On the Existence of Fixed Points for Q-Learning and Sarsa in Partially Observable Domains

Theodore J. Perkins PERKINS@CS.UMASS.EDU
Dept. of Computer Science, Univ. of Massachusetts Amherst, 140 Governor's Drive, Amherst, MA 01003 USA

Mark D. Pendrith PENDRITH@ERG.SRI.COM
SRI International, 333 Ravenswood Ave, Menlo Park, CA 94025 USA

Abstract

Model-free, action-value based reinforcement learning algorithms such as Q-Learning and Sarsa(λ) are well-suited to solving Markovian decision problems. For partially observable Markovian decision problems, however, such algorithms are less reliable. Their convergence properties have been questioned over the years, and several examples have been developed showing situations in which Q-Learning, Sarsa(λ), and related algorithms provably cannot converge. In this paper, we show that such convergence problems can stem from discontinuous action selection strategies, as were employed in all of the counterexamples. Discontinuous action selection strategies can result in a lack of fixed points in the space of action-value functions, making convergence impossible. We prove that, for a general class of POMDPs, if an agent employs any *continuous* action selection strategy, such as softmax, then action-value and policy fixed points are guaranteed to exist.

1. Introduction

Reinforcement learning algorithms that approximate value functions, such as Q-Learning, Sarsa(λ), and TD(λ), are well-suited to Markovian environments. If a separate value is stored for each state of the environment, or for each state-action pair, then these algorithms converge to correct/optimal values in both policy evaluation and control tasks (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). Even when values cannot be explicitly stored for every state, and a generalizing function approximator is used to represent the value function, some theoretical guarantees remain (Tsitsiklis & Van Roy, 1997a; Tsitsiklis & Van Roy, 1997b), and there have been notable empirical successes (e.g., Tesauro, 1994; Crites & Barto, 1998).

In a partially observable Markov decision process (POMDP), an agent receives, on each time step, an observation which does not uniquely identify the state of the agent's environment. It is natural to try to treat such observations as if they were states of the environment. One could apply, for example, Q-Learning or Sarsa to learn an observation-action value function (instead of a state-action value function) and choose a policy based on those values (Kaelbling et al., 1996).

Unfortunately, there are some serious drawbacks to this approach. One is that the agent may not be able to perform optimally. Q-Learning and Sarsa are normally used to learn stationary, deterministic policies. In Markov decision problems (MDPs), such policies map each state of the environment to an action for the agent to take. It can be shown that no other kind of policy (stochastic and/or non-stationary) can gain more reward than the optimal stationary, deterministic policy (Bertsekas, 1995). But this is not true in POMDPs. That is, for many POMDPs there are stochastic/non-stationary policies which outperform any deterministic mapping from observations to actions (Kaelbling et al., 1996).

A second objection is that Q-Learning and Sarsa may not learn the best observation to action mapping. So, not only is the class of policies that these algorithms can learn limited, but they may not even find the best policies within that class (see, e.g., Pendrith and McGarity (1998) for several examples). Lastly, and of greatest relevance to this paper, several example POMDPs have been constructed on which Q-Learning and Sarsa provably cannot converge (Gordon, 1996; Perkins, 2001).

Despite these difficulties, direct application of value function-based reinforcement learning algorithms retains some appeal. The approach is simple, and empirically it has been found that some algorithms, such as Sarsa(λ), can be quite robust to partial observability (Loch & Singh, 1998). Alternatively, augmenting an agent's representation of the environment to include prior observations (McCallum, 1995), or allowing the agent to keep an internal memory, often reduces the problems associated with partial observability. It would be useful if a sound theory could be developed of how and when value function-based reinforcement learning algorithms can be safely applied in partially observable domains.

In this paper, we take a step in this direction by arguing that the convergence difficulties for value function-based approaches may have been overstated. All of the counterexamples to convergence that have been proposed rely on ϵ-greedy action selection, for which the action selection probabilities are discontinuous in the action values. By analyzing one counterexample, we illustrate how this method of action selection may result in the non-existence of stochastic fixed points of the agent's action value function, which leads to the observed convergence difficulties. We also prove, however, that for a general class of POMDPs, if an agent uses any action selection method that is *continuous* in the action values, stochastic fixed points in the space of action value functions are guaranteed to exist. This result addresses one serious theoretical obstacle to applying algorithms such as Q-Learning and Sarsa to POMDPs.

