The Least Mean Square Family

William A. Sethares

October 8, 1999

In the beginning, Gauss created Least Squares, and he saw that it was good. So he said unto his algorithm, "be fruitful, and multiply." And it was so. Least Squares begat the Least Mean Squares (LMS), whose years were plenty. And Least Mean Squares begat Normalized LMS whose fame was projected throughout the land. And Normalized LMS begat a Leaky LMS, whose offspring were the Signed LMS and the Quantized LMS. And these children of the Sign proliferated. This is the story of LMS and its children.

1 Introduction

In chapter 2, the Least Mean Square (LMS) algorithm was introduced as a way to recursively adjust the parameters $\theta(k)$ of a linear filter with the goal of minimizing the error between a given desired signal and the output of the linear filter. LMS is one of many related algorithms which are appropriate for this task and a whole family of algorithms has been developed which can address a variety of problem settings, computational restrictions, and minimization criteria. This chapter begins by deriving the LMS as an instantaneous approximation to the steepest descent minimization of a cost function $V(\theta(k))$, which results in simple recursive scheme of the form

$$\left\{ \begin{array}{c} new \\ parameter \\ estimate \end{array} \right\} = \left\{ \begin{array}{c} old \\ parameter \\ estimate \end{array} \right\} + \{stepsize\} \left\{ \begin{array}{c} new \\ information \end{array} \right\} \tag{1}$$

where the *new information* is a function of the past and present inputs (often concatenated into a vector called the *regressor* vector) as well as the error between the output of the linear filter and the desired signal. Some general observations are made regarding the positive (and negative) aspects of the performance of LMS, and a variety of "children of LMS" are introduced as attempts to alleviate certain problems or to fine tune some aspect of the algorithms performance. These algorithms have the general recursive form

$$\theta(k) = \theta(k-1) + \mu(k-1)F(\phi(k-1))g(e(k)) \tag{2}$$

where $\theta(k)$ represents the new parameter estimate at time k, $F(\cdot) : \Re^m \rightarrow \Re$ and $g(\cdot) : \Re \rightarrow \Re$ are functions of the regressor vector $\phi(k)$ and the error signal $e(k)$, respectively, and the stepsize $\mu(k-1)$ is a user choosable parameter that may vary with time.

All of the algorithms of the "LMS Family" are special cases of (2). This usage is somewhat unfortunate, however, because many of these variants of the LMS algorithm do not actually minimize the least mean square error. The signed-error variant, for instance, tends to minimize the absolute value of the error. With a leakage parameter, the algorithm tends to minimize a linear combination of the least mean square error and the squared error away from some nominal θ^0. Other variants do not admit a minimization interpretation at all. A large body of literature has been devoted to the analysis of the behaviour of the various members of the LMS family. Sections 3, 4 and 5 discuss three major branches of this investigation.

The first of these methods, an attempt to develop a "statistical theory of adaptation" was initiated by Widrow and his co-authors in [52], [53]. By examining the expected behaviour of the algorithm under a variety of assumptions on the input and desired signals, they develop useful guidelines for implementing the algorithms (including optimal choice of stepsizes) and derive expressions describing the statistical performance of the algorithms.

A second analytical technique, the "deterministic approach," treats the parameter error update equation as a nonlinear dynamical system, and uses the tools of stability theory to examine the convergence and stability characteristics of the algorithms. Observe that (2) achieves an "averaged equilibrium" whenever $avg\{F(\phi)g(e)\} = 0$. Assuming that there is some ideal system θ^* that can exactly match the dynamics of the desired signal, suitable conditions (known as "persistence of excitation" conditions) on the character of the input sequences can be derived which imply that the parameter estimates $\theta(k)$ converge to this ideal θ^*, and that this convergence occurs in a neighborhood of the desired equilibrium. Interestingly, the PE conditions are different for the various members of the LMS family. For instance, there are inputs for which the signed error algorithm converges yet the signed-regressor variant diverges. The ideas of total stability extend these convergence/divergence results to the more realistic "nonideal" scenario, when disturbances are present, that is, when the desired input/output mapping cannot be represented exactly as a linear system parameterized by θ^*.

The third analytical technique is the "stochastic approximation" approach, pioneered by L. Ljung in [38], in which the parameter error update equation is examined indirectly by

studying a related Ordinary Differential Equation (ODE). In particular, local stability of the ODE implies weak convergence of the algorithm. Though the analysis in [38] requires a vanishing stepsize (where $\mu \to 0$ as time $\to \infty$), this restriction may be removed as shown in [33] and [4]. Relating the motion of the parameter error to an associated *forced* differential equation allows the convergent (stationary) distributions to be expressed in a concrete manner. Think of it this way: the generic behaviour of LMS and its variants is that the parameter estimates converge to a region about their final value value, and then "rattle around" this value due to unavoidable noises and disturbances. The beauty of the stochastic approximation approach is that this rattling behaviour can often be described in terms of specific probability distributions.

While other ways of understanding the various members of the LMS family exist, these three are (so far) the most widely known and the most powerful. The analytical techniques are complementary, and each offers unique insights into the behaviours and performance of the algorithms. Section 6 shows how to apply the various analytical techniques to the children of LMS, and section 7 wraps up the discussion by posing a number of open questions.

2 LMS and its Children

The Least Mean Square (LMS) Algorithm, popularized by Widrow in [52] and [53], has become one of the standard techniques of adaptive filtering. The LMS algorithm is a form of steepest (or gradient) descent that attempts to minimize a cost function $V(\theta(k))$ at each time step k by a suitable choice of the parameter vector $\theta(k)$. The strategy is to update the parameter estimate proportional to the instantaneous gradient value $\frac{dV(\theta(k-1))}{d\theta(k-1)}$, that is,

$$\theta(k) = \theta(k-1) - \mu \frac{dV(\theta(k-1))}{d\theta(k-1)} \tag{3}$$

where μ is a small positive stepsize, and the minus sign insures that the parameter estimates descend (rather than climb) the error surface. (Throughout this chapter, the time index k is used for discrete time processes while the variable t is reserved for continuous time processes.)

If the adaptive filter has a linear structure, then its output can be expressed as

$$\hat{y}(k) = \phi^T(k-1)\theta(k-1) \tag{4}$$

where $\phi(k-1)$ is a vector of past and present inputs (and possibly past outputs). For instance, given an input sequence $u(k-1)$, the input vector is $\phi(k-1) = (u(k-1), u(k-2), ..., u(k-m))^T$. Choosing the cost function to be one half the square of the error between the output of the adaptive filter and the desired signal

$$V(\theta(k-1)) = \frac{1}{2}(y(k) - \hat{y}(k))^2 = \frac{1}{2}(y(k) - \phi^T(k-1)\theta(k-1))^2, \tag{5}$$

the gradient is

$$\frac{dV(\theta(k-1))}{d\theta(k-1)} = -(y(k) - \hat{y}(k))\phi(k). \tag{6}$$

The LMS algorithm is then

$$\theta(k) = \theta(k-1) + \mu\phi(k-1)(y(k) - \hat{y}(k)). \tag{7}$$

If there is a fixed vector θ^* such that the desired signal $y(k)$ is generated from a linear system with parameterization θ^*

$$y(k) = \phi^T(k-1)\theta^* + \xi(k) \tag{8}$$

where $\xi(k)$ represents the unmodelable noise component in the desired signal, then (7) can be rewritten

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) - \mu\phi(k-1)\phi^T(k-1)\tilde{\theta}(k-1) + \mu\phi(k-1)\xi(k) \tag{9}$$

$$= (I - \mu\phi(k-1)\phi^T(k-1))\tilde{\theta}(k-1) + \mu\phi(k-1)\xi(k) \tag{10}$$

where $\tilde{\theta}(k) = \theta^* - \theta(k)$ is the parameter estimate error. This is called the *error system* and is primarily useful for analysis, since the behaviour of the parameter error $\tilde{\theta}(k)$ about the origin describes exactly the behaviour of the parameter estimates $\theta(k)$ about the true parameterization θ^*.

The LMS algorithm (7) has been successfully used in numerous applications throughout the years [54], [28], [30], and it has been analyzed extensively [6], [8]. Some notable aspects of its performance are

- LMS tends to reject noisy data due to the smoothing action of the small stepsize parameter μ.
- LMS can track slowly time varying systems, and is often useful in nonstationary environments.
- The LMS error function has a unique global minimum, and hence the algorithm does not tend to get stuck at undesirable local minima.
- LMS is computationally simple (m multiplies and m adds per iteration) and memory efficient (only one m-vector must be stored).
- The convergence of LMS is often slow (it may take hundreds or thousands of iterations to converge from an arbitrary initialization).
- LMS is susceptible to problems with noise during periods when the input fails to excite all the modes of the system.

Successful algorithms tend to breed closely related variants which attempt to alleviate problems or to fine tune some aspect of performance. LMS is no exception. These variants range from the Normalized LMS (designed to speed convergence) to leakage (which combats potential numerical problems during periods of high noise or low excitation) to the signed algorithms (which further simplify the numerical requirements) to the dual sign algorithm (and other quantized versions which attempt to simplify the numerics without sacrificing convergence speed) to "high order" algorithms (which minimize l^p norms for p greater than 2) to the median LMS and other order statistic algorithms (which attempt to optimize LMS for use in impulsive environments).

2.1 Normalized LMS

The desire to have fast convergence from an arbitrary initial state requires a large stepsize. This conflicts with the desire to have significant smoothing of the noise signal in steady state, which requires a small stepsize. An obvious algorithm modification uses a large stepsize initially and then switches to a small stepsize when in the region of the correct solution. The

Normalized LMS (NLMS) algorithm provides one way to automate this choice of varying stepsize.

It is easy to see from (10) that the stepsize μ must be less than $\frac{2}{\phi^T(k-1)\phi(k-1)}$ at each time step k, or instability may result since the term $(I - \mu\phi(k-1)\phi^T(k-1))$ will be an expansion rather than a contraction, and the solution of the difference equation (10) will tend to diverge. At each time step, the algorithm moves a distance $\mu(y(k) - \hat{y}(k))$ in the $\phi(k-1)$ direction. An "optimal" distance to move would be to set $\mu(k-1) = \frac{1}{\phi^T(k-1)\phi(k-1)}$, since then the term $(I - \mu\phi(k-1)\phi^T(k-1)) = (I - \frac{\phi(k-1)\phi^T(k-1)}{\phi^T(k-1)\phi(k-1)})$ is maximally contractive (with an eigenvalue exactly equal to zero) in the direction of the eigenvector $\phi(k-1)$. As a practical matter, the stepsize is often set to $\mu(k-1) = \frac{\mu}{1+\mu\phi^T(k-1)\phi(k-1)}$ where μ is chosen small enough to encourage smoothing in the steady state and the 1 avoids division by zero in the event that $\phi^T(k-1)\phi(k-1) = 0$. This leads to the update

$$\theta(k) = \theta(k-1) + \frac{\mu\phi(k-1)(y(k) - \hat{y}(k))}{1 + \mu\phi^T(k-1)\phi(k-1)}. \tag{11}$$

Thus, rather than taking small steps as in (7), the parameter estimates of (11) are projected onto the subspace complementary to $\phi(k-1)$. Consequently, the Normalized LMS is also called the "projection algorithm."

2.2 Leakage

The possibility of sensitivity to roundoff errors and other parasitic disturbances exists because the LMS update equation (7) is essentially an integrator. The introduction of a small *leakage* parameter $\lambda \in (0, 1)$

$$\theta(k) = (1 - \lambda)\theta(k-1) + \mu\phi(k-1)(y(k) - \hat{y}(k)) \tag{12}$$

can guard against such numerical problems. The effect of λ on the behaviour of the algorithm is seen most clearly by transforming (12) into its error system,

$$\tilde{\theta}(k) = [(1 - \lambda)I - \mu\phi(k-1)\phi^T(k-1)]\tilde{\theta}(k-1) + \mu\phi(k-1)\xi(k) + \lambda\theta^* \tag{13}$$

which should be compared to (10). For any bounded regressor $\phi(k-1)$ and disturbance $\xi(k)$, the stepsize μ can be chosen small enough so that the bracketed term in (13) is exponentially

contractive and the error system is bounded input bounded output stable. Thus leakage provides an exponential "safety net" from which the parameter estimates cannot escape. The price of this extra degree of stability is that the estimates will be biased away from their true values, that is, $\tilde{\theta}(k) = 0$ is no longer a solution to (13), even in the absence of disturbances. This bias will be proportional to λ and to the unknown θ^*.

