q)} \quad (\text{large} \;\; \mu) \tag{50}$$

(3). For Gaussian white-input signals $\left(\mathbf{R} = \sigma_u^2\mathbf{I}\right)$, using A.4 instead of A.3, we obtain the following new expressions

$$\zeta^{\text{LMMN}} = \frac{\mu^{-1}\text{Tr}(\mathbf{M}) + \mu\left(2+\bar{\delta}^2\right)\sigma_d^2\text{Tr}(\mathbf{R}^q) + \mu\text{M}\sigma_{u^{q^a}}^2}{2b - \mu(\text{N}+2)\sigma_{u^{q^c}}^2)} \quad (\text{Guassian}) \tag{51}$$

and

$$\zeta^{\text{LMF}} = \frac{\mu^{-1}\text{Tr}(\mathbf{M}) + 3\mu\sigma_d^2\text{Tr}(\mathbf{R}^q) + \mu\text{M}\sigma_{u^q}^2\xi_6^{\tilde{v}}}{6\sigma_{\tilde{v}}^2 - 15\mu(\text{N}+2)\sigma_{u^q}^2\xi_4^{\tilde{v}}} \quad (\text{Guassian}) \tag{52}$$

where $\sigma_{u^q}^2 = \sigma_u^2 + \sigma_d^2$.

For the case of complex-valued data, we replace e^3 by $e\,|\,e\,|^2$ and assume the noise is circular, i.e., $E(v^2(i)) = 0$. Then repeating the above arguments we find that the three expressions (43), (49), and (51) are still valid but with b and c replaced by

$$b' = \delta + 2\bar{\delta}\sigma_v^2$$
$$c' = \delta^2 + 8\delta\bar{\delta}\sigma_v^2 + 9\bar{\delta}\xi_4^v$$

Corresponding expressions for the LMF algorithm can be obtained by setting $\delta = 0$. Here, we may also add that more precise values for μ_0 and ζ_0 can be obtained by minimizing the more general expressions for ζ over μ.

Figure 3 compares the simulation and theoretical results of the steady-state MSE of the LMMN algorithm, with $\delta = 0.5$, for a large range of μ and two values of the wordlength. In the simulations, the unknown system weight vector $\mathbf{w}^0$ is of length 10 and the elements of the input vector, $\mathbf{u}_i$, are white Gaussian of unit variance. The plant noise is chosen to be a linear combination of normally and uniformly distributed independent random variables of variances $\sigma_\text{n}^2 = 10^{-6}$ and $\sigma_\text{u}^2 = 10^{-2}/12$, respectively. Each simulation result is the steady state statistical average of 50 runs, with up to 20,000 iterations in each run. We can see from the figure that the theoretical and experimental MSE are in good match.

4.4. The sign algorithm

For the sign algorithm (SA), we have

$$f_e(i) = \text{sign}[e(i)]$$

In this case, relation (18) leads to the equality:

$$2\mu E[e_a(i)\;\text{sign}(e_a(i) + v(i))] = \mu^2\text{Tr}(\mathbf{R}) \tag{53}$$

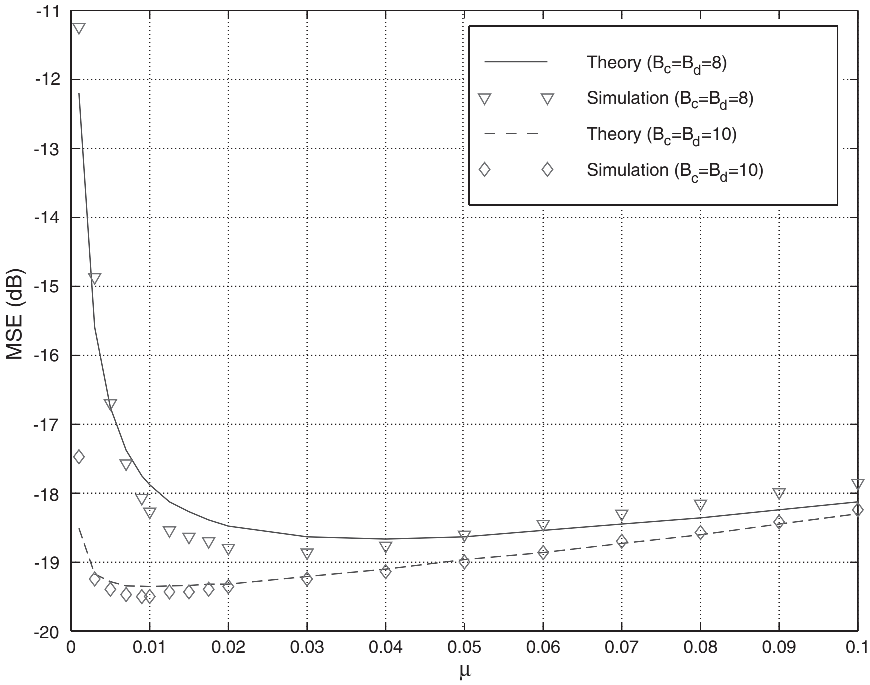

Figure 3. Theory and simulation MSE for the LMMN vs μ.

