Journal of Machine Learning Research 5 (2003) 1-25 Submitted 4/02; Revised 9/02; Published 12/03

Learning Rates for Q-learning

Eyal Even-Dar EVEND@CS.TAU.AC.IL
Yishay Mansour MANSOUR@CS.TAU.AC.IL
School of Computer Science
Tel-Aviv University
Tel-Aviv, 69978, Israel

Editor: Peter Bartlett

Abstract

In this paper we derive convergence rates for Q-learning. We show an interesting relationship between the convergence rate and the learning rate used in Q-learning. For a polynomial learning rate, one which is $1/t^\omega$ at time t where $\omega \in (1/2,1)$, we show that the convergence rate is polynomial in $1/(1-\gamma)$, where γ is the discount factor. In contrast we show that for a linear learning rate, one which is $1/t$ at time t, the convergence rate has an exponential dependence on $1/(1-\gamma)$. In addition we show a simple example that proves this exponential behavior is inherent for linear learning rates.

Keywords: Reinforcement Learning, Q-Learning, Stochastic Processes, Convergence Bounds, Learning Rates.

1. Introduction

In Reinforcement Learning, an agent wanders in an unknown environment and tries to maximize its long term return by performing actions and receiving rewards. The challenge is to understand how a current action will affect future rewards. A good way to model this task is with Markov Decision Processes (MDP), which have become the dominant approach in Reinforcement Learning (Sutton and Barto, 1998, Bertsekas and Tsitsiklis, 1996).

An MDP includes states (which abstract the environment), actions (which are the available actions to the agent), and for each state-action pair, a distribution of next states (the states reached after performing the action in the given state). In addition there is a reward function that assigns a stochastic reward for each state and action. The *return* combines a sequence of rewards into a single value that the agent tries to optimize. A *discounted return* has a parameter $\gamma \in (0,1)$ where the reward received at step k is discounted by γ^k.

One of the challenges of Reinforcement Learning is when the MDP is not known, and we can only observe the trajectory of states, actions and rewards generated by the agent wandering in the MDP. There are two basic conceptual approaches to the learning problem. The first is model based, where we first reconstruct a model of the MDP, and then find an optimal policy for the approximate model. The second approach is implicit methods that update the information after each step, and based on this derive an estimate to the optimal policy. The most popular of those methods is Q-learning (Watkins, 1989).

Q-learning is an off-policy method that can be run on top of any strategy wandering in the MDP. It uses the information observed to approximate the optimal function, from which one can construct the optimal policy. There are various proofs that Q-learning does converge to the optimal Q function, under very mild conditions (Bertsekas and Tsitsiklis, 1996, Tsitsiklis, 1994, Watkins and Dyan, 1992, Littman and Szepesvári, 1996, Jaakkola et al., 1994, Borkar and Meyn, 2000). The conditions have to do with the exploration policy and the learning rate. For the exploration one needs to require that each state action be performed infinitely often. The learning rate controls how fast we modify our estimates. One expects to start with a high learning rate, which allows fast changes, and lowers the learning rate as time progresses. The basic conditions are that the sum of the learning rates goes to infinity (so that any value could be reached) and that the sum of the squares of the learning rates is finite (which is required to show that the convergence is with probability one).

We build on the proof technique of Bertsekas and Tsitsiklis (1996), which is based on convergence of stochastic iterative algorithms, to derive convergence rates for Q-learning. We study two models of updating in Q-learning. The first is the synchronous model, where all state action pairs are updated simultaneously. The second is the asynchronous model, where at each step we update a single state action pair. We distinguish between two sets of learning rates. The most interesting outcome of our investigation is the relationship between the form of the learning rates and the rate of convergence. A linear learning rate is of the form $1/t$ at time t, and a polynomial learning rate, which is of the form $1/t^{\omega}$, where $\omega \in (1/2, 1)$ is a parameter.

We show for synchronous models that for a polynomial learning rate the convergence rate is polynomial in $1/(1-\gamma)$, while for a linear learning rate the convergence rate is exponential in $1/(1-\gamma)$. We also describe an MDP that has exponential behavior for a linear learning rate. The lower bound simply shows that if the initial value is one and all the rewards are zero, it takes $O((1/\varepsilon)^{1/(1-\gamma)})$ updates, using a linear learning rate, until we reach a value of ε.

