Adaptive Algorithms with Nonlinear Data and Error Functions

William A. Sethares, *Member, IEEE*

***Abstract*—Using the tools of nonlinear system theory, we examine several common nonlinear variants of the LMS algorithm and derive a persistence of excitation criteria for local exponential stability. The condition is tight when the inputs are periodic, and a generic counterexample is demonstrated which gives (local) instability for a large class of such nonlinear versions of LMS, specifically, those which utilize a nonlinear data function. The presence of a nonlinear error function is found to be relatively benign in that it does not affect the stability of the error system. Rather, it defines the cost function the algorithm tends to minimize. Specific examples include the dead zone modification, the cubed data nonlinearity, the cubed error nonlinearity, the signed regressor algorithm, and a single layer version of the backpropagation algorithm.**

I. Introduction

The generic adaptive algorithm generates new parameter estimates as the sum of the old parameter estimates plus some function of the input data multiplied by some function of the error signal. In the least mean square (LMS) algorithm [1], both the data and error functions are unity, but various other functions have recently come under scrutiny [2]. In [3] and [4], for instance, the data function is chosen to be the signum function. In [5], a cubic function is considered. Other algorithms, such as backpropagation [6], use other nonunity functions. This paper treats these various cases in terms of a generic adaptive update form, and, using the tools of nonlinear systems theory (linearization and averaging), finds conditions under which the various update strategies can be expected to succeed in their identification task.

The conditions are stated in terms of a persistence of excitation which, in the ideal case (with no disturbances), must be satisfied in order to guarantee exponential convergence of the parameter estimates to their true values. When bounded disturbances are present, the convergence is to a small region about the true value. The excitation conditions involve the nonlinear functions of the data and the error signal, but the nonlinearities enter in different ways. It is shown that sign preserving error nonlinearities are essentially benign in terms of stability of the adaptive system, while even the best behaved data nonlinearities can cause stability problems. A generic counter example is presented which shows that for a very large class of data nonlinearities (including all "reasonable" ones), there are inputs that lead to (locally) unstable performance of the adaptive system.

These results have a simple geometrical interpretation in terms of descending an error surface. LMS is well known to be an approximate gradient descent method utilizing the squared error as a cost function [7]. At each update instant, the vector of input data points in the "downhill" direction, while the error signal scales the motion in that direction. The effect of a nonlinearity on the data vector is to cause motion in a direction that is not necessarily "downhill." It is not surprising that for certain data sequences, this misalignment from the actual gradient direction can cause the algorithm to climb, rather than descend the error surface. The effect of an error nonlinearity is subtler. It changes the cost function that will be minimized. Thus the presence of sign preserving error nonlinearities is transparent in terms of system stability, though the various nonlinearities behave somewhat differently in terms of convergence rate and minimization properties.

The next section formulates the problem and gives several examples in which nonlinearities in the data and/or error signals arise. Section III reviews certain key results on which the succeeding analyses are based. Section IV demonstrates the persistence of excitation condition for the generic LMS with nonlinearities. Section V interprets these conditions and shows a generic counterexample to stability whenever the data nonlinearity is nontrivial. Error nonlinearities are shown to be benign in terms of fulfillment of the persistence of excitation condition. Section VI examines several concrete examples, states the relevant excitation conditions and provides examples of and counterexamples to stability. Conclusions are presented in the final section.

II. Problem Formulation

The parameter update for the LMS adaptive algorithm is [7]

$$\hat{W}_{k+1} = \hat{W}_k + \mu X_k e_k \tag{2.1}$$

where $\hat{W}_k$ is the filter weight vector at time k, e_k is the scalar error sequence, and μ is the step size. X_k is the "regressor" vector

$$(x_k, x_{k-1}, \cdots, x_{k-N+1})^T$$

Manuscript received April 12, 1990; revised August 26, 1991.

The author is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706.

IEEE Log Number 9201585.

of input data x_k, and N is the number of taps in the filter. When the parameter update contains data and/or error nonlinearities, (2.1) becomes

$$\hat{W}_{k+1} = \hat{W}_k + \mu F(X_k) g(\hat{W}_k, X_k) \quad (2.2)$$

where F: $\boldsymbol{R}^N \rightarrow \boldsymbol{R}^N$ usually consists of N copies of the scalar real valued function f, and $g(\cdot)$ is typically a scalar function of the error e_k. The nonlinearities are memoryless, sign preserving, and are often assumed to be odd. The sign preservation property is crucial because it allows the scheme to descend (rather than climb) the error surface. The odd assumption is less important, but is common in practice since it weights positive and negative data equally.

A typical analytic trick is to suppose that there is a set of parameters W^* that will cause zero error. Introducing the parameter error vector $W_k = W^* - \hat{W}_k$ shifts the equilibrium to the origin, and the evolution of the parameter estimate error is

$$W_{k+1} = W_k - \mu F(X_k) g(e_k) \quad (2.3)$$

where $e_k = W_k^T X_k$ is the error between the output of the adaptive filter and the (fictitious) system W^*. Convergence of the weights to their optimum value $\hat{W}_k = W^*$ is thus equivalent to convergence of the parameter estimate error W_k to 0.