By assuming that $e(i)$ and $v(i)$ are real-valued jointly Gaussian [25], and by using A.1 and Price's theorem[§§] 7 [26], we obtain

$$E[e_a(i)\ \mathrm{sign}(e_a(i)+v(i))] = \sqrt{\frac{2}{\pi}}\frac{E\big(|e_a(i)|^2\big)}{\sqrt{\sigma_v^2+E\big(|e_a(i)^2|\big)}}$$

Substituting into (53), we get [27]

$$\mu\sqrt{\frac{8}{\pi}}\frac{E\big(|e_a(i)|^2\big)}{\sqrt{\sigma_v^2+E\big(|e_a(i)|^2\big)}} = \mu^2\mathrm{Tr}(\mathbf{R})$$

Using the assumption that the quantization of the estimation error does not introduce any errors in its sign [14], replacing $\mathbf{u}_i$ by $\mathbf{u}_i^q$, and adding Tr(M) to the RHS, the quantized energy relation is given by

$$\mu\sqrt{\frac{8}{\pi}}\frac{E\big(|e_a(i)|^2\big)}{\sqrt{\sigma_v^2+E\big(|e_a(i)|^2\big)}} = \mathrm{Tr}(\mathbf{M})+\mu^2\mathrm{Tr}(\mathbf{R}^q)$$

[§§] For two jointly Gaussian real-valued random variables x and y, we have: $E\,(x\ \mathrm{sign}(y)) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma_v}E(xy)$.

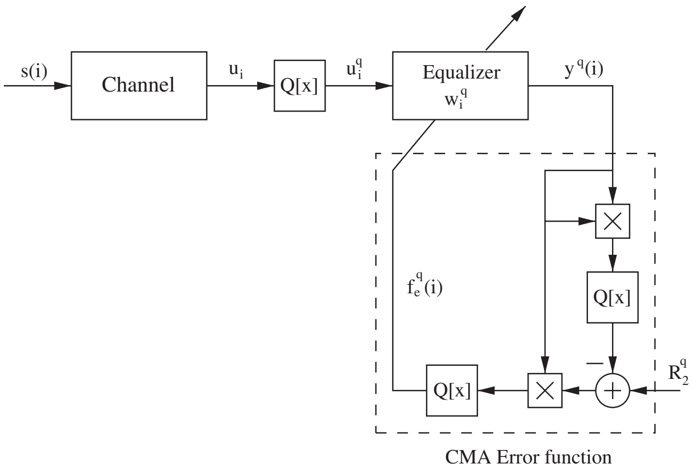

Figure 4. CMA quantization model.

Solving for $E\left(|e_a(i)|^2\right)$, we find that

$$\zeta^{\mathrm{SA}} = \frac{\alpha}{2}\left(\alpha + \sqrt{\alpha^2 + 4\sigma_v^2}\right) \tag{54}$$

where $\alpha = \sqrt{\frac{\pi}{8}}\left(\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\mathrm{Tr}(\mathbf{R}^q)\right)$ Corresponding values of μ_0^{SA} and ζ_0^{SA} are given by

$$\mu_0^{\mathrm{SA}} = \sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\mathrm{Tr}(\mathbf{R}^q)}} \tag{55}$$

$$\zeta_0^{\mathrm{SA}} = \frac{\pi}{4}\mathrm{Tr}(\mathbf{M})\mathrm{Tr}(\mathbf{R}^q) + \left[\frac{\pi}{2}\sigma_v^2\mathrm{Tr}(\mathbf{M})\mathrm{Tr}(\mathbf{R}^q) + \frac{\pi^2}{16}\mathrm{Tr}^2(\mathbf{M})\mathrm{Tr}^2(\mathbf{R}^q)\right]^{1/2} \tag{56}$$

These results are the same expressions obtained in Reference. [14] by additionally using the independence assumptions.