An alternate way to look at (13) is to suppose that the true parameter θ^* is known to lie near some nominal value θ^0. Such *a priori* knowledge can be incorporated into the algorithm by considering the cost function

$$V(\theta(k-1)) = \frac{1}{2}(y(k) - \hat{y}(k))^2 + \frac{\lambda}{2\mu}(\theta(k-1) - \theta^0)^T(\theta(k-1) - \theta^0). \tag{14}$$

The gradient of $V(\theta(k-1))$ is

$$\frac{dV(\theta(k-1))}{d\theta(k-1)} = -(y(k) - \hat{y}(k))\phi(k-1) + \frac{\lambda}{\mu}\theta(k-1) - \frac{\lambda}{\mu}\theta^0 \tag{15}$$

and the gradient algorithm that tends to minimize this cost is

$$\theta(k) = (1-\lambda)\theta(k-1) + \mu\phi(k-1)(y(k) - \hat{y}(k)) + \lambda\theta^0. \tag{16}$$

Thus the leakage algorithm (12) introduced above in an ad hoc manner as a method to combat potential numerical problems is identical to the algorithm (16) in the special case when the nominal value θ^0 is assumed to be the origin.

2.3 Dead Zone

Small errors may reflect disturbances or noises, or may result from numerical problems. Large errors, on the other hand, are likely to be caused by poor parameter estimates. A member of the LMS family designed to combat numerical problems from small error signals forbids updates when the error signal is below some user defined threshold. The dead zone nonlinearity

$$g(x) = \left\{ \begin{array}{ll} x - d & x > d > 0 \\ 0 & -d < x < d \\ x + d & x < -d \end{array} \right\}, \tag{17}$$

when applied to the error signal, converts the LMS update (7) to

$$\theta(k) = \theta(k-1) + \mu\phi(k-1)g(y(k) - \hat{y}(k)). \tag{18}$$

As with leakage, this variant can be viewed as a modification to the cost function. Let

$$V(\theta(k-1)) = \left\{ \begin{array}{lll} \frac{1}{2}(y(k) - \hat{y}(k))^2 - d(y(k) - \hat{y}(k)) & if & y(k) - \hat{y}(k) > d \\ 0 & & -d \leq y(k) - \hat{y}(k) \leq d \\ \frac{1}{2}(y(k) - \hat{y}(k))^2 + d(y(k) - \hat{y}(k)) & if & y(k) - \hat{y}(k) < -d \end{array} \right\}. \tag{19}$$

Then $\frac{dV}{d\theta} = \phi g(y - \hat{y})$, and (18) is the gradient algorithm that minimizes this cost function V.

2.4 Signed Error LMS

Although LMS is computationally quite simple, there are always applications for which even m multiplies are too many. It is reasonable to suppose that as long as the correct gradient direction is maintained, the exact length of the stepsize is unimportant. This suggests the use of $sgn(y-\hat{y})$ in place of the $(y-\hat{y})$ term in (7), and gives the signed error (SE) algorithm

$$\theta(k) = \theta(k-1) + \mu\phi(k-1)sgn(y(k) - \hat{y}(k)) \tag{20}$$

where

$$sgn(x) = \left\{ \begin{array}{ll} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{array} \right\} \tag{21}$$

If μ is chosen to be a multiple of 2, then the term $\mu\phi(k-1)sgn(y(k) - \hat{y}(k))$ can be computed directly with bit shifts, and *no* multiplications are necessary, significantly reducing the computational burden of adaptation. What is sacrificed in terms of performance for this simplification?

In one sense, it is obvious that the SE algorithm (20) will tend to converge slower than the LMS algorithm (at least in the initial phase when the parameter estimates are poor) because the motion of the parameter updates will be smaller for a given stepsize μ. On the other hand, for small errors, the SE algorithm will tend to react faster. The simplest way to compare the two is to compare their cost functions. Consider

$$V(\theta(k-1)) = |y(k) - \hat{y}(k)|. \tag{22}$$

Then

$$\frac{dV(\theta(k-1))}{\theta(k-1)} = -\phi(k-1)sgn(y(k)-\hat{y}(k)) \tag{23}$$

modulo some ambiguity at $y(k) = \hat{y}(k)$. Accordingly, the SE can be viewed as an approximate gradient descent method that attempts to minimize the least absolute value of the error. In no sense, then, is the SE a "degraded version of LMS" as is sometimes stated. Rather, it is a valid minimization scheme operating optimally on its own cost function. For a problem with disturbances that consist of large outliers, the SE might well tend to return a parameter estimate with smaller variance than LMS, which exaggerates the importance of outliers.

2.5 Signed Regressor LMS

An alternative way to reduce the numerical complexity is to apply the signum function element by element to the regressor vector ϕ, leading to the signed regressor (SR) variant of LMS

$$\theta(k) = \theta(k-1) + \mu\ sgn(\phi(k-1))(y(k)-\hat{y}(k)). \tag{24}$$

As in the previous section, if μ is chosen to be a power of 2, the update term can be calculated by replacing multiplications with bit shifts. Moreover, the SR algorithm has the capability to react quickly to large errors, unlike the SE algorithm. The SR algorithm was first proposed by Moschner in [42] and has been implemented in several successful applications [18]. Claasen and Mecklenbrauker noted in [17] that the direction of the update can be significantly different from the true gradient direction (e.g., the vector (100, 0.01, -0.01) points in a radically different direction from its signed version (1,1,-1)), and hypothesized that this might cause the algorithm to climb, rather than descend the gradient. This was debated in [5] where it was shown that on the average, updates tend to proceed in a reasonable approximation to the downhill direction, at least when the inputs are Gaussian. Then, in [48], it was shown that certain classes of inputs can actually cause divergence of the parameter estimates to infinity (at least theoretically). Though such inputs are somewhat unlikely in a typical application, situations exist where they may cause catastrophic failure of the adaptive element. Consequently, more information about the environment in which

the algorithm will operate is necessary before applying the SR algorithm than in using some of the other variants. These divergence examples also explain why no gradient minimization interpretation is possible for the SR algorithm... there is no sensible cost function $V(\theta)$ which allows infinite θ as its minimum solution.

2.6 Sign-Sign LMS

Another way of reducing the numerical complexity of LMS is to incorporate the signum function on both the error and the regressor

$$\theta(k) = \theta(k-1) + \mu \; sgn(\phi(k-1))sgn(y(k) - \hat{y}(k)), \tag{25}$$

which is often called the sign-sign (SS) variant of LMS. Again, no multiplications are necessary, and even the n additions can be simplified by judicious choice of μ. This is the oldest of the variants of LMS, first used by Lucky in 1966 [39]. A recent Adaptive Differential Pulse Code Modulation (ADPCM) standard [14] utilizes signum functions on both the regressor and the error signals (the standard incorporates several other interesting features as well). Despite its status as the first born of the signed children of LMS, it has remained the least understood, probably due to the extra complexity of its twin nonlinearities. The enigma of the SS algorithm has recently begun to clear. An elegant example in [19] shows that the algorithm can diverge if its input is suitably pathological. This sparked a flurry of activity attempting to precisely define the class of signals which could cause the algorithm to misbehave [46], [19]. This has finally been resolved in [13] and [19], and will be discussed further in the examples.

2.7 Quantized State LMS

The signed variants of LMS succeeded in reducing the (already low) numerical complexity of LMS in exchange for an even slower convergence rate. Is there a way to maintain the numerical simplicity of the signed algorithms without this sacrifice?

One of the simpler proposals in this direction is the "dual sign" algorithm of [35] which utilizes two stepsizes: a large stepsize μ_L in the initial phase when the error is large, and a

small stepsize μ_S in the converged phase to encourage sufficient smoothing. If both μ_L and μ_S are powers of 2, then there is no significant increase in the numerical complexity over the signed algorithms. One reasonable generalization of this idea is [46]

$$\theta(k) = \theta(k-1) + \mu\ Q_1(\phi(k-1))Q_2(y(k) - \hat{y}(k)), \tag{26}$$

where Q_1 and Q_2 are quantization functions applied to the regressor and error signals respectively. This generalization transforms the choice of μ_L and μ_S (and the error values at which they switch) to a choice of the appropriate quantization functions. As will be shown, the introduction of such quantization functions does not add significant complexity to the analysis of the algorithm (over the SS algorithm) except that the issue of how to chose optimal $Q's$ must be addressed.

2.8 Least Mean Fourth

Closely related to LMS are algorithms intended to minimize higher powers of the error

$$V(\theta(k-1)) = (y(k) - \hat{y}(k))^q \tag{27}$$

for integer powers of q. The case $q = 4$ has been explicitely used in [20], and is often called the Least Mean Fourth (LMF) algorithm. For even q, it is easy to derive the algorithm

$$\theta(k) = \theta(k-1) + \mu\phi(k-1)(y(k) - \hat{y}(k))^{q-1} \tag{28}$$

which will tend to minimize the least mean q^{th} estimates of θ. For large q, one should expect numerical problems when dealing with large errors, though this effect can be ameliorated by chosing a judicious normalization of the stepsize.

2.9 Median LMS

The performance of LMS and its children often degrades badly when subjected to input signals that are corrupted by impulsive noise. One modification designed to combat this problem is the median LMS [55]

$$\theta(k) = \theta(k-1) + \mu med_3\{\phi(k-1)e(k), \phi(k-2)e(k-1), \phi(k-3)e(k-2)\} \tag{29}$$

where the "median" function med_3 is applied element by element to the update vectors ϕe. For example, if the numbers a, b and c are ordered from smallest to largest, then $med_3\{a, b, c\} = b$. Of course, median functions of all different lengths, or other order statistic functions may be used as well. The median function is interesting because it tends to reject single occurences of large spikes of noise, and these spikes are not passed into the parameter estimates.

3 Expected Behaviour Approach

This section briefly describes how we expect the children of LMS to behave, and highlights some of the benefits and pitfalls of the "expected value" approach.

In an off-line technique, when the parameter estimates can be calculated in closed form, no questions of stability or convergence occur. For instance, with the Least Squares approach, the parameter estimate is calculated

$$\theta = (\Phi^T \Phi)^{-1} \Phi^T Y \tag{30}$$

where $\Phi = [\phi(k-1), \phi(k-2), ..., \phi(k-m)]$, $Y = [y(k), y(k-1), ..., y(k-m+1)]^T$, and the solution θ is the vector that minimizes the summed least square error. In adaptive on-line schemes such as LMS, however, the parameter estimates are made via a recursion like (7), (20), (24), or (18), and it becomes crucial to determine the behaviour of the recursion as it evolves in time. Typically, the parameter estimates begin at some setting, move slowly to a region about some final value, and then "bounce around" this final value. But what can be said concretely about this behaviour?

One approach is to assume that the parameter estimates are already in the converged region (that they are stationary), and that the input $\phi(k)$ is stationary. Taking the expected value of both sides of (7) gives

$$E[\theta(k)] = E[\theta(k-1)] + \mu E[\phi(k-1)(y(k) - \hat{y}(k))]. \tag{31}$$

From the stationarity assumptions, $E[\theta(k)] = E[\theta(k-1)]$, and (31) can be rewritten

$$E[\phi(k-1)y(k)] = E[\phi(k-1)\phi^T(k-1)\theta(k-1)]. \tag{32}$$

If $\phi(k-1)$ is statistically independent of $\theta(k-1)$, then $E[\phi\phi^T\theta] = E[\phi\phi^T]E[\theta]$ and (32) can be solved for $E[\theta]$ assuming that $E[\phi\phi^T]$ is invertible

$$E[\theta(k-1)] = [E[\phi\phi^T]]^{-1}E[\phi(k-1)y(k)]. \tag{33}$$

This is formally analogous to (30) and provides evidence that the LMS algorithm (7) is likely to return the same answer, on average, as the Least Square method.

Unfortunately, the independence assumption used to derive (33) from (32) is virtually always false. To see this, recall that $\phi(k-1) = [u(k-1), u(k-2), ..., u(k-m)]^T$ is typically a regressor vector of past inputs $u(k-i)$. Hence $\phi(k-2) = [u(k-2), u(k-3), ..., u(k-m-1)]^T$. Since $\theta(k-1)$ is explicitely a function of $\phi(k-2)$ (from (7)), $\theta(k-1)$ and $\phi(k-1)$ cannot be independent, except perhaps in the scalar, one parameter case. Nevertheless, this is a very common assumption, since it often leads to useful guidelines for implementation issues. An important attempt to justify this assumption formally, based on a small stepsize assumption μ, can be found in [41].