Such F's and g's can arise in several ways: as a natural consequence of a gradient minimization, as an attempt to lessen the numerical complexity of the algorithm, or as designer choice based on some *a priori* knowledge of the problem setting. Some concrete examples of F's and g's are:

Example 1: Suppose it is desired to minimize the cost function $J = e_k^4$ using a gradient procedure. Minimizing J by choice of W_k gives the algorithm

$$W_{k+1} = W_k - \mu \frac{dJ}{dW_k} \quad (2.4)$$

where dJ/dW_k can be easily calculated as $4e_k^3 X_k$. Thus F is the identity and $g(e) = e^3$. (Typically, the constant is absorbed into the stepsize parameter μ.) Somewhat more generally, attempting to minimize $J = |e_k|^\rho$ for arbitrary $\rho \neq 1$ leads to the nonlinear error function $g(e) = e^{\rho - 1}$. Note that LMS is the special case $\rho = 2$.

Example 2 (LMS with Dead Zone): One common modification to adaptive algorithms, especially in the adaptive control context [8], is to incorporate the error nonlinearity

$$g(e) = \begin{pmatrix} e - d & \text{if } e > d \\ e + d & e < -d \\ 0 & \text{otherwise} \end{pmatrix} \quad \text{for some } d > 0.$$

This g is insensitive to small errors, which are sometimes presumed to be due to noise.

Example 3 (Signed Regressor LMS [3], [9]): In order to simplify the number of computations per iteration of the adaptive algorithm (for higher throughput), one might consider an algorithm in which only the sign of the data is used, replacing the multiplications inherent in any implementation of LMS with bit shift operations. This is

$$W_{k+1} = W_k - \mu \operatorname{sgn}(X_k) e_k. \quad (2.5)$$

Hence g is the identity and F is the element by element signum function.

Example 4 (Cubic Data Multiplier): This variant is first suggested in [5], where it is noted that a cubic multiplier gives increased weight to large inputs rather than emphasizing small data as does the signed regressor algorithm. The algorithm is defined by an F which is composed of N copies of $f(x) = x^3$, and $g(e) = e$.

Example 5 (Backpropagation [6]): The backpropagation algorithm is the extension of LMS to the situation where a smooth nonlinearity $h(\cdot)$ is placed at the output of the linear combiner. (We consider here only the updates for the "output layer," those which can be directly compared to an error signal.) A cost function $J = \Sigma e^2$ is to be minimized via a gradient descent procedure as in (2.4). The sum is taken over all output units. For each unit, the algorithm updates the weights entering the unit by

$$W_{k+1} = W_k + \mu [h(X_k^T W^*) - h(X_k^T \hat{W}_k)] h'(X_k^T \hat{W}_k) X_k \quad (2.6)$$

where h' indicates the derivative of h with respect to its argument. Thus F is the identity, and $g(\hat{W}_k, X_k) = [h(X_k^T W^*) - h(X_k^T \hat{W}_k)] h'(X_k^T \hat{W}_k)$ is a function of both the error (between W^* and $\hat{W}_k$) and the current estimate $\hat{W}_k$.

Section VI returns to each of these examples and provides concrete situations in which these algorithms are locally stable and/or locally unstable.

III. Background

The key ideas which will be used are linearization, the slow time variation lemma [10], averaging [11], and total stability [12]. Linearization is used to examine the stability of the algorithm operating in a region about its equilibrium. This linearization is time varying (due to the data signal), and a slow time variation result can be used (the slowness is a consequence of the small step size μ) to relate the stability of the time varying system to the stability of the related frozen systems. Averaging is used to derive conditions under which the frozen systems are locally exponentially stable. The total stability theorem then translates the exponential stability result into robustness of the adaptive system to small disturbances, including small measurement noises, small nonlinearities, and slow parameter variation.

A. Linearization

Consider the discrete time system

$$z_{k+1} = \boldsymbol{F}(k, z_k) \quad (3.1)$$

where z_k is a state vector in $\boldsymbol{R}^N$, and $\boldsymbol{F}$ is a vector function $\boldsymbol{R}^N \rightarrow \boldsymbol{R}^N$ defining the evolution of the state. The states z^*

for which $F(k, z^*) = z^*$ for all k are the equilibria of (3.1), which we may assume without loss of generality to be located at the origin. F is linearized at the equilibrium $z^* = 0$ via the Jacobian $A_k = DF|_{z^*=0}$. The linearization theorem (Lyapunov's indirect method [13]) asserts that the behavior of (3.1) near z^* is dictated by the behavior of the related linear system

$$y_{k+1} = A_k y_k \tag{3.2}$$

that is, if (3.2) is exponentially asymptotically stable (e.a.s), then (3.1) is also e.a.s. The theorem holds assuming that A_k is bounded, and assuming that the norm of the difference $F(k, z) - A_k z$ is uniformly bounded in time. Formally, this requires that

$$\lim_{\|z\| \to 0} \max_k \frac{\|F(k, z) - A_k z\|}{\|z\|} = 0 \tag{3.3}$$

which essentially guarantees that time variation in the nonlinear terms of the Taylor series does not become arbitrarily large as time progresses.

B. Slow Time Variation and Averaging

The task of showing stability for the adaptive system is therefore translated to the simpler problem of finding conditions under which the linear, time-varying system (3.2) is e.a.s. One approach is to use the "slow time variation lemma" of [10] which asserts that if the change in A is slow enough (that is, $\|A_{k+1} - A_k\|$ is small), then exponential stability of each A_p (uniformly in p) is enough to imply e.a.s. of the time-varying system (3.2).

