Weak Convergence and Local Stability Properties of Fixed Step Size Recursive Algorithms

James A. Bucklew, *Member, IEEE*, Thomas G. Kurtz, *Senior Member, IEEE*, and William A. Sethares, *Member, IEEE*

***Abstract*—A recursive equation which subsumes several common adaptive filtering algorithms is analyzed for general stochastic inputs and disturbances by relating the motion of the parameter estimate errors to the behavior of an unforced deterministic ordinary differential equation (ODE). Local stability of the ODE implies long term stability of the algorithm while instability of the differential equation implies nonconvergence of the parameter estimates. The analysis does not require continuity of the update equation, and the asymptotic distribution of the parameter trajectories for all stable cases (under some mild conditions) is shown to be an Ornstein–Uhlenbeck process. The ODE's describing the motion of several common adaptive filters are examined in some simple settings, including the least mean square (LMS) algorithm and all three of its signed variants (the signed regressor, the signed error, and the sign–sign algorithms). Stability and instability results are presented in terms of the eigenvalues of a correlation-like matrix. This generalizes known results for LMS, signed regressor and signed error LMS, and gives new stability criteria for the sign–sign algorithm. The ability of the algorithms to track moving parameterizations 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 (under similar conditions) an Ornstein–Uhlenbeck process, whose properties can be described in a straightforward manner.**

***Index Terms*— Weak convergence, adaptive filters, recursive algorithms.**

I. Introduction

As applications of adaptive filtering, communication, control, and identification methods have grown [14], [17], [18], [19], [41], so have the number of adaptive algorithms [1], [6], [13], [16], [26], [30], [40]. Some are proposed because of their convergence properties, some because of their numerical simplicity, and others because of their noise rejection capabilities. The general recursive form

$$W_{k+1} = W_k + \mu H(W_k, Y_k, U_{k+1}) \tag{1}$$

captures most of these algorithms by suitable choice of $H(\cdot)$. In (1), W_k represents the parameter estimate errors, Y_k is some function of the inputs, U_k is a disturbance process that represents all nonidealities such as measurement and modeling errors, and μ is a small positive constant stepsize. Convergence of the process W_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 behavior of W_k arise immediately.

- Under what conditions is the process stable?
- When does there exist an asymptotic (as $k \to \infty$) distribution for W_k and how can it be characterized?

Let us define a time scaled continuous time version of (1) as

$$W_\mu(t) = W_{[t/\mu]}, \tag{2}$$

where $[z]$ represents the integer part of z.

We address our questions by relating the behavior of the scaled adaptive algorithm (2) for small μ to the behavior of the associated deterministic ordinary differential equation (ODE)

$$W(t) = W_0 + \int_0^t \hat{H}(W(s))\, ds, \tag{3}$$

where $\hat{H}(\cdot)$ is a smoothed version of $H(\cdot, \cdot, \cdot)$.

The question of when time scaled versions of the W_k process converge (as $\mu \to 0$) to $W(t)$ has been investigated by a number of researchers both for fixed μ and for the time varying stepsize cases of stochastic approximation. Many of the original notions (done in the stochastic approximations context) are due to Ljung [27] though the present approach is probably closest in spirit to [25]. (Some other important references for the fixed μ case are [4], [21], [23].) We use similar arguments although our presentation is somewhat simpler than theirs. We discuss their work at the end of the next section. The field of stochastic approximations has a long history. A good introduction to the field is given in the seminal book [39], while more modern expositions can be found in the book by Benveniste *et al.*, [5] or in any of the books by Kushner (e.g., [21], [22]).

As a byproduct of our methodology, we are able to prove the stronger almost sure convergence of the algorithms, a new result for the fixed μ algorithms.[1]

The two questions about (1) then translate into analogous questions concerning (3).

- Under what conditions is the ODE stable or unstable?

Manuscript received October 12, 1990; revised August 5, 1992. This work was presented in part at the International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, March 1992.

J. A. Bucklew and W. A. Sethares are with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Johnson Drive, Madison, WI 53706.

T. G. Kurtz is with the Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706.

IEEE Log Number 9207694.

[1] In some of the simplest no disturbance cases, [35] and [36] give (via product of random matrices type arguments) probability one convergence results of the trajectories. These results are quite different from ours, since we employ ergodic arguments to get the strong limits.

- How closely does the algorithm (1) track the behavior of the ODE (3)?

If the ODE is locally stable, then the algorithm (1) is stable at least over long time periods (indicating probable success of the adaptive scheme), while if (3) is unstable, then (1) is also unstable, and the adaptive algorithm fails. For instance, (letting $Y = X$ where X is the vector input process) it is well known [10], [40] that if the correlation matrix $E\{XX^T\}$ 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. The same matrix $E\{XX^T\}$ appears in our analysis as the linearization of $\hat{H}(W)$. 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, $E\{XX^T\}$ is always at least nonnegative definite, and the instability cannot occur.

Certain of the variants of LMS are not so fortunate. The analogous condition for the signed regressor algorithm, (letting $Y = (X, \operatorname{sgn}(X))$ for instance, requires that $E\{\operatorname{sgn}(X)X^T\}$ be positive definite [34]. As before, this same matrix appears in the present analysis as the linearization of $\hat{H}(W)$. Analogously, positive definiteness of $E\{\operatorname{sgn}(X)X^T\}$ implies stability, while a negative eigenvalue implies instability. In this case, there are nontrivial input distributions that cause instability of the associated ODE, and hence of the signed regressor algorithm.

Our methods allow us to derive analogous stability/instability conditions (often called "persistence of excitation" conditions) for algorithms which have not been previously amenable to analysis. Chief among these are the sign-sign variant of LMS [11], [12], [26]. Conditions on the input and disturbance sequences are derived which guarantee stability of the error system. Failing these conditions, the error system degenerates into (local) instability.

The relation between the adaptive algorithm (1) and the ODE (3) may be thought of as a type of "law of large numbers." To investigate how close the behavior of the algorithm is to the deterministic trajectory of the ODE, one desires a corresponding "central limit theorem." Consider the time scaled process $W_\mu(t)$. In Section II, the martingale central limit theorem is exploited to show that the error process

$$V_\mu(\cdot) = \frac{1}{\sqrt{\mu}}(W_\mu(\cdot) - W(\cdot)) \tag{4}$$

converges to a forced ODE that is driven by a sum of independent Brownian motions. Under mild assumptions on the input and disturbance processes, the limit distribution is an Ornstein–Uhlenbeck process, with known mean and variance.

In practical terms, this convergence has two major implications. First, for a given algorithm, it is easy to calculate the parameters of the steady state 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. In Section III, for instance, the mean and variance of the convergent distributions for the four signed variants of LMS are calculated when the disturbance has a density that is symmetric and zero-mean, and the input is zero-mean, independent, and identically distributed. (We do not assume that the regressor vector is independent.) Assuming fulfillment of the stability conditions, the variance of the marginal Gaussian distributions can be adjusted by choice of the stepsize μ. Alternatively, one can choose a desired final variance, and then choose the stepsize for each algorithm to achieve that variance. A fair comparison of the convergence speed of the algorithms can then be made. We performed numerical experiments to demonstrate how the comparison might actually be made.

Finally, we study the ability of certain types of adaptive algorithms to track a slowly moving parameterization. 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 parametrization. Again, the asymptotics prove to be an Ornstein–Uhlenbeck 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.

The second section demonstrates the two theorems which show that the behavior of the adaptive algorithm (1) is tied inextricably to an appropriate ODE (3). The first theorem derives the relevant ODE and the second gives bounds (in terms of an asymptotic distribution) on the difference between the parameter estimate errors and the deterministic trajectory of the ODE.

The third section presents a tutorial explanation of several common adaptive algorithms, derives the appropriate ODE, and then examines the stability properties of the ODE. For some algorithms, (LMS and the signed error algorithm), the ODE is virtually always locally stable, while others (signed regressor and sign–sign) can be locally unstable for nontrivial inputs and disturbances.

The fourth section compares the present results to previous stochastic and deterministic work. A "fair" comparison of convergence speed of the algorithms is conducted for one special case, and tracking properties of the various algorithms are compared under the assumption that the desired parameterization is slowly varying.

II. Theoretical Development

This section presents the limit theorems which relate the behavior of the algorithm (1) to an ODE (3). The update term $H(\cdot)$ in (1) has three arguments:

- W_k is the parameter estimate error;
- Y_k is a function of the inputs to the algorithm;
- U_{k+1} is the present disturbance term.

Let $\{W_k, Y_k, U_k\}$ be a random sequence defined on some probability space (Ω, F, P) and taking values in $\Re^d \times E_1 \times E_2$, where d is the number of adaptive parameters, and E_1 and E_2 are some measurable state spaces on which Y_k and U_k evolve. $\{W_k, Y_k, U_k\}$ is adapted to a filtration $\{\mathcal{F}_k\}$, (usually one takes $\mathcal{F}_k$ = the σ-algebra generated by the random variables $(W_l, Y_l, U_l)_{l=-\infty}^{k}$). Let $\mathcal{P}(A)$ denote the collection of probability measures on the space A. We assume the following.

C.1) $\{Y_k\}$ is stationary, ergodic,[2] and there is a sequence of i.i.d. E_3-valued random variables $\{\psi_k\}$, independent of $\{Y_k\}$, and a measurable function $q : \Re^d \times E_1 \times E_3 \to E_2$ such that $U_{k+1} = q(W_k, Y_k, \psi_k)$ and W_0 is independent of $\{(Y_k, \psi_k)\}$. Define $P(U_{k+1} \in C|\mathcal{F}_k) = P(q(W_k, Y_k, \psi_k) \in C|\mathcal{F}_k) = \eta(W_k, Y_k, C)$ and assume that H is integrable with respect to $\eta(w, y, \cdot)$ for each $\{w, y\} \in \Re^d \times E_1$.

