A Unified Analysis of Value-Function-Based Reinforcement-Learning Algorithms

Csaba Szepesvári
Research Group on Artificial Intelligence
"József Attila" University
Szeged 6720, Aradi vrt tere 1.
Hungary
szepes@sol.cc.u-szeged.hu

Michael L. Littman
Department of Computer Science
Duke University
Durham, NC 27708-0129
mlittman@cs.duke.edu

October 27, 1998

Abstract

Reinforcement learning is the problem of generating optimal behavior in a sequential decision-making environment given the opportunity of interacting with it. Many algorithms for solving reinforcement-learning problems work by computing improved estimates of the optimal value function. We extend prior analyses of reinforcement-learning algorithms and present a powerful new theorem that can provide a unified analysis of value-function-based reinforcement-learning algorithms. The usefulness of the theorem lies in how it allows the convergence of a complex asynchronous reinforcement-learning algorithm to be proven by verifying that a simpler synchronous algorithm converges. We illustrate the application of the theorem by analyzing the convergence of Q-learning, model-based reinforcement learning, Q-learning with multi-state updates, Q-learning for Markov games, and risk-sensitive reinforcement learning.

1 Introduction

A reinforcement learner interacts with its environment and is able to improve its behavior from experience. Different reinforcement-learning problems are defined by different objective criteria and by different types of information available to the decision maker (learner). In spite of these differences, many different reinforcement-learning problems can be solved by a *value-function*-based approach. Here, the decision maker keeps an estimate of the value of the objective criteria starting from each state in the environment, and these estimates are updated in light of new experience. Many algorithms of this type of have been proven to converge asymptotically to optimal value estimates, which in turn can be used to generate optimal behavior. Introductions to reinforcement learning can be found in an article by Kaelbling, Littman, & Moore (1996) and books by Sutton & Barto (1998) and Bertsekas & Tsitsiklis (1996).

This paper provides a unified framework for analyzing a variety of reinforcement-learning algorithms, in the form a powerful new convergence theorem. The usefulness of the theorem lies in how it allows the convergence of a complex asynchronous reinforcement-learning algorithm to be proven by verifying that a simpler synchronous algorithm converges. Section 2 states the theorem and Section 3 applies the theorem to a collection of reinforcement-learning algorithms, including Q-learning, model-based reinforcement learning, Q-learning with multi-state updates, Q-learning for Markov games, and risk-sensitive reinforcement learning. Section A then proves the theorem, providing detailed descriptions of the mathematical techniques employed.

1.1 Reinforcement Learning

The most commonly analyzed reinforcement-learning algorithm is Q-learning (Watkins & Dayan 1992). Typically, an agent following the Q-learning algorithm interacts with an environment defined as a finite Markov decision process (MDP), with the objective of minimizing total discounted expected cost (or maximizing total expected discounted reward). A finite MDP environment consists of a finite set of states $\mathcal{X}$, finite set of actions $\mathcal{A}$, transition function $\Pr(y|x,a)$ (for $x, y \in \mathcal{X}$, $a \in \mathcal{A}$), and expected cost function $c(x,a,y)$ (for $x, y \in \mathcal{X}$, $a \in \mathcal{A}$). At each discrete moment in time, the decision maker is in some state $x \in \mathcal{X}$, known to the decision maker. It chooses an action $a \in \mathcal{A}$, and issues it to the environment, resulting in a state transition to $y \in \mathcal{X}$ with probability $\Pr(y|x,a)$. It is charged an expected immediate cost of $c(x,a,y)$, and the process repeats. The decision maker's performance is measured with respect to a discount factor $0 \leq \gamma < 1$; the decision maker seeks to choose actions to minimize $E[\sum_{t=0}^{\infty} \gamma^t c_t]$, where c_t is the immediate cost received on discrete time step t.

Consider a finite MDP with let the objective criterion of minimizing total discounted expected cost. The optimal value function v^*, as is well known (Puterman 1994), is the fixed point of the optimal value operator $T : B(\mathcal{X}) \to B(\mathcal{X})$,

$$(Tv)(x) = \min_{a \in \mathcal{A}} \sum_{y \in \mathcal{X}} \Pr(y|x,a)\big(c(x,a,y) + \gamma v(y)\big), \tag{1}$$

$0 \leq \gamma < 1$, where $\Pr(y|x,a)$ is the probability of going to state y from state x when action a is used, $c(x,a,y)$ is the cost of this transition and γ is the discount factor. It is also well known that greedy policies with respect to

v^* are optimal; that is, always choosing the action $a \in \mathcal{A}$ that minimizes $\sum_{y \in \mathcal{X}} \Pr(y|x,a)(c(x,a,y) + \gamma v^*(y))$ results in optimal performance. The defining assumption of reinforcement learning (RL) is that the probability transition function and cost functions are unknown, so the optimal value operator T is also unknown. Methods for RL can be divided into two parts: value-function based, when v^* is found by some fixed-point computation, and policy-iteration based. Here, we will be concerned only with the first class of methods (policy-iteration-based RL algorithms do not appear to be amenable to the methods of this article). In the class of value-function-based algorithms, an estimate of the optimal value function is built gradually from the decision maker's experience and sometimes this estimate is used for control.

To define how a value-function-based RL algorithm works, assume we have an MDP and that the decision maker has access to unbiased samples from $\Pr(\cdot|x,a)$ and c; we assume that when the system's state-action transition is (x,a,y), the decision maker receives a random value c, called the reinforcement signal, whose expectation is $c(x,a,y)$. In a *model-based approach*, a decision maker approximates the transition and cost functions as p and c, uses the estimated values (p_t, c_t) to approximate T (the optimal value operator given in Equation (1)) by $T_t = T(p_t, c_t)$, and then uses the operator sequence T_t to build an estimate of v^*. In a *model-free approach*, such as Q-learning (Watkins 1989), for example, the decision maker directly estimates v^* without ever estimating p or c. We describe an abstract version of Q-learning next, as it provides a framework and vocabulary for summarizing the majority of our results.

Q-learning proceeds by estimating the function $Q^* = \mathcal{Q}v^*$, where $(\mathcal{Q}f)(x,a) = \sum_{y \in \mathcal{X}} \Pr(y|x,a)\big(c(x,a,y) + \gamma f(y)\big)$ is the cost-propagation operator. Q-learning explicitly represents values for state-action pairs: the function $Q^*(x,a)$ is the total discounted expected cost received by starting in state x, choosing action a once, then choosing optimal actions in all succeeding states. The idea behind the estimation procedure is the following: from the optimality equation $v^* = Tv^*$ it follows that Q^* is the fixed point of the operator $\tilde{T}$, defined as

$$\begin{aligned} (\tilde{T}Q)(x,a) &= \sum_{y \in \mathcal{X}} \Pr(y|x,a)\big(c(x,a,y) + \gamma \min_{b \in \mathcal{A}} Q(y,b)\big) \\ &= (\mathcal{Q}\mathcal{N}Q)(x,a), \end{aligned}$$

where $\mathcal{N} : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ is the minimization operator: $(\mathcal{N}Q)(x) = \min_{a \in \mathcal{A}} Q(x,a)$. For any function Q, $\tilde{T}Q$ is easily approximated by averaging; consider the sequence Q_t defined recursively by

$$Q_{t+1}(x,a) = \qquad (2)$$
$$\begin{cases} \left(1 - \frac{1}{n_t(x,a)}\right) Q_t(x,a) + \frac{1}{n_t(x,a)}\left(c_t + \gamma(\mathcal{N}Q)(x_{t+1})\right), & \text{if } (x,a) = (x_t, a_t); \\ Q_t(x,a), & \text{otherwise,} \end{cases}$$

where $n_t(x,a)$ is the number of times the state-action pair (x,a) was visited by the process (x_t, a_t) before time t plus one, and (x_t, c_t) is a Markov process (given a rule for selecting the sequence of actions, a_t) with transition laws given by $\Pr(x_{t+1}|x_t, a_t)$, $E[c_t|x_t, a_t, x_{t+1}] = c(x_t, a_t, x_t)$ and $\mathrm{Var}[c_t|x_t, a_t, x_{t+1}] < \infty$. The above iteration can be put in the more compact form

$$Q_{t+1} = T_t(Q_t, Q), \tag{3}$$

where T_t is a sequence of appropriately defined random operators:

$$(T_t(Q_t, Q))(x, a) =$$
$$\begin{cases} \left(1 - \frac{1}{n_t(x,a)}\right) Q_t(x, a) + \frac{1}{n_t(x,a)} \left(c_t + \gamma(\mathcal{N}Q)(x_{t+1})\right), & \text{if } (x, a) = (x_t, a_t); \\ Q_t(x, a), & \text{otherwise.} \end{cases}$$

Thus, we can compute $\tilde{T}Q$ for any fixed function Q using experience: define $Q_0 = Q$, $Q_{t+1} = T_t(Q_t, Q)$ for $t > 0$, then $Q_t \to \tilde{T}Q$. Convergence follows easily from the law of large numbers since, for any fixed pair (x, a), the values $Q_t(x, a)$ are simple time averages of $c_t + \gamma(\mathcal{N}Q)(x_{t+1})$ for the appropriate time steps when $(x, a) = (x_t, a_t)$. This is akin to the process of using RL to compute an improved approximation of Q^* from a fixed function Q.

The approximation of $Q^* = \tilde{T}Q^*$ comes, then, from the "optimistic" (in the sense of Bertsekas & Tsitsiklis (1996)) replacement of Q in the above iteration by Q_t. That is, we are trying to apply the operator $\tilde{T}$ to a moving target. The corresponding process, called Q-learning (Watkins & Dayan 1992), is

$$\hat{Q}_{t+1} = T_t(\hat{Q}_t, \hat{Q}_t). \tag{4}$$

Whereas the converge of Q_t given by Equation (3) is a simple consequence of stochastic approximation, the convergence $\hat{Q}_t$ given by Equation (4), Q-learning, is not so straightforward. Specifically, notice that the componentwise investigation of the process of Equation (4) is no longer possible since $\hat{Q}_{t+1}(x, a)$ depends on the values of $\hat{Q}_t$ at state-action pairs different from (x, a)—not like the case of Q_{t+1} and Q_t in Equation (3).

Interestingly, a large number of algorithms that can be viewed as methods for finding the fixed point of an operator T by defining an appropriate sequence of random T_t operators. For these definitions, the sequence of functions as defined in Equation (3) converges to $\tilde{T}Q$ for all functions Q. Our main result is, then, that under certain additional conditions on T_t, the iteration in Equation (4) will converge to the fixed point of $\tilde{T}$. In this way, we will be able to prove the convergence of a wide range of reinforcement-learning algorithms all at once. For example, we will get a convergence proof for Q-learning (Section 3.1), adaptive real-time dynamic programming (Barto, Bradtke, & Singh 1995) (the iteration $v_{t+1} = T(p_t, c_t)v_t$ outlined earlier), model-based reinforcement learning (Section 3.2), Q-learning with multi-state updates (Section 3.3), Q-learning for Markov games (Section 3.4), risk-sensitive reinforcement learning (Section 3.5), and many other related algorithms.

2 The Convergence Theorem

Most learning algorithms are, at their heart, fixed-point computations. This is because their basic structure is to apply an update rule repeatedly to seek a situation where learning is no longer possible nor desired. At this point, the learned information would be at a fixed point—additional applications of the update rule have no effect on the representation of the learned information.

In this section, we present a convergence theorem for a particular class of fixed-point computations that are particularly relevant to reinforcement learning. It may also have broader application in the analysis of learning algorithms, but we restrict our attention to reinforcement learning here.

2.1 Definitions and Theorem

Let $T : \mathcal{B} \to \mathcal{B}$ be an arbitrary operator, where $\mathcal{B}$ is a normed vector space with norm $\|.\|$.[1] Let $\mathcal{T} = (T_0, T_1, \ldots, T_t, \ldots)$ be a sequence of random operators, T_t mapping $\mathcal{B} \times \mathcal{B}$ to $\mathcal{B}$. We investigate the conditions under which the iteration $f_{t+1} = T_t(f_t, f_t)$ can be used to find the fixed point of T, provided that $\mathcal{T} = (T_0, T_1, \ldots, T_t, \ldots)$ approximates T in the sense defined next.

Definition 1 *Let $F \subseteq \mathcal{B}$ be a subset of $\mathcal{B}$ and let $\mathcal{F}_0 : F \to 2^{\mathcal{B}}$ be a mapping that associates subsets of $\mathcal{B}$ with the elements of F. If, for all $f \in F$ and all $m_0 \in \mathcal{F}_0(f)$, the sequence generated by the recursion $m_{t+1} = T_t(m_t, f)$ converges to Tf in the norm of $\mathcal{B}$ with probability 1, then we say that $\mathcal{T}$* approximates T *for initial values from $\mathcal{F}_0(f)$ and on the set $F \subseteq \mathcal{B}$. Further, we say that $\mathcal{T}$* approximates T at a certain point $f \in \mathcal{B}$ *and for initial values from $F_0 \subseteq \mathcal{B}$ if $\mathcal{T}$ approximates T on the singleton set $\{f\}$ and the initial value mapping $\mathcal{F}_0 : F \to \mathcal{B}$ defined by $\mathcal{F}_0(f) = F_0$.*

We will also make use of the following definition.

Definition 2 *The subset $F \subseteq \mathcal{B}$ is* invariant under $T : \mathcal{B} \times \mathcal{B} \to \mathcal{B}$ *if, for all $f, g \in F$, $T(f, g) \in F$. If $\mathcal{T}$ is an operator sequence as above, then F is said to be* invariant under $\mathcal{T}$ *if for all $i \geq 0$ F is invariant under T_i.*

In many applications, it is only necessary to consider the unrestricted case in which $F = \mathcal{B}$ and $\mathcal{F}_0(f) = \mathcal{B}$ for all $f \in \mathcal{B}$. For notational clarity in such cases, the set F and mapping $\mathcal{F}_0$ will not be explicitly mentioned. The general form of the definition is important in the analysis of $\hat{Q}$-learning in Section 3.5, where the approximation property of the T_t operators hold only for a limited class of functions, in particular, for the non-overestimating ones. Thus, these definitions make it possible to express the fact that T_t approximates T only for functions in F in the space of all functions $\mathcal{B}$ and restricted to initial configurations in $\mathcal{F}_0(F)$.

The following theorem is our main result. We use the notation "w.p.1" to mean "with probability 1."

[1] In the applications below, $\mathcal{B}$ is usually the space of uniformly bounded functions over a given set, the appropriate norm being the supremum norm: $\mathcal{B} = \{f : X \to \mathbb{R} : \|f\| = \sup_{x \in X} f(x) < \infty\}$.