Unfortunately, the A_k matrices from the adaptive systems of interest are virtually never exponentially stable due to the structure of the problem. This implies that the desired systems A_p fail to be e.a.s. The approach of [11] takes a time average of (3.2), defining the "sliding average"

$$\bar{A}_k(m) = \frac{1}{m} \sum_{i=1}^{m} A_{k+i-1}. \tag{3.4}$$

If the eigenvalues of $\bar{A}_k(m)$ are (uniformly in k) less than one in magnitude for some m, and if the $\bar{A}_k(m)$ vary slow enough, then the averaged system

$$\bar{y}_{k+1} = \bar{A}_k \bar{y}_k \tag{3.5}$$

and the related (3.2) are both e.a.s. Fortunately, the sliding averages can be exponentially stable even when the A_k's are not.

C. Total Stability

The final step in the argument is to relax the assumption that there are no disturbances. The total stability theorem of [12] relates the behavior of the unforced system (3.1) to the behavior of

$$\bar{z}_{k+1} = F(k, \bar{z}_k) + G(k, \bar{z}_k) \tag{3.6}$$

where G is some small disturbance term that may depend on the state. Assuming that F is Lipshitz continuous, the difference between the state z of (3.1) and the state $\bar{z}$ of (3.6) can be bounded when F is known to be e.a.s. by requiring that G be suitably small and that the initial difference is small. Formally, for every ϵ, there is a δ_1 and δ_2 such that $\|\bar{z}_0 - z_0\| < \delta_1$ and $\|G(k, \bar{z}_k)\| < \delta_2$ for every k imply that $\|z_k - \bar{z}_k\| < \epsilon$ for every k. Thus, the system no longer converges to its equilibrium, rather, it converges to a ball about the equilibrium and then "rattles around." This disturbance term can be used to formally consider measurement disturbances, small nonlinearities, slow time variation of the parameters, and other small "nonidealities" that may arise.

IV. Persistence of Excitation for LMS with Nonlinearities

The above ideas can be used to examine the stability of the generic adaptive algorithm

$$W_{k+1} = W_k - \mu F(X_k) g(W_k^T X_k). \tag{4.1}$$

A. Assumptions

The following assumptions are made about the nonlinear functions F and g:

a) F and g are sign preserving,
b) F and g are memoryless,
c) $g(\cdot)$ is differentiable at the origin.

In addition, whenever convenient we shall assume

d) F: $\boldsymbol{R}^N \to \boldsymbol{R}^N$ consists of N copies of the scalar real valued function f,
e) $F(X)$ does not vanish as $X \to \pm\infty$,
f) F is piecewise continuous.

Assumption a is fundamental in the sense that if F or g were not sign preserving, this is equivalent to designing an algorithm to climb rather than to descend the error surface. This is also equivalent to reversing the sign of the step size μ. Assumption b is implicit in the formulation of F and g as functions of their specified arguments, but it is worthwhile noting because there is, perhaps, some interest in considering functions with memory. The linear case with memory is dealt with in [10] via techniques similar to those used here, and others have attacked this situation in other ways, see [14] and [15]. Assumption c assures that the linearization step is possible. Note that no differentiability (or continuity) is required on F, nor on g anywhere but at the origin. In certain of the examples below, d, e, and/or f are assumed, though this is usually more for notational convenience than out of any real necessity. Most of the nonlinear variants of LMS of which we're aware fulfill these requirements, though the "signed error" algorithm where $g(e) = \text{sgn}(e)$ fails condition c.

B. Linearization

Define the vectors $W_k = (w_k^1, w_k^2, \cdots, w_k^N)^T$ and $X_k = (x_k^1, x_k^2, \cdots, x_k^N)^T$, and the vector function $F(X_k) = (f_1(X_k), f_2(X_k), \cdots, f_N(X_k))^T$. Typically, X_k consists of a "regressor" vector of time shifted versions of a scalar sequence x_k, that is, $x_k^i = x_{k-1}^{i-1}$ for $i = 2, N$, but this is not necessary. Identify the function F of (3.1) with the right-hand side of (4.1), and let

$$H_k = F(X_k) g(W_k^T X_k) = \begin{pmatrix} f_1(X_k) & g(x_k^1 w_k^1 + x_k^2 w_k^2 + \cdots + x_k^N w_k^N) \\ f_2(X_k) & g(x_k^1 w_k^1 + x_k^2 w_k^2 + \cdots + x_k^N w_k^N) \\ & \cdots \\ f_N(X_k) & g(x_k^1 w_k^1 + x_k^2 w_k^2 + \cdots + x_k^N w_k^N) \end{pmatrix}. \tag{4.2}$$

Then the Jacobian can be calculated as

$$\frac{dH_k}{dW_k} = \begin{Bmatrix} f_1(X_k) x_k^1 g'(X_k^T W_k) & f_1(X_k) x_k^2 g'(X_k^T W_k) & \cdots & f_1(X_k) x_k^N g'(X_k^T W_k) \\ f_2(X_k) x_k^1 g'(X_k^T W_k) & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ f_N(X_k) x_k^1 g'(X_k^T W_k) & \cdots & \cdots & f_N(X_k) x_k^N g'(X_k^T W_k) \end{Bmatrix}. \tag{4.3}$$

When evaluated at the equilibrium $W^* = 0$, this simplifies to

$$B_k = \left.\frac{dH_k}{dW_k}\right|_{W^*=0} = g'(0) F(X_k) X_k^T \tag{4.4}$$

and the linearized system is

$$y_{k+1} = (I - \mu B_k) y_k. \tag{4.5}$$

The linearization result shows that if (4.5) is exponentially stable, then the original nonlinear system (4.1) is also exponentially stable.

