Machine Learning, 16, 185–202 (1994)

Asynchronous Stochastic Approximation and Q-Learning

JOHN N. TSITSIKLIS jnt@athena.mit.edu
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139

Editor: Richard Sutton

Abstract. We provide some general results on the convergence of a class of stochastic approximation algorithms and their parallel and asynchronous variants. We then use these results to study the Q-learning algorithm, a reinforcement learning method for solving Markov decision problems, and establish its convergence under conditions more general than previously available.

Keywords: Reinforcement learning, Q-learning, dynamic programming, stochastic approximation

1. Introduction

This paper is motivated by the desire to understand the convergence properties of Watkins' (1992) Q-learning algorithm. This is a reinforcement learning method that applies to Markov decision problems with unknown costs and transition probabilities; it may also be viewed as a direct adaptive control mechanism for controlled Markov chains (Sutton, Barto & Williams, 1992).

In Q-learning, transition probabilities and costs are unknown but information on them is obtained either by simulation or by experimenting with the system to be controlled; see (Barto, Bradtke & Singh, 1991) for a nice overview and discussion of the different ways that Q-learning can be applied. Q-learning uses simulation or experimental information to compute estimates of the expected cost-to-go (the value function of dynamic programming) as a function of the initial state. Furthermore, the algorithm is recursive and each new piece of information is used for computing an additive correction term to the old estimates. As these correction terms are random, Q-learning has the same general structure as stochastic approximation algorithms. In this paper, we combine ideas from the theory of stochastic approximation and from the convergence theory of parallel asynchronous algorithms, to develop the tools necessary to prove the convergence of Q-learning.

Stochastic approximation algorithms often have a structure such as

$$x_i := x_i + \alpha(F_i(x) - x_i + w_i)$$

where $x = (x_1, \ldots, x_n) \in \Re^n$, $F_1, \ldots, F_n$ are mappings from $\Re^n$ into $\Re$, w_i is a random noise term and α is a small, usually decreasing, stepsize. The Q-learning algorithm, to be described in more detail in Section 7, is precisely of this form, with the mapping $F = (F_1, \ldots, F_n)$ being closely related to the dynamic programming operator associated with a Markov decision problem.

The convergence of Q-learning has been proved by Watkins and Dayan (1992) for discounted Markov decision problems, as well as for undiscounted problems, under the assumption that all policies eventually lead to a zero-cost absorbing state. It was assumed, in addition, that the costs per stage are bounded random variables. The proof by Watkins and Dayan uses a very clever argument. On the other hand, it does not exploit the connection with stochastic approximation and runs into certain difficulties if some of the assumptions are weakened.

In this paper, we provide a new proof of the results of Watkins and Dayan (1992). In addition, our method of proof allows us to extend these results in several directions. In particular, we can prove convergence for undiscounted problems without assuming that all policies must lead to a zero-cost absorbing state; we allow the costs per stage to be unbounded random variables; we allow the decision on which action to simulate next to depend on past experience, and, finally, we consider the case of parallel implementation that allows for the use of outdated information, as in the asynchronous model of Bertsekas (1982) and Bertsekas and Tsitsiklis (1989).

To the best of our knowledge, the convergence of Q-learning does not follow from the available convergence theory for stochastic approximation algorithms. For this reason, our first step is to extend the classical theory. We briefly explain the technical reasons for doing so. A classical method for proving convergence of stochastic approximation is based on the supermartingale convergence theorem and exploits the expected reduction of a smooth Lyapunov function such as the Euclidean norm (Poljak & Tsypkin, 1973). However, for the case of Q-learning, we face the problem that the dynamic programming operator does not always have the necessary properties. Indeed, the dynamic programming operator, for discounted problems, is a contraction only with respect to the ℓ_∞ norm and the classical theory does not apply easily to this case; for undiscounted problems, it is not a contraction with respect to any norm. Another method for establishing convergence is based on "averaging" techniques that lead to an ordinary differential equation (Kushner & Clark, 1978). While this method is very powerful, some of the required assumptions might not be natural in certain contexts. For example, in the case of Q-learning, we would have to require that there exist well-defined average frequencies under which the different state-action pairs are being simulated. The method that we develop in this paper is based on the asynchronous convergence theory of Bertsekas (1982) and Bertsekas *et al.* (1989), suitably modified so as to allow for the presence of noise. There have been some earlier works on the convergence of asynchronous stochastic approximation methods, but their results do not apply to the models considered here: the results of Tsitsiklis, Bertsekas and Athans (1986) involve a smooth Lyapunov function, the results of Kushner and Yin (1987, 1987) rely on the averaging approach, and the assumptions of Li and Basar (1987) are too strong for our purposes.

During the writing of this paper, we learned that other authors (T. Jaakkola, M.I. Jordan, S. Singh) have also been developing convergence proofs for Q-learning that exploit the connection with stochastic approximation.

The rest of the paper is organized as follows. In Section 2, we present the algorithmic model to be employed and our assumptions, and state our general results on stochastic approximation algorithms. Section 3 contains an elementary result on stochastic approx-

imation that we will be using in our proofs. Sections 4, 5, and 6 contain the proofs of the results of Section 2. (Sections 3, 4, 5, and 6 can be skipped without loss of continuity.) Section 7 applies the theory of Section 2 to Q-learning. Section 8 contains some concluding comments.

2. Model and assumptions

In this section, we describe the algorithmic model to be employed and state the assumptions that will be imposed. The model is presented for the most general case which allows for a number of parallel processors who may be updating based on outdated information. In most respects, the model is the one in Chapter 6 of Bertsekas *et al.* (1989), except for the presence of noise.

The algorithm consists of noisy updates of a vector $x \in \Re^n$, for the purpose of solving a system of equations of the form $F(x) = x$. Here F is assumed to be a mapping from $\Re^n$ into itself. Let $F_1, \ldots, F_n : \Re^n \mapsto \Re$ be the corresponding component mappings; that is, $F(x) = (F_1(x), \ldots, F_n(x))$ for all $x \in \Re^n$.

