SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 447–469

THE O.D.E. METHOD FOR CONVERGENCE OF STOCHASTIC APPROXIMATION AND REINFORCEMENT LEARNING*

V. S. BORKAR† AND S. P. MEYN‡

Abstract. It is shown here that stability of the stochastic approximation algorithm is implied by the asymptotic stability of the origin for an associated ODE. This in turn implies convergence of the algorithm. Several specific classes of algorithms are considered as applications. It is found that the results provide (i) a simpler derivation of known results for reinforcement learning algorithms; (ii) a proof for the first time that a class of asynchronous stochastic approximation algorithms are convergent without using any a priori assumption of stability; (iii) a proof for the first time that asynchronous adaptive critic and Q-learning algorithms are convergent for the average cost optimal control problem.

Key words. stochastic approximation, ODE method, stability, asynchronous algorithms, reinforcement learning

AMS subject classifications. 62L20, 93E25, 93E15

PII. S0363012997331639

1. Introduction.

The stochastic approximation algorithm considered in this paper is described by the d-dimensional recursion

$$X(n+1) = X(n) + a(n)\big[h\big(X(n)\big) + M(n+1)\big], \quad n \geq 0, \tag{1.1}$$

where $X(n) = [X_1(n), \ldots, X_d(n)]^T \in \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$, and $\{a(n)\}$ is a sequence of positive numbers. The sequence $\{M(n) : n \geq 0\}$ is uncorrelated with zero mean.

Though more than four decades old, the stochastic approximation algorithm is now of renewed interest due to novel applications to reinforcement learning [20] and as a model of learning by boundedly rational economic agents [19]. Traditional convergence analysis usually shows that the recursion (1.1) will have the desired asymptotic behavior provided that the iterates remain bounded with probability one, or that they visit a prescribed bounded set infinitely often with probability one [3, 14]. Under such stability or recurrence conditions one can then approximate the sequence $\boldsymbol{X} = \{X(n) : n \geq 0\}$ with the solution to the ordinary differential equation (ODE)

$$\dot{x}(t) = h\big(x(t)\big) \tag{1.2}$$

with identical initial conditions $x(0) = X(0)$.

The recurrence assumption is crucial, and in many practical cases this becomes a bottleneck in applying the ODE method. The most successful technique for establishing stochastic stability is the stochastic Lyapunov function approach (see, e.g.,

[*] Received by the editors December 17, 1997; accepted for publication (in revised form) February 22, 1999; published electronically January 11, 2000.
http://www.siam.org/journals/sicon/38-2/33163.html

[†] School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai 400005, India (borkar@tifr.res.in). The research of this author was supported in part by the Department of Science and Technology (Government of India) grant III5(12)/96-ET.

[‡] Department of Electrical and Computer Engineering and the Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (s-meyn@uiuc.edu). The research of this author was supported in part by NSF grant ECS 940372 and JSEP grant N00014-90-J-1270. This research was completed while this author was a visiting scientist at the Indian Institute of Science under a Fulbright Research Fellowship.

[14]). One also has techniques based upon the contractive properties or homogeneity properties of the functions involved (see, e.g., [20] and [12], respectively).

The main contribution of this paper is to add to this collection another general technique for proving stability of the stochastic approximation method. This technique is inspired by the fluid model approach to stability of networks developed in [9, 10], which is itself based upon the multistep drift criterion of [15, 16]. The idea is that the usual stochastic Lyapunov function approach can be difficult to apply due to the fact that time-averaging of the noise may be necessary before a given positive valued function of the state process will decrease towards zero. In general such time-averaging of the noise will require infeasible calculation. In many models, however, it is possible to combine time-averaging with a limiting operation on the magnitude of the initial state, to replace the stochastic system of interest with a simpler deterministic process.

The scaling applied in this paper to approximate the model (1.1) with a deterministic process is similar to the construction of the fluid model of [9, 10]. Suppose that the state is scaled by its initial value to give $\widetilde{X}(n) = X(n)/\max(|X(0)|, 1)$, $n \geq 0$. We then scale time to obtain a continuous function $\phi : \mathbb{R}_+ \to \mathbb{R}^d$ which interpolates the values of $\{\widetilde{X}(n)\}$. At a sequence of times $\{t(j) : j \geq 0\}$ we set $\phi(t(j)) = \widetilde{X}(j)$, and for arbitrary $t \geq 0$, we extend the definition by linear interpolation. The times $\{t(j) : j \geq 0\}$ are defined in terms of the constants $\{a(j)\}$ used in (1.1). For any $r > 0$, the scaled function $h_r : \mathbb{R}^d \to \mathbb{R}^d$ is given by

$$h_r(x) = h(rx)/r, \quad x \in \mathbb{R}^d. \tag{1.3}$$

Then through elementary arguments we find that the stochastic process $\boldsymbol{\phi}$ approximates the solution $\widehat{\phi}$ to the associated ODE

$$\dot{x}(t) = h_r\bigl(x(t)\bigr), \quad t \geq 0, \tag{1.4}$$

with $\widehat{\phi}(0) = \phi(0)$ and $r = \max(|X(0)|, 1)$.

With our attention on stability considerations, we are most interested in the behavior of $\boldsymbol{X}$ when the magnitude of the initial condition $|X(0)|$ is large. Assuming that the limiting function $h_\infty = \lim_{r\to\infty} h_r$ exists, for large initial conditions we find that ϕ is approximated by the solution ϕ^∞ of the limiting ODE

$$\dot{x}(t) = h_\infty\bigl(x(t)\bigr), \tag{1.5}$$

where again we take identical initial conditions $\phi^\infty(0) = \phi(0)$.

Thus, for large initial conditions all three processes are approximately equal,

$$\phi \approx \widehat{\phi} \approx \phi^\infty.$$

Using these observations we find in Theorem 2.1 that the stochastic model (1.1) is stable in a strong sense provided the origin is asymptotically stable for the limiting ODE (1.5). Equation (1.5) is precisely the fluid model of [9, 10].

Thus, the major conclusion of this paper is that the ODE method can be extended to establish both the stability *and* convergence of the stochastic approximation method, as opposed to only the latter. The result [14, Theorem 4.1, p. 115] arrives at a similar conclusion: if the ODE (1.2) possesses a “global” Lyapunov function with bounded partial derivatives, then this will serve as a stochastic Lyapunov function, thereby establishing recurrence of the algorithm. Though similar in flavor, there are significant differences between these results. First, in the present paper we consider a scaled ODE, not the usual ODE (1.2). The former retains only terms with dominant growth and is frequently simpler. Second, while it is possible that the stability of the scaled ODE and the usual one go hand in hand, this does not imply that a Lyapunov function for the latter is easily found. The reinforcement learning algorithms for ergodic-cost optimal control and asynchronous algorithms, both considered as applications of the theory in this paper, are examples where the scaled ODE is conveniently analyzed.

Though the assumptions made in this paper are explicitly motivated by applications to reinforcement learning algorithms for Markov decision processes, this approach is likely to find a broader range of applications.

The paper is organized as follows. The next section presents the main results for the stochastic approximation algorithm with vanishing stepsize or with bounded, nonvanishing stepsize. Section 2 also gives a useful error bound for the constant stepsize case and briefly sketches an extension to asynchronous algorithms, omitting details that can be found in [6]. Section 3 gives examples of algorithms for reinforcement learning of Markov decision processes to which this analysis is applicable. The proofs of the main results are collected together in section 4.

2. Main results. Here we collect together the main general results concerning the stochastic approximation algorithm. Proofs not included here may be found in section 4.

We shall impose the following additional conditions on the functions $\{h_r : r \geq 1\}$ defined in (1.3) and the sequence $\boldsymbol{M} = \{M(n) : n \geq 1\}$ used in (1.1). Some relaxations of assumption (A1) are discussed in section 2.4.

(A1) The function h is Lipschitz, and there exists a function $h_\infty : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\lim_{r\to\infty} h_r(x) = h_\infty(x), \quad x \in \mathbb{R}^d.$$

Furthermore, the origin in $\mathbb{R}^d$ is an asymptotically stable equilibrium for the ODE (1.5).

(A2) The sequence $\{M(n), \mathcal{F}_n : n \geq 1\}$, with $\mathcal{F}_n = \sigma(X(i), M(i), i \leq n)$, is a martingale difference sequence. Moreover, for some $C_0 < \infty$ and any initial condition $X(0) \in \mathbb{R}^d$,

$$\mathsf{E}\big[\|M(n+1)\|^2 \mid \mathcal{F}_n\big] \leq C_0\big(1 + \|X(n)\|^2\big), \quad n \geq 0.$$

The sequence $\{a(n)\}$ is deterministic and is assumed to satisfy one of the following two assumptions. Here TS stands for "tapering stepsize" and BS for "bounded stepsize."

(TS) The sequence $\{a(n)\}$ satisfies $0 < a(n) \leq 1,\ n \geq 0$, and

$$\sum_n a(n) = \infty, \qquad \sum_n a(n)^2 < \infty.$$

(BS) The sequence $\{a(n)\}$ satisfies for some constants $1 > \overline{\alpha} > \underline{\alpha} > 0$,

$$\underline{\alpha} \leq a(n) \leq \overline{\alpha}, \quad n \geq 0.$$

2.1. Stability and convergence. The first result shows that the algorithm is stabilizing for both bounded and tapering step sizes.

THEOREM 2.1. *Assume that* (A1) *and* (A2) *hold. Then we have the following:*

(i) *Under* (TS), *for any initial condition* $X(0) \in \mathbb{R}^d$,

$$\sup_n \|X(n)\| < \infty \qquad \textit{almost surely (a.s.).}$$

(ii) *Under* (BS) *there exist* $\alpha^* > 0$ *and* $C_1 < \infty$ *such that for all* $0 < \overline{\alpha} < \alpha^*$ *and* $X(0) \in \mathbb{R}^d$,

$$\limsup_{n\to\infty} \mathsf{E}\big[\|X(n)\|^2\big] \le C_1. \qquad \square$$

An immediate corollary to Theorem 2.1 is convergence of the algorithm under (TS). The proof is a standard application of the Hirsch lemma (see [11, Theorem 1, p. 339] or [3, 14]), but we give the details below for sake of completeness.

