# OPTIMUM ERROR NONLINEARITIES FOR LONG ADAPTIVE FILTERS

TAREQ Y. AL-NAFFOURI[1] AND ALI H. SAYED[2]

[1] Information Systems Lab
Electrical Engineering Department
Stanford University, CA 94305

[2]Electrical Engineering Department
University of California, Los Angeles, CA 90095

## ABSTRACT

In this paper, we consider the class of adaptive filters with error nonlinearities. In particular, we derive an expression for the optimum nonlinearity that minimizes the steady-state error and attains the limit mandated by the Cramer-Rao bound of the underlying estimation process.

## 1. INTRODUCTION

The least-mean-squares (LMS) algorithm is a popular adaptive algorithm because of its simplicity and robustness. Many LMS-like algorithms have been suggested and analyzed in the literature with the aim of retaining the desirable properties of LMS and simultaneously offsetting some of its limitations. Of particular importance is the class of least-mean-squares algorithms with error nonlinearities. Some of the most common nonlinearities are tabulated in Table 1.

Our aim in this paper is to derive an expression for the optimum nonlinearity that minimizes the steady-state mean-square error. We arrive at this nonlinearity by first deriving a closed form expression for the steady-state error for a general error nonlinearity. Subsequently, we choose the optimum nonlinearity which reduces the steady-state error to that mandated by the Cramer-Rao bound of the underlying estimation process.

## 2. ADAPTIVE ALGORITHMS WITH ERROR NONLINEARITY

An adaptive filter in a system identification setting uses input regressor (row) vectors $\mathbf{u}_i$ and noisy output data $d(i) = \mathbf{u}_i w^o + v(i)$ to estimate the unknown (column) vector $w^o$. Many adaptive filters are special cases of the following general class of algorithms:

$$w_{i+1} = w_i + \mu f[e(i)]\mathbf{u}_i^T, \quad i \geq 0 \tag{1}$$

The work of T. Y. Al-Naffouri was partially supported by a fellowship from King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. The work of A. H. Sayed was also partially supported by the National Science Foundation under award ECS-9820765.

Table 1: *Examples for $f[e(i)]$* .

| ALGORITHM | ERROR NONLINEARITIES $f[e(i)]$ |
|---|---|
| LMS | $e(i)$ |
| LMF | $e^3(i)$ |
| LMF family | $e^{2k+1}(i)$ |
| LMMN | $ae(i) + be^3(i)$ |
| Sign error | $\text{sign}[e(i)]$ |
| Sat. nonlin. | $\int_0^{e(i)} \exp\left(-\frac{z^2}{2\sigma_z^2}\right) dz$ |

where $w_i$ is the estimate of $w$ at time $i$, $\mu$ is the step size,

$$e(i) \triangleq d(i) - \mathbf{u}_i w_i = \mathbf{u}_i w^o - \mathbf{u}_i w_i + v(i) \tag{2}$$

is the estimation error, and $f[e(i)]$ is a scalar function of the error $e(i)$. Mean-square analysis of adaptive filters is most conveniently carried out in terms of the weight-error vector $\tilde{w}_i = w^o - w_i$ and the a-priori and a-posteriori errors

$$e_a(i) \triangleq \mathbf{u}_i \tilde{w}_i \quad e_p(i) \triangleq \mathbf{u}_i \tilde{w}_{i+1} \tag{3}$$

The adaptive filtering equations can be reformulated in terms of these quantities as

$$\tilde{w}_{i+1} = \tilde{w}_i - \mu f[e(i)]\mathbf{u}_i^T \tag{4}$$
$$e(i) = e_a(i) + v(i) \tag{5}$$

Using this parameterization, it it straightforward to show that the estimation errors are related by

$$\boxed{e_p(i) = e_a(i) - \mu \|\mathbf{u}_i\|^2 f[e(i)]} \tag{6}$$

and that the energies of the various errors are conserved according to

$$\boxed{\|\tilde{w}_{i+1}\|^2 + \frac{|e_a(i)|^2}{\|\mathbf{u}_i\|^2} = \|\tilde{w}_i\|^2 + \frac{|e_p(i)|^2}{\|\mathbf{u}_i\|^2}} \tag{7}$$

Both of these relationships are exact and apply for the class of adaptive filters (1)-(2). They were derived in [1] and used subsequently as a unifying tool in a variety of studies on different aspects of adaptive filtering (see, e.g., [2, 3] for results on steady-state performance, and [4, 5, 6, 7], where studies on stability and transient performances have been pursued to great advantage).

