ON THE SELECTION OF OPTIMAL NONLINEARITIES FOR STOCHASTIC GRADIENT ADAPTIVE ALGORITHMS

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

[1]Electrical Engineering Department
Stanford University, CA 94305

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

ABSTRACT

This paper derives an expression for the optimal error nonlinearity in adaptive filter design. Using an energy conservation relation, and some typical assumptions, the choice of the error function is optimized by minimizing the mean-square deviation subject to a fixed rate of convergence. The resulting optimal choice is shown to subsume earlier results as special cases.

1. INTRODUCTION

An important performance measure of adaptive filtering algorithms is the mean-square-deviation (MSD), which relates to the steady-state covariance matrix of the weight error vector. In this paper we study the problem of optimally designing the nonlinear error function in an adaptive filter update in order to minimize the MSD. This is an issue that has attracted some attention, as can be seen from the titles of the early references [1]–[3]. We derive an expression for the optimal nonlinearity in terms of the probability density function of the noise sequence. The result subsumes earlier expressions as special cases, and it is obtained by relying on a fundamental energy (conservation) relation that leads to simplifications both in the presentation and the derivation (cf. [4]–[6]).

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

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

where $v(i)$ accounts for measurement noise and modeling errors, $\mathbf{u}_i$ denotes a nonzero *row* input (regressor) vector, and $\boldsymbol{w}^o$ is an unknown *column* vector that we wish to estimate. Several algorithms of the stochastic gradient type have been developed for this purpose in the literature. They can be regarded as special cases of the general update scheme

$$\boldsymbol{w}_{i+1} \quad = \quad \boldsymbol{w}_i + \mu\, \mathbf{u}_i^T\, f[e(i)] \ , \tag{2}$$

where $\boldsymbol{w}_i$ is an estimate for $\boldsymbol{w}^o$ at iteration i, μ is the step-size, and $f[e(i)]$ denotes a generic scalar function of the so-called output estimation error, defined by $e(i) = d(i) - \mathbf{u}_i \boldsymbol{w}_i$. Different choices for $f[\cdot]$ result in different adaptive algorithms. Table 1 defines $f[\cdot]$ for many famous special cases of (2).

Table 1: *Examples for* $f[\cdot]$.

ALGORITHM	$f[e(i)]$
LMS	$e(i)$
NLMS	$e(i)/\|\mathbf{u}_i\|^2$
LMF family	$e^{2k+1}(i)$
LMMN	$\delta e(i) + (1-\delta)e^3(i)$
SA	$\text{sign}[e(i)]$

The weight-error vector, $\tilde{\boldsymbol{w}}_i = \boldsymbol{w}^o - \boldsymbol{w}_i$, can be easily seen to satisfy the recursion

$$\begin{aligned} \tilde{\boldsymbol{w}}_{i+1} &= \tilde{\boldsymbol{w}}_i - \mu \mathbf{u}_i^T f[e(i)] && (3) \\ e(i) &= e_a(i) + v(i), && (4) \end{aligned}$$

where $e_a(i)$ denotes the a priori estimation error $e_a(i) = \mathbf{u}_i \tilde{\boldsymbol{w}}_i$. We also define the a posteriori estimation error $e_p(i) = \mathbf{u}_i \tilde{\boldsymbol{w}}_{i+1}$. If we multiply (3) from the left by $\mathbf{u}_i$ then we can readily see that $e_p(i)$ and $e_a(i)$ are related via

$$e_p(i) = e_a(i) - \frac{\mu}{\bar{\mu}(i)} f[e(i)], \tag{5}$$

where we defined $\bar{\mu}(i) = 1/\|\mathbf{u}_i\|^2$.

This work was partially supported by the Army Research Office DAAH04-96-1-0176-P00005 and by the National Science Foundation under award ECS-9820765. The work of T. Y. Al-Naffouri was also partially supported by King Fahd University of Petroleum and Minerals, Saudi Arabia.

