STOCHASTIC ANALYSIS OF GRADIENT ADAPTIVE IDENTIFICATION OF NONLINEAR SYSTEMS WITH MEMORY

N. J. Bershad[1], P. Celka[2], and J. M. Vesin[3]

[1]Electrical and Computer Engineering Department, University of California, Irvine, CA, 92697, U.S.A., bershad@ece.uci.edu, (714)-824-6709 (fax-2321).

[2,3]Signal Processing Laboratory, Swiss Federal Institute of Technology (EPFL), CH - 1015, Lausanne Switzerland, celka@lts.epfl.ch, (41 21) 693 2605 (fax-7600).

ABSTRACT

This paper analyzes the statistical behavior of a sequential gradient search adaptive algorithm for identifying an unknown nonlinear system comprised of a discrete-time linear system H followed by a zero-memory nonlinearity g(.). The LMS algorithm first estimates H. The weights are then frozen. Recursions are derived for the mean and fluctuation behavior of LMS which agree with Monte Carlo simulations. When the nonlinearity is modelled by a scaled error function, the second part of the gradient scheme is shown to correctly learn the scale factor and the error function scale factor. Mean recursions for the scale factors show good agreement with Monte Carlo simulations.

1. INTRODUCTION

Gradient search algorithms (i.e. the LMS algorithm and its variants) are often used for the identification of unknown systems. Usually, the unknown system is linear with memory and the algorithm inputs are noise-corrupted samples of the unknown system output and noise-free samples of the unknown system input. Many researchers have studied the statistical behavior of such systems [1,2]. Relatively little analysis has been done when the system to be identified is nonlinear with memory. Much of this analysis is related to nonlinear system identification using neural networks [3-5].

This paper investigates the statistical behavior of a sequential gradient search adaptive algorithm for identifying an unknown nonlinear system comprised of a discrete-time linear system H followed by a zero-memory nonlinearity g(.). Furthermore, both the input and output of the unknown system are corrupted by additive independent noises. Gaussian models are used for all inputs. Fig. 1 shows the specific nonlinear system identification problem. The input to the unknown system is comprised of the sum of two zero-mean independent gaussian white sequences x(n) and $n_I(n)$ with variances σ_x^2 and σ_I^2, respectively. The unknown system output is obscured by a third independent zero-mean gaussian sequence $n_O(n)$ with variance σ_O^2. The nonlinearity g(.) is due to inherent system nonlinearities such as the sigmoidal threshold function in neural networks or amplifier saturation in satellite communication networks, to name a few.

Recursions are derived for the mean and fluctuation behavior for linear adaptation (LMS) and for the mean for the nonlinear adaptation. Since the filter structure of LMS is linear, LMS is not able to identify the nonlinear portion of the unknown system. Surprisingly, LMS identifies a scaled linear part of the unknown system . The scaling depends upon the unknown nonlinearity and the number of training samples. The weights are then frozen at the end of the first adaptation phase. If the nonlinearity is modelled by a scaled error function , the second part of the sequential identification scheme identifies the first phase scale factor and the error function scale factor. This second phase adaptation scheme is shown in Fig. 2.

2. ANALYSIS - LINEAR ADAPTATION

2.1 LMS Algorithm

The weight recursion for the LMS weight vector is given by [1,2]

$$W(n+1) = W(n) + \mu e(n) Y(n) \tag{1}$$

where $Y(n) = X(n) + N_I(n)$, $X^T(n) = [x(n), x(n-1).... \quad x(n-N+1)]$, $N_I^T(n) = [n_I(n), n_I(n-1).... \quad n_I(n-N+1)]$, N = number of adaptive filter taps and

$$e(n) = g\left[H^T X(n)\right] + n_O(n) - W(n)^T Y(n) \tag{2}$$

2.2 Mean Behavior of (1)

Averaging both sides of (1) and using the standard LMS assumption that the weights at time n are statistically independent of the inputs at time n yields

$$E\left[W(n+1)\right] = \left[I - \mu R_{YY}\right] E\left[W(n)\right] + \mu E\left\{g\left[H^T X(n)\right] Y(n)\right\} \tag{3}$$