## 3. STEADY-STATE BEHAVIOR

We start with the following form of the energy relation

$$\begin{aligned} E\left[\|\tilde{\boldsymbol{w}}_{i+1}\|^2\right] &= E\left[\|\tilde{\boldsymbol{w}}_i\|^2\right] - 2\mu E\left[e_a(i)f[e(i)]\right] \\ &\quad +\mu^2 E\left[\|\mathbf{u}_i\|^2 f^2[e(i)]\right] \end{aligned} \quad (8)$$

obtained by averaging (7) and replacing the posteriori error $e_p(i)$ by the equivalent expression (6). Assuming that the filter is stable, it should eventually reach its steady-state wherein $E\left[\|\tilde{\boldsymbol{w}}_{i+1}\|^2\right] = E\left[\|\tilde{\boldsymbol{w}}_i\|^2\right]$ as $i \to \infty$, so that

$$\lim_{i\to\infty} E\left[e_a(i)f[e(i)]\right] = \frac{\mu}{2}\lim_{i\to\infty} E\left[\|\mathbf{u}_i\|^2 f^2[e(i)]\right] \quad (9)$$

To proceed further, we need to evaluate the two expectations in (9). This prompts us to introduce (a realistic) independence assumption on the noise

**AN.** The noise sequence $\{v(i)\}$ is independent, identically distributed, and independent of the input sequence $\{\mathbf{u}_i\}$.

and the following asymptotic assumptions:

**AG.** The filter is long enough such that $e_a(i)$ is Gaussian.

**AU.** The random variables $\|\mathbf{u}_i\|^2$ and $f^2[e(i)]$ are asymptotically uncorrelated, i.e.

$$\lim_{i\to\infty} E\left[\|\mathbf{u}_i\|^2 f^2[e(i)]\right] = E\left[\|\mathbf{u}_i\|^2\right]\lim_{i\to\infty} E\left[f^2[e(i)]\right] \quad (10)$$

**Remark**. Assumptions AG and AU act in harmony in that both get more realistic as the filter gets longer (and hence the title of the paper). Assumption AG is justified for long filters by the central limit theorem. Assumption AU has the same spirit as the independence assumption but is weaker.[1] It is justified for long filters by an ergodic argument on $\|\mathbf{u}_i\|^2$ which then behaves like a scaled second moment of the input.

Now, using AU, we can write

$$\lim_{i\to\infty} E\left[\|\mathbf{u}_i\|^2 f^2[e(i)]\right] = E\left[\|\mathbf{u}_i\|^2\right]\lim_{i\to\infty} E\left[f^2[e(i)]\right]$$

By employing assumptions AN and AG, we can show that the expectations $E[e_a(i)f[e(i)]]$ of (9) and $E[f^2[e(i)]]$ of (11) can both be expressed in terms of the second moment $E[e_a^2(i)]$. This motivates the defining relations

$$\boxed{\begin{aligned} h_U\left[E[e_a^2(i)]\right] &\triangleq E\left[f^2[e(i)]\right] \\ h_G\left[E[e_a^2(i)]\right] &\triangleq E\left[e_a(i)f[e(i)]\right] \end{aligned}} \quad (11)$$

[1]The independence assumption AI states that the input regressors $\{\mathbf{u}_i\}$ form an independent and identically distributed sequence. This assumption is heavily resorted to in the adaptive filtering literature.