C. Slow Time Variation and Averaging

Note that by choosing the step size parameter μ small, the time variation of the transition matrix $(I - \mu B_k)$ is slowed. In fact, as $\mu \to 0$, $\|(I - \mu B_{k+1}) - (I - \mu B_k)\| \to 0$. Consequently, the exponential stability of the time-varying linearized system can be translated via the slow time variation lemma to the exponential stability of the frozen (or time invariant) systems $(I - \mu B_p)$, for each p.

Unfortunately, due to the structure of B_k as a scaled product of two vectors, each B_k has rank at most 1, and so has $N - 1$ zero eigenvalues. This implies that $(I - \mu B_k)$ has $N - 1$ unity eigenvalues, and hence is not exponentially stable. Define the sliding average $\bar{B}_k(m)$ over the time window m as in (3.4). Then the averaging theorem demonstrates that exponential stability of

$$\bar{y}_{k+1} = (I - \mu \bar{B}_k(m)) \bar{y}_k$$

implies exponential stability of (4.5), and hence (4.1). Define the excitation matrix

$$M_t = \sum_{k=1}^{t} F(X_k) X_k^T \tag{4.6}$$

which, for t-periodic inputs is equal to the sliding average. Then the magnitude of all eigenvalues of $(I - \mu M_t)$ can be guaranteed less than one as long as M_t has no eigenvalues with positive real part, and as long as μ is chosen small enough. Gathering the above results together shows the following theorem.

Theorem: Consider the algorithm (4.1) with t-periodic input data X_k and nonlinear elements F and g, under assumptions b and c. If there are $\alpha > 0$ and $\beta > 0$ such that

$$\beta > g'(0) \operatorname{Re} \lambda_i(M_t) > \alpha \qquad \text{for every } i \tag{4.7}$$

then there is a μ^* such that for every μ in $(0, \mu^*)$, the algorithm (4.1) is locally exponentially stable about its equilibrium $W^* = 0$. Conversely, if $g'(0) \operatorname{Re} \lambda_i(M_t)$ is negative for some i, then the algorithm (4.1) is locally unstable about its equilibrium at $W^* = 0$.

(The notation $\operatorname{Re} \lambda_i(M)$ means the real part of the ith eigenvalue of the matrix M.)

Remarks:

a) Local exponential stability of the algorithm implies that the parameter estimate error W_k converges to 0 if it is initialized in some region about 0. Convergence of the parameter estimate error to zero is equivalent to the convergence of the parameter estimates $\hat{W}_k$ to their true values W^*. Local instability implies that there are arbitrarily small perturbations that can drive the parameter estimates away from W^*. This does not necessarily imply divergence to infinity of the parameter estimates.

b) The condition (4.7) is called the persistence of excitation (PE) condition for the LMS algorithm with nonlinearities. Note that the condition involves the input data sequence X_k as well as the data nonlinearity F and the derivative of the error nonlinearity g at the origin.

c) The importance of the sign preservation property of F is apparent from the persistence of excitation condition, since if F reverses the sign of the data, then the right-hand inequality of (4.7) fails. Similarly, $g'(0)$ must be positive.

d) If $g'(0) = \infty$ then assumption c and the left-hand inequality of (4.7) fails. In particular, this averaging approach is inapplicable to the signed error algorithm with $g(e) = \operatorname{sgn}(e)$. An extended Lyapunov approach can be found in [9].

e) The convergence rate of the averaged system (4.6) (and hence the convergence rate of the algorithm (4.1)) is proportional to the size of the real part of the smallest eigenvalue of (4.7). Thus, given an input sequence X_k, if α is chosen as large as possible, the convergence rate is dictated by α. Since $g'(0)$ is directly proportional to α, increasing the slope of g near the origin will tend to increase the convergence rate, if other parameters are held fixed, provided that the left-hand inequality in (4.7) is not violated.

f) The periodicity assumption is not necessary, and can be relaxed to "almost periodic" inputs as in [16] at the expense of a large amount of technical detail.

g) The fact that (4.7) depends on the function g only at the origin emphasizes the local nature of the results; initial conditions must be chosen so that g remains in a small ball about 0.

Suppose that $g'(0) = 0$, as occurs in example 2 and in example 1 for certain ρ. Then the right-hand side of the persistence of excitation condition (4.7) fails, and the algorithm is not exponentially stable about $W^* = 0$. If, however, g is nondecreasing, continuous, and differentiable at the endpoints of some region R, then there is hope that the parameter estimate errors will converge to the region R rather than to W^* itself. To make this notion more precise, consider the following definition.

Definition: The system $x_{k+1} = f(k, x_k)$ is said to be (uniformly) locally exponentially stable to the compact region R contained in B if $\exists \gamma \in (0, 1)$ and an $N > 0$ such that $\forall X_0 \in B$, $d(X_k, R) < N \|X_0\| \gamma^k \forall k$, where the distance from the point X_k to the set R is defined as $d(X_k, R) = \min_{r \in R} \|X_k - r\|$.