4.5. The CMA

We now study the finite precision performance of the well-known constant modulus algorithm (CMA), whose error function is given by [28]

$$f_e(i) = y(i)\left[R_2 - |y(i)^2\right] \tag{57}$$

In this case, we use a modified version of the quantization model used in the previous section. This model is shown in Figure 4. The quantized error function of the CMA is given, from (57), as

$$\begin{aligned} f_e^q(i) &= Q\left[y^q(i)\left[R_2^q - Q\left[|y^q(i)|^2\right]\right]\right] \\ &= y^q(i)\left[R_2^q - |y^q(i)|^2 + e_1(i)\right] + e_2(i) \end{aligned} \tag{58}$$

where $e_1(i)$ and $e_2(i)$ are two quantization errors of variance $\sigma_1^2 = \sigma_2^2 = \sigma_d^2$. Furthermore, we now define the a priori and a posteriori estimation errors by

$$e_a(i) = s(i-D)e^{j\theta} - y^q(i) = \mathbf{u}_i^q \mathbf{w}_i^0 - \mathbf{u}_i^q \mathbf{w}_i = \mathbf{u}_i^q \tilde{\mathbf{w}}_i$$

$$e_p(i) = \mathbf{u}_i^q(\tilde{\mathbf{w}}_{i+1} - \mathbf{m}_i)$$

Unfortunately, for the case of the CMA, we cannot express $f_e(i)$ in the form given by (16). Thus, we need to solve the quantized energy relation (14) for the CMA recursion. For mathematical tractability of the analysis, we impose the following two reasonable assumptions in steady-state ($i \to \infty$) —for more motivation and explanation on these two assumptions, see References [8, 29]:

A.5. The transmitted signal $s(i-D)$ and the estimation error $e_a(i)$ are independent in steady-state so that $E(s^*(i-D)e_a(i)) = 0$, since $s(i-D)$ is assumed zero mean.

A.6. The scaled regressor energy $\mu^2\|\mathbf{u}_i\|^2$ is independent of $y^q(i)$ in steady-state.

We consider first the case of real-valued data $\{s(\cdot),\ \ y^q(\cdot),\ \ \mathbf{u}_i\}$. In this case, we can assume that the zero forcing response (i.e., the convolution of the channel and the equalizer) h_D that the adaptive equalizer attempts to achieve can be of either form $h_D = \pm[0,\ \ldots,\ 0,\ 1,\ 0,\ \ldots,\ 0]$.

In the following, we continue with the choice $h_D = [0,\ \ldots,\ 0,\ 1,\ 0,\ \ldots,\ 0]$, which yields $e_a(i) = s(i-D) - y^q(i)$. A similar analysis holds for the case $h_D = [0,\ \ldots,\ 0,\ -1,\ 0,\ \ldots,\ 0]$.

Substituting (58) into (14), we obtain

$$\begin{aligned} E(\bar{\mu}(i)|e_a(i)|^2) = {} & \mathrm{Tr}(\mathbf{M}) \\ & + E\left(\bar{\mu}(i)\left|e_a(i) - \frac{\mu}{\bar{\mu}(i)}\big(y^q(i)\big[R_2^q - \big(y^q(i)\big)^2 + e_1(i)\big)\big] + e_2(i)\big)\right|^2\right) \end{aligned} \tag{59}$$

We write more compactly

$$e_a \triangleq e_a(i),\quad \bar{\mu} \triangleq \bar{\mu}(i),\quad y \triangleq y^q(i),\quad \mathbf{u}^q \triangleq \mathbf{u}_i^q,\quad s \triangleq s(i-D),\quad e_1 \triangleq e_1(i),\quad e_2 \triangleq e_2(i)$$

for $i \to \infty$, so that (59) becomes, after expanding,

$$2\mu E(e_a y\ \ [R_2^q - y^2 + e_1]\ \ +\ \ e_a e_2) = \mathrm{Tr}(\mathbf{M}) + \mu^2 E\big((\|\mu^q\|^2\big(y\big[R_2^q - y^2 + e_1\big] + e_2)^2\big)\big)$$

Using this equality we can now obtain an expression for the steady-state MSE, $E\left(e_a^2\right)$. Replacing y by $s - e_a$, using assumptions **A.1**, **A.2**, **A.5**, and A.6 and neglecting $2\mu E\left(e_a^4\right)$ for sufficiently small μ and small e_a^2, it is straightforward to show that the steady-state MSE can be approximated by

$$\zeta^{\mathrm{CMA}} \approx \frac{\mathrm{Tr}(\mathbf{M})/\mu + \mu E\left(s^2 R_2^{q2} - 2R_2^q s^4 + s^2\sigma_1^2 + s^6 + \sigma_2^2\right)E\left(\|\mathbf{u}^q\|^2\right)}{2E\left(3s^2 - R_2^q\right)}$$

