Proceedings of the 37th IEEE
Conference on Decision & Control
Tampa, Florida USA • December 1998

Stability and Convergence of Stochastic Approximation using the O.D.E. Method

V.S. Borkar[1] S.P. Meyn[2]

Abstract

It is shown here that stability of the stochastic approximation algorithm is implied by the asymptotic stability of the origin for an associated o.d.e. 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.

1 Introduction

The stochastic approximation algorithm considered in this paper is described by the d-dimensional recursion given for $n \geq 0$ by

$$X(n+1) = X(n) + a(n)[h(X(n)) + M(n+1)], \quad (1)$$

where $X(n) = [X_1(n), \cdots, X_d(n)]^T \in \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$, 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) 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]. 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 (o.d.e.)

$$\dot{x}(t) = h(x(t)) \quad (2)$$

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

The recurrence assumption is crucial, but unfortunately there is no general scheme for verifying this, only a repertoire of special techniques, each with its own domain of applicability. To mention just two, one has techniques based upon the contractive properties or homogeneity properties of the functions involved (see, e.g., [20] and [13] 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 [10], [11], 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) with a deterministic process is similar to the construction of the fluid model of [10, 11]. 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 inter-

[1] Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India (borkar@csa.iisc.ernet.in). Work supported in part by the Dept. of Science and Technology (Govt. of India) grant no.III5(12)/96-ET.

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

0-7803-4394-8/98 $10.00 © 1998 IEEE

polation. The times $\{t(j) : j \geq 0\}$ are defined in terms of the constants $\{a(j)\}$ used in (1). For any $r > 0$ the scaled function $h_r : \mathbb{R}^d \to \mathbb{R}^d$ is given for $x \in \mathbb{R}^d$ by

$$h_r(x) = h(rx)/r. \tag{3}$$

Then through elementary arguements we find that the stochastic process ϕ approximates the solution $\widehat{\phi}$ to the associated o.d.e.

$$\dot{x}(t) = h_r(x(t)), \tag{4}$$

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 o.d.e.

$$\dot{x}(t) = h_\infty(x(t)). \tag{5}$$

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

So, 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) is stable in a strong sense provided the origin is asymptotically stable for the limiting o.d.e. (5). Equation (5) is precisely the fluid model of [10, 11].

Thus, the major conclusion of this paper is that the o.d.e. method can be extended to establish both the stability *and* convergence of the stochastic approximation method, as opposed to only the latter. 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.

2 Main Results

Here we collect together the main general results concerning the stochastic approximation algorithm. Proofs not included here may be found in [5].

We shall impose the following additional conditions on the functions $\{h_r : r \geq 1\}$ defined in (3), and the sequence $\boldsymbol{M} = \{M(n) : n \geq 1\}$ used in (1). Some relaxations of the assumption (A1) are discussed in [5].

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

$$\lim_{r\to\infty} h_r(x) = h_\infty(x).$$

Furthermore, the origin in $\mathbb{R}^d$ is an asymptotically stable equilibrium for the o.d.e. (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 $X(0) \in \mathbb{R}^d$, $n \geq 0$,

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

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

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

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

$$\underline{\alpha} \leq a(n) \leq \overline{\alpha}.$$

2.1 Stability and convergence

The first result shows that the algorithm is stabilizing for both the bounded and tapering step size algorithm.

Theorem 2.1 *Assume that (A1), (A2) hold. Then,*

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

$$\sup_n \|X(n)\| < \infty \qquad a.s.$$

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

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

□

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 Theorem 1, pp.339, [12], or a complete proof in [5]).