Note that this minimum exists when R is compact, and that the definition reduces to the standard definition of (local, uniform) exponential stability when R consists of an isolated equilibrium. The following corollary simply extends the theorem to include the case of convergence to a region, rather than a point.

Corollary: Consider the algorithm (4.1) with t-periodic input data X_k and nonlinear elements F and g under assumptions b and c. Suppose further that g is nondecreasing and continuous in a region $R = [-r, r]$, that $g'(0) = 0$, that $g'(r)$ and $g'(-r)$ exist and are positive, and that $\exists \alpha > 0$ and $\beta > 0$ such that $\beta > \text{Re } \lambda_i(M_t) > \alpha \forall i$. Then there is a μ^* such that for every $\mu \in (0, \mu^*)$, the algorithm is locally exponentially stable to the region R.

D. Total Stability

The final step is to remove the "ideal" assumption, and to suppose that some small nonidealities are present. The $\boldsymbol{F}$ and $\boldsymbol{G}$ of (3.6) may be related to the LMS with nonlinearities by identifying the state $\bar{z}_k$ with the parameter estimate errors W_k, and $\boldsymbol{G}$ with the disturbance term. Assuming that the input data fulfills the PE condition (4.7), then the homogeneous system (4.1) (and (3.1) with $\boldsymbol{F}$ identified as the right-hand side of (4.1)) is exponentially stable. Consequently, the total stability theorem asserts that for small disturbances $\boldsymbol{G}$, the perturbed system will remain within an ϵ ball about the origin. This has several implications:

1) Robustness to small measurement noises. Suppose that a bounded measurement disturbance η_k corrupts the prediction error e_k. Then $\boldsymbol{G}(k, W_k) = \mu F(X_k)[g(e_k + \eta_k) - g(e_k)]$, and the norm of $\boldsymbol{G}$ can be bounded in terms of μ, $\|F\|$, $\|X_k\|$, and the smoothness of g. Hence, if $\|\eta_k\|$ is small so that $\|\boldsymbol{G}\| < \delta_2$, the total stability theorem of Section III applies, showing that an algorithm that is exponentially stable cannot be destabilized by arbitrarily small measurement biases or inaccuracies.

2) Robustness to undermodeling. Suppose that the N-dimensional W^* is only an approximation to the "true" plant, which is $N + M$ dimensional. If this undermodeling is not too severe (if there is an N-dimensional W^* that is a good approximation to the true plant), then the algorithm retains stability. In this case, η_k represents the difference between the output of the true $N + M$ dimensional plant and the output due to W^*. As in (1), if this η_k is small, then the perturbed system is stable.

3) Robustness to small nonlinearities. Suppose that the linear W^* is only an approximation to the "real" plant which contains small nonlinearities. If η_k represents the output due to these nonlinearities, and if this is kept small, then the algorithm retains stability.

The above three robustness results are related in that the nonidealities enter as an additive disturbance corrupting the prediction error.

4) Robustness to slow time variations. The "real" plant may actually vary with time. If these time variations are slow enough, then the exponentially stable algorithm will track the motion and remain stable. Let W_k^* represent the time-varying plant, and suppose that $\|W_{k+1}^* - W_k^*\|$ is small. The error system is becomes

$$W_{k+1} = W_k - \mu F(X_k) g(W_k^T X_k) + (W_{k+1}^* - W_k^*). \quad (4.8)$$

Letting $\boldsymbol{G} = (W_{k+1}^* - W_k^*)$, and bounding the rate of variation by $\|\boldsymbol{G}\| < \delta_2$ shows that the algorithm retains stability.

V. Interpretation of the Excitation Conditions

This section examines the persistence of excitation (PE) condition for LMS with nonlinearities in several ways. First, it is compared with the standard PE for LMS condition, and it is shown to be strictly more difficult to fulfill the PE for nonlinear LMS than PE for (linear) LMS. Second, a generic counterexample is provided, showing that whenever F is nonlinear, there are input sequences that will fail the PE condition and destabilize the algorithm. These are interpreted in terms of a misalignment from any reasonable "gradient" direction. Error nonlinearities enter in a relatively benign fashion.

The standard PE condition for LMS [17] (without nonlinearities), when excited by t-periodic inputs X_k, is that there exist $\alpha > 0$ and $\beta > 0$ such that

$$\beta I > \sum_{k=1}^{t} X_k X_k^T > \alpha I. \quad (5.1)$$

As above, this implies local exponential stability of the error system. Since the matrix in (5.1) is symmetric, all eigenvalues are real, and the notation ">" means positive definite. How does (5.1) compare to (4.7)?

Lemma 1: Suppose that (4.7) holds for a given F, g, and input sequence X_k, and that F fulfills assumption e of Section IV. Then (5.1) also holds.

Proof: By contradiction. There are two possibilities:

i) If (5.1) fails the upper bound, then X_k must be diverging. By assumption e, this implies that $F(X_k)X_k^T$ must diverge. Hence (4.7) is unbounded.

ii) If (5.1) fails the lower bound for every positive α, then there must be a zero eigenvalue of $\Sigma\, XX^T$. Consequently, there must be a nonzero eigenvector v such that $v^T X_k = 0$ for every k. This implies that $(\Sigma\, F(X)X^T)v = 0$, and so there is a zero eigenvalue of the matrix in (4.7). ■

This says that if the nonlinear LMS algorithm is persistently excited (4.7), then the standard LMS algorithm is also persistently excited (5.1). We will show that for a large class of F, the reverse implication is false, by constructing particular input sequences X_k for which $\Sigma\, F(X)X^T$ has negative eigenvalues. These input sequences cause unstable parameter updates. The following pair of technical lemmas will be useful in constructing such X_k.