THEOREM 3 *Let $\mathcal{X}$ be an arbitrary set and assume that $\mathcal{B}$ is the space of bounded functions over $\mathcal{X}$, $B(\mathcal{X})$, i.e., $T : B(\mathcal{X}) \to B(\mathcal{X})$. Let v^* be a fixed point of T and let $\mathcal{T} = (T_0, T_1, \ldots)$ approximate T at v^* and for initial values from $\mathcal{F}_0(v^*)$, and assume that $\mathcal{F}_0$ is invariant under $\mathcal{T}$. Let $V_0 \in \mathcal{F}_0(v^*)$, and define $V_{t+1} = T_t(V_t, V_t)$. If there exist random functions $0 \le F_t(x) \le 1$ and $0 \le G_t(x) \le 1$ satisfying the conditions below w.p.1, then V_t converges to v^* w.p.1 in the norm of $B(\mathcal{X})$:*

1. *for all U_1 and $U_2 \in \mathcal{F}_0$, and all $x \in \mathcal{X}$,*

$$|T_t(U_1, v^*)(x) - T_t(U_2, v^*)(x)| \le G_t(x)|U_1(x) - U_2(x)|;$$

2. *for all U and $V \in \mathcal{F}_0$, and all $x \in \mathcal{X}$,*

$$|T_t(U, v^*)(x) - T_t(U, V)(x)| \le F_t(x)(\|v^* - V\| + \lambda_t),$$

where $\lambda_t \to 0$ w.p.1. as $t \to \infty$;

3. *for all $k > 0$, $\Pi_{t=k}^{n} G_t(x)$ converges to zero uniformly in x as $n \to \infty$; and,*

4. *there exists $0 \le \gamma < 1$ such that for all $x \in \mathcal{X}$ and large enough t,*

$$F_t(x) \le \gamma(1 - G_t(x)).$$

Note that from the conditions of the theorem and the additional condition that T_t approximates T at every function $V \in B(\mathcal{X})$, it follows that T is a contraction operator at v^* with index of contraction γ (that is, T is a pseudo-contraction at v^* in the sense of Bertsekas & Tsitsiklis (1989)).[2]

One of the most noteworthy aspects of this theorem is that it shows how to reduce the problem of approximating v^* to the problem of approximating T at a particular point V (in particular, it is enough that T can be approximated at v^*); in many cases, the latter is much easier to achieve and also to prove. For example, the theorem makes the convergence of Q-learning a consequence of the classical Robbins-Monro theory (Robbins & Monro 1951).

Conditions 1, 2, and 3 are standard for this type of result; the first two are Lipschitz conditions on the two parameters of the operator sequence $\mathcal{T} = (T_0, T_1, \ldots)$ and Condition 3 is a learning-rate condition.

[2]The proof of this goes as follows: Let $V, U_0, V_0 \in B(\mathcal{X})$ be arbitrary and let $U_{t+1} = T_t(U_t, V)$ and $V_{t+1} = T_t(V_t, v^*)$. Let $\delta_t(x) = |U_t(x) - V_t(x)|$. Then, using Conditions 1 and 2 of Theorem 3, we get that $\delta_{t+1}(x) \le G_t(x)\delta_t(x) + \gamma(1 - G_t(x))\|V - v^*\|$. By Condition 3, $\prod_{t=0}^{\infty} G_t(x) = 0$, and, thus, $\limsup_{t\to\infty} \delta_t(x) \le \gamma\|V - v^*\|$ (see, e.g., the proof of Lemma 12 of Section A.1). Since T_t approximates T at v^* and also at V, we have that $U_t \to TV$ and $V_t \to Tv^*$ w.p.1. Thus, δ_t converges to $\|TV - Tv^*\|$ w.p.1 and, thus, $\|TV - Tv^*\| \le \gamma\|V - v^*\|$ holds w.p.1. However, this equation contains only non-random objects and thus it must hold everywhere or nowhere. Note that if Condition 1 were not restricted to v^*, then following this argument we would get that T is a contraction with index γ.

The most restrictive of the conditions of the theorem is Condition 4, which links the values of $G_t(x)$ and $F_t(x)$ through some quantity $\gamma < 1$. If it were somehow possible to update the values synchronously over the entire state space, i.e., if $V_{t+1}(x)$ depended on $V_t(x)$ only, then the process would converge to v^* even when $\gamma = 1$ provided that it were still the case that $\prod_{t=n}^{\infty}(F_t + G_t) = 0$ ($n \geq 0$) uniformly in x. In the more interesting asynchronous case, when $\gamma = 1$, the long-term behavior of V_t is not immediately clear; it may even be that V_t converges to something other than v^* or that it diverges depending on the strictness of the inequalities of Condition 4 and Inequality (22) (see Section A). The requirement that $\gamma < 1$ insures that the use of outdated information in the asynchronous updates does not cause a problem in convergence.

Note that this theorem relates to results from standard stochastic approximation, but extends them in a useful way. In particular, stochastic approximation is traditionally concerned with the problem of solving for some value under the assumption that the observed values are corrupted by a source of noise. The algorithms then need to find the sought value while canceling noise, often via some form of averaging. The general convergence theorem of this paper is not directly related to averaging out noise, but it includes this as possibility (for example, when used with noisy processes such as Q-learning in Section 3.1). In this sense, this work extends the general area of stochastic approximation by relating it to the contraction properties and fixed-point computations central to dynamic programming. In addition, the present emphasis is on asynchronous processes—more precisely, to unbalanced asynchronous processes where the update rate of different components is not fixed nor does it converge to a distribution over the components under which each component has a positive probability (assuming a finite number of components). This latter type of process can be handled using ODE (ordinary differential equation) methods (Kushner & Yin 1997), although this is not the approach taken here.

It would be possible, nevertheless, to extend the theorem such that in the Lipschitz conditions we used a conditional expectation with respect to an appropriate sequence of σ-fields, which are different from the usual history spaces; we intentionally did not move in this more direction to keep the audience a bit broader.

Section A provides all the necessary pieces for proving Theorem 3. Readers interested primarily in applications can skip the majority of this material, instead focusing on the applications presented next in Section 3. Before covering applications, we present another useful result.

2.2 Relaxation Processes

In this section, we prove a corollary of Theorem 3 for relaxation processes of the form

$$V_{t+1}(x) = (1 - f_t(x))V_t(x) + f_t(x)[P_t V_t](x), \tag{5}$$

where $0 \leq f_t(x) \leq 1$ is a relaxation parameter converging to zero and the sequence $P_t : B(\mathcal{X}) \to B(\mathcal{X})$ is a randomized version of an operator T in the sense that the "averages"

$$U_{t+1}(x) = (1 - f_t(x))U_t(x) + f_t(x)[P_t V](x) \tag{6}$$

converge to TV w.p.1, where $V \in B(\mathcal{X})$. A number of reinforcement-learning algorithms, such as Q-learning with single, or multi-state updates (Section 3.3), take the form of this process, which makes it worthy of study. It is important to note that while $V_{t+1}(x)$ depends on $V_t(y)$ for all $y \in \mathcal{X}$ since $P_t V_t$ depends on all the components of V_t, $U_{t+1}(x)$ depends only on $U_t(x)$, $x \in \mathcal{X}$: the different components are decoupled. This greatly simplifies the proof of convergence of Equation (6). Usually, the following so-called conditional averaging lemma is used to show that the process of Equation (6) converges to TV.

LEMMA 4 (CONDITIONAL AVERAGING LEMMA) *Let $\mathcal{F}_t$ be an increasing sequence of σ-fields, let $0 \le \alpha_t$ and w_t be random variables such that α_t and w_{t-1} are $\mathcal{F}_t$ measurable. Assume that the following hold w.p.1: $E[w_t|\mathcal{F}_t, \alpha_t \neq 0] = A$, $E[w_t^2|\mathcal{F}_t] < B < \infty$, $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < C < \infty$ for some $B, C > 0$. Then, the process*

$$Q_{t+1} = (1 - \alpha_t)Q_t + \alpha_t w_t$$

converges to A w.p.1.

Note that this lemma generalizes the Robbins-Monro Theorem in that, here, α_t is allowed to depend on the past of the process, which will prove to be essential in our case. It is also less general than the Robbins-Monro Theorem since $E[w_t|\mathcal{F}_t, \alpha_t \neq 0]$ is not allowed to depend on Q_t. The proof of this Lemma can be found in Appendix C.

COROLLARY 5 *Consider the process generated by the iteration of Equation (5), where $0 \le f_t(x) \le 1$. Assume that the process defined by*

$$U_{t+1}(x) = (1 - f_t(x))U_t(x) + f_t(x)[P_t v^*](x) \tag{7}$$

converges to v^* w.p.1. Assume further that the following conditions hold:

1. *there exist number $0 < \gamma < 1$ and a sequence $\lambda_t \ge 0$ converging to zero w.p.1 such that $\|P_t V - P_t v^*\| \le \gamma \|V - v^*\| + \lambda_t$ holds for all $V \in B(\mathcal{X})$;*
2. *$0 \le f_t(x) \le 1$, $t \ge 0$ and $\sum_{t=1}^{n} f_t(x)$ converges to infinity uniformly in x as $n \to \infty$.*

Then, the iteration defined by Equation (5) converges to v^* w.p.1.

Note that if $f_t(x) \to 0$ uniformly in x and w.p.1 then the condition $f_t(x) \le 1$ is automatically satisfied for large enough t.

Proof. Let the random operator sequence $T_t : B(\mathcal{X}) \times B(\mathcal{X}) \to B(\mathcal{X})$ be defined by

$$T_t(U, V)(x) = (1 - f_t(x))U(x) + f_t(x)[P_t V](x).$$

We know T_t approximates T at v^*, since, by assumption, the process defined in Equation (7) converges to TV for all $V \in B(\mathcal{X})$. Moreover, observe that V_t as defined by Equation (5) satisfies $V_{t+1} = T_t(V_t, V_t)$. Because of Assumptions 1 and 2, it can be readily verified that the Lipschitz coefficients $G_t(x) = 1 - f_t(x)$, and $F_t(x) = \gamma f_t(x)$ satisfy the rest of the conditions of Theorem 3, and this yields that the process V_t converges to v^* w.p.1. $\square$

Note that, although a large number of processes of interest admit this relaxation form, there are some important exceptions. In Sections 3.2 and 3.5, we will deal with some processes that are not of the relaxation type and we will show that Theorem 3 still applies; this shows the broad utility of the convergence theorem. Another class of exceptions are formed by processes when P_t involves some additive, zero-mean, finite conditional variance noise-term that disrupts the pseudo-contraction property (see Condition 1 above) of P_t. (As we will see, this is not the case for many well-known algorithms.) With some extra work, Corollary 5 can be extended to work in these cases. As a result, a proposition almost identical to Theorem 1 of Jaakkola, Jordan, & Singh (1994) can be deduced.[3] These extensions, however, are not needed for the applications presented in this paper and introduce unneeded complications. These extensions are needed, and have been made, in the convergence analysis of SARSA (Singh *et al.* 1998). See also the work of Szepesvári (1998).

3 Analysis of Reinforcement-Learning Algorithms

In this section, we apply the results described in Section 2 to prove the convergence of a variety of reinforcement-learning algorithms.

3.1 Q-learning

In Section 1.1, we presented the Q-learning algorithm, but we repeat this definition here for the convenience of the reader. Consider an MDP with the expected

[3] The proof of this rewrites the relaxed process P_t as the sum of "noise only" (r_t) and "noise-free" ($\hat{P}_t = E[P_t|\text{history}]$) processes as was done by Jaakkola, Jordan, & Singh (1994). This is possible because of the additive structure of the process. If $Var[r_t|\text{history}]$ is bounded independently of t, then the averaging lemma (Lemma 4) yields the convergence of the process to the right values. However, the uniform bound on the variance is too restrictive, since we need to deal with the case in which the variance of the noise grows with the relaxed process V_t defined by Equation (5) and is bounded only by $Var[r_t|\text{history}] \le C(1 + \|V_t - v^*\|)^2$. This case is reduced to the bounded-noise case by breaking the noise r_t into the sum of two parts: $r_t = s_t + s_t\|U_t - v^*\|$, where s_t is defined exactly by this identity and, thus, s_t has bounded variance (and zero mean). Now, the whole process is broken up into three parts: the first part is "noise free," the second is just driven by $s_t\|U_t - v^*\|$, and the third is driven by s_t. We know the third part goes to zero, but it is far from immediate that the second part converges to zero. This is proved using the Rescaling Lemma (Lemma 14) by considering the first two parts together; the processes that are kept bounded will converge to zero. The main difficulty of the whole proof is that it is the property $E[s_t|\text{history}] = 0$ that makes these processes converge to the right values, and so the previously used machinery of taking the absolute value and estimating cannot work in this case, since in general $E[\,|s_t|\,|\,\text{history}] > 0$.

total-discounted cost criterion and with discount factor $0 \leq \gamma < 1$. Assume that at time t we are given a 4-tuple of experience $\langle x_t, a_t, y_t, c_t \rangle$, where $x_t, y_t \in \mathcal{X}$, $a_t \in \mathcal{A}$ and $c_t \in \mathbb{R}$ are the decision maker's actual and next states, the decision maker's action, and a randomized cost received at step t, respectively. We assume that the following holds on $\langle x_t, a_t, y_t, c_t \rangle$.

ASSUMPTION 3.1 (SAMPLING ASSUMPTIONS) Consider a finite MDP, $(\mathcal{X}, \mathcal{A}, c)$, where $\Pr(y|x,a)$ are the transition probabilities and $c(x,a,y)$ are the immediate costs. Let $\{(x_t, a_t, y_t, c_t)\}$ be a fixed stochastic process, and let $\mathcal{F}_t$ be an increasing sequence of σ-fields (the history spaces) for which $\{x_t, a_t, y_{t-1}, c_{t-1}, \ldots, x_0\}$ are measurable (x_0 can be random). Assume that the following hold:

1. $\Pr(y_t = y | x = x_t, a = a_t, \mathcal{F}_t) = \Pr(y|x,a)$,
2. $E[c_t | x = x_t, a = a_t, y = y_t, \mathcal{F}_t] = c(x,a,y)$ and $Var[c_t | x_t, a_t, y_t, \mathcal{F}_t]$ is bounded independently of t, and
3. y_t and c_t are independent given the history $\mathcal{F}_t$.

Note that one may set $x_{t+1} = y_t$, which corresponds to the situation in which the decision maker gains its experiences in a real system; this is in contrast to Monte-Carlo simulations, in which $x_{t+1} = y_t$ does not necessarily hold. The Q-learning algorithm is given by

$$Q_{t+1}(x,a) = (1 - \alpha_t(x,a))Q_t(x,a) + \alpha_t(x,a)\left(c_t + \gamma \min_b Q_t(y_t, b)\right), \qquad (8)$$

where $\alpha_t(x,a) = 0$ unless $(x,a) = (x_t, a_t)$; it is intended to approximate the optimal Q function Q^* of the MDP. Note that as only one component of $\alpha_t(\cdot,\cdot)$ differs from zero, only one component of $Q_t(\cdot,\cdot)$ is "updated" in each step; the resulting process is called an asynchronous process, as opposed to a synchronous process, when, in Equation (8), $\alpha_t(x,a)$ would be independent of (x,a), while c_t would depend on it: $c_t = c_t(x,a)$. The convergence of the synchronous process follows from standard stochastic approximation arguments. Theorem 3 (and Corollary 5) show that the convergence can be extended to the asynchronous process. In particular, we have the following theorem (see also the related theorems of Watkins & Dayan (1992), Jaakkola, Jordan, & Singh (1994), and Tsitsiklis (1994)).