$\nu_Y \in \mathcal{P}(E_1)$ will denote the distribution of Y_k. Define

$$\overline{H}(w, y) = \int_{E_2} H(w, y, u)\eta(w, y, du). \tag{5}$$

The probability distribution η of the disturbance term is used in (5) to smooth out, or average, H through the action of the integral. Of most significance for the present purpose is that $\overline{H}$ can be continuous even when H is not.[3]

C.2) $\overline{H}$ is continuous in (w, y), and for $K \in \Re^+$

$$E\left\{\sup_{w:|w|\le K} |H(w, Y_k, q(w, Y_k, \psi_k))|\right\} < \infty,$$
$$E\left\{\sup_{w:|w|\le K} |\overline{H}(w, Y_k)|\right\} < \infty. \tag{6}$$

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 $\overline{H}$ continuous. Just as $\overline{H}$ averages H, the distribution of Y_k is used to average $\overline{H}$ over the inputs Y_k, and the doubly averaged quantity

$$\hat{H}(w) = \int \overline{H}(w, y)\nu_Y(dy) \tag{7}$$

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

We now carefully state the mathematical framework in which our work is imbedded. [8], [15] are comprehensive references for most of the mathematical constructs mentioned in this paper. Let (E, r) denote a metric space with associated Borel field $\mathcal{B}(E)$. $D_E[0, \infty)$ is the space of right continuous functions with left limits mapping from the interval $[0, \infty)$ into E. We assume that $D_E[0, \infty)$ is endowed with the Skorohod topology.

Let $\{X_\alpha\}$ (where α ranges over some index set) be a family of stochastic processes with sample paths in $D_E[0, \infty)$ and let $\{P_\alpha\} \subset \mathcal{P}(D_E[0, \infty))$ be the family of associated probability distributions (i.e., $P_\alpha(B) = P\{X_\alpha \in B\}$ for all $B \in \mathcal{B}(E)$). We say that $\{X_\alpha\}$ is relatively compact if $\{P_\alpha\}$ is relatively compact in the space of probability measures $\mathcal{P}(D_E[0, \infty))$ endowed with the topology of weak convergence. $\Rightarrow$ will always denote weak convergence, $\to$ unless otherwise stated will denote convergence under the appropriate metric.

[2]Stationarity and ergodicity imply that $\sum_{k=0}^{[t/\mu]} \mu 1_B(Y_k) \to t \times \nu_Y(B)$ a.s., where ν_Y denotes the (asymptotic) distribution of the $\{Y_k\}$. This convergence is the essential assumption needed about the $\{Y_k\}$ sequence. Hence, some sort of asymptotic stationarity/ergodicity could be assumed.

[3]This idea has been previously exploited, see e.g., [4], [5], [21], [24].

Theorem 1: Let $W_\mu(t) = W_{[t/\mu]}$, and for $K \in \Re^+$, define $\tau_\mu^K = \inf\{t : |W_\mu(t)| \ge K\}$, and $W_\mu^{\tau_\mu^K}(\cdot) = W_\mu(\cdot \wedge \tau_\mu^K)$ define the "stopped" process. Assume C.1, C.2, and that $W_\mu(0) \to w_0$ in probability as $\mu \to 0$. Then for each K, $\{W_\mu^{\tau_\mu^K}, \mu > 0\}$ is relatively compact, and every limit point (as $\mu \to 0$) satisfies

$$W(t) = w_0 + \int_0^t \hat{H}(W(s))\, ds, \tag{8}$$

for $t < \tau^K = \inf\{t : |W(t)| \ge K\}$.

C.2a) Define $\check{H}(w, y, z) = H(w, y, q(w, y, z))$. Let $Q = \{(w, y, z) : \check{H}$ is continuous at $(w, y, z)\}$. Assume that $\int\int I_Q(w, y, z)\nu_Y(dy)\nu_\psi(dz) = 1$, for every w, and for $K \in \Re^+$

$$E\left\{\sup_{w:|w|\le K} |H(w, Y_k, q(w, Y_k, \psi_k))|\right\} < \infty,$$
$$E\left\{\sup_{w:|w|\le K} |\overline{H}(w, Y_k)|\right\} < \infty. \tag{9}$$

Note that in C.2a) the continuity assumption on $\overline{H}$ in C.2) has been replaced by another type of continuity assumption.

Corollary 1: Assume conditions C.1), C.2a), that $W_\mu(0) \to w_0$ almost surely, and that the solution of (8) is unique. Then, $\{W_\mu^{\tau_\mu^K}\}$ converges almost surely to W^{τ^K}. The theorem and corollary are proven in the Appendix.

The stopping time τ_μ^K measures how long it takes the time scaled process $W_\mu(t)$ to reach K in magnitude. The stopped process $\{W_\mu^{\tau_\mu^K}(t)\}$ is defined to be equal to $W_\mu(t)$ 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 $K \in \Re^+$, every possible sequence (as $\mu \to 0$) of the stopped process $\{W_\mu^{\tau_\mu^K}(t)\}$ contains a weakly convergent subsequence, and that every limit of these subsequences is a process that satisfies the ODE (8), at least up until the stopping time. *If the solution to the differential equation is unique, then the sequence actually converges in probability (not just has a weakly convergent subsequence).* The limiting quantity (the solution of the ODE) is continuous. The Skorohod topology for continuous functions corresponds to uniform convergence on bounded time intervals. Hence, convergence in probability means that for every $T > 0$, $\epsilon > 0$, $\lim_{\mu\to 0} P\big(\sup_{0\le t<\tau^K\wedge T} |W_\mu^{\tau_\mu^K}(t) - W(t)| > \epsilon\big) = 0$.

One should also note that if no solution of the ODE becomes unbounded in finite time, then we can guarantee that $\tau_\mu^K \to \infty$ as $K \to \infty$. We may then assert that $\{W_\mu\}$ is relatively compact without needing to restrict our attention to the "stopped processes."

We apply theorem 1 to some common adaptive algorithms in the next section. Note that the theorem is a form of "law of large numbers" where the time scaled process $W_\mu(t)$ plays the role of "observations" and the convergent process $W(t)$ plays the role of the "expected value" to which the $W_\mu(t)$ converge as the number of observations $\left[\frac{t}{\mu}\right]$ increases. To investigate

how this convergence occurs, the corresponding "central limit theorem" describes the weak convergence of the error process

$$V_\mu(t) = \frac{1}{\sqrt{\mu}}(W_\mu(t) - W(t)), \tag{10}$$

where the scaling factor $\frac{1}{\sqrt{\mu}}$ expands V_μ to compensate for the time compression of $W_\mu(t)$. 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 $(H - \overline{H})$ while the other $(\hat{H} - \overline{H})$ accounts for the error when averaging over the inputs.

Let $G(w,y,u) = (H(w,y,u) - \overline{H}(w,y))(H(w,y,u) - \overline{H}(w,y))^T$ be the matrix that represents the deviation of H from its smoothed version $\overline{H}$. If H is square integrable with respect to $\eta(w,y,\cdot)$ for each pair $(w,y) \in \mathfrak{R}^d \times E_1$, we can define a smoothed version of G as

$$\overline{G}(w,y) = \int_{E_2} G(w,y,u)\eta(w,y,du). \tag{11}$$

Averaging over all inputs yields

$$\hat{G}(w) = \int \overline{G}(w,y)\nu_Y(dy). \tag{12}$$

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

C.3) H is square integrable with respect to $\eta(w,y,\cdot)$ for each pair $(w,y) \in \mathfrak{R}^d \times E_1$. $\overline{H}$ is differentiable as a function of w, $\overline{G}$ and $\partial_w \overline{H}$ are continuous, and for $K \in \mathfrak{R}^+$

$$E\left\{\sup_{w:|w|\le K} |H(w, Y_k, q(w, Y_k, \psi_k))|^2\right\} < \infty$$
$$E\left\{\sup_{w:|w|\le K} |\overline{G}(w, Y_k)|\right\} < \infty$$
$$E\left\{\sup_{w:|w|\le K} |\partial_w \overline{H}(w, Y_k)|\right\} < \infty.$$

Note that C.3) implies $\hat{H}$ is locally Lipschitz (in fact continuously differentiable), so the solution of (8) is unique and hence, $V_\mu(t)$ is well defined (on any interval of which the solution of the ODE is bounded). For simplicity (so we do not have to stop our process outside of a compact set), we assume that the solution exists for all $t \ge 0$. Define

$$\tilde{M}_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (H(W_k, Y_k, U_{k+1}) - \overline{H}(W_k, Y_k))\sqrt{\mu} \tag{13}$$

and

$$L_\mu(t) = \sum_{k=0}^{[t/\mu]-1} \left(\overline{H}(W(k\mu), Y_k) - \hat{H}(W(k\mu))\right)\sqrt{\mu}. \tag{14}$$

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

C.4) $L_\mu \Rightarrow L$, where L is a zero-mean Brownian motion.

Theorem 2: Assume C.1)–C.4), that $W_\mu(0) \to w_0$ in probability, that the solution of (8) exists for all $t \ge 0$, and that $V_\mu(0) \to v_0$ in probability as $\mu \to 0$. Then, $\tilde{M}_\mu \Rightarrow \tilde{M}$ where $\tilde{M}$ is a mean zero Brownian motion independent of L with

$$E\left\{\tilde{M}(t)\tilde{M}(t)^T\right\} = \int_0^t \hat{G}(W(s))\,ds$$

and $V_\mu \Rightarrow V$ satisfying

$$V(t) = v_0 + \tilde{M}(t) + L(t) + \int_0^t \partial_w \hat{H}(W(s))V(s)\,ds. \tag{15}$$

The theorem is proved in the Appendix.

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 FIR adaptive filters. Let W_k^* denote the time varying "correct" filter coefficients that the adaptive filter is attempting to track, and let $\hat{W}_k$ be the parameter estimates. The parameter estimate error is then $W_k = W_k^* - \hat{W}_k$, which evolves according to

$$W_{k+1} = W_k + \mu H(W_k, Y_k, U_{k+1}) + (W_{k+1}^* - W_k^*). \tag{16}$$

Clearly, some restrictions must be placed on the possible motion of the filter W^*. One possibility is to assume the following.[4]

C.5) $W_k^* = \Phi(k\mu)$ where Φ is a differentiable function with derivative denoted by ϕ.