THEOREM 2.2. *Suppose that* (A1), (A2), *and* (TS) *hold and that the ODE* (1.2) *has a unique globally asymptotically stable equilibrium* x^*. *Then* $X(n) \to x^*$ *a.s. as* $n \to \infty$ *for any initial condition* $X(0) \in \mathbb{R}^d$.

Proof. We may suppose that $X(0)$ is deterministic without any loss of generality so that the conclusion of Theorem 2.1 (i) holds that the sample paths of $\boldsymbol{X}$ are bounded with probability one. Fixing such a sample path, we see that $\boldsymbol{X}$ remains in a bounded set H, which may be chosen so that $x^* \in \operatorname{int}(H)$.

The proof depends on an approximation of $\boldsymbol{X}$ with the solution to the primary ODE (1.2). To perform this approximation, first define $t(n) \uparrow \infty$, $T(n) \uparrow \infty$ as follows: Set $t(0) = T(0) = 0$ and for $n \ge 1$, $t(n) = \sum_{i=0}^{n-1} a(i)$. Fix $T > 0$ and define inductively

$$T(n+1) = \min\big\{t(j) : t(j) > T(n) + T\big\}, \quad n \ge 0.$$

Thus $T(n) = t(m(n))$ for some $m(n) \uparrow \infty$ and $T \le T(n+1) - T(n) \le T + 1$ for $n \ge 0$. We then define two functions from $\mathbb{R}_+$ to $\mathbb{R}^d$:

(a) $\{\psi(t), t > 0\}$ is defined by $\psi(t(n)) = X(n)$ with linear interpolation on $[t(n), t(n+1)]$ for each $n \ge 0$.

(b) $\{\widehat{\psi}(t), t > 0\}$ is piecewise continuous, defined so that, for any $j \ge 0$, $\widehat{\psi}$ is the solution to (1.2) for $t \in [T(j), T(j+1))$, with the initial condition $\widehat{\psi}(T(j)) = \psi(T(j))$.

Let $\epsilon > 0$ and let $B(\epsilon)$ denote the open ball centered at x^* of radius ϵ. We may then choose the following:

(i) $0 < \delta < \epsilon$ such that $x(t) \in B(\epsilon)$ for all $t \ge 0$ whenever $x(\,\cdot\,)$ is a solution of (1.2) satisfying $x(0) \in B(\delta)$.

(ii) $T > 0$ so large that for any solution of (1.2) with $x(0) \in H$ we have $x(t) \in B(\delta/2)$ for all $t \ge T$. Hence, $\widehat{\psi}(T(j)-) \in B(\delta/2)$ for all $j \ge 1$.

(iii) An application of the Bellman Gronwall lemma as in Lemma 4.6 below that leads to the limit

$$\big\|\psi(t) - \widehat{\psi}(t)\big\| \to 0 \qquad \text{a.s.,} \quad t \to \infty. \tag{2.1}$$

Hence we may choose $j_0 > 0$ so that we have

$$\big\|\psi\big(T(j)-\big) - \widehat{\psi}\big(T(j)-\big)\big\| \le \delta/2, \quad j \ge j_0.$$

Since $\psi(\,\cdot\,)$ is continuous, we conclude from (ii) and (iii) that $\psi(T(j)) \in B(\delta)$ for $j \geq j_0$. Since $\widehat{\psi}(T(j)) = \psi(T(j))$, it then follows from (i) that $\widehat{\psi}(t) \in B(\epsilon)$ for all $t \geq T(j_0)$. Hence by (2.1),

$$\limsup_{t\to\infty} \|\psi(t) - x^*\| \leq \epsilon \qquad \text{a.s.}$$

This completes the proof since $\epsilon > 0$ was arbitrary. □

We now consider (BS), focusing on the absolute error defined by

$$e(n) := \|X(n) - x^*\|, \qquad n \geq 0. \tag{2.2}$$

Theorem 2.3. *Assume that* (A1), (A2), *and* (BS) *hold, and suppose that* (1.2) *has a globally asymptotically stable equilibrium point x^*.*

Then for any $0 < \alpha \leq \alpha^*$, where α^* is introduced in Theorem 2.1 (ii),

(i) *for any $\epsilon > 0$, there exists $b_1 = b_1(\epsilon) < \infty$ such that*

$$\limsup_{n\to\infty} \mathsf{P}\big(e(n) \geq \epsilon\big) \leq b_1\overline{\alpha};$$

(ii) *if x^* is a globally exponentially asymptotically stable equilibrium for the ODE* (1.2)*, then there exists $b_2 < \infty$ such that for every initial condition $X(0) \in \mathbb{R}^d$,*

$$\limsup_{n\to\infty} \mathsf{E}\big[e(n)^2\big] \leq b_2\overline{\alpha}. \qquad \square$$

2.2. Rate of convergence. A uniform bound on the mean square error $\mathsf{E}[e(n)^2]$ for $n \geq 0$ can be obtained under slightly stronger conditions on $\boldsymbol{M}$ via the theory of ψ-irreducible Markov chains. We find that this error can be bounded from above by a sum of two terms: the first converges to zero as $\alpha \downarrow 0$, while the second decays to zero exponentially as $n \to \infty$.

To illustrate the nature of these bounds, consider the linear recursion

$$X(n+1) = X(n) + \alpha\big[-\big(X(n) - x^*\big) + W(n+1)\big], \quad n \geq 0,$$

where $\{W(n)\}$ is independently and identically distributed (i.i.d.) with mean zero and variance σ^2. This is of the form (1.1) with $h(x) = -(x - x^*)$ and $M(n) = W(n)$. The error $e(n+1)$ defined in (2.2) may be bounded as follows:

$$\begin{aligned}\mathsf{E}\big[e(n+1)^2\big] &\leq \alpha^2\sigma^2 + (1-\alpha)^2\mathsf{E}\big[e(n)^2\big] \\ &\leq \alpha\sigma^2/(2-\alpha) + \exp(-2\alpha n)\mathsf{E}\big[e(0)^2\big], \quad n \geq 0.\end{aligned}$$

For a deterministic initial condition $X(0) = x$ and any $\epsilon > 0$, we thus arrive at the formal bound,

$$\mathsf{E}[e(n)^2 \mid X(0) = x] \leq B_1(\alpha) + B_2\big(\|x\|^2 + 1\big)\exp\big(-\epsilon_0(\alpha)n\big), \tag{2.3}$$

where B_1, B_2, and ϵ_0 are positive-valued functions of α. The bound (2.3) is of the form that we seek: the first term on the right-hand side (r.h.s.) decays to zero with α, while the second decays exponentially to zero with n. However, the rate of convergence for the second term becomes vanishingly small as $\alpha \downarrow 0$. Hence to maintain a small probability of error the variable α should be neither too small nor too large. This recalls the well-known trade-off between mean and variance that must be made in the application of stochastic approximation algorithms.

A bound of this form carries over to the nonlinear model under some additional conditions. For convenience, we take a Markov model of the form

$$X(n+1) = X(n) + \alpha\bigl[h\bigl(X(n)\bigr) + m\bigl(X(n), W(n+1)\bigr)\bigr], \tag{2.4}$$

where again $\{W(n)\}$ is i.i.d. and also independent of the initial condition $X(0)$. We assume that the functions $h : \mathbb{R}^d \to \mathbb{R}^d$ and $m : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d$ are smooth (C^1) and that assumptions (A1) and (A2) continue to hold. The recursion (2.4) then describes a Feller–Markov chain with stationary transition kernel to be denoted by P.

Let $V\colon \mathbb{R}^d \to [1,\infty)$ be given. The Markov chain $\boldsymbol{X}$ with transition function P is called *V-uniformly ergodic* if there is a unique invariant probability π, an $R < \infty$, and $\rho < 1$ such that for any function g satisfying $|g(x)| \le V(x)$,

$$\bigl|\mathsf{E}\bigl[g\bigl(X(n)\bigr) \mid X(0) = x\bigr] - \mathsf{E}_\pi\bigl[g\bigl(X(n)\bigr)\bigr]\bigr| \le RV(x)\rho^n, \qquad x \in \mathbb{R}^d, \quad n \ge 0, \tag{2.5}$$

where $\mathsf{E}_\pi[g(X(n))] = \int g(x)\,\pi(dx)$, $n \ge 0$.

The following result establishes bounds of the form (2.3) using V-ergodicity of the model. Assumptions (2.6) and (2.7) below are required to establish ψ-irreducibility of the model in Lemma 4.10.

There exists a $w^* \in \mathbb{R}^q$ with $m(x^*, w^*) = 0$, and for a continuous function $p : \mathbb{R}^q \to [0,1]$ with $p(w^*) > 0$,

$$\mathsf{P}\bigl(W(1) \in A\bigr) \ge \int_A p(z)dz, \quad A \in \mathcal{B}(\mathbb{R}^q). \tag{2.6}$$

The pair of matrices (F, G) is controllable with

$$F = \frac{d}{dx}h(x^*) + \frac{\partial}{\partial x}m(x^*, w^*) \quad \text{and} \quad G = \frac{\partial}{\partial w}m(x^*, w^*). \tag{2.7}$$

THEOREM 2.4. *Suppose that* (A1), (A2), (2.6), *and* (2.7) *hold for the Markov model* (2.4) *with $0 < \alpha \le \alpha^*$. Then the Markov chain $\boldsymbol{X}$ is V-uniformly ergodic, with $V(x) = \|x\|^2 + 1$, and we have the following bounds:*

(i) *There exist positive-valued functions A_1 and ϵ_0 of α and a constant A_2 independent of α, such that*

$$\mathsf{P}\bigl\{e(n) \ge \epsilon \mid X(0) = x\bigr\} \le A_1(\alpha) + A_2\bigl(\|x\|^2 + 1\bigr)\exp\bigl(-\epsilon_0(\alpha)n\bigr).$$

The functions satisfy $A_1(\alpha) \to 0$, $\epsilon_0(\alpha) \to 0$ as $\alpha \downarrow 0$.