0-7803-6293-4/00/$10.00©2000 IEEE

The mean-square deviation (MSD) of an adaptive algorithm is defined as the steady-state value

$$\text{MSD} = \lim_{i\to\infty} \mathrm{E}\,\|\tilde{\boldsymbol{w}}_i\|^2.$$

The MSD clearly depends on the choice of the nonlinear error function f in (2). Let $p(x)$ denote the pdf of the noise signal $v(i)$. It was argued in [3] that, for sufficiently small step-sizes, the choice

$$f_{opt}(x) = -\frac{p'(x)}{p(x)} \tag{6}$$

is optimal in the sense that it leads to an asymptotically efficient estimator (i.e., one that achieves the Cramer-Rao bound). This expression shows that different noise distributions can lead to different choices for f; a remark that is consistent with related studies in the literature for some adaptive algorithms with nonlinear updates such as the LMF algorithm [7]. The result (6) does not show any dependence on the distribution of the regression vector $\mathbf{u}_i$. In [8], on the other hand, a calculus of variations argument was used to argue that, for the special case of an iid Gaussian input sequence $\{\mathbf{u}_i\}$ with variance $\sigma_u^2 I$, the optimal nonlinearity takes the form

$$f_{opt,G}(x) = -\frac{q'(x)}{q(x) + \mu\sigma_u^2 q''(x)}, \tag{7}$$

where $q(e)$ now denotes the pdf of the output error $e(i)$. We shall comment on this expression further ahead — see (26). In the sequel, we shall derive the optimal nonlinearity for a general input sequence under some assumptions (also used in [3, 8]); our expression turns out to subsume the nonlinearities (6) and (7), as well as the algorithms of Table 1, as special cases.

2. FOUR ASSUMPTIONS

To make our analysis tractable, we shall make the following reasonable and often used assumptions:

A1. The noise sequence $\{v(i)\}$ is assumed to be a zero-mean iid process with symmetric and unimodal pdf, and to be independent of the sequence $\{\mathbf{u}_i\}$.

A2. The error-nonlinearity function f is restricted to be sufficiently smooth, odd-symmetric, and sign preserving.

The convenience of these assumptions can be understood by noting that they imply that the variables

$$\{\ f[v(i)],\ f''[v(i)],\ f[v(i)]f'[v(i)],\ f'[v(i)]f''[v(i)]\ \}$$

all have zero-mean.

A3. The a priori error $e_a(i)$ is small enough in steady-state so that its higher than second order powers could be neglected.

Assumption A3 will be quite useful when we expand f in a third order Taylor series:

$$\begin{aligned} f[e(i)] &= f[e_a(i) - (-v(i))] \\ &= -f[v(i)] + f'[v(i)]e_a(i) \\ &\quad -(1/2)f''[v(i)](e_a(i))^2 \\ &\quad +(1/6)f'''[\eta](e_a(i))^3, \end{aligned} \tag{8}$$

for some η between 0 and $e_a(i)$. For it will enable us to neglect higher powers of $e_a(i)$ when operating on f.

We further assume that (this is known as the independence assumptions — see though Remark 5 in Sec. 4.1):

A4. At steady-state, the regressor vector $\mathbf{u}_i$ and the weight-error vector $\tilde{\boldsymbol{w}}_i$ are independent.

3. SECOND-ORDER ANALYSIS

The derivation in this section is simplified by relying on the feedback approach of [4]-[6], which notes that by computing the energies of both sides of (3) we obtain the energy conservation relation

$$\|\tilde{\boldsymbol{w}}_{i+1}\|^2 + \bar{\mu}(i)|e_a(i)|^2 = \|\tilde{\boldsymbol{w}}_i\|^2 + \bar{\mu}(i)|e_p(i)|^2 \tag{9}$$

or, equivalently, using (5),

$$\|\tilde{\boldsymbol{w}}_{i+1}\|^2 = \|\tilde{\boldsymbol{w}}_i\|^2 - 2\mu e_a(i) f[e(i)] + \mu^2\|\mathbf{u}_i\|^2 f^2[e(i)]. \tag{10}$$

No approximations or assumptions are needed to establish this relation; the relation holds irrespective of assumptions A1–A4!