Let $\mathcal{N}$ be the set of nonnegative integers. We employ a discrete "time" variable t, taking values in $\mathcal{N}$. This variable need not have any relation with real time; rather, it is used to index successive updates. Let $x(t)$ be the value of the vector x at time t and let $x_i(t)$ denote its ith component. Let T^i be an infinite subset of $\mathcal{N}$ indicating the set of times at which an update of x_i is performed. We assume that

$$x_i(t+1) = x_i(t), \qquad t \notin T^i. \tag{1}$$

Regarding the times that x_i is updated, we postulate an update equation of the form

$$x_i(t+1) = x_i(t) + \alpha_i(t)(F_i(x^i(t)) - x_i(t) + w_i(t)), \qquad t \in T^i. \tag{2}$$

Here, $\alpha_i(t)$ is a stepsize parameter belonging to $[0, 1]$, $w_i(t)$ is a noise term, and $x^i(t)$ is a vector of possibly outdated components of x. In particular, we assume that

$$x^i(t) = (x_1(\tau_1^i(t)), \ldots, x_n(\tau_n^i(t))), \qquad t \in T^i, \tag{3}$$

where each $\tau_j^i(t)$ is an integer satisfying $0 \leq \tau_j^i(t) \leq t$. If no information is outdated, we have $\tau_j^i(t) = t$ and $x^i(t) = x(t)$ for all t; the reader may wish to think primarily of this case. For an interpretation of the general case, see Bertsekas *et al.* (1989). In order to bring Eqs. (1) and (2) into a unified form, it is convenient to assume that $\alpha_i(t)$, $w_i(t)$, and $\tau_j^i(t)$ are defined for every i, j, and t, but that $\alpha_i(t) = 0$ and $\tau_j^i(t) = t$ for $t \notin T^i$.

We will now continue with our assumptions. All variables introduced so far ($x(t)$, $\tau_j^i(t)$, $\alpha_i(t)$, $w_i(t)$) are viewed as random variables defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and the assumptions deal primarily with the dependencies between these random variables. Our assumptions also involve an increasing sequence $\{\mathcal{F}(t)\}_{t=0}^{\infty}$ of subfields of $\mathcal{F}$. Intuitively, $\mathcal{F}(t)$ is meant to represent the history of the algorithm up to, and including the point at which the stepsizes $\alpha_i(t)$ for the tth iteration are selected, but just before the noise term $w_i(t)$ is generated. Also, the measure-theoretic terminology that "a random variable Z is

$\mathcal{F}(t)$-measurable" has the intuitive meaning that Z is completely determined by the history represented by $\mathcal{F}(t)$.

The first assumption, which is the same as the *total asynchronism* assumption of Bertsekas *et al.* (1989), guarantees that even though information can be outdated, any old information is eventually discarded.

Assumption 1. For any i and j, $\lim_{t\to\infty} \tau_j^i(t) = \infty$, with probability 1.

Our next assumption refers to the statistics of the random variables involved in the algorithm.

Assumption 2.

a) $x(0)$ is $\mathcal{F}(0)$-measurable;

b) For every i and t, $w_i(t)$ is $\mathcal{F}(t+1)$-measurable.

c) For every i, j, and t, $\alpha_i(t)$ and $\tau_j^i(t)$ are $\mathcal{F}(t)$-measurable.

d) For every i and t, we have $E[w_i(t) \mid \mathcal{F}(t)] = 0$.

e) There exist (deterministic) constants A and B such that

$$E[w_i^2(t) \mid \mathcal{F}(t)] \leq A + B \max_j \max_{\tau \leq t} |x_j(\tau)|^2, \qquad \forall i, t.$$

Assumption 2 allows for the possibility of deciding whether to update a particular component x_i at time t, based on the past history of the process. In this case, the stepsize $\alpha_i(t)$ becomes a random variable. However, part (c) of the assumption requires that the choice of the components to be updated must be made without anticipatory knowledge of the noise variables w_i that have not yet been realized. This is trivially satisfied if the sets T^i and the stepsizes $\alpha_i(t)$ are deterministic but this would be too restrictive as we argue in Section 7.

The next assumption concerns the stepsize parameters and is standard for stochastic approximation algorithms.

Assumption 3.

a) For every i,

$$\sum_{t=0}^{\infty} \alpha_i(t) = \infty, \qquad \text{w.p.1.} \tag{4}$$

b) There exists some (deterministic) constant C such that for every i,

$$\sum_{t=0}^{\infty} \alpha_i^2(t) \leq C, \qquad \text{w.p.1.} \tag{5}$$

Finally, we introduce a few alternative assumptions on the structure of the iteration mapping F. We first need some notation: if $x, y \in \Re^n$, the inequality $x \leq y$ is to be interpreted as $x_i \leq y_i$ for all i. Furthermore, for any positive vector $v = (v_1, \ldots, v_n)$, we define a norm $\|\cdot\|_v$ on $\Re^n$ by letting

$$\|x\|_v = \max_i \frac{|x_i|}{v_i}, \qquad x \in \Re^n. \tag{6}$$

Notice that in the special case where all components of v are equal to 1, $\|\cdot\|_v$ is the same as the maximum norm $\|\cdot\|_\infty$.

Assumption 4.

a) The mapping F is monotone; that is, if $x \leq y$, then $F(x) \leq F(y)$.

b) The mapping F is continuous.

c) The mapping F has a unique fixed point x^*.

d) If $e \in \Re^n$ is the vector with all components equal to 1, and r is a positive scalar, then

$$F(x) - re \leq F(x - re) \leq F(x + re) \leq F(x) + re. \tag{7}$$

Assumption 5. There exists a vector $x^* \in \Re^n$, a positive vector v, and a scalar $\beta \in [0, 1)$, such that

$$\|F(x) - x^*\|_v \leq \beta \|x - x^*\|_v, \qquad \forall x \in \Re^n. \tag{8}$$

Assumption 6. There exists a positive vector v, a scalar $\beta \in [0, 1)$, and a scalar D such that

$$\|F(x)\|_v \leq \beta \|x\|_v + D, \qquad \forall x \in \Re^n. \tag{9}$$

We now state the main results of this paper. Theorem 1 provides conditions for $x(t)$ to be bounded. Theorems 2 and 3 deal with convergence under Assumptions 4 and 5, respectively.

THEOREM 1. *Let Assumptions* 1, 2, 3, *and* 6 *hold. Then, the sequence* $x(t)$ *is bounded, with probability* 1.

THEOREM 2. *Let Assumptions* 1, 2, 3, *and* 4 *hold. Furthermore, suppose that* $x(t)$ *is bounded with probability* 1. *Then,* $x(t)$ *converges to* x^*, *with probability* 1.

THEOREM 3. *Let Assumptions* 1, 2, 3, *and* 5 *hold. Then,* $x(t)$ *converges to* x^*, *with probability* 1.

3. Preliminaries

In this section, we state a well known result that will be needed later.