(ii) *If in addition the ODE* (1.2) *is exponentially asymptotically stable, then the stronger bound* (2.3) *holds, where again $B_1(\alpha) \to 0$, $\epsilon_0(\alpha) \to 0$ as $\alpha \downarrow 0$, and B_2 is independent of α.*

Proof. The V-uniform ergodicity is established in Lemma 4.10.

From Theorem 2.3 (i) we have, when $X(0) \sim \pi$,

$$\mathsf{P}_\pi\bigl(e(n) \ge \epsilon\bigr) = \mathsf{P}_\pi\bigl(e(0) \ge \epsilon\bigr) \le b_1\overline{\alpha},$$

and hence from V-uniform ergodicity,

$$\begin{aligned}\mathsf{P}\bigl(e(n) \ge \epsilon \mid X(0) = x\bigr) &\le \mathsf{P}_\pi\bigl(e(n) \ge \epsilon\bigr) + \bigl|\mathsf{P}\bigl(e(n) \ge \epsilon \mid X(0) = x\bigr) - \mathsf{P}_\pi\bigl(e(n) \ge \epsilon\bigr)\bigr| \\ &\le b_1\alpha + RV(x)\rho^n, \quad n \ge 0.\end{aligned}$$

This and the definition of V establishes (i). The proof of (ii) is similar.

The fact that $\rho = \rho_\alpha \to 1$ as $\alpha \downarrow 0$ is discussed in section 4.3. □

2.3. The asynchronous case. The conclusions above also extend to the model of asynchronous stochastic approximation analyzed in [6]. We now assume that each component of $X(n)$ is updated by a separate processor. We postulate a set-valued process $\{Y(n)\}$ taking values in the set of subsets of $\{1, 2, \ldots, d\}$, with the interpretation: $Y(n) = \{\text{indices of the components updated at time } n\}$. For $n \geq 0$, $1 \leq i \leq d$, define

$$\nu(i, n) = \sum_{m=0}^{n} I\bigl\{i \in Y(m)\bigr\},$$

the number of updates executed by the ith processor up to time n. A key assumption is that there exists a deterministic $\Delta > 0$ such that for all i,

$$\liminf_{n\to\infty} \frac{\nu(i, n)}{n} \geq \Delta \qquad \text{a.s.}$$

This ensures that all components are updated comparably often. Furthermore, if

$$N(n, x) = \min\left\{m > n : \sum_{k=n+1}^{m} a(n) > x\right\}$$

for $x > 0$, the limit

$$\mathrm{lim}_{n\to\infty} \frac{\sum_{k=v(i,n)}^{v(i,N(n,x))} a(k)}{\sum_{k=v(j,n)}^{v(j,N(n,x))} a(k)}$$

exists a.s. for all i, j.

At time n, the kth processor has available the following data:

(i) Processor (k) is given $\nu(k, n)$, but it may not have n, the "global clock."

(ii) There are interprocessor communication delays $\tau_{kj}(n), 1 \leq k, j \leq d, n \geq 0$, so that at time n, processor (k) may use the data $X_j(m)$ only for $m \leq n - \tau_{kj}(n)$.

We assume that $\tau_{kk}(n) = 0$ for all n and that $\{\tau_{kj}(n)\}$ have a common upper bound $\overline{\tau} < \infty$ ([6] considers a slightly more general situation).

To relate the present work to [6], we recall that the "centralized" algorithm of [6] is

$$X(n+1) = X(n) + a(n) f\bigl(X(n), W(n+1)\bigr),$$

where $\{W(n)\}$ are i.i.d. and $\{f(\cdot, y)\}$ are uniformly Lipschitz. Thus $F(x) := \mathsf{E}[f(x, W(1))]$ is Lipschitz. The correspondence with the present set up is obtained by setting $h(x) = F(x)$ and

$$M(n+1) = f\bigl(X(n), W(n+1)\bigr) - F\bigl(X(n)\bigr)$$

for $n \geq 0$. The asynchronous version then is

$$\begin{aligned} X_i(n+1) = X_i(n) + a\bigl(\nu(i,n)\bigr) f\bigl(X_1(n - \tau_{i1}(n)\bigr), X_2\bigl(n - \tau_{i2}(n)\bigr), \\ \ldots, X_d\bigl(n - \tau_{id}(n)\bigr), W(n+1)) I\bigl\{i \in Y(n)\bigr\}, \quad n \geq 0, \end{aligned} \tag{2.8}$$

for $1 \leq i \leq d$. Note that this can be executed by the ith processor without any knowledge of the global clock which, in fact, can be a complete artifice as long as causal relationships are respected.

The analysis presented in [6] depends upon the following additional conditions on $\{a(n)\}$:

(i) $a(n+1) \leq a(n)$ eventually;

(ii) for $x \in (0,1)$, $\sup_n a([xn])/a(n) < \infty$;

(iii) for $x \in (0,1)$,

$$\left(\sum_{i=0}^{[xn]} a(i)\right) \Bigg/ \left(\sum_{i=0}^{n} a(i)\right) \to 1,$$

where $[\,\cdot\,]$ stands for "the integer part of $(\,\cdot\,)$."

A fourth condition is imposed in [6], but this becomes irrelevant when the delays are bounded. Examples of $\{a(n)\}$ satisfying (i)–(iii) are $a(n) = 1/(n+1)$ or $1/(1 + n\log(n+1))$.

As a first simplifying step, it is observed in [6] that $\{Y(n)\}$ may be assumed to be singletons without any loss of generality. We shall do likewise. What this entails is simply unfolding a single update at time n into $|Y(n)|$ separate updates, each involving a single component. This blows up the delays at most d-fold, which does not affect the analysis in any way.

The main result of [6] is the analog of our Theorem 2.2 *given* that the conclusions of our Theorem 2.1 hold. In other words, stability implies convergence. Under (A1) and (A2), our arguments above can be easily adapted to show that the conclusions of Theorem 2.2 also hold for the asynchronous case. One argues exactly as above and in [6] to conclude that the suitably interpolated and rescaled trajectory of the algorithm tracks an appropriate ODE. The only difference is a scalar factor $1/d$ multiplying the r.h.s. of the ODE (i.e., $\dot{x}(t) = (1/d)h(x(t))$). This factor, which reflects the asynchronous sampling, amounts to a time-scaling that does not affect the qualitative behavior of the ODE.

THEOREM 2.5. *Under the conditions of Theorem* 2.2 *and the above hypotheses on* $\{a(n)\}$, $\{Y(n)\}$, *and* $\{\tau_{ij}(n)\}$, *the asynchronous iterates given by* (3.7) *remain a.s. bounded and (therefore) converge to* x^* *a.s.* □

2.4. Further extensions.

Although satisfied in all of the applications treated in section 3, in some other models assumption (A1) that $h_r \to h_\infty$ pointwise may be violated. If this convergence does not hold, then we may abandon the fluid model and replace (A1) by

(A1′) The function h is Lipschitz, and there exists $T > 0$, $R > 0$ such that

$$|\widehat{\phi}(t)| \leq \frac{1}{2}, \quad t \geq T,$$

for any solution to (1.4) with $r \geq R$ and with initial condition satisfying $|\widehat{\phi}(0)| = 1$.

Under the Lipschitz condition on h, at worst we may find that the pointwise limits of $\{h_r : r \geq 1\}$ will form a family Λ of Lipschitz functions on $\mathbb{R}^d$. That is, $h_\infty \in \Lambda$ if and only if there exists a sequence $\{r_i\} \uparrow \infty$ such that

$$h_{r_i}(x) \to h_\infty(x), \quad i \to \infty,$$

where the convergence is uniform for x in compact subsets of $\mathbb{R}^d$. Under (A1′) we then find, using the same arguments as in the proof of Lemma 4.1, that the family Λ is uniformly stable.

LEMMA 2.6. *Under* (A1′) *the family of ODEs defined via* Λ *is uniformly exponentially asymptotically stable in the following sense. For some* $b < \infty$, $\delta > 0$, *and any solution* ϕ^∞ *to the ODE* (1.5) with $h_\infty \in \Lambda$,

$$|\phi^\infty(t)| \le be^{-\delta t}|\phi^\infty(0)|, \quad t \ge 0. \qquad \square$$

Using this lemma the development of section 4 goes through with virtually no changes, and hence Theorems 2.1–2.5 are valid with (A1) replaced by (A1′).

Another extension is to broaden the class of scalings. Consider a nonlinear scaling defined by a function $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $g(r)/r \to \infty$ as $r \to \infty$, and suppose that $h_r(\,\cdot\,)$ redefined as $h_r(x) = h(rx)/g(r)$ satisfies

$$h_r(x) \to h_\infty(x) \text{ uniformly on compacts as } r \to \infty.$$

Then, assuming that the a.s. boundedness of rescaled iterates can be separately established, a completely analogous development of the stochastic algorithm is possible. An example would be a "stochastic gradient" scheme, where $h(\,\cdot\,)$ is the gradient of an even degree polynomial, with degree, say, $2n$. Then $g(r) = r^{2n-1}$ will do. We do not pursue this further because the reinforcement learning algorithms we consider below do conform to the case $g(r) = r$.

3. Reinforcement learning.

As both an illustration of the theory and an important application in its own right, in this section we analyze reinforcement learning algorithms for Markov decision processes. The reader is referred to [4] for a general background of the subject and to other references listed below for further details.

3.1. Markov decision processes.

We consider a Markov decision process $\boldsymbol{\Phi} = \{\Phi(t) : t \in \mathbb{Z}\}$ taking values in a finite state space $S = \{1, 2, \ldots, s\}$ and controlled by a control sequence $\boldsymbol{Z} = \{Z(t) : t \in \mathbb{Z}\}$ taking values in a finite action space $A = \{a_0, \ldots, a_r\}$. We assume that the control sequence is *admissible* in the sense that $Z(n) \in \sigma\{\Phi(t) : t \le n\}$ for each n. We are most interested in stationary policies of the form $Z(t) = w(\Phi(t))$, where the *feedback law* w is a function $w\colon S \to A$. The controlled transition probabilities are given by $p(i, j, a)$ for $i, j \in S, a \in A$.