It is then easy to show that (8) can be replaced by

$$dW(t) = -\phi dt + \hat{H}(W(t))\,dt$$

or

$$W(t) = w_0 - \int_0^t \phi(s)\,ds + \int_0^t \hat{H}(W(t))\,dt. \tag{17}$$

The implications of (17), in terms of the tracking capabilities of the various adaptive algorithms, are briefly discussed in Section IV-C.

Kushner and Shwartz [25] contains the results of the LMS Algorithm in Section III-D under different hypotheses on the input sequence $\{X_k\}$. They do not require stationarity and ergodicity (or asymptotic stationarity and ergodicity), but do place assumptions on $E[X_{k+l}|\mathcal{F}_k]$ as a function of k and l. They characterize the limit as a solution of a martingale problem in the sense of Stroock and Varadhan [37]. Their approach would also apply to the Signed Regressor Algorithm in Section III-C and to any other algorithm that has the linear form

$$W_{k+1} = W_k + \mu(A_k W_k + B_k).$$

The results of Kurtz and Protter [20] provide an alternative approach to these linear algorithms which would cover both the Kushner and Shwartz results and the linear examples in the present paper.

[4] This is a deterministic parameterization. [36] does a stochastic one and gives a discussion of other choices made by other researchers.

Theorem 3: Let

$$A_\mu(t) = \sum_{k=0}^{[t/\mu]-1} \mu A_k, \qquad B_\mu(t) = \sum_{k=0}^{[t/\mu]-1} \mu B_k,$$

and suppose that there exists a matrix A and a vector B such that for each $t > 0$

$$\sup_{s\le t} |A_\mu(s) - As| \to 0, \qquad \sup_{s\le t} |B_\mu(s) - Bs| \to 0$$

in probability, and for each t

$$\left\{ \sum_{k=0}^{[t/\mu]-1} \mu(|A_k| + |B_k|), 0 < \mu < 1 \right\}$$

is stochastically bounded (e.g., has uniformly bounded expectations). Note that these conditions hold if the sequence $\{(A_k, B_k)\}$ is stationary and ergodic and $E[|A_k| + |B_k|] < \infty$. If $W_\mu(0) \to w_0$, then $W_\mu \Rightarrow W$ where W satisfies

$$W(t) = w_0 + \int_0^t (AW(s) + B)\, ds.$$

Define $M_\mu^A(t) = \frac{1}{\sqrt{\mu}}(A_\mu(t) - At)$ and $M_\mu^B(t) = \frac{1}{\sqrt{\mu}}(B_\mu(t) - Bt)$, and suppose that $M_\mu^A \Rightarrow M^A$ and $M_\mu^B \Rightarrow M^B$. If $V_\mu(0) \to v_0$, then $V_\mu \Rightarrow V$ satisfying

$$V(t) = v_0 + \int_0^t \Big(AV(s) + M^A(s)\dot{W}(s)\Big)\, ds + M^A(t)W(t) + M^B(t).$$

The theorem is proved in the Appendix.

Another related class of theorems is surveyed in the book by Benveniste, Métivier, and Priouret [5]. In their setting, the explicit assumption of stationarity (or asymptotic stationarity) is replaced by the assumption that $\{(X_k, Y_k)\}$ is a Markov chain, and averaging properties for Markov chains are exploited to obtain the desired limit.

III. Examples

This section applies the theorems to a handful of adaptive algorithms; the sign–sign algorithm, the signed error algorithm, the signed regressor algorithm, and LMS. The strategy in each example is:

- define appropriate Y_k (input) and H (update term);
- find the unforced ODE (1) by calculating the smoothed versions $\overline{H}$ and $\hat{H}$;
- check local stability of the ODE by linearizing $\hat{H}$ about the equilibrium $W = 0$ (recall that $W = 0$ precisely when the algorithm has achieved its optimum performance);
- examine the forced ODE (15) to determine the steady state distribution of the algorithm.

In the various examples, we impose some common additional assumptions on the input and disturbance processes. These are not required by the theory. Rather, they are a way to find relatively simple expressions for the stability/instability of the ODE, and for the mean and variance of the corresponding Ornstein–Uhlenbeck process. These assumptions are the following.

E1) $\{U_k\}$ is a zero-mean i.i.d. symmetric (hence $\eta(0) = 1/2$) sequence with probability distribution $\eta(\cdot)$ and bounded, continuous density $f_u(\cdot)$ with $f_u(0) > 0$. The sequence $\{X_k\}$ is a stationary, ergodic sequence (with finite mean and covariance) of $\mathfrak{R}^d$ valued random variables independent of $\{U_k\}$.

E2) Assumption E1) holds and the components of $X_j = (X_{j1}, X_{j2}, \cdots, X_{jd})^T$ are i.i.d. symmetric, mean zero, variance σ_x^2 random variables for all $j \in \mathcal{Z}$.

These, of course, are a very restrictive set of assumptions. However, they will allow us to compare in a common setting the local stability/limiting distribution behavior of the four algorithms.

A. Sign–Sign Algorithm

The sign–sign algorithm [26], prized for its computational simplicity, has seen a resurgence of interest since its incorporation in a recent CCIT standard [11] for adaptive differential pulse code modulation. Despite some efforts [3], [12], a clear and simple test for stability of the algorithm has been elusive. The algorithm is

$$W_{k+1} = W_k - \mu \operatorname{sgn}(X_k) \operatorname{sgn}\big(X_k^T W_k + U_k\big), \qquad (18)$$

where W_k is the parameter estimate error, X_k is a regressor of past inputs, $\operatorname{sgn}(X)$ applied to a vector is an element by element operation, and U_k is a disturbance term.

Suppose that the $\{U_k\}$ sequence satisfies E1). Define $y = (x \operatorname{sgn}(x))$ or $Y_k = (X_k, \operatorname{sgn}(X_k))$. Then,

$$-\overline{H}(w,y) = \operatorname{sgn}(x) \int \operatorname{sgn}\big(x^T w + u\big) f_u(u) du = \operatorname{sgn}(x)\big(1 - 2\eta\big(-x^T w\big)\big)$$

is continuous in (w, y). Thus, conditions C.1) and C.2) (and, hence, Theorem 1) hold.

Let $F(\cdot)$ denote the distribution function of X_1. Then,

$$-\hat{H}(w) = \int\int \operatorname{sgn}(x) \operatorname{sgn}\big(x^T w + u\big) f_u\, du\, dF(x) = \int \operatorname{sgn}(x)\big[1 - 2\eta\big(-x^T w\big)\big]\, dF(x).$$

Since f_u is bounded, we can show $\hat{H}$ is globally Lipshitz. Hence there exists a unique solution to the ODE that does not become unbounded in finite time. Therefore, we do not need to work with the "stopped" processes. To linearize $\hat{H}$, take the derivative with respect to w. This gives

$$-\frac{\partial}{\partial w}\hat{H}(w) = \Big(\hat{h}_{jk}(w)\Big),$$

where

$$\hat{h}_{jk}(w) = \int \operatorname{sgn}(x_j)\big(2f_u\big(-x^T w\big)\big) x_k\, dF(x).$$

This can be evaluated at the equilibrium $w = \underline{0}$

$$-\frac{\partial}{\partial w}\hat{H}(0) = 2f_u(0)E\{\operatorname{sgn}(X_1)X_1^T\}.$$

For the "central limit theorem" results, note that

$$\begin{aligned}\overline{G}(w,y) &= \int (\operatorname{sgn}(x)\operatorname{sgn}(x^Tw+u) - \operatorname{sgn}(x)\\ &\quad\cdot \int \operatorname{sgn}(x^Tw+d)\,d\eta(d))\\ &\quad\cdot (\operatorname{sgn}(x)\operatorname{sgn}(x^Tw+u) - \operatorname{sgn}(x)\\ &\quad\cdot \int \operatorname{sgn}(x^Tw+d)\,d\eta(d))^T\,d\eta(u)\\ &= \operatorname{sgn}(x)\operatorname{sgn}(x^T)\left(1-\left(1-2\eta(-w^Tx)\right)^2\right).\end{aligned}$$

Assume E2) holds, then $E\{\operatorname{sgn}(X_1X_1^T)\} = I$. Hence,

$$\begin{aligned}\hat{G}(w) &= E\left\{\operatorname{sgn}(X_1X_1^T)\left(1-\left(1-2\eta(-X_1^Tw)\right)^2\right)\right\}\\ &= I - E\left\{\operatorname{sgn}(X_1X_1^T)\left(1-2\eta(-X_1^Tw)\right)^2\right\}\end{aligned}$$

or $\hat{G}(0) = (1-(1-2\eta(0))^2)I = I$.

Recall that the Brownian driving term $L(t)$ is the limit of $L_\mu(t)$ of (14). At the equilibrium $w = 0$, $\overline{H}(0,Y_k) = -\operatorname{sgn}(X_k)(1-2\eta(0)) = 0$. Hence, $\hat{H}(0) = 0$, which implies that $L_\mu(t) = 0 \Rightarrow L(t) = 0$.

Hence, from (15), the limiting stochastic differential equation is

$$V(t) = v_0 + \tilde{M}(t) - 2f_\mu(0)E\{\operatorname{sgn}(X_1)X_1^T\}\int_0^t V(s)\,ds. \tag{19}$$

Under assumption E2), the $V(\cdot)$ process "decouples" into n independent components $V(t) = (V_1(t), V_2(t), \cdots, V_n(t))^T$ where

$$\begin{aligned}V_i(t) &= v_{0i} + \tilde{m}(t) - 2f_u(0)\\ &\quad\cdot E\{X_{1i}\operatorname{sgn}(X_{1i})\}\int_0^t V_i(s)\,ds.\end{aligned}$$

This is the general form of an Ornstein–Uhlenbeck random process. Define $\alpha = 2f_u(0)E\{X_1\operatorname{sgn}(X_1)\}$ and $\sigma^2 = 1$. Then, $V_i(t)$ is an asymptotically stationary Gaussian random process with mean zero, variance $\frac{\sigma^2}{2\alpha}$ and autocorrelation function $R_v(\tau) = E\{V_i(t+\tau)V_i(t)\} = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$.