THEOREM 6 *Consider Q-learning in a finite* MDP *where the sequence* $\langle x_t, a_t, y_t, c_t \rangle$ *satisfies Assumption 3.1. Assume that the learning rate sequence* α_t *satisfies the following:*

1. $0 \leq \alpha_t(z,a)$, $\sum_{t=0}^{\infty} \alpha_t(z,a) = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2(z,a) < \infty$, *and both hold uniformly and hold w.p.1, and*
2. $\alpha_t(x,a) = 0$ *if* $(x,a) \neq (x_t, a_t)$ *w.p.1.*

Then, the values defined by Equation (8) converge to the optimal Q function Q^* *w.p.1.*

Proof. The proof relies on the observation that Q-learning is a relaxation process, so we may apply Corollary 5.[4] We identify the state set $\mathcal{X}$ of Corollary 5 by the set of possible state-action pairs $\mathcal{X} \times \mathcal{A}$. If we let

$$f_t(x,a) = \begin{cases} \alpha_t(x,a), & \text{if } (x,a) = (x_t,a_t); \\ 0, & \text{otherwise,} \end{cases}$$

and

$$(P_t Q)(x,a) = c_t + \gamma \max_{b \in \mathcal{A}} Q(y_t, b)$$

(P_t does not depend on a), then we see that Conditions 1 and 2 of Corollary 5 on f_t and P_t are satisfied because of our Condition 1 ($\|\alpha_t(\cdot,\cdot)\| \to 0, t \to \infty$ w.p.1, so, for large enough t, $f_t(\cdot) \le 1$.) So, it remains to prove that for a fixed function $Q \in B(\mathcal{X} \times \mathcal{A})$, the process

$$\hat{Q}_{t+1}(x,a) = (1 - \alpha_t(x,a))\hat{Q}_t(x,a) + \alpha_t(x,a)\left(c_t + \gamma \min_b Q(y_t,b)\right) \quad (9)$$

converges to TQ, where T is defined by

$$(TQ)(x,a) = \sum_{y \in \mathcal{X}} \Pr(y|x,a)\left(c(x,a,y) + \gamma \min_b Q(y,b)\right). \quad (10)$$

Using the conditional averaging lemma (Lemma 4), this is straightforward. First, observe that the different components of $\hat{Q}_t$ are decoupled, i.e., $\hat{Q}_{t+1}(x,a)$ does not depend on $\hat{Q}_t(x',a')$ and vice versa whenever $(x,a) \neq (x',a')$. Thus, it is sufficient to prove the convergence of the one-dimensional process $\hat{Q}_t(x,a)$ to $(TQ)(x,a)$ for any fixed pair (x,a). So, pick up any such pair (x,a) and identify Q_t of Lemma 4 with $\hat{Q}_t(x,a)$ defined by Equation (9). Let $\mathcal{F}_t$ be the σ-field that is adapted to

$$(x_t, a_t, \alpha_t(x,a), y_{t-1}, c_{t-1}, x_{t-1}, a_{t-1}, \alpha_{t-1}(x,a), y_{t-2}, c_{t-2}, \ldots, x_0, a_0),$$

if $t \ge 1$ and let $\mathcal{F}_0$ be adapted to (x_0, a_0), $\alpha_t = \alpha_t(x,a)$, $w_t = c_t + \gamma \min_b Q(y_t, b)$. The conditions of Lemma 4 are satisfied, namely,

1. $\mathcal{F}_t$ is an increasing sequence of σ-fields by its definition;
2. $0 \le \alpha_t \le 1$ by the same property of $\alpha_t(x,a)$ (Condition 1 of Theorem 6);
3. α_t and w_{t-1} are $\mathcal{F}_t$ measurable because of the definition of $\mathcal{F}_t$;
4. $E[w_t|\mathcal{F}_t, \alpha_t \neq 0] = E[c_t + \gamma \min_b Q(y_t,b)|\mathcal{F}_t] = \sum_{y \in \mathcal{X}} \Pr(y|x,a)(c(x,a,y) + \gamma \min_b Q(y,b)) = (TQ)(x,a)$ because of the first part of Condition 2;
5. $E[w_t^2|\mathcal{F}_t]$ is uniformly bounded because y_t can take on finite values since, by assumption, $\mathcal{X}$ is finite, the bounded variance of c_t given the past (see the second part of Condition 2) and the independence of c_t and y_t (Condition 3);

[4] Alternatively, one could directly apply Theorem 3, but we felt it more convenient to introduce Corollary 5 for use here and later.

6. $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ (Condition 1).

Thus, we get that $\hat{Q}_{t+1}(x, a)$ converges to $E[w_t|\mathcal{F}_t, \alpha_t \neq 0] = (TQ)(x, a)$, which proves the theorem. $\square$

The proof of the convergence of Q-learning provided by Theorem 6, while not particularly simpler than earlier proofs, does serve as an example of how Theorem 3 (specifically, Corollary 5) can be used to prove the convergence of a reinforcement-learning algorithm. Similar arguments appear in later sections in proofs of several novel theorems.

To reiterate, our approach attempts to decouple the difficulties related to estimation (learning the correct values) from those of asynchronous updates, which is inherent when control and learning are interleaved. This means that, besides checking some obvious conditions, the convergence proofs for Q-learning and other algorithms reduce to the proof that a one-dimensional version of the learning rule (the estimation part) works as intended.

3.2 Model-Based Reinforcement Learning

Q-learning shows that optimal value functions can be estimated without ever explicitly learning the transition and cost functions; however, estimating these functions can make more efficient use of experience at the expense of additional storage and computation (Moore & Atkeson 1993). The parameters of the functions can be learned from experience by keeping statistics for each state-action pair on the expected cost and the proportion of transitions to each next state. In model-based reinforcement learning, the transition and cost functions are estimated on line, and the value function is updated according to the approximate dynamic-programming operator derived from these estimates. Interestingly, although this process is not of the relaxation form, still Theorem 3 implies their convergence for a wide variety of models and methods. In order to capture this generality, let us introduce a class of generalized MDPs. In generalized MDPs (Szepesvári & Littman 1996), the cost-propagation operator $\mathcal{Q}$ takes the special form

$$(\mathcal{Q}V)(x, a) = \bigoplus_{y \in \mathcal{X}}^{(x,a)} \left(c(x, a, y) + \gamma V(y)\right).$$

Here, $\bigoplus^{(x,a)} f(\cdot)$ might take the form $\sum_{y \in \mathcal{X}} \Pr(y|x, a) f(y)$, which corresponds to the case of expected total-discounted cost criterion, or it may take the form

$$\max_{y:\Pr(y|x,a)>0} f(y),$$

which corresponds to the case of the risk-averse worst-case total discounted cost criterion. One may easily imagine a heterogeneous criterion, when $\bigoplus^{(x,a)}$ would be of the expected-value form for some (x, a) pairs, while it would be of the worst-case criterion form for other pairs expressing a state-action dependent risk attitude of the decision maker. In general, we require only that the operation

$\bigoplus^{(x,a)} : B(\mathcal{X}) \to \mathbb{R}$ be a non-expansion with respect to the supremum-norm, i.e., that

$$\left| \bigoplus^{(x,a)} f(\cdot) - \bigoplus^{(x,a)} g(\cdot) \right| \leq \|f - g\|$$

for all $f, g \in B(\mathcal{X})$. Earlier work (Littman & Szepesvári 1996; Szepesvári & Littman 1996) provides an in-depth discussion of non-expansion operators.

As was noted above, in model-based reinforcement learning, the transition and cost functions are estimated by some quantities c_t and p_t. As long as every state-action pair is visited infinitely often, there are a number of simple methods for computing c_t and p_t that converge to the true functions. Model-based reinforcement-learning algorithms use the latest estimates of the model-parameters (e.g. c_t and p_t) to approximate operator $\mathcal{Q}$, and in particular operator $\bigoplus$. In some cases, a bit of care is needed to insure that $\bigoplus_t$, the latest estimate of $\bigoplus$, converges to $\bigoplus$, however (here, convergence should be understood in the sense that $\| \bigoplus_t f - \bigoplus f \| \to 0, t \to \infty$ holds for all $f \in B(\mathcal{X})$). There is no problem with expected-cost models; here the convergence of p_t to the transition function guarantees the convergence of $\bigoplus_t^{(x,a)} f = \sum_{y \in \mathcal{X}} p_t(x,a,y) f(y)$ to $\bigoplus$. For worst-case-cost models, it is necessary to approximate the transition function in a way that insures that the set of y such that $p_t(x,a,y) > 0$ converges to the set of y such that $\Pr(y|x,a) > 0$. This can be accomplished easily, however, by setting $p_t(x,a,y) = 0$ if no transition from x to y under a has been observed.

In this framework, the adaptive real-time dynamic-programming algorithm (Barto, Bradtke, & Singh 1995) takes the form

$$V_{t+1}(x) = \begin{cases} \min_{a \in \mathcal{A}} \bigoplus_t{}^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V_t(\cdot) \right), & \text{if } x \in \tau_t \\ V_t(x), & \text{otherwise,} \end{cases} \tag{11}$$

where $c_t(x,a,y)$ is the estimated cost-function and τ_t is the set of states updated at time step t. This algorithm is called "real time" if the decision maker encounters its experiences in the real system and $x_t \in \tau_t$, where x_t denotes the actual state of the decision maker at time step t, i.e., the value of the actual state is always updated.

Theorem 7 *Consider a finite* MDP *and, for any pair* $(x,a) \in \mathcal{X} \times \mathcal{A}$, *let* $\bigoplus_t^{(x,a)}, \bigoplus : B(\mathcal{X}) \to \mathbb{R}$. *Assume that the following hold w.p.1:*

1. $\bigoplus_t \to \bigoplus$ *in the sense that*

$$\lim_{t \to \infty} \max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \left| \bigoplus_t^{(x,a)} f(\cdot) - \bigoplus^{(x,a)} f(\cdot) \right| = 0$$

   for all functions f.

2. $\bigoplus_t^{(x,a)}$ *is a non-expansion for all* $(x,a) \in \mathcal{X} \times \mathcal{A}$ *and* t.

3. $c_t(x,a,y)$ *converges to* $c(x,a,y)$ *for all* (x,a,y).

4. $0 \leq \gamma < 1$.

5. Every state x is updated infinitely often (i.o.), that is, $x \in \tau_t$ i.o. for all $x \in \mathcal{X}$.

Then, V_t defined in Equation (11) converges to the fixed point of the operator $T : B(\mathcal{X}) \to B(\mathcal{X})$, where

$$(TV)(x) = \min_{a \in \mathcal{A}} \bigoplus_{y \in \mathcal{X}}^{(x,a)} \left(c(x,a,y) + \gamma V(y) \right).$$

Proof. We apply Theorem 3. Let the appropriate approximate dynamic-programming operator sequence $\{T_t\}$ be defined by

$$T_t(U,V)(x) = \begin{cases} \min_{a \in \mathcal{A}} \bigoplus_t^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V(\cdot) \right), & \text{if } x \in \tau_t \\ U(x), & \text{otherwise.} \end{cases}$$

Now, we prove that T_t approximates T.[5] Let $x \in \mathcal{X}$ and let $U_{t+1} = T_t(U_t, V)$. Then, $U_{t+1}(x) = U_t(x)$ if $x \notin \tau_t$. Since, in the other case, when $x \in \tau_t$, $U_{t+1}(x)$ does not depend on U_t and, since $x \in \tau_t$ i.o., it is sufficient to show that $D_t = |\min_{a \in \mathcal{A}} \bigoplus_t^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V(\cdot) \right) - (TV)(x)|$ converges to zero as $t \to \infty$. Now,

$$\begin{aligned} D_t &\leq \max_{a \in \mathcal{A}} \left| \bigoplus_t^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V(\cdot) \right) - \bigoplus^{(x,a)} \left(c(x,a,\cdot) + \gamma V(\cdot) \right) \right| \\ &\leq \max_{a \in \mathcal{A}} \left| \bigoplus_t^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V(\cdot) \right) - \bigoplus_t^{(x,a)} \left(c(x,a,\cdot) + \gamma V(\cdot) \right) \right| \\ &\quad + \max_{a \in \mathcal{A}} \left| \bigoplus_t^{(x,a)} \left(c(x,a,\cdot) + \gamma V(\cdot) \right) - \bigoplus^{(x,a)} \left(c(x,a,\cdot) + \gamma V(\cdot) \right) \right| \\ &\leq \max_{a \in \mathcal{A}} \max_{y \in \mathcal{X}} |c_t(x,a,y) - c(x,a,y)| \\ &\quad + \max_{a \in \mathcal{A}} \left| \bigoplus_t^{(x,a)} \left(c(x,a,\cdot) + \gamma V(\cdot) \right) - \bigoplus^{(x,a)} \left(c(x,a,\cdot) + \gamma V(\cdot) \right) \right|, \end{aligned}$$

where we made use of the triangle inequality and Condition 2. The first term on the right-hand side converges to zero because of our Condition 3, while the second term converges to zero because of our Condition 1. This, together with Condition 5 implies that $D_t \to 0$, which, since $x \in \mathcal{X}$ was arbitrary, shows that T_t indeed approximates T.

Returning to checking the conditions of Theorem 3, we find that the functions

$$G_t(x) = \begin{cases} 0, & \text{if } x \in \tau_t; \\ 1, & \text{otherwise,} \end{cases}$$

[5] Note that $U_{t+1} = T_t(U_t, V)$ can be viewed as a composite of two converging processes and, thus, Theorem 15 of Section A.3 could easily be used to prove that $U_t \to TV$. Here, we give another direct argument.

and

$$F_t(x) = \begin{cases} \gamma, & \text{if } x \in \tau_t; \\ 0, & \text{otherwise,} \end{cases}$$

satisfy the remaining conditions of Theorem 3, as long as $\bigoplus_t$ is a non-expansion for all t (which holds by Condition 2), each x is included in the τ_t sets infinitely often (this is required by Condition 3 of Theorem 3), and the discount factor γ is less than 1 (see Condition 4 of Theorem 3). But, these hold by our Conditions 5 and 4, respectively, and, therefore, the proof is complete. $\square$

This theorem generalizes the results of Gullapalli & Barto (1994), which deal only with the expected total-discounted cost criterion, i.e., when

$$\bigoplus_{y \in \mathcal{X}}^{(x,a)} f(y) = \sum_{y \in \mathcal{X}} \Pr(y|x,a) f(y).$$

Note that in the above argument, $\min_{a \in \mathcal{A}}$ could have been replaced by any other non-expansion operation (this holds also for the other algorithms presented in this article). As a consequence of this, model-based methods can be used to find optimal policies in MDPs, alternating Markov games, Markov games (Littman 1994), risk-sensitive models (Heger 1994), and exploration-sensitive (i.e., SARSA) models (John 1994; Rummery & Niranjan 1994). Also, if we fix $c_t(x,a,y) = c(x,a,y)$ and $p_t(x,a,y) = \Pr(y|x,a)$ for all t, $x, y \in \mathcal{X}$ and $a \in \mathcal{A}$, this result implies that asynchronous dynamic programming converges to the optimal value function (Barto, Sutton, & Watkins 1989; Bertsekas & Tsitsiklis 1989; Barto, Bradtke, & Singh 1995).

3.3 Q-learning With Multi-State Updates

Ribeiro (1995) argued that the use of available information in Q-learning is inefficient: in each step it is only the actual state and action whose Q value is re-estimated. The training process is local both in space and time. If some *a priori* knowledge of the "smoothness" of the optimal Q value is available, then one can make the updates of Q-learning more efficient by introducing a so-called "spreading mechanism," which updates the Q values of state-action pairs in the vicinity of the actual state-action pair as well.

The rule studied by Ribeiro is as follows: let Q_0 be arbitrary and

$$\begin{aligned} Q_{t+1}(z,a) &= (1 - \alpha_t(z,a) s(z,a,x_t)) Q_t(z,a) + \\ &\quad \alpha_t(z,a) s(z,a,x_t) \left(c_t + \gamma \min_a Q_t(y_t,a) \right), \end{aligned} \tag{12}$$

where $\alpha_t(z,a) \geq 0$ is the learning rate associated with the state-action pair (z,a), which is 0 if $a \neq a_t$, $s(z,a,x)$ is a fixed "similarity" function satisfying $0 \leq s(z,a,x)$, and $\langle x_t, a_t, y_t, c_t \rangle$ is the experience of the decision maker at time t. The difference between the above and the standard Q-learning rule is that here we may allow $\alpha_t(z,a) \neq 0$ even if $x_t \neq z$, i.e., the values of states different from the state actually experienced may be updated, too. The similarity function

$s(z,a,x)$ weighs the relative strength at which updates occur at z when state x is experienced. (One could also use a similarity that extends spreading over actions, or time. The similarity could be made time-dependent by making it converge to the Dirac-delta function at an appropriate rate. In this way, convergence to the optimal Q-function could be recovered (Ribeiro & Szepesvári 1996). For simplicity, we do not consider these cases here.)

Our aim here is to show that, under the appropriate conditions, this learning rule converges; also, we will be able to derive a bound on how far the limit values of this rule are from the optimal Q function of the underlying MDP.