Let $c : S \times A \to R$ be the one-step cost function, and consider first the infinite horizon discounted cost control problem of minimizing over all admissible $\boldsymbol{Z}$ the total discounted cost

$$J(i, \boldsymbol{Z}) = \mathsf{E}\left[\sum_{t=0}^{\infty} \beta^t c\big(\Phi(t), Z(t)\big) \mid \Phi(0) = i\right],$$

where $\beta \in (0, 1)$ is the discount factor. The minimal value function is defined as

$$V(i) = \min J(i, \boldsymbol{Z}),$$

where the minimum is over all admissible control sequences $\boldsymbol{Z}$. The function V satisfies the dynamic programming equation

$$V(i) = \min_a \left[c(i, a) + \beta \sum_j p(i, j, a)V(j)\right], \qquad i \in S,$$

and the optimal control minimizing J is given as the stationary policy defined through the feedback law w^* given as any solution to

$$w^*(i) := \arg\min_a \left[c(i, a) + \beta \sum_j p(i, j, a)V(j)\right], \quad i \in S.$$

The value iteration algorithm is an iterative procedure to compute the minimal value function. Given an initial function $V_0 \colon S \to \mathbb{R}_+$ one obtains a sequence of functions $\{V_n\}$ through the recursion

$$V_{n+1}(i) = \min_a \Bigl[c(i,a) + \beta \sum_j p(i,j,a)V_n(j)\Bigr], \quad i \in S, \quad n \geq 0. \tag{3.1}$$

This recursion is convergent for any initialization $V_0 \geq 0$. If we define *Q-values* via

$$Q(i,a) = c(i,a) + \beta \sum_j p(i,j,a)V(j), \qquad i \in S, \quad a \in A,$$

then $V(i) = \min_a Q(i,a)$ and the matrix Q satisfies

$$Q(i,a) = c(i,a) + \beta \sum_j p(i,j,a) \min_b Q(j,b), \qquad i \in S, \quad a \in A.$$

The matrix Q can also be computed using the equivalent formulation of value iteration,

$$Q_{n+1}(i,a) = c(i,a) + \beta \sum_j p(i,j,a) \min_b Q_n(j,b), \qquad i \in S, \quad a \in A, \quad n \geq 0, \tag{3.2}$$

where $Q_0 \geq 0$ is arbitrary.

The value iteration algorithm is initialized with a function $V_0 \colon S \to \mathbb{R}_+$. In contrast, the *policy iteration algorithm* is initialized with a feedback law w^0 and generates a sequence of feedback laws $\{w^n : n \geq 0\}$. At the nth stage of the algorithm a feedback law w^n is given and the value function for the resulting control sequence $\boldsymbol{Z}^n = \{w^n(\Phi(0)), w^n(\Phi(1)), w^n(\Phi(2)), \dots\}$ is computed to give

$$J_n(i) = J\bigl(i, \boldsymbol{Z}^n\bigr), \qquad i \in S.$$

Interpreted as a column vector in $\mathbb{R}^s$, the vector J_n satisfies the equation

$$\bigl(I - \beta P_n\bigr) J_n = c_n, \tag{3.3}$$

where the $s \times s$ matrix P_n is defined by $P_n(i,j) = p(i,j,w^n(i))$, $i,j \in S$, and the column vector c_n is given by $c_n(i) = c(i, w^n(i))$, $i \in S$. Equation (3.3) can be solved for fixed n by the "fixed-policy" version of value iteration given by

$$J_n(i+1) = \beta P_n J_n(i) + c_n, \quad i \geq 0, \tag{3.4}$$

where $J_n(0) \in \mathbb{R}^s$ is given as an initial condition. Then $J_n(i) \to J_n$, the solution to (3.3), at a geometric rate as $i \to \infty$.

Given J_n, the next feedback law w^{n+1} is then computed via

$$w^{n+1}(i) = \arg\min_a \Bigl[c(i,a) + \beta \sum_j p(i,j,a) J_n(j)\Bigr], \quad i \in S. \tag{3.5}$$

Each step of the policy iteration algorithm is computationally intensive for large state spaces since the computation of J_n requires the inversion of the $s \times s$ matrix $I - \beta P_n$.

In the average cost optimization problem one seeks to minimize over all admissible $\boldsymbol{Z}$,

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}\bigl[c\bigl(\Phi(t), Z(t)\bigr)\bigr]. \tag{3.6}$$

The policy iteration and value iteration algorithms to solve this optimization problem remain unchanged with three exceptions. One is that the constant β must be set equal to unity in (3.1) and (3.5). Second, in the policy iteration algorithm the value function J_n is replaced by a solution J_n to Poisson's equation

$$\sum p\big(i,j,w^n(i)\big)J_n(j) = J_n(i) - c\big(i,w^n(i)\big) + \eta_n, \quad i \in S,$$

where η_n is the steady state cost under the policy w^n. The computation of J_n and η_n again involves matrix inversions via

$$\pi_n\big(I - P_n + ee'\big) = e', \qquad \eta_n = \pi_n c_n, \qquad \big(I - P_n + ee'\big)J_n = c_n,$$

where $e \in \mathbb{R}^s$ is the column vector consisting of all ones and the row vector π_n is the invariant probability for P_n. The introduction of the outer product ensures that the matrix $(I - P_n + ee')$ is invertible, provided that the invariant probability π_n is unique.

Lastly, the value iteration algorithm is replaced by the "relative value iteration," where a common scalar offset is subtracted from all components of the iterates at each iteration (likewise for the Q-value iteration). The choice of this offset term is not unique. We shall be considering one particular choice, though others can be handled similarly (see [1]).

3.2. *Q*-learning. If the matrix Q defined in (3.2) can be computed via value iteration or some other scheme, then the optimal control is found through a simple minimization. If transition probabilities are unknown so that value iteration is not directly applicable, one may apply a stochastic approximation variant known as the *Q-learning algorithm* of Watkins [1, 20, 21]. This is defined through the recursion

$$Q_{n+1}(i,a) = Q_n(i,a) + a(n)\Big[\beta \min_b Q_n(\Psi_{n+1}(i,a),b) + c(i,a) - Q_n(i,a)\Big],$$

$i \in S, a \in A$, where $\Psi_{n+1}(i,a)$ is an independently simulated S-valued random variable with law $p(i,\cdot,a)$.

Making the appropriate correspondences with our set up, we have $X(n) = Q_n$ and $h(Q) = [h_{ia}(Q)]_{i,a}$ with

$$h_{ia}(Q) = \beta \sum_j p(i,j,a) \min_b Q(j,b) + c(i,a) - Q(i,a), \qquad i \in S, \quad a \in A.$$

The martingale is given by $M(n+1) = [M_{ia}(n+1)]_{i,a}$ with

$$M_{ia}(n+1) \\ = \beta\left(\min_b Q_n(\Psi_{n+1}(i,a),b) - \sum_j p(i,j,a)\Big(\min_b Q_n(j,b)\Big)\right), \quad i \in S, \quad a \in A.$$

Define $F(Q) = [F_{ia}(Q)]_{i,a}$ by

$$F_{ia}(Q) = \beta \sum_j p(i,j,a) \min_b Q(j,b) \ + \ c(i,a).$$

Then $h(Q) = F(Q) - Q$ and the associated ODE is

$$\dot{Q} = F(Q) - Q := h(Q). \tag{3.7}$$

The map $F : \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is a contraction with respect to the max norm $\|\cdot\|_\infty$. The global asymptotic stability of its unique equilibrium point is a special case of the results of [8]. This $h(\,\cdot\,)$ fits the framework of our analysis, with the (i,a)th component of $h_\infty(Q)$ given by

$$\beta \sum_j p(i,j,a) \min_b Q(j,b) - Q(i,a), \quad i \in S, \quad a \in A.$$

This also is of the form $h_\infty(Q) = F_\infty(Q) - Q$ where $F_\infty(\,\cdot\,)$ is an $\|\cdot\|_\infty$- contraction, and thus the asymptotic stability of the unique equilibrium point of the corresponding ODE is guaranteed (see [8]). We conclude that assumptions (A1) and (A2) hold, and hence Theorems 2.1–2.4 also hold for the Q-learning model.

3.3. Adaptive critic algorithm. Next we shall consider the *adaptive critic algorithm*, which may be considered as the reinforcement learning analog of policy iteration (see [2, 13] for a discussion). There are several variants of this, one of which, taken from [13], is as follows. For $i \in S$, we define

$$V_{n+1}(i) = V_n(i) + b(n)\big[c\big(i, \psi_n(i)\big) + \beta V_n\big(\Psi_n\big(i, \psi_n(i)\big)\big) - V_n(i)\big], \tag{3.8}$$

from which the policies are updated according to

$$\begin{aligned}(3.9)\quad &\widehat{w}_{n+1}(i)\\ &= \Gamma\left\{\widehat{w}_n(i) + a(n)\sum_{\ell=1}^{r}\Big(\big[c(i,a_0) + \beta V_n\big(\eta_n(i,a_0)\big)\big] - [c(i,a_\ell) + \beta V_n(\eta_n(i,a_\ell))]e_\ell\Big)\right\}.\end{aligned}$$

Here $\{V_n\}$ are s-vectors and for each $i, \{\widehat{w}_n(i)\}$ are r-vectors lying in the simplex $\{x \in \mathbb{R}^r \mid x = [x_1,\ldots,x_r], x_i \geq 0, \sum_i x_i \leq 1\}$. $\Gamma(\,\cdot\,)$ is the projection onto this simplex. The sequences $\{a(n)\}, \{b(n)\}$ satisfy

$$\sum_n a(n) = \sum_n b(n) = \infty, \qquad \sum_n \big(a(n)^2 + b(n)^2\big) < \infty, \qquad a(n) = o\big(b(n)\big).$$

The rest of the notation is as follows. For $1 \leq \ell \leq r, e_\ell$ is the unit r-vector in the ℓth coordinate direction. For each i, n, $w_n(i) = w_n(i,\cdot)$ is a probability vector on A defined by the following. For $\widehat{w}_n(i) = [\widehat{w}_n(i,1),\ldots,\widehat{w}_n(i,r)]$,

$$w_n(i,a_\ell) = \begin{cases} \widehat{w}_n(i,\ell) & \text{for } \ell \neq 0,\\ 1 - \displaystyle\sum_{j\neq 0} \widehat{w}_n(i,j) & \text{for } \ell = 0.\end{cases}$$

Given $w_n(i)$, $\psi_n(i)$ is an A-valued random variable independently simulated with law $w_n(i)$. Likewise, $\Psi_n(i,\psi_n(i))$ are S-valued random variables which are independently simulated (given $\psi_n(i)$) with law $p(i,\cdot,\psi_n(i))$ and $\{\eta_n(i,a_\ell)\}$ are S-valued random variables independently simulated with law $p(i,\cdot,a_\ell)$, respectively.

