ANALYSIS OF STOCHASTIC GRADIENT IDENTIFICATION OF POLYNOMIAL NONLINEAR SYSTEMS WITH MEMORY

P. Celka[†], N.J. Bershad[§] and J.M. Vesin[†]

[†] Department of Electrical Engineering, Signal Processing Laboratory,
Swiss Federal Institute of Technology
[§]Electrical and Computer Engineering Department,
University of California, Irvine

ABSTRACT

This paper present analytical, numerical and experimental results for a stochastic gradient adaptive scheme which identifies a polynomial-type nonlinear system with memory for noisy output observations. The analysis includes the computation of the stationary points, the mean square error surface, and the mean behavior of the algorithm for Gaussian data. Monte Carlo simulations confirm the theoretical predictions which show a small sensitivity to the observation noise.

1. INTRODUCTION

Much research has been performed on the nonlinear system identification problem for many years. Many methods have been developed devoted to this task [1, 2].

Recently, Bershad et al. have analyzed the stochastic gradient (SG) adaptive identification of Wiener systems (a linear filter followed by a zero-memory nonlinear function [3]) with noisy input and output measurements. The analytical results were obtained by modeling a smooth threshold type nonlinearity with an Erf function with input and output scaling factors. Many nonlinear systems can be modeled globally using this familly of nonlinear functions. However, often the nonlinear system operates in a small region about a bias point. It is simpler to study the identification behavior of these systems using a linear filter followed by a limited Taylor series expansion around the bias point. The unknown system output is obscured by some noise n_o. The identification of polynomial-type nonlinear systems with memory can be handled with this model. Identification is performed in two steps: 1) linear filter identification using the LMS algorithm, 2) polynomial nonlinearity identification using a SG algorithm. Since the unknown coefficients of the polynomial nonlinearity are linearly embedded in the model, LMS algorithms can be used to identify both the linear and the polynomial coefficients. Their statistical behaviors are studied. Recursions for the mean polynomial coefficients is obtained. Monte Carlo simulation is provided for 0dB signal to noise ratio in order to support the theory.

2. SOME PRELIMINARY RESULTS

2.1. Linear filter estimation

The unknown system structure consists of an N'th order linear time-invariant system $\mathbf{H}$ followed by a zero-memory nonlinearity $h(.)$. $h(.)$ is assumed to be represented by a P'th order Taylor expansion ($P > 0$) $f(x, \mathbf{a})$ around a bias point (see Figure 1).

This system is a Wiener-type block structure [4]. Consider the identification of a local expansion of the nonlinearity about the bias point. The input and output sequences, x and y, are composed of L samples with power σ_x^2 and σ_y^2 respectively. The output is corrupted by additive noise n_o with power σ_o^2. The signals x and n_o are zero-mean Gaussian processes, independent of each other. Hence,

$$y(n) = f(z(n), \mathbf{a}) + n_o(n) \qquad (1)$$

with $f(x, \mathbf{a}) = \sum_{k=0}^{P} a_k \; x^k$. The signal $z(n) = \mathbf{H}^T \mathbf{X}(n)$ is the linear filter output with power σ_z^2, $\mathbf{X}(n)^T$ is the input vector, and $\mathbf{a}^T = [a_0 \ldots a_P]$. The identification goal is to estimate the $N + P + 1$ parameters of $\mathbf{H}$ and $\mathbf{a}$. A recent pa-

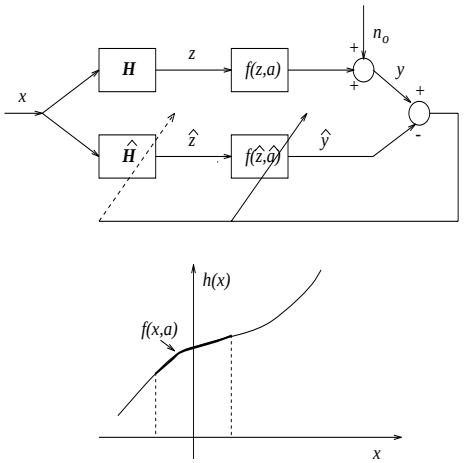

Figure 1: Adaptive identification scheme

per [3] has shown that the linear part $\mathbf{H}$ of the system can be identified to within a scale factor using the LMS algorithm. The weight vector $\mathbf{W}(n)$ recursion is given by

