Mixed Time Scale Recursive Algorithms

James. A. Bucklew* and Thomas G. Kurtz**

*Department of Electrical and Computer Engineering
**Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706-1691 USA

Corresponding author: J.A. Bucklew (bucklew@engr.wisc.edu, 608-262-3760)
Abstract

In this paper we investigate the behavior of certain types of mixed time scale adaptive algorithms. These systems comprise a “fast” or quickly changing algorithm mutually coupled to a “slow” or slowly changing algorithm. They arise naturally in a variety of adaptive environments such as in IIR system identification, the training of recurrent neural networks, decision feedback equalization, and others. (These algorithms (despite their title) should not be confused with the mixed time scales of wavelet transforms or other algorithms associated with multiresolution signal processing.) We give conditions for when the system can be analyzed from the framework of a simpler “frozen state” system. This analysis extends some of the previous work of V. Solo and his coworkers.

EDICS 4.3.1,4.3.2,6.1.1,6.1.4

Permission granted to publish this abstract separately.
1 Introduction

Many recursive algorithms have the general form

\[ W_{k+1} = W_k + \mu \times \text{“gradient”} \times \text{“error”} \]  

(1)

where the \( \{W_k\} \) are a sequence of some desired parameter estimates, the gain, \( \mu \) is usually some small constant (or possibly a sequence of constants going to zero), the “gradient” is some vector valued function of the data, and the “error” is some quantity that tells us how close we are getting to the desired performance. In many problems of practical interest, the gradient and error terms may arise from another set of recursive equations that depend upon old or previous values of \( \{W_k\} \). In many applications these equations vary much more rapidly than the adaptive estimator (with its small \( \mu \) term). Thus there are fast time scale equations coupled to the adaptation equation (a slow time scale equation), and hence the name mixed time scale algorithms. These algorithms will be the subject of this paper. We warn the reader that these algorithms should not be confused with the mixed time scales of wavelet transforms or other algorithms associated with multiresolution time-frequency analysis. The name “mixed time scale” is not our choice; we merely follow the lead of previous literature on this subject.

Many researchers have considered algorithms of this type in various settings, most notably the work of Ljung, Beneveniste, and their coworkers. Much of our notation and impetus to study these algorithms comes from [7] a superb introduction and reference book for those interested in averaging techniques in the study of signal processing algorithms. This book gives quite a lot of history and insight into the current state of the art for these mixed time scale problems.

In section 2, we develop the model to be considered and discuss the contributions of other researchers. In section 3, we present various examples of our analysis techniques. Section 4 is devoted to discussion and conclusions.
2 Development

Let \((S, r)\) denote a complete, separable, metric space (i.e. a Polish space) with associated Borel field \(\mathcal{B}(S)\), and let \(D_S[0, \infty)\) be the space of right continuous functions with left limits mapping the interval \([0, \infty)\) into \(S\). We let \(C_S[0, \infty)\) denote the subspace of continuous functions. We assume that \(D_S[0, \infty)\) is endowed with its usual topology \(\mathcal{B}(D_S[0, \infty))\) (this is called the Skorohod topology, see appendix for more definitions and a discussion).

Let \(\{X_\alpha\}\) (where \(\alpha\) ranges over some index set) be a family of stochastic processes with sample paths in \(D_S[0, \infty)\), and let \(\{P_\alpha\} \subset \mathcal{P}(D_S[0, \infty))\) be the family of associated probability distributions (i.e. \(P_\alpha(B) = P\{X_\alpha \in B\}\) for all \(B \in \mathcal{B}(D_S[0, \infty))\)). We say that \(\{X_\alpha\}\) is relatively compact if \(\{P_\alpha\}\) is relatively compact in the space of probability measures \(\mathcal{P}(D_S[0, \infty))\) endowed with the topology of weak convergence (see appendix for definition). (A set is said to be relatively compact if every sequence contained in the set has a convergent subsequence.) A reference for all the mathematical terms and probabilistic constructs used in this section and in the appendix is [4]. For related work on a similar problem, please see [2, 3].

We consider a recursive algorithm of the general form

\[
W_{k+1} = W_k + \mu H(W_k, U_{k+1}, \xi_{k+1}) \tag{2}
\]

\[
U_{k+1} = F(U_k, W_k, \eta_{k+1}) \tag{3}
\]

where \(W_k \in \mathbb{R}^d, U_k \in E_U, \{(\xi_k, \eta_k)\}\) is a Markov chain with state space \(E = E_\xi \times E_\eta\) and transition function \(q(x, y, A), E_U, E_\xi, E_\eta\) are Polish spaces and \(H\) and \(F\) are measurable functions of their arguments.