where $R_{YY} = E[Y(n)Y^T(n)] = (\sigma_x^2 + \sigma_I^2)I$. The second expectation can be evaluated using Bussgang's Thm. [6], yielding a mean weight recursion whose solution is given by

$$E\left[W(n)\right] = \left[I - \mu R_{YY}\right]^n E\left[W(0)\right] + \mu E\left\{g'\left[H^T X(n)\right]\right\} \sum_{p=0}^{n-1} \left[I - \mu R_{YY}\right]^{n-1-p} R_{XX} H \tag{4}$$

where $R_{XX} = E[X(n)X^T(n)] = \sigma_x^2 I$. The steady-state solution to (4) is

$$\lim_{n\to\infty} E\left[W(n)\right] = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_I^2} E\left\{g'\left[H^T X(n)\right]\right\} H. \tag{5}$$

Thus, the mean weights of the LMS algorithm converge to a scaled version of the linear portion of the unknown channel.

2.3 Wiener MSE

The optimum Wiener filter for this problem satisfies the orthogonality condition $E\left[e(n)Y(n)\right] = \underline{0}$ with

$$W_O = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_I^2} E\left\{g'\left[H^T X(n)\right]\right\} H \tag{6}$$

Comparing (5) and (6), the LMS algorithm converges to the Wiener filter on average. The Wiener MSE ξ_O is

$$\xi_O = \sigma_O^2 + E\left[g^2(r)\right] - \frac{\sigma_x^4}{\sigma_x^2 + \sigma_I^2} E^2\left\{g'(r)\right\} H^T H \tag{7}$$

where $r = H^T X(n)$.

2.4 Misadjustment Error

An exact recursion is derived for the fluctuation behavior of the weights which can be used to evaluate the increase in MSE due to the adaptation. Let $V(n) = W(n) - W_0$. Then (1) can be written as

$$V(n+1) = \left[I - \mu Y(n)Y(n)^T\right]V(n) + \mu\varepsilon_W(n)Y(n) \qquad (8)$$

where $\varepsilon_W(n) = g\left[H^T X(n)\right] + n_0(n) - W_O^T Y(n)$ is the Wiener filter error. Averaging (8) and noting that $\varepsilon_W(n)$ is orthogonal to the data vector $Y(n)$, yields

$$E[V(n)] = \left[1 - \mu(\sigma_x^2 + \sigma_I^2)\right]^n V(0) \qquad (9)$$

Post-multplying (8) by its transpose and averaging yields a recursion for the covariance matrix of $V(n)$, $K_{VV}(n)$,

$$\begin{aligned} K_{VV}(n+1) &= K_{VV}(n) - \mu R_{YY}K_{VV}(n) - \mu K_{VV}(n)R_{YY} \\ &+ \mu E\left[\varepsilon_W(n)Y(n)V^T(n)\left\{I - \mu Y(n)Y(n)^T\right\}\right] \\ &+ \mu E\left[\varepsilon_W(n)Y(n)V^T(n)\left\{I - \mu Y(n)Y(n)^T\right\}\right]^T \\ &+ \mu^2 E\left[Y(n)Y(n)^T K_{VV}(n)Y(n)Y(n)^T\right] \\ &+ \mu^2 E\left[\varepsilon_W^2(n)Y(n)Y(n)^T\right] \end{aligned} \qquad (10)$$

$\varepsilon_W(n)$ is non-Gaussian because $g(.)$ is nonlinear. The orthogonality of $\varepsilon_W(n)$ and $Y(n)$ is not sufficient for independence. Hence, the expectations involving $\varepsilon_W(n)$ in (10) are new and must be evaluated. Let

$A = E[z^3 g(z)]/Var^3(z) - 3E[zg(z)]/Var^2(z)$,

$B = \{E[z^2 g^2(z)]/Var(z) - E[g^2(z)]\}/Var(z)$ and $C = \sigma_x^2 + \sigma_I^2$.