THEOREM 8 *Consider the learning rule of Equation (12), assume that the sampling conditions of Assumption 3.1 are satisfied, and further assume that*

1. *the states, x_t, are sampled from a probability distribution $p^\infty \in \Pi(\mathcal{X})$*
2. $0 \le s(z,a,\cdot)$ *and* $s(z,a,z) \ne 0$,
3. $\alpha_t(z,a) = 0$ *if* $a \ne a_t$, *and* $0 \le \alpha_t(z,a)$, $\sum_{t=0}^\infty \alpha_t(z,a) = \infty$, $\sum_{t=0}^\infty \alpha_t^2(z,a) < \infty$.

Then, Q_t, as given by Equation (12), converges to the fixed point of the operator $\hat{T} : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$,

$$(\hat{T}Q)(z,a) = \sum_{x\in\mathcal{X}} \hat{s}(z,a,x) \sum_{y\in\mathcal{X}} \Pr(y|x,a)\left(c(x,a,y) + \gamma \min_b Q(y,b)\right), \qquad (13)$$

where

$$\hat{s}(z,a,x) = \frac{s(z,a,x)p^\infty(x)}{\sum_y s(z,a,y)p^\infty(y)}.$$

Proof. Note that $\hat{T}$ as defined is a contraction with index γ since $\sum_x \hat{s}(z,a,x) = 1$ for all (z,a). Since the process of Equation (12) is of the relaxation type, we apply Corollary 5. As in the proof of the convergence of Q-learning in Theorem 6, we identify the state set $\mathcal{X}$ of Corollary 5 by the set of possible state-action pairs $\mathcal{X}\times\mathcal{A}$. We let

$$(P_tQ)(x,a) = c_t + \gamma \max_{b\in\mathcal{A}} Q(y_t,b),$$

but now we set $f_t(z,a) = s(z,a,x_t)\alpha_t(z,a)$. The conditions on f_t and P_t are satisfied by Condition 2, and the conditions on the learning rates $\alpha_t(x,a)$ are also satisfied (in particular, $\|\alpha_t(\cdot,\cdot)\| \to 0, t\to\infty$ w.p.1, so $f_t(\cdot) \le 1$ for large enough t), so it remains to prove that for a fixed function $Q \in B(\mathcal{X}\times\mathcal{A})$, the process

$$\begin{aligned} Q_{t+1}(z,a) &= (1-\alpha_t(z,a)s(z,a,x_t))Q_t(z,a) + \\ &\quad \alpha_t(z,a)s(z,a,x_t)\left(c_t + \gamma \min_b Q(y_t,b)\right), \end{aligned} \qquad (14)$$

converges to $\hat{T}Q$. We apply a modified form of the conditional averaging lemma (Lemma 4), which concerns processes of the form $Q_{t+1} = (1-\alpha_t s_t)Q_t + \alpha_t s_t w_t$ and is presented and proved in Appendix C as Lemma 20. This lemma states that, under some bounded-variance conditions, Q_t converges to $E[s_t w_t|\mathcal{F}_t]/E[s_t|\mathcal{F}_t]$, where $\mathcal{F}_t$ is an increasing sequence of σ-fields that is adapted to $\{s_{t-1}, w_{t-1}, \alpha_t\}$. In our case, let $\mathcal{F}_t$ of Lemma 20 be the σ-field generated by

$$(a_t, \alpha_t(x,a), y_{t-1}, c_{t-1}, x_{t-1}, \ldots, a_1, \alpha_1(x,a), y_0, c_0, x_0, a_0, \alpha_0(x,a)))$$

if $t \geq 1$ and let $\mathcal{F}_0$ be adapted to $(a_0, \alpha_0(x,a))$. Easily,

$$(\hat{T}Q)(z,a) = \frac{E[s(z,a,x_t)(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b))|\mathcal{F}_t, \alpha_t(z,a) \neq 0]}{E[s(z,a,x_t)|\mathcal{F}_t, \alpha_t(z,a) \neq 0]}.$$

$E[s^2(z,a,x_t)(c_t + \gamma \min_a Q(y_t,a))^2|x_t, \mathcal{F}_t] < B < \infty$ for some $B > 0$ by Conditions 2 and 3. Moreover, $E[s(z,a,x_t)|\mathcal{F}_t] = \sum_{x\in\mathcal{X}} p^\infty(x) s(z,a,x) > 0$ by Conditions 1 and 2, and $E[s^2(z,a,x_t)|\mathcal{F}_t] = \sum_{x\in\mathcal{X}} p^\infty(x) s^2(z,a,x) < \hat{B} < \infty$, for some $\hat{B} > 0$, by the finiteness of $\mathcal{X}$. Finally, $\alpha_t(z,a)$ obviously satisfies the assumptions of Lemma 20 and, therefore, all the conditions of the quoted lemma are satisfied. So, $Q_t(z,a)$, defined by Equation (14), converges to $(\hat{T}Q)(z,a)$. $\square$

Note that if we set $s(z,a,x) = 1$ if and only if $z = x$ and $s(z,a,x) = 0$ then Equation (12) becomes the same as the Q-learning update rule of Equation (8). However, the condition on the sampling of x_t is quite strict, so Theorem 8 is less general than Theorem 6.

It is interesting and important to ask how close is $\hat{Q}^*$, the fixed point of $\hat{T}$ where $\hat{T}$ is defined by Equation (13), to the "true" optimal Q^*, which is the fixed point of T defined by Equation (10). The following proposition (related to Theorem 6.2 of Gordon (1995)) answers this question in the general case. The specific case we are concerned with here comes from taking the operator F to be

$$(FQ)(z,a) = \sum_{x\in\mathcal{X}} \hat{s}(z,a,x) Q(x,a).$$

PROPOSITION 9 *Let $\mathcal{B}$ be a normed vector space, $T : \mathcal{B} \to \mathcal{B}$ be a contraction and $F : \mathcal{B} \to \mathcal{B}$ be a non-expansion. Further, let $\hat{T} : \mathcal{B} \to \mathcal{B}$ be defined by $\hat{T}Q = F(TQ)$, $Q \in \mathcal{B}$. Let Q^* be the fixed point of T and $\hat{Q}^*$ be the fixed point of $\hat{T}$. Then,*

$$\|\hat{Q}^* - Q^*\| \leq \frac{2\inf_Q\{\, \|Q - Q^*\| \,:\, FQ = Q \,\}}{1-\gamma}. \tag{15}$$

Proof. Let Q denote an arbitrary fixed point of F.[6] Then, since $\|T\hat{Q}^* - Q^*\| = \|T\hat{Q}^* - TQ^*\| \leq \gamma\|\hat{Q}^* - Q^*\|$, $\|\hat{Q}^* - Q^*\| = \|FT\hat{Q}^* - Q^*\| \leq \|FT\hat{Q}^* - Q\| + \|Q - Q^*\| = \|FT\hat{Q}^* - FQ\| + \|Q - Q^*\| \leq \|T\hat{Q}^* - Q\| + \|Q - Q^*\| \leq \|T\hat{Q}^* - Q^*\| + 2\|Q - Q^*\| \leq \gamma\|\hat{Q}^* - Q^*\| + 2\|Q - Q^*\|$. Rearranging the terms and taking the infimum over the possible Qs yields the bound of Inequality (15). $\square$

[6] If F does not have a fixed point, then the infimum is infinity, so the proposition is still correct (trivially).

Inequality (15) helps us to define the spreading coefficients $s(z,a,x)$. Specifically, let $n>0$ be fixed and let

$$s(z,a,x)=\begin{cases}1, & \text{if } i/n \le Q^*(z,a), Q^*(x,a) < (i+1)/n \text{ for some } i;\\ 0, & \text{otherwise,}\end{cases} \tag{16}$$

then we get that the learned Q function is within $1/n$ of the optimal Q function Q^*.[7] Of course, the problem with this definition is that we do not know in advance the optimal Q function, so we can't define $s(z,a,x)$ precisely as shown in Equation (16). However, the above example gives us a guideline for how to define a "good" spreading function (by good here, we mean that the error introduced by the spreading function is kept as small as possible): $s(z,a,x)$ should be small (zero) for states z and x for which $Q^*(z,a)$ and $Q^*(x,a)$ differ substantially, otherwise $s(z,a,x)$ should take on larger values. In other words, it is a good idea to define $s(z,a,x)$ as the degree of expected difference between $Q^*(z,a)$ and $Q^*(x,a)$.

Note that the above learning process is closely related to learning on aggregated states (Bertsekas & Castañon 1989; Schweitzer 1984; Singh, Jaakkola, & Jordan 1995). An aggregated state is simply a subset $\mathcal{X}_i$ of $\mathcal{X}$. The idea is that the size of the Q table (which stores the $Q_t(x,a)$ values) could be reduced if we assigned a common value to all of the states in the same aggregated state $\mathcal{X}_i$. By defining the aggregated states $\{\mathcal{X}_i\}_{i=1,2,\dots,n}$ in a clever way, one may achieve that the common value assigned to the states in $\mathcal{X}_i$ are close to the actual values of the states. In order to avoid ambiguity, the aggregated states should be disjoint, i.e., $\{\mathcal{X}_i\}$ should form a partitioning of $\mathcal{X}$. For convenience, let us introduce the equivalence relation "$\approx$" among states with the definition that $x \approx y$ if and only if x and y are elements of the same aggregated state.

Now, observe that if we set $s(z,a,x)=1$ if and only if $z \approx x$ and $s(z,a,x)=0$ otherwise, then, by iterating Equation (12), the values of any two state-action pairs will be equal when the corresponding states are in the same aggregated states. In mathematical terms, $Q_t(x,a)=Q_t(z,a)$ will hold for all x,z with $x \approx z$, i.e., Q_t is compatible with the "$\approx$" relation. Of course, this holds only if the initial estimate Q_0 is compatible with the "$\approx$" relation, too. The compatibility of the estimates with the partitioning enables us to rewrite Equation (12) in terms of the indices of the aggregated states:

$$Q_{t+1}(i,a)=\begin{cases}(1-\alpha_t(i,a))Q_t(i,a) & \\ \quad +\alpha_t(i,a)\left(c_t+\gamma \min_a Q_t(i(y_t),a)\right), & \text{if } i(x_t)=i, a_t=a;\\ Q_t(i,a), & \text{otherwise.}\end{cases} \tag{17}$$

Here, $i(z)$ stands for the index of the aggregated state to which z belongs. Then, we have the following:

PROPOSITION 10 *Let* $\underline{n}=\{1,2,\dots,n\}$ *and let* $\tilde{T}: B(\underline{n}\times\mathcal{A}) \to B(\underline{n}\times\mathcal{A})$ *be*

[7] The $s(z,a,x)$ function can also be defined in terms of the absolute difference of $Q^*(z,a)$ and $Q^*(x,a)$. This may lead to better approximation bounds, but it doesn't allow us to develop the "equivalence class" discussion later in this section.

given by

$$(\tilde{T}\tilde{Q})(i,a) = \sum_{x\in\mathcal{X}_i, y\in\mathcal{X}} P(\mathcal{X}_i,x)\,\Pr(y|x,a)\left(c(x,a,y)+\gamma\min_b \tilde{Q}(i(y),b)\right),$$

where $P(\mathcal{X}_i,x) = p^\infty(x)/\sum_{y\in\mathcal{X}_i} p^\infty(y)$*. Then, under the conditions of Theorem 8,* $Q_t(i,a)$ *defined by Equation (17) converges to the fixed point of* $\tilde{T}$*.*

Proof. Since $\tilde{T}$ is a contraction, its fixed point is well defined. The proposition follows from Theorem 8.[8] Indeed, let $Q_0(x,a) = Q(i(x),a)$ for all (x,a) pair. Then, Theorem 8 yields that $Q_t(x,a)$ converges to $\hat{Q}^*(x,a)$, where $\hat{Q}^*$ is the fixed point of operator $\hat{T}$. Observe that $\hat{s}(z,a,x) = 0$ if $z \not\approx x$ and $\hat{s}(z,a,x) = P(\mathcal{X}_i(z),x)$ if $z \approx x$. The properties of $\hat{s}$ yield that if Q is compatible with the partitioning (i.e., if $Q(x,a) = Q(z,a)$ if $x \approx z$), then $\hat{T}Q$ will also be compatible with the partitioning, since the right-hand side of the following equation depends only on the index of z and $\tilde{Q}(i,b)$, which is the common Q value of state-action pairs for which the state is the element of $\mathcal{X}_i$:

$$\begin{aligned}(\hat{T}Q)(z,a) &= \sum_{x\in\mathcal{X}_{i(z)}, y\in\mathcal{X}} P(\mathcal{X}_{i(z)},x)\,\Pr(y|x,a)\left(c(x,a,y)+\gamma\min_b Q(y,b)\right)\\ &= \sum_{x\in\mathcal{X}_{i(z)}, y\in\mathcal{X}} P(\mathcal{X}_{i(z)},x)\,\Pr(y|x,a)\left(c(x,a,y)+\gamma\min_b \tilde{Q}(i(y),b)\right).\end{aligned}$$

Since $\hat{T}$ is compatible with the partitioning, its fixed point must be compatible with the partitioning, and, further, the fixed point of $\tilde{T}$ and that of $\hat{T}$ are equal when we identify functions of $B(\mathcal{X}\times\mathcal{A})$ that are compatible with the given partitioning with the corresponding functions of $B(\underline{n}\times\mathcal{A})$ in the natural way. Putting the above pieces together yields that Q_t as defined in Equation (17) converges to the fixed point of $\tilde{T}$. □

Note that Inequality (15) still gives an upper bound for the largest difference between $\hat{Q}^*$ and Q^*, and Equation (16) defines how a $1/n$-precise partitioning should ideally look.

The above results can be trivially extended to the case in which the decision maker follows a fixed stationary policy that guarantees that every state-action pair is visited infinitely often and that there exists a non-vanishing limit probability distribution over the states $\mathcal{X}$. However, if the actions that are chosen depend on the estimated Q_t values, then there does not seem to be any simple way to ensure the convergence of Q_t unless randomized policies are used during learning whose rate of change is slower than that of the estimation process (Konda & Borkar 1997).

[8] Note that Corollary 5 could also be applied directly to this rule. Another way to deduce the above convergence result is to consider the learning rule over the aggregated states as a standard Q-learning rule for an induced MDP whose state space is $\{\mathcal{X}_1,\ldots,\mathcal{X}_n\}$, whose transition probabilities are $p(\mathcal{X}_i,a,\mathcal{X}_j) = \sum_{x\in\mathcal{X}_i,y\in\mathcal{X}_j} p^\infty(x)\Pr(y|x,a)$ and whose cost-structure is $c(\mathcal{X}_i,a,\mathcal{X}_j) = \sum_{x\in\mathcal{X}_i,y\in\mathcal{X}_j} p^\infty(x)\Pr(y|x,a)c(x,a,y)$.

Other extensions of the results of this section are to the case in which the spreading function s decays to one that guarantees convergence to an optimal Q function, and the case in which learned values are a function of the chosen exploratory actions (the so-called SARSA algorithm) (John 1994; Rummery & Niranjan 1994; Singh & Sutton 1996; Singh *et al.* 1998).

3.4 Q-learning for Markov Games

In an MDP, a single decision maker selects actions to minimize its expected discounted cost in a stochastic environment. A generalization of this model is the alternating Markov game, in which two players, the maximizer and the minimizer, take turns selecting actions—the minimizer tries to minimize its expected discounted cost, while the maximizer tries to maximize the cost to the other player. The update rule for alternating Markov games is a simple variation of Equation (11) in which a max replaces a min in those states in which the maximizer gets to choose the action; this makes the optimality criterion discounted minimax optimality. Theorem 7 implies the convergence of Q-learning for alternating Markov games because min and max are both non-expansions (Littman 1996).