Theorem 2.2 *Suppose that (A1), (A2), (TS) hold and that the o.d.e. (2) has a unique globally asymptotically stable equilibrium* x^*. *Then* $X(n) \to x^*$ *a.s. as* $n \to \infty$ *for any initial condition* $X(0) \in \mathbb{R}^d$. □

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

$$e(n) := \|X(n) - x^*\|, \qquad n \geq 0. \tag{6}$$

Theorem 2.3 *Assume that (A1), (A2) and (BS) hold, and suppose that (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*
$$\liminf_{n\to\infty} \mathsf{P}(e(n) \geq \epsilon) \leq b_1\overline{\alpha}.$$

(ii) *If x^* is a globally exponentially asymptotically stable equilibrium for the o.d.e. (2), then there exists $b_2 < \infty$ such that for every initial condition $X(0) \in \mathbb{R}^d$,*
$$\limsup_{n\to\infty} \mathsf{E}[e(n)^2] \leq b_2\overline{\alpha}.$$

□

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 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

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

where $\{W(n)\}$ is i.i.d. with mean zero, and variance σ^2. This is of the form (1) with $h(x) = -(x - x^*)$ and $M(n) = W(n)$. The error $e(n+1)$ efined in (6) may be bounded as follows:

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

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

$$\mathsf{E}[e(n)^2 \mid X(0) = x] \leq B_1(\alpha)+B_2(\|x\|^2+1)\exp(-\epsilon_0(\alpha)n) \tag{7}$$

where B_1, B_2 and ϵ_0 are positive-valued functions of α. The bound (7) is of the form that we seek: the first term on the 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 tradeoff 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 (see [5]).

2.2 The asynchronous case

The conclusions above also extend to the model of asynchronous stochastic approximation analysed in [7]. 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 $\{1, 2, \cdots, d\}$, with the interpretation: $Y(n) = \{$ indices of the components updated at time $n\}$. For $n \geq 0, 1 \leq i \leq d$, define

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

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

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

This ensures that all components are updated comparably often.

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 $\overline{\tau} < \infty$ ([7] considers a slightly more general situation.)

To relate the present work to [7], we recall that the 'centralized' algorithm of [7] is

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

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

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

for $n \geq 0$. The asynchronous version then is $X_i(n+1) =$

$$\begin{aligned}&X_i(n) + a(\nu(i,n))f(X_1(n-\tau_{i1}(n)), \cdots \\ &\cdots, X_d(n-\tau_{id}(n)), W(n+1))I\{i \in Y(n)\},\end{aligned} \tag{8}$$

for $1 \leq i \leq d$. Note that this can be executed by the i-th 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 [7] 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 $x \in (0,1)$,

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

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

A fourth condition is imposed in [7], but this becomes irrelevant when the delays are bounded. Examples of $\{a(n)\}$ satisfying the (i)–(iii) are $a(n) = 1/(n+1)$, or $1/(1+n\log(n+1))$.

As a first simplifying step, it is observed in [7] 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 [7] 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 [7] to conclude that the suitably interpolated and rescaled trajectory of the algorithm tracks an appropriate o.d.e.. The only difference is a scalar factor $1/d$ multiplying the r.h.s. of the o.d.e. (i.e., $\dot{x}(t) = \frac{1}{d}h(x(t))$). This factor, which reflects the asynchronous sampling, amounts to a time scaling that does not affect the qualitative behavior of the o.d.e.

Theorem 2.4 *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 (10) remain a.s. bounded and (therefore) converge to x^* a.s.* □

3 Reinforcement learning

As both an illustration of the theory and an important application in its own right, in this section we analyse 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 $\mathbf{\Phi} = \{\Phi(t) : t \in \mathbb{Z}_+\}$ taking values in a finite state space $S = \{1, 2, \cdots, s\}$ and controlled by a control sequence $\boldsymbol{Z} = \{Z(t) : t \in \mathbb{Z}_+\}$ taking values in a finite action space $A = \{a_0, \cdots, a_r\}$. We assume that the control sequence is *admissible* in the sense that $Z(n) \in \sigma\{\Phi(t) : t \leq n\}$ 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 $w: S \to A$. 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

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

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

$$V(i) = \min_a \Big[c(i,a) + \beta \sum_j p(i,j,a)V(j)\Big],$$

$i \in S$, and the optimal control minimizing J is given as the stationary policy defined through the feedback law w^* given as any solution to