Near convergence, (10) becomes [7]

$$\begin{aligned} K_{VV}(n+1) &= \left(1 - 2\mu C + 2\mu^2 C^2\right)K_{VV}(n) + \mu^2 C^2 tr[K_{VV}(n)] \\ &+ \mu^2\left\{C\xi_O I + \sigma_x^4 HH^T + \left\{\begin{bmatrix} B - 2E^2[g'(z)] - \\ 2E[g'(z)]\dfrac{\sigma_x^4}{C}A \end{bmatrix} H^T H\right\}\right\} \end{aligned} \qquad (11)$$

Taking the trace of (11), solving the recursion and assuming the additional components of the trace due to the nonlinearity $g(z)$ are negligible yields the result for the linear case,

$$\lim_{n\to\infty} tr\left[K_{VV}(n)\right] = \frac{\mu N\xi_O}{\left(2 - (N+2)\mu(\sigma_x^2 + \sigma_I^2)\right)} \qquad (12)$$

However, in (12), ξ_O is bounded away from the noise floor σ_O^2 because of the nonlinearity. Hence, the weight fluctuations are larger than when identifying a linear system with memory. Since ξ_O is bounded away from σ_O^2 and $K_{VV}(n)$ is of order μ for large n, one can write $K_{VV}(n) = E[V(n)]E[V(n)]^T$ for small n. For large n, the MSE is dominated by the mis-match terms. Then,

$$\begin{aligned} E\left[e^2(n)\right] &= \sigma_O^2 + E\left[g^2[H^T X(n)]\right] - \\ &\left(\sigma_x^2 + \sigma_I^2\right)\left[W_O^T W_O - E[V(n)]^T E[V(n)]\right] \end{aligned} \qquad (13)$$

Using (9) in (13) for $W(0) = \underline{0}$, yields

$$E\left[e^2(n)\right] = \sigma_O^2 + E\left[g^2(r)\right] - C\left(1 - \left\{1 - \mu C\right\}^{2n}\right)W_O^T W_O \qquad (14)$$

Hence for practical purposes, when $g(z)$ is nonlinear, the fluctuation effects of $W(n)$ upon the MSE is negligible.

Thus, $W(n) = \alpha(n)H$ where

$$\alpha(n) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_I^2}E\left\{g'\left[H^T X(n)\right]\right\} \times \left[1 - \left\{1 - \mu(\sigma_x^2 + \sigma_I^2)\right\}^n\right] \qquad (15)$$

and holds for sufficiently small μ. After n iterations, $W(n)$ in Fig. 1 will be replaced by $\alpha(n)H$, a deterministic scaled version of the linear portion of the unknown channel. The adaptive filter weights are now frozen in time.

3. ANALYSIS - NONLINEAR ADAPTATION

3.1 Nonlinear Adaptation Algorithm

Consider the model in Fig. 2 for learning the nonlinear portion of the channel $g(z)$. α_f is the scale factor in (15) after the weights are frozen. Since $g(z)$ is unknown, α_f is an unknown parameter. The factor k is selected so that the nonlinearity input and output powers are independent of σ. Thus, $(SNR)_0 = \sigma_x^2 H^T H / \sigma_0^2$ is fixed as σ varies. Since

$$E\left[z^2(n)\right] = \sigma_x^2 H^T H, \qquad k^2 = \sigma_x^2 H^T H / E\left[g^2(z)\right] \qquad (16)$$

The factor k is also unknown since $g(z)$ is unknown. However, the shape of $g(z)$ is assumed known. Thus, a zero memory nonlinear system ID problem has been defined. It consists of adapting b_1 and b_2 to the two unknown parameters α_f and k. $b_1(n)$ and $b_2(n)$ will be adapted using a gradient descent algorithm.

The error is given by

$$e(n) = kg(r) + n_2(n) - b_2(n)g\left[b_1(n)\alpha_f H^T Y(n)\right] \qquad (17)$$

The MSE is

$$\begin{aligned} E[e^2(n)] &= \sigma_O^2 + \sigma_x^2 + \\ &b_2^2(n)E\left[g^2\left\{\alpha_f b_1(n)\left(r + H^T n_I(n)\right)\right\}\right] \\ &-2b_2(n)k E\left[g(r)g\left\{\alpha_f b_1(n)\left(r + H^T n_I(n)\right)\right\}\right] \end{aligned} \qquad (18)$$