$$\mathbf{W}(n+1) = \mathbf{W}(n) + \mu e_H(n)\mathbf{X}(n) \qquad (2)$$

where $e_H(n) = y(n) - \mathbf{W}(n)^T\mathbf{X}(n)$. Let $E[x]$ denote the expectation of x. The statistical analysis results for slow learning (small μ) are summarized here [3] and applied to this problem. The LMS mean weight vector $E[\mathbf{W}(n)]$ converges to the optimum Wiener filter $\hat{\mathbf{H}}$, i.e.

$$\lim_{n\to\infty} E[\mathbf{W}(n)] = \alpha\mathbf{H} = \hat{\mathbf{H}} \qquad (3)$$

where $\alpha = E\left[(\partial f(x,\mathbf{a})/\partial x|_{x=z(n)}\right]$. The value of α depends on $\mathbf{H}$ and $\mathbf{a}$ (both unknown parameter sets) through the function $f(x,\mathbf{a})$. Evaluating α with $x = z(n) = \mathbf{H}^T\mathbf{X}(n)$ for a Gaussian random variable x, yields

$$\alpha = \sum_{k=0}^{\lfloor (P-1)/2 \rfloor} a_{2k+1}(2k+1)(2k-1)!!\sigma_z^{2k} \qquad (4)$$

where $\lfloor k \rfloor$ stands for the smallest integer near k. The weight vector $\mathbf{W}(n)$ displays fluctuations around its mean, i.e. misadjustment error. This leads to some problems for the analysis of the nonlinearity learning. For instance, the output $\hat{z}(n)$ is given by $\hat{z}(n) = \mathbf{W}(n)^T\mathbf{X}(n)$. This is somewhat different from $E[\hat{z}(n)] = \hat{\mathbf{H}}^T\mathbf{X}(n) = \alpha\mathbf{H}^T\mathbf{X}(n)$. Nevertheless, under the slow convergence hypothesis (small μ), the output $\hat{z}(n)$ can be written approximately as [3]

$$\hat{z}(n) \approx \alpha(n)\mathbf{H}^T\mathbf{X}(n) \qquad (5)$$

where $\alpha(n) = \alpha\,(1-(1-\mu\sigma_x^2)^n)$. $\alpha(n)$ converges towards α as n goes to infinity. $\hat{z}(n)$ converges towards a scaled version of the output of the filter $\mathbf{H}$. The asymptotic fluctuation behavior can be computed using the results in [3]. The Wiener mean square error (MSE) ξ_H is given by $\xi_H \approx \sigma_o^2 + E\left[f^2\,(z(n),\mathbf{a})\right] - \alpha^2\sigma_z^2$ where the fluctuation behavior of $\mathbf{W}$ is neglected. Let $\mathbf{V}_H(n) = \mathbf{W}(n) - \hat{\mathbf{H}}$ be the weight error about the optimum Wiener filter. It is shown in [3] that

$$\begin{aligned}\lim_{n\to\infty} tr\left[\mathbf{V}_H(n)\mathbf{V}_H^T(n)\right] &= \frac{\mu}{(2-(N+2)\mu\sigma_x^2)} \\ &\times \left[N\xi_H + \sigma_z^2\,\Gamma\right] \qquad (6)\end{aligned}$$

with $\Gamma = (2+B) - 2\alpha\left(2\alpha + \sigma_z^2 A\right)$. The Γ factor takes into account the effect of the nonlinearity. Equation (6) indicates that the fluctuations are proportional to μ for any f. Hence, the approximation in ξ_H is valid for small μ. The factors A and B in Γ (introduced in [3]) can be evaluated using Bussgang's Theorem. The MSE in the linear filter learning phase is given by

$$E\left[e_H^2(n)\right] = \xi_H + \sigma_x^2 tr\left[\mathbf{V}_H(n)\mathbf{V}_H^T(n)\right] \qquad (7)$$

If $f(z,\mathbf{a}) = z$, then the MSE reduces to $E\left[e_H^2(n)\right] = \sigma_o^2$ because $A = 0$, $B = 2$, and $\alpha = 1$. If μ is chosen small, the MSE in equation (7) is dominated by ξ_H. Thus, the MSE depends only on the shape of the nonlinearity and not on the adaptive learning process. The nonlinear contribution to the misadjustment error is due to Γ.