Markov games are a generalization of both MDPs and alternating Markov games in which the two players simultaneously choose actions at each step in the process (Owen 1982; Littman 1994). The basic model is defined by the tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{B}, \Pr(\cdot|\cdot,\cdot), c \rangle$ (states, min actions, max actions, transitions, and costs) and discount factor γ. As in alternating Markov games, the optimality criterion is one of discounted minimax optimality, but because the players move simultaneously, the Bellman equations take on a more complex form:

$$v^*(x) = \min_{\rho \in \Pi(\mathcal{A})} \max_{b \in \mathcal{B}} \sum_{a \in \mathcal{A}} \rho(a) \left(c(x,(a,b)) + \gamma \sum_{y \in \mathcal{X}} \Pr(y|x,(a,b)) v^*(y) \right). \quad (18)$$

In these equations, $c(x,(a,b))$ is the immediate cost for the minimizer for taking action $a \in \mathcal{A}$ in state x at the same time the maximizer takes action $b \in \mathcal{B}$, $\Pr(y|x,(a,b))$ is the probability that state y is reached from state x when the minimizer takes action a and the maximizer takes action b, and $\Pi(\mathcal{A})$ represents the set of discrete probability distributions over the set $\mathcal{A}$. The sets $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{B}$ are finite.

Optimal policies are in equilibrium, meaning that neither player has any incentive to deviate from its policy as long as its opponent adopts its policy. In every Markov game, there is a pair of optimal policies that are stationary (Owen 1982). Unlike MDPs and alternating Markov games, the optimal policies are sometimes stochastic; there are Markov games in which no deterministic policy is optimal (the classic playground game of "rock, paper, scissors" is of this type). The stochastic nature of optimal policies explains the need for the optimization over probability distributions in the Bellman equations, and stems from the fact that players must avoid being "second guessed" during action selection. An equivalent set of equations to Equation (18) can be written with a stochastic choice for the maximizer, and also with the roles of the minimizer and maximizer reversed.

The obvious way to extend Q-learning to Markov games is to define the cost-propagation operator $\mathcal{Q}$ analogously to the case of MDPs from the fixed-point Equation (18). This yields the definition $\mathcal{Q} : B(\mathcal{X}) \to B(\mathcal{X} \times \Pi(\mathcal{A}))$ as

$$(\mathcal{Q}V)(x,\rho) = \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) \left(c(x,(a,b)) + \gamma \sum_{y\in\mathcal{X}} \Pr(y|x,(a,b))V(y) \right).$$

Note that $\mathcal{Q}$ is a contraction with index γ.

Unfortunately, because $Q^* = \mathcal{Q}v^*$ would be a function of an infinite space (all discrete probability distributions over the action space), we have to choose another representation. If we redefine $\mathcal{Q}$ to map functions over $\mathcal{X}$ to functions over the finite space $\mathcal{X} \times (\mathcal{A} \times \mathcal{B})$:

$$[\mathcal{Q}V](x,(a,b)) = \left(c(x,(a,b)) + \gamma \sum_{y\in\mathcal{X}} \Pr(y|x,(a,b))V(y) \right)$$

then, for $Q^* = \mathcal{Q}v^*$, the fixed-point Equation (18) takes the form

$$v^*(y) = \min_{\rho\in\Pi(\mathcal{A})} \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) Q^*(y,(a,b)).$$

Applying $\mathcal{Q}$ on both sides yields

$$Q^*(x,(a_0,b_0)) = c(x,(a_0,b_0)) + \gamma \sum_{y\in\mathcal{X}} \Pr(y|x,(a_0,b_0)) \min_{\rho\in\Pi(\mathcal{A})} \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) Q^*(y,(a,b)).$$

The corresponding Q-learning update rule (Littman 1994) given the step t experience $\langle x_t, a_t, b_t, y_t, c_t \rangle$ has the form

$$Q_{t+1}(x_t,(a_t,b_t)) = \tag{19}$$
$$(1-\alpha_t(x_t,(a_t,b_t)))Q_t(x_t,(a_t,b_t)) + \alpha_t(x_t,(a_t,b_t)) \left(c_t + \gamma \left(\bigotimes Q_t \right)(y_t) \right),$$

where

$$\left(\bigotimes Q \right)(y) = \min_{\rho\in\Pi(\mathcal{A})} \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) Q(y,(a,b))$$

and the values of Q_t not shown in Equation (19) are left unchanged.

This update rule is identical to Equation (8), except that actions are taken to be simultaneous pairs for both players. The results of Section 3.1 prove that this rule converges to the optimal Q function under the proper sampling conditions. It is worth noting that similar results could also be derived by extending previous Q-learning convergence proofs.

In general, it is necessary to solve a linear program to compute $(\bigotimes Q)(y)$. It is possible that Theorem 3 can be combined with the results of Vrieze & Tijs (1982) on solving Markov games by "fictitious play" to prove the convergence of a linear-programming-free version of Q-learning for Markov games. Hu & Wellman (1998) extended the results of this section to non-zero-sum games.

3.5 Risk-Sensitive Reinforcement Learning

The optimality criterion for MDPs in which only the *worst* possible value of the next state makes a contribution to the value of a state is called the worst-case total discounted cost criterion. An optimal policy under this criterion is one that avoids states for which a bad outcome is possible, even if it is not probable; for this reason, the criterion has a risk-averse quality to it. Following Heger (1994), this can be expressed by changing the expectation operator of MDPs used in the definition of the cost-propagation operator $\mathcal{Q}$ to

$$(\mathcal{Q}V)(x,a) = \max_{y:\Pr(y|x,a)>0} \left(c(x,a,y) + \gamma V(y)\right).$$

The argument in Section 3.2 shows that model-based reinforcement learning can be used to find optimal policies in risk-sensitive models, as long as the transition probabilities are estimated in a way that preserves its zero vs. non-zero nature in the limit. Analogously, a Q-learning-like algorithm, called $\hat{Q}$-learning (Q-hat learning) can be shown and will be shown here to converge to optimal policies. In essence, the learning algorithm uses an update rule that is quite similar to the rule in Q-learning with a max replacing exponential averaging and no learning rate, but has the additional requirement that the initial Q function be set optimistically; that is, $Q_0(x,a) \le Q^*(x,a)$ for all x and a.[9] Like Q-learning, this learning algorithm is a generalization of the LRTA* algorithm of Korf (1990) to stochastic environments.

THEOREM 11 *Assume that both $\mathcal{X}$ and $\mathcal{A}$ are finite. Let*

$$Q_{t+1}(x,a) = \begin{cases} \max\left(Q_t(x,a), c_t + \gamma \min_{b\in\mathcal{A}} Q_t(y_t,b)\right); & \text{if } (x,a) = (x_t,a_t);\\ Q_t(x,a); & \text{otherwise,}\end{cases}$$

where $\langle x_t, a_t, y_t, c_t\rangle$ is the experience of the decision maker at time t, y_t is selected at random according to $\Pr(\cdot|x,a)$, and c_t is a random variable satisfying the following condition: If $t_n(x,a,y)$ is the subsequence of ts for which $(x,a,y) = (x_t,a_t,y_t)$, then $c_{t_n(x,a,y)} \le c(x,a,y)$ and $\limsup_{n\to\infty} c_{t_n(x,a,y)} = c(x,a,y)$ w.p.1. Then, Q_t converges to $Q^* = \mathcal{Q}v^*$ provided that $Q_0 \le Q^*$ and every state-action pair is updated infinitely often.

Proof. The proof is another application of Theorem 3, but here the definition of the appropriate operator sequence T_t needs some more care. Let the set of "critical states" for a given (x,a) pair be given by

$$\mathcal{M}(x,a) = \left\{ y \in \mathcal{X} \;\middle|\; \Pr(y|x,a) > 0,\ Q^*(x,a) = c(x,a,y) + \gamma \min_{b\in\mathcal{A}} Q^*(y,b) \right\}.$$

[9] The necessity of this condition is clear since in the $\hat{Q}$-learning algorithm we need to estimate the operator $\max_{y:\Pr(y|x,a)>0}$ from the observed transitions, and the underlying iterative method is consistent with $\max_{y:\Pr(y|x,a)>0}$ only if the initial estimate is overestimating. Since we require only that T_t approximates T at Q^*, it is sufficient for the initial value of the process to satisfy $Q_0 \le Q^*$. Note that $Q_0 = -M/(1-\gamma)$ satisfies this condition, where $M = \max_{(x,a,y)} c(x,a,y)$.

The set $\mathcal{M}(x,a)$ is non-empty, since $\mathcal{X}$ is finite. Since the costs c_t satisfy $c_{t_n(x,a,y)} \le c(x,a,y)$ and

$$\limsup_{n\to\infty} c_{t_n(x,a,y)} = c(x,a,y),$$

we may also assume (by possibly redefining $t_n(x,a,y)$ to become a subsequence of itself) that

$$\lim_{n\to\infty} c_{t_n(x,a,y)} = c(x,a,y). \tag{20}$$

Now, let $T(x,a,y) = \{t_k(x,a,y) \,|\, k \ge 0\}$ and $T(x,a) = \cup_{y\in\mathcal{M}(x,a)} T(x,a,y)$. Consider the following sequence of random operators:

$$T_t(Q',Q)(x,a) = \begin{cases} \max\left(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b), Q'(x,a)\right); & \text{if } t \in T(x,a), \\ Q'(x,a); & \text{otherwise,} \end{cases}$$

and the sequence $Q'_0 = Q_0$ and $Q'_{t+1} = T_t(Q'_t, Q'_t)$ with the set of possible initial values taken from

$$\mathcal{F}_0 = \{Q \in B(\mathcal{X}\times\mathcal{A}) \,|\, Q(x,a) \le Q^*(x,a) \text{ for all } (x,a) \in \mathcal{X}\times\mathcal{A}\}.$$

Clearly, $\mathcal{F}_0$ is invariant under T_t. We claim that it is sufficient to consider the convergence of Q'_t. Since there are no more updates (increases of value) in the sequence Q'_t than in Q_t, we have that $Q^* \ge Q_t \ge Q'_t$ and, thus, if Q'_t converged to Q^*, then necessarily so did Q_t. It is immediate that T_t approximates T at Q^* (since w.p.1 there exist an infinite number of $t > 0$ such that $t \in T(x,a)$), and also that we can safely define the Lipschitz function

$$G_t(x,a) = \begin{cases} 0; & \text{if } (x,a) = (x_t,a_t) \text{ and } y_t \in \mathcal{M}(x,a), \\ 1; & \text{otherwise,} \end{cases}$$

since $T_t(Q,Q^*)(x,a) = Q^*(x,a)$ if $(x,a) = (x_t,a_t)$ and $y_t \in \mathcal{M}(x,a)$.

Now, let us bound the quantity $|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)|$. For this, assume first that $t \in T(x,a)$. This means that $(x,a) = (x_t,a_t)$ and $y_t \in \mathcal{M}(x,a)$. Since $Q_0 \in \mathcal{F}_0$ and $\mathcal{F}_0$ is invariant we may assume that the functions Q, Q' below satisfy $Q, Q' \le Q^*$ (they over non-overestimating):

$$\begin{aligned} &|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)| \\ &\quad\le\quad (c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b)) - \max(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b), Q'(x,a)) \\ &\quad\le\quad \Big(c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b)\Big) - \Big(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b)\Big) \\ &\quad\le\quad \gamma\|Q^* - Q\| + |c(x,a,y_t) - c_t|. \end{aligned} \tag{21}$$

We have used the fact that $T_t(Q',Q^*)(x,a) \ge T_t(Q',Q)(x,a)$ (since T_t is monotone in its second variable) and that

$$\begin{aligned} T_t(Q',Q^*)(x,a) &\le \max\left(c(x,a,y_t) + \gamma \min_{b\in A} Q^*(y_t,b), Q'(x,a)\right) \\ &= c(x,a,y_t) + \gamma \min_{b\in A} Q^*(y_t,b) \end{aligned}$$

since $y_t \in \mathcal{M}(x,a)$ and $Q' \leq Q^*$.

Let $\sigma_t(x,a) = |c(x,a,y_t) - c_t|$. Note that by Equation (20),

$$\lim_{t\to\infty, t\in T(x,a)} \sigma_t(x,a) = 0$$

w.p.1. In the other case (when $t \notin T(x,a)$),

$$|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)| = 0.$$

Therefore,

$$|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)| \leq F_t(x,a)(\|Q - Q^*\| + \lambda_t),$$

where

$$F_t(x,a) = \begin{cases} \gamma; & \text{if } t \in T(x,a), \\ 0; & \text{otherwise,} \end{cases}$$

and $\lambda_t = \sigma_t(x_t,a_t)/\gamma$ if $t \in T(x,a)$, and $\lambda_t = 0$, otherwise. Thus, we get that Condition 2 of Theorem 3 is satisfied since λ_t converges to zero w.p.1 (which holds because there are only a finite number of (x,a) pairs).

Condition 3 of the same theorem is satisfied if and only if $t \in T(x,a)$ i.o. But, this must hold due to the assumptions on the sampling of (x_t,a_t) and y_t, and since $\Pr(y|x,a) > 0$ for all $y \in \mathcal{M}(x,a)$. Finally, Condition 4 is satisfied, since for all t, $F_t(x) = \gamma(1 - G_t(x))$, and so Theorem 3 yields that $\hat{Q}$-learning converges to Q^* w.p.1. □

In this section, we have proven Theorem 11 concerning the convergence of $\hat{Q}$-learning under a worst-case total discounted cost criterion, first stated by Heger (1994). Note that, once again, this process is not of the relaxation type (that is, Equation (5)) but Theorem 3 still applies to it.

Another interesting thing to note is that, in spite of the absence of any learning rate sequence, $\hat{Q}$-learning converges. It does require that the initial Q function be set optimistically, however.

4 Conclusions

This article presents and proves a general convergence theorem useful for analyzing reinforcement-learning algorithms. This theorem enables proofs of convergence of some learning algorithms outside of the scope of the earlier theorems; novel results include the convergence of reinforcement-learning algorithms in game environments and under a risk-sensitive assumption. At the same time, the theorem enables the derivation of the earlier general convergence results. However, the generality of these earlier results is not always needed—as for Q-learning—and the present approach shows simple ways to prove the convergence of practical algorithms. The purpose of the theorem is to extract the basic tools needed to prove convergence and decouple difficulties rising from stochasticity and asynchronousness: The theorem enables the treatment of non-stochastic algorithms like asynchronous value iteration, along with stochastic ones (Q-learning) with asynchronous components. (Synchronous stochastic algorithms are subject of standard stochastic approximation theory.) Note also that the methods developed in this paper can be used to obtain an asymptotic convergence rate results for averaging-type asynchronous algorithms (Szepesvári 1997).

Similarly to Jaakkola, Jordan, & Singh (1994) and Tsitsiklis (1994), we develop the connection between stochastic approximation theory and reinforcement learning in MDPs. Our work is similar in structure and spirit to that of Jaakkola, et al. We believe the form of Theorem 3 makes it particularly convenient for proving the convergence of reinforcement-learning algorithms; our theorem reduces the proof of the convergence of an asynchronous process to a simpler proof of convergence of a corresponding synchronized one. This idea enables us to prove the convergence of asynchronous stochastic processes whose underlying synchronous process is not of the Robbins-Monro type (e.g., risk-sensitive MDPs, model-based algorithms, etc.) in a unified way.

There are many areas of interest in the theory of reinforcement learning that we would like to address in future work. The results in this paper concern reinforcement-learning in discounted models ($\gamma < 1$), and there are important non-contractive reinforcement-learning scenarios, for example, reinforcement learning under an average-reward criterion (Schwartz 1993; Mahadevan 1996).

In principle, the analysis of actor-critic-type learning algorithms (Williams & Baird 1993; Konda & Borkar 1997) could benefit from the type of convergence results developed in this paper. Our early attempts to apply these techniques to actor-critic learning have been unsuccessful, however. The fact that the space of policies is not continuous presents serious difficulties for the type of metric-space arguments used here, and we have yet to find a way to achieve the required contraction properties in the policy-update operators.