2. POMDPs and Reactive Policies

We assume that the agent's environment is modeled as a POMDP with finite state set S and finite observation set O. When the environment is in state s, the agent observes o with probability $P_{s,o}$. We assume that for all $o \in O$, $P_{s,o}$ is positive for some s. When the agent observes o, it chooses an action from a finite set $A(o)$. When the environment is in state s and the agent chooses action a, the agent receives a stochastic reward which has mean $r_{s,a}$ and the environment transitions to state s' with probability $P_{s,a,s'}$. For $t \in \{1, 2, 3, \ldots\}$ we use s_t, o_t, a_t, and r_t to denote, respectively, the state of the environment at time t, the agent's observation at time t, the action the agent takes at time t, and the reward the agent receives at time t. Because of additional assumptions introduced below, all of our results are independent of how the initial state of the POMDP is determined. For simplicity, we may assume that s_1 is a fixed state. Rewards are discounted, with discount rate $\gamma \in [0, 1)$.

A stochastic, stationary, reactive policy, π, maps each observation o to a probability distribution over $A(o)$. $\pi(o, a)$ is the probability that the agent takes action a after observing o. Let Π denote the set of policies that place positive probability on every action available in every observation. We assume that the POMDP is ergodic in the sense that every $\pi \in \Pi$ induces a stationary distribution over S that is positive everywhere and can be written as $P_S^\pi(s) = \mathrm{clim}_{t\to\infty}\Pr(s_t = s) > 0$ for all $s \in S$. By "clim" we mean the Cesaro limit: $\mathrm{clim}_{t\to\infty}\Pr(s_t = s) = \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t}\Pr(s_\tau = s)$.

3. Agent Architecture

We consider one of the simplest possible designs for a reinforcement learning agent in a partially-observable environment. We assume that the agent maintains an action-value function, a table, which associates a real value, $Q(o, a)$, to each $o \in O$ and $a \in A(o)$. The agent updates Q according to a rule, U, based on experience in the POMDP. For example, the Q-Learning rule states that when the agent observes o, takes action a, receives reward r, and then observes o', the agent should perform the update:

$$Q(o, a) \leftarrow (1 - \alpha)Q(o, a) + \alpha(r + \gamma \max_{a' \in A(o')} Q(o', a')) \ ,$$

where α is a learning rate parameter. Other updating rules of interest include Sarsa, Sarsa(λ), Q(λ), and Advantage Learning (Sutton & Barto, 1998; Baird, 1994).

We assume that the agent chooses actions stochastically, with probabilities depending solely on its action values. Specifically, we assume that the agent follows an exploration strategy, X, which maps each possible action value function to a stochastic, stationary, reactive policy $\pi = X(Q)$. We assume that $X(Q) \in \Pi$ for all Q. This definition of exploration strategy encompasses, for example, agents behaving according to a fixed stochastic policy, agents using ϵ-greedy action selection with fixed ϵ, and agents using softmax (a.k.a. Boltzmann, Gibbs) action selection with a fixed temperature. These are three of the most common action selection strategies employed.

For us, then, an agent is primarily characterized by its updating rule, U, and its exploration strategy, X.

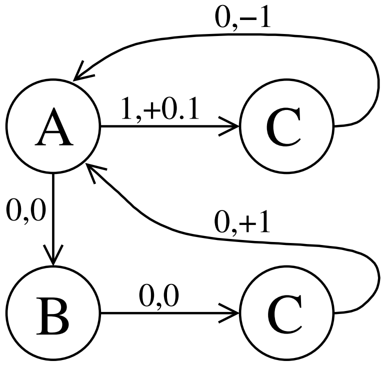

Figure 1. Example POMDP.

4. An Example Demonstrating the Problem with Discontinuous Exploration Strategies

Let us motivate our analyses with an example. The POMDP we consider is depicted in Figure 1. The four circles represent the four possible states of the environment. The letters inside the circles indicate what the agent observes when the environment is in that state. Thus, the two states on the right appear the same to the agent. The arcs between the states correspond to actions, each of which results in a fixed immediate reward and a fixed next state. The arcs are labelled by an action number and the immediate reward for taking that action.