Practically speaking, this means that for fixed t and small μ we have the approximation $V_\mu(t) = \frac{1}{\sqrt{\mu}}(W_\mu(t) - W(t)) \approx V(t)$, where $V(t)$ has a $N\left(0, \frac{\sigma^2}{2\alpha}\right)$ density, and $W(t) \approx 0$. Hence, $W_\mu(t) = W_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu}{4f_u(0)E\{X_1\operatorname{sgn}(X_1)\}})$ density.

B. Signed Error Algorithm

The signed error algorithm [16], [38] is similar to (18) but with the sgn function applied only to the error term

$$W_{k+1} = W_k - \mu X_k\operatorname{sgn}(X_k^TW_k + U_k). \tag{20}$$

Emulating this derivation (with $y = x$ or $Y_k = X_k$), it is easy to see that the corresponding linearization is (under E1)–E2))

$$\begin{aligned}-\frac{\partial}{\partial w}\hat{H}(0) &= 2f_u(0)E\{X_1X_1^T\}\\ &= 2f_u(0)\sigma_x^2 I.\end{aligned}$$

Note that $\hat{H}$ is again globally Lipschitz. The "central limit" results are also analogous (under E1)–E2)), with $\overline{G}(w,y) = xx^T(1-(1-2\eta(-x^Tw))^2)$, $\hat{G}(w) = \sigma_x^2 I$. Again, $\overline{H}(0,Y_k) = \hat{H}(0,Y_k) = 0 \Rightarrow L(t) = 0$.

Hence, the limiting stochastic differential equation (15) becomes (at the equilibrium)

$$V(t) = v_0 + \tilde{M}(t) - 2f_u(0)\sigma_x^2 I\int_0^t V(s)\,ds.$$

Define $\alpha = 2f_u(0)\sigma_x^2$ and $\sigma^2 = \sigma_x^2$. Hence, as before we have, $R_v(\tau) = E\{V_i(t+\tau)V_i(t)\} = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$ and $W_\mu(t) = W_{[t/\mu]}$ has (approximately) a $N(0,\mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu}{4f_u(0)})$ density.

C. Signed Regressor Algorithm

Applying the sgn function to only the regressor vector X_k yields [30], [34],

$$W_{k+1} = W_k - \mu\operatorname{sgn}(X_k)(X_k^TW_k + U_k). \tag{21}$$

Under E1), we obtain with ($y = (x, \operatorname{sgn}(x))$ or $Y_k = (X_k, \operatorname{sgn}(X_k))$)

$$\begin{aligned}-\hat{H}(w) &= \int\int \operatorname{sgn}(x)(x^Tw+u)f_u(u)\,du\,dF(x)\\ &= \int \operatorname{sgn}(x)(x^Tw)\,dF(x)\end{aligned}$$

(note $\hat{H}$ is linear and hence Lipschitz) and

$$-\frac{\partial}{\partial w}\hat{H}(w) = E\{\operatorname{sgn}(X_1)X_1^T\}. \tag{22}$$

The "central limit" results follow quite easily also. Under E2) it is straightforward to verify that $\overline{G}(w,y) = \operatorname{sgn}(x)\operatorname{sgn}(x^T)\sigma_u^2$, $\hat{G}(w) = I\sigma_u^2$, and $L(t) = 0$. Then, (15) becomes (at the equilibrium)

$$V(t) = v_0 + \tilde{M}(t) - E\{\operatorname{sgn}(X_1)X_1^T\}\int_0^t V(s)\,ds.$$

Let $\alpha = E\{\operatorname{sgn}(X_{1i})X_{1i}\}$ and $\sigma^2 = \sigma_u^2$. Then $R_v(\tau) = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$, and $W_{[t/\mu]}$ has (approximately) a $N(0,\mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu\sigma_u^2}{2E\{X_1\operatorname{sgn}(X_1)\}})$ density.

D. LMS Algorithm

Probably the most studied adaptive algorithm is the least mean square algorithm [40]

$$W_{k+1} = W_k - \mu X_k(X_k^TW_k + U_k) \tag{23}$$

Under E1)

$$\hat{H}(w) = -E\{X_1 X_1^T\}w = -\sigma_x^2 I w\,, \tag{24}$$

which is linear (therefore $\hat{H}$ is Lipschitz) and hence,

$$\frac{\partial}{\partial w}\hat{H}(w) = -\sigma_x^2 I\,.$$

For the "central limit theory" we may easily verify (under E2)) $\overline{G} = \sigma_u^2 x x^T$, $\hat{G}(w) = \sigma_x^2 I \sigma_u^2$, $L(t) = 0$, Equation (15) becomes

$$V(t) = v_0 + \tilde{M}(t) - \sigma_x^2 I \int_0^t V(s)\,ds\,.$$

Define $\alpha = \sigma_x^2$, $\sigma^2 = \sigma_x^2\sigma_u^2$. Then $R_v(\tau) = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$ and $W_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu\sigma_u^2}{2})$ for its stationary density. Of course, this most famous of algorithms has been treated elegantly by other researchers, most notably H. Kushner, A. Benveniste, and their coworkers.

IV. Discussion

This section compares the adaptive algorithms with previous stochastic and deterministic analyses, and compares the algorithms with each other in terms of convergence speed and tracking ability.

A. Comparison with Previous (Stochastic) Results

For LMS, the fact that $E\{XX^T\}$ positive definite implies convergence in distribution is well known [40] though it appeared that the limiting distribution (as $k \to \infty$) was strongly dependent on the input distribution [9]. Theorem 2 demonstrates that the limiting distribution is approximately Gaussian irrespective of the input, assuming sufficiently smooth disturbances, mixing, and sufficiently small stepsize. This result was foreshadowed in [7] (under the condition that the inputs are Gaussian), and the result is implicit in [4] and [25].

The signed regressor algorithm was shown to be locally stable in [34] if all eigenvalues of $E\{\text{sgn}(X)X^T\}$ have positive real parts, and instability was conjectured if an eigenvalue has negative real parts. As shown in Section III-C, this instability conjecture is true, at least locally. Examples of nontrivial stochastic processes for which $E\{\text{sgn}(X)X^T\}$ has negative eigenvalues were calculated in [34]. Such inputs destabilize the sign regressor algorithm. When the inputs cause the algorithm to be stable, Theorem 2 describes the limiting distributions.

The signed error algorithm was shown in [16] (in certain cases) to converge in distribution to the optimal solution plus a term dependent on the stepsize when the inputs are jointly Gaussian. Theorem 1 states a more general stability criterion, and Theorem 2 characterizes the limiting distribution concretely.

The sign–sign algorithm has been shown locally stable when the inputs are independent and Gaussian [3], but more general results are unavailable. Section III-A ties the stability properties of the sign–sign algorithm to the stability properties of the sign regressor algorithm. Thus the examples of [34] are also examples of stability and instability for the sign–sign algorithm.

B. Comparison with Deterministic Results

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 [2]. Speaking loosely, the deterministic "persistence of excitation condition" [10] tends to function analogously to the conditions derived here via linearization of $\hat{H}$. For example, the LMS (signed regressor) algorithm is exponentially stable when all eigenvalues of $\sum XX^T$ ($\sum \text{sgn}(X)X^T$) have positive real parts, which clearly parallels the conditions on $E\{XX^T\}$ ($E\{\text{sgn}(X)X^T\}$).

The first example of instability in an FIR adaptive filter was given in [33]. The first example that did not violate the persistence of excitation condition $\sum XX^T$ presented a period three input sequence that drives the parameter estimates of the sign–sign algorithm to infinity [12]. This spurred activity to try and determine the class of signals that stabilize and destabilize the various signed algorithms, and answers were found for LMS [10], signed regressor [34], and signed error [32]. Lacking, however, was a condition for the sign–sign algorithm.

Consider the twelve periodic input sequence $\{3, -1, -1, 3, -1, -1, 3, -1, -1, 3, -1, -7\}$. (Note that this input does not satisfy our assumptions of ergodicity. It is easy to check though that the proof of Theorem 1 is still valid for inputs of this type.) This destabilizes the three dimensional sign–sign algorithm much as the example in [12], but all eigenvalues of $\sum\{\text{sgn}(X)X^T\}$ have positive real parts. Hence, this input stabilizes the sign regressor algorithm [34]. Thus, positive definiteness of $\sum\{\text{sgn}(X)X^T\}$ 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 $E\{\text{sgn}(X)X^T\}$ are positive. The explanation of this apparent contradiction is simple, though somewhat surprising. Throughout this paper, we have assumed that the disturbance term is "smooth" enough to "average out" the discontinuities. An identically zero disturbance does not give any 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, we simulated again the sign–sign algorithm with the same 12 periodic sequence just given, this time adding a small disturbance. The algorithm stabilized, converging to a small ball about the optimal parameterization.

Similarly, for the signed error algorithm, the smoothing effect of the disturbance is necessary to demonstrate the stability of the equilibrium at $W = 0$. Deterministically (and without disturbances), the equilibrium is unstable (in the sense of Lyapunov), though it can be shown [32] that the algorithm is totally stable (convergent to a ball about the origin). The strength of the present approach is that the characteristics of the "convergent ball" can be precisely described

as the parameters of the Ornstein–Uhlenbeck distribution of Theorem 2.

C. Convergence and Tracking of LMS and Variants

One implication of Theorem 2 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. Assume conditions E1) and E2). Then in the examples we showed that the asymptotic distribution of all the algorithms is approximately $N(0, \frac{\sigma^2\mu}{2\alpha})$ where

- sign–sign: $\sigma^2 = 4\big(\eta(0) - \eta^2(0)\big)$ and $\alpha = 2f_u(0)\, E\{\mathrm{sgn}(X)X^T\}_{ii}$;
- signed error: $\sigma^2 = 4\sigma_x^2(\eta(0) - \eta^2(0))$ and $\alpha = 2f_u(0)\sigma_x^2$;
- signed regressor: $\sigma^2 = \sigma_u^2$ and $\alpha = E\{\mathrm{sgn}(X)X^T\}_{ii}$;
- LMS: $\sigma^2 = \sigma_x^2\sigma_u^2$ and $\alpha = \sigma_x^2$;

where $E\{\mathrm{sgn}(X)X^T\}_{ii}$ represents a diagonal term of the matrix $E\{\mathrm{sgn}(X)X^T\}$. Suppose the input is i.i.d. uniform $[-0.5, 0.5]$ (which fulfills both stability criteria $E\{XX^T\}$ and $E\{\mathrm{sgn}(X)X^T\}$), the distribution of the disturbance is $0.1^*N(0,1)$, 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.