Another possible direction for future research is to apply the modern ODE (ordinary differential equation) theory of stochastic approximations. If one is given a definite exploration strategy, then this theory may yield results about convergence, speed of convergence, finite sample size effects, optimal exploration, limiting distribution of Q-values, etc.

The presented mathematical tools help us to understand how reinforcement-learning problems can be attacked in a well-motivated way and pave the way to more general and powerful algorithms.

Acknowledgements

This work was supported in part by grants NSF-IRI-97-02576-CAREER (Littman), and OTKA Grant No. F20132 (Szepesvári) and a grant by the Hungarian Ministry of Education under contract number FKFP 1354/1997 (Szepesvári).

A Proof of the Convergence Theorem

This section proves Theorem 3 (Section 2.1).

Let U_0 be a value function in $\mathcal{F}_0(v^*)$ and let $U_{t+1} = T_t(U_t, v^*)$. Since T_t approximates T at v^*, U_t converges to $Tv^* = v^*$ w.p.1 uniformly over $\mathcal{X}$. We will show that $\|U_t - V_t\|$ converges to zero w.p.1, which implies that V_t converges to v^*. Let

$$\delta_t(x) = |U_t(x) - V_t(x)|$$

and let

$$\Delta_t(x) = |U_t(x) - v^*(x)|.$$

We know that $\Delta_t(x)$ converges to zero because U_t converges to v^*.

By the triangle inequality and the conditions on T_t (invariance of $\mathcal{F}_0$ and the Lipschitz conditions), we have

$$\begin{aligned}
\delta_{t+1}(x) &= |U_{t+1}(x) - V_{t+1}(x)| \\
&= |T_t(U_t, v^*)(x) - T_t(V_t, V_t)(x)| \\
&\leq |T_t(U_t, v^*)(x) - T_t(V_t, v^*)(x)| + |T_t(V_t, v^*)(x) - T_t(V_t, V_t)(x)| \\
&\leq G_t(x)|U_t(x) - V_t(x)| + F_t(x)(\|v^* - V_t\| + \lambda_t) \\
&= G_t(x)\delta_t(x) + F_t(x)(\|v^* - V_t\| + \lambda_t) \\
&\leq G_t(x)\delta_t(x) + F_t(x)(\|v^* - U_t\| + \|U_t - V_t\| + \lambda_t) \\
&\quad = G_t(x)\delta_t(x) + F_t(x)(\|\delta_t\| + \|\Delta_t\| + \lambda_t).
\end{aligned} \tag{22}$$

It is not difficult to prove that a process δ_t satisfying Inequality (22) converges to zero when, in Inequality (22), the "perturbation term" $\|\Delta_t\| + \lambda_t$ equals zero for all $t \geq 0$. This is shown in Lemma 12 in Section A.1 below. The problem of transferring this proof to the general case when $\|\Delta_t\| + \lambda_t > 0$ is that the boundedness of δ_t cannot be checked directly. However, the proof still applies for a modified process $\hat{\delta}_t$, which is the version of δ_t kept bounded by rescaling it; i.e., $\hat{\delta}_t$ is defined in the same way as δ_t, but whenever $\|\hat{\delta}_t\|$ grows above a fixed limit $C > 0$, we rescale it (by multiplying it appropriately) so that $\|\hat{\delta}_t\| \leq C$ holds for all $t \geq 0$. In Section A.2, we prove that it is indeed sufficient that $\hat{\delta}_t$ converges to zero since δ_t is a homogeneous process, i.e., it can be written in the form $\delta_{t+1} \leq G_t(\delta_t, \|\Delta_t\| + \lambda_t)$ such that $\beta G_t(x, y) = G_t(\beta x, \beta y)$ holds for all $\beta > 0$. Finally, still in Section A.2, we finish the proof of Theorem 3 by showing that $\hat{\delta}_t$ converges to zero (Lemma 16).

It is interesting to note the connection between this last lemma and the general problem of unboundedness of stochastic approximation processes. When using the ODE technique, it is typical that probability one convergence can be proved only when the boundedness of the process is proven beforehand (Benveniste, Métivier, & Priouret 1990). Then, the boundedness is shown using other techniques. As such, this lemma may also find some applications in standard stochastic approximation. Another way to cope with unboundedness, known as the projection technique, is advocated by Kushner & Clark (1978), Ljung (1977), and others. This technique modifies the original process in a way that its boundedness is guaranteed. It is interesting to note that the proof of the lemma below shows that if one of the artificially bound-kept process converges (to zero), then so does the original, under the additional assumptions of the lemma.

Note that our results, most importantly in the proof of Lemma 16, use the methods of Jaakkola, Jordan, & Singh (1994); our theorem illustrates the strength of their approach.

A.1 Convergence in the Perturbation-Free Case

First, we prove our version of Lemma 2 of Jaakkola, Jordan, & Singh (1994), which concerns the convergence of the above process δ_t from the process of Inequality (22) in the perturbation-free case. Note that both our assumptions and our proof are slightly different from theirs—we make some further comments on this after the proof.

LEMMA 12 *Let $\mathcal{Z}$ be an arbitrary set and consider the random sequence*

$$x_{t+1}(z) = G_t(z)x_t(z) + F_t(z)\|x_t\|, z \in \mathcal{Z} \tag{23}$$

where $x_1, F_t, G_t \geq 0$ are random processes, and $\|x_1\| < C < \infty$ w.p.1 for some $C > 0$. Assume that for all k $\lim_{n\to\infty} \prod_{t=k}^{n} G_t(z) = 0$ uniformly in z w.p.1 and $F_t(z) \leq \gamma(1 - G_t(z))$ for some $0 \leq \gamma < 1$ w.p.1. Then, $\|x_t\|$ converges to 0 w.p.1.

Proof. We will prove that for each $\varepsilon, \delta > 0$ there exist an index $M = M(\varepsilon, \delta) < \infty$ (possibly random, see Appendix B) such that

$$\Pr\left(\sup_{t\geq M} \|x_t\| < \delta\right) > 1 - \varepsilon. \tag{24}$$

Fix arbitrary $\varepsilon, \delta > 0$ and a sequence of numbers $p_1, \ldots, p_t, \ldots$ satisfying $0 < p_t < 1$ to be chosen later.

We have that

$$\begin{aligned} x_{t+1}(z) &= G_t(z)x_t(z) + F_t(z)\|x_t\| \\ &\leq G_t(z)\|x_t\| + F_t(z)\|x_t\| \\ &= (G_t(z) + F_t(z))\|x_t\| \\ &\leq \|x_t\|, \end{aligned}$$

since, by assumption, $G_t(z) + F_t(z) \leq G_t(z) + \gamma(1 - G_t(z)) \leq 1$. Thus, we have that $\|x_{t+1}\| \leq \|x_t\|$ for all t and, particularly, $\|x_t\| \leq C_1 = \|x_1\|$ holds for all t. Consequently, the process

$$y_{t+1}(z) = G_t(z)y_t(z) + \gamma(1 - G_t(z))C_1, \tag{25}$$

with $y_1 = x_1$, estimates the process $\{x_t\}$ from above: $0 \leq x_t \leq y_t$ holds for all t. The process y_t converges to γC_1 w.p.1 uniformly over $\mathcal{Z}$. (Subtract γC_1 from both sides to get $(y_{t+1}(z) - \gamma C_1) = G_t(z)(y_t(z) - \gamma C_1)$. Now, convergence of $\|y_t - \gamma C_1\|$ follows since $\lim_{n\to\infty} \prod_{t=k}^{n} G_t(z) = 0$ uniformly in z). Therefore,

$$\limsup_{t\to\infty} \|x_t\| \leq \gamma C_1$$

w.p.1. Thus, there exists an index, say M_1, for which if $t > M_1$ then $\|x_t\| \leq (1+\gamma)/2\, C_1$ with probability p_1. Assume that up to some index $i \geq 1$ we have found numbers M_i such that when $t > M_i$ then

$$\|x_t\| \leq \left(\frac{1+\gamma}{2}\right)^i C_1 = C_{i+1} \tag{26}$$

holds with probability $p_1 p_2 \ldots p_i$. Now, let us restrict our attention to those events for which Inequality (26) holds. Then, we see that the process

$$\begin{aligned} y_{M_i} &= x_{M_i} \\ y_{t+1}(z) &= G_t(z) y_t(z) + \gamma(1 - G_t(z)) C_{i+1},\ t \geq M_i \end{aligned}$$

bounds x_t from above from the index M_i. Now, the above argument can be repeated to obtain an index M_{i+1} such that Inequality (26) holds for $i+1$ with probability $p_1 p_2 \ldots p_i p_{i+1}$.

Since $(1+\gamma)/2 < 1$, there exists an index k for which $((1+\gamma)/2)^k C_1 < \varepsilon$. Then, we get that Inequality (24) is satisfied when we choose $p_1, \ldots, p_k$ in a way that $p_1 p_2 \ldots p_k \geq 1 - \varepsilon$ and we set $M = M_k$ (where M_k will depend upon $p_1, p_2, \ldots, p_k$). □

A significant contrast between Lemma 12 and the results of Jaakkola, Jordan, & Singh (1994) lies in the use of the constants F_t and G_t. Jaakkola et al. relate these quantities through their conditional expectations ($E[F_t|P_t] \leq \gamma(1 - E[G_t|P_t])$, where P_t is the history of the process), whereas our result uses the relation $F_t \leq \gamma(1 - G_t)$. Ours is a stronger assumption, but it has the advantage of simplifying the mathematics while still being sufficient for a wide range of applications. If only the conditional expectations are related, then two additional assumptions are needed, namely that

$$\begin{aligned} \lim_{N\to\infty} \|\sum_{t=0}^{N} F_t^2\| &= 0, \quad \text{and} \\ \lim_{n\to\infty} \|\sum_{t=0}^{N} G_t^2\| &= 0 \end{aligned} \tag{27}$$

w.p.1 and a version of the conditional averaging lemma (Lemma 4, presented in Section 2.2) can be used to show the convergence of $\|x_t\|$ to zero. Note that F_t and G_t correspond to the Lipschitz functions of Theorem 3, respectively. In some of the applications (see Sections 3.2 and 3.5), the appropriate Lipschitz constants do not satisfy this assumption (Equation 27), but Condition 4 is satisfied in all the applications. These applications include the model-based and risk-sensitive

RL algorithms. Note that our approach still requires the above assumptions in the proof of Q-learning (Section 3.1).

When the process of Equation (23) is subject to decaying perturbations, say ε_t (see, e.g., the process of Inequality (22)), then the proof no longer applies. The problem is that $\|x_t\| \le \|x_1\|$ (or $\|x_{M+t}\| \le \|x_M\|$, for large enough M) can no longer be ensured without additional assumptions. For $x_{t+1}(z) \le \|x_t\|$ to hold, we would need that $\gamma\varepsilon_t \le (1-\gamma)\|x_t\|$, but if $\liminf_{t\to\infty} \|x_t\| = 0$ (which, in fact, is a consequence of what should be proved), then we could not check this relation *a priori*. Thus, we choose another way to prove that the perturbed process converges to zero. Notice that the key idea in the above proof is to bound x_t by y_t. This can be done if we assume that x_t is kept bounded artificially, e.g., by scaling. The next subsection shows that such a change of x_t does not effect its convergence properties.

A.2 The Rescaling of Two-Variable Homogeneous Processes

The next lemma is about two-variable homogeneous processes, that is, processes of the form

$$x_{t+1} = G_t(x_t, \varepsilon_t), \tag{28}$$

where $G_t : \mathcal{B} \times \mathcal{B} \to \mathcal{B}$ is a homogeneous random function ($\mathcal{B}$ denotes a normed vector space, as before), i.e.,

$$G_t(\beta x, \beta\varepsilon) = \beta G_t(x, \varepsilon) \tag{29}$$

holds for all $\beta > 0$, x and ε.[10] We are interested in the question of whether x_t converges to zero or not. Note that when the inequality defining δ_t (Inequality (22)) is an equality, it becomes a homogeneous process in the above sense. The lemma below says that, under additional technical conditions, it is enough to prove the convergence of a modified process *that is kept bounded by rescaling* to zero, namely the process

$$y_{t+1} = \begin{cases} G_t(y_t, \varepsilon_t), & \text{if } \|G_t(y_t, \varepsilon_t)\| \le C; \\ C\, G_t(y_t, \varepsilon_t)/\|G_t(y_t, \varepsilon_t)\|, & \text{otherwise,} \end{cases} \tag{30}$$

where $C > 0$ is an arbitrary fixed number.

We denote the solution of Equation (28) corresponding to the initial condition $x_0 = w$ and the sequence $\varepsilon = \{\varepsilon_k\}$ by $x_t(w, \varepsilon)$. Similarly, we denote the solution of Equation (30) corresponding to the initial condition $y_0 = w$ and the sequence ε by $y_t(w, \varepsilon)$.

DEFINITION 13 *We say that the process x_t is* insensitive to finite perturbations *of ε if it holds that if $x_t(w, \varepsilon)$ converges to zero then so does $x_t(w, \varepsilon')$, where*

[10] Jaakkola, Jordan, & Singh (1994) considered a question similar to that investigated in our Lemma 14 for the case of *single-variable* homogeneous processes, which would correspond to the case when $\varepsilon_t = 0$ for all $t \ge 0$ (see Equation (28)). The single-variable case follows from our result. The extension to two variables is needed in our proof of the lemma in Section A.3.

$\varepsilon'(\omega)$ *is an arbitrary sequence that differs only in a finite number of terms from* $\varepsilon(\omega)$*, where the bound on the number of differences is independent of* ω*. Further, we say that the process* x_t *is* insensitive to scaling of ε by numbers smaller than 1*, if for all random* $0 < c \leq 1$ *it holds that if* $x_t(w, \varepsilon)$ *converges to zero then so does* $x_t(w, c\varepsilon)$*.*

LEMMA 14 (RESCALING LEMMA) *Let us fix an arbitrary positive number* C *and an arbitrary* w_0 *and sequence* ε*. Then, a homogeneous process* $x_t(w_0, \varepsilon)$ *converges to zero w.p.1 provided that (i)* x_t *is insensitive to finite perturbations of* ε*; (ii)* x_t *is insensitive to the scaling of* ε *by numbers smaller than one and (iii)* $y_t(w_0, \varepsilon)$ *converges to zero.*

Proof. We state that

$$y_t(w, \varepsilon) = x_t(d_t w, c_{t\cdot}\varepsilon) \tag{31}$$

for some sequences $\{c_{t\cdot}\}$ and $\{d_t\}$, where $c_{t\cdot} = (c_{t0}, c_{t1}, \ldots, c_{ti}, \ldots)$, $\{c_{t\cdot}\}$ and $\{d_t\}$ satisfy $0 < d_t, c_{ti} \leq 1$, and $c_{ti} = 1$ if $i \geq t$. Here, the product of the sequences $c_{t\cdot}$ and ϵ should be understood to be componentwise: $(c_{t\cdot}\epsilon)_i = c_{ti}\epsilon_i$. Note that $y_t(w, \varepsilon)$ and $x_t(w, \varepsilon)$ depend only on $\varepsilon_0, \ldots, \varepsilon_{t-1}$. Thus, it is possible to prove Equation (31) by constructing the appropriate sequences c_t and d_t.

Set $c_{0i} = d_i = 1$ for all $i = 0, 1, 2, \ldots$. Then, Equation (31) holds for $t = 0$. Let us assume that $\{c_i, d_i\}$ is defined in a way that Equation (31) holds for t. Let S_t be the "scaling coefficient" of y_t at step $(t+1)$ ($S_t = 1$ if there is no scaling, otherwise $0 < S_t < 1$ with $S_t = C/\|G_t(y_t, \varepsilon_t)\|$):

$$\begin{aligned} y_{t+1}(w, \varepsilon) &= S_t G_t(y_t(w, \varepsilon), \varepsilon_t) \\ &= G_t(S_t y_t(w, \varepsilon), S_t \varepsilon_t) \\ &= G_t(S_t x_t(d_t w, c_t \varepsilon), S_t \varepsilon_t). \end{aligned}$$

We claim that

$$S x_t(w, \varepsilon) = x_t(Sw, S\varepsilon) \tag{32}$$

holds for all w, ε and $S > 0$.