We are concerned with the behavior of the algorithm when \(\mu\) is small. In particular,
if we define $W_\mu(t) = W[t/\mu]$, we would like to derive a limit $W = \lim_{\mu \to 0} W_\mu$. Writing

$$W_\mu(t) = W_\mu(0) + \int_0^{[t/\mu]} H(W_\mu(s), U_{[s/\mu]+1}, \xi_{[s/\mu]+1}) ds,$$

we will see that under modest assumptions on $H$, we can characterize the limiting behavior of $W_\mu$, provided we can characterize the limit of the occupation measure:

$$\Gamma_\mu(A \times B \times C \times [0, t]) = \mu \sum_{k=1}^{[t/\mu]} I_{A \times B \times C}(U_k, \xi_k, \eta_k).$$

2.1 Limit Theorem

Define

$$\hat{H}(w, u, x, y) = \int_E H(w, F(u, w, y'), x') q(x, y, dx' \times dy').$$

Let $\{\mathcal{F}_k\}$ be the filtration of $\sigma-$algebras generated by $(W_k, U_k, \xi_k, \eta_k)$. Then we may write

$$\hat{H}(W_k, U_k, \xi_k) = E[H(W_k, U_{k+1}, \xi_{k+1})|\mathcal{F}_k]$$

which can be considered to be a version of $H$ that is smoothed by the distribution of the $U_{k+1}$ and $\xi_{k+1}$. This smoothed version is often differentiable even if $H$ itself is discontinuous.

Consider the following technical assumptions:

A1) $\hat{H}(w, u, x, y)$ is a bounded continuous function on $\mathbb{R}^d \times E_U \times E$.

A2)

$$\mu^2 \sum_{k=1}^{[t/\mu]} E[(H(W_k, U_{k+1}, \xi_{k+1}) - \hat{H}(W_k, U_k, \xi_k, \eta_k))^2] \to 0 \text{ as } \mu \to 0 \ \forall \ t \in [0, \infty).$$

A3) $W_\mu(0) \to w_0 \in \mathbb{R}^d$ (as $\mu \to 0$) in probability.
Theorem 2.1: Under A1-A4, \( \{(W_\mu, \Gamma_\mu)\} \) is relatively compact and every possible limit point \((W, \Gamma)\) satisfies
\[
\Gamma(du \times dx \times dy \times ds) = \gamma_s(du \times dx \times dy)ds
\]
for some \( \mathcal{P}(E_u \times E) \)-valued process \( \gamma_s \) and
\[
W(t) = w_0 + \int_{E_u \times E \times [0,t]} \hat{H}(W(s), u, x, y)\gamma_s(du \times dx \times dy)ds. \tag{6}
\]
The proof is given in the appendix.

Remark 1: \((W, \Gamma)\) may still be random. \( W \) of course would be a random process in \( C_{\mathbb{R}^d}[0, \infty) \). We will look at conditions under which \( \gamma_s \) is a deterministic function of \( W(s) \). When \( \gamma_s \) is deterministic, Eq.(6) is (the integrated form of) an ODE. If that ODE has a unique solution, then \( W \) is deterministic also. This would the “standard” situation that one hopes for when trying to analyze these adaptive algorithms via averaging methods.

Remark 2: The statement that \( \Gamma(du \times dx \times dy \times ds) = \gamma_s(du \times dx \times dy)ds \) is a statement that the time marginal measure is converging to Lebesgue measure. This is immediate from Eq. (5), where one can see that the time marginal measures of \( \{\Gamma_\mu\} \) are scaled counting measures which of course converge to Lebesgue measure.

Remark 3: If the sum in A2 converges to zero for one value of \( t \), call it \( t_1 \) then clearly for all \( t \leq t_1 \) the sum converges to zero. Therefore we only require the sum to converge
to zero for a sequence of $t$ going to infinity.

**Remark 4:** Conditions A1-A2 depend directly on the $\{W_k\}$ values themselves which we are trying to obtain information about in the first place. Hence, it may at first appear very difficult to verify them. Note however that they hold in a variety of situations; for example if the function $H$ is bounded.

Verifying A4 and identifying the limiting measure is the computational and theoretical key to the understanding of this model. The next section is devoted to developing a framework to understand and verify this important assumption.

## 2.2 The Occupation Measure

### 2.2.1 Relative Compactness Conditions for the Occupation Measure

Note that we may use Eq. (5) to rewrite Eq. (4) as

$$W_\mu(t) = W_\mu(0) + \int_{E_u \times E_\xi \times [0,t]} H(W_\mu(s-), u, x, y) \Gamma_\mu(du \times dx \times dy \times ds).$$

It is reasonable to assume that the Markov chain $(\xi, \eta)$ is ergodic, and we denote its stationary probability measure by $\pi$. Our first problem is to specify conditions under which the sequence of random measures $\{\Gamma_\mu\}$ is relatively compact, which is equivalent, under the ergodicity assumption on $(\xi, \eta)$, to the relative compactness of the marginal measure

$$\Gamma_\mu^u(A) = \mu \sum_{k=1}^{[t/\mu]} I_A(U_k).$$

(We can see from Eq. (5) that $\{\Gamma_\mu\}$ are measures on the product space $E_u \times E_\xi \times E_\eta$.