Note that only the upper-left-hand state offers any choice of action. Under ϵ-greedy action selection, then, the policy an agent follows depends solely on the relationship between $Q(A,0)$ and $Q(A,1)$. Let the set of all action value functions be partitioned into three sets:

$$\begin{aligned} \Theta_1 &= \{Q : Q(A,0) > Q(A,1)\} \ , \\ \Theta_2 &= \{Q : Q(A,0) = Q(A,1)\} \ , \\ \Theta_3 &= \{Q : Q(A,0) < Q(A,1)\} \ . \end{aligned}$$

Let π_1, π_2, and π_3 be the corresponding ϵ-greedy policies, so that:

$$\begin{aligned} &\pi_1(A,0) = 1-\epsilon \ , && \pi_1(A,1) = \epsilon \ , \\ &\pi_2(A,0) = \tfrac{1}{2} \ , && \pi_2(A,1) = \tfrac{1}{2} \ , \\ &\pi_3(A,0) = \epsilon \ , && \pi_3(A,1) = 1-\epsilon \ . \end{aligned}$$

Now, suppose the agent follows the fixed policy π_1 and updates its action value function using the Q-Learning rule. What action values should be learned? Observation-action pair $(A,0)$ is always followed by zero reward and observation B. Thus the agent should learn $Q(A,0) = \gamma Q(B,0)$. By similar reasoning, $Q(A,1)$ should equal $0.1 + \gamma Q(C,0)$, and $Q(B,0)$ should equal $\gamma Q(C,0)$. The case is most interesting for observation C. When the agent acts according to π_1, C is followed by $+1$ reward $1-\epsilon$ fraction of the time, because that is the fraction of time the agent takes action 0 from observation A. ϵ fraction of the time, a reward of -1 follows C. In either case, the next observation is A, so the agent should learn the observation-action value $Q(C,0) = (1-2\epsilon) + \gamma \max_{a\in\{0,1\}} Q(A,a)$.

Suppose $\gamma = 0.9$ and $\epsilon = 0.01$. Solving the above system of equations, one finds that the action value function the agent should learn if it behaves according to π_1 is:

$$\begin{aligned} Q_1(A,0) &= 4.56 \ , & Q_1(A,1) &= 5.17 \ , \\ Q_1(B,0) &= 5.07 \ , & Q_1(C,0) &= 5.63 \ . \end{aligned}$$

One can also work out that the action values the agent should learn if it follows policy π_2 are:

$$\begin{aligned} Q_2(A,0) &= 0.384 \ , & Q_2(A,1) &= 0.527 \ , \\ Q_2(B,0) &= 0.426 \ , & Q_2(C,0) &= 0.474 \ . \end{aligned}$$

And for π_3:

$$\begin{aligned} Q_3(A,0) &= -2.93 \ , & Q_3(A,1) &= -3.15 \ , \\ Q_3(B,0) &= -3.25 \ , & Q_3(C,0) &= -3.62 \ . \end{aligned}$$

These calculations reveal the difficulty. $Q_1 \in \Theta_3$ and $Q_2 \in \Theta_3$, but $Q_3 \in \Theta_1$. Whatever the agent's action values, the ϵ-greedy exploration policy forces it to learn action values in a different region! What does this mean? It does not mean that it is impossible for the action values to converge. They may converge to the boundary between the regions, forever chattering between them. It does mean that it is impossible for the agent's behavior, the policy it follows, to ever settle down to a fixed policy. The agent's action values are forced to move from one region to another indefinitely, and so the agent's policy must switch indefinitely among π_1, π_2, and π_3, never settling on any of them. The reader may verify that a similar situation occurs if one considers Sarsa updates instead of Q-Learning updates. In Section 8, we present simulations in this domain, comparing how the action values change under ϵ-greedy action selection and under softmax action selection, a continuous exploration strategy.

5. Fixed Points

In our analyses, we discuss three types of stochastic fixed points, which we summarize in Figure 2. The first describes an action value function being a fixed point with respect to an update rule and a policy. Suppose that an agent follows policy π and updates its action value function according to U. Intuitively, we want to say that Q is a stochastic fixed point if the "expected update" to Q is zero. That is, Q is a stochastic

Q fixed w.r.t. U and π	$\equiv$	$Q = \mathcal{Q}_U(\pi)$
Q fixed w.r.t. U and X	$\equiv$	$Q = \mathcal{Q}_U(X(Q))$
π fixed w.r.t. U and X	$\equiv$	$\pi = X(\mathcal{Q}_U(\pi))$

Figure 2. Three types of fixed points.

fixed point if, over the long-term, the changes to Q recommended by U average out to zero. We formally define the expected update for Q-Learning and Sarsa individually below. For now, let us suppose that for a given U and π there always exists a unique action value function that is a stochastic fixed point. We write this fixed point as $Q = \mathcal{Q}_U(\pi)$.

One may also think of $\mathcal{Q}_U(\pi)$ as the action value function to which Q converges if the agent behaves according to π and updates Q according to U with appropriately decreasing learning rates. In general, the existence of a fixed point, Q, does not guarantee that the agent's action value function converges to Q. It turns out, however, that this *is* true for Q-Learning and Sarsa updating under the assumptions we have made. That is, if the agent behaves according to π, its action values do converge to $\mathcal{Q}_U(\pi)$ (Singh et al., 1994).