The stochastic gradient search algorithm for $b_1(n)$, $b_2(n)$ is

$$b_1(n+1) = b_1(n) + \mu e(n) b_2(n) \alpha_f H^T Y(n) \times g'\left[b_1(n)\alpha_f H^T Y(n)\right] \quad (19)$$

$$b_2(n+1) = b_2(n) + \mu\, e(n)\, g\left[\alpha_f b_1(n) H^T Y(n)\right]$$

3.2 Specification of Nonlinearity

In order to proceed further with the analysis, g(z) must be chosen,

$$g(z) = \int_0^z \exp(-u^2 / 2\sigma^2)\, du \quad (20)$$

This function is a reasonable model for saturation type nonlinearities. A scaled version has been successfully used for modelling saturation type nonlinearities in such applications as the threshold function in neural networks [3] and limiters in satellite amplifiers [5]. By varying the parameter σ, g(z) can range from a linear device ($\sigma \to \infty$) to a scaled hard limiter sgn(z) ($\sigma \to 0$).

3.3 MSE Surface

Let $a = \sigma^2 + \sigma_x^2 H^T H$, $b = (\sigma_x^2 + \sigma_I^2) H^T H$ and $c = \sigma^2 + \alpha_f^2 b_1^2 b$. Using (20) in (19) yields,

$$E[e^2(n)] = \sigma_o^2 + \sigma_x^2 + b_2^2(n)\sigma^2 \sin^{-1}\left[\frac{\alpha_f^2 b_1^2(n) b}{\alpha_f^2 b_1^2(n) b + \sigma^2}\right] - 2 b_2(n) k \sigma^2 \sin^{-1}\left[\frac{\alpha_f b_1(n) \sigma_x^2 H^T H}{\sqrt{ac}}\right] \quad (21)$$

The MSE surface has a global minimum [7].

3.4 Transient Mean Weight Behavior

Eq. (21) can be used to find recursions for the transient mean behavior of (19). For small μ, the fluctuations of $b_1(n)$ and $b_2(n)$ about their respective means can be neglected. Thus, using the partial derivatives of the MSE surface (18) and averaging (19) yields

$$\bar{b}_1(n+1) = \bar{b}_1(n) + \mu \frac{\alpha_f \sigma^4 \bar{b}_2(n)}{\left(\sigma^2 + \alpha_f^2 \bar{b}_1^2(n) b\right)} \times \left\{ \frac{k \sigma_x^2 H^T H}{\sqrt{a\bar{c} - \left(\alpha_f \bar{b}_1(n) \sigma_x^2 H^T H\right)^2}} - \frac{\bar{b}_1(n) \bar{b}_2(n) \alpha_f b}{\sqrt{\bar{c}^2 - \left(\alpha_f^2 \bar{b}_1^2(n) b\right)^2}} \right\} \quad (22)$$

$$\bar{b}_2(n+1) = \bar{b}_2(n) + \mu \sigma^2 k \sin^{-1}\left[\frac{\alpha_f \bar{b}_1(n) \sigma_x^2 H^T H}{\sqrt{a\bar{c}}}\right] - \mu \sigma^2 \bar{b}_2(n) \sin^{-1}\left[\frac{\alpha_f^2 \bar{b}_1^2(n) b}{\alpha_f^2 \bar{b}_1^2(n) b + \sigma^2}\right] \quad (23)$$

where $\bar{b}_1(n) = E\left[b_1(n)\right]$, $\bar{b}_2(n) = E\left[b_2(n)\right]$ and $\bar{c} = E[c]$.

4. COMPARISON OF MONTE CARLO SIMULATIONS AND THEORY

A number of assumptions were made in Section 2. These led to the conclusion that the linear adaptive filter can identify the linear portion of the unknown channnel to within a scale factor. Monte Carlo simulations (100 runs) have been performed to verify this. The filter H was a normalized time-delayed raised-cosine with 13 taps. The other parameters were $\mu = .001$, $\sigma_o = 0.1$, $\sigma_x = 1$, $\sigma_I = 1$, and $\sigma = 1$. Fig. 3 compares the time averaged (uniform weighting of 10 adjacent sample points) simulated MSE with the MSE predicted by (15). Fig. 4 shows the mean behavior of the weights as predicted by (4). Fig. 5 shows the Monte Carlo simulations. Comparison of the MC simulations and theory show excellent agreement. Fig. 6 shows the weights after 1000 iterations.