The ergodicity assumption gives us convergence (and hence relative compactness) of two of these marginals, leaving us only with the question of the relative compactness of the “third” marginal measures, $\Gamma_\mu^u(\cdot)$. We then use the fact that the relative compactness
of the marginal random measures implies relative compactness of the joint random
measures (for more details see Corollary 1.2 of [5]).) Stochastic boundedness of \{U_k\}
would imply relative compactness for \{\Gamma^u_{\mu}\}. (A sequence of \(E_\mu\)-valued random
variables is stochastically bounded or tight if for every \(\epsilon > 0\), there exists a compact set \(K \subset E\)
such that \(\sup_k P(U_k \not\in K) < \epsilon\). See [1] for a classical discussion of the equivalence
between tightness, stochastic boundedness, and weak compactness.)

We consider a Lyapunov function approach and assume that \(\Psi \geq 0\) has compact
level sets (that is, \(\{u : \Psi(u) \leq m\}\) is compact for each \(m\)) and satisfies
\[
\Psi(F(u, w, y)) \leq e^{-\alpha(w, y)} \Psi(u) + \beta(w, y),
\]
for some \(\alpha : \mathbb{R}^d \times \mathcal{E}_\eta \to \mathbb{R}\) and \(\beta : \mathbb{R}^d \times \mathcal{E}_\eta \to [0, \infty)\). We then have
\[
\Psi(U_{k+1}) \leq e^{-\alpha(W_k, \eta_{k+1})} \Psi(U_k) + \beta(W_k, \eta_{k+1}) \leq e^{-\alpha(W_k, \eta_{k+1})} \Psi(U_{k-1}) + e^{-\alpha(W_k, \eta_{k+1})} \Psi(U_0) + \sum_{i=1}^{k+1} e^{-\alpha(W_i, \eta_{i+1})} \beta(W_{i-1}, \eta_i)
\]
This inequality gives stochastic boundedness for \(\{U_k\}\) under a variety of conditions, the
simplest being \(\beta(w, y) \leq \bar{\beta}(y)\) for some \(\bar{\beta}\) satisfying \(\int_{\mathcal{E}_\eta} \bar{\beta}d\pi_{\eta} < \infty\) and \(\alpha(w, y) \geq \epsilon > 0\)
for some fixed \(\epsilon > 0\). (\(\pi_{\eta}\) is the marginal distribution of the \(\eta_k\) under \(\pi\).) The next
lemma gives somewhat more subtle conditions under which stochastic boundedness will
hold. For simplicity, we assume that the conditions hold uniformly in \(w \in \mathbb{R}^d\). As with
other conditions, this uniformity could be relaxed by applying a stopping argument.

**Lemma 2.1** Assume that \(\{\eta_k\}\) is stationary and ergodic. Let \(\Psi \geq 0\) have compact level
sets and satisfy \(\Psi(F(u, w, y)) \leq e^{-\alpha(w, y)} \Psi(u) + \beta(w, y)\). Suppose that there exists \(\bar{\alpha}\)
satisfying \(\alpha(w, y) \geq \bar{\alpha}(y)\), for all \(w \in \mathbb{R}^d\), and
\[
\int_{E_\eta} \bar{\alpha}(y) \pi_{\eta}(dy) > 0,
\]
and that there is a convex, increasing function \( \varphi \) satisfying \( \lim_{r \to \infty} \varphi(r) = \infty \) such that

\[
\sup_i E[\varphi(\beta(W_{i-1}, \eta_i))] < \infty.
\]

Then \( \{U_k\} \) is stochastically bounded and hence the sequence of occupation measures \( \{\Gamma_\mu\} \) is relatively compact.

**Proof.** By (7), we have

\[
\Psi(U_{k+1}) \leq e^{-\sum_{i=0}^{k} \tilde{\alpha}(\eta_{i+1}) \Psi(U_0)} + \sum_{i=1}^{k+1} e^{-\sum_{i=1}^{k} \tilde{\alpha}(\eta_{i+1}) \beta(W_{i-1}, \eta_i)}.
\]

(8)

We can assume the \( \{\eta_k\} \) is defined for all \(-\infty < k < \infty\). By the ergodic theorem

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=k-n}^{k} \tilde{\alpha}(\eta_i) = \int_{E_\eta} \tilde{\alpha}(y) \pi_\eta(dy) > 0.
\]

and, consequently, for \( 0 < \delta < \int_{E_\eta} \tilde{\alpha} d\pi_\eta \),

\[
J_k = 0 \land \inf_{n \geq 0} \sum_{i=k-n}^{k} (\tilde{\alpha}(\eta_i) - \delta) > -\infty.
\]