An alternative approach [53], [28], is to retain the dynamics of the adapted system by defining a parameter error vector $\tilde{\theta}(k)$ as in (10) but without assuming that steady state has been achieved. Taking the expected value of both sides of (10), and assuming that

- the input is stationary
- the disturbance term $\xi(k)$ is independent of the input $\phi(k-1)$, and $E[\phi(k-1)\xi(k)] = 0$
- $\phi(k-1)$ is independent of $\tilde{\theta}(k-1)$,

the error system (10) can be rewritten

$$E[\tilde{\theta}(k)] = (I - \mu E[\phi(k-1)\phi^T(k-1)])E[\tilde{\theta}(k-1)]. \tag{34}$$

Since the auto correlation matrix $E[\phi\phi^T]$ is symmetric and nonnegative definite due to its structure as an outer product of ϕ with itself, it can be diagonalized $E[\phi\phi^T] = QDQ^T$ where $QQ^T = I$ and $D = diag(d_1, d_2, ..., d_n)$ is a diagonal matrix with all entries real and nonnegative. Thus

$$E[\tilde{\theta}(k)] = (QQ^T - \mu QDQ^T)E[\tilde{\theta}(k-1)] \tag{35}$$

$$= Q(I - \mu D)Q^T E[\tilde{\theta}(k-1)]. \tag{36}$$

Defining a transformed (expected) parameter error vector $\theta_Q(k) = Q^T E[\tilde{\theta}(k)]$, and multiplying both sides of (36) by Q^T yields

$$\theta_Q(k) = (I - \mu D)\theta_Q(k-1). \tag{37}$$

Since D is diagonal, this is simply m copies of the scalar equation

$$\phi_i(k) = (1 - \mu d_i)\phi_i(k-1) \tag{38}$$

which decreases exponentially to zero as long as $\mu d_i < 2$. Consequently, if μ is chosen small enough so that $\mu < \frac{2}{\lambda_{max}(R)}$ (where $\lambda_{max}(R)$ indicates the largest eigenvalue of R), then all modes of (37) are stable and $\theta_Q(k) \to 0$ as $k \to \infty$. This implies that $E[\tilde{\theta}(k)] \to 0$, which in turn implies that $E[\theta(k)] \to \theta^*$, at a rate proportional to the size of the eigenvalues of the correlation matrix.

Many useful results are possible from this style of analysis,

- Expressions for "misadjustment" (ratio of excess mean squared error to the minimum mean square error) can be derived as in [28].
- Time constants of convergence rates can be shown to be proportional to the magnitude of the eigenvalues of $E[\phi\phi^T]$, and to the eigenvalue "spread", (the ratio of the largest to the smallest eigenvalue). See [53].
- In some cases (especially when the inputs are Gaussian [6]), expressions for second and higher order moments are feasible.
- The analysis applies (via an extension of Prices theorem [43]) to the signed algorithms [40], and to other algorithms which incorporate quantization functions in their error update [23].
- The method can be extended to examine the nonstationary case (when θ^* itself is time varying) or when the statistics of ϕ are changing with time, see [24] and [35].
- Optimum stepsizes can often be calculated in terms of (possibly) available quantities as in [24] and [15].

Despite the fact that this style of analysis is nonrigorous in a formal mathematical sense (for instance, the independence assumption on $\theta(k-1)$ and $\phi(k-1)$), there is quite close agreement between the conclusions of the analysis, simulations of the algorithms, and the behaviour of the algorithms in applications. Indeed, it has taken more careful analysts over a decade to rigorously verify but a small part of the conclusions of this rougher analysis. While most of the results of the careful analysis were anticipated, there are situations in which the conclusions of this nonrigorous analysis can be misleading. Possibly the most dramatic of these differences involves the actual stability of certain of the variants of LMS, especially those which manipulate the regressor vector so that the update can point away from the "downhill" gradient direction.

Consider the signed regressor algorithm (24). Following the technique and assumptions of (31) to (33) yields

$$E[\theta(k-1)] = [E[sgn(\phi)\phi^T]]^{-1} E[sgn(\phi)y(k))]. \tag{39}$$

Consequently, one expects that, as long as the matrix $E[sgn(\phi)\phi^T]$ is invertible, $E[\theta(k-1)]$ takes on a single well defined value.

On the other hand, following the techniques and assumptions of (34) to (37) yields

$$E[\tilde{\theta}(k)] = (I - \mu E[sgn(\phi)\phi^T]) E[\tilde{\theta}(k-1)]. \tag{40}$$

The matrix $E[sgn(\phi)\phi^T]$ is not symmetric, and may have real or complex eigenvalues, with positive or negative real parts. Suppose that the stochastic process $\phi(k)$ is chosen so as to cause $E[sgn(\phi)\phi^T]$ to have an eigenvalue with a negative real part. Then there is a direction (the eigenvector associated with this negative eigenvalue) in which the scalar analog (38) is exponentially unstable, since $d_i < 0$. This implies that the parameter estimates will be driven away from the "correct" solution (39).

Clearly, the conclusion of (39) is diametrically opposed to the conclusion of (40). Hopefully, the careful reader will spot the reason for this discrepancy... the assumption of stationarity in the θ process in (33) and (39) is tantamount to an assumption of stability of (34) and (40). The intent here is to highlight the need for a more careful analysis, in which the ramifications of the various assumptions are pursued vigorously. The next two sections

present two approaches to a more rigorous analysis, the "deterministic approach" and the "stochastic approximation" approach.

4 The Deterministic Approach

The deterministic approach uses the tools of nonlinear system theory to examine LMS and its children. The generic adaptive update form (2) can be interpreted as the state equation of a nonlinear and time varying system. This system can be linearized and averaged to derive conditions under which the various algorithms 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 conditions guarantee convergence to a small region about the true value. The excitation conditions involve the nonlinear functions of the data and the error signal, F and g of (2), but the nonlinearities enter in different ways. Sign preserving *error* nonlinearities are essentially benign in terms of stability of the adaptive system, while even modest data nonlinearities can cause stability problems.

These results have a simple geometrical interpretation in terms of descending an error surface. Recall that LMS is an approximate gradient descent method utilizing the squared error as a cost function. 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 (such as in the signed regressor (24), the sign-sign (25), or the quantized state (26) algorithms) is to cause motion in a direction that is not necessarily "downhill." Is it 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 (such as in the dead zone (18), signed error (20), or the least mean pth (28) algorithms) is subtler. It changes the cost function that will be minimized. Each of the latter three algorithms has a simple interpretation as an approximate gradient method on some cost surface. 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.

4.1 Analytical Background

The key ideas of the deterministic approach are linearization, the slow time variation lemma [49], averaging [10], and total stability [27]. Linearization is used to examine the stability of the algorithm (2) 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 to relate the stability of the time varying system to the stability of the related frozen systems. The slowness is a consequence of the small value which the stepsize μ is assumed to have. 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.

4.1.1 Linearization

Consider the discrete time system

$$z(k) = \mathcal{F}(k-1, z(k-1)) \tag{41}$$

where $z(k)$ is a state vector in $\Re^m$, and $\mathcal{F}$ is a vector function $\Re^m \rightarrow \Re^m$ defining the evolution of the state. The states z^* for which $\mathcal{F}(k, z^*) = z^*$ for all k are the equilibria of (41), which we may assume without loss of generality to be located at the origin. $\mathcal{F}$ is linearized at the equilibrium $z^* = 0$ via the Jacobian $A(k) = D\mathcal{F}|_{z^*=0}$. The linearization theorem (Lyapunov's indirect method [51]) asserts that the behaviour of (41) near z^* is dictated by the behaviour of the related linear system

$$y(k) = A(k-1)y(k-1), \tag{42}$$

that is, if the linearized state equation (42) is exponentially asymptotically stable (e.a.s), then (41) is also e.a.s. The theorem holds assuming that $A(k)$ is bounded, and assuming

that the norm of the difference $\mathcal{F}(k,z) - A(k)z$ is uniformly bounded in time. Formally, this requires that

$$lim_{||z||\to 0} max_k \frac{||\mathcal{F}(k,z) - A(k)z||}{||z||} = 0, \tag{43}$$

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

4.1.2 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 (42) is e.a.s. One approach is to use the "slow time variation lemma" of [49] which asserts that if the change in A is slow enough (that is, $||A(k) - A(k-1)||$ is small), then exponential stability of each $A(j)$ (uniformly in j) is enough to imply e.a.s. of the time-varying system (42).

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 $A(j)$ fail to be e.a.s. An alternate approach [10] is to take the time average of (42) and to define the *sliding average*

$$\bar{A}(k,m) = \frac{1}{m}\sum_{i=1}^{m} A(k+i). \tag{44}$$

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 it can be shown that the averaged system

$$\bar{y}(k) = \bar{A}(k-1)\bar{y}(k-1) \tag{45}$$

and the related (42) are both e.a.s. Fortunately, the sliding averages can be exponentially stable even when the $A(k)$'s are not.

4.1.3 Total Stability

The final step in the argument is to relax the assumption that there are no disturbances. The total stability theorem of [27] relates the behaviour of the unforced system (41) to the behaviour of

$$\bar{z}(k) = \mathcal{F}(k-1, \bar{z}(k-1)) + \mathcal{G}(k-1, \bar{z}(k-1)) \tag{46}$$

where $\mathcal{G}$ is some small disturbance term that may depend on the state. Assuming that $\mathcal{F}$ is Lipshitz continuous, the difference between the state z of (41) and the state $\bar{z}$ of (46) can be bounded when $\mathcal{F}$ is known to be e.a.s. by requiring that $\mathcal{G}$ 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 $||\mathcal{G}(k, \bar{z}(k))|| < \delta_2$ 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.

4.2 Persistence of Excitation

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

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) - \mu F(\phi(k-1))g(\tilde{\theta}^T(k-1)\phi(k-1)) \tag{47}$$

which is derived from (2) by the introduction of the parameter estimate error $\tilde{\theta}(k) = \theta^* - \theta(k)$. 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) $g(.)$ is differentiable at the origin.

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 stepsize 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 [49] via similar techniques to those used here, and others have attacked this situation in other ways, see [36] and [37]. 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. Most of the nonlinear variants of LMS fulfill these requirements, though the "signed error" algorithm (20) fails condition (c).

4.2.1 Linearization

Define the vectors $\tilde{\theta}(k) = [\theta_1(k), \theta_2(k), ..., \theta_m(k)]^T$, $\phi(k) = [x_1(k), x_2(k), ..., x_m(k)]^T$, and the vector function $F(\phi(k)) = [f_1(\phi(k)), f_2(\phi(k)), ..., f_m(\phi(k))]^T$. Typically, $\phi(k)$ consists of a "regressor" vector of time shifted versions of a scalar sequence $x(k)$, that is, $x_i(k) = x_{i-1}(k-1)$ for i=2,m, but this is not necessary. Identify the function F of (41) with the right hand side of (47), and let

$$H(k) = F(\phi(k))g(\tilde{\theta}^T(k)\phi(k)) = \begin{pmatrix} f_1(\phi(k))g(\theta_1(k)x_1(k) + \theta_2(k)x_2(k) + \theta_m(k)x_m(k)) \\ f_2(\phi(k))g(\theta_1(k)x_1(k) + \theta_2(k)x_2(k) + \theta_m(k)x_m(k)) \\ \vdots \\ f_n(\phi(k))g(\theta_1(k)x_1(k) + \theta_2(k)x_2(k) + \theta_m(k)x_m(k)) \end{pmatrix}. \tag{48}$$

Then the Jacobian $\frac{dH(k)}{d\tilde{\theta}(k)}$ can be calculated as

$$\begin{pmatrix} f_1(\phi(k))\theta_1(k)g'(\tilde{\theta}^T(k)\phi(k)) & f_1(\phi(k))\theta_2(k)g'(\tilde{\theta}^T(k)\phi(k)) & \cdots & f_1(\phi(k))\theta_m(k)g'(\tilde{\theta}^T(k)\phi(k)) \\ f_2(\phi(k))\theta_1(k)g'(\tilde{\theta}^T(k)\phi(k)) & f_2(\phi(k))\theta_2(k)g'(\tilde{\theta}^T(k)\phi(k)) & \cdots & f_2(\phi(k))\theta_m(k)g'(\tilde{\theta}^T(k)\phi(k)) \\ \vdots & \vdots & \vdots & \vdots \\ f_n(\phi(k))\theta_1(k)g'(\tilde{\theta}^T(k)\phi(k)) & f_n(\phi(k))\theta_2(k)g'(\tilde{\theta}^T(k)\phi(k)) & \cdots & f_n(\phi(k))\theta_m(k)g'(\tilde{\theta}^T(k)\phi(k)) \end{pmatrix}. \tag{49}$$

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

$$B(k) = \frac{dH(k)}{d\tilde{\theta}(k)}|_{\theta^*=0} = g'(0)F(\phi(k))\phi^T(k), \tag{50}$$

and the linearized system is

$$y(k) = (I - \mu B(k-1))y(k-1). \tag{51}$$

The linearization result shows that if (51) is exponentially stable, then the original nonlinear system (47) is also exponentially stable.