Our next definition describes an action value function being fixed with respect to an exploration strategy rather than just a fixed policy. We say Q is a fixed point with respect to updating rule U and exploration strategy X if $Q = \mathcal{Q}_U(X(Q))$. Intuitively, Q is a fixed point if an agent with that action value function can follow its exploration strategy, X, and the updating rule recommends no change to Q on average.

Lastly, we define π to be a fixed point with respect to U and X if $\pi = X(\mathcal{Q}_U(\pi))$. Action value fixed points exist if and only if policy fixed points exist. If $Q = \mathcal{Q}_U(X(Q))$, then $X(Q) = X(\mathcal{Q}_U(X(Q)))$, so $\pi = X(Q)$ is a policy fixed point. Likewise, if $\pi = X(\mathcal{Q}_U(\pi))$, then $\mathcal{Q}_U(\pi) = \mathcal{Q}_U(X(\mathcal{Q}_U(\pi)))$, so $Q = \mathcal{Q}_U(\pi)$ is an action value fixed point. For an arbitrary POMDP, U, and X, there may be no fixed points or there may be infinitely many fixed points. The main positive result of this paper is to show that, for a general class of POMDPs, if an agent uses Q-Learning or Sarsa updates and follows a continuous exploration strategy, then there will always be at least one action value function and corresponding policy that are fixed points.

Theorem 1 *For any ergodic POMDP, updating rule U, and exploration strategy X, if the mappings $\mathcal{Q}_U$ and X are continuous in their inputs, and if $\mathcal{Q}_U(\pi)$ is bounded for all $\pi \in \Pi$, then there exists at least one action value function and one policy that are fixed points with respect to U and X.*

Proof: Since $\mathcal{Q}_U(\pi)$ is bounded for all $\pi \in \Pi$, there is some $M > 0$ such that for all $\pi \in \Pi$, o, and $a \in A(o)$, $|\mathcal{Q}_U(\pi)(o,a)| \leq M$. Let $\Theta^M = \{Q : |Q(o,a)| \leq M \text{ for all } o \text{ and } a\}$. Note that an action value function can be viewed as an m-dimensional real-valued vector, where m is the number of observation-action pairs. Thus, Θ^M is a closed, bounded, convex subset of $\Re^m$. Because $\mathcal{Q}_U$ and X are continuous, the composition, $\mathcal{Q}_U \circ X$ is a continuous mapping from Θ^M into itself. The existence of an action value fixed point, $Q = \mathcal{Q}_U(X(Q))$, follows from Brouwer's Fixed Point Theorem.[1] $\pi = X(Q)$ is a policy fixed point with respect to U and X. $\square$

It is usually obvious whether or not X is continuous. For example, softmax action selection is continuous in the action values, whereas ϵ-greedy action selection is not. Determining whether $\mathcal{Q}_U$ is well-defined, continuous, and bounded for any given U is less obvious. We do this for Q-Learning and Sarsa below.

6. Existence of Fixed Points for Q-Learning

The essence of our argument is that, although the agent's environment is partially observable, because it is ergodic, under any fixed policy there are limiting "empirical rewards" following each observation and action and limiting "empirical probabilities" that one observation will follow another. Thus, the expected update to an observation-action value turns out to be just the expected update to a state-action value of a related Markov decision process. This establishes the well-definedness and continuity of the $\mathcal{Q}_U$ mapping for the Q-Learning update rule, which we denote $\mathcal{Q}_{QL}$.

Suppose that the agent behaves according to a fixed $\pi \in \Pi$. As noted in Section 2, the ergodicity assumption implies that π induces a stationary distribution over S, $P^\pi_S(s) = \text{clim}_{t\to\infty}\Pr(s_t = s)$, which is positive for all s. Thus, we can also define a stationary distribution over O as $P^\pi_O(o) = \text{clim}_{t\to\infty}\Pr(o_t = o) = \text{clim}_{t\to\infty}\sum_s P_{s,o}\Pr(s_t = s) = \sum_s P_{s,o}P^\pi_S(s)$. Because we assume that, for all o, $P_{s,o}$ is positive for some s, P^π_O is positive for all o.

[1] Brouwer's Fixed Point Theorem states that a continuous mapping from a closed, bounded, convex subset of $\Re^m$ into itself has a fixed point (Cairns, 1968).