Note that the distribution of \( J_k \) does not depend on \( k \), and that

\[
\sum_{i=k-n+1}^{k} \tilde{\alpha}(\eta_i) \geq J_k + n\delta \quad \forall \text{ positive integer } n.
\]

Let

\[
H_k = \frac{\sum_{i=1}^{k+1} e^{-(k-i+1)\delta} \beta(W_{i-1}, \eta_i)}{\sum_{i=1}^{k+1} e^{-(k-i+1)\delta}}
\]

By (8), we have

\[
\Psi(U_{k+1}) \leq e^{-J_{k+1}} \left( e^{-\delta} \Psi(U_0) + \frac{1}{1 - e^{-\delta}} H_k \right),
\]

and

\[
P\{\Psi(U_{k+1}) > 2C^2\} \leq P\{-J_{k+1} > \log C\} + P\{\Psi(U_0) > C\} + P\{H_k > (1 - e^{-\delta})C\}. \quad (9)
\]
The first two terms on the right do not depend on \( k \) and can be made arbitrarily small by making \( C \) large. To estimate the last term, observe that by Jensen’s inequality (or just by invoking convexity)

\[
E[\varphi(H_k)] \leq \sum_{i=1}^{k+1} e^{-(k+i+1)\delta} E[\varphi(\beta(W_{i-1}, \eta_i))] \leq \sup_i E[\varphi(\beta(W_{i-1}, \eta_i))].
\]

Consequently, the last term on the right of (9) is bounded by

\[
\frac{\sup_i E[\varphi(\beta(W_{i-1}, \eta_i))]}{\varphi((1-e^{-\delta})C)},
\]

which can also be made small by making \( C \) large. We conclude that \( \{\Psi(U_k)\} \) is stochastically bounded and hence that \( \{U_k\} \) is.

\[\square\]

### 2.2.2 Characterizing the Limits of the Occupation Measure Sequence

Once relative compactness of \( \{\Gamma_\mu\} \) is established, the limits must be characterized.

**Lemma 2.2** Suppose \((W_\mu, \Gamma_\mu)\) is relatively compact. Then any limit point \((W, \Gamma)\) must satisfy

\[
\int_{E_\mu} \int_{E_\psi} g(u, x, y, W(s)) \gamma_s(du \times dx \times dy) = \int_{E_\mu} \int_{E_\psi} g(F(u, W(s), y'), x', y', W(s)) q(x, y, dx' \times dy') \gamma_s(du \times dx \times dy)
\]

(10)

for almost every \( s \).

**Proof.** Note that for any \( g \in \tilde{O}(E_\mu \times E_\psi \times E_\eta \times \mathbb{R}^d) \)

\[
\int_{E_\mu \times E \times [0,t]} g(u, x, y, W_\mu(s-)) \Gamma_\mu(du \times dx \times dy \times ds) = \mu \sum_{k=1}^{\lfloor t/\mu \rfloor} g(U_k, \xi_k, \eta_k, W_{k-1}) = \mu \sum_{k=1}^{\lfloor t/\mu \rfloor} g(F(U_{k-1}, W_{k-1}, \eta_k), \xi_k, \eta_k, W_{k-1})
\]

10
Suppose that for each $w$ only on the bounded function $g$.

\[ L_\mu(t) = \mu \sum_{k=1}^{[t/\mu]} \int g(F(U_{k-1}, W_{k-1}, \eta_k), \xi_k, \eta_k, W_{k-1}) - E[g(F(U_{k-1}, W_{k-1}, \eta_k), \xi_k, \eta_k, W_{k-1})| \mathcal{F}_{k-1})] + \mu \sum_{k=1}^{[t/\mu]} g(F(U_{k-1}, W_{k-1}, y'), x', y', W_{k-1})q(\xi_{k-1}, \eta_{k-1}, dx', dy') \]

\[ = L_\mu(t) + \mu \sum_{k=1}^{[t/\mu]} \int g(F(U_{k-1}, W_{k-1}, y'), x', y', W_{k-1})q(\xi_{k-1}, \eta_{k-1}, dx', dy') \]

\[ = L_\mu(t) + \mu \sum_{k=1}^{[t/\mu]} \int g(F(U_{k-1}, W_{k-1}, y'), x', y', W_{k-1})q(x, y, dx' \times dy') \Gamma(du \times dx \times dy \times ds) \]

\[ - \mu g(F(U_{[t/\mu]}, W_{[t/\mu]}, \eta_{[t/\mu]+1}), \xi_{[t/\mu]+1}, \eta_{[t/\mu]+1}, W_{[t/\mu]}) + \mu g(F(U_0, W_0, \eta_1), \xi_1, \eta_1, W_0). \]

