Probability in the Engineering and Informational Sciences, **14**, 2000, 243–258. Printed in the U.S.A.

A LEARNING ALGORITHM FOR DISCRETE-TIME STOCHASTIC CONTROL

V. S. BORKAR
School of Technology and Computer Science
Tata Institute of Fundamental Research
Homi Bhabha Road
Mumbai 400005, India
E-mail: borkar@tifr.res.in

A simulation-based algorithm for learning good policies for a discrete-time stochastic control process with unknown transition law is analyzed when the state and action spaces are compact subsets of Euclidean spaces. This extends the Q-learning scheme of discrete state/action problems along the lines of Baker [4]. Almost sure convergence is proved under suitable conditions.

1. INTRODUCTION

Reinforcement learning has been a popular paradigm in Artificial Intelligence for capturing learning situations in which the environment offers some reward/penalty for the agent's actions. This is in contrast with unsupervised learning, where there is no such feedback, and supervised learning, where the environment provides exact error measurements. From its traditional application domains like game-playing automata, reinforcement learning has been picked up by control engineers as a framework for designing simulation-based algorithms for learning optimal or near-optimal control policies. It is particularly appealing for the control of complex systems such as large communication networks, where exact modeling of the system dynamics is infeasible or expensive; but a simulation device is available which can simulate typical transitions of the system at low or moderate cost. Among various such schemes that have been proposed, one of the front-runners has been the Q-learning algorithm of Watkins [22]. Taking a cue (pun intended) from the value iteration algorithm of

© 2000 Cambridge University Press 0269-9648/00 $12.50

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

Markov decision theory, it defines "Q-values" as the argument of the minimizer in the dynamic programming equations. Assuming that the state and action spaces are finite, the Q-values are seen to form a vector indexed by state–action pairs. These Q-vectors satisfy a dynamic programming equation similar to that satisfied by the value function. Viewing this equation as a fixed-point equation, $x = F(x)$, one can conceive of a standard fixed-point iteration, $x(n + 1) = F(x(n))$, for finding the Q-vector, just like its illustrious predecessor, the value iteration.

This is fine as long as one knows the conditional probabilities required for the conditional averaging operation implicit in the dynamic programming operator $F(\cdot)$. Q-Learning enters the picture when these are unknown. It replaces the conditional average by actual evaluation at a single simulated transition at each step. The average in question is then "seen" by the algorithm by incorporating features of the classical stochastic approximation algorithm whose averaging properties are well known.

The algorithm and its variants have been analyzed by Watkins and Dayan [23], Tsitsiklis [21], and Abounady, Bertsekas, and Borkar [1,2]. Extensive accounts of reinforcement learning appear in Bertsekas and Tsitsiklis [6] and Sutton and Barto [19]. For background in Markov decision theory and stochastic approximation required here, see respectively, Puterman [18] and Benveniste, Metivier, and Priouret [5].

Although these works confine themselves largely to discrete state and action spaces, there has been some interest in extending them to continuous state and action spaces; see, for example, the recent theses of Baker [4] and Boussios [12]. The present work is in the spirit of [4]. An immediate problem with reinforcement learning algorithms for continuous state and action spaces is that they require one to update iteratively a function on these spaces based only on point measurements. Baker [4] handles this problem by updating the function not just at the point of measurement but everywhere in its domain, albeit weighed by an appropriate "kernel" function. The latter is taken to be a smooth approximation of Dirac delta, which it approaches asymptotically on a slower timescale. Baker [4] views this scheme as a stochastic approximation algorithm in the Hilbert space of functions square-integrable with respect to a base measure and uses techniques of linear functional analysis to analyze its convergence. Thus, the analysis is useful for the linear case that occurs when one aims to estimate the cost for a fixed control policy or for the linear component of an overall nonlinear "actor–critic" scheme (Konda and Borkar [15]). The Q-learning algorithm, however, involves a minimization operation at each iterate, which makes it nonlinear. For many cases of interest, the nonlinear operator in question turns out to be a contraction w.r.t. the sup norm or a weighted version thereof, as in the usual value iteration. Our approach analyzes the full nonlinear case exploiting this contractivity. Another methodological difference with Baker [4] is that whereas his analysis uses the traditional "martingale" approach extended to the Hilbert space framework, ours is in the tradition of the ODE (ordinary differential equation) approach pioneered by Ljung [16]. Specifically, we adapt the techniques developed by Borkar [8,9], Bhatnagar and Borkar [7], Borkar and Meyn [10], and Borkar and Soumyanath [11] for finite-dimensional algorithms.

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

The article is organized as follows. The next section briefly recalls the original Q-learning algorithm and introduces the algorithm under study, along with the key assumptions. Section 3 presents some preliminary results leading to the characterization of a desirable property of the control process, which we include in our assumptions. Section 4 analyzes some associated ODEs as a prelude to our "ODE approach." Section 5 proves the main convergence result assuming boundedness of iterates, the latter being established in Section 6. Section 7 concludes with some brief remarks about future possibilities. Throughout the exposition, whenever the techniques for finite-dimensional algorithms are being adapted with little change, we are brief with details, pointing out in each case the appropriate reference.

2. THE LEARNING ALGORITHM

We begin by recalling briefly the usual Q-learning algorithm for Markov chains. Let $\{X_n\}$ be a controlled Markov chain on a finite state space S with a finite action space A and transition probabilities $p : S \times A \times S \to [0,1]$. Thus,

$$P(X_{n+1} = j/X_i, Z_i, i \le n) = p(X_n, Z_n, j), \qquad n \ge 0,$$

for an A-valued control process $\{Z_n\}$. The objective is to minimize over all such $\{Z_n\}$ the infinite horizon discounted cost

$$J(\{Z_n\}, i) = E\left[\sum_{m=0}^{\infty} \beta^m k(X_m, Z_m)/X_0 = i\right],$$

where $\beta \in (0,1)$ is a prescribed discount factor and $k : S \times A \to R$ is a prescribed "running cost" function. Define the value function $V : S \to R$ by

$$V(i) = \min_{\{Z_n\}} J(\{Z_n\}, i), \qquad i \in S.$$

$V(\cdot)$ satisfies the dynamic programming equation [18]

$$V(i) = \min_u \left(k(i,u) + \beta \sum_j p(i,u,j) V(j)\right), \qquad i \in S.$$