Monte Carlo simulations (10 runs) of (19) and the theoretical behavior predicted by (22) and (23) have shown excellent agreement [7] for $b_1(0) = b_2(0) = 1$, $\mu = .01$, $\sigma_o = .1$, $\sigma_x = 1$, $\sigma_I = \text{sqrt}(.1)$ and $\sigma = 1$, sqrt(.1), sqrt (10).

5. RESULTS AND CONCLUSIONS

This paper has investigated the statistical behavior of a gradient adaptive scheme for identifying an unknown parameterized nonlinear system (i.e. the shape of the nonlinearity is assumed known) with memory. New recursions were derived for the mean and fluctuation behavior of the LMS algorithm and for the mean behavior of the nonlinear gradient algorithm. The deterministic recursions accurately predicted the behavior of Monte Carlo simulations.

REFERENCES

1. S. Haykin, *Adaptive Filter Theory* , Prentice-Hall, Englewood-Cliffs, N.J., 1991.
2. B. Widrow and S. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood-Cliffs, N.J., 1985.
3. N. J. Bershad, J. J. Shynk and P.L. Feintuch, "Statistical Analysis of the Single-Layer Back-Propagation Algorithm, Parts I and II," IEEE Trans. on Signal Processing, Vol. SP-41, No. 2, pp. 573-591, Feb. 1993.
4. N. J. Bershad, M. Ibnkahla and F. Castanie, "Statistical Analysis of a Two-Layer Backpropagation Algorithm used for Modelling Memoryless Channels: the Single Neuron Case," IEEE Trans. on Signal Processing, Vol. SP-45, No. 3, pp.747-756, March 1997.
5. M. Ibnkahla, N. J. Bershad, J. Sombrin and F. Castanie, "Neural Network Modelling and Identification of Nonlinear Channels with Memory: Algorithms, Applications and Analytic Models," IEEE Trans. on Signal Processing -Special Issue on Applications of Neural Networks to Signal Processing, November 1997 (to be published).
6. J. J. Bussgang, "Cross-Correlation Functions of Amplitude Distorted Gaussian Signals, Tech. Rpt. 216, Res. Lab. Electron, M.I.T. Cambridge, MA, March, 1952.
7. N. J. Bershad, P. Celka and J. M. Vesin, "Stochastic Analysis of Gradient Adaptive Identification of Nonlinear Systems with Memory for Gaussian Data and Noisy Input and Output Measurements ", submitted to IEEE Trans. on Signal Processing, August 1997.

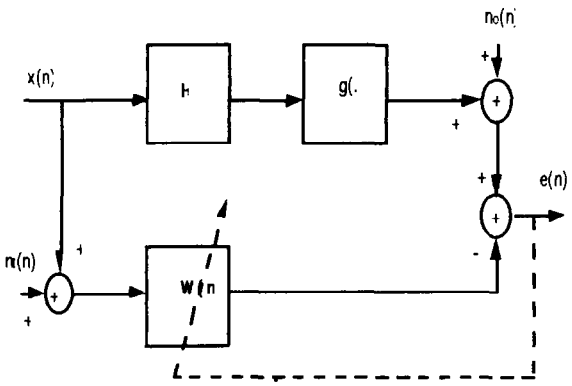

Fig. 1 - Nonlinear System Identification with a Linear Adaptive Filter.

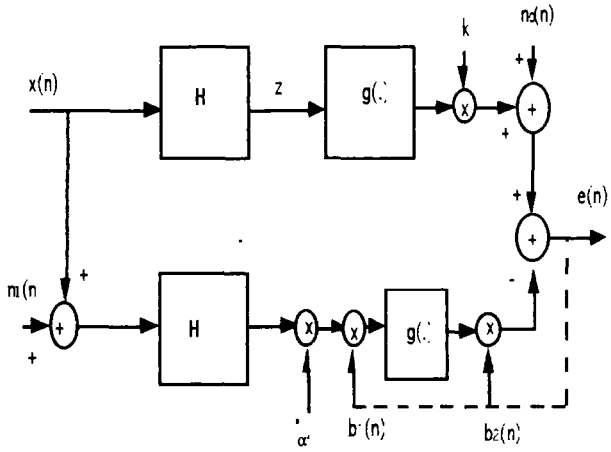

Fig. 2 - Nonlinear System Identification with a frozen linear filter and adaptive scaling for the nonlinearity.