$L_\mu(t)$ is a martingale with $\sup_{t' \leq t} E[L_\mu(t')^2] \leq \mu C_g t$ (where $C_g$ is a constant that depends only on the bounded function $g$). Hence, $L_\mu \to 0$ as $\mu \to 0$. Since $(W_\mu, \Gamma_\mu)$ is relatively compact, it follows that any limit point $(W, \Gamma)$ of $\{(W_\mu, \Gamma_\mu)\}$ must satisfy (taking the limit of the first and last terms)

\[ \int_{E_U \times E} g(u, x, y, W(s)) \Gamma(du \times dx \times dy \times ds) = \int_{E_U \times E} g(F(u, W(s), y'), x', y', W(s))q(x, y, dx' \times dy') \Gamma(du \times dx \times dy \times ds). \]

Since the “time marginal” of $\Gamma$ is Lebesgue measure, we can write $\Gamma(du \times dx \times dy \times ds) = \gamma_s(du \times dx \times dy)ds$. Substituting this last expression into the last equation we get the lemma statement.

**Lemma 2.3** Consider the “frozen state” model

\[ \hat{U}_{k+1} = F(\hat{U}_k, w, \eta_{k+1}). \]

Suppose that for each $w$ the random process $(\hat{U}_k, \xi_k, \eta_k)$ has a unique stationary distribution, $\hat{\pi}_w$. Then if $(W, \Gamma)$ is a limit point of $(W_\mu, \Gamma_\mu)$, we have

\[ \gamma_s = \hat{\pi}_{W(s)}. \]

**Proof.** Let $f$ by any bounded measurable function $f : E_U \times E \to \mathbb{R}$. Define an operator $T_w$ that maps this class of functions into itself by $T_w f(u, x, y) = \int g(F(u, w, y'), x', y')q(x, y, dx' \times dy') \Gamma(du \times dx \times dy \times ds)$.
Note that for each \( w \), \( T_w \) is the transition operator for the frozen state Markov chain \( (\hat{U}_k, \xi_k, \eta_k) \). Since it is the stationary distribution of the Markov chain, \( \hat{\pi}_w \) satisfies

\[
\int_{E_U \times E} f(u, x, y) \hat{\pi}_w(du \times dx \times dy) = \int_{E_U \times E} T_w f(u, x, y) \hat{\pi}_w(du \times dx \times dy),
\]

for all bounded measurable functions \( f \). By the uniqueness of \( \hat{\pi} \) and Eq. (10) of the lemma we have that

\[
\gamma_s = \hat{\pi}_{W(s)}.
\]

The lemma states that the averaging can be considered to be done by the “frozen state” process as the heuristic explanation of the behavior of these algorithms suggests. Finding \( \hat{\pi}_w \) for every \( w \) is very often a formidable computational task whether done analytically or computationally. However it is usually quite feasible to perform the averaging on the frozen state process via Monte Carlo averaging and compute an approximation to the limiting ODE directly. We say more about this technique in the following sections.

### 3 Applications

The theory developed here is applicable to a variety of so-called mixed time scale problems [7]. The model considered by these authors has a certain linear-type structure for the fast process:

\[
U_{k+1} = A(W_k)U_k + h(k + 1, W_k) + \mu g(k + 1, W_k, \mu)
\]

where \( h(k, \cdot), g(k, \cdot) \) are stochastic, and the matrix \( A(\cdot) \) satisfies a variety of stability and Lipshitz continuity assumptions. This model allows them to consider many interesting problems including the standard recursive algorithm (Mixed Time Scale LMS) used for IIR filter system identification.
With the structure considered in this paper we are able to analyze other types of mixed time scale problems. We present a sequence of three examples. In all of the examples considered, assumptions A1-A4 will follow quite readily from boundedness of the $H, F$ functions. Also, in all of the examples, the frozen state process is a simple irreducible non-null persistent Markov chain (for each $w$) (a fact that we will not prove formally in the examples) and hence will have unique stationary distributions (for each $w$). The resulting ODE’s can then be argued to be Lipshitz (which we won’t formally do either) and hence the solution to the ODE’s will be unique.

**Example 1.** Consider the following nonlinear system:

\[
\begin{align*}
y_{k+1} &= \theta_1 w_k + \theta_2 w_{k-1} \\
z_{k+1} &= \text{sgn}(y_{k+1} + n_{k+1})
\end{align*}
\]

where $\{w_k\}$ is an i.i.d. sequence. $\{n_k\}$ is also an i.i.d. noise sequence (independent of $\{w_k\}$) of unknown variance. Thus the random process $z$ is merely the hard-limited output of the noisy observations of a moving average filter. From knowledge of the inputs $\{w_k\}$ and outputs $\{z_k\}$ we wish to estimate $\theta_1, \theta_2$. A moment’s reflection shows that we can only estimate these up to a scale factor. Thus without loss of generality, we take $\theta_1 = 1, \theta_2 = \theta$ and use the following system to estimate $\theta$.