Lemma 2: Suppose that f: $\boldsymbol{R} \to \boldsymbol{R}$ is sign preserving, piecewise continuous, and is not a linear function. Then there are positive x, y, z such that

$$x > y + z \quad \text{and} \quad f(x) < f(y) + f(z) \tag{5.2}$$

or there are positive x, y, z such that

$$x < y + z \quad \text{and} \quad f(x) > f(y) + f(z). \tag{5.3}$$

Proof: By contradiction. We show that if both (5.2) and (5.3) fail, then f must actually be linear. Suppose that for every positive x, y, z, $x > y + z$ implies that $f(x) > f(y) + f(z)$ and $x < y + z$ implies that $f(x) < f(y) + f(z)$. Pick a point x at which f is continuous, and let y and z be any two points with $x = y + z$. Let $\boldsymbol{x}_k \to x$ from below, and $\bar{x}_k \to x$ from above. Then for every k, $f(\boldsymbol{x}_k) < f(y) + f(z)$ and $f(\bar{x}_k) > f(y) + f(z)$. The continuity of f at x implies that $f(x) = f(y) + f(z)$. If f is actually continuous everywhere, this implies linearity, and gives the contradiction. If f is only piecewise continuous, then f can be approximated arbitrarily closely by a continuous function $\bar{f}$. The same argument repeated for $\bar{f}$ gives the desired contradiction. □

Lemma 3: With $f(x)$ as in Lemma 2, for any integer n, there are positive x and y such that

$$x > ny \quad \text{and} \quad f(x) < nf(y) \tag{5.4}$$

or

$$x < ny \quad \text{and} \quad f(x) > nf(y). \tag{5.5}$$

Proof: Note that Lemma 2 holds for $y = z$, and iterate the argument n times. □

By choosing values of x (and y) based on the function F and the dimension N, it is possible to construct periodic input sequences that will destabilize the nonlinear LMS algorithm for virtually any data nonlinearity F.

Lemma 4: Suppose that F: $\boldsymbol{R}^N \to \boldsymbol{R}^N$ (with $N > 2$) fulfills assumptions a, b, d and f, and that F is not a linear map. Then there is an N periodic data sequence X_k, $k = 1, 2, \cdots, N$ such that $\text{Re}\, \lambda_i(M_N) < 0$, where M_N is the excitation matrix (4.6).

Proof: By construction. Any symmetric matrix M_N will have a negative eigenvalue if there is a V such that

$$V^T M_N V < 0. \tag{5.6}$$

We construct a matrix M_N of the form (4.6), find an appropriate V, and then verify that $M_N^T = M_N$. From the definition of M_N, (5.6) can be rewritten

$$V^T M_N V = V^T(\Sigma\, F(X)X^T)V = \Sigma\, V^T F(X) X^T V < 0. \tag{5.7}$$

One way this could occur is if $V^T F(X_k)$ has an opposite sign from $X_k^T V$ for every k. Let $V = (1, 1, \cdots, 1)^T$. To construct such an X_k sequence, let $X_k = (x_k^1, x_k^2, \cdots, x_k^N)^T$ and $F(X_k) = (f(x_k^1), f(x_k^2), \cdots, f(x_k^N))^T$. Choose X_1 such that

$$|x_1^1| > |x_1^2| + |x_1^3| + \cdots + |x_1^N| \quad \text{with } x_1^1 > 0,\ x_1^i = y < 0,\quad i = 2, \cdots, N$$

and

$$|f(x_1^1)| < |f(x_1^2)| + |f(x_1^3)| + \cdots + |f(x_1^N)| \tag{5.8}$$

or choose X_1 such that

$$|x_1^1| < |x_1^2| + |x_1^3| + \cdots + |x_1^N| \quad \text{with } x_1^1 > 0,\ x_1^i = y < 0,\quad i = 2, \cdots, N$$

and

$$|f(x_1^1)| > |f(x_1^2)| + |f(x_1^3)| + \cdots + |f(x_1^N)|. \tag{5.9}$$

Note that either (5.8) or (5.9) is always possible by Lemma 3, provided $f(\cdot)$ is not linear. By the sign preservation property, $f(x_1^1) > 0$ and $f(y) < 0$ for $i = 2, \cdots, N$. In either case, $X_1^T V$ is the sum of elements of X_1, $V^T F(X_1)$ is the sum of elements of $F(X_1)$, and $\text{sgn}\,(X_1^T V) = -\text{sgn}\,(V^T F(X_1))$. Now construct an N-periodic X_k sequence by $X_{k+1} = QX_k$ where Q is the permutation matrix

$$\begin{pmatrix} 010 \cdots 0 \\ 001 \cdots 0 \\ 0001 \cdots 0 \\ \cdots \\ 100 \cdots 0 \end{pmatrix}.$$