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

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

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}[c(\Phi(t), Z(t))]. \tag{9}$$

3.2 Q-learning

If we define Q*-values* via

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

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

$$Q(i,a) = c(i,a) + \beta \sum_j p(i,j,a) \min_b Q(j,b),$$

for $i \in S, a \in A$. If the matrix Q 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 $Q_{n+1}(i,a) =$

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

$i \in S, a \in A$, where $W_{n+1}(i,a)$ 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

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

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

$$\beta\Big(\min_b Q_n(W_{n+1}(i,a),b)-\sum_j p(i,j,a)\Big(\min_b Q_n(j,b)\Big)\Big),$$

$i \in S, a \in A$. 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 o.d.e. is

$$\dot{Q} = F(Q) - Q := h(Q). \qquad (10)$$

The map $F : \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is a contraction w.r.t. the max norm $\|\cdot\|_\infty$. The global asymptotic stability of its unique equilibrium point is a special case of the results of [9]. This $h(\cdot)$ fits the framework of our analysis, with the (i,a)-th component of $h_\infty(Q)$ given by

$$\beta\sum_j p(i,j,a)\min_b Q(j,b) - Q(i,a),$$

$i \in S, a \in A$. This also is of the form $h_\infty(Q) = F_\infty(Q) - Q$ where $F_\infty(\cdot)$ is an $\|.\|_\infty$- contraction, and thus the asymptotic stability of the unique equilibrium point of the corresponding o.d.e. is guaranteed (see [9]). We conclude that assumptions (A1) and (A2) hold, and hence also Theorems 2.1–?? 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, 14] for a discussion). There are several variants of this, one of which, taken from [14], is as follows: For $i \in S$, $V_{n+1}(i) =$

$$V_n(i) + b(n)[c(i,\psi_n(i)) + \beta V_n(W_n(i,\psi_n(i))) - V_n(i)], \qquad (11)$$

and $\widehat{\pi}_{n+1}(i) =$

$$\Gamma\Big\{\widehat{\pi}_n(i) + \sum_{\ell=1}^{r} a(n)[c(i,a_0) - c(i,a_\ell) \qquad (12)$$

$$+\beta[V_n(\eta_n(i,a_0)) - V_n(\eta_n(i,a_\ell))]e_\ell\Big\}. \qquad (13)$$

Here $\{V_n\}$ are s-vectors and for each $i, \{\widehat{\pi}_n(i)\}$ are r-vectors lying in the simplex $\{x \in \mathbb{R}^r \mid x = [x_1,\cdots,x_r], x_i \geq 0, \sum_i x_i \leq 1\}$. $\Gamma(\cdot)$ is the projection onto this simplex. The sequences $\{a(n)\}, \{b(n)\}$ satisfy $\sum_n a(n) = \sum_n b(n) = \infty, \sum_n a(n)^2 < \infty, \sum_n b(n)^2 < \infty, a(n) = o(b(n))$. The rest of the notation is as follows: For $1 \leq \ell \leq r, e_\ell$ is the unit r-vector in the ℓ-th coordinate direction.

For each $i, n, \pi_n(i) = \pi_n(i,.)$ is a probability vector on A defined by: For $\widehat{\pi}_n(i) = [\widehat{\pi}_n(i,1),...,\widehat{\pi}_n(i,r)]$,

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

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

The different choices of stepsize schedules for the two iterations (11), (13) induces the 'two time-scale' effect discussed in [6]. Thus 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 sheme for minimization over A. See [14] for details.