The four algorithms were initialized at $W = 0$, and each was computed for 1 million iterations. Fig. 1 is the "simulated density" constructed by counting the number of times $|W|$ falls into bins of width 0.1. Note that the final distributions are virtually identical to each other despite the fact that these are not particularly "small" values of the stepsize μ. Indeed, this illustrates our assertion that the results are not limited to "vanishing" μ.

Fig. 2 shows a time plot for the same inputs, disturbances, and stepsizes, with an initialization at $W = 20$. The four trajectories converge to the same process with variance 0.0025, as before, though the speed of convergence varies with the algorithm used. Not surprisingly, 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 import of the present work in this regard is that it shows how to fairly conduct such a study, thus 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 (17). This ODE is forced by the $\int \phi$, which represents the motion of the parameters that the algorithm is trying to identify. The term $\int \hat{H}(W)$ 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 moving parameterization W^*. Since (17) is essentially the same in all four cases (except for the details

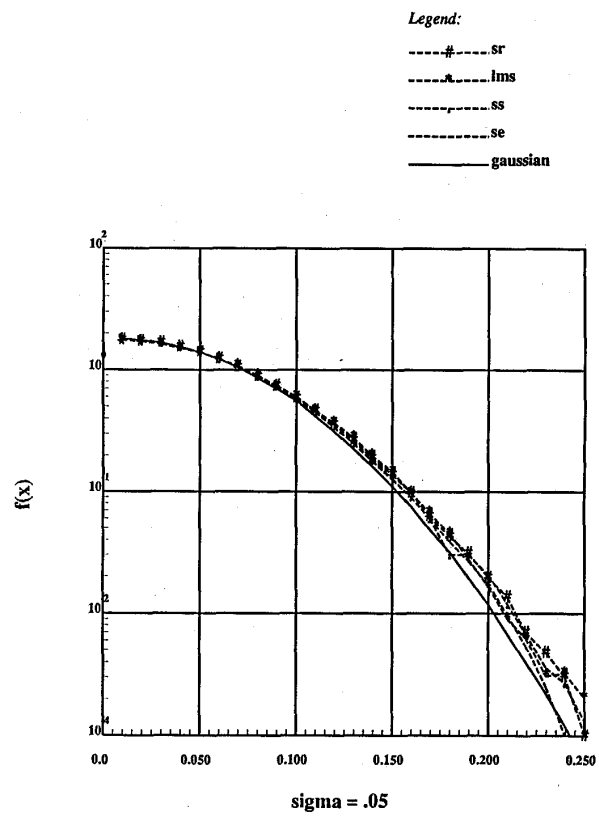

Fig. 1. Predicted and actual error densities.

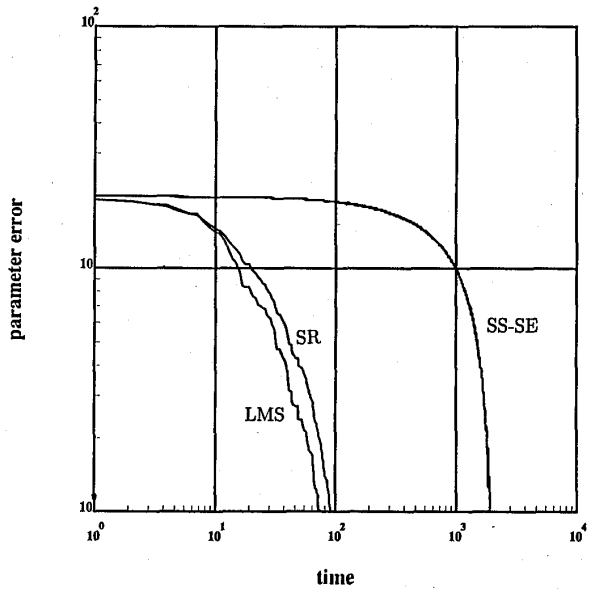

Fig. 2. "Fair" convergence behavior simulation.

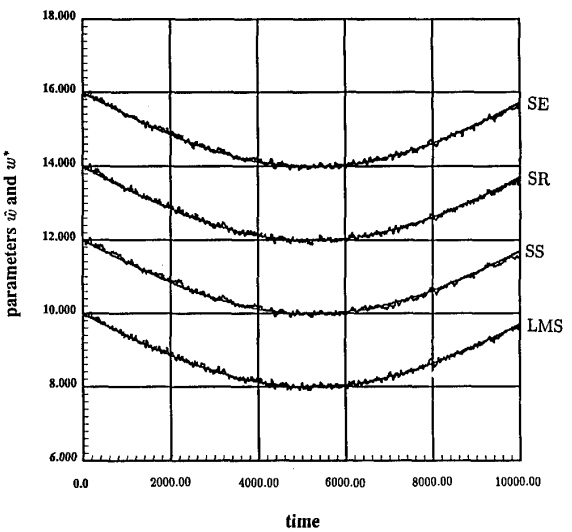

Fig. 3. Tracking behavior simulation.

of the linearization), this implies that all four algorithms have roughly the same performance in terms of tracking ability, presuming the motion of W^* is slow enough.

To illustrate this, Fig. 3 shows a time varying W^* superimposed on plots of all four algorithms. As before, all four are driven by uniform $[-0.5, 0.5]$ inputs, the distribution of the disturbance is $0.1^* N(0,1)$ and stepsizes are chosen so that all algorithms have a final variance of 0.0025. It is impossible to distinguish the four plots visually. The mean-squared error $W_k^* - \hat{W}_k$ was summed over the 10 000 iterations, giving

- 27.15 for LMS,
- 27.07 for signed error,
- 28.65 for signed regressor,
- 30.79 for sign sign,

which are the same to within experimental accuracy, and fairly close to the theoretical value of 25. We suspect the minor differences are due to the fairly large stepsizes involved. Typical applications of adaptive algorithms generally presume that W^* is slowly varying. As shown above, this implies that it does not matter which algorithm is used in terms of tracking ability. Differences in tracking performance would undoubtedly arise, however, when the motion of W^* becomes large. In this case, algorithms which converge faster will likely have an advantage over those (such as sign–sign) that have a bounded rate of change.

V. Conclusion

The behavior of four common adaptive filters has been examined by relating the motion of the parameter estimate errors to a deterministic ordinary differential equation. Appropriate stability and instability conditions are derived based on the linearization of a smoothed version of the error update. The steady-state distributions are shown to be Ornstein–Uhlenbeck under mild conditions on the input and disturbance processes. A method of choosing the stepsize in terms of the desired final variance of the Ornstein–Uhlenbeck is introduced, allowing a fair comparison of the convergence speed of the algorithms. When used in a nonstationary environment (when the desired parameterization is changing), the four algorithms are shown to track the moving parameterization equally well, assuming the motion is slow enough. Clearly, one could hope to analyze other adaptive algorithms (such as the dual sign LMS, output error identification, various adaptive controllers, the median LMS, and decision directed equalization schemes) in a similar vein. Conditions for stability and instability, and rules that relate stepsizes to the parameters of the convergent distributions should be obtainable.

Appendix
Proofs of Main Theorems

This appendix gives the proofs of the major results of Section II. A common reference for the mathematical framework needed to read this appendix may be found in [8], [15]. Let (S, d) be a complete separable metric space and let $\mathcal{M}(S)$ be the space of finite measures on S with the weak topology. Let $\mathcal{L}(S)$ be the space of measures on $[0,\infty) \times S$ such that for every $\mu \in \mathcal{L}(S)$, $u([0,t] \times S) < \infty$ for each $t \geq 0$. For $\mu \in \mathcal{L}(S)$, let μ^t denote the restriction of μ to $[0,t] \times S$. Let r_t denote the Prohorov metric on $\mathcal{M}([0,t] \times S)$ and define $\hat{r}$ on $\mathcal{L}(S)$ by

$$\hat{r}(\mu,\nu) = \int_0^\infty \exp(-t) 1 \wedge r_t(\mu^t, \nu^t) dt .$$

$\overline{C}(A)$ is defined as the space of all bounded continuous functions on the metric space A. For a metric space, E, let $D_E[0,\infty)$ be the space of right continuous E-valued functions with left limits endowed with the Skorohod topology. See [15] for definitions and properties of this space.

We first state some preliminary notions from the theory of relative compactness of random processes and functions in $D_E[0,\infty)$.

Lemma 1: Let $\{(x_n, \mu_n)\} \subset D_E[0,\infty) \times \mathcal{L}(S)$, and $(x_n, \mu_n) \to (x, \mu)$. Let $h \in \overline{C}(E \times S)$. Define

$$u_n(t) = \int_{[0,t]\times S} h(x_n(s), y) \mu_n(ds \times dy),$$
$$u(t) = \int_{[0,t]\times S} h(x(s), y) \mu(ds \times dy) .$$

Let $z_n(t) = \mu_n([0,t] \times S)$ and $z(t) = \mu([0,t] \times S)$.

a) If x is continuous on $[0,t]$ and $\lim_{n\to\infty} z_n(t) = z(t)$, then $\lim_{n\to\infty} u_n(t) = u(t)$.

b) If $(x_n, z_n, \mu_n) \to (x, z, \mu)$ in $D_{E\times\Re}[0,\infty) \times \mathcal{L}(S)$, then $(x_n, z_n, u_n, \mu_n) \to (x, z, u, \mu)$ in $D_{E\times\Re\times\Re}[0,\infty) \times \mathcal{L}(S)$. In particular, $\lim_{n\to\infty} u_n(t) = u(t)$ holds at all points of continuity of z.

c) The continuity assumption on h can be replaced by the assumption that h is continuous ν_t-a.e. for each t, where $\nu_t \in \mathcal{M}(E \times S)$ is the measure determined by $\nu_t(A \times B) = \mu\{(s,y) : x(s) \in A, s \leq t, y \in B\}$.

d) In both a) and b), the boundedness assumption on h can be replaced by the assumption that there exists a nonnegative convex function ζ on $[0,\infty)$ satisfying

$\lim_{r\to\infty} \zeta(r)/r = \infty$ such that

$$\sup_n \int_{[0,t]\times S} \zeta(|h(x_n(s),y)|)\mu_n(ds\times dy) < \infty, \quad (25)$$

for each $t > 0$.