LEMMA 1. *Let $\{\mathcal{F}(t)\}$ be an increasing sequence of σ-fields. For each t, let $\alpha(t)$, $w(t-1)$, and $B(t)$ be $\mathcal{F}(t)$-measurable scalar random variables. Let C be a deterministic constant. Suppose that the following hold with probability* 1*:*

a) $E[w(t) \mid \mathcal{F}(t)] = 0$;

b) $E[w^2(t) \mid \mathcal{F}(t)] \leq B(t)$;

c) $\alpha(t) \in [0, 1]$;

d) $\sum_{t=0}^{\infty} \alpha(t) = \infty$;

e) $\sum_{t=0}^{\infty} \alpha^2(t) \leq C$.

Suppose that the sequence $\{B(t)\}$ is bounded with probability 1*. Let $W(t)$ satisfy the recursion*

$$W(t+1) = (1 - \alpha(t))W(t) + \alpha(t)w(t).$$

Then $\lim_{t\to\infty} W(t) = 0$, with probability 1.

Proof: For the case where the sequence $B(t)$ is bounded by a deterministic constant, we are dealing with the classical stochastic gradient algorithm for minimizing a quadratic cost function and its convergence is well known; for example, see Poljak *et al.* (1973).

For every positive integer k, we define $\tau_k = \min\{t \geq 0 \mid B(t) \geq k\}$, with the understanding that $\tau_k = \infty$ if $B(t) < k$ for all k. We define $w^k(t) = w(t)$ if $t < \tau_k$ and $w^k(t) = 0$, otherwise. Let $W^k(t)$ be defined by letting $W^k(0) = W(0)$ and $W^k(t+1) = (1 - \alpha(t))W^k(t) + \alpha(t)w^k(t)$. Since $E[(w^k(t))^2 \mid \mathcal{F}(t)] \leq k$ for all t, we see that $W^k(t)$ converges to zero, with probability 1, for every k. On the other hand, since $B(t)$ is bounded, there exists some k such that $W^k(t) = W(t)$ for all t, and this implies that $W(t)$ also converges to zero. ■

4. Proof of Theorem 1

In this and in all subsequent proofs, we assume that we have already discarded a suitable set of measure zero, so that we do not need to keep repeating the qualification "with probability 1."

We assume that all components of the vector v in Assumption 6 are equal to 1. (The case of a general positive weighting vector v can be reduced to this special case by a suitable coordinate scaling.) In particular, there exists some $\beta \in [0, 1)$ and some D such that

$$\max_i |F_i(x)| \leq \beta \max_i |x_i| + D, \qquad \forall x \in \Re^n. \tag{10}$$

It follows from Eq. (10) that there exist $\gamma \in [0,1)$ and $G_0 > 0$ such that

$$|F_i(x)| \leq \gamma \max\left\{ \max_j |x_j|, G_0 \right\}, \qquad \forall x \in \Re^n, \ \forall i. \tag{11}$$

(Any $\gamma \in [0,1)$ and $G_0 > 0$ satisfying $\beta G_0 + D \leq \gamma G_0$ will do.) Let us also fix $\epsilon > 0$ so that $\gamma(1+\epsilon) = 1$.

Let

$$M(t) = \max_{\tau \leq t} \|x(\tau)\|_\infty. \tag{12}$$

We define a sequence $\{G(t)\}$, recursively, as follows. Let $G(0) = \max\{M(0), G_0\}$. Assuming that $G(t)$ has already been defined, let $G(t+1) = G(t)$ if $M(t+1) \leq (1+\epsilon)G(t)$. If $M(t+1) > (1+\epsilon)G(t)$, then let $G(t+1) = G_0(1+\epsilon)^k$ where k is chosen so that

$$G_0(1+\epsilon)^{k-1} < M(t+1) \leq G_0(1+\epsilon)^k.$$

A key consequence of our definitions is that

$$M(t) \leq (1+\epsilon)G(t), \qquad \forall t \geq 0, \tag{13}$$

and

$$M(t) \leq G(t) \qquad \text{if } G(t-1) < G(t). \tag{14}$$

It is easily seen that $M(t)$ and $G(t)$ are $\mathcal{F}(t)$-measurable.

In order to motivate our next step, note that as long as there is possibility that $x(t)$ is unbounded, $E[w_i^2(t) \mid \mathcal{F}(t)]$ could also be unbounded (cf. Assumption 2) and results such as Lemma 1 are inapplicable. To circumvent this difficulty, we will work in terms of a suitably scaled version of $w_i(t)$ whose conditional variance will be bounded.

We define

$$\tilde{w}_i(t) = \frac{w_i(t)}{G(t)}, \qquad \forall t \geq 0,$$

which is $\mathcal{F}(t+1)$-measurable. Assumption 2 implies that

$$E[\tilde{w}_i(t) \mid \mathcal{F}(t)] = \frac{E[w_i(t) \mid \mathcal{F}(t)]}{G(t)} = 0,$$

and

$$\begin{aligned} E[\tilde{w}_i^2(t) \mid \mathcal{F}(t)] &= \frac{E[w_i^2(t) \mid \mathcal{F}(t)]}{G^2(t)} \leq \frac{A + BM^2(t)}{G^2(t)} \\ &\leq \frac{A + B(1+\epsilon)^2 G^2(t)}{G^2(t)} \leq K, \qquad \forall t \geq 0, \end{aligned}$$

where K is some deterministic constant.

For any i and $t_0 \geq 0$, we define $\tilde{W}_i(t_0;t_0) = 0$ and

$$\tilde{W}_i(t+1;t_0) = (1-\alpha_i(t))\tilde{W}_i(t;t_0) + \alpha_i(t)\tilde{w}_i(t), \qquad t \geq t_0.$$

LEMMA 2. *For every $\delta > 0$, there exists some T such that $|\tilde{W}_i(t;t_0)| \leq \delta$, for every t and t_0 satisfying $T \leq t_0 \leq t$.*

Proof: By Lemma 1, we obtain $\lim_{t\to\infty} \tilde{W}_i(t;0) = 0$. For every $t \geq t_0$, we have

$$\tilde{W}_i(t;0) = \left[\prod_{\tau=t_0}^{t-1}(1-\alpha_i(\tau))\right]\tilde{W}_i(t_0;0) + \tilde{W}_i(t;t_0),$$

which implies that $|\tilde{W}_i(t;t_0)| \leq |\tilde{W}_i(t;0)| + |\tilde{W}_i(t_0;0)|$. The result follows by letting T be large enough so that $|\tilde{W}_i(t;0)| \leq \delta/2$ for every $t \geq T$. ■

Suppose now, in order to derive a contradiction, that $x(t)$ is unbounded. Then, Eqs. (12) and (13) imply that $G(t)$ converges to infinity, and Eq. (14) implies that the inequality $M(t) \leq G(t)$ holds for infinitely many different values of t. In view of Lemma 2, we conclude that there exists some t_0 such that $M(t_0) \leq G(t_0)$ and

$$|\tilde{W}_i(t;t_0)| \leq \epsilon, \qquad \forall t \geq t_0, \ \forall i. \tag{15}$$

The lemma that follows derives a contradiction to the unboundedness of $G(t)$ and concludes the proof of the theorem.