The boundedness of $\{\widehat{\pi}_n\}$ is guaranteed by the projection $\Gamma(\cdot)$. For $\{V_n\}$, the fact that $b(n) = o(a(n))$ allows one to treat $\widehat{\pi}_n(i)$ as constant, say $\overline{\pi}(i)$ - see, e.g., [14]. The appropriate o.d.e. then turns out to be

$$\dot{v} = G(v) - v := h(v) \qquad (14)$$

where the ith component of $G : \mathbb{R}^s \to \mathbb{R}^s$ is defined by:

$$G_i(x) = \sum_\ell \overline{\pi}(i,a_\ell)\Big[\beta\sum_j p(i,j,a_\ell)x_j + c(i,a_\ell)\Big] - x_i.$$

Once again, $G(\cdot)$ is an $\|\cdot\|_\infty$-contraction and it follows from the results of [9] that (14) 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 i-th component is

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

We see that G_∞ is also a $\|\cdot\|_\infty$- contraction and the global asymptoyic stability of the origin for the corresponding limiting o.d.e. follows as before from the results of [9].

Analogous results have been obtained for the average cost optimal control problem. Asynchronous versions of all the above can be written down along the lines of (10). Then by Theorem 2.4, 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], [13], [20]. For asynchronous algorithms, it is proved for the discounted cost only [1], [14], [20]. See [5] for further details.

References

[1] ABOUNADI, J., BERTSEKAS, D., BORKAR, V.S., Learning algorithms for Markov decision processes with average cost, Lab. for Info. and Decision Systems, M.I.T., 1996, (Draft report).

[2] BARTO, A.G., SUTTON, R.S., ANDERSON, C.W., Neuron-like elements that can solve difficult learning control problems, IEEE Trans. on Systems, Man and Cybernetics 13 (1983), 835-846.

[3] BENVENISTE, A., METIVIER, M., PRIOURET, P., Adaptive Algorithms and Stochastic Approximations, Springer Verlag, Berlin-Heidelberg, 1990.

[4] BERTSEKAS, D., TSITSIKLIS, J., Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

[5] Borkar, V., MEYN, S., The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning. To appear, SIAM J. Control and Optim. 1998. See also http://black.csl.uiuc.edu:80/~meyn

[6] BORKAR, V.S., Stochastic approximations with two time scales, Systems and Control Letters 29 (1997), 291-294.

[7] BORKAR, V.S., Asynchronous stochastic approximation, to appear in SIAM J. Control and Optim. 1998.

[8] BORKAR, V.S., Recursive self-tuning control of finite Markov chains, Applicationes Mathematicae 24 (1996), 169-188.

[9] BORKAR, V.S., SOUMYANATH, K., An analog scheme for fixed point computation, Part I: Theory, IEEE Trans. Circuits and Systems I. Fundamental Theory and Appl. 44 (1997), 351-354.

[10] DAI, J.G., On the positive Harris recurrence for multiclass queueing networks: a unified approach via fluid limit models, Ann. Appl. Prob. 5 (1995), 49-77.

[11] DAI, J.G., MEYN, S.P., Stability and convergence of moments for multiclass queueing networks via fluid limit models, IEEE Trans. Automatic Control 40 (1995), 1889-1904.

[12] HIRSCH, M.W., Convergent activation dynamics in continuous time networks, Neural Networks 2 (1989), 331-349.

[13] JAAKOLA, T., JORDAN, M.I., SINGH, S.P., On the convergence of stochastic iterative dynamic programming algorithms, Neural Computation 6, (1994), 1185-1201.

[14] KONDA, V.R., BORKAR, V.S., Actor-critic type learning algorithms for Markov decision processes, submitted.

[15] MALYSHEV, V.A., MEN'SIKOV, M.V., Ergodicity, continuity and analyticity of countable Markov chains, Trans. Moscow Math. Soc. 1 (1982), 1-48.

[16] MEYN, S.P., TWEEDIE, R.L., Markov Chains and Stochastic Stability, Springer Verlag, London, 1993.

[17] MEYN, S.P., TWEEDIE, R.L., Computable bounds for convergence rates of Markov chains, Annals of Applied Probability, **4**, 1994.

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

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

[20] TSITSIKLIS, J., Asynchronous stochastic approximation and Q-learning, Machine Learning 16 (1994), 195-202.

[21] WATKINS, C.J.C.H., DAYAN, P., Q-learning, Machine Learning 8 (1992) 279-292.