Proof of Lemma 1: Let $h \in \overline{C}(E\times S)$. By assumption, the μ_n converge and hence, are tight. Therefore, for each $\epsilon > 0$ and $t > 0$, there exists a compact $K \subset S$ with $\sup_n \mu_n([0,t]\times K^c) \le \epsilon$. If x is continuous, then

$$\lim_{n\to\infty} \sup_{y\in K, s\le t} |h(x_n(s),y) - h(x(s),y)| = 0$$

and if $z_n(t) \to z(t)$, it follows that

$$\overline{\lim}_{n\to\infty} \left| \int_{[0,t]\times S} h(x_n(s),y)\mu_n(ds\times dy) - \int_{[0,t]\times S} h(x(s),y)\mu(ds\times dy) \right| \le 2\|h\|\epsilon,$$

which verifies part a).

If $(x_n, z_n) \to (x,z)$ in the Skorohod topology, then there exist continuous, strictly increasing mappings η_n of $[0,\infty)$ onto $[0,\infty)$ such that $\eta_n(t) \to t$ for each t and $(x_n\circ\eta_n, z_n\circ\eta_n) \to (x,z)$ uniformly on bounded intervals. Define $\tilde{\mu}_n$ so that $\tilde{\mu}_n([0,t]\times M) = \mu_n([0,\eta_n(t)]\times M)$ for any measurable set M, and observe that $\tilde{\mu}_n \to \mu$ in $\mathcal{L}(S)$. But the uniformity of the convergence of $x_n\circ\eta_n$ to x and $z_n\circ\eta_n$ to z implies

$$\begin{aligned} &\int_{[0,\eta_n(t)]\times S} h(x_n(s),y)\mu_n(ds\times dy) \\ &\quad = \int_{[0,t]\times S} h(x_n\circ\eta_n(s),y)\tilde{\mu}_n(ds\times dy) \\ &\quad \to \int_{[0,t]\times S} h(x(s),y)\mu(ds\times dy), \end{aligned} \quad (26)$$

for each fixed t. We want to show that the convergence is uniform on bounded intervals. Let $\tilde{u}_n(t)$ denote the integral on the left. It is sufficient to show that for any sequence satisfying $t_n \to t$, $\tilde{u}_n(t_n) - u(t_n) \to 0$. But this convergence holds if for any sequence satisfying $t_n \ge t$ and $t_n \to t$, we have $\tilde{u}_n(t_n) \to u(t)$ and for any sequence satisfying $t_n < t$ and $t_n \to t$, we have $\tilde{u}_n(t_n) \to u(t^-)$. Since for all r, s, $|\tilde{u}_n(s) - \tilde{u}_n(t)| \le \|h\| \mid z_n\circ\eta_n(s) - z_n\circ\eta_n(t)|$, the pointwise convergence of $\tilde{u}_n$ and the uniformity of the convergence of $z_n\circ\eta_n$ imply, in the first case, that

$$\begin{aligned} \overline{\lim}_{n\to\infty} |\tilde{u}_n(t_n) - u(t)| &= \overline{\lim}_{n\to\infty} |\tilde{u}_n(t_n) - \tilde{u}_n(t)| \\ &\le \overline{\lim}_{n\to\infty} \|h\| \cdot |z_n\circ\eta_n(t_n) - z_n\circ\eta_n(t)| \\ &\le \overline{\lim}_{n\to\infty} \|h\| \cdot |z(t_n) - z(t)| \\ &= 0, \end{aligned}$$

and in the second case, that

$$\begin{aligned} \overline{\lim}_{n\to\infty} |\tilde{u}_n(t_n) - u(t^-)| &= \overline{\lim}_{n\to\infty} \lim_{\epsilon\to 0} |\tilde{u}_n(t_n) - \tilde{u}_n(t-\epsilon)| \\ &\le \overline{\lim}_{\epsilon\to 0} \lim_{n\to\infty} \|h\| \cdot |z_n\circ\eta_n(t_n) - z_n\circ\eta_n(t-\epsilon)| \\ &= 0, \end{aligned}$$

which completes the proof of part b).

Define $\nu_t^n(A\times B) = \tilde{\mu}_n\{(s,y) : x_n\circ\eta_n(s) \in A, s \le t, y \in B\}$. Then, for a fixed t, the fact that (26) holds for each $h \in \overline{C}(E\times S)$ is just the assertion that $\nu_t^n \Rightarrow \nu_t$. We claim that this convergence is uniform on bounded time intervals. If not, then there exists a bounded sequence t_n and an $\epsilon > 0$ such that $\overline{\lim}_{n\to\infty} \rho(\nu_{t_n}^n, \nu_{t_n}) \ge \epsilon$ where ρ is the Prohorov metric on $\mathcal{M}(E\times S)$. As in the proof of part b), without loss of generality, we can assume that $t_n \to t$ and that either $t_n \ge t$ for all n or $t_n < t$ for all n. In the first case, $\nu_{t_n} \Rightarrow \nu_t$ and the uniformity of the convergence in (26) for $h \in \overline{C}(E\times S)$ implies that $\nu_{t_n}^n \Rightarrow \nu_t$, so $\rho(\nu_{t_n}^n, \nu_{t_n}) \to 0$. Similarly, in the second case, $\nu_{t_n} \Rightarrow \nu_{t-}$ and it follows that $\nu_{t_n}^n \Rightarrow \nu_{t-}$, so again $\rho(\nu_{t_n}^n, \nu_{t_n}) \to 0$. Note that if h is ν_t almost surely continuous then it is ν_{t-} almost surely continuous. For $t_n \to t$ with $t_n \ge t$, the continuous mapping theorem gives $\tilde{u}_n(t_n) \to u(t)$; and $u(t_n) \to u(t)$; for $t_n \to t$ with $t_n < t$, $u_n(t_n) \to u(t^-)$ and $u(t_n) \to u(t^-)$. Part c) then follows as in the proof of part b).

Dropping the boundedness assumption on h and assuming (25), part d) follows by approximating h by $h_c = c \wedge ((-c)\vee h)$. □

Let us define

$$\rho_\delta(x,T) = \inf_{\{t_i\}} \max_i \sup_{t_i\le s\le t<t_{i+1}} |x(s) - x(t)|,$$

where $\{t_i\}$ ranges over all partitions of the form $0 = t_0 < t_1 < \cdots < t_{n-1} < T \le t_n$ with $\min_{1\le i\le n}(t_i - t_{i-1}) > \delta$ and $n \ge 1$.

Lemma 2: A sequence $\{x_n\} \subset D_E[0,\infty)$ is relatively compact, if and only if the following two conditions hold.

a) For every rational $t \ge 0$, there exists a compact set Γ_t such that $x_n(t) \in \Gamma_t$ for all n.

b) For each $T > 0$,

$$\lim_{\delta\to 0} \overline{\lim}_{n\to\infty} \rho_\delta(x_n, T) = 0.$$

Proof of Lemma 2: See [15, pp. 123–125]. □

Let (E,r) be a metric space, $\mathcal{B}(E)$ the Borel σ-algebra of subsets of E. Given $\epsilon > 0$, define for all $F \in \mathcal{B}(E)$, $F^n = \{x \in E : \inf_{y\in F} r(x,y) < \eta\}$.

Theorem 4: Let (E,r) be complete and separable, and let $\{X_n\}$ be a family of processes with sample paths in $D_E[0,\infty)$. Suppose we have the following.

a) For every $\eta > 0$ and rational $t \ge 0$, there exists a compact set $\Gamma_{\eta,t} \subset E$ such that

$$\underline{\lim}_{n\to\infty} P(X_n(t) \in \Gamma_{\eta,t}^\eta) \ge 1 - \eta.$$

b) For each $\eta > 0$ and $T > 0$, there exist $\delta > 0$ such that

$$\overline{\lim}_{n\to\infty} P(\rho_\delta(X_n, T) \ge \eta) \le \eta.$$

Then, $\{X_n\}$ is relatively compact.

Proof of Theorem 4: See [15, pp. 128–129]. □

Lemma 3: Let $\{X_n\}$ be a sequence of processes with sample paths in $D_{\mathfrak{R}^d}[0,\infty)$. Suppose for each $T > 0$,

$$\overline{\lim}_{n\to\infty} \sup_{t\le T} |X_n(t)| < \infty \quad \text{a.s.} \tag{27}$$

and for each $T > 0$ and $\delta > 0$,

$$\lim_{\delta\to 0} \overline{\lim}_{n\to\infty} \rho_\delta(X_n, T) = 0 \quad \text{a.s.} \tag{28}$$

Then, $\{X_n\}$ is relatively compact.

Proof of Lemma 3: Since $\sup_{t\le T} |X_n(t)| < \infty$ and $\lim_{\delta\to 0} \rho_\delta(X_n, T) = 0$ for each fixed n, (27) and (28) imply

$$\sup_n \sup_{t\le T} |X_n(t)| < \infty \quad \text{a.s.}$$

and

$$\lim_{\delta\to 0} \sup_n \rho_\delta(X_n, T) = 0 \quad \text{a.s.}$$

But almost sure boundedness implies stochastic boundedness and convergence almost surely implies convergence in probability, so conditions a) and b) of Theorem 4 hold and the lemma follows. □

Proof of Theorem 1: For simplicity, assumed that C.2) holds with $\{w : |w| \le K\}$ replaced by $\mathfrak{R}^d$. The more general result can be obtained by first multiplying H (and hence, $\overline{H}$) by a continuous function with compact support which is identically 1 on $\{w : |w| \le K\}$. Define

$$M_\mu(t) = \sum_{k=0}^{[t/\mu]-1} \big(H(W_k, Y_k, U_{k+1}) - \overline{H}(W_k, Y_k)\big)\mu \tag{29}$$

and for $B \in \mathcal{B}(E_1)$, define

$$\Gamma_\mu([0,t]\times B) = \sum_{k=0}^{[t/\mu]-1} 1_B(Y_k)\mu$$

and note that the ergodic theorem implies $\Gamma_\mu \to m \times \nu_Y$ almost surely.