Ergodicity also implies the existence of limiting expected rewards,

$$
\begin{aligned}
r^\pi_{o,a} &= \operatorname*{clim}_{t\to\infty} E\{r_t \mid o_t=o,\ a_t=a\} \\
&= \operatorname*{clim}_{t\to\infty} \sum_{s\in S} \left(\begin{array}{c} E\{r_t \mid s_t=s,\ o_t=o,\ a_t=a\} \\ *\Pr(s_t=s \mid o_t=o,\ a_t=a) \end{array} \right) \\
&= \operatorname*{clim}_{t\to\infty} \sum_{s\in S} \left(\begin{array}{c} E\{r_t \mid s_t=s,\ a_t=a\} \\ *\Pr(s_t=s \mid o_t=o) \end{array} \right) \\
&= \operatorname{clim}_{t\to\infty} \sum_{s\in S} \left(\begin{array}{c} E\{r_t \mid s_t=s,\ a_t=a\} \\ * \dfrac{\Pr(o_t=o \mid s_t=s)\Pr(s_t=s)}{\Pr(o_t=o)} \end{array} \right) \\
&= \sum_{s\in S} r_{s,a} \frac{P_{s,o} P^\pi_S(s)}{P^\pi_O(o)} \ ,
\end{aligned}
$$

and transition probabilities,

$$
\begin{aligned}
P^\pi_{o,a,o'} &= \operatorname*{clim}_{t\to\infty} \Pr(o_{t+1}=o' \mid o_t=o,\ a_t=a) \\
&= \operatorname*{clim}_{t\to\infty} \sum_{s\in S} \left(\begin{array}{c} \Pr(o_{t+1}=o' \mid s_t=s,\ o_t=o,\ a_t=a) \\ *\Pr(s_t=s \mid o_t=o,\ a_t=a) \end{array} \right) \\
&= \operatorname*{clim}_{t\to\infty} \sum_{s\in S} \left(\begin{array}{c} \Pr(o_{t+1}=o' \mid s_t=s,\ a_t=a) \\ *\Pr(s_t=s \mid o_t=o) \end{array} \right) \\
&= \sum_{s\in S} \left(\sum_{s'\in S} P_{s,a,s'} P_{s',o'} \right) \left(\frac{P_{s,o} P^\pi_S(s)}{P^\pi_O(o)} \right) \ .
\end{aligned}
$$

Recall that the Q-Learning update is: $Q(o_t,a_t) \leftarrow (1-\alpha)Q(o_t,a_t)+\alpha(r_t+\gamma \max_{a\in A(o_{t+1})} Q(o_{t+1},a))$. We define the expected Q-Learning update to the action value function of an agent behaving according to π as:

$$
\begin{aligned}
\mathrm{EU}^\pi_{QL}(o,a) &= \operatorname{clim}_{t\to\infty} E\{r_t \mid o_t=o,\ a_t=a\} \\
&\quad + \gamma \sum_{o'} \left(\begin{array}{c} \Pr(o_{t+1}=o' \mid o_t=o,\ a_t=a) \\ *\max_{a'} Q(o',a') \end{array} \right) \\
&= r^\pi_{o,a} + \gamma \sum_{o'} P^\pi_{o,a,o'} \max_{a'} Q(o',a') \ .
\end{aligned}
$$

Q is a fixed point with respect to Q-Learning updates and π if $Q = \mathrm{EU}^\pi_{QL}$. That is, Q is a fixed point when:

$$
Q(o,a) = r^\pi_{o,a} + \gamma \sum_{o'} P^\pi_{o,a,o'} \max_{a'} Q(o',a') \ ,
$$

for all o and a.

Theorem 2 *For any ergodic POMDP, $\mathcal{Q}_{QL}(\pi)$ is well-defined and bounded for all $\pi \in \Pi$ and is continuous in π.*

Proof: $\mathcal{Q}_{QL}(\pi)$ is obviously the optimal action value function for an MDP with states o, rewards $r^\pi_{o,a}$, and transition probabilities $P^\pi_{o,a,o'}$. Optimal action values are known to exist and be unique under the assumptions we have made (Bertsekas, 1995). For boundedness, note that for any π, o, and a, $r^\pi_{o,a} \leq r_{\max} = \max_{s,a} |r_{s,a}|$. Since $\gamma < 1$, such optimal action values are bounded in magnitude by $\frac{1}{1-\gamma} r_{\max}$ (Sutton & Barto, 1998). Optimal action values are also known to be continuous in the MDP's rewards and transition probabilities. By inspection, $r^\pi_{o,a}$ and $P^\pi_{o,a}$ are continuous in P^π_S. De Farias and Van Roy (2000) showed that P^π_S is continuous in π. So $\mathcal{Q}_{QL}(\pi)$ is continuous in π. □

Corollary 1 *For any ergodic POMDP and any continuous exploration strategy, X, there exists at least one action value function and corresponding policy that are fixed points with respect to Q-Learning updates and X.*

This follows directly from Theorems 1 and 2.