To see why this is based on policy iteration, recall that policy iteration alternates between two steps. One step solves the linear system of (3.3) to compute the fixed-policy value function corresponding to the current policy. We have seen that solving (3.3) can be accomplished by performing the fixed-policy version of value iteration given in (3.4). The first step (3.8) in the above iteration is indeed the "learning" or "simulation-based stochastic approximation" analog of this fixed-policy value iteration. The second step in policy iteration updates the current policy by performing an appropriate minimization. The second iteration (3.9) is a particular search algorithm for computing this minimum over the simplex of probability measures on A. This search algorithm is by no means unique; the paper [13] gives two alternative schemes. However, the first iteration (3.8) is common to all.

The different choices of stepsize schedules for the two iterations (3.8) and (3.9) induces the "two time-scale" effect discussed in [5]. The first iteration sees the policy computed by the second as nearly static, thus justifying viewing it as a fixed-policy iteration. In turn, the second sees the first as almost equilibrated, justifying the search scheme for minimization over A. See [13] for details.

The boundedness of $\{\widehat{w}_n\}$ is guaranteed by the projection $\Gamma(\,\cdot\,)$. For $\{V_n\}$, the fact that $b(n) = o(a(n))$ allows one to treat $\widehat{w}_n(i)$ as constant, say, $\overline{w}(i)$; see, e.g., [13]. The appropriate ODE then turns out to be

$$\dot{v} = G(v) - v := h(v), \tag{3.10}$$

where $G : \mathbb{R}^s \to \mathbb{R}^s$ is defined by

$$G_i(x) = \sum_\ell \overline{w}(i, a_\ell) \left[\beta \sum_j p(i, j, a_\ell) x_j + c(i, a_\ell) \right] - x_i, \quad i \in S.$$

Once again, $G(\,\cdot\,)$ is an $\|\cdot\|_\infty$-contraction and it follows from the results of [8] that (3.10) is globally asymptotically stable. The limiting function $h_\infty(x)$ is again of the form $h_\infty(x) = G_\infty(x) - x$ with $G_\infty(x)$ defined so that its ith component is

$$\sum_\ell \overline{w}(i, a_\ell) \left[\beta \sum_j p(i, j, a_\ell) x_j \right] - x_i.$$

We see that G_∞ is also a $\|\cdot\|_\infty$-contraction and the global asymptotic stability of the origin for the corresponding limiting ODE follows as before from the results of [8].

3.4. Average cost optimal control. For the average cost control problem, we impose the additional restriction that the chain $\mathbf{\Phi}$ has a *unique* invariant probability measure under any stationary policy so that the steady state cost (3.6) is independent of the initial condition.

For the average cost optimal control problem, the Q-learning algorithm is given by the recursion

$$Q_{n+1}(i, a) = Q_n(i, a) + a(n) \Big(\min_b Q_n(\Psi_n(i, a), b) + c(i, a) - Q_n(i, a) - Q_n(i_0, a_0) \Big),$$

where $i_0 \in S$, $a_0 \in A$ are fixed a priori. The appropriate ODE now is (3.7) with $F(\,\cdot\,)$ redefined as $F_{ia}(Q) = \sum_j p(i, j, a) \min_b Q(j, b) + c(i, a) - Q(i, a) - Q(i_0, a_0)$. The global asymptotic stability for the unique equilibrium point for this ODE has been established in [1]. Once again this fits our framework with $h_\infty(x) = F_\infty(x) - x$ for F_∞ defined the same way as F, except for the terms $c(\cdot, \cdot)$ which are dropped. We conclude that (A1) and (A2) are satisfied for this version of the Q-learning algorithm.

Another variant of Q-learning for average cost, based on a "stochastic shortest path" formulation, is presented in [1]. This also can be handled similarly.

In [13], three variants of the adaptive critic algorithm for the average cost problem are discussed, differing only in the $\{\widehat{w}_n\}$ iteration. The iteration for $\{V_n\}$ is common to all and is given by

$$V_{n+1}(i) = V_n(i) + b(n)\big[c\big(i, \psi_n(i)\big) + V_n\big(\Psi_n\big(i, \psi_n, (i)\big)\big) - V_n(i) - V_n(i_0)\big], \quad i \in S,$$

where $i_0 \in S$ is a fixed state prescribed beforehand. This leads to the ODE (3.10) with G redefined as

$$G_i(x) = \sum_\ell \overline{w}(i, a_\ell)\left(\sum_j p(i, j, a_\ell)x_j + c(i, a_\ell)\right) - x_i - x_{i_0}, \quad i \in S.$$

The global asymptotic stability of the unique equilibrium point of this ODE has been established in [7]. Once more, this fits our framework with $h_\infty(x) = G_\infty(x) - x$ for G_∞ defined just like G, but without the $c(\cdot,\cdot)$ terms.

Asynchronous versions of all the above can be written down along the lines of (3.7). Then by Theorem 2.5, they have bounded iterates a.s. The important point to note here is that to date, a.s. boundedness for Q-learning and adaptive critic is proved by other methods for centralized algorithms [1, 12, 20]. For asynchronous algorithms, it is proved for discounted cost only [1, 13, 20] or by introducing a projection to enforce stability [14].

4. Derivations.

Here we provide proofs for the main results given in section 2. Throughout this section we assume that (A1) and (A2) hold.

4.1. Stability.

The functions $\{h_r, r \geq 1\}$ and the limiting function h_∞ are Lipschitz with the same Lipschitz constant as h under (A1). It follows from Ascoli's theorem that the convergence $h_r \to h_\infty$ is uniform on compact subsets of $\mathbb{R}^d$. This observation is the basis of the following lemma.

LEMMA 4.1. *Under* (A1)*, the ODE* (1.5) *is globally exponentially asymptotically stable.*

Proof. The function h_∞ satisfies

$$h_\infty(cx) = ch_\infty(x), \qquad c > 0, \quad x \in \mathbb{R}^d.$$

Hence the origin $\theta \in \mathbb{R}^d$ is an equilibrium for (1.5), i.e., $h_\infty(\theta) = \theta$. Let $B(\epsilon)$ be the closed ball of radius ϵ centered at θ with ϵ chosen so that $x(t) \to \theta$ as $t \to \infty$ uniformly for initial conditions in $B(\epsilon)$. Thus there exists a $T > 0$ such that $\|x(T)\| \leq \epsilon/2$ whenever $\|x(0)\| \leq \epsilon$. For an arbitrary solution $x(\,\cdot\,)$ of (1.5), $y(\,\cdot\,) = \epsilon x(\,\cdot\,)/\|x(0)\|$ is another, with $\|y(0)\| = \epsilon$. Hence $\|y(T)\| < \epsilon/2$, implying $\|x(T)\| \leq \frac{1}{2}\|x(0)\|$. The global exponential asymptotic stability follows. ∎

With the scaling parameter r given by $r(j) = \max(1, \|X(m(j))\|)$, $j \geq 0$, we define three piecewise continuous functions from $\mathbb{R}_+$ to $\mathbb{R}^d$ as in the introduction:

(a) $\{\phi(t) : t \geq 0\}$ is an interpolated version of $\boldsymbol{X}$ defined as follows. For each $j \geq 0$, define a function ϕ_j on the interval $[T(j), T(j+1)]$ by

$$\phi_j\big(t(n)\big) = X(n)/r(j), \quad m(j) \leq n \leq m(j+1),$$

with $\phi_j(\,\cdot\,)$ defined by linear interpolation on the remainder of $[T(j), T(j+1)]$ to form a piecewise linear function.

We then define ϕ to be the piecewise continuous function

$$\phi(t) = \phi_j(t), \qquad t \in \big[T(j), T(j+1)\big), \quad j \geq 0.$$

(b) $\{\widehat{\phi}(t) : t \geq 0\}$ is continuous on each interval $[T(j), T(j+1))$, and on this interval it is the solution to the ODE

$$\dot{x}(t) = h_{r(j)}\bigl(x(t)\bigr), \tag{4.1}$$

with initial condition $\widehat{\phi}(T(j)) = \phi(T(j))$, $j \geq 0$.

(c) $\{\phi^\infty(t) : t \geq 0\}$ is also continuous on each interval $[T(j), T(j+1))$, and on this interval it is the solution to the "fluid model" (1.5) with the same initial condition

$$\phi^\infty\bigl(T(j)\bigr) = \widehat{\phi}\bigl(T(j)\bigr) = \phi\bigl(T(j)\bigr) \quad j \geq 0.$$

Boundedness of $\widehat{\phi}(\,\cdot\,)$ and $\phi^\infty(\,\cdot\,)$ is crucial in deriving useful approximations.

LEMMA 4.2. *Under* (A1) *and* (A2) *and either* (TS) *or* (BS), *there exists* $\bar{C} < \infty$ *such that for any initial condition* $X(0) \in \mathbb{R}^d$

$$\widehat{\phi}(t) \leq \bar{C} \quad \textit{and} \quad \phi^\infty(t) \leq \bar{C}, \qquad t \geq 0.$$

Proof. To establish the first bound use the Lipschitz continuity of h to obtain the bound

$$\frac{d}{dt}\bigl\|\widehat{\phi}(t)\bigr\|^2 = 2\widehat{\phi}(t)^T h_{r(j)}\bigl(\widehat{\phi}(t)\bigr) \leq C\bigl(\bigl\|\widehat{\phi}(t)\bigr\|^2 + 1\bigr), \quad T(j) \leq t < T(j+1),$$

where C is a deterministic constant, independent of j. The claim follows with $\bar{C} = 2\exp((T+1)C)$ since $\|\widehat{\phi}(T(j))\| \leq 1$. The proof of the second bound is therefore identical. □

The following version of the Bellman Gronwall lemma will be used repeatedly.