This result implies that the steady-state MSE is composed of two terms. The first term decreases with μ and increases with the multiplication error variance $\mathrm{Tr}(\mathbf{M})$. The second term increases with μ and the received signal variance, $E\left(\|\mathbf{u}^q\|^2\right)$. Thus, unlike the stationary case (see, e.g. References [8, 29]), the steady-state MSE is not a monotonically increasing function of μ. We can also see that in the noiseless case, and for non-constant modulus data $\{s(\cdot)\}$, there exists a finite optimal value of the step size, μ_0, that minimizes the above expression for the steady-state

MSE, which is given by

$$\mu_0^{\mathrm{CMA}} = \sqrt{\mathrm{Tr}(\mathbf{M})/\left[E\left(s^2R_2^q - 2R_2^q s^4 + s^2\sigma_1^2 + s^6 + \sigma_2^2\right)E\left(\|\mathbf{u}^q\|^2\right)\right]}$$

where $E\left(\|\mathbf{u}^q\|^2\right) = E\left(\mathbf{u}_{\boldsymbol{i}}^{q*}\mathbf{u}_{\boldsymbol{i}}^q\right) = E\left(\|\mathbf{u}_i\|^2\right) + N\sigma_d^2$. This expression shows that μ_0 decreases with the signal variance, $E\left(\|\mathbf{u}^q\|^2\right)$, and increases with the multiplication error variance Tr($\mathbf{M}$). The corresponding minimum value of the steady-state MSE is then given by

$$\zeta_0^{\mathrm{CMA}} = \frac{\sqrt{\mathrm{Tr}(\mathbf{M})E\left(s^2R_2^{q2} - 2R_2^q s^4 + s^2\sigma_1^2 + s^6 + \sigma_2^2\right)E\left(\|\mathbf{u}^q\|^2\right)}}{E\left(3s^2 - R_2^q\right)}$$

Here, we may add that for complex-valued data, the steady-state MSE will have a different expression than that in the real-valued case. Following the same derivation, and assuming signal constellations that satisfy the circularity condition $E\left(s^2(i)\right) = 0$, in addition to the condition $E\left(2|s(i)|^2 - R_2\right) > 0$ (both of which hold for most constellations [28]), we can show that the steady-state MSE for complex-valued data, and for sufficiently small step-sizes, can be approximated by

$$\zeta^{\mathrm{CMA}} \approx \frac{\sqrt{\mathrm{Tr}(\mathbf{M})/\mu + \mu E\left(|s|^2R_2^{q2} - 2R_2^q|s|^4 + |s|^2\sigma_1^2 + |s|^6 + \sigma_2^2\right)E\left(\|\mathbf{u}^q\|^2\right)}}{2E\left(2|s|^2 - R_2^q\right)}$$

In this case, the optimum value of the algorithm step size still has the same value as in the real-valued data case, while the minimum achievable steady-state MSE is given by

$$\zeta_0^{\mathrm{CMA}} = \frac{\sqrt{\mathrm{Tr}(\mathbf{M})E\left(|s|^2R_2^{q2} - 2R_2^q|s|^4 + |s|^2\sigma_1^2 + |s|^6 + \sigma_2^2\right)E\left(\|\mathbf{u}^q\|^2\right)}}{E\left(2|s|^2 - R_2^q\right)}$$

Finally, we may add that, for the infinite precision case $\left(\sigma_c^2 = \sigma_d^2 = 0\right)$, the expressions for the steady-state MSE reduce to the expressions obtained in References [8, 29].

We now provide some simulation results that compare the experimental performance with the one predicted by the derived expressions. The channel considered in this simulation is given by c = [0.1; 0.3, 1, −0.1, 0.5, 0.2]. A 4-tap FIR filter is used as a $\frac{T}{2}$-fractionally spaced quantized equalizer, with $Bc = Bd = 8$, and 9. In this simulation, the transmitted signal was 6-PAM, $s(i) \in \{1, 0.6, 0.2, -0.2, -0.6, -1\}$ with $E\left(s^6\right) = 0.3489$, $E\left(s^4\right) = 0.3771$, $E\left(s^2\right) = 0.4667$, and $R_2 = 0.808$. The value of $\|\mathbf{u}_i\|^2$ is the norm of the received signal vector. The value of $E\left(\|\mathbf{u}_i\|^2\right)$ was computed as the average over 10,000 realizations of $\|\mathbf{u}_i\|^2$. The value of experimental MSE was obtained as the average over 100 repeated runs. Figures 5 and 6 are plots of the experimental MSE and the theoretical MSE versus the step-size μ for $B_c = B_d = 8$ and 9 bits, respectively. It can be seen from the figure that the theoretical results reasonably match the experimental results. We can also see that, for $B_c = B_d = 8$ bits, the experimental MSE reaches a minimum value of −30.13 dB, which corresponds to an optimal value of μ equal to 1.5×10^{-2}, while our theory predicted a minimum achievable MSE of −30.38 dB at $\mu_0 = 0.94 \times 10^{-2}$. For $B_c = B_d = 9$ bits, the experimental MSE reaches a minimum value of −32.11 dB, which corresponds to an optimal value of μ equal to 10^{-2}. On the other hand, our theory predicted a minimum achievable MSE of -33.38 dB at $\mu_0 = 0.47 \times 10^{-2}$.