To proceed, we introduce the Taylor series expansion (8) of $f[e(i)]$, which yields

$$\begin{aligned} \|\tilde{\boldsymbol{w}}_{i+1}\|^2 &\simeq \|\tilde{\boldsymbol{w}}_i\|^2 - 2\mu e_a(i)\left(-f + f'e_a(i) - \frac{1}{2}f''e_a^2(i)\right) \\ &+ \mu^2\|\mathbf{u}_i\|^2\left(f^2 - 2ff'e_a(i) + (f'^2 + ff'')e_a^2(i)\right). \end{aligned} \tag{11}$$

For notational convenience, the argument $v(i)$ of f and its derivatives has been suppressed. Upon taking the expectations of both sides and invoking A1 we get

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

Finally, by using the independence assumption A4 we can write

$$\mathrm{E}[e_a^2(i)] \simeq \mathrm{E}\left[\|u_i\|^2\right] \mathrm{E}\left[\|\tilde{\boldsymbol{w}}_i\|^2\right], \qquad (12)$$

and, following [7] and [9], we also have

$$\mathrm{E}\left[e_a^2(i)\|\mathbf{u}_i\|^2\right] \simeq \mathrm{E}\left[\|u_i\|^4\right] \mathrm{E}\left[\|\tilde{\boldsymbol{w}}_i\|^2\right]. \qquad (13)$$

Upon substituting (12) and (13) into (11), we get a first-order difference recursion for the mean-square deviation (MSD) of the form

$$\mathrm{E}\left[\|\tilde{\boldsymbol{w}}_{i+1}\|^2\right] = a\,\mathrm{E}\left[\|\tilde{\boldsymbol{w}}_i\|^2\right] + b, \qquad (14)$$

where the coefficients $\{a, b\}$ are given by

$$a = 1 - 2\mu\,\mathrm{E}[f']\,\mathrm{E}\left[\|\mathbf{u}_i\|^2\right] + \mu^2\,\mathrm{E}\left[f'^2 + ff''\right]\mathrm{E}\left[\|\mathbf{u}_i\|^4\right]$$

$$b = \mu^2\,\mathrm{E}[f^2]\,\mathrm{E}[\|\mathbf{u}_i\|^2].$$

The moments $\mathrm{E}[\|\mathbf{u}_i\|^2]$ and $\mathrm{E}[\|\mathbf{u}_i\|^4]$ are assumed time-invariant.

4. SYNTHESIS OF THE OPTIMUM NONLINEARITY

We can now use our analysis results to optimize the choice of the error nonlinearity. By inspecting (14), we notice that the quantity a controls the transient behavior of (14) and (together with b) controls its steady-state value, given by $\frac{b}{1-a}$. We can optimize the choice of f by minimizing the steady-state value $E\left[\|\tilde{\boldsymbol{w}}_\infty\|^2\right]$ for a fixed transient behavior, or, equivalently,

$$\min_{f \leq f_{max}} \int_{-\infty}^{\infty} f^2(x)p(x)dx \qquad (15)$$

subject to

$$\int_{-\infty}^{\infty} \left[f'(x) - \mu\lambda\left(f(x)f''(x) + (f'(x))^2\right)\right]p(x)dx = C, \qquad (16)$$

where $p(v)$ is the pdf of the noise sequence $\{v(i)\}$,

$$\lambda = \frac{1}{2}\frac{\mathrm{E}\left[\|\mathbf{u}_i\|^4\right]}{\mathrm{E}\left[\|\mathbf{u}_i\|^2\right]} \qquad (17)$$

and where C is a constant. The constraint $f \leq f_{max}$ on f is included to ensure that f is bounded for finite values of its argument. Just like μ, f_{max} is a design parameter that is mandated by the practical application.