4.2.2 Slow Time Variation and Averaging

Note that by choosing the stepsize parameter μ small, the time variation of the transition matrix $(I - \mu B(k-1))$ is slowed. In fact, as $\mu \to 0$, $||(I - \mu B(k)) - (I - \mu B(k-1))|| \to 0$. 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 $(I - \mu B(j))$, for each j.

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 $m-1$ zero eigenvalues. This implies that $(I - \mu B(k))$ has $m-1$ unity eigenvalues, and hence is not exponentially stable. To overcome this, define the sliding average $\bar{B}(k,l)$ over the time window l as in (44). Then the averaging theorem demonstrates that exponential stability of

$$\bar{y}(k) = (I - \mu \bar{B}(k-1))\bar{y}(k-1) \tag{52}$$

implies exponential stability of (51), and hence (47). Define the *excitation matrix*

$$M_s = \sum_{k=1}^{s} F(\phi(k))\phi^T(k), \tag{53}$$

which, for s-periodic inputs is equal to the sliding average, as is done in [46]. Then the magnitude of all eigenvalues of $(I - \mu M_s)$ can be guaranteed less than one as long as M_s has all eigenvalues with positive real part, and as long as μ is chosen small enough. Gathering the above results together shows

Theorem 4.1 (Persistence of Excitation Theorem) *Consider the algorithm (47) with s-periodic input data $\phi(k)$ and nonlinear elements F and g, under assumptions (b) and (c). If there are $\alpha > 0$ and $\beta > 0$ such that*

$$\beta > g'(0) Re\lambda_i(M_s) > \alpha \qquad for\ every\ i, \tag{54}$$

then there is a μ^* such that for every μ in $(0, \mu^*)$, the algorithm (47) is locally exponentially stable about its equilibrium $\theta^* = 0$. Conversely, if $g'(0) Re\lambda_i(M_s)$ is negative for some i, then the algorithm (47) is locally unstable about its equilibrium at $\theta^* = 0$.

[The notation $Re\lambda_i(M)$ means the real part of the ith eigenvalue of the matrix M.] An input which fulfills (54) for a particular algorithm is said to be *persistently exciting* for the algorithm. Equivalently, the algorithm is *persistently excited* by the input.

Some remarks:

(a) Local exponential stability of the algorithm implies that the parameter estimate error

$\tilde{\theta}(k)$ 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 $\theta(k)$ to their true values θ^*. Local instability implies that there are arbitrarily small perturbations that can drive the parameter estimates away from θ^*. This does not necessarily imply divergence to infinity of the parameter estimates.

(b) The condition (54) is called the *persistence of excitation* (PE) condition for the LMS algorithm with nonlinearities F and g. Note that the condition involves the input data sequence $\phi(k)$ 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 (54) fails. Similarly, $g'(0)$ must be positive.

(d) If $g'(0) = \infty$ then assumption (c) and the left hand inequality of (54) fails. In particular, this averaging approach is inapplicable to the signed error algorithm with $g(e) = sgn(e)$. An extended Lyapunov approach for this algorithm can be found in [46].

(e) The convergence rate of the averaged system (52) (and hence the convergence rate of the algorithm (47)) is proportional to the size of the real part of the smallest eigenvalue of (53). Thus, given an input sequence $\phi(k)$, 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 (54) is not violated.

(f) The periodicity assumption is not necessary, and can be relaxed to "almost periodic" inputs as in [3] at the expense of a large amount of technical detail.

(g) The fact that (54) 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 this ball about the origin.

Suppose that $g'(0) = 0$, as occurs in the dead zone algorithm (18), in the least mean q^{th} algorithms (28) and in the quantized state algorithms (26) for certain quantization functions Q_2. Then the right hand side of the persistence of excitation condition (54) fails, and the algorithm is not exponentially stable about $\theta^* = 0$. If, however, g is nondecreasing, con-

tinuous, 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 θ^* itself. To make this notion more precise, consider the following definition.

Definition 4.1 *The system $x(k) = f(k-1, x(k-1))$ is said to be (uniformly) locally exponentially stable to the compact region R contained in B if there exists a $\gamma \in (0,1)$ 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 4.1 *: Consider the algorithm (47) with s-periodic input data $\phi(k)$ 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 there are $\alpha > 0$ and $\beta > 0$ such that $\beta > Re\lambda_i(M_s) > \alpha \ \forall i$. Then there is a μ^* such that for every $\mu \in (0, \mu^*)$, the algorithm is locally exponentially stable to the region R.*

4.2.3 Total Stability

The final step is to remove the "ideal" assumption, and to suppose that some small nonidealities are present. The $\mathcal{F}$ and $\mathcal{G}$ of (46) may be related to the various versions of LMS by identifying the state $\bar{z}(k)$ with the parameter estimate errors $\tilde{\theta}(k)$, and $\mathcal{G}$ with the disturbance term. Assuming that the input data fulfills the PE condition (54), then the homogeneous system (47) (and (41) with $\mathcal{F}$ identified as the right hand side of (47)) is exponentially stable. Consequently, the total stability theorem asserts that for small disturbances $\mathcal{G}$, 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 $\xi(k)$ corrupts the prediction error $e(k) = y(k) - \hat{y}(k)$. Then $\mathcal{G}(k-1, \tilde{\theta}(k-1)) = \mu F(\phi(k-1))[g(e(k)+\xi(k)) - g(e(k))]$, and the norm of $\mathcal{G}$ can be bounded in terms of μ, $||F||$, $||\phi(k-1)||$, and the smoothness of g. Hence, if $||\xi(k)||$ is small enough so that $||\mathcal{G}|| < \delta_2$, the total stability theorem shows that an algorithm that is exponentially stable cannot be destabilized by arbitrarily small measurement biases or inaccuracies.

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

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

(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 $\theta^*(k)$ represent the time varying plant, and suppose that $||\theta^*(k) - \theta^*(k-1)||$ is small. The error system becomes

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) - \mu F(\phi(k-1)) g(\tilde{\theta}^T(k-1)\phi(k-1)) + \theta^*(k) - \theta^*(k-1) \qquad (55)$$

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

4.2.4 Interpretation of the Excitation Conditions

This section compares the persistence of excitation (PE) condition for LMS with nonlinearities F and g to the PE condition for LMS by showing that it is strictly more difficult to fulfill the PE condition for the nonlinear variants of LMS than the PE condition for (linear) LMS. The standard PE condition for LMS [9] (without nonlinearities), when excited by s-periodic

inputs $\phi(k)$, is that there exist $\alpha > 0$ and $\beta > 0$ such that

$$\beta I > \sum_{k=1}^{s} \phi(k)\phi^T(k) > \alpha I. \tag{56}$$

As above, this implies local exponential stability of the error system. Since the matrix in (56) is symmetric, all eigenvalues are real, and the notation ">" means positive definite. How does (56) compare to (54)?

Lemma 4.1 *: Suppose that (54) holds for a given F, g, and input sequence $\phi(k)$, and suppose that $F(\phi)$ does not vanish as $\phi \to \infty$. Then (56) also holds.*

Proof: By contradiction. There are two possibilities:
(i) If (56) fails the upper bound, then $\phi(k)$ must be diverging. By assumption, this implies that $F(\phi(k))\phi^T(k)$ must diverge. Hence the left hand inequality of (54) is violated.
(ii) If (56) fails the lower bound for every positive α, then there must be a zero eigenvalue of $\sum_{k=1}^{s} \phi(k)\phi^T(k)$. Consequently, there must be a nonzero eigenvector v such that $v^T\phi(k) = 0$ for every k. This implies that $(\sum_{k=1}^{s} F(\phi(k))\phi^T(k))v = 0$, and so there is a zero eigenvalue of the matrix in (54).

This says that if the nonlinear LMS algorithm is persistently excited (54), then the standard LMS algorithm is also persistently excited (56). For a large class of F, the reverse implication is false, since one can easily construct examples for which (54) has eigenvalues with negative real parts. Such examples can be found in [48] and a generic counterexample is constructed in [44] to demonstrate that whenever F is nonlinear, there are input sequences that will fail the PE condition and destabilize the algorithm.

5 The Stochastic Approximation Approach

The stochastic approximation approach to the analysis of adaptive algorithms was pioneered by L. Ljung in [38], and has been extended over the years by several researchers, most notably Kushner [34] and Benveniste [4]. The approach relates the motion of the parameter estimate errors of the algorithms to the behaviour of an unforced deterministic ordinary differential equation (ODE) by showing that local stability of the ODE implies weak convergence of

the algorithm. More recent is the observation that the ODE can be unstable, which implies nonconvergence of the parameter estimates [13]. These results are stated in terms of the eigenvalues of a correlation-like matrix which may be thought of as the stochastic analog of the persistence of excitation condition of the previous section. When the recursion is stable, the asymptotic distribution of the parameter trajectories is a Guass-Markov process under very general assumptions on the statistics of the inputs and disturbances. It is not necessary to assume independence of $\tilde{\theta}(k-1)$ and $\phi(k-1)$.

The ability of the algorithms to track moving parameterizations (when θ^* is a function of time) can be analyzed in a similar manner, by relating the time varying system to a *forced* ODE. The asymptotic distribution about the forced ODE is again a Guass-Markov process, whose properties can be described in a straightforward manner. This allows a comparison of the various adaptive algorithms in terms of their convergence and tracking ability.

The analysis is carried out by examining the general recursive form

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) - \mu H(\tilde{\theta}(k-1), \phi(k-1), \xi(k)) \tag{57}$$

which is a slight rewriting of (2) and (47) that captures the children of LMS by suitable choice of $H(\cdot)$. As before, $\tilde{\theta}(k)$ represents the parameter estimate errors, $\phi(k)$ is (usually) a vector of inputs, $\xi(k)$ is a disturbance process that represents all nonidealities such as measurement and modeling errors, and μ is a small positive constant stepsize. As in the previous section, convergence of the process $\tilde{\theta}(k)$ to a stationary distribution about zero is equivalent to convergence of the adaptive filter parameter estimates to a region about their optimal values. Two important questions concerning the behaviour of $\tilde{\theta}(k)$ arise:

- Under what conditions is the process stable?
- When do stationary distributions for $\tilde{\theta}(k)$ exist, and how can these stationary distributions be characterized?

One way to answer these questions is to relate the behaviour of the adaptive algorithm (57) for small μ to the behaviour of the associated deterministic ordinary differential equation (ODE)

$$\tilde{\theta}(t) = \tilde{\theta}(0) - \int_0^t \hat{H}(\tilde{\theta}(s))ds \tag{58}$$

where $\hat{H}(\cdot)$ is a smoothed version of $H(\cdot,\cdot,\cdot)$. The fundamental issue is to determine the relationship between $\tilde{\theta}(k)$ and $\tilde{\theta}(t)$. Recall that processes using the time index k are discrete, while processes with time index t are continuous. The two questions about (57) translate into analogous questions concerning (58).

- Under what conditions is the ODE stable or unstable?
- How closely does the algorithm (57) track the behaviour of the ODE (58)?

If the ODE is stable, then the algorithm (57) is stable (indicating probable success of the adaptive scheme), while if (58) is unstable, then (57) is also unstable, and the adaptive algorithm fails. For instance, if the correlation matrix of the input process $E[\phi\phi^T]$ is positive definite, then (for small enough μ) the parameter estimate errors of the LMS algorithm converge in distribution to a region about the origin ([53], [8]). The same matrix $E[\phi\phi^T]$ appears here as the linearization of $\hat{H}(\tilde{\theta})$, and is called the *stochastic excitation matrix.* Positive definiteness of this matrix implies local stability of the ODE, while a negative eigenvalue would imply local instability. Of course, due to its structure as a correlation matrix, $E[\phi\phi^T]$ is always at least nonnegative definite, and the instability cannot occur.