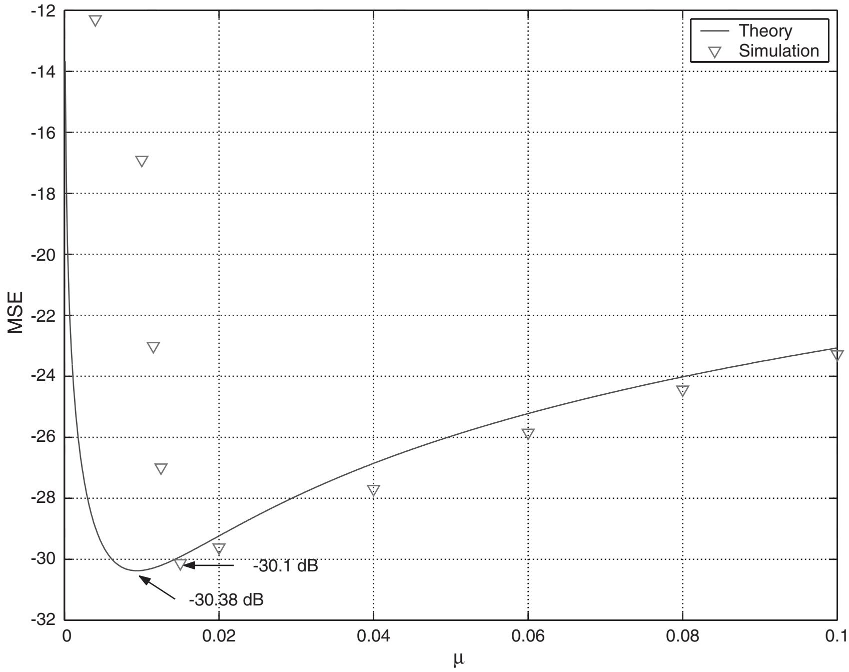

Figure 5. Theoretical and simulation MSE of CMA for $B_c = B_d = 8$.

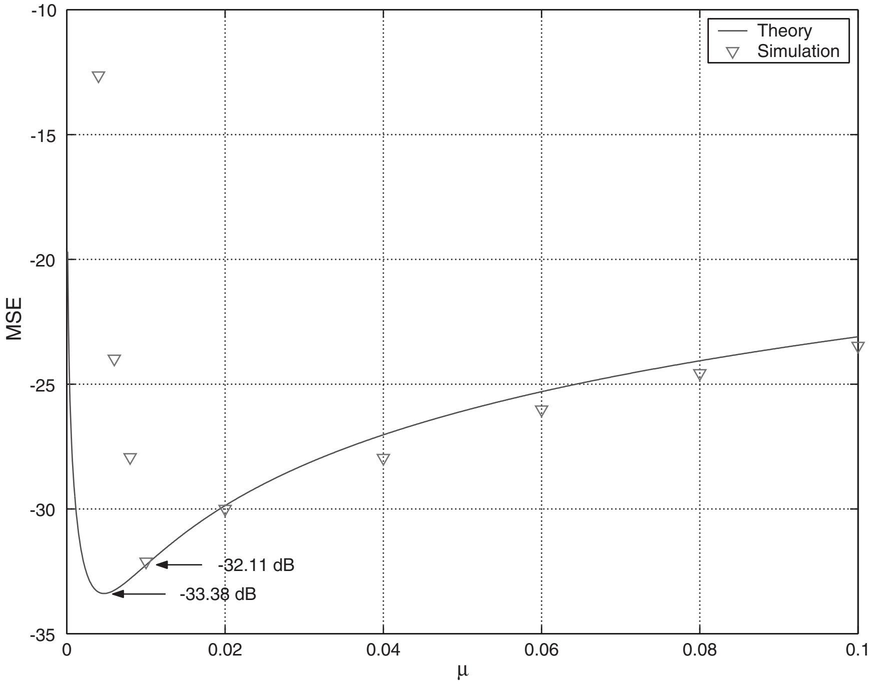

Figure 6. Theoretical and simulation MSE of CMA for $B_c = B_d = 9$.