Now, let us consider bounding the increments of $W_\mu(t)$:

$$\begin{aligned}\sup_{s\le h} |W_\mu(t+s) - W_\mu(t)| &\le \sum_{k=[t/\mu]+1}^{[(t+h)/\mu]} \sup_w |H(w, Y_k, q(w, Y_k, \psi_k))|\mu \\ &\xrightarrow{\text{a.s.}} hE[\sup_w |H(w, Y_k, q(w, Y_k, \psi_k))|] = Ch\,, \end{aligned} \tag{30}$$

by the ergodic theorem.

This bound will be uniform for $t \le T$ almost surely. To see this, for a fixed t, let l be such that $lh \le t < (l+1)h$ and $Nh \le T < (N+1)h$. Then, $|W_\mu(t+s) - W_\mu(t)| \le |W_\mu(t) - W_\mu(lh)| + |W_\mu(t+s) - W_\mu(lh)| \le \sup_{0\le s\le h} |W_\mu(lh+s) - W_\mu(lh)| + \sup_{0\le s\le 2h} |W_\mu(lh+s) - W_\mu(lh)|$. Hence, $\sup_{0\le s\le h} |W_\mu(t+s) - W_\mu(t)| \le \sup_{0\le s\le h} |W_\mu(lh+s) - W_\mu(lh)| + \sup_{0\le s\le 2h} |W_\mu(lh+s) - W_\mu(lh)|$. Therefore, $\sup_{0\le t\le T} \sup_{0\le s\le h} |W_\mu(t+s) - W_\mu(t)| \le \max_{1\le l\le N+1}[\sup_{0\le s\le h} |W_\mu(lh+s) - W_\mu(lh)| + \sup_{0\le s\le 2h} |W_\mu(lh+s) - W_\mu(lh)|]$. Therefore,

$$\overline{\lim}_{\mu\to 0} \sup_{t\le T} \sup_{s\le h} |W_\mu(t+s) - W_\mu(t)| \le Ch\,. \tag{31}$$

Lemma 3 shows that $\{W_\mu\}$ is relatively compact.

By a similar argument, $\{M_\mu\}$ is also relatively compact. By the definition of $\overline{H}$, M_μ is a martingale. Condition C.2) ensures that for each t, the total variations up to time t, $\{T_t(M_\mu)\}$ is bounded in L_1. (Note that $|M_\mu(t)| \le \sum_{k=0}^{[t/\mu]-1} (\sup_w |H(w, Y_k, q(w, Y_k, \psi_k))| + \sup_w |\overline{H}(w, Y_k)|)\mu$.) The ergodic theorem and C.2) imply that the right-hand side of the previous inequality is a uniformly integrable sequence (in μ) of random variables. Hence, $\{M_\mu(t)\}$ is a uniformly integral sequence of random variables. This uniform integrability implies that as $\mu \to 0$, any limit point of $\{M_\mu\}$ is a martingale. Since any limit point will be continuous and of finite variation, $M_\mu \Rightarrow 0$.

We now may write

$$W_\mu(t) = W_\mu(0) + M_\mu(t) + \int_{[0,t]\times E_1} \overline{H}\big(W_\mu(s^-), y\big)\Gamma_\mu(ds \times dy)\,.$$

$\{W_\mu\}$ and $\{M_\mu\}$ are almost surely relatively compact sequences of functions in $D_{\mathfrak{R}^d}[0,t]$. We note that all possible limit points of the summands being continuous (almost surely) is sufficient to assert that the relative compactness of the summands implies relative compactness of the sum. The theorem (with $\{w : |w| \le K\}$ replaced by $\mathfrak{R}^d$) now follows by invoking Lemma 1a).

We now wish to prove the assertion that any limit of $W_\mu^{\tau^K}$ will satisfy the ODE for $t < \tau^K$. First, note that along some sequence $W_{\mu_n}^{\tau^{K_{\mu_n}}}$ is converging in probability to a solution of the ODE, W, i.e., $\lim_{\mu_n\to 0} P(\sup_{0\le t\le \tau^{K_{\mu_n}}\wedge T} |W_{\mu_n}(t) - W(t)| > \epsilon) = 0$. Now, for some ω, μ_n such that $\sup_{0\le t\le \tau^{K_{\mu_n}}\wedge T} |W_{\mu_n}(t) - W(t)| \le \epsilon$, we have $T \wedge \tau^{K-\epsilon} \le T \wedge \tau_{\mu_n}^K$. Hence, $\lim_{\mu_n\to 0} P(\sup_{0\le t\le \tau^{K-\delta}\wedge T} |W_{\mu_n}(t) - W(t)| > \epsilon) = 0$, for any $\delta > \epsilon$. Hence, we have that the limit satisfies the ODE for $t < \tau^{K-\delta}$. Since $W(t)$ is continuous, $\lim_{\delta\to 0} \tau^{K-\delta} = \tau^K$. Since $\delta > 0$ is arbitrary, the final assertion is verified. □

Proof of Corollary 1: Under the assumptions of the corollary, define $\hat{\Gamma}_{\mu,t} \in \mathcal{M}\big(\mathfrak{R}^d \times E_1 \times E_3\big)$ by

$$\hat{\Gamma}_{\mu,t}(A\times B\times C) = \sum_{k=0}^{[t/\mu]-1} I_A(W_k) I_B(Y_k) I_C(\psi_k)\mu$$

and note that

$$W_\mu(t) = W_\mu(0) + \int_{\mathfrak{R}^d\times E_1\times E_2} \check{H}(x,y,z)\hat{\Gamma}_{\mu,t}(dx\times dy\times dz)\,.$$

Define $\hat{\Gamma}^0_{\mu,t}$ by $\hat{\Gamma}^0_{\mu,t}(B\times C) = \hat{\Gamma}_{\mu,t}(\mathfrak{R}^d \times B \times C)$. Note that the ergodic theorem implies $\hat{\Gamma}^0_{\mu,t} \Rightarrow t\cdot \nu_Y \times \nu_\psi$ uniformly in $t \le T$ almost surely. Fix an ω for which (30) holds and

$\hat{\Gamma}^0_{\mu,t} \to t \cdot \nu_Y \times \nu_\psi$. By (30) and Lemma 2, the sequence of functions $\{W_\mu(\cdot,\omega)\}$ is relatively compact as a sequence of functions in $D_{\mathfrak{R}^d}[0,\infty)$. Along a subsequence of μ's such that $W_\mu(\cdot,\omega) \to w(\cdot)$, Lemma 1 implies $\hat{\Gamma}_{\mu,t}(\omega) \to \hat{\Gamma}_t$ given by

$$\hat{\Gamma}_t(A \times B \times C) = \int_{[0,t]\times E_1 \times E_3} I_A(w(s)) I_B(y) I_C(z)\, ds \times d\nu_Y(y) \times d\nu_\psi(z).$$

Under the stated assumptions, $\bar{H}$ is continuous $\hat{\Gamma}_t$-almost surely, and the last part of Lemma A.1 implies that

$$w(t) = w_0 + \int_0^t \hat{H}(w(s))\, ds.$$

Consequently, if the solution of this equation is unique there is only one possible limit point for $\{W_\mu(\cdot,\omega)\}$ and the assertion of almost sure convergence follows. □

Proof of Theorem 2: With Γ_μ defined as in the proof of Theorem 1, now define

$$Z_\mu(t) = \tilde{M}_\mu(t)\tilde{M}_\mu(t)^T - \int_{[0,t]\times E_1} \overline{G}\big(W_\mu(s^-), y\big)\Gamma_\mu(ds \times dy).$$

Observe that the independence of $\{\psi_k\}$ and $\{Y_k\}$ implies that $\tilde{M}_\mu$ and Z_μ are martingales with respect to the filtration given by $\mathcal{G}^\mu_t = \mathcal{F}_{[t/\mu]} \vee \sigma(L_\mu)$ (where $\sigma(L_\mu)$ is the σ-algebra generated by L_μ for all time). Theorem 1 implies that $\{W_\mu\}$ is converging in probability. Hence, there is a subsequence converging almost surely. Along this subsequence, Lemma 1 implies

$$\int_{[0,t]\times E_1} \overline{G}\big(W_\mu(s^-), y\big)\Gamma_\mu(ds \times dy) \to \int_0^t \hat{G}(W(s))\, ds. \tag{32}$$

Hence, we have convergence of these integrals in probability. (32) and C.3) assure that $\{\tilde{M}_\mu\}$ satisfies the conditions of the martingale central limit theorem (e.g., [15, Theorem 7.1.4]).[5] The convergence of $\{\tilde{M}_\mu\}$ to a Brownian motion (i.e., a continuous process with independent, Gaussian increments) follows.