Certain of the children of LMS are not so fortunate. The analogous stochastic excitation condition for the signed regressor algorithm, for instance, requires that all eigenvalues of $E[sgn(\phi)\phi^T]$ have positive real parts [48]. As before, this same matrix appears in the present analysis as the linearization of $\hat{H}(\tilde{\theta})$, and positivity of the real parts of the eigenvalues of of $E[sgn(\phi)\phi^T]$ implies stability, while an eigenvalue with negative real part implies instability. In this case, there are nontrivial input distributions which cause instability of the associated ODE, and hence of the signed regressor algorithm.

The relation between the adaptive algorithm (57) and the ODE (58) may be thought of as a type of "law of large numbers." To investigate how close the behaviour of the algorithm is to the deterministic trajectory of the ODE, one desires a corresponding "central limit theorem." Consider the time scaled process $\tilde{\theta}_{[\frac{t}{\mu}]}(t)$ where $[z]$ represents the integer part of z. The martingale central limit theorem can be exploited to show that the error process

$$V_\mu = \frac{1}{\sqrt{\mu}}(\tilde{\theta}_{[\frac{t}{\mu}]}(t) - \tilde{\theta}(t)) \tag{59}$$

converges to a forced ODE that is driven by a sum of independent mean zero Brownian motions. Significantly, under mild assumptions on the input and disturbance processes, the limiting distribution is a Guass-Markov process, with known mean and variance.

In practical terms, this convergence has two implications. First, for a given algorithm, it is easy to calculate the parameters of the convergent distribution in terms of the properties of the inputs and disturbances, and hence to give a measure of the performance of the algorithm. Second, this allows a fair comparison between competing adaptive schemes by calculating the mean and variance of the convergent distributions for the various algorithms. Equivalently, this allows a fair comparison of the convergence speed of the various algorithms.

Finally, the ability of the adaptive algorithms to track a slowly moving parameterization θ^* can be examined by following essentially the same program as above. The asymptotic distributions of the appropriate error process can be related to a forced ODE, where the forcing term is directly related to the motion of the underlying parameterization, and the asymptotics once again prove to be a Guass-Markov process. In contrast to the convergence speed, there is little difference between the various algorithms in terms of their ability to track slowly moving targets.

5.1 Theoretical Development

This subsection presents the limit theorems which relate the behaviour of the adaptive algorithm (57) to the ODE (58), basically following the presentation in [13]. The adaptive update term $H(\cdot)$ in (57) has three arguments

- $\tilde{\theta}(k-1)$ is the parameter estimate error
- $\phi(k-1)$ is the input to the adaptive filter
- $\xi(k)$ is the disturbance term

The process $\{\tilde{\theta}(k-1), \phi(k-1), \xi(k)\}$ takes on values in $\Re^m \times E_1 \times E_2$, where m is the number of adaptive parameters, and E_1 and E_2 are the appropriate state spaces on which $\phi(k)$ and $\xi(k)$ evolve. $\{\tilde{\theta}(k-1), \phi(k-1), \xi(k)\}$ is adapted to a filtration $\{\mathcal{F}_k\}$, (usually one takes $\mathcal{F}_k$ = the σ-algebra generated by the random variables $(\tilde{\theta}(l-1), \phi(l-1), \xi(l))_{l=-\infty}^{k}$). We assume

that there exists a transition function $\eta(\tilde{\theta}, \phi, C)$ such that $P(\xi(k) \in C|\mathcal{F}_k) = \eta(\tilde{\theta}(k), \phi(k), C)$ and that H is integrable with respect to $\eta(\tilde{\theta}, \phi, \cdot)$ for each $(\tilde{\theta}, \phi) \in \Re^m \times E_1$. Define

$$\bar{H}(\tilde{\theta}, \phi) = \int_{E_2} H(\tilde{\theta}, \phi, \xi)\eta(\tilde{\theta}, \phi, d\xi). \tag{60}$$

The probability distribution η of the disturbance term is used in (60) to smooth out, or average, H through the action of the integral. Of most significance for the present purpose is that $\bar{H}$ can be continuous even when H is not.

We assume the following:

C.1 $\{\phi(k)\}$ is stationary and ergodic, there is a sequence of i.i.d. E_3-valued random variables $\{\psi(k)\}$, independent of $\{\phi(k)\}$, and a measurable function $q : \Re^d \times E_1 \times E_3 \to E_2$ such that $\xi(k) = q(\tilde{\theta}(k-1), \phi(k-1), \psi(k))$, and $\tilde{\theta}(0)$ is independent of $\{(\phi(k-1), \psi(k))\}$. $\nu_\phi \in \mathcal{P}(E_1)$ will denote the distribution of $\phi(k-1)$.

C.2 $\bar{H}$ is continuous, and for compact $K \subset \Re^d$

$$E[\sup_{\tilde{\theta} \in K} |H(\tilde{\theta}, \phi(k-1), q(\tilde{\theta}, \phi(k-1), \psi(k)))|] < \infty, E[\sup_{\tilde{\theta} \in K} |\bar{H}(\tilde{\theta}, \phi(k-1)|] < \infty. \tag{61}$$

Note that there are no assumptions on the autocorrelations of the inputs or disturbances. H is allowed to be discontinuous, provided that the expectation over η is smooth enough to make $\bar{H}$ continuous. Just as $\bar{H}$ averages H, the distribution of $\phi(k)$ is used to average $\bar{H}$ over the inputs $\phi(k)$, and the doubly averaged quantity

$$\hat{H}(\tilde{\theta}) = \int \bar{H}(\tilde{\theta}, \phi)\nu_\phi(d\phi) \tag{62}$$

is the key ingredient in the ODE and to the questions of stability.

Theorem 5.1 (Stochastic Excitation Theorem) *Let $\tilde{\theta}_\mu(t) = \tilde{\theta}([t/\mu])$, and for compact $K \subset \Re^m$, define $\tau_\mu^K = inf\{t : \tilde{\theta}_\mu(t) \notin K\}$. Denote the minimum of a and b as $a \wedge b$ and let $\tilde{\theta}_\mu^{\tau_\mu^K}(\cdot) = \tilde{\theta}_\mu(\cdot \wedge \tau_{\mu_K})$ define the "stopped" process. Assume C1 and C2, and that $\tilde{\theta}_\mu(0) \to \tilde{\theta}_0$ as $\mu \to 0$. Then for each K, $\{\tilde{\theta}_\mu^{\tau_\mu^K}, \mu > 0\}$ is relatively compact, and every limit point (as $\mu \to 0$) satisfies*

$$\tilde{\theta}(t) = \tilde{\theta}_0 - \int_0^t \hat{H}(\tilde{\theta}(s))ds \tag{63}$$

for $t < \tau^K = inf\{t : \tilde{\theta}(t) \notin K\}$.

Corollary 5.1 *Define* $\check{H}(\tilde{\theta}, \phi, z) = H(\tilde{\theta}, \phi, q(\tilde{\theta}, \phi, z))$. *Let* $C = \{(\tilde{\theta}, \phi, z) : \check{H}$ *is continuous at* $(\tilde{\theta}, \phi, z)\}$, *and let* I_C *be the indicator function of the set* C. *If* $\int\int I_C(\tilde{\theta}, \phi, z)\nu_\phi(d\phi)\nu_\psi(dz) = 1$, *for every* $\tilde{\theta}$, *then the assumption of the continuity of* $\bar{H}$ *can be dropped, and if in addition, the solution of (63) is unique, the convergence of* $\tilde{\theta}_\mu$ *to* $\tilde{\theta}$ *is almost sure.*

The theorem and corollary are proven in [13].

Consider the various elements of this theorem. A new process $\tilde{\theta}_\mu(t)$ is defined as a time scaled version of the original $\tilde{\theta}(k)$ for each stepsize μ. The time scaling compresses the original process variation into a smaller time frame. In effect, the gross motion of the parameter estimate errors of $\tilde{\theta}_\mu$ remains unchanged for various μ. Large μ imply larger steps of $\tilde{\theta}_\mu$, but fewer steps are taken. Smaller μ implies smaller updates of $\tilde{\theta}_\mu$, but more steps are taken. Thus the result is applicable to reasonable stepsizes due to the time rescaling, even though it is exact only asymptotically in μ.

The stopping time τ_μ^K measures how long it takes the time scaled process $\tilde{\theta}_\mu(t)$ to reach the edge of some closed and bounded set K. The stopped process $\{\tilde{\theta}_\mu^{\tau_\mu^K}(t)\}$ is defined to be equal to $\tilde{\theta}_\mu(t)$ from time zero to the stopping time τ_μ^K and is then held constant for all $t > \tau_\mu^K$. The theorem asserts that for any given compact set K, every possible sequence (as $\mu \to 0$) of the stopped process $\{\tilde{\theta}_\mu^{\tau_\mu^K}(t)\}$ contains a convergent subsequence, and that every limit of these subsequences is a process that satisfies the ODE (63), at least up until the stopping time. If the solution to the differential equation is unique, then the sequence actually converges (not just has a convergent subsequence).

The stability of the ODE can be determined by linearizing (63) about $\tilde{\theta} = 0$, that is, by calculating $M = \frac{d\hat{H}(\tilde{\theta})}{d\tilde{\theta}}|_{\tilde{\theta}=0}$. If all eigenvalues of the resulting matrix M have positive real parts, then the ODE is exponentially stable (note the minus sign before the integral in (63)) while if some eigenvalue has a negative real part, then the ODE is unstable. These stability and instability results translate directly into convergence and divergence results for the algorithms.

The ramifications of the stochastic excitation theorem for particular adaptive algorithms will be examined in the next section. Note that the theorem is a form of "law of large numbers" where the time scaled process $\tilde{\theta}_\mu(t)$ plays the role of "observations" and the convergent process $\tilde{\theta}(t)$ plays the role of the "expected value" to which the $\tilde{\theta}_\mu(t)$ converge as the number

of observations $\frac{t}{\mu}$ increases. To investigate how this convergence occurs, the corresponding "central limit theorem" describes the weak convergence of the error process

$$V_\mu(t) = \frac{1}{\sqrt{\mu}}(\tilde{\theta}_\mu(t) - \tilde{\theta}(t)), \tag{64}$$

where the scaling factor $\frac{1}{\sqrt{\mu}}$ expands V_μ to compensate for the time compression of $\tilde{\theta}_\mu(t)$. The next theorem shows that the error process V_μ converges to a forced ODE that is driven by the sum of two independent, mean zero Brownian motions. One driving term accounts for the error introduced by the smoothing with the disturbance $(H - \bar{H})$ while the other $(\hat{H} - \bar{H})$ accounts for the error when averaging over the inputs.

To understand this, imagine that the solution trajectory of the ODE is a smooth curve in $\Re^m$. The stochastic excitation theorem asserts that trajectories of the algorithm tend to follow this curve, though any particular trajectory will make occasional excursions, which wiggle about the curve. The central limit theorem below describes how this wiggling occurs by showing that it can be described in terms of a stationary Guass-Markov process. This is what is meant by the statement that the algorithm converges to a stationary distribution about the solution of the ODE.

Assume that H is square integrable with respect to $\eta(\tilde{\theta}, \phi, \cdot)$ for each pair $(\tilde{\theta}, \phi) \in \Re^d \times E_1$. Let $G(\tilde{\theta}, \phi, \xi) = (H(\tilde{\theta}, \phi, \xi) - \bar{H}(\tilde{\theta}, \phi))(H(\tilde{\theta}, \phi, \xi) - \bar{H}(\tilde{\theta}, \phi))^T$ be the matrix that represents the deviation of H from its smoothed version $\bar{H}$, and define a smoothed version of G as

$$\bar{G}(\tilde{\theta}, \phi) = \int_{E_2} G(\tilde{\theta}, \phi, \xi)\eta(\tilde{\theta}, \phi, d\xi). \tag{65}$$

Averaging over all inputs yields

$$\hat{G}(\tilde{\theta}) = \int \bar{G}(\tilde{\theta}, \phi)\nu_\phi(d\phi). \tag{66}$$

The various G's play a similar role in the central limit theorem that the H's play in the previous theorem. In addition to C.1 and C.2, we make the further assumptions:

C.3 $\bar{H}$ is differentiable as a function of $\tilde{\theta}$, $\bar{G}$ and $\partial_{\tilde{\theta}}\bar{H}$ are continuous, and for compact $K \subset \Re^d$

$$E[\sup_{\tilde{\theta} \in K} |H(\tilde{\theta}, \phi(k), q(\tilde{\theta}, \phi(k), \psi(k)))|^2] < \infty$$
$$E[\sup_{\tilde{\theta} \in K} |\bar{G}(\tilde{\theta}, \phi(k))|] < \infty$$
$$E[\sup_{\tilde{\theta} \in K} |\partial_{\tilde{\theta}}\bar{H}(\tilde{\theta}, \phi(k))|] < \infty$$

Note that C.3 implies $\hat{H}$ is locally Lipschitz (in fact continuously differentiable), so the solution of (63) is unique and hence $V_\mu(t)$ is well defined. For simplicity (so we don't have to stop our process outside of a compact set), we assume that the solution exists for all $t \geq 0$. Define

$$\tilde{M}_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (H(\tilde{\theta}(k-1), \phi(k-1), \xi(k)) - \bar{H}(\tilde{\theta}(k-1), \phi(k-1)))\sqrt{\mu} \tag{67}$$

and

$$L_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (\bar{H}(\tilde{\theta}(k\mu), \phi(k)) - \hat{H}(\tilde{\theta}(k\mu)))\sqrt{\mu}. \tag{68}$$

There are a variety of different conditions (for example, mixing conditions on $\{\phi(k)\}$) that imply $\{L_\mu\}$ converges in distribution to a (time inhomogeneous) Brownian motion. We simply assume this convergence.