The different behavior might be explained by the asymptotic behavior of $\sum_t \alpha_t$, one of the conditions that ensure that Q-learning converges from any initial value. In the case of a linear learning rate we have that $\sum_{t=1}^{T} \alpha_t = O(\ln(T))$, whereas using polynomial learning rate it behaves as $O(T^{1-\omega})$. Therefore, using polynomial learning rate each value can be reached by polynomial number of steps and using linear learning rate each value requires exponential number of steps.

The convergence rate of Q-learning in a batch setting, where many samples are averaged for each update, was analyzed by Kearns and Singh (1999). A batch setting does not have a learning rate and has much of the flavor of model based techniques, since each update is an average of many samples. A run of a batch Q-learning is divided into phases, at the end of each phase an update is made. The update after each phase is reliable since it averages many samples.

The convergence of Q-learning with linear learning rate was studied by Szepesvari (1998) for special MDPs, where the next state distribution is the same for each state. (This setting is much closer to the PAC model, since there is no influence between the action performed and the states reached, and the states are i.i.d distributed). For this model Szepesvari (1998) shows a convergence rate, which is exponential in $1/(1-\gamma)$. Beleznay et al. (1999) give an exponential lower bound in the number of the states for undiscounted return.

2. The Model

We define a Markov Decision process (MDP) as follows

Definition 1 *A Markov Decision process (MDP) M is a 4-tuple (S,U,P,R), where S is a set of the states, U is a set of actions ($U(i)$ is the set of actions available at state i), $P_{i,j}^{M}(a)$ is the transition* *probability from state i to state j when performing action $a \in U(i)$ in state i, and $R_M(s,a)$ is the reward received when performing action a in state s.*

We assume that $R_M(s,a)$ is non-negative and bounded by R_{max}, i.e., $\forall s,a: \; 0 \leq R_M(s,a) \leq R_{max}$. For simplicity we assume that the reward $R_M(s,a)$ is deterministic, however all our results apply when $R_M(s,a)$ is stochastic.

A strategy for an MDP assigns, at each time t, for each state s a probability for performing action $a \in U(s)$, given a history $F_{t-1} = \{s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}, r_{t-1}\}$ which includes the states, actions and rewards observed until time $t-1$. A policy is memory-less strategy, i.e., it depends only on the current state and not on the history. A deterministic policy assigns each state a unique action.

While following a policy π we perform at time t action a_t at state s_t and observe a reward r_t (distributed according to $R_M(s,a)$), and the next state s_{t+1} (distributed according to $P^M_{s_t,s_{t+1}}(a_t)$). We combine the sequence of rewards to a single value called the return, and our goal is to maximize the return. In this work we focus on *discounted return*, which has a parameter $\gamma \in (0,1)$, and the discounted return of policy π is $V^\pi_M = \sum_{t=0}^\infty \gamma^t r_t$, where r_t is the reward observed at time t. Since all the rewards are bounded by R_{max} the discounted return is bounded by $V_{max} = \frac{R_{max}}{1-\gamma}$.

We define a value function for each state s, under policy π, as $V^\pi_M(s) = E[\sum_{i=0}^\infty r_i \gamma^i]$, where the expectation is over a run of policy π starting at state s. We define a state-action value function $Q^\pi_M(s,a) = R_M(s,a) + \gamma \sum_{\bar{s}} P^M_{s,\bar{s}}(a) V^\pi_M(\bar{s})$, whose value is the return of initially performing action a at state s and then following policy π. Since $\gamma < 1$ we can define another parameter $\beta = (1-\gamma)/2$, which will be useful for stating our results. (Note that as β decreases V_{max} increases.)