Define Q-values $\{Q(i,u)\}$ by $Q(i,u) = k(i,u) + \beta \sum_j p(i,u,j) V(j)$. Then

$$Q(i,u) = k(i,u) + \beta \sum_j p(i,u,j) \min_b Q(j,b). \tag{1}$$

This suggests the Q-value iteration for solving (1):

$$Q^{n+1}(i,u) = k(i,u) + \beta \sum_j p(i,u,j) \min_b Q^n(j,b), \qquad i \in S, u \in A, n \ge 0.$$

If $p(\cdot)$ is unknown, one may "see" the conditional average w.r.t. $p(\cdot)$ by means of averaging properties of the stochastic approximation algorithm [5]. This is achieved by using the following stochastic approximation variant of the above algorithm:

$$Q^{n+1}(i,u) = (1 - a(n)) Q^n(i,u) + a(n)(\beta \min_b Q^n(\zeta_n(i,u), b) + k(i,u)). \tag{2}$$

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

Here, $\{a(n)\} \subset (0,1)$ satisfy $\sum_n a(n) = \infty$, $\sum_n a(n)^2 < \infty$ (i.e., the usual stochastic approximation stepsize schedule), and $\zeta_n(i,u)$ is an S-valued random variable independent of all other random variables realized up to time n, with law $p(i,u,\cdot)$. (This is presumably generated by a simulation device.) Iteration (2) is the Q-learning algorithm or, to be precise, the synchronous version thereof. The so-called asynchronous version is also of interest and has been studied in the aforementioned references.

Let $S \subset R^d$ and $U \subset R^r$ (for $d, r \geq 1$) be closures of bounded open sets with smooth boundaries in the respective spaces. Our control system will be an S-valued controlled Markov process $\{X_n\}$ given by

$$X_{n+1} = f(X_n, Z_n, \xi_{n+1}), \qquad n \geq 0,$$

where $\{Z_n\}$ is a U-valued control process and $\{\xi_n\}$ are i.i.d. R^s-valued ($s \geq 1$) "noise" random variables. The map $f: S \times U \times R^s \to S$ is assumed to be measurable. Denote the controlled transition kernel of $\{X_n\}$ by $p(x,u,dy): (x,u) \in S \times U \to \mathcal{P}(S)$ (the space of probability measures on S with Prohorov topology; we use analogous notation for other spaces of probability measures throughout.) This map is assumed to be continuous; that is, for $f \in C(S)$,

$$\int f(y)p(x,u,dy) = E[\, f(X_{n+1})/X_n = x,\, Z_n = u]$$

is continuous in (x,u). The aim is to minimize over admissible $\{Z_n\}$ the cost

$$J(\{Z_n\}, x) = E\left[\sum_{m=0}^{\infty} \beta^m k(X_m, Z_m)/X_0 = x\right]$$

for β as before and $k \in C(S \times U)$. The value function $V(x) = \inf_{\{Z_n\}} J(\{Z_n\}, x)$, $x \in S$, then satisfies the dynamic programming equation

$$V(x) = \min_u \left(k(x,u) + \beta \int p(x,u,dy)V(y) \right),$$

whence the Q-function defined by $Q^*(x,u) = k(x,u) + \beta \int p(x,u,dy)V(y)$ satisfies

$$Q^*(x,u) = k(x,u) + \beta \int p(x,u,dy) \min_z Q^*(x,z).$$

Baker's Q-learning algorithm (or, rather, a variation of it) is given by

$$\begin{aligned} Q^{n+1}(x,u) = {} & Q^n(x,u) + a(n) g_{e(n)}(x, X_n; u, Z_n) \\ & \times (k(X_n, Z_n) + \beta \min_z Q^n(X_{n+1}, z) - Q^n(X_n, Z_n)) \end{aligned} \tag{3}$$

and

$$e(n+1) = e(n) + b(n)h(e(n)), \qquad e(0) > 0, \tag{4}$$

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

for $n \geq 0$. Here, $\sum_n a(n) = \sum_n b(n) = \infty$, $\sum_n a(n)^2$, $\sum_n b(n)^2 < \infty$, $a(n+1)/a(n) \to 1$, and $b(n) = o(a(n))$. The map $h : R \to R$ is Lipschitz and chosen so that the ordinary differential equation

$$\dot{x}(t) = h(x(t)) \tag{5}$$

has the origin as its unique globally asymptotically stable equilibrium point. (We shall further qualify the choice of $(h,\{a(n)\})$ later in the proof of our main result, Theorem 5.1, the idea being that $e(n)$ should tend to zero "sufficiently slowly" in a precise sense.) Finally, for $\epsilon > 0$, $g_\epsilon : S^2 \times U^2 \to [0,\infty)$ is a smooth function satisfying

$$\int_{S\times U} g_\epsilon(x,x';y,y')\,dx'\,dy' = 1$$

and such that for $f \in C(S \times U)$,

$$\int_{S\times U} g_\epsilon(x,x';y,y')\,f(x',y')\,dx'\,dy' \to f(x,y)$$

as $\epsilon \to 0$. (That is, they are appropriate "smooth approximations" of the Dirac delta. One can use any standard family thereof that is nonnegative, with due modification near the boundary of $S \times U$.)

The interpretation of (3) and (4) is as follows. The first iteration updates the current estimate Q^n of Q based on the currently observed transition from X_n to X_{n+1} under the control action Z_n. It does so throughout its domain using the "kernel" $g_{e(n)}(\cdot, X_n; \cdot, Z_n)$ as the weighing function. The second iteration in turn ensures that $\{e(n)\}$ "tracks" the trajectories of (5), in particular converging to zero asymptotically. This decrease to zero is at a slower timescale than that of (3) because of the condition $b(n) = o(a(n))$.