Then $\text{sgn}\,(X_k^T V) = -\text{sgn}\,(V^T F(X_k))$ for every k, which implies (5.7) holds. To see that the M_N constructed is actually symmetric, observe that $X_1 = (x_1^1, y, \cdots, y)$ and $F(X_1) = (f(x_1^1), f(y), \cdots, f(y))$. Thus, M_N is a matrix in which all diagonal elements are equal to $x_1^1 f(x_1^1) + (N - 1)yf(y)$ and all off-diagonal elements are equal to $x_1^1 f(y) + yf(x_1^1) + (N - 2)yf(y)$. ■

Remarks:

a) Note that this result is not true for $N < 3$. For $N = 1$, the sign preservation of F is enough to guarantee that fulfillment of (5.1) implies fulfillment of the persistence of excitation condition for nonlinear LMS (4.7). For $N = 2$, a monotonicity assumption on F is enough to insure that the choice of X_1 in (5.8) and (5.9) cannot be made. Most applications require many more than 2 parameters.

b) The possibility of instability with nonlinear F is understandable from a gradient point of view since virtually any cost function leads to a gradient that points in the direction of X. By manipulating X in such a way as to change its direction "often enough," the algorithm can be made to climb, rather than to descend the error surface. This was the inspiration for the construction in Lemma 4.

c) As one might expect, if F is "nearly" linear, then such examples of instability will be highly unlikely to occur. In fact, if the maximum error between $F(X)$ and X can be bounded by some δ_1, then one can find a δ_2 such that if α of (5.1) is greater than δ_2, then the algorithm retains stability. This can be shown as in [18].

This is a "generic" counterexample showing that it is strictly more difficult to insure that the nonlinear algorithm "works" than to insure that LMS "works." Moreover, this instability is due solely to the presence of the data nonlinearity F, and not to the error nonlinearity g. Indeed, suppose that $F(X) = X$ is the identity map. Then (4.7) is a scaled version of (5.1) with scaling factor $g'(0)$. As long as $g'(0)$ is finite and positive, $g(\cdot)$ does not affect the stability of the algorithm. The primary effect of $g(\cdot)$ in the ideal case is to speed up or slow down the asymptotic convergence. Stated simply, it is fine to manipulate the error term as long as the sign is preserved, but it is dangerous to tamper with the gradient calculation.

The error nonlinearity $g(\cdot)$ can often be viewed as determining the function of the error that the algorithm is attempting to minimize. In example 1, for instance, the error function $g(e) = e^{\rho - 1}$ (for $\rho \neq 1$) corresponds to a minimization problem with cost function $J = |e_k|^\rho$. Similarly, whenever $f(X) = X$ and $g(\cdot)$ is integrable with respect to its argument, the algorithm tends to minimize $\int g(\cdot)$.

VI. Examples

Several concrete examples of the stability and instability results are presented, demonstrating that many nonlinear versions of LMS can be analyzed via the present techniques.

Example 1 Revisited: The algorithm (2.4) is designed to minimize $J = e_k^4$ using a gradient procedure. Since $F(X) = X$ and $g(x) = e^3$, $g'(0) = 0$, and the persistence of excitation condition for this algorithm is $\Sigma\, XX^T$ (by the corollary). Convergence is exponential to the region $[-r, r]$, and $r > 0$ can be chosen arbitrarily small. For the more general situation when $g(e) = e^{\rho - 1}$, the stability depends on the value of ρ. For $\rho > 1$, the corollary shows stability about the region $[-r, r]$ as above. For $\rho < 1$, $g'(0) = \infty$. The origin is unstable, and trajectories are repelled.

Example 2 Revisited: LMS with the dead zone modification also falls under the conditions of the corollary. Thus, if $\Sigma\, XX^T$ has all positive eigenvalues, the algorithm will converge to $[-r, r]$ for any $r > d$, where d is the dead zone parameter.

Example 3 Revisited: The persistence of excitation condition for the signed regressor algorithm is $\Sigma \operatorname{sgn}(X)X^T$ by the theorem. Lemma 4 gives a way of constructing input sequences X_k which cause the algorithm to be locally unstable. For instance, for $X \in \mathbf{R}^3$, the three periodic input $[3, -1, -1]$ fulfills condition (5.8) of Lemma 4 and it is easy to verify that the excitation matrix has a negative eigenvalue. Simulations of this algorithm with this input "diverge" (give overflow errors) from any initial condition (other than the unstable equilibrium itself).

Example 4 Revisited: The theorem shows that the persistence of excitation condition for the cubed data algorithm is $\Sigma\, X^3X^T$. Although many inputs will cause convergence of this algorithm, Lemma 3 allows construction of simple 3 periodic inputs (for the $X \in \mathbf{R}^3$ case) that cause instability. One such example is $[1.5, -1, -1]$, which fulfills condition (5.9) of Lemma 4 and has a negative eigenvalue. As in the previous example, simulations of the data cubed algorithm with this input grow rapidly.