LEMMA 3. *Suppose that $x(t)$ is unbounded. Then, for every $t \geq t_0$, we have $G(t) = G(t_0)$. Furthermore, for every i we have*

$$\begin{aligned} -G(t_0)(1+\epsilon) &\leq -G(t_0) + \tilde{W}_i(t;t_0)G(t_0) \leq x_i(t) \\ &\leq G(t_0) + \tilde{W}_i(t;t_0)G(t_0) \leq G(t_0)(1+\epsilon). \end{aligned}$$

Proof: The proof proceeds by induction on t. For $t = t_0$, the result is obvious from $|x_i(t_0)| \leq M(t_0) \leq G(t_0)$ and $\tilde{W}_i(t_0;t_0) = 0$. Suppose that the result is true for some t. We then use the induction hypothesis and Eq. (11) to obtain

$$\begin{aligned} x_i(t+1) &= (1-\alpha_i(t))x_i(t) + \alpha_i(t)F_i(x^i(t)) + \alpha_i(t)w_i(t) \\ &\leq (1-\alpha_i(t))(G(t_0) + \tilde{W}_i(t;t_0)G(t_0)) + \alpha_i(t)\gamma G(t_0)(1+\epsilon) \\ &\quad +\alpha_i(t)\tilde{w}_i(t)G(t_0) \\ &= G(t_0) + ((1-\alpha_i(t))\tilde{W}_i(t;t_0) + \alpha_i(t)\tilde{w}_i(t))G(t_0) \\ &= G(t_0) + \tilde{W}_i(t+1;t_0)G(t_0). \end{aligned}$$

A symmetrical argument also yields $-G(t_0) + \tilde{W}_i(t+1;t_0)G(t_0) \leq x_i(t+1)$. Using Eq. (15), we obtain $|x_i(t+1)| \leq G(t_0)(1+\epsilon)$ which also implies that $G(t+1) = G(t_0)$. ■

5. Proof of Theorem 2

Recall that e stands for the vector with all components equal to 1. Let r be a large enough scalar so that $x^* - re \leq x(t) \leq x^* + re$ for all t. [Such a scalar exists by the boundedness assumption on $x(t)$ but is a random variable because $\sup_t \|x(t)\|_\infty$ could be different for different sample paths.] Let $L^0 = (L_1^0, \ldots, L_n^0) = x^* - re$ and $U^0 = (U_1^0, \ldots, U_n^0) = x^* + re$. Let us define two sequences $\{U^k\}$ and $\{L^k\}$ in terms of the recursions

$$U^{k+1} = \frac{U^k + F(U^k)}{2}, \qquad k \geq 0, \tag{16}$$

and

$$L^{k+1} = \frac{L^k + F(L^k)}{2}, \qquad k \geq 0. \tag{17}$$

LEMMA 4. *For every $k \geq 0$, we have*

$$F(U^k) \leq U^{k+1} \leq U^k,$$

and

$$F(L^k) \geq L^{k+1} \geq L^k.$$

Proof: The proof is by induction on k. Notice that, by Assumption 4(d) and the fixed point property of x^*, we have $F(U^0) = F(x^* + re) \leq F(x^*) + re = x^* + re = U^0$. Using the definition of U^1, we obtain $F(U^0) \leq U^1 \leq U^0$. Suppose that the result is true for some k. The inequality $U^{k+1} \leq U^k$ and the monotonicity of F yield $F(U^{k+1}) \leq F(U^k)$. Equation (16) then implies that $U^{k+2} \leq U^{k+1}$. Furthermore, since U^{k+2} is the average of $F(U^{k+1})$ and U^{k+1}, we also obtain $F(U^{k+1}) \leq U^{k+2}$. The inequalities for L^k follow by a symmetrical argument. ■

LEMMA 5. *The sequences $\{U^k\}$ and $\{L^k\}$ converge to x^*.*

Proof: We first prove, by induction, that $U^k \geq x^*$ for all k. This is true for U^0, by definition. Suppose that $U^k \geq x^*$. Then, by monotonicity, $F(U^k) \geq F(x^*) = x^*$, from which the inequality $U^{k+1} \geq x^*$ follows. Therefore, the sequence $\{U^k\}$ is bounded below. Since this sequence is monotonic (Lemma 4), it converges to some limit U. Using the continuity of F, we must have $U = (U + F(U))/2$, which implies that $U = F(U)$. Since x^* was assumed to be the unique fixed point of F, it follows that $U = x^*$. Convergence of L^k to x^* follows from a symmetrical argument. ■

We will now show that for every k, there exists some time t_k such that

$$L^k \leq x(t) \leq U^k, \qquad \forall t \geq t_k. \tag{18}$$

(The value of t_k will not be the same for different sample paths and is therefore a random variable.) Once this is proved, the convergence of $x(t)$ to x^* follows from Lemma 5. For $k = 0$, Eq. (18) is certainly true, with $t_0 = 0$, because of the way that U^0 and L^0 were defined. We continue by induction on k. We fix some k and assume that there exists some t_k so that Eq. (18) holds. Let t'_k be such that for every $t \geq t'_k$, and every i, j, we have $\tau_j^i(t) \geq t_k$. Such a t'_k exists because of Assumption 1. (For the case where no outdated values of the components of x are used and $\tau_j^i(t) = t$, we may simply let $t'_k = t_k$.) In particular, we have

$$L^k \leq x^i(t) \leq U^k, \qquad \forall t \geq t'_k. \tag{19}$$

Let $W_i(0) = 0$ and

$$W_i(t+1) = (1 - \alpha_i(t))W_i(t) + \alpha_i(t)w_i(t).$$

We then have $\lim_{t\to\infty} W_i(t) = 0$ (cf. Lemma 1). For any time t_0, we also define $W_i(t_0; t_0) = 0$ and

$$W_i(t+1; t_0) = (1 - \alpha_i(t))W_i(t; t_0) + \alpha_i(t)w_i(t), \qquad t \geq t_0. \tag{20}$$

Following the same argument as in the proof of Lemma 2, we see that for every t_0, we have $\lim_{t\to\infty} W_i(t; t_0) = 0$.