C.4 $L_\mu \Rightarrow L$.

Theorem 5.2 (Central Limit Theorem) *Assume C.1-C.4, that $\tilde{\theta}_\mu(0) \to \tilde{\theta}_0$, that the solution of (63) exists for all $t \geq 0$, and that $V_\mu(0) \to v_0$. Then $\tilde{M}_\mu \Rightarrow \tilde{M}$ where $\tilde{M}$ is a mean zero Brownian motion independent of L with*

$$E[\tilde{M}(t)\tilde{M}(t)^T] = \int_0^t \hat{G}(\tilde{\theta}(s))ds$$

and $V_\mu \Rightarrow V$ satisfying

$$V(t) = v_0 + \tilde{M}(t) + L(t) - \int_0^t \partial_{\tilde{\theta}} \hat{H}(\tilde{\theta}(s))V(s)ds \tag{69}$$

This theorem is taken from [13] where a proof can be found.

These results can be extended in a variety of directions with little or no change in the hypotheses. For example, consider the asymptotics of the "tracking problem" for adaptive filters. Let $\tilde{\theta}^*(k)$ denote the time varying "correct" filter coefficients that the adaptive filter is attempting to track, and let $\theta(k)$ be the parameter estimates. The parameter estimate error is then $\tilde{\theta}(k) = \tilde{\theta}^*(k) - \theta(k)$, which evolves according to

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) - \mu H(\tilde{\theta}(k-1), \phi(k-1), \xi(k)) + (\tilde{\theta}^*(k) - \tilde{\theta}^*(k-1)). \tag{70}$$

Clearly, some restrictions must be placed on the possible motion of the filter $\tilde{\theta}^*$. One possibility is to assume

C.5 $\tilde{\theta}^*(k) = \Psi(k\mu)$ where Ψ is a differentiable function with derivative denoted by ψ. It is then easy to show that Eq.(63) can be replaced by

$$d\tilde{\theta}(t) = -\psi dt - \hat{H}(\tilde{\theta}(t))dt$$

or

$$\tilde{\theta}(t) = \tilde{\theta}_0 - \int_0^t \psi(s)ds - \int_0^t \hat{H}(\tilde{\theta}(t))dt \tag{71}$$

The implications of (71), in terms of the tracking capabilities of the various adaptive algorithms, are explored in the next section. Note that in books such as [26], the martingale convergence theorem is viewed as an alternative to the ODE approach. Here, (and in [13] and [4]), the two approaches are married.

6 Examples, Comparisons, and Discussion

Several of the children of LMS are examined concretely in light of the deterministic and stochastic approaches of the previous sections. The intent is to compare the algorithms with each other and to compare the types of conclusions possible from the analytical methods. Progress in the analysis of adaptive algorithms has often alternated between the deterministic and stochastic realms. The deterministic approach typically assumes that the disturbances are identically zero, proves an exponential stability result, and then uses some form of total stability to guarantee robustness to disturbances. Speaking loosely, the deterministic "persistence of excitation condition" tends to function analogously to the stochastic excitation conditions derived via linearization of $\hat{H}$. For example, the LMS algorithm is exponentially stable when $\sum \phi\phi^T$ is positive definite, which clearly parallels the stochastic excitation condition which requires that $E[\phi\phi^T]$ be positive definite. One need only replace the Greek letter $\sum$ with the Roman letter E!

What is the relation between these conditions? The deterministic condition (even when relaxed for general, non-periodic inputs) is strictly stronger than the stochastic condition,

since it requires uniform positive definiteness of $\sum_s \phi\phi^T$ over *every* window of some length s. Most stochastic processes will fail this for some windows, albeit with small probability for large s. Thus the hypotheses required to demonstrate exponential stability via the deterministic approach are stronger. So, too, are the conclusions. Even when stable, there exist sample paths that cause the parameter errors to attain arbitrarily large values under the stochastic assumptions, though these events are of vanishing probability [45]. In contrast, the total stability conclusions of the deterministic approach are absolute; the parameter errors *never* leave the appropriate δ ball. The disadvantage is that it is hard to determine, a priori, a reasonable bound for δ, and it is impossible to say anything at all about the behaviour of the parameter estimates inside the δ ball. In contrast, the stochastic method gives a stationary distribution that describes how the parameter errors "rattle around" their converged values. This distribution can often be calculated, although it is only exact asymptotically (as $\mu \to 0$). To apply the deterministic method:

- Determine the appropriate F and g (47).
- Derive the relevant PE condition (53).
- Determine stability/instability and convergence rates based on the PE condition and the type of inputs expected in the given application.

To apply the stochastic approximation approach

- Define appropriate $\phi(k)$ (input) and H (update term).
- Find the unforced ODE (57) by calculating the smoothed versions $\bar{H}$ and $\hat{H}$.
- Check local stability of the ODE by linearizing $\hat{H}$ about the equilibrium $\tilde{\theta} = 0$ (recall that $\tilde{\theta} = 0$ precisely when the algorithm has achieved its optimum performance)
- Examine the forced ODE (69) to determine the convergent distribution of the algorithm.

6.1 LMS

Analysis of LMS does not require all the machinery of the last two sections, but it is an important special case. In the absence of disturbances, the deterministic approach shows that the LMS parameter estimates converge exponentially to θ^* if the stepsize is chosen "small enough", and if the excitation matrix $M = \sum \phi\phi^T$ (53) is positive definite. Convergence occurs at a rate proportional to the smallest eigenvalue α of M. When disturbances are present, the convergence is to a δ ball about θ^*, where the size of the ball is proportional to the disturbance $\xi(k)$, and inversely proportional to α and μ. Unfortunately, there is no easy way to determine the constants of proportionality or to state explicitely how large "small" can be.

The condition that $E[\phi\phi^T]$ positive definite implies convergence in distribution is well known [53] though it appeared that the limiting distribution was strongly dependent on the input distribution [8]. Our theorem demonstrates that the limiting distribution is closely approximated by a Guass-Markov process irrespective of the input, assuming sufficiently smooth disturbances, mixing, and sufficiently small stepsize. This result was foreshadowed in [6] (under the condition that the inputs are Gaussian), and the result is implicit in [4] and [34].

From (7) and (62), the smoothed version of H can be shown to be

$$\hat{H}(\tilde{\theta}) = E[\phi\phi^T]\tilde{\theta}, \tag{72}$$

assuming that the disturbances are zero mean. Since (72) is already linear, its "linearization" is

$$\frac{\partial}{\partial\tilde{\theta}}\hat{H}(\tilde{\theta}) = E[\phi\phi^T]. \tag{73}$$

For the "central limit theory" we may easily verify $\bar{G} = \sigma_u^2\phi\phi^T$, $\hat{G}(\tilde{\theta}) = \sigma_{\phi^2} I\sigma_u^2$, and $L(t) = 0$. Define $\alpha = \sigma_{\phi^2}$ and $\sigma^2 = \sigma_{\phi^2}\sigma_u^2$. Then the stationary distribution of $\tilde{\theta}_{[t/\mu]}$ is (approximately) $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu\sigma_u^2}{2})$.

6.2 Normalized LMS

The simplest way to analyze the NLMS algorithm is to rewrite (11) as

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) - \mu\phi(k-1)\phi^T(k-1)\tilde{\theta}(k). \tag{74}$$

as is done in [31]. Note that (74) has an "a posteriori" error $\phi^T(k-1)\tilde{\theta}(k)$ in the rightmost term in contrast to the more common "a priori" error $\phi^T(k-1)\tilde{\theta}(k-1)$ of (10). Nonetheless, all the analysis of the previous sections applies without change to (74) because the a priori and aposteriori parameter errors are virtually the same for small μ. As expected, the conditions for stability are that $\sum \phi\phi^T$ ($E[\phi\phi^T]$) be positive definite for the deterministic (stochastic) analysis.

6.3 Leakage

Although leakage is one of the most used variants of LMS it is surprisingly difficult to analyze its behaviour precisely. At a global level, it is easy to see that (12) is bounded input, bounded output stable for small μ, since all eigenvalues of $(1-\lambda)I$ are strictly within the unit circle. The problem arises because the equilibrium of the system is not independent of the inputs. Consider a scalar version of (13) with disturbance $\xi(k) = 0$ and with constant input $\phi(k) = \phi$

$$\tilde{\theta}(k) = [(1-\lambda) - \mu\phi^2]\tilde{\theta}(k-1) + \lambda\theta^*, \tag{75}$$

which has an equilibrium at $\tilde{\theta} = \frac{\lambda\theta^*}{\lambda+\mu\phi^2}$. Clearly, this equilibrium changes for different values of ϕ. Moreover, $\tilde{\theta}$ is biased away from zero, and hence $\theta(k)$ is biased away from the desired parameterization θ^*. Because of this dependence of the equilibrium point on the input, it is difficult to carry out the linearization in either the deterministic or stochastic approaches. It remains an open issue how to deal with this situation.

6.4 Dead Zone

The dead zone algorithm (18) cannot have an asymptotically stable equilibrium at the origin because the dead zone nonlinearity $g(x)$ of (17) is insensitive to small perturbations of x about 0, that is, (18) implies that the algorithm behaves like $\theta(k) = \theta(k-1)$ near 0. Rather,

the algorithm is locally exponentially stable to the region $R = [-d, d]$ as long as $\sum \phi\phi^T$ is positive definite. As usual, the stochastic analog requires that $E[\phi\phi^T]$ be positive definite.

6.5 Signed Error

The sign error algorithm fails assumption (c) of the deterministic approach (which requires differentiability of g at the origin), and the linearization and averaging approach fails. In fact, the equilibrium at $\tilde{\theta} = 0$ is unstable in the sense of Lyapunov since the *sgn* function has "infinite gain". However, a different line of deterministic reasoning, an extended Lyapunov approach, can be used as in [46] to demonstrate that the algorithm is totally stable (convergent to a small ball about the origin) when $\sum \phi\phi^T$ is positive definite.

Using an expected value approach (and assuming the independence of $\tilde{\theta}(k-1)$ from $\phi(k-1)$), [25] shows that the signed error algorithm converges in distribution to the optimal solution plus a term dependent on the stepsize when the inputs are jointly Gaussian. The stochastic approximation approach does not require this independence assumption, nor does it require differentiability of H because the disturbance ξ smooths H enough so that $\hat{H}$ can be differentiated. Suppose that the probability distribution function η is absolutely continuous with density f_ξ. Then conditions C.1 and C.2 (and hence the theorem) hold, and the corresponding linearization is

$$\frac{\partial}{\partial\tilde{\theta}}\hat{H}(0) = 2f_\xi(0)E[\phi\phi^T]. \tag{76}$$

The signum function has essentially been "smoothed away" by the averaging effect of the disturbance. The central limit results follow easily, with $\bar{G}(\tilde{\theta}, \phi) = \phi\phi^T(1-(1-2\eta(-\phi^T\tilde{\theta}))^2)$, $\hat{G}(\tilde{\theta}) = \sigma_\phi^2 I$. Since $\bar{H}(0, \phi(k)) = \hat{H}(0, \phi(k)) = 0$, $L(t) = 0$, and the limiting stochastic differential equation (69) is

$$V(t) = v_0 + \tilde{M}(t) - 2f_\xi(0)E[\phi\phi^T]\int_0^t V(s)ds. \tag{77}$$

Again under the assumptions of an independent input sequence and symmetric noise, $E[\phi\phi^T] = \sigma_\phi^2 I$, $\eta(0) = 1/2$. Now define $\alpha = 2f_\xi(0)\sigma_\phi^2$ and $\sigma^2 = \sigma_\phi^2$. Hence, $\tilde{\theta}_\mu(t) = \tilde{\theta}_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu}{4f_\xi(0)})$ density.