\[
\begin{align*}
\hat{y}_{k+1} &= w_k + \hat{\theta}_k w_{k-1} \\
\hat{z}_{k+1} &= \text{sgn}(\hat{y}_{k+1}) \\
\hat{\theta}_{k+1} &= \hat{\theta}_k + \mu (z_{k+1} - \hat{z}_{k+1}) w_{k-1}
\end{align*}
\]

Thus from Equations (2-3), we have the correspondences,

\[
U_k = \begin{pmatrix} z_k \\ \hat{z}_k \end{pmatrix}
\]
\( \eta_{k+1} = \xi_{k+1} = (w_k, w_{k-1}, n_{k+1}) \), and \( W_k = \hat{\theta}_k \). A2 follows since \( \{z_k, \hat{z}_k\} \) are bounded processes and \( \{w_i\} \) is i.i.d. A3 will follow since we will assume a common starting point. A4 follows since \( F(\cdot, \cdot, \cdot) \) is bounded (the \( \{U_k\} \) sequence is the output of sign functions). A1 follows due to the smoothing of the expectation over the \( \{n_k\} \) process.

In Figure 1, the solid line shows a typical plot of the algorithm behavior for 10000 iterations with \( \mu = 0.01 \), \( \{w_k\} \) i.i.d uniform \([-0.5, 0.5]\), \( \{n_k\} \) i.i.d. normal with mean zero and standard deviation 0.1, and \( \theta = -0.5 \). The dotted line is an approximation to the limiting differential equation obtained by Monte Carlo averaging to obtain the needed expectations. To compute the “real times” of the differential equation, note that we can take \( t_0 \), the initial time to be zero. \( t_f \), the final time, is thus \( \mu \times \) number of iterations which in this case gives \( t_f = 100 \). To find an approximation to the limiting differential equation, we compute \( \theta(t + \Delta) = \theta(t) + \Delta f(\theta(t)) \). We chose to construct the ODE over a 1000 points on the interval \([t_0, t_f]\) or \( \Delta = 0.1 \). \( f(\theta(t)) \) was computed by Monte Carlo averaging 1000 samples of the “frozen state” algorithm, that is with the “hat” parameters frozen or fixed at \( \theta(t) \). The residual non-smoothness in the dotted curve is due to the residual noise in the Monte Carlo averaging.

In this problem we can derive the form of the ODE in closed form. Let us define

\[
G(W(t)) = \int_{E_U \times E} \hat{H}(W(t), u, x, y) \gamma(t) (du \times dx \times dy).
\]

Thus the integral equation in Theorem 2.1 can be written as the ODE

\[
\frac{dW(t)}{dt} = G(W(t)).
\]

In this example, the limiting measure \( \gamma_s \) is found by simple substitution. We can then easily obtain

\[
G(W) = E[(\text{sgn}(w_k + \theta w_{k-1} + n_{k+1}) - \text{sgn}(w_k + W w_{k-1}))w_{k-1}]
\]

\[
= -\frac{W}{6} - 2 \int_{-0.5}^{0.5} \int_{-0.5}^{0.5} wF_n\left(\frac{w}{2} - y\right)dydw
\]

14
Figure 1: Typical and Predicted Algorithm Behavior

\[ F_n(\cdot) = - \frac{W}{6} - 2C \]

where \( F_n(\cdot) \) is the cumulative distribution function of the \( \{n_k\} \) i.i.d. sequence of random variables and \( C \) is the constant defined by the double integral. Thus we see that the (unique) solution to the limiting ODE is an exponentially fast convergence to the limiting value \( -12C \). In our setting \( -12C \approx .4996 \). Thus our estimate is biased. This bias is more extreme with greater variance for the distribution \( F_n(\cdot) \). The circles plotted on the graph correspond to the “exact” solutions of the ODE. One can see that there is very close agreement with the “frozen state” solution.

**Example 2.** Let us now consider a non-linear autoregressive example. With the correspondence \( U_k = (z_k, z_{k-1}, \hat{z}_k, \hat{z}_{k-1}) \), the assumptions A1-A4 follow even more simply
than in the previous example due to the boundedness of the $H$ and $F$ functions of the algorithm.