We also define a sequence $X_i(t)$, $t \geq t'_k$, by letting $X_i(t'_k) = U_i^k$ and

$$X_i(t+1) = (1 - \alpha_i(t))X_i(t) + \alpha_i(t)F_i(U^k), \qquad t \geq t'_k. \tag{21}$$

LEMMA 6.

$$x_i(t) \leq X_i(t) + W_i(t; t'_k), \qquad \forall t \geq t'_k.$$

Proof: The proof proceeds by induction on t. For $t = t'_k$, Eq. (18) yields $x_i(t'_k) \leq U_i^k$ and, by definition, we have $U_i^k = X_i(t'_k) + W_i(t'_k; t'_k)$. Suppose that the result is true for some t. Then, Eqs. (2), (19), (21), and (20) imply that

$$\begin{aligned} x_i(t+1) &\leq (1 - \alpha_i(t))(X_i(t) + W_i(t; t'_k)) + \alpha_i(t)F_i(U^k) + \alpha_i(t)w_i(t) \\ &= X_i(t+1) + W_i(t+1; t'_k). \end{aligned}$$

■

Let δ_k be equal to the minimum of $(U_i^k - F_i(U^k))/4$, where the minimum is taken over all i for which $U_i^k - F_i(U^k)$ is positive. Clearly, δ_k is well-defined and positive unless $U^k = F(U^k)$. But in the latter case, we must have $U^k = x^* = U^{k+1}$ and the inequality $x(t) \leq U^k$ implies that $x(t) \leq U^{k+1}$ and there is nothing more to be proved. We therefore assume that δ_k is well-defined and positive.

Let t''_k be such that $t''_k \geq t'_k$,

$$\prod_{\tau=t'_k}^{t''_k-1} (1 - \alpha_i(\tau)) \leq \frac{1}{4}$$

and

$$W_i(t; t'_k) \leq \delta_k$$

for all $t \geq t''_k$ and all i. Such a t''_k exists because Assumption 3(a) implies that

$$\prod_{\tau=t'_k}^{\infty} (1 - \alpha_i(\tau)) = 0$$

and because $W_i(t; t'_k)$ converges to zero, as discussed earlier.

LEMMA 7. *We have $x_i(t) \leq U_i^{k+1}$, for all i and $t \geq t''_k$.*

Proof: Fix some i. If $U_i^{k+1} = U_i^k$, the inequality $x_i(t) \leq U_i^{k+1}$ follows from Eq. (18). We therefore concentrate on the case where $U_i^{k+1} < U_i^k$. Equation (21) and the relation $X_i(t'_k) = U_i^k$ imply that $X_i(t)$ is a convex combination of U_i^k and $F_i(U^k)$. Furthermore, the coefficient of U_i^k is equal to $\prod_{\tau=t'_k}^{t-1}(1 - \alpha_i(\tau))$, which is no more than $1/4$ for $t \geq t''_k$. It follows that

$$X_i(t) \leq \frac{1}{4}U_i^k + \frac{3}{4}F_i(U^k) = \frac{1}{2}U_i^k + \frac{1}{2}F_i(U^k) - \frac{1}{4}(U_i^k - F_i(U^k)) \leq U_i^{k+1} - \delta_k.$$

This inequality, together with the inequality $W_i(t; t'_k) \leq \delta_k$ and Lemma 6, imply that $x_i(t) \leq U_i^{k+1}$ for all $t \geq t''_k$. ∎

By an entirely symmetrical argument, we can also establish that $x_i(t) \geq L_i^{k+1}$ for all t greater than some t'''_k. This proves Eq. (18) for $k + 1$, concludes the induction, and completes the proof of the theorem.

6. Proof of Theorem 3

Without loss of generality, we assume that $x^* = 0$; this can be always accomplished by translating the origin of the coordinate system. Furthermore, as in the proof of Theorem 1, we assume that all components of the vector v in Assumption 5 are equal to 1. Notice that Theorem 1 applies and establishes that $x(t)$ is bounded.

Theorem 1 states that there exists some (generally random) D_0 such that $\|x(t)\|_\infty \leq D_0$, for all t. Fix some $\epsilon > 0$ such that $\beta(1 + 2\epsilon) < 1$. We define

$$D_{k+1} = \beta(1 + 2\epsilon)D_k, \qquad k \geq 0.$$

Clearly, D_k converges to zero.

Suppose that there exists some time t_k such that $\|x(t)\|_\infty \leq D_k$ for all $t \geq t_k$. We will show that this implies that there exists some time t_{k+1} such that $\|x(t)\|_\infty \leq D_{k+1}$ for all $t \geq t_{k+1}$. This will complete the proof of convergence of $x(t)$ to zero.

Let $W_i(0) = 0$ and

$$W_i(t+1) = (1-\alpha_i(t))W_i(t) + \alpha_i(t)w_i(t). \tag{22}$$

We then have $\lim_{t\to\infty} W_i(t) = 0$, (cf. Lemma 1). For any time t_0, we also define $W_i(t_0;t_0) = 0$ and

$$W_i(t+1;t_0) = (1-\alpha_i(t))W_i(t;t_0) + \alpha_i(t)w_i(t), \qquad t \geq t_0. \tag{23}$$

Following the same argument as in the proof of Lemma 2, we see that for every $\delta > 0$, there exists some T such that $|W_i(t;t_0)| \leq \delta$ for all $t_0 \geq T$ and $t \geq t_0$.

Let $\tau_k \geq t_k$ be such that $|W_i(t;\tau_k)| \leq \beta\epsilon D_k$ and $\|x^i(t)\|_\infty \leq D_k$ for all $t \geq \tau_k$ and all i. As discussed earlier, the first requirement will be eventually satisfied. The same is true for the second requirement, because of Assumption 1. We define $Y_i(\tau_k) = D_k$ and

$$Y_i(t+1) = (1-\alpha_i(t))Y_i(t) + \alpha_i(t)\beta D_k, \qquad t \geq \tau_k. \tag{24}$$

LEMMA 8.

$$-Y_i(t) + W_i(t;\tau_k) \leq x_i(t) \leq Y_i(t) + W_i(t;\tau_k), \qquad \forall t \geq \tau_k. \tag{25}$$

Proof: We use induction on t. Since $Y_i(\tau_k) = D_k$ and $W_i(\tau_k;\tau_k) = 0$, the result is true for $t = \tau_k$. Suppose that Eq. (25) holds for some t. We then have

$$\begin{aligned} x_i(t+1) &\leq (1-\alpha_i(t))(Y_i(t) + W_i(t;\tau_k)) + \alpha_i(t)\beta D_k + \alpha_i(t)w_i(t) \\ &= Y_i(t+1) + W_i(t+1;\tau_k). \end{aligned}$$

A symmetrical argument yields $-Y_i(t+1) + W_i(t+1;\tau_k) \leq x_i(t+1)$ and the inductive proof is complete. ■

It is evident from Eq. (24) and Assumption 3(a) that $Y_i(t)$ converges to βD_k as $t \to \infty$. This fact, together with Eq. (25) yields

$$\limsup_{t\to\infty} |x_i(t)| \leq \beta(1+\epsilon)D_k < D_{k+1},$$

and the proof is complete.