6.6 Signed Regressor

The original analysis of the signed regressor algorithm [42] assumed that the entries of the input are Gaussian, and assumed the independence of the input $\phi(k-1)$ and the parameter estimates $\theta(k-1)$. Combined with the small stepsize assumptions, this is enough to demonstrate mean convergence of the parameter estimates. The present approach removes these assumptions and sharpens the results.

The deterministic approach shows that if all eigenvalues of $\sum sgn(\phi)\phi^T$ have positive real parts, then the algorithm is exponentially stable. Though "most" sequences fulfill this condition, it is fairly easy to construct short periodic sequences that violate this condition, as is done in [48]. Such inputs cause the parameter estimates to diverge away from the optimal θ^*, no matter how small the stepsize.

The stochastic approach defines

$$\hat{H}(\tilde{\theta}) = \int\int \text{sgn}(\phi)(\phi^T\tilde{\theta} + \xi)d\eta(\xi)dF(\phi)$$

Assuming that the disturbance is symmetric and zero mean, this can be rewritten

$$= \int \text{sgn}(\phi)(\phi^T\tilde{\theta})dF(\phi).$$

which can be linearized as

$$\frac{\partial}{\partial\tilde{\theta}}\hat{H}(\tilde{\theta}) = E[\text{sgn}(\phi)\phi^T]. \tag{78}$$

The signed regressor algorithm was proved stable in [48] if all eigenvalues of $E[\text{sgn}(\phi)\phi^T]$ have positive real parts, and instability was conjectured if an eigenvalue has negative real parts. The stochastic approximation approach shows that this instability conjecture is true, at least locally. Examples of nontrivial stochastic processes for which $E[\text{sgn}(\phi)\phi^T]$ has negative eigenvalues were calculated in [48]. Such inputs destabilize the sign regressor algorithm.

When the inputs cause the algorithm to be stable, the stochastic approximation theorem describes the limiting distributions. The "central limit" results define $\bar{G}(\tilde{\theta},\phi) = \text{sgn}(\phi)\text{sgn}(\phi^T)\sigma_{\xi^2}$, $\hat{G}(\tilde{\theta}) = I\sigma_{\xi^2}$, and hence $L(t) = 0$. Then Eq.(69) becomes

$$V(t) = v_0 + \tilde{M}(t) - E[\text{sgn}(\phi)\phi^T]\int_0^t V(s)ds. \tag{79}$$

Let $\alpha = E[\mathrm{sgn}(\phi_1)\phi_1]$ and $\sigma^2 = \sigma_{\xi^2}$. Then $\tilde{\theta}_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu\sigma_u^2}{2E[X_1\mathrm{sgn}(X_1)]})$ density.

6.7 Sign Sign

The first example of instability in an adaptive FIR filter demonstrated that the three periodic input sequence $\{3, -1, -1, 3, -1, -1, 3, -1, -1, ...\}$ can drive the parameter estimates of the sign-sign algorithm to infinity. This was shown by a simple inductive argument in [19], and did not give any method to distinguish between the class of signals that stabilize the algorithm from the class of signals that is destabilizing.

The stochastic approach allows such classification of signals in terms of the stochastic excitation matrix. Suppose that $\xi(k)$ is a real valued i.i.d. disturbance with probability distribution function $\eta(\cdot)$. Define $\bar{\phi} = (\phi, \mathrm{sgn}(\phi))$. Then

$$\bar{H}(\tilde{\theta}, \bar{\phi}) = \mathrm{sgn}(\phi) \int \mathrm{sgn}(\phi^T\tilde{\theta} + \xi) d\eta(\xi) = \mathrm{sgn}(\phi)(1 - 2\eta(-\phi^T\tilde{\theta})) \tag{80}$$

is continuous in $(\tilde{\theta}, \bar{\phi})$ if, for example, $\eta(\cdot)$ is absolutely continuous with density $f_{\xi(\cdot)}$. Consequently, conditions C.1 and C.2 (and hence the theorem) hold. If $F(\cdot)$ denotes the marginal distribution function of $\{X_k\}$, then

$$\hat{H}(\tilde{\theta}) = \int \mathrm{sgn}(\phi)[1 - 2\eta(-\phi^T\tilde{\theta})] dF(\phi), \tag{81}$$

which can be linearized about the equilibrium $\tilde{\theta} = 0$ as

$$\frac{\partial}{\partial\tilde{\theta}}\hat{H}(0) = 2f_\xi(0)E[\mathrm{sgn}(\phi)\phi^T]. \tag{82}$$

For the central limit results, note that

$$\bar{G}(\tilde{\theta}, \bar{\phi}) = \mathrm{sgn}(\phi)\mathrm{sgn}(\phi^T)(1 - (1 - 2\eta(-\tilde{\theta}^T\phi))^2). \tag{83}$$

For nontrivial symmetric i.i.d. inputs, $E[\mathrm{sgn}(\phi\phi^T)] = I$, and

$$\hat{G}(\tilde{\theta}) = I - E_\phi[\mathrm{sgn}(\phi\phi^T)(1 - 2\eta(-\phi^T\tilde{\theta}))^2], \tag{84}$$

or $\hat{G}(0) = (1 - (1 - 2\eta(0))^2)I = 4(\eta(0) - \eta(0)^2)I = I$ for the case of symmetric noise. Define $\alpha = 2f_\xi(0)E[\phi_1\mathrm{sgn}(\phi_1)]$ and $\sigma^2 = 1$. Practically speaking, this implies that for small

μ, the approximation $V_\mu(t) = \frac{1}{\sqrt{\mu}}(\tilde{\theta}_\mu(t) - \tilde{\theta}(t)) \approx V(t)$, where $V(t)$ has a $N(0, \frac{\sigma^2}{2\alpha})$ density, and $\tilde{\theta}(t) \approx 0$. Hence $\tilde{\theta}_\mu(t) = \tilde{\theta}_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu}{4f_\xi(0)E[\phi_1 \mathrm{sgn}(\phi_1)]})$ density.

The form of the stochastic excitation matrix $E[sgn(\phi)\phi^T]$ ties the stability properties of the sign-sign algorithm to the stability properties of the sign regressor algorithm, and it is reasonable to anticipate that a condition on the positivity of the (real parts of the) eigenvalues of $\sum sgn(\phi)\phi^T$ will be the correct deterministic criterion for stability of the sign-sign algorithm. This is, however, false. Consider the twelve periodic input sequence $\{3, -1, -1, 3, -1 - 1, 3 - 1 - 1, 3, -1, -7\}$. This can be shown (via an inductive argument) to destabilize the three dimensional sign-sign algorithm just as the example in [19], but all eigenvalues of $\sum \mathrm{sgn}(\phi)\phi^T$ have positive real parts. Hence this input stabilizes the sign regressor algorithm, but destabilizes the sign-sign algorithm. Thus $\sum \mathrm{sgn}(\phi)\phi^T$ is *not* the correct stability criterion for the deterministic sign - sign algorithm. Yet we have shown that both sign regressor and sign-sign are locally stable exactly when the real parts of the eigenvalues of $E[\mathrm{sgn}(\phi)\phi^T]$ are positive. The explanation of this apparent contradiction is simple, though somewhat surprising. Throughout the stochastic approximation approach, we have assumed that the disturbance term is "smooth" enough to "average out" the discontinuities. An identically zero disturbance does not give enough smoothing! Thus the presence of disturbances is crucial to being able to state a concise condition for the stability of the algorithm. As evidence that this is the correct interpretation, one can resimulate the sign - sign algorithm with the same 12 periodic sequence above, but adding a small disturbance. The algorithm stabilizes, converging to a small ball about the optimal parameterization. This is discussed further in [13].

6.8 Quantized State

The quantization functions Q_1 and Q_2 of the QS algorithm (26) are typically "staircase" functions which may be zero in some region about the origin, or they may be discontinuous at the origin. If Q_2 is differentiable at the origin, then the deterministic approach shows that $\sum Q_1(\phi)\phi^T$ is the appropriate persistence of excitation condition. If Q_2 is discontinuous at the origin, then an extended Lyapunov approach can be used as in [46] to show total

stability (though not exponential stability). As in the sign- sign algorithm, the stochastic approximation approach does not require continuity at the origin due to the smoothing of the disturbance. As expected, the stochastic analog of the PE condition is that all eigenvalues of $E[Q_1(\phi)\phi^T]$ have positive real parts.

6.9 Least Mean Fourth

Deriving excitation conditions for the least mean q^{th} algorithm (28) is straightforward in both the deterministic and stochastic settings. Since $F(\phi) = \phi$ and $g(e) = e^{q-1}$, $g'(0) = 0$ and the deterministic PE condition requires that $\sum \phi\phi^T$ be positive definite. By the corollary, convergence is exponential to the region $[-r, r]$, and $r > 0$ can be chosen arbitrarily small. As expected, the equivalent stochastic condition is that $E[\phi\phi^T]$ be positive definite.

6.10 Median LMS

The median LMS was included in this chapter as an example of an LMS variant that cannot be fully analyzed via any of the known methods. The approach in [55] is very much in the spirit of the "expected value" analyses, and is restricted to showing mean convergence when the input and parameter vectors are assumed independent.

The median LMS cannot be analyzed in the deterministic framework because it does not have the "nice" form of (2) which distinguishes between the function F on the regressor and the function g on the error. Moreover, the median function is not memoryless. It operates explicitly on "old" values of both the regressor and the error. The stochastic approximation approach also cannot be applied directly to the median LMS due to its memory. Of course, there is nothing inherent in the approaches that forbids the analysis of such nonlinearities with memory, but a significant effort may be required to extend the approaches to handle updates containing medians and other order statistic nonlinearities.

6.11 Convergence and Tracking of LMS and Variants

One implication of the central limit theorem is that the signed variants of LMS converge to a Gaussian distribution with known mean and variance. A fair comparison of the convergence

speed of the algorithms can be made by adjusting the stepsize so that the final distributions of all four algorithms are identical, and to then explore the convergence rates of the algorithms. Suppose the disturbance is mean zero and symmetric with distribution $\eta(\cdot)$ and density $f_\xi(\cdot)$, and that the input is zero mean i.i.d. (Note we do not assume that the regressor vector $\phi(k)$ is i.i.d.) Then the discussion of the previous sections shows that the convergent distribution of the algorithms is $N(0, \frac{\sigma^2\mu}{2\alpha})$ where

- sign - sign: $\sigma^2 = 4(\eta(0) - \eta^2(0))$ and $\alpha = 2f_u(0)E[\text{sgn}(\phi)\phi^T]_{ii}$
- signed error: $\sigma^2 = 4\sigma_{\phi^2}(\eta(0) - \eta^2(0))$ and $\alpha = 2f_u(0)\sigma_{\phi^2}$
- signed regressor: $\sigma^2 = \sigma_u^2$ and $\alpha = E[\text{sgn}(\phi)\phi^T]_{ii}$
- LMS: $\sigma^2 = \sigma_{\phi^2}\sigma_u^2$ and $\alpha = \sigma_{\phi^2}$

where $E[\text{sgn}(\phi)\phi^T]_{ii}$ represents a diagonal term of the matrix $E[\text{sgn}(\phi)\phi^T]$. Suppose the input is i.i.d. uniform [-0.5,0.5] (which fulfills both stability criteria $E[\phi\phi^T]$ and $E[\text{sgn}(\phi)\phi^T]$), the distribution of the disturbance is (1/10)N(0,1), and the desired variance is 0.0025. This can be achieved by choosing

- $\mu = 0.01$ for sign - sign
- $\mu = 0.04$ for sign error
- $\mu = \frac{\sqrt{2\pi}}{20}$ for signed regressor
- $\mu = \frac{\sqrt{2\pi}}{5}$ for LMS

Using these four values, all four signed algorithms have the same convergent distribution. This was verified experimentally in figure 1 (all figures are taken from [13]) which shows the four simulated densities for the LMS, Signed Regressor, Signed Error, and Sign-Sign algorithms. The data was gathered over one million iterations, and corresponds remarkably well with the predicted gaussian density. The theory asserts that these simulated densities must converge as μ vanishes; the figure is striking because the stepsizes are not "small" compared to the sizes typically used in applications.