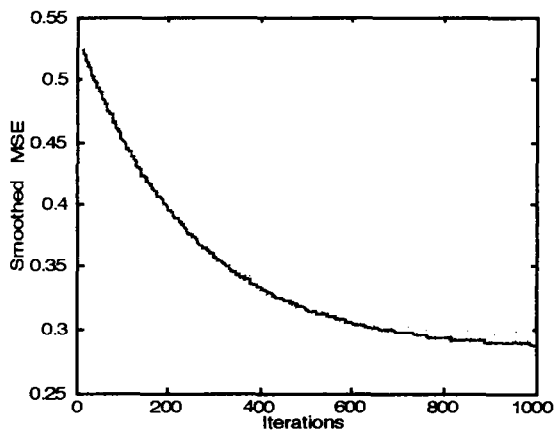

Fig. 3 - Smoothed MSE vs. Iterations for μ =.001, σ_O = 0.1 , σ_X = 1, σ_I =1, σ = 1.

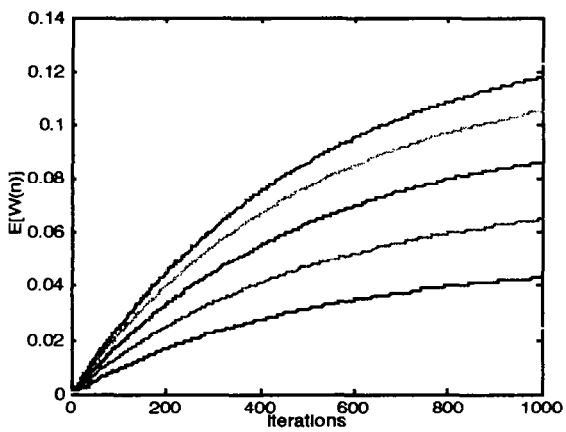

Fig. 4 - Mean Weight vs. Iterations for μ =.001, σ_O = 0.1 , σ_X = 1, σ_I =1, σ = 1.

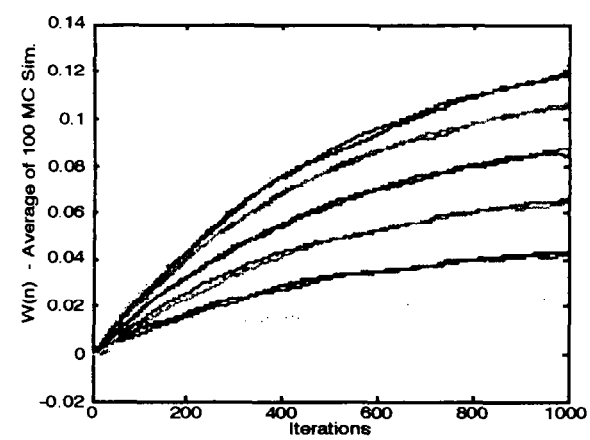

Fig. 5 - Average Weight (100 MC) vs. Iterations for μ =.001, σ_O = 0.1 , σ_X = 1, σ_I =1, σ =1.

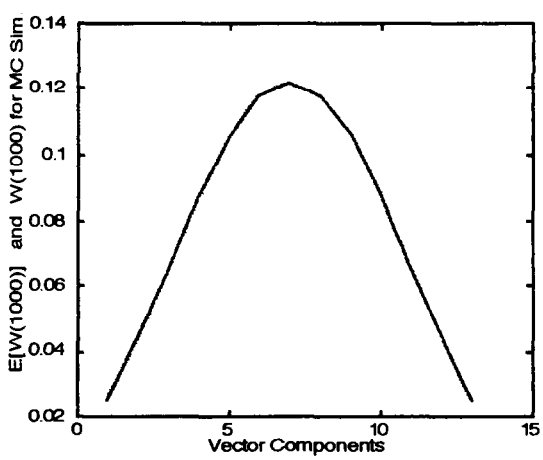

Fig. 6 - E[W(1000)] and W(1000) for 100 MC Sim.