and, together with (11), enables us to rewrite (9) as

$$\lim_{i\to\infty} E\left[e_a^2(i)\right] \triangleq \frac{\mu}{2}\mathrm{Tr}\left(\boldsymbol{R}\right)\frac{\lim_{i\to\infty} h_U\left[E[e_a^2(i)]\right]}{\lim_{i\to\infty} h_G\left[E[e_a^2(i)]\right]}$$

Now denote the mean-square error by $S \triangleq \lim_{i\to\infty} E\left[e_a^2(i)\right]$. Since both $h_U$ and $h_G$ are analytic in their arguments, we have

$$\lim_{i\to\infty} h_U\left[E[e_a^2(i)]\right] = h_U[S] \quad \lim_{i\to\infty} h_G\left[E[e_a^2(i)]\right] = h_G[S]$$

This means that the MSE satisfies the nonlinear relationship

$$\boxed{S = \tfrac{\mu}{2}\mathrm{Tr}(R)\tfrac{h_U[S]}{h_G[S]}} \quad (12)$$

For a given error nonlinearity $f$, we can evaluate $h_U$ and $h_G$ and subsequently solve for the MSE (see [6] for specific examples).

## 4. OPTIMUM CHOICE OF THE NONLINEARITY

In this section, we build upon the second-order analysis performed above to optimize the choice of the error nonlinearity $f$. To this end, consider expression (12) for the mean-square written in a more explicit form [2]

$$S = \frac{\mu}{2}\mathrm{Tr}(\boldsymbol{R})\frac{E[f^2[e(i)]]}{E[f'[e(i)]]} \quad (13)$$

The mean-square error is fundamentally lower-bounded by the Cramer-Rao bound $\alpha$ of the underlying estimation process (viz., the problem of estimating the random quantity $\mathbf{u}_i\boldsymbol{w}^o$ by using $\mathbf{u}_i\boldsymbol{w}_i$). We can thus write

$$\frac{E[f^2[e(i)]]}{E[f'[e(i)]]} \geq \frac{2}{\mu\mathrm{Tr}(\boldsymbol{R})}\alpha = \alpha' \quad (14)$$

Now let $p_e$ denote the pdf of $e(i)$. We claim that the nonlinearity

$$\boxed{f[e(i)] = -\alpha'\tfrac{p_e'[e(i)]}{p_e[e(i)]}} \quad (15)$$

attains the lower bound on the MSE and hence is optimum. To see this, let's evaluate the numerator and denominator of (14) for this choice of $f$. Using integration by parts, we can write

$$\begin{aligned} E[f'[e(i)]] &= \int_{-\infty}^{\infty} f'[e(i)]p_e[e(i)]\,de(i) \\ &= f[e(i)]p_e[e(i)]|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f[e(i)]p_e'[e(i)]\,de(i) \end{aligned}$$

For the choice (15) of $f$, this reads

$$\begin{aligned} E[f'[e(i)]] &= -\alpha'\, p_e'[e(i)]|_{-\infty}^{\infty} + \alpha'\int_{-\infty}^{\infty}\frac{(p_e'[e(i)])^2}{p_e[e(i)]}\,de(i) \\ &= \alpha'\int_{-\infty}^{\infty}\frac{(p_e'[e(i)])^2}{p_e[e(i)]}\,de(i) \end{aligned} \quad (16)$$

[2]Assuming $f$ is differentiable, we can show by Price theorem that $h_G$ takes the alternative form $h_G = E[e_a(i)f[e(i)]] = E[f'[e(i)]]$. This, together with (11) and (12), yields (13).

assuming that $p_e'$ decays to zero as $e(i)$ approaches $\pm\infty$. Now for the same choice of $f$, we have

$$E\left[f^2[e(i)]\right] = (\alpha')^2 \int_{-\infty}^{\infty} \left(\frac{p_e'[e(i)]}{p_e[e(i)]}\right)^2 p_e[e(i)]\, de(i)$$

$$= (\alpha')^2 \int_{-\infty}^{\infty} \frac{(p_e'[e(i)]))^2}{p_e[e(i)]}\, de(i)$$

i.e. $\frac{E[f^2[e(i)]]}{E[f'[e(i)]]} = \alpha'$, as required.