Let π^* be an optimal policy, which maximizes the return from any start state. (It is well known that there exists an optimal strategy, which is a deterministic policy (Puterman., 1994).) This implies that for any policy π and any state s we have $V^{\pi^*}_M(s) \geq V^\pi_M(s)$, and $\pi^*(s) = argmax_a(R_M(s,a) + \gamma(\sum_{s'} P^M_{s,s'}(a) \max_b Q(s',b))$. The optimal policy is also the only fixed point of the operator, $(TQ)(s,a) = R_M(s,a) + \gamma \sum_{s'} P_{s,s'}(a) \max_b Q(s',b)$. We use V^*_M and Q^*_M for $V^{\pi^*}_M$ and $Q^{\pi^*}_M$, respectively. We say that a policy π is an ε-approximation of the optimal policy if $\|V^*_M - V^\pi_M\|_\infty \leq \varepsilon$.

For a sequence of state-action pairs let the *covering time*, denoted by L, be an upper bound on the number of state-action pairs starting from any pair, until all state-action appear in the sequence. Note that the covering time can be a function of both the MDP and the sequence or just of the sequence. Initially we assume that from any start state, within L steps all state-action pairs appear in the sequence. Later, we relax the assumption and assume that with probability at least $\frac{1}{2}$, from any start state in L steps all state-action appear in the sequence. In this paper, the underlying policy generates the sequence of state action pairs.

The Parallel Sampling Model, $PS(M)$, as was introduced by Kearns and Singh (1999). The $PS(M)$ is an ideal exploration policy. A single call to $PS(M)$ returns for every pair (s,a) the next state s', distributed according to $P^M_{s,s'}(a)$ and a reward r distributed according to $R_M(s,a)$. The advantage of this model is that it allows to ignore the exploration and to focus on the learning. In some sense $PS(M)$ can be viewed as a perfect exploration policy.

Notations: The notation $g = \tilde{\Omega}(f)$ implies that there are constants c_1 and c_2 such that $g \geq c_1 f \ln^{c_2}(f)$. All the norms $\|\cdot\|$, unless otherwise specified, are L_∞ norms, i.e., $\|(x_1,\ldots,x_n)\| = \max_i x_i$.

3. Q-learning

The Q-learning algorithm (Watkins, 1989) estimates the state-action value function (for discounted return) as follows:

$$Q_{t+1}(s,a) = (1-\alpha_t(s,a))Q_t(s,a) + \alpha_t(s,a)(R_M(s,a) + \gamma \max_{b\in U(s')} Q_t(s',b)), \tag{1}$$

where s' is the state reached from state s when performing action a at time t. Let $T^{s,a}$ be the set of times, where action a was performed at state s, then $\alpha_t(s,a) = 0$ for $t \notin T^{s,a}$. It is known that Q-learning converges to Q^* if each state action pair is performed infinitely often and $\alpha_t(s,a)$ satisfies for each (s,a) pair: $\sum_{t=1}^{\infty} \alpha_t(s,a) = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2(s,a) < \infty$ (Bertsekas and Tsitsiklis, 1996, Tsitsiklis, 1994, Watkins and Dyan, 1992, Littman and Szepesvári, 1996, Jaakkola et al., 1994).

Q-learning is an asynchronous process in the sense that it updates a single entry each step. Next we describe two variants of Q-learning, which are used in the proofs. The first algorithm is *synchronous Q-learning*, which performs the updates by using the $PS(M)$. Specifically:

$$\begin{aligned} \forall s,a &: \quad Q_0(s,a) = C \\ \forall s,a &: \quad Q_{t+1}(s,a) = (1-\alpha_t^{\omega})Q_t(s,a) + \alpha_t^{\omega}(R_M(s,a) + \gamma \max_{b\in U(\bar{s})} Q_t(\bar{s},b)), \end{aligned}$$

where $\bar{s}$ is the state reached from state s when performing action a and C is some constant. The learning rate is $\alpha_t^{\omega} = \frac{1}{(t+1)^{\omega}}$, for $\omega \in (1/2, 1]$. We distinguish between a *linear learning rate*, which is $\omega = 1$, and a *polynomial learning rate*, which is $\omega \in (\frac{1}{2}, 1)$.