Table II. Expressions for EMSE in a quantized environment.

Algorithm	EMSE
LMS (small μ)	$\frac{1}{2}\mathrm{Tr}(\mathbf{M})/\frac{\mu}{2}\sigma_v^2\mathrm{Tr}(\mathbf{R}^q)$
NLMS	$\dfrac{\mathrm{Tr}(\mathbf{M})/\mu + \mu\sigma_v^2\mathrm{Tr}(\mathbf{R}^q)}{2-\mu\mathrm{Tr}(\mathbf{R}^q)}$
LMF (real, small μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu\sigma_v^2 E\left(\frac{1}{\lVert\mathbf{u}_i^q\rVert^2}\right)}{2-\mu\mathrm{Tr}(\mathbf{R}^q)}$
LMF (real, large μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + 3\mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu\xi_6^{\tilde{v}}\;\;\mathrm{Tr}(\mathbf{R}^q)}{6\sigma_{\bar{v}}^2 - 15\mu\xi_4^{\tilde{v}}\mathrm{Tr}(\mathbf{R}^q)}$
LMF (complex, small μ)	$\frac{1}{2}\mathrm{Tr}(\mathbf{M})/\mu + \frac{\mu}{2}\left(\dfrac{\xi_{\tilde{v}}^6 + 3\sigma_d^2}{2\sigma_{\tilde{v}}^2}\right)\mathrm{Tr}(\mathbf{R}^q)$
LMF (complex, large μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + 3\mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu\xi_6^{\tilde{v}}\;\;\mathrm{Tr}(\mathbf{R}^q)}{4\sigma_{\bar{v}}^2 - 9\mu\xi_4^{\tilde{v}}\mathrm{Tr}(\mathbf{R}^q)}$
LMMN (real, small μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu(2+\bar{\delta})\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu a\;\;\mathrm{Tr}(\mathbf{R}^q)}{2b}$
LMMN (real, large μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + 2\mu\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu a\;\;\mathrm{Tr}(\mathbf{R}^q)}{2b-\mu c\mathrm{Tr}(\mathbf{R}^q)}$
LMMN (complex, small μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu(2+\bar{\delta})\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu a'\;\;\mathrm{Tr}(\mathbf{R}^q)}{2b'}$
LMMN (complex, large μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu(2+\bar{\delta})\sigma_d^2\mathrm{Tr}(\mathbf{R}^q) + \mu a'\;\;\mathrm{Tr}(\mathbf{R}^q)}{2b'-\mu c'\mathrm{Tr}(\mathbf{R}^q)}$
SA	$\dfrac{\alpha}{2}\left(\alpha + \sqrt{\alpha^2 + 4\sigma_v^2}\right)$
CMA (real, small μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M})/\mu + \mu E\left(s^2R_2^{q2} - 2R_2^q s^4 + s^2\sigma_1^2 + s^6 + \sigma_2^2\right)E\left(\lVert\mathbf{u}^q\rVert^2\right)}{2E\left(3s^2 - R_2^q\right)}$
CMA (complex, small μ)	$\dfrac{\mu^{-1}\mathrm{Tr}(\mathbf{M}) + \mu E\left(\lvert s\rvert^2R_2^{q2} - 2R_2^q\lvert s\rvert^4 + \lvert s\rvert^2\sigma_1^2 + \lvert s\rvert^6 + \sigma_2^2\right)E\left(\lVert\mathbf{u}^q\rVert^2\right)}{2E\left(2\lvert s\rvert^2 - R_2^q\right)}$

Here, we note that the experimental value results validate that the steady-state MSE is not a monotonically increasing function of μ, as predicted by our analytical results. Furthermore, the experimental values of the minimum achievable MSE match reasonably well the analytical values. Thus, the derived results for the minimum MSE can be reliable in predicting the best steady-state performance, which the CMA can achieve for a given word-length. However, the

Table III. Optimum algorithm step-size and minimum EMSE.