In a similar fashion to the argument given in the proof of Theorem 1 for the uniform integrability of the $\{M_\mu(t)\}$ sequence, C.2) and C.3) imply that $\{\tilde{M}_\mu(t)\}$ and $\{Z_\mu(t)\}$ are uniformly integrable sequences. This fact and the fact that $\tilde{M}_\mu$ and Z_μ are $\{\mathcal{G}^\mu_t\}$-martingales implies that the limit $\tilde{M}$ of $\{\tilde{M}_\mu(t)\}$ and

$$Z(t) = \tilde{M}(t)\tilde{M}(t)^T - \int_0^t \hat{G}(W(s))\, ds \tag{33}$$

are martingales with respect to the filtration given by $\mathcal{G}_t = \sigma(\tilde{M}(s) : s \le t) \vee \sigma(L)$. Denote the distribution of $\tilde{M}$, which is uniquely determined by $\hat{G}$, by $P_{\hat{G}}$. Let $(\Omega, \mathcal{F}, P)$ denote the probability space on which $\tilde{M}$ and L are defined. Suppose that N is a nonnegative, Borel function defined on $C_{\mathfrak{R}^d}[0,\infty)$ such that $E^P[N(L)] = 1$. Define the probability measure Q on $\mathcal{F}$ by $dQ = N(L)\, dP$. Then, $\tilde{M}$ and Z are still $\{\mathcal{G}_t\}$-martingales under Q (that is, on the probability space $(\Omega, \mathcal{F}, Q)$), so $\tilde{M}$ has distribution $P_{\hat{G}}$ under Q. Consequently, for any bounded, Borel function F on $C_{\mathfrak{R}^d}[0,\infty)$

$$\begin{aligned} E^P\Big[F\Big(\tilde{M}\Big)N(L)\Big] &= E^Q\Big[F\Big(\tilde{M}\Big)\Big] \\ &= \int F(y) P_{\hat{G}}(dy) \\ &= E^P\Big[F\Big(\tilde{M}\Big)\Big] \\ &= E^P\Big[F\Big(\tilde{M}\Big)\Big] E^P[N(L)]. \end{aligned}$$

Since N can be any nonnegative, bounded Borel function normalized so that $E^P[N(L)] = 1$, the desired independence follows.

The process V_μ satisfies

$$\begin{aligned} V_\mu(t) = V_\mu(0) &+ \tilde{M}_\mu(t) + L_\mu(t) \\ &+ \int_{[0,t]\times E_1} \frac{1}{\sqrt{\mu}} \big(\overline{H}\big(W_\mu(s^-), y\big) - \overline{H}(W(s), y)\big) \\ &\cdot \Gamma_\mu(ds \times dy) \\ &+ \left[\sum_{\mu=0}^{[t/\mu]-1} \hat{H}(W(k\mu))\mu - \int_0^t \hat{H}(W(s))\, ds\right] \frac{1}{\sqrt{\mu}}. \end{aligned}$$

The first three terms on the right of (34) of this equation converge by hypothesis and the convergence of $\{\tilde{M}_\mu\}$; the fourth term is asymptotic to

$$R_\mu(t) = \int_{[0,t]\times E_1} \partial_w \overline{H}(W(s), y) V_\mu\big(s^-\big)\Gamma_\mu(ds \times dy) \tag{34}$$

and the fifth term goes to zero by the differentiability of $\hat{H}$ and W. For compact $K \subset \mathfrak{R}^d$, let $\tau^\mu_K = \inf\{t : V_\mu(t) \notin K\}$.

We now assert that $\{R_\mu(\cdot \wedge \tau^\mu_K)\}$ is relatively compact. Note that the processes are stopped at the time $\{V_\mu\}$ becomes too large. Therefore, V_μ is bounded in the integral defining the stopped R_μ. Hence, we may bound the increments of the stopped R_μ via C.3) and the convergence of Γ_μ. We then use an identical argument to the one used for the relative compactness of W_μ in the proof of Theorem 1 to obtain the relative compactness of $\{R_\mu(\cdot \wedge \tau^\mu_K)\}$.

Since all terms on the right of (34) are asymptotically almost surely continuous, the relative compactness of $\{V_\mu(\cdot \wedge \tau^\mu_K)\}$ follows from the relative compactness of the individual terms on the right. Finally, any limit of V_k of $\{V_\mu(\cdot \wedge \tau^\mu_K)\}$ must satisfy (15) up to $\tau_K = \inf\{t : V(t) \notin K\}$. Since the solution of (15) exists and is unique for all $t \ge 0$, $V_K(t) = V(t)$ for $t \le \tau_K$ and $\tau_K \to \infty$ as K increases to $\mathfrak{R}^d$. It follows that $V_\mu \Rightarrow V$. □

Proof of Theorem 3: A consequence of the hypothesis is that $\max_{k \le [t/\mu]} \mu |A_k| \to 0$. This observation ensures that $\{A_\mu\}$ satisfies condition C2.2(i) of [20]. Observing that

$$W_\mu(t) = W_\mu(0) + \int_0^t dA_\mu(s) W_\mu\big(s^-\big) + B_\mu(t),$$

[5] The main additional condition is that $\lim_{\mu\to 0} E\{\sup_{t\le T} |\tilde{M}_\mu(t) - \tilde{M}_\mu(t^-)|^2\} = 0$, which is true in our case.

the first conclusion follows by theorem 5.4 of [20]. The second conclusion follows by the same theorem after writing

$$V_\mu(t) = V_\mu(0) + \int_0^t dA_\mu(s)V_\mu(s-) + \int_0^t M_\mu^A(s)\dot{W}(s)\,ds + M_\mu^A(t)W(t) + M_\mu^B(t). \quad \square$$

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," *Automat. Remote Contr.*, vol. 25, no. 12, pp. 822–830, 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, and B. D. Reidle, *Stability of Adaptive Systems: Passivity and Averaging Analysis.* Cambridge, MA: 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," in *Proc. 26th Ann. Allerton Conf. Commun., Contr., and Comput.*, Sept. 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. Automat. Contr.*, vol. AC-25, no. 6, pp. 1042–1058, Dec. 1980.

[5] A. Benveniste, M. Metivier, P. Priouret, *Adaptive Algorithms and Stochastic Approximations.* New York: Springer-Verlag, 1990.

[6] N. J. Bershad, "Comments on 'Comparison of the convergence of two algorithms for adaptive FIR digital filters,'" *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 1604–1606, Dec. 1985.

[7] N. J. Bershad and L. Z. Qu, "On the probability density function of the complex scalar LMS adaptive weights," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, no. 1, pp. 43–56, Jan. 1989.

[8] P. Billingsley, *Convergence of Probability Measures.* New York: John Wiley, 1968.

[9] R. R. Bitmead, "Convergence in distribution of LMS-type adaptive parameter estimates," *IEEE Trans. Automat. Contr.*, vol. AC-28, no. 1, pp. 54–60, Jan. 1983.

[10] ——, "Persistence of excitation conditions and the convergence of adaptive systems," *IEEE Trans. Inform. Theory,* vol. IT-30, no. 2, pp. 183–191, Mar. 1984.

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

[12] S. Dasgupta and C. R. Johnson, Jr., "Some comments on the behavior of sign–sign adaptive identifiers," *Syst. Contr. Lett.*, vol. 7, pp. 75–82, Apr. 1986.

[13] D. L. Duttweiler, "Adaptive filter performance with nonlinearities," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-30, pp. 578–586, Aug. 1982.

[14] L. J. Eriksson, M. C. Allie, and R. A. Greiner, "The selection and application of an IIR adaptive filter for use in active sound attentuation," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-35, no. 4, pp. 437–438, Apr. 1987.

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

[16] A. Gersho, "Adaptive filtering with binary reinforcement," *IEEE Trans. Inform. Theory,* vol. IT-30, no. 2, pp. 191–198, Mar. 1984.

[17] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction, and Control.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[18] S. Haykin, *Adaptive Filter Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1986.

[19] N. S. Jayant and P. Knoll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[20] T. G. Kurtz and P. Protter, "Weak limit theorems for stochastic integrals and stochastic differential equations," *Ann. Probab.*, vol. 19, pp. 1035–1070, 1991.

[21] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes.* Cambridge, MA: MIT Press Series in Signal Processing, Optimization, and Control, 1984.

[22] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* New York: Springer-Verlag, 1978.

[23] H. J. Kushner and H. Huang, "Asymptotic properties of stochastic approximations with constant coefficients," *SIAM J. Contr. Optimizat.*, vol. 19, no. 1, pp. 87–105, Jan. 1981.

[24] H. J. Kushner and A. Shwartz, "An invariant meausre approach to the convergence of stochastic approximations with state dependent noise," *SIAM J. Contr. Optimizat.*, vol. 22, no. 1, pp. 13–27, Jan. 1984.

[25] ——, "Weak convergence and asymptotic properties of adaptive filters with constant gains," *IEEE Trans. Inform. Theory,* vol. IT-30, no. 2, pp. 177–182, Mar. 1984.

[26] R. W. Lucky, "Techniques for adaptive equalizaiton of digital communication systems," *Bell Syst. Tech. J.*, vol. 45, pp. 225–286, Feb. 1966.

[27] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-22, no. 4, pp. 551–575, Aug. 1977.

[28] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Trans. Inform. Theory,* vol. IT-30, no. 2, pp. 140–150, Mar. 1984.

[29] ——, "Theoremes de convergence presque sure pour une classe d'algorithmes stochastiques a pas decroissants," *Prob. Th. Rel. Fields,* vol. 74, pp. 403–428, 1987.

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

[31] G. Pflug, "Stochastic minimization with constant step-size: Asymptotic laws," *SIAM J. Contr. Optimizat.*, vol. 24, pp. 655–666, 1986.

[32] W. A. Sethares and C. R. Johnson, Jr., "A comparison of two quantized state adaptive algorithms," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, no. 1, pp. 138–143, Jan. 1989.

[33] W. A. Sethares, D. A. Lawrence, C. R. Johnson, Jr., and R. R. Bitmead, "Parameter drift in LMS adaptive filters," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-34, no. 4, pp. 868–880, Aug. 1986.

[34] W. A. Sethares, I. M. Y. Mareels, B. D. O. Anderson, C. R. Johnson, Jr., "Excitation conditions for sign-regressor LMS," *IEEE Trans. Circuits Syst.*, vol. 35, no. 6, pp. 613–625, June 1988.

[35] D. H. Shi and F. Kozin, "On almost sure convergence of adaptive algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 471–474, 1986.

[36] V. Solo, "The limiting behavior of LMS," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, pp. 1909–1922, Dec. 1989.

[37] D. W. Strook and S. R. S. Varadhan, *Multidimensional Diffusion Processes.* Berlin: Springer-Verlag, 1979.

[38] N. A. Verhoeckx and T. A. C. M. Claasen, "Some considerations on the design of adaptive filters with the sign algorithm," *IEEE Trans. Commun.*, vol. COM-32, no. 3, pp. 258–267, Mar. 1984.

[39] M. T. Wasan, *Stochastic Approximations.* Cambridge: Cambridge Univ. Press, 1969.

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

[41] B. Widrow and S. D. Stearns, *Adaptive Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1985.