$\{Z_n\}$ is said to be a Markov control if it is conditionally independent of X_m, Z_m, $m < n$, and $\{\xi_i\}$, given X_n, and, furthermore, the regular conditional law of Z_n given X_n is independent of n. Denoting this law by a measurable map $v : S \to \mathcal{P}(U)$, we refer to $v(\cdot)$ itself as the Markov control by standard abuse of terminology. Under this control, $\{X_n\}$ is a Markov process with transition kernel $p_v(x,dy) = \int v(x,du) \times p(x,u,dy)$. We further assume that this process is ergodic with a unique invariant continuous probability density $\varphi_v(\cdot)$ (w.r.t. the Lebesgue measure on S) satisfying

$$0 < c \leq \varphi_v(\cdot) \leq C < \infty$$

for constants c and C independent of the choice of v. We associate with the Markov control v the corresponding "ergodic occupation measure" $\nu_v \in \mathcal{P}(S \times U)$ defined by $\nu_v(dx,du) = \varphi_v(x)v(x,du)\,dx$. We shall further say that v is a *fair* Markov control if $v(x,du) = \eta_v(x,u)\,du$ for some continuous $\eta_v(\cdot,\cdot)$ satisfying

$$\epsilon_0^{-1} \geq \eta_v(\cdot,\cdot) \geq \epsilon_0 > 0, \text{ a.e.},$$

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

for a prescribed $\epsilon_0 > 0$. Our hypothesis on $\{Z_n\}$ will involve this notion of fairness. As the statement of this hypothesis itself involves some technicalities, it is postponed until the next section. Suffice it to say that, intuitively, fairness in this sense ensures that the process (X_n, Z_n) samples $S \times U$ sufficiently well.

Our proof technique will be as follows. We view $\{a(n)\}$ as nonuniform time steps; that is, we define time instants $t(0) = 0$, $t(n) = \sum_{i=1}^{n} a(i), n \geq 1$, and a continuous piecewise linear process $\bar{q}(\cdot) : [0,\infty) \to C(S \times U)$ by the statement that $\bar{q}(\cdot)$ is linear on each $[t(n), t(n+1)]$ with $\bar{q}(t(n)) = Q^n, n \geq 0$. One may view $\bar{q}(\cdot)$ as a noisy Euler approximation (albeit with nonuniform step sizes) of an associated ODE in $C(S \times U)$. This approximation can then be shown to get better with time, allowing us to infer stability and convergence of $\bar{q}(\cdot)$ and therefore of $\{Q^n\}$ from the corresponding properties of the ODE.

3. THE CONTROL PROCESS

In this section, we make precise the condition on $\{Z_n\}$ alluded to earlier. Recall our definition of the ergodic occupation measure ν_v associated with a Markov control v.

LEMMA 3.1: *$H = \{\nu_v : v$ Markov$\}$ is compact convex in $\mathcal{P}(S \times U)$.*

PROOF: Let $\nu \in H$ be disintegrated as $\nu(dx,dy) = \phi(dx)\psi(x,dy)$ for $\phi \in \mathcal{P}(S)$ and $\psi : S \to \mathcal{P}(U)$. Then, it is characterized by the equation

$$\int f(x)\phi(dx) = \iint f(y)p(x,u,dy)\psi(x,du)\phi(dx), \qquad f \in C(S).$$

This equation is conserved under convergence in $\mathcal{P}(S \times U)$, so H is closed. Since S and U are compact, so is $\mathcal{P}(S \times U)$ by Prohorov's theorem. Hence, H is compact. If ν_1 and ν_2 satisfy the above equation, so does any convex combination thereof. Hence, H is convex. ■

Now, define a $\mathcal{P}(S \times U)$-valued process μ_t, $t \geq 0$, by

$$\mu_t = \delta_{(X_n, Z_n)} \quad \text{for } t \in [t(n), t(n+1)), \qquad n \geq 0,$$

where $\delta_{(x,z)}$ is the Dirac measure at (x,z). View $\mu_\cdot$ as a random variable taking values in $\mathcal{U} = \{$measurable maps $[0,\infty) \to \mathcal{P}(S \times U)\}$, topologized with the coarsest topology that renders continuous the maps $u(\cdot) \in \mathcal{U} \to \int_0^T \omega(t) \int f \, du(t) \, dt$ for all $T > 0$, $\omega \in L_2[0,T]$, and f in a countable dense subset of $C(S \times U)$. It is easy to verify that $\mathcal{U}$ is compact metrizable, therefore Polish. In particular, as $t \to \infty$, $\mu_{t+\cdot}$ will converge to a sample-path-dependent compact subset of $\mathcal{U}$. The proof of the next lemma adapts the arguments of [7] to continuous state space.

LEMMA 3.2: *Almost surely, every limit point $\mu^*_\cdot$ of $\mu_{t+\cdot}$ in $\mathcal{U}$ as $t \to \infty$ satisfies $\mu^*_t \in H$ for a.e. t.*

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

PROOF: Let $\{f_i\}$ be countable dense in $C(S)$. Fix $j \geq 1$ and define

$$M_0 = 0, \qquad M_n = \sum_{i=1}^{n} a(i)(f_j(X_{i+1}) - \int f_j(y)p(X_i, Z_i, dy)), \qquad n \geq 1,$$

and $\mathcal{F}_n = \sigma(X_i, Z_i, \xi_i, i \leq n)$, $n \geq 0$. Then, $\{M_n, \mathcal{F}_n\}$ is a zero mean square-integrable martingale. Its quadratic variation is $O(\sum_{i=1}^{n} a(i)^2)$, therefore convergent a.s. By Proposition VII.2.3(c) of Neveu [17], $\{M_n\}$ converges a.s. Define

$$\widetilde{M}_n = \sum_{i=1}^{n} a(i)\left(f_j(X_i) - \int f_j(y)p(X_i, Z_i, dy)\right), \qquad n \geq 1.$$

Let $s > 0$. For $n \geq 1$, let $s(n) = \min\{m > n : \sum_{i=n}^{m} a(i) \geq s\}$. Then, $M_{s(n)} - M_n \to 0$ a.s., hence so does $(M_{s(n)} - M_n)/s$. Since $a(n+1)/a(n) \to 1$, so does

$$\frac{\widetilde{M}_{s(n)} - \widetilde{M}_n}{s}$$
$$= \frac{1}{s}\left(\int_{t(n)}^{t(n)+s} \iint f_j(x)\mu_t(dx, du)\, dt - \int_{t(n)}^{t(n)+s} \iint f_j(z)p(x, u, dz)\mu_t(dx, du)\, dt\right)$$
$$+ o(1).$$

This holds for all $\{f_i\}$ outside a common zero probability set. Since $\{f_i\}$ is a convergence-determining class for $\mathcal{P}(S)$, it follows that, almost surely, any limit point $\mu^*_\cdot$ of $\mu_{t+\cdot}$ must satisfy the following: For $t \geq 0$, $i \geq 1$, $s > 0$,

$$\frac{1}{s}\int_t^{t+s} \iint f_i(x)\mu^*_y(dx, du)\, dy = \frac{1}{s}\int_t^{t+s} \iint f_i(z)p(x, u, dz)\mu^*_y(dx, du)\, dy.$$