LEMMA 4.3.

(i) *Suppose* $\{\alpha(n)\}$, $\{A(n)\}$ *are nonnegative sequences and* $\beta > 0$ *such that*

$$A(n+1) \leq \beta + \sum_{k=0}^{n} \alpha(k)A(k), \quad n \geq 0.$$

Then for all $n \geq 1$,

$$A(n+1) \leq \exp\left(\sum_{k=1}^{n} \alpha(k)\right)\bigl(\alpha(0)A(0) + \beta\bigr).$$

(ii) *Suppose* $\{\alpha(n)\}$, $\{A(n)\}$, $\{\gamma(n)\}$ *are nonnegative sequences such that*

$$A(n+1) \leq \bigl(1 + \alpha(n)\bigr)A(n) + \gamma(n), \quad n \geq 0.$$

Then for all $n \geq 1$,

$$A(n+1) \leq \exp\left(\sum_{k=1}^{n} \alpha(k)\right)\bigl(\bigl(1 + \alpha(0)\bigr)A(0) + \beta(n)\bigr),$$

where $\beta(n) = \sum_0^n \gamma(k)$.

Proof. Define $\{R(n)\}$ inductively by $R(0) = A(0)$ and

$$R(n+1) = \beta + \sum_{k=0}^{n} \alpha(k)R(k), \quad n \geq 0.$$

A simple induction shows that $A(n) \le R(n),\ \ n \ge 0$. An alternative expression for $R(n)$ is

$$R(n) = \left(\prod_{k=1}^{n} (1 + \alpha(k) \right) \bigl(\alpha(0)A(0) + \beta\bigr).$$

The inequality (i) then follows from the bound $1 + x \le e^x$.

To see (ii) fix $n \ge 0$ and observe that on summing both sides of the bound

$$A(k+1) - A(k) \le \alpha(k)A(k) + \gamma(k)$$

over $0 \le k \le \ell$ we obtain for all $0 \le \ell < n$,

$$A(\ell+1) \le A(0) + \beta(n) + \sum_{k=0}^{\ell} \alpha(k)A(k).$$

The result then follows from (i). □

The following lemmas relate the three functions $\phi(\,\cdot\,)$, $\widehat{\phi}(\,\cdot\,)$, and $\phi^\infty(\,\cdot\,)$.

LEMMA 4.4. *Suppose that* (A1) *and* (A2) *hold. Given any $\epsilon > 0$, there exist $T, R < \infty$ such that for any $r > R$ and any solution to the ODE* (1.4) *satisfying $\|x(0)\| \le 1$, we have $\|x(t)\| \le \epsilon$ for $t \in [T, T+1]$.*

Proof. By global asymptotic stability of (1.5) we can find $T > 0$ such that $\|\phi^\infty(t)\| \le \epsilon/2$, $t \ge T$, for solutions $\phi^\infty(\,\cdot\,)$ of (1.5) satisfying $\|\phi^\infty(0)\| \le 1$.

With T fixed, choose R so large that $|\widehat{\phi}(t) - \phi^\infty(t)| \le \epsilon/2$ whenever $\widehat{\phi}$ is a solution to (1.4) satisfying $\widehat{\phi}(0) = \phi^\infty(0)$; $|\widehat{\phi}(0)| \le 1$; and $r \ge R$. This is possible since, as we have already observed, $h_r \to h_\infty$ as $r \to \infty$ uniformly on compact sets. The claim then follows from the triangle inequality. □

Define the following: For $j \ge 0$, $m(j) \le n < m(j+1)$,

$$\widetilde{X}(n) := X(n)/r(j),$$
$$\widetilde{M}(n+1) := M(n+1)/r(j),$$

and for $n \ge 1$,

$$\xi(n) := \sum_{m=0}^{n-1} a(m)\widetilde{M}(m+1).$$

LEMMA 4.5. *Under* (A1), (A2), *and either* (TS) *or* (BS), *for each initial condition $X(0) \in \mathbb{R}^d$ satisfying $\mathsf{E}[\|X(0)\|^2] < \infty$, we have the following:*

(i) $\sup_{n \ge 0} \mathsf{E}[\|\widetilde{X}(n)\|^2] < \infty$.

(ii) $\sup_{j \ge 0} \mathsf{E}[\|X(m(j+1))/r(j)\|^2] < \infty$.

(iii) $\sup_{j \ge 0, T(j) \le t \le T(j+1)} \mathsf{E}[\|\phi(t)\|^2] < \infty$.

(iv) *Under* (TS) *the sequence $\{\xi(n), \mathcal{F}_n\}$ is a square integrable martingale with*

$$\sup_{n \ge 0} \mathsf{E}[\|\xi(n)\|^2] < \infty.$$

Proof. To prove (i) note first that under (A2) and the Lipschitz condition on h there exists $C < \infty$ such that for all $n \ge 1$,

$$\mathsf{E}\bigl[\|X(n)\|^2 \mid \mathcal{F}_{n-1}\bigr] \le \bigl(1 + Ca(n-1)\bigr)\|X(n-1)\|^2 + Ca(n-1), \quad n \ge 0. \tag{4.2}$$

It then follows that for any $j \geq 0$ and any $m(j) < n < m(j+1)$,

$$\mathsf{E}\big[\|\widetilde{X}(n)\|^2 \mid \mathcal{F}_{n-1}\big] \leq \big(1 + Ca(n-1)\big)\|\widetilde{X}(n-1)\|^2 + Ca(n-1),$$

so that by Lemma 4.3 (ii), for all such n,

$$\begin{aligned}\mathsf{E}\big[\|\widetilde{X}(n+1)\|^2\big] &\leq \exp\big(C(T+1)\big)\big(2\mathsf{E}\big[\|\widetilde{X}(m(j))\|^2\big] + C(T+1)\big)\\ &\leq \exp\big(C(T+1)\big)\big(2 + C(T+1)\big).\end{aligned}$$

Claim (i) follows, and claim (ii) follows similarly. We then obtain claim (iii) from the definition of $\phi(\,\cdot\,)$. From (i), (ii), and (A2), we have $\sup_n \mathsf{E}[\|\widetilde{M}(n)\|^2] < \infty$. Using this and the square summability of $\{a(n)\}$ assumed in (TS), the bound (iv) immediately follows. ◻

LEMMA 4.6. *Suppose* $\mathsf{E}[\|X(0)\|^2] < \infty$. *Under* (A1), (A2), *and* (TS), *with probability one,*

(i) $\|\phi(t) - \widehat{\phi}(t)\| \to 0$ *as* $t \to \infty$,

(ii) $\sup_{t\geq 0}\|\phi(t)\| < \infty$.

Proof. Express $\widehat{\phi}(\,\cdot\,)$ as follows: For $m(j) \leq n < m(j+1)$,

$$\begin{aligned}\widehat{\phi}(t(n+1)-) &= \widehat{\phi}(T(j)) + \sum_{i=m(j)}^{n}\int_{t(i)}^{t(i+1)} h_{r(j)}\big(\widehat{\phi}(s)\big)ds\\ &= \widehat{\phi}(T(j)) + \epsilon_1(j) + \sum_{i=m(j)}^{n} a(i)h_{r(j)}\big(\widehat{\phi}(t(i))\big),\end{aligned} \tag{4.3}$$

where $\epsilon_1(j) = O(\sum_{i=m(j)}^{m(j+1)} a(i)^2) \to 0$ as $j \to \infty$. The "$-$" covers the case where $t(n+1) = t(m(j+1)) = T(j+1)$.

We also have by definition

$$\phi\big(t(n+1)-\big) = \phi\big(T(j)\big) + \sum_{i=m(j)}^{n} a(i)\big[h_{r(j)}\big(\phi\big(t(i)\big)\big) + \widetilde{M}(i+1)\big]. \tag{4.4}$$

For $m(j) \leq n \leq m(j+1)$, let $\varepsilon(n) = \|\phi(t(n)-) - \widehat{\phi}(t(n)-)\|$. Combining (4.3), (4.4), and the Lipschitz continuity of h, we have

$$\varepsilon(n+1) \leq \varepsilon\big(m(j)\big) + \epsilon_1(j) + \|\xi(n+1) - \xi(m(j))\| + C\sum_{i=m(j)}^{n} a(i)\varepsilon(i),$$

where $C < \infty$ is a suitable constant. Since $\varepsilon(m(j)) = 0$, we can use Lemma 4.3 (i) to obtain

$$\varepsilon(n) \leq \exp\big(C(T+1)\big)\big(\epsilon_1(j) + \epsilon_2(j)\big), \qquad m(j) \leq n \leq m(j+1),$$

where $\epsilon_2(j) = \max_{m(j)\leq n\leq m(j+1)}\|\xi(n+1) - \xi(m(j))\|$. By (iv) of Lemma 4.5 and the martingale convergence theorem [18, p. 62], $\{\xi(n)\}$ converges a.s.; thus $\epsilon_2(j) \to 0$ a.s., as $j \to \infty$. Since $\epsilon_1(j) \to 0$ as well,

$$\sup_{m(j)\leq n\leq m(j+1)}\big\|\phi(t(n)-) - \widehat{\phi}(t(n)-)\big\| = \sup_{m(j)\leq n\leq m(j+1)}\varepsilon(n) \to 0$$

as $j \to \infty$, which implies the first claim.

Result (ii) then follows from Lemma 4.2 and the triangle inequality. ∎

LEMMA 4.7. *Under* (A1), (A2), *and* (BS), *there exists a constant* $C_2 < \infty$ *such that for all* $j \geq 0$,

(i) $\sup_{j\geq 0, T(j)\leq t\leq T(j+1)} \mathsf{E}[\|\phi(t) - \widehat{\phi}(t)\|^2 \mid \mathcal{F}_{n(j)}] \leq C_2\overline{\alpha}$,

(ii) $\sup_{j\geq 0, T(j)\leq t\leq T(j+1)} \mathsf{E}[\|\phi(t)\|^2 \mid \mathcal{F}_{n(j)}] \leq C_2$.