Alg.	μ_0	ζ
LMS	$\frac{1}{\sigma_{\bar{v}}}\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\mathrm{Tr}(\mathbf{R}^q)}}$	$\sigma_{\bar{v}}\sqrt{\mathrm{Tr}(\mathbf{M})\mathrm{Tr}(\mathbf{R}^q)}$
NLMS	$\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)+\sigma_{\bar{v}}^2 E\left(\frac{1}{\lVert\mathbf{u}_i^q\rVert^2}\right)}}$	$\frac{\sqrt{\mathrm{Tr}(\mathbf{M})\left[\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)+\sigma_{\bar{v}}^2 E\left(\frac{1}{\lVert\mathbf{u}_i^q\rVert^2}\right)\right]}}{E\left(\frac{1}{\lVert\mathbf{u}_i^q\rVert^2}\right)}$
LMF	$\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\left[3\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)+\xi_6^{\tilde{v}}\mathrm{Tr}(\mathbf{R}^q)\right]}}$	$\frac{1}{3\sigma_v^2}\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\left[3\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)+\xi_6^{\tilde{v}}\mathrm{Tr}(\mathbf{R}^q)\right]}}$
LMMN	$\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\left[\left(2+\bar{\delta}\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)+a\;\;\mathrm{Tr}(\mathbf{R}^q)\right]}}$	$\frac{1}{b}\sqrt{\mathrm{Tr}(\mathbf{M})\left[\left(2+\bar{\delta}\right)\sigma_d^2\mathrm{Tr}(\mathbf{R}^q)+a\;\;\mathrm{Tr}(\mathbf{R}^q)\right]}$
SA	$\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{\mathrm{Tr}(\mathbf{R}^q)}}$	$\frac{\pi}{4}\mathrm{Tr}(\mathbf{M})\mathrm{Tr}(\mathbf{R}^q)$ $+\left[\frac{\pi}{2}\sigma_v^2\mathrm{Tr}(\mathbf{M})\mathrm{Tr}(\mathbf{R}^q)+\frac{\pi^2}{16}\mathrm{Tr}^2(\mathbf{M})(\mathbf{R}^q)\right]^{1/2}$
CMA	$\sqrt{\frac{\mathrm{Tr}(\mathbf{M})}{E\left(s^2R_2^{q2}-2R_2^q s^4+s^2\sigma_1^2+s^6+\sigma_2^2\right)E\left(\lVert\mathbf{u}^q\rVert^2\right)}}$	$\frac{\mathrm{Tr}(\mathbf{M})E\left(s^2R_2^{q2}-2R_2^q s^4+s^2\sigma_1^2+s^6+\sigma_2^2\right)E\left(\lVert\mathbf{u}^q\rVert^2\right)}{E\left(3s^2-R_2^q\right)}$

experimental values for the optimum step size are lower than the corresponding predicted analytical values. This is due to quantizer underflow effects that were not taken into consideration in our quantization model. Thus, a more conservative (larger) design value for μ_0 should be taken into consideration to account for this effect.

4.6. Summary of results

Table II summarizes the derived expressions for the finite precision steady-state EMSE for several of the algorithms of Table I. For the first three EMSE results in the table, the derivation provided in the article is different from prior approaches in that it relies on energy conservation arguments. In the SA case, the EMSE expression has been obtained without relying on the independence assumption. All other EMSE expressions in the table appear to be new. Table III lists the derived expressions for the optimum step-size of each algorithm and the corresponding minimum achievable EMSE.

5. CONCLUSIONS

In this paper, we developed a framework for the finite precision steady-state analysis of adaptive filtering algorithms that, in our opinion, facilitates the derivation of earlier results and also leads to some new results, especially for adaptive algorithms with complex update equations.

One of the main features of the framework developed herein is that its starting point is the fundamental energy (or variance) relation (15). Comparing this relation to the corresponding infinite precision variance relation (18), we can extend the infinite precision results to the finite precision case with minimal effort for a large class of adaptive algorithms. In the general case, the finite precision variance relation could be solved to arrive at the desired EMSE. We could also find expressions for the optimum algorithm step-size that minimizes the EMSE and the corresponding minimum achievable EMSE, for each algorithm.

REFERENCES

1. Widrow B, Stearns S D. *Adaptive Signal Processing*. Prentice-Hall: Englewood cliffs, NJ, 1985.
2. Haykin S. *Adaptive Filter Theory (3rd edn)* Prentice-Hall: Englewood cliffs, NJ, 1996.
3. Macchi O. *Adaptive Processing: The LMS Approach with Applications in Transmission*. Wiley, New York, 1995.
4. Widrow B, McCool J, Larimore M G, Johnson C R. Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proceedings of the IEEE*, vol. 46, no. 8, August 1976; 1151–1162.
5. Gitlin R D, Mazo J E, Taylor M G. On the design of gradient algorithms for digitally implemented adaptive filter. IEEE *Transactions on Circuit Theory* 1973; **20:**125–136.
6. Bershad N J. Analysis of the normalized LMS algorithm with Gaussian inputs. *IEEE Transactions of Acoustics, Speech and Signal Processing*, 1986; **34:**793–806.