2.2. Scaling property for the nonlinearity

The parameter vector $\mathbf{a}$ is unknown. Thus, α is also unknown. It is not possible to identify α and $\mathbf{a}$ independently. Indeed, denoting $\mathbf{Z}^T = (1, z, z^2, \ldots, z^P)$, we can write $f(z,\mathbf{a}) = \mathbf{a}^T\mathbf{Z}$. $f(z,\mathbf{a})$ is identified using a similar polynomial-type function $f(\hat{z},\hat{\mathbf{a}})$ (see Figure 1) given by $f(\hat{z},\hat{\mathbf{a}}) = \sum_{k=0}^{P}\hat{a}_k\,\hat{z}^k$ with $\hat{\mathbf{a}}^T = (\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3)$. Introduce the diagonal matrix $G = diag(1, \alpha, \ldots, \alpha^P)$ composed of the $P+1$ powers of α. Using equations (5), and the matrix G, $f(\hat{z},\hat{\mathbf{a}})$ be rewritten approximatly as $f(\hat{z},\hat{\mathbf{a}}) \approx \hat{\mathbf{a}}^T\,G\,\mathbf{Z} = f(z, G\hat{\mathbf{a}})$. Perfect identification occurs when $f(\hat{z},\hat{\mathbf{a}}) = f(z,\mathbf{a})$, and thus $\hat{\mathbf{a}} =$

$G^{-1}\mathbf{a}$. This implies only $\hat{\mathbf{a}}^T = (a_0, a_1/\alpha, \ldots, a_P/\alpha^P)$ can be estimated. Thus, it is impossible to identify α and $\mathbf{a}$ independently.

3. LEARNING THE NONLINEARITY

3.1. Global stationary points

The N coefficients of the linear part of the model $\hat{\mathbf{H}}$ have been estimated in section 2.1. Now the $P+1$ parameters $\mathbf{a}$ are estimated using a stochastic gradient learning algorithm. The estimate of $\mathbf{a}$, $\hat{\mathbf{a}}$, converges in some statistical sense towards the minimum of the mean square error surface [5]

$$\xi_f = E\left[e_f^2(n)\right] = \xi_f(\mathbf{a}) \tag{8}$$

The residual error $e_f(n)$ is now given by $e_f(n) = y(n) - f(\hat{z}(n), \hat{\mathbf{a}})$ or by

$$e_f(n) \approx (\mathbf{a} - G\hat{\mathbf{a}})^T\, \mathbf{Z}(n) + n_o(n) \tag{9}$$

where $\hat{\mathbf{Z}}$ is approximated by $G\mathbf{Z}$. The parameter vector $\hat{\mathbf{a}}$ is not time dependent. Equation (8) defines a $P+1$ dimensional surface. Define the parameter deviation vector $\mathbf{V}_f = \hat{\mathbf{a}} - G^{-1}\mathbf{a}$, and assuming that $n_o(n)$ and $\mathbf{Z}(n)$ are independent, equation (8) for the MSE surface becomes

$$\xi_f = \sigma_o^2 + \mathbf{V}_f^T E\left[Q(n)\right] \mathbf{V}_f \tag{10}$$

where $Q(n) = [\hat{\mathbf{Z}}(n)][\hat{\mathbf{Z}}(n)]^T \approx [G\mathbf{Z}(n)][G\mathbf{Z}(n)]^T$ is a square semi-definite positive symmetric real valued matrix. Equation (10) defines a $P+1$ quadratic error surface. This surface is elliptic and possesses a unique global minimumat $\mathbf{V}_f^* = 0$. The stochastic gradient algorithm will converge to this minimum under some stability conditions. The matrix $E\left[Q(n)\right]$ plays the role of the input vector covariance matrix R in standard linear LMS theory [5].