### 4.1. Further simplifications

To determine the nonlinearity (15), we need to evaluate $\alpha'$ and determine the pdf $p_e$ at each time instant. Upon substituting the optimum nonlinearity (15) into the adaptation equation (1), the constant $\alpha'$ appears multiplied by the step size $\mu$ – a design parameter that is usually varied; hence, $\alpha'$ can be absorbed into $\mu$. Thus, the optimum nonlinearity simplifies to

$$\boxed{f[e(i)] = -\frac{p_e'[e(i)]}{p_e[e(i)]}} \tag{17}$$

Furthermore, since $e(i)$ is the sum of the independent variables $e_a(i)$ and $v(i)$, its pdf is the convolution of their pdf's, i.e.,

$$\begin{aligned} p_e[e(i)] &= p_{e_a}[e(i)] \star p_v[e(i)] \\ &= \frac{1}{\sqrt{2\pi\sigma_{e_a}^2}} e^{-\frac{e^2}{2\sigma_{e_a}^2}} \star p_v[e(i)] \end{aligned} \tag{18}$$

where the second line follows from the Gaussian assumption AG on $e_a(i)$, and $\sigma_{e_a}^2$ denotes the variance of $e_a(i)$. Thus, modeling $p_e$ reduces to the simpler task of modeling the noise statistics and tracking the time variations of $\sigma_{e_a}^2$. In our simulations, we estimate $\sigma_{e_a}^2$ by first estimating the variance of $e(i)$ using a window of (the 4 most recent) samples of $e(i)$ and subsequently calculate the estimate $\hat{\sigma}_{e_a(i)}^2$ from

$$\hat{\sigma}_{e_a(i)}^2 = \hat{\sigma}_{e(i)}^2 - \sigma_v^2 \tag{19}$$

To avoid malfunctioning of the algorithm, we enforce the assignment $\hat{\sigma}_e^2 = c$ whenever the windowed estimate falls outside the interval $[a, b]$; the three constants $a, b$, and $c$ should be specified by the designer.

#### Remarks

1. The derivation of the optimum nonlinearity blends smoothly with the stability[3] and steady-state analyses in that it relies on the same set of assumptions, and is also obtained as a fallout of the same energy conservation approach.

[3]Refer to [5], [6], and [7] to see how the energy relation can be used to establish stability under the assumptions of this paper. Stability analysis and optimum design share another feature in that they both rely on the fundamental limit set by the Cramer-Rao bound.

2. Also note that no heavy machinery is appealed to in developing the optimum nonlinearity. In particular, we avoid the variational approaches that are usually employed in literature in designing optimum adaptation schemes (see [8, 11]).
3. The nonlinearity (17) is derived under simpler assumptions compared to what is available in literature. For instance, we employ a weaker version of the independence assumption (compare with [8] and [9], [11], [10]) and make no restriction on the color or statistics of the input (compare with [8] and [11]). The nonlinearity (17) also applies irrespective of the noise statistics or whether its pdf is symmetric or not (contrary to what is assumed in [8] and [9]).
4. More importantly, perhaps, we avoid the need for any linearization arguments making the nonlinearity (17) accurate over all stages of adaptation. In contrast, the optimum nonlinearity

$$f[e(i)] = -\frac{p_v'[e(i)]}{p_v[e(i)]} \tag{20}$$

of [9], rederived in [8] using linearization arguments, is only accurate in the final stages of adaptation. In fact, the more accurate expression (17) collapses to the nonlinearity (20) as the filter reaches its steady-state.