To solve this optimization problem, we consider the associated Lagrangian

$$L(f) = \int_{-\infty}^{\infty} f^2(x)p(x)dx \; + \qquad (18)$$

$$\gamma \int_{-\infty}^{\infty} \left[f'(x) - \mu\lambda\left(f(x)f''(x) + (f'(x))^2\right)\right]p(x)dx$$

where γ is a Lagrange multiplier. While the calculus of variations is the standard method for solving similar problems, sufficient conditions for optimality are often difficult to check. We can instead arrive at the optimum solution by using more elementary methods. To do this, notice first that $ff'' + f'^2 = (ff')' = \frac{1}{2}(f^2)''$. Using this fact, coupled with a repeated use of integration by parts, and by further invoking the symmetry properties of f and p, we get

$$\begin{aligned} L(f) &= 2\int_0^{\infty} f\left((p - \frac{\gamma\mu\lambda}{2}p'')f - \gamma p'\right)dx \\ &+ 2\gamma \lim_{x\to\infty}\left(fp - \frac{\mu\lambda}{2}((f^2)'p + f^2p')\right) \\ &= 2\int_0^{\infty} f\left((p - \frac{\gamma\mu\lambda}{2}p'')f - \gamma p'\right)dx, \quad (19) \end{aligned}$$

where the second equality follows from the boundedness of f. Minimizing $L(f)$ now amounts to minimizing the integrand of (19) at each point. It can be shown that the multiplier γ must be negative and that

$$f_{opt}(x) = \min\left\{\frac{\frac{\gamma}{2}p'(x)}{p(x) - \frac{\gamma}{2}\mu\lambda p''(x)},\ f_{max}\right\} \qquad (20)$$

if $p(x) - \frac{1}{2}\gamma\mu\lambda p''(x) > 0$. Otherwise, $f_{opt}(x) = f_{max}$. For small μ, we essentially have

$$f_{opt}(x) = \frac{\frac{\gamma}{2}p'(x)}{p(x) - \frac{\gamma}{2}\mu\lambda p''(x)}, \qquad (21)$$

which is dependent on γ. We can determine γ from the constraint (16), or we can do away with it by substituting (21) into the adaptation equation (3) where we notice that $\frac{\gamma}{2}$ and μ always appear multiplied by each other. Thus, $-\frac{\gamma}{2}$ can be absorbed into the design parameter μ and the optimum nonlinearity effectively becomes one of the general form:

$$\boxed{f_{opt}(x) = -\frac{p'(x)}{p(x)+\mu\lambda p''(x)}} \qquad (22)$$

4.1. Relation to other optimum nonlinearities

The above optimal nonlinearity subsumes as special cases several of the nonlinearities that have already appeared in literature. This is summarized in the following remarks.

1. Since $\mathbf{u}_i$ and $\tilde{w}_i$ are independent (by the independence assumption A4), we could have carried out the second-order analysis and the optimization conditioned on the data vector $\mathbf{u}_i$. The resulting nonlinearity would have been

$$f_{opt|\mathbf{u}_i}(x) \quad = \quad -\frac{p'(x)}{p(x)+\frac{\mu}{2}\|\mathbf{u}_i\|^2 p''(x)}, \quad (23)$$

which incorporates data normalization similar to that of the NLMS.

2. For small μ, the optimum nonlinearity reduces to (6), which was obtained in [3] for the arbitrary input case. Note however that, more generally, the optimal choice (22) is also dependent on the input data statistics (through the variable $\lambda = \mathrm{E}[\|\mathbf{u}_i\|^4]/2\,\mathrm{E}[\|\mathbf{u}_i\|^2]$), as we expect it to be.

3. If the input sequence is iid with fourth order moment ξ_u^4 and variance σ_u^2, then

$$\lambda = \frac{\mathrm{E}\,\|\mathbf{u}_i\|^4}{2\,\mathrm{E}\,\|\mathbf{u}_i\|^2} = \frac{\xi_u^4+(M-1)\sigma_u^4}{2\sigma_u^2} \qquad (24)$$

and the optimum nonlinearity simplifies to

$$f_{opt}(x) = -\frac{p'(x)}{p(x)+\mu\frac{\xi_u^4+(M-1)\sigma_u^4}{2\sigma_u^2}p''(x)}. \qquad (25)$$

Here M is the filter order. This is the same nonlinearity obtained in [10] for an iid input.