7. Existence of Fixed Points for Sarsa

For Sarsa, a similar argument works. According to the Sarsa rule, the value of one observation-action pair, (o,a), is updated based on an immediate reward and the value of the next observation pair, (o',a'). One can define limiting empirical probabilities that (o,a) is followed by (o',a'), which define a Markov chain evolving on a state set consisting of observation-action pairs. Formally, let π be fixed, and let P^π_S, P^π_O, and $r^\pi_{o,a}$ be as above. We define:

$$
\begin{aligned}
P^\pi_{(oa),(o'a')} &= \operatorname*{clim}_{t\to\infty} \Pr(o_{t+1}=o',\ a_{t+1}=a' \mid o_t=o,\ a_t=a) \\
&= \operatorname*{clim}_{t\to\infty} \sum_{s\in S} \left(\begin{array}{c} \Pr(o_{t+1}=o',\ a_{t+1}=a' \\ \mid s_t=s,\ o_t=o,\ a_t=a) \\ *\Pr(s_t=s \mid o_t=o,\ a_t=a) \end{array} \right) \\
&= \sum_{s\in S} \left(\sum_{s'\in S} P_{s,a,s'} P_{s',o'} \pi(o',a') \right) \frac{P_{s,o} P^\pi_S(s)}{P^\pi_O(o)} \ .
\end{aligned}
$$

Recall that the Sarsa update is $Q(o_t,a_t) \leftarrow (1-\alpha)Q(o_t,a_t) + \alpha(r_t + \gamma Q(o_{t+1},a_{t+1}))$. We define the expected Sarsa update to the action value function of an agent behaving according to π as:

$$
\begin{aligned}
\mathrm{EU}^\pi_{Sarsa}(o,a) &= \operatorname*{clim}_{t\to\infty} E\{r_t \mid o_t=o,\ a_t=a\} \\
&\quad + \gamma \sum_{o',a'} \left(\begin{array}{c} \Pr(o_{t+1}=o',\ a_{t+1}=a' \\ \mid o_t=o,\ a_t=a) \\ *Q(o',a') \end{array} \right) \\
&= r^\pi_{o,a} + \gamma \sum_{o',a'} P^\pi_{(oa),(o'a')} Q(o',a') \ .
\end{aligned}
$$

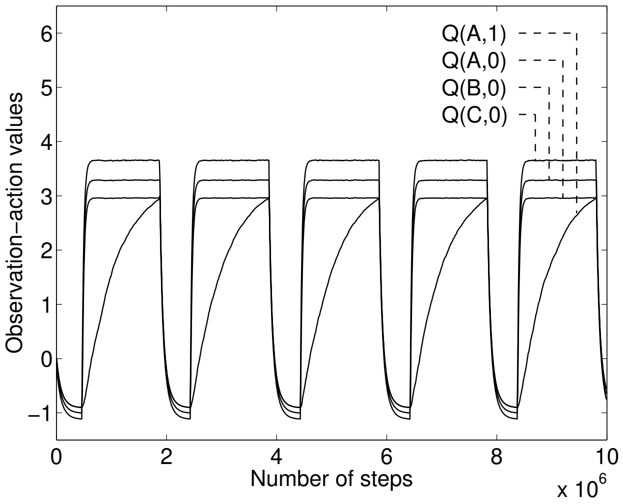

Figure 3. Evolution of observation-action values under ϵ-greedy action selection.

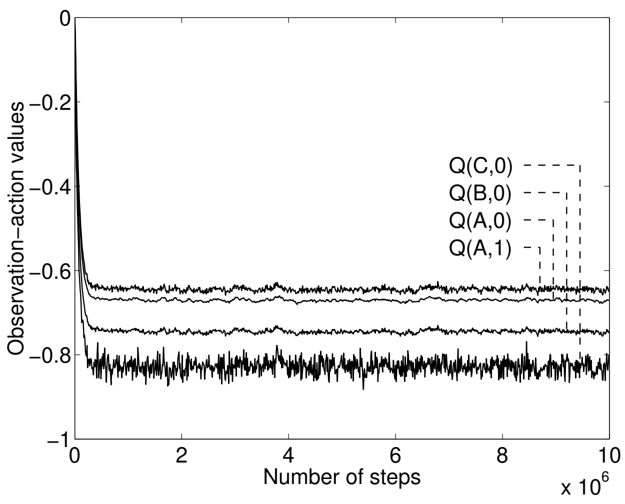

Figure 4. Evolution of observation-action values under modified softmax action selection.