7. The convergence of Q-learning

We consider a Markov decision problem defined on a finite state space S. For every state $i \in S$, there is a finite set $U(i)$ of possible control actions and a set of nonnegative scalars $p_{ij}(u)$, $u \in U(i)$, $j \in S$, such that $\sum_{j\in S} p_{ij}(u) = 1$ for all $u \in U(i)$. The scalar $p_{ij}(u)$ is interpreted as the probability of a transition to j, given that the current state is i and control u is applied. Furthermore, for every state i and control u, there is a random variable c_{iu} which represents the one-stage cost if action u is applied at state i. We assume that the variance of c_{iu} is finite for every i and $u \in U(i)$.

A *stationary policy* is a function π defined on S such that $\pi(i) \in U(i)$ for all $i \in S$. Given a stationary policy, we obtain a discrete-time Markov chain $s^\pi(t)$ with transition probabilities

$$\Pr(s^\pi(t+1) = j \mid s^\pi(t) = i) = p_{ij}(\pi(i)).$$

Let $\beta \in [0,1]$ be a discount factor. To any stationary policy π and initial state i, the cost-to-go V_i^π is defined by

$$V_i^\pi = \limsup_{T\to\infty} E\left[\sum_{t=0}^{T} \beta^t c_{s^\pi(t),\pi(s^\pi(t))} \mid s^\pi(0) = i\right].$$

The optimal cost-to-go function V^* is defined by

$$V_i^* = \inf_\pi V_i^\pi, \qquad i \in S.$$

The Markov decision problem is to evaluate the function V^*. (Once this is done, an optimal policy is easily determined.)

Markov decision problems are easiest when the discount factor β is strictly smaller than 1. For the undiscounted case ($\beta = 1$), we will assume throughout that there is a cost-free state, say state 1, which is absorbing; that is, $p_{11}(u) = 1$ and $c_{1u} = 0$ for all $u \in U(1)$. The objective is then to reach that state at minimum expected cost. We say that a stationary policy is *proper* if the probability of being at the absorbing state converges to 1 as time converges to infinity; otherwise, we say that the policy is *improper*. The following assumption is natural for undiscounted problems.

Assumption 7.

a) There exists at least one proper stationary policy.

b) Every improper stationary policy yields infinite expected cost for at least one initial state.

We define the dynamic programming operator $T{:}\Re^{|S|} \mapsto \Re^{|S|}$, with components T_i, by letting

$$T_i(V) = \min_{u\in U(i)} \left\{ E[c_{iu}] + \beta \sum_{j\in S} p_{ij}(u) V_j \right\}.$$

It is well known that if $\beta < 1$, then T is a contraction with respect to the norm $\|\cdot\|_\infty$ and V^* is its unique fixed point. If $\beta = 1$, then T is not, in general, a contraction. However, it is still true that the set $\{V \in \Re^{|S|} \mid V_1 = 0\}$ contains a unique fixed point of T and this fixed point is equal to V^*, as long as Assumption 7 holds (Bertsekas *et al.*, 1989, 1991).

The Q-learning algorithm is a method for computing V^* based on a reformulation of the Bellman equation $V^* = T(V^*)$. We provide a brief description of the algorithm. Let $P = \{(i,u) \mid i \in S,\ u \in U(i)\}$ be the set of all possible state-action pairs and let n be its cardinality. We use a discrete index variable t in order to count iterations. After t

iterations, we have a vector $Q(t) \in \Re^n$, with components $Q_{iu}(t)$, $(i,u) \in P$, which we update according to the formula

$$Q_{iu}(t+1) = Q_{iu}(t) + \alpha_{iu}(t)\left[c_{iu} + \beta \min_{v \in U(s(i,u))} Q_{s(i,u),v}(t) - Q_{iu}(t)\right]. \tag{26}$$

Here, each $\alpha_{iu}(t)$ is a nonnegative stepsize coefficient which is set to zero for those $(i,u) \in P$ for which Q_{iu} is not to be updated at the current iteration. Furthermore, c_{iu} is a random sample of the immediate cost if action u is applied at state i. Finally, $s(i,u)$ is a random successor state which is equal to j with probability $p_{ij}(u)$. It is understood that all random samples that are drawn in the course of the algorithm are drawn independently.

We now argue that the Q-learning algorithm has the form of Eq. (2). Let F be the mapping from $\Re^n$ into itself with components F_{iu} defined by

$$F_{iu}(Q) = E[c_{iu}] + \beta E\left[\min_{v \in U(s(i,u))} Q_{s(i,u),v}\right], \tag{27}$$

and note that

$$E\left[\min_{v \in U(s(i,u))} Q_{s(i,u),v}\right] = \sum_{j \in S} p_{ij}(u) \min_{v \in U(j)} Q_{jv}.$$

It is not hard to see that if a vector Q is a fixed point of F, then the vector with components $V_i = \min_{u \in U(i)} Q_{iu}$ is a fixed point of T. In view of Eq. (27), Eq. (26) can be written as

$$Q_{iu}(t+1) = Q_{iu}(t) + \alpha_{iu}(t)[F_{iu}(Q(t)) - Q_{iu}(t) + w_{iu}(t)],$$

where

$$w_{iu}(t) = c_{iu} - E[c_{iu}] + \min_{v \in U(s(i,u))} Q_{s(i,u),v}(t) - E\left[\min_{v \in U(s(i,u))} Q_{s(i,u),v}(t) \mid \mathcal{F}(t)\right]. \tag{28}$$

[The expectation in the expression $E[\min_{v \in U(s(i,u))} Q_{s(i,u),v}(t) \mid \mathcal{F}(t)]$ is with respect to $s(i,u)$.]

We now discuss the meaning of the various assumptions of Section 2 in the context of the Q-learning algorithm. Assumption 1 is satisfied in the special case where $\tau_j^i(t) = t$, which is what was implicitly assumed in Eq. (26), but can be also satisfied even if we allow for outdated information. The latter case could be of interest if the Q-learning algorithm were to be implemented in a massively parallel machine, with different processors carrying out updates of different components of Q, possibly using outdated information on some of the components of Q.