4. If we further restrict the input to be iid Gaussian, then $\xi_u^4 = 3\sigma_u^4$ and the optimum nonlinearity (22) becomes

$$f_{opt}(x) = -\frac{p'(x)}{p(x)+\mu\frac{M+2}{2}\sigma_u^2 p''(x)}. \qquad (26)$$

When $e_a(i)$ is small enough that $e(k) \simeq v(k)$, the nonlinearity becomes essentially (7) which was derived in [8] by a conditional analysis approach.

5. We could have instead approached the problem of optimally designing the nonlinearity f by studying the mean-square error (MSE) rather than the mean-square deviation, where the MSE is defined as

$$\mathsf{MSE} = \sigma_v^2 + \lim_{i\to\infty} \mathrm{E}\,|e_a(i)|^2. \qquad (27)$$

As argued in [6], the evaluation of the MSE can be carried out by relying on a weaker set of assumptions than those used in the earlier sections. For example, the independence assumption A4 could be removed. We shall pursue this extension elsewhere.

6. Notice finally that the sign algorithm turns out to be the optimum algorithm to use in the presence of Laplacian noise. This can be seen by substituting $p(x) = \frac{1}{2}e^{-|x|}$ in the expression for the optimum nonlinearity (20). Moreover, it was argued in [11] for the choice (6) that by expanding $p(x)$ in an Edgeworth series, the other (non)linearities of Table 1 turn out to be approximations of (6). Similar analysis can be extended to the more general optimal nonlinearity (22) of this paper.

5. REFERENCES

[1] R. L. Stratonovich, Is there a theory of synthesis of optimal adaptive self-learning and self-organizing systems? *Automat. Remote Contr. (USSR)*, **29**(1), 1968.

[2] Y. Tsypkin, Is there a theory of optimal adaptive systems? *Automat. Remote Contr. (USSR)*, **29**(1), 1968.

[3] B. Polyak and Y. Tsypkin, Adaptive estimation algorithms (convergence, optimality, stability), *Avtomatika i Telemekhanika*, No. 3, pp. 71-84, March 1979.

[4] 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.

[5] M. Rupp and A. H. Sayed, A time-domain feedback analysis of filtered-error adaptive gradient algorithms, *IEEE Transactions on Signal Processing*, vol. 44, no. 6, pp. 1428–1439, Jun. 1996.

[6] N. R. Yousef and A. H. Sayed, A unified approach to the steady-state and tracking analyses of adaptive filtering algorithms, *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Process.*, vol. 2, pp. 699-703, Antalya, Turkey, June, 1999.

[7] E. Walach and B. Widrow, The least-mean-fourth (LMF) adaptive algorithm and its family, *IEEE Trans. Inf. Theory*, vol. IT-30, pp. 275-283, Feb. 1984.

[8] S. C. Douglas and T. H.-Y. Meng, Stochastic gradient adaptation under general error criteria, *IEEE Trans. Signal Process.*, vol. 42, no. 6, pp. 1352-1365, June 1994.

[9] N. J. Bershad, On the optimum data nonlinearity in LMS adaptation, *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-34, no. 1, pp. 69-76, Feb. 1986.

[10] T. Y. Al-Naffouri, *Adaptive Filtering Using the Least-Mean Mixed-Norms Algorithm and its Application to Echo Cancellation*, M.Sc. Thesis, Electrical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, July 1997.

[11] T. Y. Al-Naffouri, A. Zerguine, and M. Bettayeb, A unifying view of error nonlinearities in LMS adaptation, *Proc. ICASSP*, pp. 1697-1700, Seattle, May 12-15, 1998.