\[
    z_{j+1} = \text{sgn}(a_1 z_j + a_2 z_{j-1} + y_{j+1})
\]

\[
    \hat{z}_{j+1} = \text{sgn}(\hat{a}_{1,j} \hat{z}_j + \hat{a}_{2,j-1} \hat{z}_{j-1} + y_{j+1})
\]

\[
    \begin{pmatrix}
        \hat{a}_{1,j+1} \\
        \hat{a}_{2,j+1}
    \end{pmatrix} =
    \begin{pmatrix}
        \hat{a}_{1,j} \\
        \hat{a}_{2,j}
    \end{pmatrix} +
    \mu
    \begin{pmatrix}
        \hat{z}_j \\
        \hat{z}_{j-1}
    \end{pmatrix}
    (z_{j+1} - \hat{z}_{j+1})
\]

where $\{y_j\}$ is an i.i.d. sequence of inputs and $a_1, a_2$ are the desired system parameters to be identified from the outputs $\{z_j\}$. In Figure 2, we see a sample plot of 5000 iterations of the algorithm output for the case $a_1 = .45, a_2 = .3, \mu = .01$ and the input $\{y_j\}$ is an i.i.d. sequence of uniform random variables on $[-1.5, 1.5]$. In Figure 3, we see the trajectory of the ODE over the same time period, again with Monte Carlo averaging (1000 samples) to obtain the needed expectations. The ODE is indeed converging to the correct values of $(a_1, a_2) = (.45, .3)$. 

16
Example 3. We now consider an algorithm for training recurrent neural networks, the so-called real-time recurrent learning (RTRL) algorithm first proposed in [8] though we follow the notation and discussion given in [[6], pp. 521-526]. We consider the problem of predicting a simple first order autoregressive process $x_n$ with a 4 neuron recurrent neural net with a logistic functions for the nonlinearity. The relevant equations for the learning algorithm and inputs are as follows.

$$x_{n+1} = \alpha x_n + z_k.$$  

where \{z_n\} is assumed to be i.i.d. with mean $m$. The output of the $j^{th}$ neuron at time $n+1$ is given by

$$y_j(n+1) = \phi(\sum_{i=1}^{6} w_{ji}(n)u_i(n)), \quad j = 1, \ldots, 4,$$
where \( \phi(v) = 1/(1 + \exp(-v)) \) is the logistic function. We define an augmented input vector as
\[
u_i(n) = \begin{cases} 
y_i(n) & i = 1, \ldots, 4 
x_n & i = 5 
-1 & i = 6
\end{cases}
\]
where we see we have the 4 outputs being fed back as inputs, along with the current value of the time series and \(-1\) is a threshold provision for the neurons. We now define the adaptive algorithm as
\[
\pi^j_{kl}(n+1) = y_j(n+1)(1-y_j(n+1))\left[\sum_{i=1}^{4} w_{ji} \pi^i_{kl}(n) + \delta_{kj} u_l(n)\right] \text{ and } \pi^j_{kl}(0) = 0
\]
for \( j = 1, \ldots, 4 \) \( k = 1, \ldots, 4 \) \( l = 1, \ldots, 6 \)
\[
e(n) = x(n+1) - y_1(n)
\]
\[
\Delta w_{kl}(n) = e(n) \pi^1_{kl}(n) \text{ for } k = 1, \ldots, 4 \text{ and } l = 1, \ldots, 6
\]
\[
w_{kl}(n+1) = w_{kl}(n) + \mu\Delta w_{kl}(n) \text{ for } k = 1, \ldots, 4 \text{ and } l = 1, \ldots, 6
\]

The architecture we have chosen for this prediction problem turns out to be a poor one. The algorithm basically converges to choosing the mean of the time series instead of the optimal \( \hat{x}_{n+1} = \alpha x_n + m \). Unfortunately neither using more lag terms of the time series nor more neurons improves the situation. Be that as it may, we are interested in analyzing the performance of the adaptive algorithm.

In Fig. 4 the dotted line gives a typical trajectory of one of the 24 parameters \((w_{2,2})\) of the actual algorithm (1000 iterations with \( \mu = .01, \alpha = .5 \)) and a random starting point. The solid line is the computed via Monte Carlo ODE solution \((t_0 = 0, t_f = 10, \Delta = .1)\) starting from the same random initial condition.
4 Discussion

This paper has studied algorithms of the general form described in Eqs. [2 - 3]. The intuitive idea of how to analyze the small $\mu$ asymptotics of this system is clear. One supposes that you can consider the “frozen” version of Eq. [3] i.e. one sets $W_k = w$. The frozen system will hopefully have a unique stationary distribution $\pi_w$ for every possible value of $w$. One then would suppose that this stationary distribution can be used in a standard averaging analysis of Eq. [2].

In this paper we have tried to show that this intuitive notion for how this system behaves can under certain conditions be true. If these conditions are satisfied, then many more analysis and simulation options would be available to the algorithm user and developer. For example, as done in the examples, the user can freeze the system and average via Monte Carlo to get an idea of the limiting differential equation behavior.
and the associated stable points of the algorithm and the ODE. It can also be shown that the algorithm designer may investigate suspected stable points by analytically or computationally solving for the stationary distribution of the frozen state process at those points and being able to prove rigorously that these points are stable.