The *asynchronous Q-learning algorithm*, is simply regular Q-learning as define in (1), and we add the assumption that the underlying strategy has a covering time of L. The updates are as follows:

$$\begin{aligned} \forall s,a &: \quad Q_0(s,a) = C \\ \forall s,a &: \quad Q_{t+1}(s,a) = (1-\alpha_t^{\omega}(s,a))Q_t(s,a) + \alpha_t^{\omega}(s,a)(R_M(s,a) + \gamma \max_{b\in U(\bar{s})} Q_t(\bar{s},b)) \end{aligned}$$

where $\bar{s}$ is the state reached from state s when performing action a and C is some constant. Let $\#(s,a,t)$ be one plus the number of times, until time t, that we visited state s and performed action a. The learning rate $\alpha_t^{\omega}(s,a) = \frac{1}{[\#(s,a,t)]^{\omega}}$, if $t \in T^{s,a}$ and $\alpha_t^{\omega}(s,a) = 0$ otherwise. Again, $\omega = 1$ is a linear learning rate, and $\omega \in (\frac{1}{2}, 1)$ is a polynomial learning rate.

4. Our Main Results

Our main results are upper bounds on the convergence rates of Q-learning algorithms and showing their dependence on the learning rate. The basic case is the synchronous Q-learning. We show that for a polynomial learning rate we have a complexity, which is polynomial in $1/(1-\gamma) = 1/(2\beta)$. In contrast, we show that linear learning rate has an exponential dependence on $1/\beta$. Our results exhibit a sharp difference between the two learning rates, although they both converge with probability one. This distinction, which is highly important, can be observed only when we study the convergence rate, rather than convergence in the limit.

The bounds for asynchronous Q-learning are similar. The main difference is the introduction of a covering time L. For polynomial learning rate we derive a bound polynomial in $1/\beta$, and for linear learning rate our bound is exponential in $\frac{1}{\beta}$. We also show a lower bound for linear learning rate, which is exponential in $\frac{1}{\beta}$. This implies that our upper bounds are tight, and that the gap between the two bounds is real.

We first prove the results for the synchronous Q-learning algorithm, where we update all the entries of the Q function at each time step, i.e., the updates are synchronous. The following theorem derives the bound for polynomial learning rate.

Theorem 2 *Let Q_T be the value of the synchronous Q-learning algorithm using polynomial learning rate at time T. Then with probability at least $1-\delta$, we have that $||Q_T - Q^*|| \leq \varepsilon$, given that*

$$T = \Omega\left(\left(\frac{V_{max}^2 \ln\left(\frac{|S|\,|A|V_{max}}{\delta\beta\varepsilon}\right)}{\beta^2\varepsilon^2}\right)^{\frac{1}{\omega}} + \left(\frac{1}{\beta}\ln\frac{V_{max}}{\varepsilon}\right)^{\frac{1}{1-\omega}}\right)$$

The above bound is somewhat complicated. To simplify, assume that ω is a constant and consider first only its dependence on ε. This gives us $\Omega((\ln(1/\varepsilon)/\varepsilon^2)^{1/\omega} + (\ln(1/\varepsilon))^{1/(1-\omega)})$, which is optimized when ω approaches one. Considering the dependence only on β, recall that $V_{max} = R_{max}/(2\beta)$, therefore the complexity is $\tilde{\Omega}(1/\beta^{4/\omega} + 1/\beta^{1/(1-\omega)})$ which is optimized for $\omega = 4/5$. The following theorem bounds the time for linear learning rate.

Theorem 3 *Let Q_T be the value of the synchronous Q-learning algorithm using linear learning rate at time T. Then for any positive constant ψ with probability at least $1-\delta$, we have $||Q_T - Q^*|| \leq \varepsilon$, given that*

$$T = \Omega\left(((2+\psi)^{\frac{1}{\beta}\ln(\frac{V_{max}}{\varepsilon})}\frac{V_{max}^2 \ln\left(\frac{|S|\,|A|V_{max}}{\delta\beta\psi\varepsilon}\right)}{(\psi\beta\varepsilon)^2}\right).$$

Next we state our results to asynchronous Q-learning. The bounds are similar to those of synchronous Q-learning, but have the extra dependency on the covering time L.