Letting $s \to 0$, we have for a.e. t,

$$\iint \left(f_i(x) - \int f_i(z)p(x, u, dz)\right)\mu^*_t(dx, du) = 0, \qquad i \geq 1$$

(i.e., $\mu^*_t \in H$). (In fact, by modifying $\mu^*_\cdot$ on a Lebesgue-null set, we may drop the qualification "a.e." altogether.) ■

Let $H_0 \subset H$ be the set of ergodic occupation measures that correspond to fair Markov controls. We shall say that $\{Z_n\}$ above is fair if, a.s., every limit point $\mu^*_\cdot$ of $\mu_{t+\cdot}$ as $t \to \infty$ is H_0-valued. Our main assumption is

$$\{Z_n\} \text{ is fair.}$$

Recall that in discrete state/action space Q-learning, some condition is always imposed to ensure that all state–action pairs are sampled sufficiently often. The above condition is in a similar spirit.

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

4. AN ASSOCIATED ODE

With an eye on the ODE approximation argument we propose to use, we shall study an associated ODE in this section. This will involve the operator $T: C(S \times U) \to C(S \times U)$ defined by

$$(Tf)(x,u) = k(x,u) + \beta \int p(x,u,dy) \min_z f(y,z), \qquad f \in C(S \times U).$$

Recall the sup norm on $C(S \times U)$ defined by $\| f \|_\infty = \sup_{x,u} | f(x,u)|$. It is easily verified that T is a $\| \cdot \|_\infty$-contraction, with

$$\| Tf_1 - Tf_2 \|_\infty \leq \beta \| f_1 - f_2 \|_\infty, \qquad f_1, f_2 \in C(S \times U). \tag{6}$$

Define $G = \{b(x,u): b(x,u) = \varphi_v(x)\eta_v(x,u)$ for some fair Markov control $v(\cdot)\}$. Consider the $C(S \times U)$-valued ODE

$$\frac{d}{dt} q_t(\cdot) = b_t(\cdot)(Tq_t(\cdot) - q_t(\cdot)), \qquad t \geq 0, \tag{7}$$

where b_t, $t \geq 0$, is a G-valued measurable process. The following lemma is adapted from its finite-dimensional counterparts in [8,11].

LEMMA 4.1: *There exists an $a > 0$ such that*

$$\| q_t - Q^* \|_\infty \leq e^{-at} \| q_0 - Q^* \|_\infty. \tag{8}$$

PROOF: If we multiply the right-hand side of (7) by a positive scalar, that is tantamount to a linear time scaling, which will not affect the claim. Thus, it suffices to consider the ODE

$$\frac{d}{dt} q_t = \tilde{b}_t(Tq_t - q_t), \tag{9}$$

where $\tilde{b}_t = b_t \epsilon_0/(2C)$. In particular, $0 < (c\epsilon_0^2/2C) \leq \tilde{b}_t \leq \frac{1}{2}$. Rewrite (9) as

$$\frac{d}{dt} q_t = F_t(q_t) - q_t,$$

where $F_t: C(S \times U) \to C(S \times U)$ is defined by

$$(F_t f)(x,u) = \tilde{b}_t(x,u)(Tf)(x,u) + (1 - \tilde{b}_t(x,u)) f(x,u).$$

It is easily verified that for each $t \geq 0$,

$$\| F_t f_1 - F_t f_2 \|_\infty \leq \alpha \| f_1 - f_2 \|_\infty, \qquad f_1, f_2 \in C(S \times U),$$

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

with $\alpha = 1 - (1-\beta)c\epsilon_0^2/2C < 1$, and that Q^* is the common unique fixed point of all $F_t, t \geq 0$. Explicit calculation leads to

$$\frac{d}{dt}\|q_t - Q^*\|_p = -\|q_t - Q^*\|_p^{1-p}(\|q_t - Q^*\|^p + \int (q_t(x) - Q^*(x))^{p-1} \times \operatorname{sgn}(q_t(x) - Q^*(x)) \times (F_t(q_t)(x) - F_t(Q^*)(x))\, dx),$$

for $p \geq 2$, where we use the identity $F_t(Q^*) = Q^*, t \geq 0$. By the Hölder inequality, the right-hand side is

$$\leq -\|q_t - Q^*\|_p + \|q_t - Q^*\|_p^{1-p}\|q_t - Q^*\|_p^{p-1}\|F_t(q_t) - F_t(Q^*)\|_p.$$

Integrate both sides from s to $t > s$ and let $p \to \infty$. Since $\|f\|_p \to \|f\|_\infty$ as $p \to \infty$ for $f \in C(S \times U)$, we have

$$\|q_t - Q^*\|_\infty - \|q_s - Q^*\|_\infty \leq -\int_s^t \|q_y - Q^*\|_\infty\, dy + \int_s^t \|F_y(q_y) - F_y(Q^*)\|_\infty\, dy$$

$$\leq -(1-\alpha)\int_s^t \|q_y - Q^*\|_\infty\, dy.$$

The claim follows. ■

Given $\tau, \delta > 0$, we call a bounded measurable $\bar{q}: [0,\infty) \to C(S \times U)$ a (τ,δ) perturbation of (7) if there exists $0 = \tau_0 < \tau_1 < \tau_2 < \cdots, \tau_{i+1} - \tau_i \geq \tau$, and G-valued processes $g_t^i, t \in [\tau_i, \tau_{i+1}]$, such that the solutions $q_t^i, t \in [\tau_i, \tau_{i+1}], i \geq 0$, of the ODE

$$\frac{d}{dt}q_t^i = g_t^i(Tq_t^i - q_t^i),\ q_{\tau_i}^i = \bar{q}(\tau_i), \qquad i \geq 0,$$

satisfy

$$\sup_{t \in [\tau_i, \tau_{i+1}]} \|q_t^i - \bar{q}(t)\|_\infty < \delta, \qquad i \geq 0. \tag{10}$$

The next lemma is inspired by Theorem 1 of Hirsch [13]. (See also Theorem 2.2 of [10].)