It is now possible to fairly compare the convergence speed and tracking abilities of the algorithms. Figure 2, for instance, shows the trajectories of the four algorithms with the same input and stepsizes as above. All four are initialized at the same "wrong" answer and converge towards zero (the desired answer). Typically, the algorithms which can respond to large errors by taking a larger step (LMS and signed regressor) converge faster than the algorithms which must react through the signum function of the error. This may not always be the case, however, since the relative performance of the algorithms may differ depending on the distributions of the input and disturbance processes. The importance of the central limit theorem in this regard is that it shows how to fairly conduct such a study, allowing a more knowledgeable choice of algorithm and stepsize for a given application setting.

A second important area in terms of performance is the algorithms ability to track a moving parameterization. Reconsider (71). This ODE is forced by the term $\int \psi$, which represents the motion of the parameters that the algorithm is trying to identify. The term $\int \hat{H}(\tilde{\theta})$ represents the exponentially stable transient part (assuming all eigenvalues of linearization have positive real parts) that dies away as the algorithm converges to a region about the current $\tilde{\theta}^*$. Since (71) is essentially the same in all four cases (except for the details of the linearization), this implies that all four algorithms have roughly the same performance in terms of tracking ability, presuming the motion of $\tilde{\theta}^*$ is slow enough. Figure 3 shows simulations of the four algorithms tracking a slowly moving parameter (the four are offset from each other so that they can be distinguished in the picture). The stepsizes are chosen as above so that the final convergent distributions match. As implied by the ODE analysis, it doesn't matter which algorithm is used when tracking a slowly moving parameterization. Differences in tracking performance would undoubtedly arise when the motion of $\tilde{\theta}^*$ becomes rapid. In this case, algorithms which converge faster will likely have an advantage over those (such as sign-sign) which have a bounded rate of change.

7 Conclusion

The three branches of analysis are complementary in the sense that they provide different insights into the behaviours of the various algorithms. The expected value approach gives

useable answers and guidelines for implementation that are often quite straightforward, especially when the inputs are Gaussian. The deterministic results provide convergence and divergence proofs in terms of the eigenvalues of the excitation matrix, and total stability in the case of bounded disturbances. The stochastic convergence theorems are weaker, but they may allow calculation of the final distribution of the parameter estimates about their optimal value in terms of the distributions of the inputs and disturbances. On the other hand, the expected value approach suffers from problems with rigor, the deterministic approach is riddled with ϵ's and δ's that are hard to quantify, and the stochastic approximation approach can become tangled with an unmanagable number of expressions for densities and distributions, all defined in terms of one another.

There are, of course, numerous other approaches to the analysis of LMS and its variants. Lyapunov theory can often be applied to show convergence (or at least stability) in the ideal, noise free case. When possible, this is a powerful method because it can give global results, rather than local like the present averaging approach. Large deviations theory can be applied to answer questions about the expected length of time until the parameter estimates reach a certain bound, and may give insights into the expected time until failure of adaptive applications which contain a feedback of the error path back into the input of the adaptive element [12], [11].

Recently, the Poisson Clumping Heuristic has been applied to describe the probability of reaching large bounds b as being distributed in a Poisson manner with parameter λ_b [45]. Together with the stochastic approximation results, this characterizes the behaviour of the parameter estimates quite fully. Near the equilibrium, the process behaves in an essentially Gaussian manner, while far from the equilibrium, it is Poisson.

Even after all these years and all these papers, there are still new questions to be asked about the LMS family. Some issues are

- How can the approaches be generalized to apply to the median and other order statistic algorithms?
- What happens to the deterministic and stochastic excitation conditions when the adaptive element is enclosed in a feedback loop?

- The extension of the approaches to the IIR (infinite impulse response) adaptive filters is sometimes straightforward, though often such extensions require new insights.
- How can the ϵ's and δ's of the deterministic approach be quantified?
- How can the bias inherent in algorithms such as the LMS with leakage be precisely analyzed?
- What is the actual excitation condition for the deterministic sign-sign algorithm?

References

[1] M. A. Aizerman, E. M. Braverman, and R. I. Rozonoer, "The method of potential functions for restoring the characteristic of a function from randomly observed points," *Automatic Remote Control,* Vol. 25, No.12, Dec 1964.

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

[3] B. D. O. Anderson, I. M. Y. Mareels, W. A. Sethares, and C. R. Johnson, "Averaging theory for sign-sign LMS," Proceedings, 26th Annual Allerton Conference on Communication, Control, and Computing, September, 1988.

[4] A. Benveniste, M. Goursat, and G. Ruget, "Analysis of stochastic approximation schemes with discontinuous and dependent forcing terms with applications to data communication algorithms," *IEEE Trans. Automatic Control,* Vol. 25, No. 6, Dec 1980, pp. 1042-1058.

[5] N. J. Bershad, "Comments on 'Comparison of the convergence of two algorithms for adaptive FIR digital filters," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 33, Dec 1985.

[6] N. J. Bershad and L.Z. Qu, "On the probability density function of the complex scalar LMS adaptive weights," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 37, No.1, Jan 1989.

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

[8] R. R. Bitmead, "Convergence in distribution of LMS-type adaptive parameter estimates," *IEEE Trans. on Automatic Control,* Vol. 28, No. 1, Jan. 1983.

[9] R. R. Bitmead, "Persistence of excitation conditions and the convergence of adaptive systems," *IEEE Trans. on Information Theory,* Vol. 30, No. 2, pp. 183-191, March 1984.

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

[11] R. R. Bitmead and P. E. Caines, "Escape time formulation of robust stochastic adaptive control," *Proc. 27th Conf. on Decision and Control,* Austin TX, Dec. 1988.

[12] T. Brennan, "Large deviation theory and the asymptotics of convergent LMS algorithms," *1990 International Symposium on Information Theory*, San Diego, CA, Jan. 1990.

[13] J. A. Bucklew, T. Kurtz, and W. A. Sethares, "Results on local stability of fixed step size recursive algorithms," *Proc. 1992 IEEE Conf. on Acoustics, Speech and Signal Proc,* San Fransisco, March 1992. Also submitted as "Local stability and tracking properties of adaptive algorithms," to the *IEEE Trans. on Information Theory.*

[14] CCIT Red Book, Recommendation G721, Tome III-3, Oct. 1984.

[15] R. Y. Chen and C. L. Wang, "On the optimal step size for the adaptive sign and LMS algorithms," *IEEE Trans. on Circuits and Systems*, June 1990.

[16] S. H. Cho and V. J. Matthews, "Tracking analysis of the sign algorithm in nonstationary environments," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol. ASSP-38, no. 12, Dec. 1990.

[17] T. A. C. M. Claasen and W. F. G. Mecklenbrauker, "Comparison of the convergence of two algorithms for adaptive FIR digital filters," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol. ASSP-29, June 1981.

[18] C. F. N. Cowan and P. M. Grant, *Adaptive Filters*, Prentice-Hall, 1985.

[19] S. Dasgupta and C. R. Johnson, Jr., "Some comments on the behavior of sign-sign adaptive identifiers," *Systems and Control Letters* Vol. 7, pp.75-82, April 1986.

[20] D. L. Duttweiler, "Adaptive filter performance with nonlinearities," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 30, Aug 1982.

[21] L. J. Eriksson, M. C. Allie, and R. A. Greiner, "The selection and application of an IIR adaptive filter for use in active sound attenuation, *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 35, No. 4, April 1987.

[22] S. Ethier and T. Kurtz, *Markov Processes - Characterization and Convergence,* Wiley-Interscience, New York, 1986.

[23] J. B. Evans, P. Xue, and B. D. Liu, "Analysis and implementation of variable step-size adaptive algorithms," To appear, *IEEE Trans. on Acoustics, Speech, and Signal Processing,*

[24] E. Eweda, "Optimum step size of sign algorithm for nonstationary adaptive filtering," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Nov. 1990.

[25] A. Gersho, "Adaptive filtering with binary reinforcement," *IEEE Trans. on Information Theory,* Vol. IT-30, No. 2, pp. 191-198, March 1984.

[26] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction, and Control,* Prentice-Hall, 1984.

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

[28] S. Haykin, *Adaptive Filter Theory,* Prentice-Hall, 1986.

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

[30] Jayant and Noll, *Digital Coding of Waveforms,* Prentice Hall, 1984.

[31] C. R. Johnson, Jr., *Lectures on Adaptive Parameter Estimation, Prentice-Hall*, 1988.

[32] T. G. Kurtz and P. Protter, "Weak limit theorems for stochastic integrals and stochastic differential equations," to appear *Annals of Probability.*

[33] H. J. Kushner and H. Huang, "Asymptotic properties of stochastic approximations with constant coefficients," *SIAM J. Control and Optimization* Vol. 19, No. 1, Jan. 1981.

[34] H. J. Kushner and A. Schwartz, "An invariant measure approach to the convergence of stochastic approximations with state dependent noise," *SIAM J. Control and Optimization,* Vol. 22, No. 1, Jan. 1984.

[35] C. P. Kwong, "Dual sign algorithm for adaptive filtering," *IEEE Trans. on Communications,* vol. COM-34, Dec. 1986.

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

[37] D. A. Lawrence and C. R. Johnson, Jr., "Recursive parameter identification algorithm stability analysis via p-sharing," *IEEE Trans. on Auto. Control*, vol. AC-25, no. 1, Jan. 1986.

[38] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. on Automatic Control* Vol. 22, No. 4, August 1977.

[39] R. W. Lucky, "Techniques for adaptive equalization of digital communication systems," *Bell Systems Technical Journal,* Vol. 45, Feb 1966.

[40] V. J. Matthews and C. H. Cho, "Improved convergence analysis of stochastic gradient adaptive filters using the sign algorithm," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol. ASSP-35, no. 4, April 1987.

[41] J. E. Mazo, "on the Independence theory of equalizer convergence," *The Bell System Technical Journal,* vol. 58, No. 5, May 1979.

[42] J. L. Moschner, "Adaptive equalization via fast quantized state methods," Tech. Dept. 6796-1, Information systems Laboratory, Stanford University, 1970.

[43] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IRE Transaction on Information Theory*, vol. IT-4, pp. 69-72, June 1958.

[44] W. A. Sethares, "Adaptive algorithms with nonlinear data and error functions," Submitted to *IEEE Trans. on Acoustics, Speech, and Signal Processing.*

[45] W. A. Sethares and J. A. Bucklew, "Excursions of adaptive algorithms via the poisson clumping heuristc," submitted to *IEEE Trans. on Acoustics, Speech, and Signal Processing.*

[46] W. A. Sethares and C. R. Johnson, Jr., "A Comparison of Two Quantized State Adaptive Algorithms," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 37, No. 1, Jan. 1989.

[47] W. A. Sethares, D. A. Lawrence, C. R. Johnson, Jr., and R. R. Bitmead, "Parameter drift in LMS adaptive filters," *IEEE Trans. on Acoustics, Speech, and Signal Proc.* Vol. 34, No. 4, August, 1986.

[48] W. A. Sethares, I. M. Y. Mareels, B. D. O. Anderson, C.R. Johnson, Jr., "Excitation Conditions for Sign-Regressor LMS," *IEEE Trans. on Circuits and Systems,* Vol. 35, No. 6, June 1988.

[49] W. A. Sethares, B. D. O. Anderson, C. R. Johnson, Jr., "Adaptive Algorithms with Filtered Regressor and Filtered Error," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 381-403, 1989.

[50] N. A. Verhoeckx and T. A. C. M. Claasen, "Some considerations on the design of adaptive filters with the sign algorithm," *IEEE Trans. on Communications,* Vol. 32, No. 3, March 1984.

[51] M Vidyasagar, *Nonlinear Systems Analysis*, Prentice Hall, 1978.

[52] B. Widrow *et al.*, "Adaptive noise cancelling: principles and applications," *Proceedings IEEE*, vol. 63, pp. 1692-1716, Dec. 1975.

[53] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proceedings of the IEEE,* Vol. 64, No. 8, pp. 1151-1162, August 1976.

[54] B. Widrow, and S. D. Stearns, *Adaptive Signal Processing,* Prentice–Hall, 1985.

[55] G. A. Williamson, P. M. Clarkson, and W. A. Sethares, "Performance characteristics of the median adaptive filter," submitted to *IEEE Trans. on Acoustics, Speech, and Signal Proc.*