5. Notice further that expression (17) for the optimum nonlinearity applies irrespective of whether the noise pdf is smooth enough (differentiable) or not. Thanks to the smoothing convolution operator of (18), we can, for example, directly calculate the optimum nonlinearity for binary and uniform noise (see examples below). This comes contrary to the nonlinearity (20) where an artificial smoothing kernel needs to be employed for such singular cases [8].

## 5. EXAMPLES

In what follows, we show how the error nonlinearity manifests itself for different noise statistics. Due to space limitations, we save the details to [6], and simply write down the form of the nonlinearity.

**Gaussian noise:** $v$ is $\mathcal{N}(0, 1)$:

$$f_{\text{opt}}[e(i)] = -\frac{p_e'[e(i)]}{p_e[e(i)]} = \frac{1}{\sigma_e^2} e(i) \tag{21}$$

**Laplacian noise:** $p_v[v] = \frac{1}{2} e^{-|v|}$:

$$f_{\text{opt}} = -\frac{e^{e(i)} g_+[e(i)] - e^{-e(i)} g_-[e(i)]}{e^{e(i)}\left(1 - \text{erf}\left[\frac{e(i)+\sigma_{e_a}^2}{\sqrt{2\sigma_{e_a}^2}}\right]\right) + e^{-e(i)}\left(1 + \text{erf}\left[\frac{e(i)-\sigma_{e_a}^2}{\sqrt{2\sigma_{e_a}^2}}\right]\right)}$$

where

$$g_\pm[e(i)] = 1 \mp \text{erf}\left[\frac{e(i) \pm \sigma_{e_a}^2}{\sqrt{2\sigma_{e_a}^2}}\right] - \sqrt{\frac{2}{\pi\sigma_{e_a}^2}}\, e^{-\frac{(e(i)\pm\sigma_{e_a}^2)^2}{2\sigma_{e_a}^2}}$$

**Binary noise** $v = \pm 1$ with equal probability:

$$f_{\text{opt}}[e(i)] = \frac{1}{\sigma_{e_a}^2}\left(e(i) - e^{-\left(\frac{e^2(i)+1}{2\sigma_{e_a}^2}\right)}\tanh\left[\frac{e(i)}{\sigma_{e_a}^2}\right]\right) \quad (22)$$

### 5.1. Simulations

Here we use simulations to illustrate the behavior of the optimum algorithm in comparison to the LMS and to the asymptotic algorithm (20). The system to be identified is an FIR channel with 15 taps normalized so that the SNR relative to the input and output is the same (10 dB in our case). The input is taken to be Gaussian while the additive output noise is assumed to be Gaussian or Laplacian. The variance $\sigma_e^2$ is estimated using the most recent 4 samples of $e(i)$, and the estimate is in turn used in (19) to estimate the variance of $e_a(i)$. Whenever the estimate $\hat{\sigma}_e^2$ falls outside the range $[\frac{\sigma_v^2}{2}, 5\sigma_v^2]$ $\left([\frac{3\sigma_v^2}{2}, 6\sigma_v^2]\right)$ in the Gaussian (Laplacian) noise case, we enforce the assignment $\hat{\sigma}_e^2 = \sigma_v^2$ $(\hat{\sigma}_e^2 = 2\sigma_v^2)$ instead. The experiment is averaged over 1000 runs.

The three algorithms are compared (Fig. 1 and Fig. 2) in terms of their learning curves; the evolution of $E[\|\tilde{w}_i\|^2]$ with time (also known as the mean-square deviation or MSD). We also plot the nonlinearities employed by the three algorithms. Since the optimum nonlinearity is *time varying* (through its dependence on $\sigma_{e_a}^2$), it has a stochastic nature. The plots thus show the optimum nonlinearities in their averaged forms.