LEMMA 4.2: *For any $\tau, \epsilon > 0$, one can find a $\delta > 0$ such that any (τ,δ) perturbation $\bar{q}(\cdot)$ of (7) converges to $B_\epsilon = \{f \in C(S \times U) : \|f - Q^*\|_\infty < \epsilon\}$.*

PROOF: Let $\epsilon_1, \delta > 0$ be such that $0 < \epsilon_1 < \epsilon_1 + 2\delta < \epsilon$ and let $R > \max(\sup_t \|\bar{q}(t) - Q^*\|_\infty, \epsilon)$. Select $\tau > 0$ such that $e^{-a\tau}R < \epsilon_1$, where $a > 0$ is as in Lemma 4.1. Using (8) for $\{q^i\}$ as in (10) above and using (10), the claim is easily verified. ■

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

Now consider a related operator $L_{\epsilon,t}: C(S \times U) \to C(S \times U)$ defined by

$$(L_{\epsilon,t} f)(x,u) = \iint b_t(x',u') g_\epsilon(x,x';u,u')(Tf(x',u') - f(x',u'))\, dx'\, du'$$

for $f \in C(S \times U)$. Under our hypotheses,

$$L_{\epsilon,t} f \to b_t(Tf - f) \tag{11}$$

in $C(S \times U)$ as $\epsilon \to 0$, for $t \geq 0$. Consider the ODEs

$$\frac{d}{dt} q_t^\epsilon = L_{\epsilon,t} q_t^\epsilon, \qquad \epsilon > 0, \tag{12}$$

$$\frac{d}{dt} \bar{q}_t = b_t(T\bar{q}_t - \bar{q}_t), \tag{13}$$

with $q_0^\epsilon = \bar{q}_0$ for all ϵ. Using (6) and (11), a standard argument shows that $q_t^\epsilon \to \bar{q}_t$ as $\epsilon \to 0$, uniformly on compact time intervals. In particular, we have the following.

LEMMA 4.3: *Given $\tau, \delta > 0$, $q_\cdot^\epsilon$ is a (τ,δ) perturbation of* (13) *for sufficiently small ϵ.*

Finally, we consider a related operator $\hat{T}_\infty: C(S \times U) \to C(S \times U)$ defined by

$$\hat{T}_\infty f(x,u) = \beta \int p(x,u,dy) \min_z f(x,z), \qquad f \in C(S \times U).$$

It is easily seen that

$$\|\hat{T}_\infty f_1 - \hat{T}_\infty f_2\|_\infty \leq \beta \|f_1 - f_2\|_\infty, \qquad f_1, f_2 \in C(S \times U),$$

and the constant function identically equal to zero is the unique fixed point of $\hat{T}_\infty$. The above development can be mimicked line by line for $\hat{T}_\infty$ in place of T, with Q^* replaced by the zero function. We shall be using this fact later.

5. CONVERGENCE OF THE ALGORITHM

We shall prove convergence of the algorithm assuming that the iterates remain bounded; that is,

$$\sup_n \|Q^n\|_\infty < \infty \text{ a.s.} \tag{14}$$

This, in turn, will be proved later. Define

$$Y_{n+1}(x,u) = g_{e(n)}(x, X_n; u, Z_n)(k(X_n, Z_n) + \beta \min_z Q^n(X_{n+1}, z) - Q^n(X_n, Z_n)),$$

$$n \geq 0.$$

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

Then,

$$E[Y_{n+1}(x,u)/\mathcal{F}_n]$$
$$= g_{e(n)}(x,X_n;u,Z_n)\left(k(X_n,Z_n)+\beta\int p(X_n,Z_n,dy)\min_z Q^n(y,z)-Q^n(X_n,Z_n)\right),$$
$$n\geq 0.$$

Let

$$\Phi_n=\sum_{i=0}^{n-1}a(i)(Y_{i+1}-E[Y_{i+1}/\mathcal{F}_i]),\qquad n\geq 1,$$

viewed as a $C(S\times U)$-valued random variable.

Assume for the time being that $e(n)=\tilde{\epsilon}$ for all n.

LEMMA 5.1: *$\{\Phi_n\}$ a.s. converges in $C(S\times U)$.*

PROOF: First, view $\{\Phi_n\}$ as a sequence of random variables in $L_2(S\times U)$. Then, it is seen to be an $L_2(S\times U)$-valued martingale w.r.t. $\{\mathcal{F}_n\}$. Define

$$\langle\Phi\rangle(n)=\sum_{i=1}^{n-1}a(i)^2E[\|Y_{i+1}-E[Y_{i+1}/\mathcal{F}_i]\|_2^2/\mathcal{F}_i],\qquad n\geq 1,$$
$$\Gamma_N=\min\{n\geq 1:\langle\Phi\rangle(n+1)\geq N\},\qquad N\geq 1,$$

Note that each Γ_N is a stopping time w.r.t. $\{\mathcal{F}_n\}$. From (14), it follows that

$$E[\|Y_{n+1}-E[Y_{n+1}/\mathcal{F}_n]\|_2^2/\mathcal{F}_n]$$
$$=\int\beta^2\left\|g_{\tilde{\epsilon}}(x,X_n;u,Z_n)\left(\min_w Q^n(y,w)-\int p(X_n,Z_n,dz)\min_w Q_n(z,w)\right)\right\|_2^2$$
$$\times\, p(X_n,Z_n,dy),$$

which is bounded uniformly in n by virtue of (14). Thus, $\{\langle\Phi\rangle(n)\}$ converges a.s. Now, for $N\geq 1$, the $L_2(S\times U)$-valued martingale $\{\Phi_{n\wedge\Gamma_N}\}$ satisfies

$$\sup_n E[\|\Phi_{n\wedge\Gamma_N}\|_2^2]=\sup_n E[\langle\Phi\rangle(n\wedge\Gamma_N)]<\infty.$$

By Proposition V.2.8, of Neveu [17], $\{\Phi_{n\wedge\Gamma_N}\}$ converges a.s. Then, $\{\Phi_n\}$ converges a.s. on $B_N=\{\langle\Phi\rangle(\infty)=\lim_{n\to\infty}\langle\Phi\rangle(n)<N\}$, hence on $\cup_N B_N$, which has probability 1.