Example 5 Revisited: This algorithm does not fulfill the exact conditions of the theorem, but a similar result is easy to derive. With $g(\cdot)$ and $h(\cdot)$ as defined in (2.6),

$$\frac{dg(\hat{W}_k, X_k)}{d\hat{W}_k} = \frac{d}{d\hat{W}_k}\,[-h(X_k^T\hat{W}_k)h'(X_k^T\hat{W}_k) + h(X_k^TW^*)h'(X_k^T\hat{W}_k)]$$
$$= [(h'(X_k^T\hat{W}_k))^2 - h(X_k^T\hat{W}_k)h''(X_k^T\hat{W}_k)]X_k - h(X_k^TW^*)h''(X_k^T\hat{W}_k)X_k.$$

Consequently, with $H_k = g(\hat{W}_k, X_k)X_k^T$,

$$\frac{dH_k}{d\hat{W}_k} = [(h(X_k^TW^*) - h(X_k^T\hat{W}_k))h''(X_k^T\hat{W}_k) - (h'(X_k^T\hat{W}_k))^2]X_kX_k^T.$$

Following the logic of Section IV shows that the appropriate persistence of excitation condition is that there are $\beta > \alpha > 0$ such that

$$\beta I > \Sigma \frac{dH_k}{d\hat{W}_k} > \alpha I \tag{6.1}$$

over the period of the input pattern X_k. That is, fulfillment of (6.1) implies local exponential stability of the algorithm.

If (5.1) fails, then (6.1) will also fail, which indicates the necessity that the data sequence X_k span the space regularly. Equation (6.1) also requires that $[(h(X_k^TW^*) -$

$h(X_k^T \hat{W}_k))h''(X_k^T \hat{W}_k) - (h'(X_k^T \hat{W}_k))^2]$ be negative. This will occur when the second term dominates, that is, when the error $h(X_k^T W^*) - h(X_k^T \hat{W}_k)$ between the "real" system W^* and the estimated system parameters $\hat{W}_k$ is small.

This emphasizes the local nature of the result, that the initial parameter estimates must not be too far from W^* in order to insure convergence. Equation (6.1) also explains the very slow observed convergence times of backpropagation algorithms, since the magnitude of $(h'(X_k^T \hat{W}_k))^2$ is small for most sigmoidal functions $h(\cdot)$. No doubt the multiple layer case can be handled similarly, though this requires suitable identifiability assumptions on the unknown system W^*.

VII. Conclusions

Persistence of excitation conditions are derived for several common nonlinear variants of the LMS algorithm. These provide conditions under which the error system is locally exponentially stable. When there is a nonlinear data function $F(\cdot)$, a generic counterexample is constructed for which the algorithms are (locally) unstable. Nonlinear error functions $g(\cdot)$ cannot affect the stability of the error system. Rather, they define the cost function the algorithm tends to minimize, and are related (through their derivative at zero) to the asymptotic convergence rate. The ideal (no disturbance) assumption is relaxed via application of the total stability theorem. Several specific examples illustrate the method.

References

[1] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, no. 8, pp. 1151-1162, Aug. 1976.

[2] N. J. Bershad, "On the optimum data nonlinearity in LMS adaptation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, Jan. 1986.

[3] J. L. Moschner, "Adaptive equalization via fast quantized state methods," Tech. Dep. 6796-1, Inform. Syst. Lab., Stanford Univ., 1970.

[4] W. A. Sethares, I. M. Y. Mareels, B. D. O. Anderson, and C. R. Johnson, Jr., "Excitation conditions for sign-regressor LMS," *IEEE Trans. Circuits Syst.*, June 1988.

[5] D. L. Duttweiler, "Adaptive filter performance with nonlinearities," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, Aug. 1982.

[6] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*. Cambridge, MA: M.I.T. Press, 1986.

[7] C. R. Johnson, Jr., *Lectures on Adaptive Parameter Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[8] P. A. Ioannou and P. V. Kokotovic, *Adaptive Systems with Reduced Models*. Springer-Verlag, 1983.

[9] W. A. Sethares and C. R. Johnson, Jr., "A comparison of two quantized state adaptive algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, Jan. 1989.

[10] W. A. Sethares, B. D. O. Anderson, C. R. Johnson, Jr., "Adaptive algorithms with filtered regressor and filtered error," *Math. Contr., Signals, Syst.*, vol. 2, pp. 381-403, 1989.

[11] R. R. Bitmead and C. R. Johnson, Jr., "Discrete averaging principles and robust adaptive identification," in *Control and Dynamic Systems: Advances in Theory and Applications*, vol. 24, C. T. Leondes, Ed. New York: Academic, 1986.

[12] W. Hahn, *Stability of Motion*. Springer-Verlag, 1967.

[13] M. Vidyasagar, *Nonlinear Systems Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[14] I. D. Landau, *Adaptive Control: The Model Reference Approach*. Marcel Dekker, 1979.

[15] D. A. Lawrence and C. R. Johnson, Jr., "Recursive parameter identification algorithm stability analysis via π-sharing," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 1, Jan. 1986.

[16] B. D. O. Anderson, R. R. Bitmead, C. R. Johnson, Jr., P. V. Kokotovic, R. L. Kosut, I. M. Y. Mareels, L. Praly, and B. D. Reidle, *Stability of Adaptive Systems: Passivity and Averaging Analysis*. New York: M.I.T. Press, 1986.

[17] R. R. Bitmead, "Persistence of excitation conditions and the convergence of adaptive systems," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 2, Mar. 1984.

[18] G. A. Williamson and W. A. Sethares, "Robustness to quantization errors in LMS adaptation via degree of excitation," in *Proc. IEEE 1990 Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990.

William A. Sethares (S'84–M'86) received the B.A. degree in mathematics from Brandeis University, Waltham, MA, and the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY.

He has worked at the Raytheon Company as a Systems Engineer and is currently on the faculty of the Department of Electrical and Computer Engineering at the University of Wisconsin in Madison. His research interests include adaptive systems in signal processing, communications, and control.