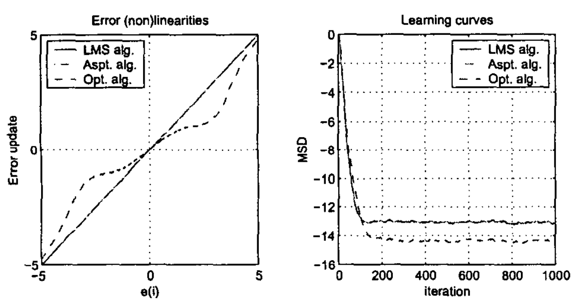


Figure 1: Error updates and learning curves for the LMS, optimum (17), and asymptotically optimum (20) algorithms (Gaussian noise case).

## 6. CONCLUSION

In this paper, we derived an expression for the optimum error nonlinearity in LMS adaptation. Starting from an energy conservation relation, we derived a closed form expression for the steady-state error for a general error nonlinearity and subsequently minimized this expression over the class of smooth nonlinearities. The nonlinearity turns out to be a function of the estimation error and the pdf of additive noise.

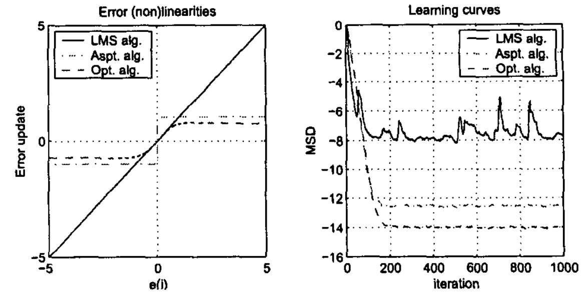


Figure 2: Error updates and learning curves for the LMS, optimum (17), and asymptotically optimum (20) algorithms (Laplacian noise case).

## 7. REFERENCES


[1] A. H. Sayed and M. Rupp, "A time-domain feedback analysis of adaptive algorithms via the small gain theorem," *Proc. SPIE*, vol. 2563, pp. 458–69, San Diego, CA, Jul. 1995.

[2] N. R. Yousef and A. H. Sayed, " A unified approach to the steady-state and tracking analyses of adaptive filters," *IEEE Trans. Signal Processing*, vol. 49, no. 2, pp. 314–324, Feb. 2001.

[3] J. Mai and A. H. Sayed, "A feedback approach to the steady-state performance of fractionally-spaced blind adaptive equalizers," *IEEE Trans. Signal Processing*, vol. 48, no. 1, pp. 80–91, Jan. 2000.

[4] T. Y. Al-Naffouri and A. H. Sayed, "Transient analysis of adaptive filters – Part I: The data nonlinearity case," *Proc. 5th IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Baltimore, Maryland, Jun. 2001.

[5] T. Y. Al-Naffouri and A. H. Sayed, "Transient analysis of adaptive filters – Part II: The error nonlinearity case," *Proc. 5th IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Baltimore, Maryland, Jun. 2001.

[6] T. Y. Al-Naffouri and A. H. Sayed, "Adaptive filters with error nonlinearities: Mean-square analysis and optimum design," to appear in *EURASIP Journal on App. Signal Processing*, Dec. 2001.

[7] T. Y. Al-Naffouri and A. H. Sayed, "Transient analysis of adaptive filters," *Proc. ICASSP*, vol. VI, pp. 3869–3872 Salt Lake City, Utah, May 2001.

[8] S. C. Douglas and T. H. -Y Meng, "Stochastic gradient adaptation under general error criterion," *IEEE Trans. Signal Processing*, vol. 42, no. 6, pp. 1335–1351, Jun. 1994.

[9] B. Polyak and Y. Tsypkin, "Adaptive estimation algorithms (convergence, optimality, stability)," *Avtomatika i Telemekhanika*, no. 3, pp. 71–84, Mar. 1979.

[10] T. Y. Al-Naffouri, A. H. Sayed, and T. Kailath, "On the selection of optimal nonlinearities for stochastic gradient adaptive algorithms," *Proc. ICASSP*, vol. I, pp. 464-467, Istanbul, Turkey, Jun. 2000.

[11] N. J. Bershad, "On the optimum data nonlinearity in LMS adaptation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, pp. 69–76, Feb. 1986.