Proof. Mimic the proof of Lemma 4.6 to obtain

$$\varepsilon(n+1) \leq \sum_{i=m(j)}^{n} Ca(i)\varepsilon(i) + \epsilon_0(j), \qquad m(j) \leq n < m(j+1),$$

where $\varepsilon(n) = \mathsf{E}[\|\phi(t(n)-) - \widehat{\phi}(t(n)-)\|^2 \mid \mathcal{F}_{m(j)}]^{1/2}$ for $m(j) \leq n \leq m(j+1)$, and the error term has the upper bound

$$|\epsilon_0(j)| = O(\overline{\alpha}),$$

where the bound is deterministic. By Lemma 4.3 (i) we obtain the bound,

$$\varepsilon(n) \leq \exp\bigl(C(T+1)\bigr)\epsilon_0(j), \qquad m(j) \leq n \leq m(j+1),$$

which proves (i). We, therefore, obtain (ii) using Lemma 4.2, (i), and the triangle inequality. ∎

Proof of Theorem 2.1. (i) By a simple conditioning argument, we may take $X(0)$ to be deterministic without any loss of generality. In particular, $\mathsf{E}[\|X(0)\|^2] < \infty$ trivially. By Lemma 4.6 (ii), it now suffices to prove that $\sup_n \|X(m(n))\| < \infty$ a.s. Fix a sample point outside the zero probability set where Lemma 4.6 fails. Pick $T > 0$ as above and $R > 0$ such that for every solution $x(\,\cdot\,)$ of the ODE (1.4) with $\|x(0)\| \leq 1$ and $r \geq R$, we have $\|x(t)\| \leq \frac{1}{4}$ for $t \in [T, T+1]$. This is possible by Lemma 4.4.

Hence by Lemma 4.6 (i) we can find an $j_0 \geq 1$ such that whenever $j \geq j_0$ and $\|X(m(j))\| \geq R$,

$$\frac{\|X(m(j+1))\|}{\|X(m(j))\|} = \phi\bigl(T(j+1) - \bigr) \leq \frac{1}{2}. \tag{4.5}$$

This implies that $\{X(m(j)) : j \geq 0\}$ is a.s. bounded, and the claim follows.

(ii) For $m(j) < n \leq m(j+1)$,

$$\begin{aligned} \mathsf{E}\bigl[\|X(n)\|^2 \mid \mathcal{F}_{m(j)}\bigr]^{1/2} &= \mathsf{E}\bigl[\|\phi\bigl(t(n) - \bigr)\|^2 \mid \mathcal{F}_{m(j)}\bigr]^{1/2}\bigl(\|X(m(j))\| \vee 1\bigr) \\ &\leq \mathsf{E}\bigl[\|\phi\bigl(t(n) - \bigr) - \widehat{\phi}\bigl(t(n) - \bigr)\|^2 \mid \mathcal{F}_{m(j)}\bigr]^{1/2}\bigl(\|X\bigl(m(j)\bigr)\| \vee 1\bigr) \\ &\quad + \mathsf{E}\bigl[\|\widehat{\phi}\bigl(t(n)-\bigr)\|^2 \mid \mathcal{F}_{m(j)}\bigr]^{1/2}\bigl(\|X(m(j))\| \vee 1\bigr). \end{aligned} \tag{4.6}$$

Let $0 < \eta < \frac{1}{2}$, and let $\alpha^* = \eta/(2C_2)$, for C_2 as in Lemma 4.7. We then obtain for $\overline{\alpha} \leq \alpha^*$,

$$\begin{aligned} \mathsf{E}\bigl[\|X(n)\|^2 \mid \mathcal{F}_{m(j)}\bigr]^{1/2} &\leq (\eta/2)\bigl(\|X\bigl(m(j)\bigr)\| \vee 1\bigr) \\ &\quad + \mathsf{E}\bigl[\|\widehat{\phi}\bigl(t(n) - \bigr)\|^2 \mid \mathcal{F}_{m(j)}\bigr]^{1/2}\bigl(\|X\bigl(m(j)\bigr)\| \vee 1\bigr). \end{aligned} \tag{4.7}$$

Choose $R, T > 0$ such that for any solution $x(\,\cdot\,)$ of the ODE (1.4), $\|x(t)\| < \eta/2$ for $t \in [T, T+1]$, whenever $\|x(0)\| < 1$ and $r \geq R$. When $\|X(m(j))\| \geq R$, we then obtain

$$\mathsf{E}\big[\big\|X\big(m(j+1)\big)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2} \leq \eta\big\|X\big(m(j)\big)\big\|, \tag{4.8}$$

while by Lemma 4.7 (ii) there exists a constant C such that the left-hand side (l.h.s.) of the inequality above is bounded by C a.s. when $\|X(m(j))\| \leq R$. Thus,

$$\mathsf{E}\big[\big\|X\big(m(j+1)\big)\big\|^2\big] \leq 2\eta^2\mathsf{E}\big[\big\|X\big(m(j)\big)\big\|^2\big] + 2C^2.$$

This establishes boundedness of $\mathsf{E}[\|X(m(j+1))\|^2]$, and the proof then follows from (4.7) and Lemma 4.2. $\square$

4.2. Convergence for (BS). LEMMA 4.8. *Suppose that* (A1), (A2), *and* (BS) *hold and that* $\overline{\alpha} \leq \alpha^*$. *Then for some constant* $C_3 < \infty$,

$$\sup_{t\geq 0}\mathsf{E}\big[\big\|\widehat{\psi}(t) - \psi(t)\big\|^2\big] \leq C_3\overline{\alpha}.$$

Proof. By (A2) and Theorem 2.1 (ii),

$$\sup_n \mathsf{E}\big[\big\|X(n)\big\|^2\big] < \infty, \quad \sup_n \mathsf{E}\big[\big\|M(n)\big\|^2\big] < \infty.$$

The claim then follows from familiar arguments using the Bellman Gronwall lemma exactly as in the proof of Lemma 4.6. $\square$

Proof of Theorem 2.3. (i) We apply Theorem 2.1 which allows us to choose an $R > 0$ such that

$$\sup_n \mathsf{P}\big(\|X(n)\| > R\big) < \overline{\alpha}.$$

Let $B(c)$ denote the ball centered at x^* of radius $c > 0$ and let $0 < \mu < \epsilon/2$ be such that if a solution $x(\,\cdot\,)$ of (1.2) satisfies $x(0) \in B(\mu)$, then $x(t) \in B(\epsilon/2)$ for $t \geq 0$. Pick $T > 0$ such that if a solution $x(\,\cdot\,)$ of (1.2) satisfies $\|x(0)\| \leq R$, then $x(t) \in B(\mu/2)$ for $t \in [T, T+1]$. Then for all $j \geq 0$,

$$\begin{aligned}\mathsf{P}\big(e\big(m(j+1)\big) \geq \mu\big) &= \mathsf{P}\Big(e\big(m(j+1)\big) \geq \mu, \big\|X\big(m(j)\big)\big\| > R\big)\\ &\qquad + \mathsf{P}\big(e\big(m(j+1)\big) \geq \mu, \|X(m(j))\| \leq R\Big)\\ &\leq \overline{\alpha} + \mathsf{P}\Big(\psi\big(T(j+1)\big) \notin B(\mu), \widehat{\psi}\big(T(j+1)\big) \in B(\mu/2)\Big)\\ &\leq \overline{\alpha} + \mathsf{P}\Big(\|\psi\big(T(j+1)\big) - \widehat{\psi}\big(T(j+1)\big)\| > \mu/2\Big)\\ &\leq O(\overline{\alpha})\end{aligned}$$

by Lemma 4.8. Then for $m(j) \leq n < m(j+1)$,

$$\begin{aligned}\mathsf{P}\big(e(n) \geq \epsilon\big) &= \mathsf{P}\big(e(n) \geq \epsilon, e\big(m(j)\big) \geq \mu\big)\\ &\qquad + \mathsf{P}\big(e(n) \geq \epsilon, e\big(m(j)\big) \leq \mu\big)\\ &\leq O(\overline{\alpha}) + \mathsf{P}\big(\psi\big(t(n)\big) \notin B(\epsilon), \widehat{\psi}\big(t(n)\big) \in B\big(\epsilon/2\big)\big)\\ &\leq O(\overline{\alpha}) + \mathsf{P}\big(\|\psi\big(t(n)\big) - \widehat{\psi}\big(t(n)\big)\| > \epsilon/2\big)\\ &\leq O(\overline{\alpha}).\end{aligned}$$

Since the bound on the r.h.s. is uniform in n, the claim follows.

(ii) We first establish the bound with $n = m(j+1)$, $j \to \infty$. We have for any j,

$$\begin{aligned} \mathsf{E}\big[e\big(m(j+1)\big)^2\big]^{1/2} &\le \mathsf{E}\big[\big\|\psi\big(T(j+1)-\big) - \widehat{\psi}\big(T(j+1)-\big)\big\|^2\big]^{1/2} \\ &\quad + \mathsf{E}\big[\big\|\widehat{\psi}\big(T(j+1)-\big) - x^*\big\|^2\big]^{1/2}. \end{aligned} \tag{4.9}$$

By exponential stability there exist $C < \infty$, $\delta > 0$ such that for all $j \ge 0$,

$$\begin{aligned} \big\|\widehat{\psi}\big(T(j+1)-\big) - x^*\big\| &\le C\exp\big(-\delta\big[T(j+1)-T(j)\big]\big)\big\|\widehat{\psi}\big(T(j)\big) - x^*\big\| \\ &\le C\exp(-\delta T)\big\|\widehat{\psi}\big(T(j)\big) - x^*\big\|. \end{aligned}$$

Choose T so large that $C\exp(-\delta T) \le \frac{1}{2}$ so that

$$\begin{aligned} \mathsf{E}\big[\big\|\widehat{\psi}\big(T(j+1)-\big) - x^*\big\|^2\big]^{1/2} &\le \frac{1}{2}\mathsf{E}\big[\big\|\widehat{\psi}\big(T(j)\big) - x^*\big\|^2\big]^{1/2} \\ &\le \frac{1}{2}\mathsf{E}\big[e\big(m(j)\big)^2\big]^{1/2} + \frac{1}{2}\mathsf{E}\big[\big\|\psi\big(T(j)\big) - \widehat{\psi}\big(T(j)\big)\big\|^2\big]^{1/2}. \end{aligned} \tag{4.10}$$

Combining (4.9) and (4.10) with Lemma 4.8 gives

$$\mathsf{E}\big[e\big(m(j+1)\big)^2\big]^{1/2} \le \frac{1}{2}\mathsf{E}\big[e\big(m(j)\big)^2\big]^{1/2} + 2\sqrt{C_3\overline{\alpha}},$$

which shows that

$$\limsup_{j\to\infty} \mathsf{E}\big[e\big(m(j)\big)^2\big] \le 16C_3\overline{\alpha}.$$

The result follows from this and Lemma 4.7 (ii). □

Proof of Theorem 2.5. The details of the proof, though pedestrian in the light of the foregoing and [6], are quite lengthy, not to mention the considerable overhead of additional notation, and are therefore omitted. We briefly sketch below a single point of departure in the proof.