This proves convergence of $\{\Phi_n\}$ as $L_2(S\times U)$-valued random variables. Since $g(\cdot)$ is smooth, so is $\Phi_n(\cdot)$ for each n and the above argument may be repeated for partial derivatives of $\{\Phi_n\}$, obtainable by replacing g by its corresponding partial derivative in the definition of Φ_n. Thus, $\{\Phi_n\}$ converge a.s. as random variables in the Sobolev space $H^m(S\times U)$ for any $m\geq 1$. Since this is continuously embedded in $C(S\times U)$ for m sufficiently large (Adams [3, Chap. VI]), the claim follows. ■

Fix a sample path for which Lemma 5.1 holds. Recall our definition of $\{t(n)\}$. Recall $\bar{q}:[0,\infty)\to C(S\times U)$ defined by the statement that $t\to\bar{q}_t$ is continuous

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

piecewise linear, and linear on each $[t(n), t(n+1)]$ with $\bar{q}_{t(n)} = Q^n$, $n \geq 0$. Fix $\tau > 0$. Define $\tau_0 = 0$, and for $n \geq 0$,

$$\tau_{n+1} = \min\{t(i) \geq \tau_n + \tau\}.$$

Thus, $\tau_n = t(i(n))$ for an increasing sequence $\{i(n)\}$ and for all n, $\tau_{n+1} \in [\tau_n + \tau, \tau_n + \tau + 1]$. Consider a subsequence of $\{\tau_n\}$ along which $\mu_{\tau_n + \cdot}$ converges to some $\mu^*_\cdot$ in $\mathcal{U}$. By abuse of notation, we shall denote this subsequence by $\{\tau_n\}$ again. Let q^n_t, $t \in [\tau_n, \tau_{n+1}]$, be the solution to (12) with $q^n_{\tau_n} = \bar{q}_{\tau_n}$, $\epsilon = \tilde{\epsilon}$, and with b_t the density of $\mu^*_{t-\tau_n}$ with respect to the Lebesgue measure on $S \times U$. (This is possible by virtue of our "fairness" assumption in Section 3.) Let $\bar{g}(x,y) = g_{\tilde{\epsilon}}(x,\cdot;y,\cdot)$.

LEMMA 5.2: $\sup_{t \in [\tau_n, \tau_{n+1}]} \|q^n_t - \bar{q}_t\|_\infty \to 0$.

PROOF: For each n and τ_m such that $t(n) \in [\tau_m, \tau_{m+1}]$,

$$\begin{aligned}
&\bar{q}_{t(n+1)} - \bar{q}_{t(n)} \\
&\quad = a(n)E[Y_{n+1}/\mathcal{F}_n] + a(n)(Y_{n+1} - E[Y_{n+1}/\mathcal{F}_n]) \\
&\quad = \int_{t(n)}^{t(n+1)} g_{\tilde{\epsilon}}(\cdot, X_n; \cdot, Z_n)((TQ^n)(X_n, Z_n) - Q^n(X_n, Z_n))\, dt + (\Phi_{n+1} - \Phi_n) \\
&\quad = \int_{t(n)}^{t(n+1)} \int \bar{g}(T\bar{q}_{t(n)} - \bar{q}_{t(n)})\, d\mu_t\, dt + (\Phi_{n+1} - \Phi_n) \\
&\quad = \int_{t(n)}^{t(n+1)} \int \bar{g}(T\bar{q}_t - \bar{q}_t)\, d\mu^*_{t-\tau_m}\, dt \\
&\qquad + \left[\int_{t(n)}^{t(n+1)} \int \bar{g}(T\bar{q}_t - \bar{q}_t)\, d\mu_t\, dt - \int_{t(n)}^{t(n+1)} \int \bar{g}(T\bar{q}_t - \bar{q}_t)\, d\mu^*_{t-\tau_m}\, dt \right] \\
&\qquad + \left[\int_{t(n)}^{t(n+1)} \int \bar{g}(T\bar{q}_{t(n)} - \bar{q}_{t(n)})\, d\mu_t\, dt - \int_{t(n)}^{t(n+1)} \int \bar{g}(T\bar{q}_t - \bar{q}_t)\, d\mu_t\, dt \right] \\
&\qquad + [\Phi_{n+1} - \Phi_n].
\end{aligned}$$

Comparing this with

$$q^m_{t(n+1)} = q^m_{t(n)} + \int_{t(n)}^{t(n+1)} L_{\tilde{\epsilon}, t} q^m_t\, dt,$$

a standard application of Gronwall inequality leads to

$$\sup_{t \in [\tau_m, \tau_{m+1}]} \|q^m_t - \bar{q}_t\|_\infty \leq \exp(K\tau)(\mathrm{I} + \mathrm{II} + \mathrm{III}),$$

for some constant $K > 0$; I, II, and III are the respective "error terms" contributed by the three terms in the square brackets above, which we describe below in detail. Of these, III is simply

$$\sup_{i(m) \leq l \leq i(m+1)} \|\Phi_l - \Phi_{i(m)}\|_\infty,$$

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

which tends to zero as $n \to \infty$ by Lemma 5.1. II turns out to be

$$\sup_{t(n)\in[\tau_m,\tau_{m+1}]} \sup_{t\in[t(n),\,t(n+1)]} \left| \int_{t(n)}^{t} \bar{g}((T\bar{q}_{t(n)} - \bar{q}_{t(n)}) - (T\bar{q}_s - \bar{q}_s))\, d\mu_s\, ds \right|.$$

Using (6) and (14), this is easily verified to be $O(\sum_{j=i(m)}^{i(m+1)} a(j)^2)$, tending to zero as $m \to \infty$. Finally, I turns out to be

$$\sup_{t\in[\tau_m,\tau_{m+1}]} \left| \int_{\tau_m}^{t} \int \bar{g}(T\bar{q}_s - \bar{q}_s)\, d\mu_s\, ds - \int_{\tau_m}^{t} \int \bar{g}(T\bar{q}_s - \bar{q}_s)\, d\mu^*_{s-\tau_m}\, ds \right|.$$