Theorem 4 *Let Q_T be the value of the asynchronous Q-learning algorithm using polynomial learning rate at time T. Then with probability at least $1-\delta$, we have $||Q_T - Q^*|| \leq \varepsilon$, given that*

$$T = \Omega\left(\left(\frac{L^{1+3\omega}V_{max}^2 \ln\left(\frac{|S|\,|A|V_{max}}{\delta\beta\varepsilon}\right)}{\beta^2\varepsilon^2}\right)^{\frac{1}{\omega}} + \left(\frac{L}{\beta}\ln\frac{V_{max}}{\varepsilon}\right)^{\frac{1}{1-\omega}}\right)$$

The dependence on the covering time, in the above theorem, is $\Omega(L^{2+1/\omega} + L^{1/(1-\omega)})$, which is optimized for $\omega \approx 0.77$. For the linear learning rate the dependence is much worse, since it has to be that $L \geq |S| \cdot |A|$, as is stated in the following theorem.

Theorem 5 *Let Q_T be the value of the asynchronous Q-learning algorithm using linear learning rate at time T. Then with probability at least $1-\delta$, for any positive constant ψ we have $||Q_T - Q^*|| \leq \varepsilon$, given that*

$$T = \Omega\left((L + \psi L + 1)^{\frac{1}{\beta}\ln\frac{V_{max}}{\varepsilon}} \frac{V_{max}^2 \ln\left(\frac{|S|\,|A|V_{max}}{\delta\beta\varepsilon\psi}\right)}{(\psi\beta\varepsilon)^2}\right)$$

The following theorem shows that a linear learning rate may require an exponential dependence on $1/(2\beta) = 1/(1-\gamma)$, thus showing that the gap between linear learning rate and polynomial learning rate is real and does exist for some MDPs.

Theorem 6 *There exists a deterministic MDP, M, such that Q-learning with linear learning rate after* $T = \Omega((\frac{1}{\varepsilon})^{\frac{1}{1-\gamma}})$ *steps has* $\|Q_T - Q_M^*\| > \varepsilon$.

5. Experiments

In this section we present experiments using two types of MDPs as well as one we call M_0, which is used in the lower bound example from Section 10. The two MDP types are the "random MDP" and "line MDP". Each type contains n states and two actions for each state.

We generate the "random MDP" as follows: For every state i, action a and state j, we assign a random number $n_{i,j}(a)$ uniformly from $[0,1]$. The probability of a transition from state i to state j while performing action a is $p_{i,j}(a) = \frac{n_{i,j}(a)}{\sum_k n_{i,k}(a)}$. The reward $R(s,a)$ is deterministic and chosen at random uniformly in the interval $[0,10]$.

For the line MDP, all the states are integers and the transition probability from state i to state j is proportional to $\frac{1}{|i-j|}$, where $i \neq j$. The reward distribution is identical to that of the random MDP. (We implemented the random function using the function rand() in C.)