5 Appendix

Let \((S,d)\) be a complete separable metric space (with metric \(d\)) and let \(\mathcal{M}(S)\) be the space of finite positive measures on \(S\). \(\bar{C}(S)\) is defined as the space of all bounded real valued continuous functions on the metric space \(S\). Consider a sequence of measures \(\{\phi_n\}\) in \(\mathcal{M}(S)\). This sequence *converges weakly* to another measure \(\phi\) in \(\mathcal{M}\) if for all \(f \in \bar{C}(S)\) we have

\[
\lim_n \int_S f d\phi_n = \int_S f d\phi.
\]

When \(S\) is \(\mathbb{R}^d\), weak convergence corresponds exactly to convergence in distribution. There exists a metric (called the Prohorov metric) for \(\mathcal{M}(S)\) under which weak convergence is equivalent to convergence under the metric.

Let \(\mathcal{L}(S)\) be the space of measures on \([0, \infty) \times S\) such that for every \(\mu \in \mathcal{L}(S)\), \(\mu([0, t] \times S) < \infty\) for each \(t \geq 0\). (We need to define \(\mathcal{L}(S)\) since our sequence of occupation measures \(\{\Gamma_\mu\}\) lies in this space.) For \(\mu \in \mathcal{L}(S)\), let \(\mu^t\) denote the restriction of \(\mu\) 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

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

Convergence under this metric in \(\mathcal{L}(S)\) is defined as weak convergence for this space.

Let \(D_S[0, \infty)\) be the space of right continuous functions with left limits mapping the interval \([0, \infty)\) into \(S\). We let \(C_S[0, \infty)\) denote the subspace of continuous functions. We may define a metric on \(D_S[0, \infty)\) under which it is complete and separable as long
as $S$ is complete and separable. Convergence under this metric (which we are not going to give explicitly) corresponds to uniform convergence in the subspace $C_S[0, \infty)$. The topology induced by this metric is called the Skorohod topology. See [4] for further definitions and properties of this space.

To say that a subset of some metric space is relatively compact means merely that for every infinite sequence, there exists a convergent subsequence.

We first state a necessary lemma.

**Lemma 5.1** Let $\{(x_n, \mu_n)\} \subset D_E[0, \infty) \times \mathcal{L}(S)$, and $(x_n, \mu_n) \to (x, \mu)$. Let $h \in \bar{C}(E \times S)$. Suppose further that $h_n$ converges uniformly to $h$ on $E \times S$. Define

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

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

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

**Proof:** The proof is given in [2].

**Proof of the theorem:** Let $M_\mu$ be the martingale defined by

$$M_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (H(W_k, U_{k+1}, \xi_{k+1}) - \hat{H}(W_k, U_k, \xi_k, \eta_k))\mu.$$  

Note that

$$W_\mu(t) = W_\mu(0) + M_\mu(t) + \sum_{k=0}^{[t/\mu]-1} \hat{H}(W_k, U_k, \xi_k, \eta_k)\mu.$$  

By Doob’s inequality

$$E[\sup_{s \leq t} |M_\mu(s)|^2] \leq 4\mu^2 \sum_{k=0}^{[t/\mu]-1} E[(H(W_k, U_{k+1}, \xi_{k+1}) - \hat{H}(W_k, U_k, \xi_k))^2].$$

The right side goes to zero by A2, so $M_\mu \to 0$. Define $V_\mu(t) = \sum_{k=0}^{[t/\mu]-1} \hat{H}(W_k, U_k, \xi_k, \eta_k)\mu$.

A1 implies that $\{V_\mu\}$ is relatively compact. This fact and A3 gives us relative compactness of $\{W_\mu\}$. (Note that any weak limit of $\{W_\mu\}$ has continuous sample paths.)

21
Note that we may write

\[ W_\mu(t) = W_\mu(0) + M_\mu(t) + \int_{E^u \times E^x \times [0,t]} \hat{H}(W_\mu(s-), u, x, y) \Gamma_\mu(du \times dx \times dy \times ds). \]

We may suppose by the Skorohod Representation Theorem that \((W_\mu, M_\mu)\) is converging (along some subsequence and possibly on some other probability space) almost surely to \((W, 0)\) where \(W\) is some random process on \(C_{\mathbb{R}^d}[0, \infty)\). For almost every \(\omega\) in that probability space, \(W_\mu(t, \omega)\) is converging (as a sequence of functions) to \(W(t, \omega) \in C_{\mathbb{R}^d}[0, \infty)\). Also for almost every \(\omega\), \(\Gamma_\mu(\omega) \to \Gamma(\omega)\), where \(\Gamma(du \times dx \times dy \times ds) = \gamma_s(du \times dx \times dy)ds\). Hence by the lemma, \(W(t, \omega)\) must satisfy the deterministic relationship

\[ W(t, \omega) = w_0 + \int_{E^u \times E^x \times [0,t]} \hat{H}(W(s, \omega), u, x, y) \gamma_s(du \times dx \times dy)ds. \]

Since we have this behavior is almost sure in this other probability space, we have the asserted relative compactness and limit point behavior in the original space. \(\Box\)

References