Note that by (14) and the Arzela–Ascoli theorem, $\{\bar{q}_{\tau_m+\cdot}\}$ is seen to be relatively compact in $C([0,\infty); C(S\times U))$. By dropping to a subsequence if necessary, we may then suppose that $\bar{q}_{\tau_m+\cdot} \to q^*$ in $C([0,\infty); C(S\times U))$. Now,

$$\begin{aligned}
&\left| \int_{\tau_m}^{t} \int \bar{g}(T\bar{q}_s - \bar{q}_s)\, d\mu_s\, ds - \int_{\tau_m}^{t} \int \bar{g}(T\bar{q}_s - \bar{q}_s)\, d\mu^*_{s-\tau_m}\, ds \right| \\
&\quad \le \left| \int_{\tau_m}^{t} \int \bar{g}(T\bar{q}_s - \bar{q}_s)\, d\mu_s\, ds - \int_{\tau_m}^{t} \int \bar{g}(Tq^*_{s-\tau_m} - q^*_{s-\tau_m})\, d\mu_s\, ds \right| \\
&\qquad + \left| \int_{\tau_m}^{t} \int \bar{g}(T\bar{q}_s - \bar{q}_s)\, d\mu^*_{s-\tau_m}\, ds - \int_{\tau_m}^{t} \int \bar{g}(Tq^*_{s-\tau_m} - q^*_{s-\tau_m})\, d\mu^*_{s-\tau_m}\, ds \right| \\
&\qquad + \left| \int_{\tau_m}^{t} \int \bar{g}(Tq^*_{s-\tau_m} - q^*_{s-\tau_m})\, d\mu_s\, ds - \int_{\tau_m}^{t} \int \bar{g}(Tq^*_{s-\tau_m} - q^*_{s-\tau_m})\, d\mu^*_{s-\tau_m}\, ds \right|.
\end{aligned}$$

Since $\mu_{\tau_m+\cdot} \to \mu^*$, $\bar{q}_{\tau_m+\cdot} \to q^*$, all three terms on the right tend to zero as $m \to \infty$. That they do so uniformly in $t \in [\tau_m, \tau_{m+1}]$ follows from the boundedness and equicontinuity of the integrals, easily verifiable using (14). Thus, $I \to 0$ as $n \to \infty$. This completes the proof. ■

COROLLARY 5.1: *Given $\epsilon > 0$, the above $\{Q^n\}$ (i.e., those corresponding to the choice $e(n) = \tilde{\epsilon} > 0$ for all n) converge to the ϵ-neighborhood of Q^* a.s. for sufficiently small $\tilde{\epsilon} > 0$.*

PROOF: Let $\tau, \delta > 0$ be such that Lemma 4.2 holds. By Lemmas 4.3 and 5.2, $\bar{q}_{t+\cdot}$ is a (τ,δ) perturbation of (7) for sufficiently large t a.s., as long as $\tilde{\epsilon} > 0$ is chosen sufficiently small. The claim follows. ■

THEOREM 5.1: *Under* (14), *the* $\{Q_n\}$ *given by* (3) *converge to* Q^* *a.s. if* $e(n) \to 0$ *sufficiently slowly*.

PROOF: We first need to verify that Lemma 5.1 continues to hold for $e(n) \to 0$. The problem arises because g_ϵ being a smooth approximation to Dirac delta in the $\epsilon \to 0$ limit, its partial derivatives blow up as $\epsilon \to 0$. Note, however, that the result of [3] cited in the proof of Lemma 5.1 uses partial derivatives up to a finite order only. Let $\epsilon \to f(\epsilon)$ denote the function that captures the maximum growth thereof in the $\epsilon \to 0$ limit (i.e., these finitely many partial derivatives do not grow in absolute value

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

faster than $f(\epsilon)$ as $\epsilon \to 0$). Then, the proof of Lemma 5.1 will be complete if $\sum a(n)^2 f(e(n))^2 < \infty$. This will be true if $e(n) \to 0$ sufficiently slowly. (This can be arranged by a suitable choice of $(h,\{a(n)\})$.) Now, since $b(n) = o(a(n))$, $e(m(n+1)) = e(m(n)) + o(1)$. Thus, the argument of Lemma 5.2 proceeds as earlier with one additional error term that also goes to zero with increasing n. Now, use the argument of [9] to conclude, in view of the fact that $e(n) \to 0$ on a slower time scale than (3). ■

Note that the qualification imposed on $(h,\{a(n)\})$ earlier depends on the choice of the family $\{g_\epsilon\}$ of smooth approximations of the Dirac delta. It may also be remarked that the adaptation of the arguments of [9] here is, in fact, much simpler than in the original because the iteration for $\{e(n)\}$ does not involve $\{Q^n\}$.

6. BOUNDEDNESS OF ITERATES

This section proves (14) using essentially the same arguments as those of [10] for the finite-dimensional case. We shall sketch the outline of the proof, the details being the same as in [10]. We assume that $e(n) \to 0$ "sufficiently slowly" in the above sense.

Suppose (14) is false. Define a piecewise continuous and piecewise linear function $\tilde{q}_t$, $t \geq 0$, taking values in $C(S \times U)$ as follows: Let $r(n) = \|q_{\tau_n}\|_\infty$. If $r(n) < 1$, set $\tilde{q}_t = \bar{q}_t$, $t \in [\tau_n, \tau_{n+1})$. If $r(n) \geq 1$, set $\tilde{q}_t = \bar{q}_t / r(n)$ for $t \in [\tau_n, \tau_{n+1})$. Define $\{Y_n\}$ and $\{\Phi_n\}$ in terms of $\tilde{q}_{t(n)}$ in place of Q^n. Then, Lemma 5.1 continues to hold. For $x \in (0,\infty]$, consider the ODEs

$$\frac{d}{dt}\,\hat{q}_t^x = b_t(\hat{T}_x \hat{q}_t^x - \hat{q}_t^x), \tag{15}$$

where, for $x > 0$, the operator $\hat{T}_x$ is defined by $\hat{T}_x(f) = T(xf)/x$.

LEMMA 6.1: *As $x \to \infty$, the solutions of* (15) *with a fixed initial condition converge to the solution corresponding to $x = \infty$ uniformly on compact time intervals, uniformly w.r.t. permissible choices of b.*

PROOF: This follows by a standard argument based on the Gronwall inequality on noting that

$$\|\hat{T}_x - \hat{T}_\infty\| \to 0$$

and

$$\|\hat{T}_x f_1 - \hat{T}_x f_2\|_\infty \leq \beta \|f_1 - f_2\|_\infty, \qquad x \in (0,\infty],\ f_1, f_2 \in C(S \times U).$$

The details are standard and are omitted. ■

Consider the ODE

$$\frac{d}{dt}\,\tilde{q}_t^n = b_t(\hat{T}_{r(n)} \tilde{q}_t^n - \tilde{q}_t^n), \qquad t \in [\tau_n, \tau_{n+1}],\ n \geq 0,$$

with $\tilde{q}_{\tau_n}^n = \tilde{q}_{\tau_n}$.