7 Yousef N R, Sayed A H. A unified approach to the steady-state and tracking analyses of adaptive filters. *IEEE Transactions on Signal Processing* 2001; **49**(2):314–324.

8. Mai J, Sayed A H. A feedback approach to the steady-state performance of fractionally spaced blind adaptive equalizers. *IEEE Transactions on Signal Processing* 200; **48**(1)**:**0–91.
9. Caraiscos C, Liu B. A roundoff error analysis of the LMS algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing* 1984; **32**(1)**:**34–41.
10. Cioffi J M. Limited-precision effects in adaptive filtering. *IEEE Transactions on Circuits and Systems* 1987; **32**(7)**:**821–833, 834–841.
11. Alexander S T. Transient weight misadjustment properties for the finite precision LMS algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing* 1987; **35**(9)**:**1250–1258.
12. Chang P S, Willson A N. A roundoff error analysis of the normalized LMS algorithm. *Proceedings of the 29th Asilomar Conference on Signals, Systems and Computers* vol. 2, October 1995; 1337–1341.
13. Eweda E, Yousef N R, El-Ramly S H. Reducing the effect of finite wordlength on the performance of an LMS adaptive filter. *Proceedings of the IEEE International Conference on Communications* vol. 2, Atlanta, GA, June 1998; 688–692,
14. Eweda E, Younis W M, El-Ramly S H. Tracking performance of a quantized adaptive filter equipped with the sign algorithm. *Signal Processing* 1998; **69**(2):157–162.
15. Sayed A H, Rupp M. A time-domain feedback analysis of adaptive algorithms via the small gain theorem. *Proceedings of the SPIE*, vol. 2563, San Diego, CA, July 1995; 458–469.
16. Rupp M, Sayed A H. A time-domain feedback analysis of filtered-error adaptive gradient algorithms. *IEEE Transactions on Signal Processing* 1996; **44**(6)**:**1428–1439.

17 Sayed A H, Rupp M. Robustness issues in adaptive filtering. *DSP Handbook*, CRC Press: Boca Raton, FL; 1998 (Chapter 20).

18. Yousef N R, Sayed A H. Ability of adaptive filters to track carrier offsets and random channel nonstationarities. *IEEE Transactions on Signal Processing* 2002; **50**(7)**:**1533–1544.
19. Al-Naffouri T Y, Sayed A H. Transient analysis of adaptive filters. *Proceedings of the ICASSP*, Salt Lake City, Utah, May 2001; **6:**3869–3872.
20. Al-Naffouri T Y, Sayed A H. Transient analysis of adaptive filters—Part II: The error nonlinearity case. *Proceedings of the 5th IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Baltimore, MD, June 2001.
21. Bershad N J, Bermudez J C M A. nonlinear analytical model for the quantized LMS algorithm-the power-of-two step size case. *IEEE Transactions on Signal Processing*, 1996; **44**(11)**:**2895–2900.
22. Pfann E, Stewart R. LMS adaptive filtering with *ΣΔ* modulated input signals. *IEEE Signal Processing Letters* 1998; **5**(4)**:**95–97.

23. Chambers J A, Tanrikulu O, Constantindes A G. Least mean mixed-norm adaptive filtering. *Electronics Letters* 1994; **30**(19)**:**1574–1575.
24. Walach E, Widrow B. The least mean fourth (LMF) adaptive algorithm and its family. *IEEE Transactions on Information Theory* 1984; **30**(2)**:**275–283.
25. Mathews V, Cho S. Improved convergence analysis of stochastic gradient adaptive filters using the sign algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 1987; **35**(4)**:**450–454.
26. Price R. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory* 1958; **4:**69–72.
27. Yousef N R, Sayed A H. Steady-state and tracking analyses of the sign algorithm without the explicit use of the independence assumption. *IEEE Signal Processing Letters* 2001; **7**(11)**:**307–309.
28. Godard D N. Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *IEEE Transactions on Communications* 1980; **28**(11)**:**1867–1875.
29. Fijakow I, Manlove C E, Johnson C R. Adaptive fractionally spaced blind CMA equalization: Excess MSE. *IEEE Transactions on Signal Processing* 1998; **46**(1)**:**227–231.