3.2. Stochastic gradient algorithm

LMS is often used for adapting the parameters of a linear tapped delayed line. It is also widely used for signal processing and system identification [5]. The LMS is also used here to estimate the coefficient vector $\mathbf{a}$ (more precisely a scaled version). The adaptive gradient recursion for $\hat{\mathbf{a}}$ is

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) - \frac{M}{2}\nabla_{\hat{\mathbf{a}}(n)}\, e_f^2(n) \tag{11}$$

where $M = diag(\mu_0, \ldots, \mu_P)$ is a diagonal matrix of positive real adaptive coefficients (step sizes). M is usually a scaled identity matrix $M = \mu I$, where μ is a scalar. The change in the parameter $\hat{a}_k(n)$ in equation (11) depends on $z(n)^k$ (see equation (12)). The kth component of $\hat{\mathbf{a}}$ is proportional to σ_z^{2k}. A diagonal step-size matrix allows for the selection of an optimal step-size for $\hat{a}_k$. The stochastic gradient recursions can be rewritten as

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) + e_f(n) M \hat{\mathbf{Z}}(n) \tag{12}$$

Not surprisingly, the recursions for $\hat{\mathbf{a}}(n)$ are linear because f is linear in $\mathbf{a}$ (or equivalently ξ_f is quadratic in $\mathbf{V}_f$). The equation (12) is used to learn $\mathbf{a}$ after the filter coefficients $\hat{\mathbf{H}}$ have been computed. Hereafter, the mean behavior of (12) is studied.

4. MEAN BEHAVIOR OF (12)

Some basic statistical properties of the set $\{\hat{z}(n), y(n), \hat{a}_k(n), n_o(n)\}$ are usually assumed in order to simplify the analysis problem. For example $\hat{z}(n)$, $\hat{a}_k(n)$ and $n_o(n)$ are assumed mutually independent random variables. This is often not true. However, this assumption has provided reasonable stability conditons for stochastic gradient algorithms. The approach here is based upon this independence assumption. This section derives the mean polynomial parameter recursions. These recursions predict the behavior of (12) in some statistical sense. The recursion for the vector $\mathbf{V}_f(n)$ is

$$\mathbf{V}_f(n+1) = (I - MQ(n))\, \mathbf{V}_f(n) - n_o(n) M \hat{\mathbf{Z}}(n) \tag{13}$$

Taking the expectation of both sides of equation (13), and using the independence assumption,

$$E\left[\mathbf{V}_f(n+1)\right] = E\left[\mathcal{T}(n)\right] E\left[\mathbf{V}_f(n)\right] \tag{14}$$

where $\mathcal{T}(n) = I - MQ(n)$. Equation (14) is linear in the $P+1$ deviation vector $E\left[\mathbf{V}_f(n)\right]$. The stationary points are the solution of $E\left[\mathbf{V}_f(n+1)\right] = E\left[\mathbf{V}_f(n)\right]$. This solution is given by $E\left[\mathbf{V}_f(n)\right]^* = 0$, and is also the minimum point of the MSE surface (see equation (10) and comments below). Equation (14) is similar to the standard LMS result for the mean weight vector (see for example [6] p. 102). However, note that the matrix $E\left[\mathcal{T}(n)\right]$ is not symmetric because M and $E\left[Q(n)\right]$ do not commute. The matrix $E\left[\mathcal{T}(n)\right]$

is usually denoted the transition matrix. The solution to (14) leads to

$$E\left[\hat{\mathbf{a}}(n)\right] = E\left[\mathcal{T}(n)\right]^n E\left[\mathbf{V}_f(0)\right] + G^{-1}\,\mathbf{a} \qquad (15)$$

The matrix G^{-1} always exists because $\alpha \neq 0$. The initial condition $\hat{\mathbf{a}}(0) = 0$ leads to $E\left[\mathbf{V}_f(0)\right] = -G^{-1}\mathbf{a}$. Convergence of the mean implies $\lim_{n\to\infty} E\left[\mathbf{V}_f(n)\right] = E\left[\mathbf{V}_f(n)\right]^* = 0$.