Q is a fixed point with respect to Sarsa updates and π if $Q = \mathrm{EU}^{\pi}_{Sarsa}$. That is, Q is a fixed point when:

$$Q(o,a) = r^{\pi}_{o,a} + \gamma \sum_{o',a'} P^{\pi}_{(oa),(o'a')} Q(o',a') \ ,$$

for all o and a.

Theorem 3 *For any ergodic POMDP, $\mathcal{Q}_{Sarsa}(\pi)$ is well-defined and bounded for all $\pi \in \Pi$ and is continuous in π.*

Proof: $\mathcal{Q}_{Sarsa}(\pi)$ is the value function for a Markov chain with states (o,a), rewards $r^{\pi}_{o,a}$, and transition probabilities $P^{\pi}_{(oa),(o'a')}$. Value functions are known to exist and be unique under the assumptions we have made (Bertsekas, 1995), and state values are bounded by $\frac{1}{1-\gamma} r_{\max} = \frac{1}{1-\gamma} \max_{s,a} |r_{s,a}|$. Values are known to be continuous in the reward and transitions probabilities of the chain, which are, by inspection, continuous in P^{π}_{S}, and thus continuous in π. $\Box$

Corollary 2 *For any ergodic POMDP and any continuous exploration strategy, X, there exists at least one action value function and corresponding policy that are fixed points with respect to Sarsa updates and X.*

This follows directly from Theorems 1 and 3.

8. Example Revisited

In this section we present simulation results in the domain introduced in Section 4. We performed two runs of Q-Learning with different exploration strategies. In one, the agent used ϵ-greedy action selection with $\epsilon = 0.01$. For the other, the agent used softmax action selection with a temperature of $T = 0.05$. Thus, at each time step, each action had probability $\exp(Q(o,a)/T)/\sum_{a'} \exp(Q(o,a')/T)$ of being chosen. For both runs, the environment started in the upper-left-hand state (A), and continued for 10 million time steps. Observation-action values were initialized to zero and were updated using a fixed learning rate of $\alpha = 0.001$.

Figures 3 and 4 display the time-evolution of the four observation-action values under the two exploration strategies. Under ϵ-greedy action selection, the action values followed a definite, cyclic pattern of rises and falls. By contrast, under the continuous exploration strategy the values quickly converged and hovered around a fixed point. Qualitatively similar results were obtained for a variety of other choices for ϵ, T, and α.

Interestingly, although the action values of the ϵ-greedy agent did not converge, for the majority of the time, $Q(A,0) > Q(A,1)$. Thus, most of the time, the ϵ-greedy agent followed the optimal policy. By contrast, the softmax learning converged to a steady state in which $Q(A,1) > Q(A,0)$, and thus performance was far from optimal. On average, the ϵ-greedy agent received 0.212 reward per time step and the softmax agent received -0.078 reward per time step. In this example, then, the non-convergent behavior of ϵ-greedy agent is associated with greater reward over time than the convergent behavior of the softmax agent. Further study is warranted to determine the generality of this phenomenon.

9. Conclusions and Future Work

We showed that if a reinforcement learning agent employs a continuous exploration strategy with Q-Learning or Sarsa update rules, then stochastic fixed points in the space of action value functions are guaranteed to exist. These results can be extended to other, similar algorithms, such as R-Learning (Schwartz, 1993), Advantage Learning (Baird, 1994), and, we believe, Sarsa(λ). Our results can also be extended to agents that condition action value estimates on portions of experience history or on internal memory, rather than on just the most recent observation. The essential restriction imposed by our present proof is that the action value function be tabular—i.e. that it maps a finite number of different "situations" to action values—so that Bellman equations can be established relating those action values. Finally, we note that we have considered continuing tasks (no terminal states) in ergodic POMDPs, but we anticipate that, under appropriate assumptions, episodic tasks can be handled by similar arguments.

An obvious next step in this line of analysis is to try to establish convergence of the action values to the fixed points. Perhaps even more important, however, is characterizing the fixed points in terms of the learning rule used and in terms of features of the domain. The example presented in this paper demonstrates that the policy fixed points created by a continuous action selection strategy may have worse average reward than is achieved by a non-converging agent. In other work, it has been shown that, for some POMDPs, the fixed points for Q-Learning or Sarsa(λ) may correspond to quite bad policies regardless of the continuity of action selection (Singh et al., 1994; Pendrith & McGarity, 1998).