For $t = 0$, this obviously holds. Assume that it holds for t. Then,

$$\begin{aligned} S x_{t+1}(w, \varepsilon) &= S G_t(x_t(w, \varepsilon), \varepsilon_t) \\ &= G_t(S x_t(w, \varepsilon), S\varepsilon_t) \\ &= G_t(x_t(Sw, S\varepsilon), S\varepsilon_t) \\ &= x_{t+1}(Sw, S\varepsilon). \end{aligned}$$

Thus,

$$y_{t+1}(w, \varepsilon) = G_t(x_t(S_t d_t w, S_t c_t \varepsilon), S_t \varepsilon_t),$$

and we see that Equation (31) holds if we define $c_{t+1,i}$ as $c_{t+1,i} = S_t c_{ti}$ if $0 \leq i \leq t$, $c_{t+1,i} = 1$ if $i > t$ and $d_{t+1} = S_t d_t$.

Thus, we get that with the sequences

$$c_{t,i} = \begin{cases} \prod_{j=i}^{t-1} S_j, & \text{if } i < t; \\ 1, & \text{otherwise,} \end{cases}$$

$d_0 = 1$, and

$$d_{t+1} = \prod_{i=0}^{t} S_i,$$

Equation (31) is satisfied for all $t \geq 0$.

Now, assume that we want to prove for a particular sequence ε and initial value w that

$$\lim_{t\to\infty} x_t(w,\varepsilon) = 0 \tag{33}$$

holds w.p.1. It is enough to prove that Equation (33) holds with probability $1-\delta$ when $\delta > 0$ is an arbitrary, sufficiently small number.

We know that $y_t(w,\varepsilon) \to 0$ w.p.1. We may assume that $\delta < C$. Then, there exists an index $M = M(\delta)$ such that if $t > M$ then

$$\Pr(\|y_t(w,\varepsilon)\| < \delta) > 1-\delta. \tag{34}$$

Now, let us restrict our attention to those events ω for which $\|y_t(w,\varepsilon(\omega))\| < \delta$ for all $t > M$: $A_\delta = \{\omega : \|y_t(w,\varepsilon)(\omega)\| < \delta\}$. Since $\delta < C$, we get that there is no rescaling after step M: $S_t(\omega) = 1$ if $t > M$. Thus, $c_{t,i} = c_{M+1,i}$ for all $t \geq M+1$ and i, and specifically $c_{t,i} = 1$ if $i, t \geq M+1$. Similarly, if $t > M$ then $d_{t+1}(\omega) = \prod_{i=0}^{M} S_i(\omega) = d_{M+1}(\omega)$. By Equation (31), we have that if $t > M$ then

$$y_t(w,\varepsilon(\omega)) = x_t(d_{M+1}(\omega)w, c_{M+1}(\omega)\varepsilon(\omega)).$$

Thus, it follows from our assumption concerning y_t that $x_t(d_{M+1}(\omega)w, c_{M+1}\varepsilon(\omega))$ converges to zero almost everywhere (a.e.) on A_δ and, consequently, by Equation (32), $x_t(w, c_{M+1}\varepsilon(\omega)/d_{M+1(\omega)})$ also converges to zero a.e. on A_δ. Since x_t is insensitive to finite perturbations, and since, in c_{M+1}, only a finite number of entries differs from 1, $x_t(w,\varepsilon(\omega)/d_{M+1}(\omega))$ also converges to zero, and, further, since $d_{M+1}(\omega) < 1$, $x_t(w,\varepsilon(\omega)) = x_t(w, d_{M+1}(\omega)(\varepsilon(\omega)/d_{M+1}(\omega)))$ converges to zero, too (x_t is insensitive to scaling of ε by d_{M+1}). All these hold with probability at least $1-\delta$, since, by Equation (34), $\Pr(A_\delta) > 1-\delta$. Since δ was arbitrary, the lemma follows. □

A.3 Convergence of Perturbed Processes

We have established that Inequality (22) converges if not perturbed. We now extend this to more general perturbed processes so we can complete the proof of Theorem 3.

For this we need a theorem that gives sufficient conditions under which the cascade of two converging processes still converges. The theorem itself is very simple (the proof requiring just elementary analysis). However, it is quite useful in the context of the current work, with applications to the convergence of both model-based reinforcement-learning in Section 3.2 and to that of the perturbed difference sequence in Lemma 16. Therefore, although, this theorem is somewhat of a digression from the main stream of the present work, it provides a convenient analysis of a common phenomenon.

THEOREM 15 *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces, $U_t : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ ($t = 0,1,2,\ldots$) be a sequence of mappings, and $\theta_t \in \mathcal{Y}$ be an arbitrary sequence. Let $\theta_\infty \in \mathcal{Y}$ and $x_\infty \in \mathcal{X}$. Consider the sequences $x_{t+1} = U_t(x_t, \theta_\infty)$, and $y_{t+1} = U_t(y_t, \theta_t)$, and suppose that x_t and θ_t converge to x_∞ and θ_∞, respectively, in the norm of the appropriate spaces.*

Let L_k^θ be the uniform Lipschitz index of $U_k(x,\theta)$ with respect to θ at θ_∞ and, similarly, let $L_k^\mathcal{X}$ be the uniform Lipschitz index of $U_k(x,\theta_\infty)$ with respect to x.[11] *Then, if the Lipschitz constants $L_t^\mathcal{X}$ and L_t^θ satisfy the relations $L_t^\theta \le C(1-L_t^\mathcal{X})$, and $\prod_{m=t}^\infty L_m^\mathcal{X} = 0$, where $C > 0$ is some constant and $t = 0,1,2,\ldots$, then $\lim_{t\to\infty} \|y_t - x_\infty\| = 0$.*

Proof. For simplicity, assume that $x_0 = y_0$; this assumption could be easily removed at the cost of additional complication. Since $\|y_t - x_\infty\| \le \|y_t - x_t\| + \|x_t - x_\infty\|$, it is sufficient to prove that $\|y_t - x_t\|$ converges to zero. Since $\|x_{t+1} - y_{t+1}\| = \|U_t(x_t,\theta_\infty) - U_t(y_t,\theta_t)\|$,

$$\begin{aligned} \|x_{t+1} - y_{t+1}\| &\le \|U_t(x_t,\theta_\infty) - U_t(y_t,\theta_\infty)\| + \|U_t(y_t,\theta_\infty) - U_t(y_t,\theta_\infty)\| \\ &\le L_t^\mathcal{X}\|x_t - y_t\| + L_t^\theta\|theta_t - \theta_\infty\|. \end{aligned}$$

Then, it is easy to prove by induction on r that

$$\|x_r - y_r\| \le \sum_{s=0}^r \|\theta_s - \theta_\infty\| L_s^\theta \prod_{t=s+1}^r L_t^\mathcal{X} \tag{35}$$

(the assumption $x_0 = y_0$ was used here). Now, fix an arbitrary positive ε. We want to prove that for r big enough, $\|x_r - y_r\| < \varepsilon$.

Using $L_s^\theta \le C(1 - L_s^\mathcal{X})$, we get from Equation (35)

$$\|x_r - y_r\| \le C \sum_{s=0}^r \|\theta_s - \theta_\infty\|(1 - L_s^\mathcal{X}) \prod_{t=s+1}^r L_t^\mathcal{X}.$$

Now, consider $S_r = \sum_{s=0}^r \|\theta_s - \theta_\infty\|(1 - L_s^\mathcal{X}) \prod_{t=s+1}^r L_t^\mathcal{X}$. Let K be big enough such that $\sup_{s>K} \|\theta_s - \theta_\infty\| < \varepsilon/(2C)$ (such a K exists since θ_s converges to θ_∞). Now, split the sum into two parts (assuming $r > K+1$):

$$\begin{aligned} S_r &= \sum_{s=0}^K \|\theta_s - \theta_\infty\|(1 - L_s^\mathcal{X}) \prod_{t=s+1}^r L_t^\mathcal{X} + \sum_{s=K+1}^r \|\theta_s - \theta_\infty\|(1 - L_s^\mathcal{X}) \prod_{t=s+1}^r L_t^\mathcal{X} \\ &\le \max_{0\le s\le K} \|\theta_s - \theta_\infty\| \sum_{s=0}^K (1 - L_s^\mathcal{X}) \prod_{t=s+1}^r L_t^\mathcal{X} + \sup_{s>K} \|\theta_s - \theta_\infty\| \sum_{s=K+1}^r (1 - L_s^\mathcal{X}) \prod_{t=s+1}^r L_t^\mathcal{X}. \end{aligned}$$

[11] That is, for all $x \in \mathcal{X}$ and $\theta \in \mathcal{Y}$ $\|U_k(x,\theta) - U_k(x,\theta_\infty)\| \le L_k^\theta\|\theta - \theta_\infty\|$ and for all $x, y \in \mathcal{X}$ $\|U_k(x,\theta_\infty) - U_k(y,\theta_\infty)\| \le L_k^\mathcal{X}\|x - y\|$.

For r big enough, the first term is easily seen to become smaller than $\varepsilon/(2C)$, since $\max_{0\le s\le K}\|\theta_s-\theta_\infty\|$ is finite and the rest is the sum of $K+1$ sequences converging to zero (since $\prod_{t=s+1}^{r} L_t^{\chi}$ converges to zero). In the second term, $\sup_{s>K}\|\theta_s-\theta_\infty\|\le\varepsilon/(2C)$, by assumption. The sum can be further bounded by above by increasing the lower bound of the summation to 0 (here, we exploited the fact that $0\le L_t^{\chi}\le 1$). The increased sum turns out to be a telescopic sum, which in turn is equal to $1-\prod_{t=0}^{r} L_t^{\chi}$. This, in fact, converges to 1, but for our purposes it is sufficient to notice that 1 upper bounds it. Thus, for r big enough, $S_r\le\varepsilon/(2C)+\varepsilon/(2C)=\varepsilon/C$ and, therefore, $\|x_r-y_r\|\le\varepsilon$, which is what was to be proven. □

Now, we are in the position to prove that Lemma 12 is immune to decaying perturbations.

LEMMA 16 *Assume that the conditions of Lemma 12 are satisfied but Equation (23) is replaced by*

$$x_{t+1}(z)=G_t(z)x_t(z)+F_t(z)(\|x_t\|+\varepsilon_t), \tag{36}$$

where $\varepsilon_t\ge 0$ and ε_t converges to zero with probability 1. Then, $x_t(z)$ still converges to zero w.p.1 uniformly over $\mathcal{Z}$.

Proof. We follow the proof of Lemma 12. First, we show that the process of Equation (36) satisfies the assumptions of the Rescaling Lemma (Lemma 14) and, thus, it is enough to consider the version of Equation (36) that is kept bounded by scaling.

First, note that x_t is a homogeneous process of the form of Equation (28) (note that Equation (29) was required to hold only for positive β). Let us prove that x_t is immune to finite perturbations of ε. To this end, assume that ε'_t differs only in a finite number of terms from ε_t and let

$$y_{t+1}(z)=G_t(z)y_t(z)+F_t(z)(\|y_t\|+\varepsilon'_t).$$

Take

$$k_t(z)=|x_t(z)-y_t(z)|.$$

Then,

$$k_{t+1}(z)\le G_t(z)k_t(z)+F_t(z)(\|k_t(z)\|+|\varepsilon_t-\varepsilon'_t|).$$

For large enough t, $\varepsilon_t=\varepsilon'_t$, so

$$k_{t+1}(z)\le G_t(z)k_t(z)+F_t(z)\|k_t(z)\|,$$

which we know to converge to zero by Lemma 12. Thus, x_t and y_t both converge or do not converge and if one converges then the other must converge to the same value.

The other requirement that we must satisfy to be able to apply the Rescaling Lemma (Lemma 14) is that x_t is insensitive to scaling of the perturbation by numbers smaller than one; let us choose a random number $0<c\le 1$ and assume

that $x_t(w,\varepsilon)$ converges to zero with probability 1. Then, since $0 \le x_t(w,c\varepsilon) \le x_t(w,\varepsilon)$, $x_t(w,c\varepsilon)$ converges to zero w.p.1, too.

Now, let us prove that the process that is obtained from x_t by keeping it bounded converges to zero. The proof is the mere repetition of the proof of Lemma 12, except a few points that we discuss now. Let us denote by $\hat{x}_t$ the process that is kept bounded and let the bound be C_1. It is enough to prove that $\|\hat{x}_t\|$ converges to zero w.p.1. Now, Equation (25) is replaced by

$$y_{t+1}(z) = G_t(z)y_t(z) + \gamma(1 - G_t(z))(C_1 + \varepsilon_t).$$

By Theorem 15, y_t still converges to γC_1, as the following bindings show: $\mathcal{X}, \mathcal{Y} := \mathbf{R}$ $\theta_t := \varepsilon_t$, $U_t(x,\theta) := G_t(z)x + \gamma(1 - G_t(z))(C_1 + \theta)$, where $z \in \mathcal{Z}$ is arbitrary. Then, $L_t^{\mathcal{X}} = G_t(z)$ and $L_t^{\theta} = \gamma(1 - G_t(z))$, satisfying the conditions of Theorem 15.

Since it is also the case that $0 \le \hat{x}_t \le y_t$, the whole argument of Lemma 12 can be repeated for the process $\hat{x}_t$, yielding that $\|\hat{x}_t\|$ converges to zero w.p.1 and, consequently, so does $\|x_t\|$. $\square$

This completes the proof of Theorem 3.

B Random Indices

Recall that, by definition, a random sequence x_t converges to zero w.p.1 if for all $\eta, \delta > 0$ there exist a finite number $T = T(\eta,\delta)$ such that $\Pr\left(\sup_{t \ge T} |x_t| \ge \delta\right) < \eta$. In this section, we address the fact that the bound T might need to be random. Note that, in the standard treatment, T is not allowed to be random. However, we show that T can be random and almost sure convergence still holds if T is almost surely bounded.

Lemma 17 *Let x_t be a random sequence. Assume that for each $\eta, \delta > 0$ there exist an almost surely finite random index $M = M(\eta,\delta)$ such that*

$$\Pr\left(\sup_{M \le t} |x_t| \ge \delta\right) < \eta. \tag{37}$$

Then, x_t converges to zero w.p.1.

Proof. Inequality (37) differs from the condition in the standard definition because M is allowed to be random.

Notice that, if $M(\omega) \le k$, then $\sup_{t \ge k} |x_t(\omega)| \le \sup_{t \ge M(\omega)} |x_t(\omega)|$ and, thus,

$$\left\{\omega \mid \sup_{t \ge k} |x_t(\omega)| \ge \delta,\ M(\omega) \le k\right\} \subseteq \left\{\omega \mid \sup_{t \ge M(\omega)} |x_t(\omega)| \ge \delta,\ M(\omega) \le k\right\}.$$

Now,

$$A \;=\; \left\{\omega \mid \sup_{t \ge k} |x_t(\omega)| \ge \delta\right\}$$

$$
\begin{aligned}
&= \Big(A \cap \{\omega \,|\, M(\omega) \le k\}\Big) \cup \Big(A \cap \{\omega \,|\, M(\omega) > k\}\Big) \\
&\subseteq \left\{\omega \,|\, \sup_{t \ge M(\omega)} |x_t(\omega)| \ge \delta,\, M(\omega) \le k\right\} \cup \{\omega \,|\, M(\omega) > k\}\,.
\end{aligned}
$$

Thus,

$$
\Pr\left(\sup_{t \ge k} |x_t| \ge \delta\right) \le \Pr\left(\sup_{t \ge M} |x_t| \ge \delta\right) + \Pr(M > k).
$$

Now, pick up an arbitrary $\delta, \eta > 0$. We want to prove that, for large enough $k > 0$, $\Pr(\sup_{t \ge k} |x_t| \ge \delta) < \eta$. Let $M_0 = M(\delta, \eta/2)$ be the random index whose existence is guaranteed by assumption and let $k = k(\varepsilon, \eta)$ be a natural number large enough such that $\Pr(M_0 > k) < \eta/2$. Such a number exists since $M_0 < \infty$ w.p.1. Then, $\Pr(\sup_{t \ge k} |x_t| \ge \delta) \le \Pr(\sup_{t \ge M} |x_t| \ge \delta) + \Pr(M_0 > k) < \eta$, showing that k is a suitable (non-random) index. □

C Convergence of Certain Stochastic Approximation Processes

In this section, we prove two useful stochastic approximation theorems, which are used in the applications involving averaging-type processes. We will make use of the following "super-martingale"-type lemma due to Robbins & Siegmund (1971).