Regarding Assumption 2, we let $\mathcal{F}(t)$ represent the history of the algorithm during the first t iterations. Parts (a) and (b) of the assumption are then automatically valid. Part (c) is quite natural: in particular, it assumes that the required samples are generated after we decide which components to update during the current iteration. Note, however, that it allows this decision to be made on the basis of past experience, past explorations, or by

simply following a simulated trajectory. This point was not appreciated in earlier proofs. In particular, the proof of Watkins *et al.* (1992) implicitly assumes that the coefficients $\alpha_{iu}(t)$ are deterministic and, therefore, does not allow for experience-driven exploration or the use of simulated trajectories. Part (d) is automatic from Eq. (28). The conditional variance of $\min_{v\in U(s(i,u))} Q_{s(i,u),v}$, given $\mathcal{F}(t)$, is bounded above by the largest possible value that this random variable could take, which is $\max_{j\in S} \max_{v\in U(j)} Q^2_{jv}(t)$. We then take the conditional variance of both sides of Eq. (28), to obtain

$$E\big[w^2_{iu}(t) \mid \mathcal{F}(t)\big] \leq \mathrm{Var}(c_{iu}) + \max_{j\in S} \max_{v\in U(j)} Q^2_{jv}(t)$$

and part (e) is also satisfied. [We have assumed here that c_{iu} is independent from $s(i,u)$. If it is not, the right-hand side in the last inequality must be multiplied by 2, but the conclusion does not change.]

Assumption 3 needs to be imposed on the stepsizes employed by the Q-learning algorithm. In particular, it requires that every state-action pair (i,u) is simulated an infinite number of times.

For discounted problems ($\beta < 1$), Eq. (27) easily yields

$$|F_{iu}(Q) - F_{iu}(Q')| \leq \beta \max_{j\in S,\ v\in U(j)} |Q_{jv} - Q'_{jv}|, \qquad \forall Q, Q'.$$

In particular, F is a contraction mapping, with respect to the maximum norm $\|\cdot\|_\infty$ and Assumption 5 is satisfied. In particular, Theorem 3 establishes convergence.

For undiscounted problems ($\beta = 1$), our assumptions on the absorbing state 1 imply that the update equation for Q_{1u} degenerates to $Q_{1u}(t+1) = Q_{1u}(t)$, for all t. We will be assuming in the sequel, that Q_{1u} is initialized at zero. This leads to an equivalent description of the algorithm in which the mappings F_{iu} of Eq. (27) are replaced by mappings $\tilde{F}_{iu}$ satisfying $\tilde{F}_{iu} = F_{iu}$ if $i \neq 1$ and $\tilde{F}_{1u}(Q) = 0$ for all $u \in U(1)$ and $Q \in \Re^n$. Let us consider the special case where every policy is proper. It is then known (Bertsekas *et al.*, 1989, 1991) that there exists a vector $v > 0$ such that T is a contraction with respect to the norm $\|\cdot\|_v$. In fact, a close examination of the proof of Bertsekas *et al.* (1989) (pp. 325–327) shows that the proof is easily extended to show that the mapping $\tilde{F}$ (with components $\tilde{F}_{iu}$) is a contraction with respect to the norm $\|\cdot\|_z$, where $z_{iu} = v_i$ for every $u \in U(i)$. Convergence then follows again from Theorem 3.

Let us now keep assuming that $\beta = 1$, but remove the assumption that all policies are proper; we only impose Assumption 7. It is then known that the dynamic programming operator T satisfies Assumption 4 (Bertsekas *et al.*, 1989, 1991) and this implies easily that $\tilde{F}$ satisfies the same assumption. However, in order to invoke Theorem 2, we must also guarantee that $Q(t)$ is bounded. We discuss later how this can be accomplished.

We summarize our discussion in the following result.

THEOREM 4. *Consider the Q-learning algorithm and let* $Q^*_{iu} = E[c_{iu}] + \beta \sum_j p_{ij}(u) V^*_j$. *Then,* $Q_{iu}(t)$ *converges to* Q^*_{iu}, *with probability* 1, *for every* i *and* u, *in each of the following cases:*

a) *If* $\beta < 1$.

b) *If* $\beta = 1$, $Q_{1u}(0) = 0$, *and all policies are proper.*

c) *If* $\beta = 1$, *Assumption 7 holds,* $Q_{1u}(0) = 0$, *and if* $Q(t)$ *is guaranteed to be bounded, with probability* 1.

We conclude by discussing further how to guarantee boundedness for undiscounted problems. One option is to enforce boundedness artifially using the "projection" method (Kushner *et al.*, 1978). With this method, one projects the vector $Q(t)$ onto a given bounded set B whenever $Q(t)$ becomes too large. We only need to choose a large enough set B so that the fixed point of $\tilde{F}$ is certain to be contained in B. This requires some prior knowledge on the Markov decision problem being solved, but such knowledge is often available. Since the projection method is a general purpose method, there is not much new that can be said here and we do not provide any further details.

The lemma that follows covers another case in which boundedness is guaranteed.

LEMMA 9. *Suppose that* $\beta = 1$, *Assumption 7 holds, and* $Q_{1u}(0) = 0$. *Furthermore, suppose that all one-stage costs* c_{iu} *are nonnegative with probability* 1, *and that* $Q(0) \geq 0$. *Then, the sequence* $\{Q(t)\}$ *generated by the Q-learning algorithm is bounded with probability* 1.

Proof: Given the assumption that $c_{iu} \geq 0$, it is evident from Eq. (26) that if the algorithm is initialized with a nonnegative vector $Q(0) \geq 0$, then $Q(t) \geq 0$ for all t. This establishes a lower bound on each $Q_{iu}(t)$.