Empirically, however, it has been found that Sarsa(λ) can often find good policies even in the face of significant partial observability (Loch & Singh, 1998). Other reinforcement learning algorithms can also do well if enough history or memory is added to the agent's value function representation. The intuition behind adding history or memory is that it leads to a "more Markovian" representation. How "degree of Markovianness" might be formalized, and how it would affect the fixed points of algorithms such as Q-Learning or Sarsa(λ), are important questions that warrant study.

Another issue to be addressed is what happens when exploration strategies vary over time. Often, the exploration strategy of an agent is not just function of the action values. In ϵ-greedy or softmax action selection, for example, ϵ or T, respectively, are sometimes taken to zero over time. Intuitively, as the agent gains confidence in its action value estimates, there is less need to test actions estimated to be suboptimal. For an agent using Sarsa updates in a Markovian environment, convergence to an optimal policy can be guaranteed only if exploration is taken to zero at an appropriate rate (Singh et al., 2000). Our results show that under softmax action selection, for example, fixed points exist for any fixed T, but it is unclear what happens if T varies as a function of time. Especially problematic is the case in which T goes to zero, because in that limit, the exploration policy becomes discontinuous in the action values. The implications of our results for exploration strategies that vary over time have yet to be fully determined.

Interestingly, continuous exploration strategies are not required by all action-value based reinforcement learning algorithms. Other work has shown that some Monte Carlo algorithms, which have no or little dependence on Markovian assumptions, have fixed points regardless of the continuity of the exploration strategy—although other aspects of exploration must be handled with care (Pendrith & McGarity, 1998; Perkins, 2001). Viewed in this light, our work shows that it is not action-value based methods per se that require continuous exploration. Rather, it is the combination of temporal-difference style updating, which depends strongly on the Markov assumption, with partial observability (violating the Markov assumption) that make discontinuous action selection dangerous.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant Nos. ECS-0070102 and ECS-9980062. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Daniel Bernstein and Doina Precup for comments on various drafts of this paper.

References

Baird, L. C. (1994). Reinforcement learning in continuous time: Advantage updating. *Proceedings of the International Conference on Neural Networks.*

Bertsekas, D. P. (1995). *Dynamic programming and optimal control, vol. 1.* Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming.* Belmont, MA: Athena Scientific.

Cairns, S. S. (1968). *Introductory topology*. New York: Ronald Press Company.

Crites, R. H., & Barto, A. G. (1998). Elevator group control using multiple reinforcement learning agents. *Machine Learning*, *33*, 235–262.

De Farias, D. P., & Van Roy, B. (2000). On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, *105*.

Gordon, G. (1996). Chattering in Sarsa(λ). CMU Learning Lab Internal Report. Available at www.cs.cmu.edu/∼ggordon.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Loch, J., & Singh, S. (1998). Using eligibility traces to find the best memoryless policy in a partially observable Markov decision process. *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann.

McCallum, A. K. (1995). *Reinforcement learning with selective perception and hidden state*. Doctoral dissertation, University of Rochester.

Pendrith, M. D., & McGarity, M. J. (1998). An analysis of direct reinforcement learning in non-Markovian domains. *Machine Learning: Proceedings of the 15th International Conference* (pp. 421–429).

Perkins, T. J. (2001). *Action value based reinforcement learning for POMDPs* (Technical Report UM-CS-2001-020). Department of Computer Science, University of Massachusetts Amherst.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. *Proceedings of the Tenth International Conference on Machine Learning* (pp. 298–305). Morgan Kaufmann.

Singh, S., Jaakkola, T., & Jordan, M. (1994). Learning without state-estimation in partially observable Markovian decision processes. *Proceedings of the Eleventh International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann.

Singh, S., Jaakkola, T., Littman, M. L., & Szepesvari, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, *38*, 287–308.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, Massachusetts: MIT Press/Bradford Books.

Tesauro, G. J. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, *6*, 215–219.

Tsitsiklis, J. N., & Van Roy, B. (1997a). Analysis of temporal-difference learning with function approximation. *Advances in Neural Information Processing Systems 9* (pp. 1075–1081). Cambridge, MA: MIT Press.

Tsitsiklis, J. N., & Van Roy, B. (1997b). Approximate solution to optimal stopping problems. *Advances in Neural Information Processing Systems 9* (pp. 1082–1088). Cambridge, MA: MIT Press.