5. NUMERICAL STUDIES

Monte Carlo simulations are presented for both the linear filter learning (2) and the nonlinear polynomial coefficient learning (12). Theoretical mean behaviors (5) and (15) of (2) and (12) respectively are compared with Monte Carlo simulations for a 0dB signal to noise ratio. P is chosen equal to 3 (third order polynomial nonlinearity) and $N = 5$ (fifth order linear filter). The input power has been fixed at $\sigma_x^2 = 4$. Equation (4) gives $\alpha = a_1 + 3a_3(H^T H)\sigma_x^2$. 100 Monte Carlo runs have been performed for this simulation. The linear filter learning gain was fixed at $\mu = 0.0005$. The step-size coefficients were fixed at $\mu_0 = 0.01, \mu_1 = 0.005, \mu_2 = 0.0005$,$\mu_3 = 0.00005$. Figure 2 shows the Monte Carlo simulation on the left and the theoretical prediction on the right.

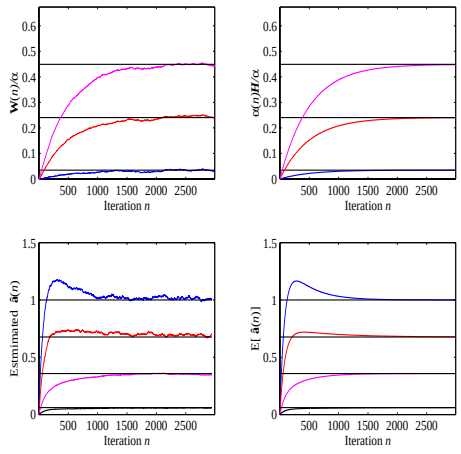

Figure 2: Monte Carlo simulation (left pannel) and theoretical prediction (right panel) of (12), SNR=0dB

The constant lines in the two first graphics show the first three linear filter weights $W_i(n)/\alpha$. The theoretical value of α has been used to verify equation (4). The scaled weights $W_i(n)/\alpha$ converge to $\mathbf{H}$. The theory agrees with the Monte Carlo simulations. On the two second graphics, the time varying curves show the four polynomial coefficient learning curves $\hat{a}_i(n)$. Each coefficient $\hat{a}_i(n)$ converge to a_i/α^i and the theoretical curves match Monte Carlo simulations. The converged coefficient values are in very good agreement with the theoretical values. The polynomial coefficient fluctuations are smaller then those of $\mathbf{W}(n)$.

6. CONCLUSIONS

This paper has investigated the identification of a Wiener-type nonlinear system using adaptive stochastic gradient algorithms. The nonlinearity was assumed to be locally expandable in a Taylor series. The linear and nonlinear polynomial parameters have been identified separately using LMS. Analysis has been performed for the mean behavior of the LMS algorithms for both linear and nonlinear coefficients. Monte Carlo simulation has confirmed the theoretical predictions.

Finally, this type of adaptive nonlinear system identification could be useful when no apriori knowledge is provided for the class of nonlinearities.

7. REFERENCES

[1] S. Chen, A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification", Int. J. Control, vol.50, pp. 1873–1896, 1989.

[2] K.H. Chon, M.J. Korenberg, and N.H. Holstein-Rathlou, "Application of fast orthogonal search to linear and nonlinear stochastic systems", Ann. Biomed. Eng., vol. 25, pp. 793–801, 1997.

[3] N.J. Bershad, P. Celka, and J.M. Vesin, "Stochastic analysis of gradient adaptive identification of nonlinear systems with memory", ICASSP'98, Seattle, vol. 1, 1998.

[4] N. Wiener, Nonlinear Problems in Random Theory, John Wiley and Sons, 1958.

[5] S. Haykin, Adaptive Filter Theory, Prentice-Hall, 1991.

[6] B. Widrow and S.D. Stearns, Adaptive Signal Processing, Prentice-Hall, 1985.