Let us now fix a proper policy π. We define a mapping F^π, with components F^π_{iu}, by letting $F^\pi_{1u}(Q) = 0$, for every $u \in U(1)$, and

$$F^\pi_{iu}(Q) = E[c_{iu}] + \sum_{j \neq 1} p_{ij}(\pi(i)) Q_{j,\pi(j)}, \qquad i \neq 1,\ u \in U(i).$$

Let P be the matrix whose (i,j)th entry is equal to $p_{ij}(\pi(i))$, for $i, j \in S$ and $i, j \neq 1$. Since policy π is proper, we see that P^t converges to 0 as t converges to infinity. Since P is a nonnegative matrix, the Perron-Frobenius theorem implies that there exists a positive vector w and some $\gamma \in [0,1)$ such that

$$\sum_{j \neq 1} p_{ij}(\pi(i)) w_j \leq \gamma w_i, \qquad \forall i \neq 1. \tag{29}$$

Therefore, for any vectors Q and Q', we have

$$\frac{|F^\pi_{iu}(Q) - F^\pi_{iu}(Q')|}{w_i} \leq \frac{1}{w_i} \sum_{j \neq 1} p_{ij}(\pi(i)) w_j \frac{|Q_{ju} - Q'_{ju}|}{w_j} \leq \gamma \max_{j \neq 1} \frac{|Q_{ju} - Q'_{ju}|}{w_j}.$$

We conclude that there exists a positive vector v such that $\|F^\pi(Q) - Q^\pi\|_v \leq \gamma \|Q - Q^\pi\|_v$ for all vectors Q, where Q^π is the unique fixed point of F^π. (In particular, $v_{iu} = w_i$ for every $u \in U(i)$.) Compare now the definition (27) with the definition of F^π to conclude that

for every vector $Q \geq 0$, we have $\tilde{F}(Q) \leq F^{\pi}(Q)$. Using this and the triangle inequality, we obtain

$$\begin{aligned}\|\tilde{F}(Q)\|_v \leq \|F^{\pi}(Q)\|_v &\leq \|F^{\pi}(Q) - F^{\pi}(Q^{\pi})\|_v + \|F^{\pi}(Q^{\pi})\|_v \\ &\leq \gamma\|Q - Q^{\pi}\|_v + \|Q^{\pi}\|_v \leq \gamma\|Q\|_v + 2\|Q^{\pi}\|_v.\end{aligned}$$

This establishes that $\tilde{F}$ satisfies Assumption 6 and the result follows from Theorem 1. ■

We close this section by pointing out that the interest in Markov decision problems for which not every stationary policy is proper is not purely academic. Consider a directed graph in which the length of every arc (i, j) is a nonnegative random variable c_{ij}. We may then be interested in the problem of finding a path, from a given origin to a given destination, with the smallest possible expected arc length. If the expected arc costs $E[c_{ij}]$ were known, this would simply be a shortest path problem. On the other hand, if the statistics of the arc costs are unknown, Q-learning can be used because shortest path problems are special cases of Markov decision problems in which every $p_{ij}(u)$ is either zero or 1. Notice that not every policy will be proper, in general: for example, a policy may choose a cycle that does not go through the destination and cycle forever around that cycle. On the other hand, Assumption 7 is equivalent to requiring that every cycle have positive expected costs and such an assumption is pretty much necessary for the shortest path to be well-posed. Once this assumption is imposed, Theorem 4 and Lemma 9 imply that Q-learning will converge.

8. Concluding remarks

We have established the convergence of Q-learning under fairly general conditions. The only technical problem that remains open is whether Assumption 7 alone is sufficient to guarantee boundedness for undiscounted problems, thus rendering Lemma 9 unnecessary.

An interesting direction for further research concerns the convergence rate of Q-learning. In some sense, Q-learning makes inefficient use of information, because each piece of information is only used once. Alternative methods, that estimate the transition probabilities, can be much faster, as demonstrated experimentally by Moore and Atkeson (1992). It is an open question whether the method of Moore *et al.* (1992) has a provably better convergence rate.

We finally point out that the tools in this paper (Theorem 3, in particular) can be used to establish that the batch version of Sutton's (1988) TD(λ) algorithm converges with probability 1, for every value of λ, thus strengthening the results of Dayan (1992) where convergence was proved for the case $\lambda = 0$. This is done by expressing TD(λ) in the form of Eq. (1) and then checking that the corresponding mapping F has the required contraction properties. A proof along such lines has also been carried out by T. Jaakkola, M.I. Jordan, and S. Singh.

Acknowledgments

This research was supported by the National Science Foundation under grant ECS-9216531, and by a grant from Siemens A.G.

The author is grateful to the reviewers for their very detailed comments.

References

Barto, A.G., Bradtke, S.J., and S.P., Singh (1991). *Real-time Learning and Control Using Asynchronous Dynamic Programming*, (Technical Report 91-57). Amherst, MA: University of Massachusetts, Computer Science Dept.

Bertsekas, D.P. (1982). Distributed Dynamic Programming. *IEEE Transactions on Automatic Control, AC-27*, 610–616.

Bertsekas, D.P. and Tsitsiklis, J.N. (1989). *Parallel and Distributed Computation: Numerical Methods*, Englewood Cliffs, NJ: Prentice Hall.

Bertsekas, D.P. and Tsitsiklis, J.N. (1991). An Analysis of Stochastic Shortest Path Problems. *Mathematics of Operations Research, 16*, 580–595.

Dayan, P. (1992). The Convergence of TD(λ) for general λ. *Machine Learning, 8*, 341–362.

Kushner, H.J. and Clark, D.S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Problems*, New York, NY: Springer Verlag.

Kushner, H.J. and Yin, G., (1987). Stochastic Approximation Algorithms for Parallel and Distributed Processing, *Stochastics, 22*, 219–250.

Kushner, H.J. and Yin, G. (1987). Asymptotic Properties of Distributed and Communicating Stochastic Approximation Algorithms. *SIAM J. Control and Optimization, 25*, 1266–1290.

Li, S. and Basar, T. (1987). Asymptotic Agreement and Convergence of Asynchronous Stochastic Algorithms. *IEEE Transactions on Automatic Control, 32*, 612–618.

Moore, A.W. and Atkeson, C.G. (1992). *Memory-based Reinforcement Learning: Converging with Less Data and Less Real Time*, preprint, July 1992.

Poljak, B.T. and Tsypkin, Y. Z. (1973). Pseudogradient Adaptation and Training Algorithms. *Automation and Remote Control, 12*, 83–94.

Sutton, R.S. (1988). Learning to Predict by the Method of Temporal Differences. *Machine Learning, 3*, 9–44.

Sutton, R.S., Barto, A.G., and Williams, R.J. (1992). Reinforcement Learning is Direct Adaptive Control. *IEEE Control Systems Magazine*, April, 19–22.

Tsitsiklis, J.N., Bertsekas, D.P., and Athans, M. (1986). Distributed Deterministic and Stochastic Gradient Optimization Algorithms. *IEEE Transactions on Automatic Control. 31*, 803–812.

Watkins, C.I.C.H. (1989). Learning from Delayed Rewards. Doctoral dissertation. University of Cambridge, Cambridge, United Kingdom.

Watkins, C.I.C.H. and Dayan, P. (1992). Q-learning. *Machine Learning, 8*, 279–292.

Received March 4, 1993
Final Manuscript September 17, 1993