In Lemma 4.6 we compare two functions $\phi(\,\cdot\,)$ and $\widehat{\phi}(\,\cdot\,)$ on the interval $[T(j), T(j+1)]$. The former in turn involved the iterates $\widetilde{X}(n)$ for $m(j) \le n < m(j+1)$ or, equivalently, $X(n)$ for $m(j) \le n < m(j+1)$. Here $X(n+1)$ was computed in terms of $X(n)$ and the "noise" $M(n+1)$. In the asynchronous case, however, the evaluation of $X_j(n+1)$ can involve $X_j(n)$ for $n - \overline{\tau} \le m \le n$, $j \ne i$. Therefore the argument leading to Lemma 4.6 calls for a slight modification. While computing $X(n), m(j) \le n < m(j+1)$, we plug into the iteration as and when required $\widetilde{X}_i(m) = X_i(m)/r(j)$. Note, however, that if the same $X_i(m)$ also features in the computation of $X_k(l)$ for $m(q) \le \ell < m(q+1)$, say, with $q \ne j$, then $\widetilde{X}_i(m)$ should be redefined there as $X_i(m)/r(q)$. Thus the definition of $\widetilde{X}_i(m)$ now becomes context-dependent.

With this minor change, the proofs of [6] can be easily combined with the arguments used in the proofs of Theorems 2.1 and 2.2 to draw the desired conclusions. □

4.3. The Markov model. The bounds that we obtain for the Markov model (2.4) are based upon the theory of ψ-irreducible Markov chains.

A subset $S \subset \mathbb{R}^d$ is called *petite* if there exists a probability measure ν on $\mathbb{R}^d$ and $\delta > 0$ such that the resolvent kernel K satisfies

$$K(x, A) := \sum_{k=0}^{\infty} 2^{-k-1}P^k(x, A) \ge \delta\nu(A), \quad x \in S,$$

for any measurable $A \subset \mathbb{R}^d$. Under assumptions (2.6) and (2.7) we show below that every compact subset of $\mathbb{R}^d$ is petite, so that $\mathbf{\Phi}$ is a ψ-irreducible *T-chain.* We refer the reader to [16] for further terminology and notation.

LEMMA 4.9. *Suppose that* (A1)*,* (A2)*,* (2.6)*, and* (2.7) *hold and that $\alpha \le \alpha^*$. Then all compact subsets of $\mathbb{R}^d$ are petite for the Markov chain $\boldsymbol{X}$, and hence the chain is ψ-irreducible.*

Proof. The conclusions of the theorem will be satisfied if we can find a function s which is bounded from below on compact sets and a probability ν such that the resolvent kernel K satisfies the bound

$$K(x, A) \ge s(x)\nu(A)$$

for every $x \in \mathbb{R}^d$ and any measurable subset $A \subset \mathbb{R}^d$. This bound is written succinctly as $K \ge s \otimes \nu$.

The first step of the proof is to apply the implicit function theorem together with (2.6) and (2.7) to obtain a bound of the form

$$P^d(x, A) = \mathsf{P}(X(d) \in A \mid X(0) = x) \ge \epsilon\nu(A), \quad x \in O,$$

where O is an open set containing $x^*, \epsilon > 0$, and ν is the uniform distribution on O. The set O can be chosen independent of α, but the constant ϵ may depend on α. For details on this construction, see Chapter 7 of [16].

To complete the proof it is enough to show that $K(x, O) > 0$. To see this, suppose that $\alpha \le \alpha^*$ and that $W(n) = w^*$ for all n. Then the foregoing stability analysis shows that $X(n) \in O$ for all n sufficiently large. Since w^* is in the support of the marginal distribution of $\{W(n)\}$, it then follows that $K(x, O) > 0$.

From these two bounds, we then have

$$K(x, A) \ge 2^{-d} \int K(x, dy) P^d(y, A) \ge 2^{-d}\epsilon K(x, O)\nu(A).$$

This is of the form $K \ge s \otimes \nu$ with s lower semicontinuous and positive everywhere. The function s is therefore bounded from below on compact sets, which proves the claim. □

The previous lemma together with Theorem 2.1 allows us to establish a strong form of ergodicity for the model.

LEMMA 4.10. *Suppose that* (A1)*,* (A2)*,* (2.6)*, and* (2.7) *hold and that $\alpha \le \alpha^*$.*

(i) *There exists a function $V_\alpha : \mathbb{R}^d \to [1, \infty)$ and constants $b, L < \infty$ and $\epsilon_0 > 0$ independent of α such that*

$$PV_\alpha(x) \le \exp(-\epsilon_0\alpha)V_\alpha(x) + bI_C(x),$$

where $C = \{x : \|x\| \le L\}$. While the function V_α will depend upon α, it is uniformly bounded as follows,

$$\gamma^{-1}(\|x\|^2 + 1) \le V_\alpha(x) \le \gamma(\|x\|^2 + 1),$$

where $\gamma \ge 1$ does not depend upon α.

(ii) *The chain is V-uniformly ergodic, with $V(x) = \|x\|^2 + 1$.*

Proof. Using (4.8) we may construct T and L independent of $\alpha \le \alpha^*$ such that

$$\mathsf{E}\big[\|X(k_0)\|^2 + 1 \mid X(0) = x\big] \le (1/2)(\|x\|^2 + 1), \quad \|x\| \ge L,$$

where $k_0 = [T/\alpha] + 1$. We now set

$$V_\alpha(x) = \alpha \sum_{k=0}^{k_0-1} \mathsf{E}\left[\left\|X(k)\right\|^2 + 1 \mid X(0) = x\right] 2^{k/k_0}.$$

From the previous bound, it follows directly that the desired drift inequality holds with $\epsilon_0 = \log(2)/T$. Lipschitz continuity of the model gives the bounds on V_α. This proves (i).

The V-uniform ergodicity then follows from Lemma 4.9 and Theorem 16.0.1 of [16]. □

We note that for small α and large x, the Lyapunov function V_α approximates V_∞ plus a constant, where

$$V_\infty(x) = \int_0^T \left(\|x(s)\|^2 + 1\right) 2^{s/T} ds; \quad x(0) = x,$$

and $x(\,\cdot\,)$ is a solution to (1.5). If this ODE is asymptotically stable then the function V_∞ is in fact a Lyapunov function for (1.5), provided $T > 0$ is chosen sufficiently large.

In [17] a bound is obtained on the rate of convergence ρ given in (2.5) for a chain satisfying the drift condition

$$PV_\alpha(x) \le \lambda V(x) + bI_C(x).$$

The bound depends on the "petiteness" of the set C and the constants $b < \infty$ and $\lambda < 1$. The bound on ρ obtained in [17] also tends to unity with vanishing α since in the preceding lemma we have $\lambda = \exp(-\epsilon_0\alpha) \to 1$ as $\alpha \to 0$. From the structure of the algorithm this is not surprising, but this underlines the fact that care must be taken in the choice of the stepsize α.

REFERENCES

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar, *Learning algorithms for Markov decision processes with average cost*, SIAM J. Control Optim., submitted.

[2] A. G. Barto, R. S. Sutton, and C. W. Anderson, *Neuron-like elements that can solve difficult learning control problems*, IEEE Trans. Systems, Man and Cybernetics, 13 (1983), pp. 835–846.

[3] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[4] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[5] V. S. Borkar, *Stochastic approximation with two time scales*, Systems Control Lett., 29 (1997), pp. 291–294.

[6] V. S. Borkar, *Asynchronous stochastic approximation*, SIAM J. Control Optim., 36 (1998), pp. 840–851.

[7] V. S. Borkar, *Recursive self-tuning control of finite Markov chains*, Appl. Math., 24 (1996), pp. 169–188.

[8] V. S. Borkar and K. Soumyanath, *An analog scheme for fixed-point computation, part* I*: Theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 351–355.

[9] J. G. Dai, *On positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 49–77.

[10] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control, 40 (1995), pp. 1889–1904.

[11] M. W. Hirsch, *Convergent activation dynamics in continuous time networks*, Neural Networks, 2 (1989), pp. 331–349.

[12] T. Jaakola, M. I. Jordan, and S. P. Singh, *On the convergence of stochastic iterative dynamic programming algorithms*, Neural Computation, 6 (1994), pp. 1185–1201.

[13] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (1999), pp. 94–123.

[14] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.

[15] V. A. Malyshev and M. V. Men'sikov, *Ergodicity, continuity and analyticity of countable Markov chains*, Trans. Moscow Math. Soc., 1 (1982), pp. 1–48.

[16] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[17] S. P. Meyn and R. L. Tweedie, *Computable bounds for geometric convergence rates of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 981–1011.

[18] J. Neveu, *Discrete Parameter Martingales*, North Holland, Amsterdam, 1975.

[19] T. Sargent, *Bounded Rationality in Macroeconomics*, Clarendon Press, Oxford, 1993.

[20] J. Tsitsiklis, *Asynchronous stochastic approximation and q-learning*, Mach. Learning, 16 (1994), pp. 195–202.

[21] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Mach. Learning, 8 (1992), pp. 279–292.