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

LEMMA 6.2: $\lim_{n\to\infty} \sup_{t\in[\tau_n,\tau_{n+1})} \|\tilde{q}_t - \tilde{q}_t^n\|_\infty = 0$.

This follows exactly along the lines of Lemma 5.2. Now, suppose that (14) were false. Fix a sample path for which $\{\Phi_n\}$ converge and $r(n) \to \infty$ along a subsequence, which we denote as $\{r(n(i))\}$. Consider the ODE

$$\frac{d}{dt} q_t^n = b_t(\hat{T}_\infty q_t^n - q_t^n), \qquad t \in [\tau_n, \tau_{n+1}),$$

with $q_{\tau_n}^n = \tilde{q}_{\tau_n}$, $n \geq 0$. By Lemmas 6.1 and 6.2,

$$\lim_{n\to\infty} \sup_{t\in[\tau_n,\tau_{n+1})} \|\tilde{q}_t - q_t^n\|_\infty = 0. \tag{16}$$

Recall the remarks at the end of Section 4. Thus, for some $a > 0$ and $n = n(i)$ large,

$$\|q_{\tau_{n+1}}^n\|_\infty \leq e^{-a\tau} \|q_{\tau_n}^n\|_\infty \leq e^{-a\tau}.$$

Choosing $\tau > 0$ suitably, the right-hand side may be made less than $\frac{1}{4}$. Then, by (16),

$$\frac{\|q_{\tau_{n+1}}\|_\infty}{\|q_{\tau_n}\|_\infty} = \frac{\|\tilde{q}_{\tau_{n+1}-}\|_\infty}{\|\tilde{q}_{\tau_n}\|_\infty} < \frac{1}{2}$$

(where $\tilde{q}_{t-}$ is defined as $\lim_{t>s\to t} \tilde{q}_s$) for $n = n(i)$ with i sufficiently large; that is, for a sufficiently large time, whenever $q.$ exits the unit ball of $C(S \times U)$, it returns toward it at an exponential rate, implying in particular that $q_t, t \geq 0$, remains bounded. We have proved the following:

LEMMA 6.3: *Inequality* (14) *holds*.

7. CONCLUDING REMARKS

We have proved a.s. convergence of a variant of Baker's Q-learning algorithm. This brings up several other issues.

1. The compactness of state and action spaces was assumed for purely technical convenience. It is important to extend these results to cases where either or both fail.
2. Each step of our algorithm involves an exact minimization of a continuous function. In practice, this can be replaced by a minimization algorithm that runs on a faster timescale (in the sense of [9]) and thereby closely tracks the minimum in question.
3. The nonlocal updation idea of Baker may be useful even in discrete state/action spaces. Because of the curse of dimensionality attached to reinforcement learning schemes like Q-learning in their pure form, various approximations have been proposed, among them schemes that update components of candidate Q-vectors not at each instant, but more sparsely [20]. These may be combined with the nonlocal updation with an appropriate weighting function, as earlier.

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press

Acknowledgment

This work was supported in part by Grant No. III 5(12)/96-ET from the Department of Science and Technology, Government of India.

References

1. Abounady, J., Bertsekas, D., & Borkar, V.S. (1998). Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms. Technical Report LIDS-P-2433, Laboratory for Information and Decision Systems, MIT, Cambridge, MA.
2. Abounady, J., Bertsekas, D., & Borkar, V.S. (1998). Learning algorithms for Markov decision processes with average cost. Technical Report LIDS-P-2434, Laboratory for Information and Decision Systems, MIT, Cambridge, MA.
3. Adams, R.A. (1975). *Sobolev spaces.* New York: Academic Press.
4. Baker, W.L. (1997). Learning via stochastic approximation in function space. [Unpublished] Ph.D. thesis, Harvard University, Cambridge, MA.
5. Benveniste, A., Metivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximation.* New York: Springer-Verlag.
6. Bertsekas, D. & Tsitsiklis, J. (1996). *Neurodynamic programming.* Belmont, MA: Athena Scientific.
7. Bhatnagar, S. & Borkar, V.S. (1997). Multiscale stochastic approximation for parametric optimization of hidden Markov models. *Probability in the Engineering and Informational Sciences* 11: 509–522.
8. Borkar, V.S. (1996). Distributed computation of fixed points of ∞-nonexpansive maps. *Proceedings of the Indian Academy of Sciences (Mathematical Sciences)* 106(3): 289–300.
9. Borkar, V.S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters* 29: 291–294.
10. Borkar, V.S. & Meyn, S.P. (1998). The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal of Control and Optimization* (to appear).
11. Borkar, V.S. & Soumyanath, K. (1997). An analog scheme for fixed point computation, Part 1: Theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* CS-44(4): 351–355.
12. Boussios, C.I. (1998). An approach for nonlinear control via approximate dynamic programming. [Unpublished] Ph.D. thesis, MIT, Cambridge, MA.
13. Hirsch, M.W. (1987). Convergent activation dynamics in continuous time networks. *Neural Networks* 2: 331–349.
14. Jaakola, T., Jordan, M., & Singh, S. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation* 6: 1185–1201.
15. Konda, V.R. & Borkar, V.S. (1998). Actor-critic type learning algorithms for Markov decision processes. *SIAM Journal of Control and Optimization* (to appear).
16. Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control* AC-22: 551–575.
17. Neveu, J. (1975). *Discrete parameter martingales.* Amsterdam: North-Holland.
18. Puterman, M.I. (1994). *Markov decision processes.* New York: Wiley.
19. Sutton, R.S. & Barto, A. (1998). *Reinforcement learning.* Cambridge, MA: MIT Press.
20. Sutton, R.S., Precup, D., & Singh, S.P. (1998). Between MDPs and semi-MDPs: Learning, planning and representing knowledge as multiple temporal scales. *Journal of A.I. Research* 1: 1–39.
21. Tsitsiklis, J.N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning* 16: 185–202.
22. Watkins, C. (1989). Learning from delayed rewards. [Unpublished] Ph.D. thesis, Cambridge University, Cambridge, UK.
23. Watkins, C. & Dayan, P. (1992). Q-learning. *Machine Learning* 8: 279–292.

https://doi.org/10.1017/S0269964800142081 Published online by Cambridge University Press