LEMMA 18 *Suppose that Z_t, B_t, C_t, D_t are finite, non-negative random variables, adapted to the σ-field $\mathcal{F}_t$, which satisfy*

$$
E[Z_{t+1} \,|\, \mathcal{F}_t] \le (1 + B_t) Z_t + C_t - D_t. \tag{38}
$$

Then, on the set $\{\sum_{t=0}^{\infty} B_t < \infty, \sum_{t=0}^{\infty} C_t < \infty\}$, we have $\sum_{t=0}^{\infty} D_t < \infty$ and $Z_t \to Z < \infty$ almost surely.

The following could be regarded as a typical Robbins-Monro stochastic approximation theorem; however, it is also motivated by Dvoretzky's theorem, resulting in a mixture of the two. The main purpose here is to provide a short proof of the conditional averaging lemma (Lemma 4, presented in Section 2.2), which itself is a very useful result in this particular form.[12]

THEOREM 19 *Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \ldots$ be an increasing sequence of σ-fields and consider the process*

$$
x_{t+1} = x_t + H_t(x_t), \quad t = 0, 1, 2, \ldots \tag{39}
$$

where $H_t(\cdot)$ is real-valued and almost surely bounded function. Assume that x_t is $\mathcal{F}_t$-measurable and let $h_t(x_t) = E[H_t(x_t) \,|\, \mathcal{F}_t]$. Assume that the following assumptions are satisfied:

[12] Interestingly, in a probabilistic setup, the convergence of the outstar-learning algorithm of Grossberg (1969) used, for example, in counter-propagation networks (Hecht-Nielsen 1991), could be analyzed directly with this type of lemma.

1. *A number x^* exists such that*

   (a) $(x - x^*)h_t(x) \le 0$ *for all* $t \ge 0$.
   and if for any fixed $\varepsilon > 0$ *we let*

   $$\overline{h}_t(\varepsilon) = \sup_{\varepsilon \le |x - x^*| \le 1/\varepsilon} \frac{h_t(x)}{x - x^*}$$

   then w.p.1

   (b) $\sum_{t=0}^{\infty} \overline{h}_t(\varepsilon) = -\infty$;

   (c) $\sum_{t=0}^{\infty} \overline{h}_t^{+}(\varepsilon) < \infty$, *where* $r^+ = (r + |r|)/2$ *as usual; and*

2. $E[H_t^2(x_t)\,|\mathcal{F}_t] \le C_t(1+(x_t-x^*)^2)$, *for some non-negative random sequence* C_t *which satisfies* $\sum_{t=1}^{\infty} C_t < \infty$ *w.p.1.*

Then, x_t *converges to* x^* *w.p.1.*

Proof.
Begin with Lemma 18. In our case, let $Z_t = (x_t - x^*)^2$. Then,

$$\begin{aligned} E[Z_{t+1}\,|\mathcal{F}_t] &\le Z_t + C_t(1 + Z_t) + 2(x_t - x^*)h_t(x_t) \\ &\le (1 + C_t)Z_t + C_t + 2(x_t - x^*)h_t(x_t) \end{aligned}$$

and, therefore, by Lemma 18 (since by assumption $C_t \ge 0$, $\sum_{t=0}^{\infty} C_t < \infty$ and $(x_t - x^*)h_t(x_t) \le 0$), $Z_t \to Z < \infty$ w.p.1 for some random variable Z and $\sum_{t=0}^{\infty}(x_t - x^*)h_t(x_t) > -\infty$. If $\infty > Z(\omega) \ne 0$ for some ω, then there exist an $\varepsilon > 0$ and $N > 0$ (which may depend on ω) such that if $t \ge N$ then $\varepsilon \le |x_t(\omega) - x^*| \le \frac{1}{\varepsilon}$. Consequently,

$$\begin{aligned} -\infty &< \sum_{s=0}^{\infty}(x_s(\omega) - x^*)h_s(x_s(\omega)) \\ &\le \sum_{s=0}^{\infty}(x_s(\omega) - x^*)^2\, \overline{h}_s(\varepsilon;\omega) \\ &\le \sum_{s=0}^{N-1}(x_s(\omega) - x^*)^2\, \overline{h}_s(\varepsilon;\omega) + \varepsilon^2 \sum_{s \ge N, \overline{h}_s(\varepsilon;\omega) \le 0} \overline{h}_s(\varepsilon;\omega) + \\ &\qquad \frac{1}{\varepsilon^2} \sum_{s \ge N, \overline{h}_s(\varepsilon;\omega) > 0} \overline{h}_s(\varepsilon;\omega) \\ &= -\infty \end{aligned}$$

by Condition 1b. This means that $\{\omega \,|\, Z(\omega) \ne 0\}$ must be a null-set, finishing the proof of the theorem. $\square$

The theorem could easily be extended to vector-valued processes. Then, the definition of $\overline{h}_t(\varepsilon)$ would become $\overline{h}_t(\varepsilon) = \sup_{\varepsilon \le \|x - x^*\|_2 \le 1/\varepsilon}(x - x^*)^T h_t(x)$, and Condition 1a becomes $(x - x^*)^T h(x) \leq 0$, but not another word of the proof needs to be changed if we define $Z_t = \|x_t - x^*\|_2^2$. Note that Theorem 19 includes as a special case (*i*) the standard Robbins-Monro process of the form $x_{t+1} = x_t + \gamma_t H(x_t, \eta_t)$, where η_t are random variables whose distributions depend only on x_t, $\gamma_t \geq 0$, $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$, and (*ii*) one form of the Dvoretzky process $x_{t+1} = T_t + \eta_t$, where $T_t = G_t(x_t - x^*) + x^*$, $E[\eta_t | G_t, \eta_{t-1}, G_{t-1}, \ldots, \eta_0, G_0] = 0$, $\sum_t E[\eta_t^2] < \infty$, $G_t \leq 1$, and $\sum_t (G_t - 1) = -\infty$.

For our purposes, however, the following simple lemma (part of this lemma appeared in Lemma 4) is sufficient.

LEMMA 20 (CONDITIONAL AVERAGING LEMMA) *Let $\mathcal{F}_t$ be an increasing sequence of σ-fields, let $0 \leq \alpha_t$, s_t and w_t be random variables such that α_t, w_{t-1} and s_{t-1} are $\mathcal{F}_t$ measurable. Assume that the following hold w.p.1: $E[s_t | \mathcal{F}_t, \alpha_t \neq 0] = \hat{A} > 0$, $E[s_t^2 | \mathcal{F}_t] < \hat{B} < \infty$, $E[s_t w_t | \mathcal{F}_t, \alpha_t \neq 0] = A$, $E[s_t^2 w_t^2 | \mathcal{F}_t] < B < \infty$, $\sum_{t=1}^{\infty} \alpha_t = \infty$, and $\sum_{t=1}^{\infty} \alpha_t^2 < C < \infty$ for some $B, C > 0$. Then, the process*

$$Q_{t+1} = (1 - s_t \alpha_t) Q_t + \alpha_t s_t w_t \tag{40}$$

converges to $A/\hat{A}$ w.p.1.

Proof. Without loss of generality, we may assume that $E[s_t | \mathcal{F}_t] = \hat{A}$ and $E[s_t w_t | \mathcal{F}_t] = A$. Rewriting the process of Equation (40) in the form of Equation (39) we get $Q_{t+1} = Q_t + \alpha_t s_t (w_t - Q_t)$ and, thus, $h_t(Q) = E[\alpha_t s_t (w_t - Q) | \mathcal{F}_t] = \alpha_t (E[s_t w_t | \mathcal{F}_t] - Q E[s_t | \mathcal{F}_t]) == \alpha_t \hat{A}(A/\hat{A} - Q)$ and $\overline{h}_t(\varepsilon) = -\alpha_t \hat{A}$ independently of ε. Thanks to the identity $|x| \leq 1 + x^2$, $|E[s_t^2 w_t | \mathcal{F}_t]| \leq E[s_t^2 |w_t| \, | \mathcal{F}_t] \leq E[s_t^2 (1 + w_t^2) \, | \mathcal{F}_t] \leq \hat{B} + B$ and making use of $|x| \leq 1 + x^2$ again, we have $E[H_t^2(Q_t) | \mathcal{F}_t] = \alpha_t^2 E[s_t^2 (w_t - Q_t)^2 | \mathcal{F}_t] \leq \alpha_t^2 (B + 2(\hat{B} + B)(1 + Q_t^2) + \hat{B} Q_t^2) \leq \alpha_t^2 C'(1 + (Q_t - A/\hat{A})^2)$ for some $C' > 0$. Thus, the lemma follows from Theorem 19. □

References

Barto, A. G.; Bradtke, S. J.; and Singh, S. P. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* 72(1):81–138.

Barto, A. G.; Sutton, R. S.; and Watkins, C. J. C. H. 1989. Learning and sequential decision making. Technical Report 89-95, Department of Computer and Information Science, University of Massachusetts, Amherst, MA. Also published in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, Michael Gabriel and John Moore, editors. The MIT Press, Cambridge, MA, 1991.

Benveniste, A.; Métivier, M.; and Priouret, P. 1990. *Adaptive Algorithms and Stochastic Approximations.* Springer Verlag, New York.

Bertsekas, D. P., and Castañon, D. A. 1989. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE Transactions on Automatic Control* 34(6):589–598.

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. 1996. *Neuro-Dynamic Programming.* Belmont, MA: Athena Scientific.

Gordon, G. J. 1995. Stable function approximation in dynamic programming. In Prieditis, A., and Russell, S., eds., *Proceedings of the Twelfth International Conference on Machine Learning*, 261–268. San Francisco, CA: Morgan Kaufmann.

Grossberg, S. 1969. Embedding fields: A theory of learning with physiological implications. *Journal of Mathematical Psychology* 6:209–239.

Gullapalli, V., and Barto, A. G. 1994. Convergence of indirect adaptive asynchronous value iteration algorithms. In Cowan, J. D.; Tesauro, G.; and Alspector, J., eds., *Advances in Neural Information Processing Systems 6*, 695–702. San Mateo, CA: Morgan Kaufmann.

Hecht-Nielsen, R. 1991. *Neurocomputing.* Addison-Wesley.

Heger, M. 1994. Consideration of risk in reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 105–111. San Francisco, CA: Morgan Kaufmann.

Hu, J., and Wellman, M. P. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In Shavlik, J., ed., *Proceedings of the Fifteenth International Conference on Machine Learning.* Morgan Kaufmann.

Jaakkola, T.; Jordan, M. I.; and Singh, S. P. 1994. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation* 6(6):1185–1201.

John, G. H. 1994. When the best move isn't optimal: Q-learning with exploration. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1464.

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

Konda, V., and Borkar, V. 1997. Learning algorithms for Markov decision processes. Submitted.

Korf, R. E. 1990. Real-time heuristic search. *Artificial Intelligence* 42:189–211.

Kushner, H., and Clark, D. 1978. *Stochastic approximation methods for constrained and unconstrained systems.* Springer-Verlag, Berlin, Heidelberg, New York.

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

Littman, M. L., and Szepesvári, C. 1996. A generalized reinforcement-learning model: Convergence and applications. In Saitta, L., ed., *Proceedings of the Thirteenth International Conference on Machine Learning*, 310–318.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 157–163. San Francisco, CA: Morgan Kaufmann.

Littman, M. L. 1996. *Algorithms for Sequential Decision Making*. Ph.D. Dissertation, Department of Computer Science, Brown University. Also Technical Report CS-96-09.

Ljung, L. 1977. Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* 22:551–575.

Mahadevan, S. 1996. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning* 22(1/2/3):159–196.

Moore, A. W., and Atkeson, C. G. 1993. Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning* 13:103–130.

Owen, G. 1982. *Game Theory: Second edition*. Orlando, Florida: Academic Press.

Puterman, M. L. 1994. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons, Inc.

Ribeiro, C., and Szepesvári, C. 1996. Q-learning combined with spreading: Convergence and results. In *Proceedings of ISRF-IEE International Conference: Intelligent and Cognitive Systems, Neural Networks Symposium*, 32–36.

Ribeiro, C. 1995. Attentional mechanisms as a strategy for generalisation in the Q-learning algorithm. In *Proceedings of ICANN'95*, volume 1, 455–460.

Robbins, H., and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22:400–407.

Robbins, H., and Siegmund, D. 1971. A convergence theorem for non-negative almost supermartingales and some applications. In Rustagi, J., ed., *Optimizing Methods in Statistics*. New York: Academic Press. 235–257.

Rummery, G. A., and Niranjan, M. 1994. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department.

Schwartz, A. 1993. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning*, 298–305. Amherst, MA: Morgan Kaufmann.

Schweitzer, P. J. 1984. Aggregation methods for large Markov chains. In Iazola, G.; Coutois, P. J.; and Hordijk, A., eds., *Mathematical Computer Performance and Reliability*. Amsterdam, Holland: Elsevier. 275–302.

Singh, S. P., and Sutton, R. S. 1996. Reinforcement learning with replacing eligibility traces. *Machine Learning* 22(1/2/3):123–158.

Singh, S.; Jaakkola, T.; Littman, M. L.; and Szepesvári, C. 1998. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*. To appear.

Singh, S.; Jaakkola, T.; and Jordan, M. 1995. Reinforcement learning with soft state aggregation. In Tesauro, G.; Touretzky, D. S.; and Leen, T. K., eds., *Advances in Neural Information Processing Systems 7*, 361–368. Cambridge, MA: The MIT Press.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. The MIT Press.

Szepesvári, C., and Littman, M. L. 1996. Generalized Markov decision processes: Dynamic-programming and reinforcement-learning algorithms. Technical Report CS-96-11, Brown University, Providence, RI.

Szepesvári, m. 1997. On the asymptotic convergence rate of Q-learning. In *Proc. of Neural Information Processing Systems.* accepted.

Szepesvári, C. 1998. *Static and Dynamic Aspects of Optimal Sequential Decision Making*. Ph.D. Dissertation, Bolyai Institute of Mathematics, "József Attila" University, Szeged 6720, Aradi vrt. tere 1, HUNGARY.

Tsitsiklis, J. N. 1994. Asynchronous stochastic approximation and Q-learning. *Machine Learning* 16(3):185–202.

Vrieze, O. J., and Tijs, S. H. 1982. Fictitious play applied to sequences of games and discounted stochastic games. *International Journal of Game Theory* 11(2):71–85.

Watkins, C. J. C. H., and Dayan, P. 1992. Q-learning. *Machine Learning* 8(3):279–292.

Watkins, C. J. C. H. 1989. *Learning from Delayed Rewards*. Ph.D. Dissertation, King's College, Cambridge, UK.

Williams, R. J., and Baird, III, L. C. 1993. Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems. Technical Report NU-CCS-93-11, Northeastern University, College of Computer Science, Boston, MA.