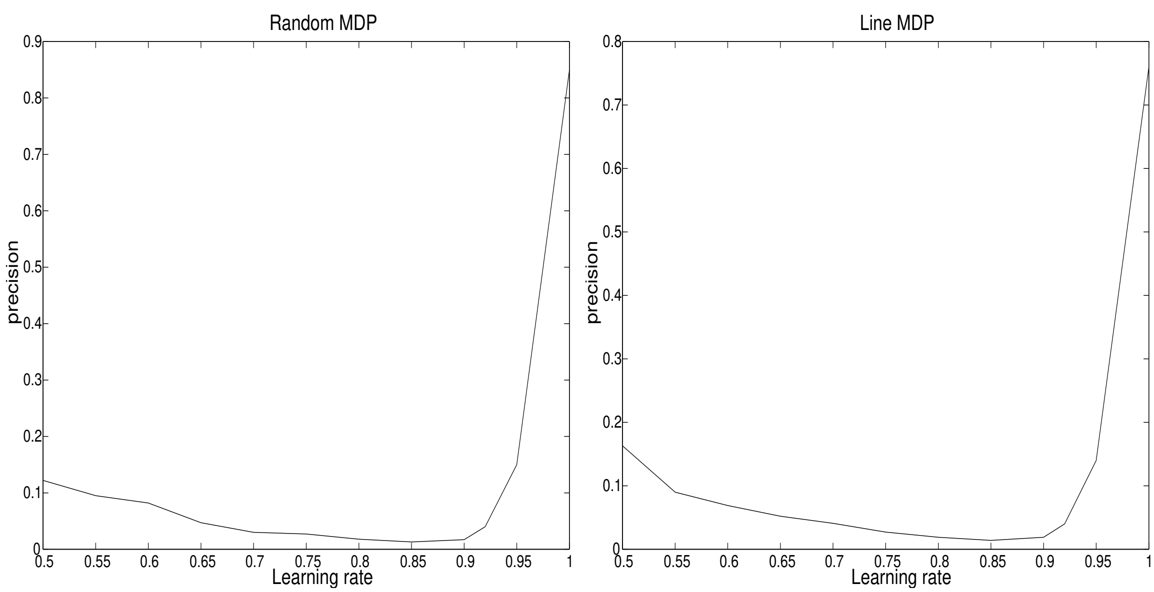

Figure 1: Example of 100 states MDP (both line and random), where the discount factor is $\gamma = 0.7$. We ran asynchronous Q-learning using a random exploration policy for 10^8 steps.

Figure 1 demonstrates the relation between the exponent of the learning rate ω and the accuracy of the model. The best experimental value for ω is about 0.85. Note that when ω approaches one (a linear learning rate), the precision deteriorates. This behavior coincides with our theoretical results on two points. First, our theoretical results predict bad behavior when the learning rate approaches one (an exponential lower and upper bound). Second, the experiments suggest an optimal value for ω of approximately 0.85. Our theoretical results derive optimal values of optimal ω for different settings of the parameters but most give a similar range. Furthermore, the two types of MDP have similar behavior, which implies that the difference between linear and polynomial learning rates is inherent to many MDPs and not only special cases (as in the lower bound example).

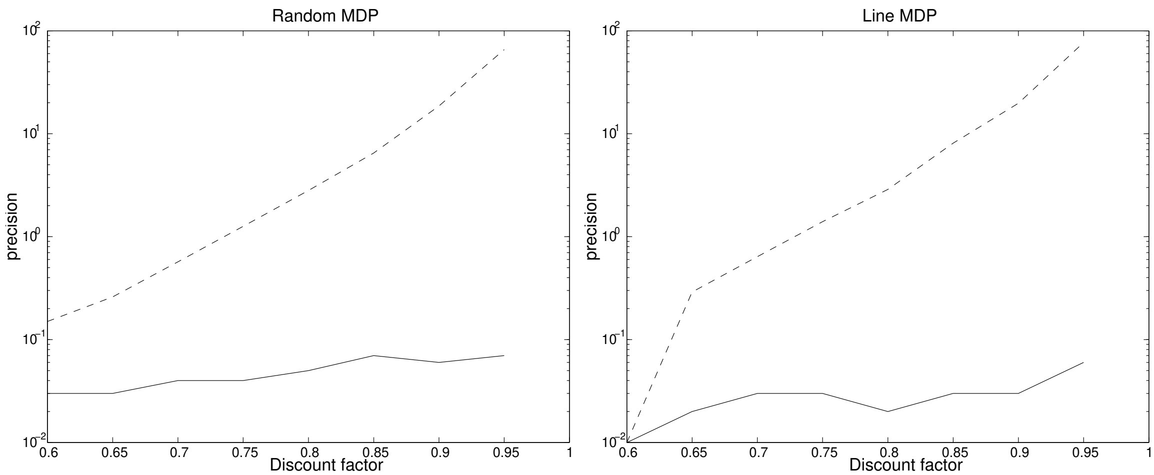

Figure 2: Example of 10 state MDPs (both random and line) using two different learning rates for Q-learning. Both use random exploration policy for 10^7 steps. The solid line is asynchronous Q-learning using $\omega = 0.7$; the dashed line is asynchronous Q-learning using a linear learning rate ($\omega = 1.0$).

Figure 2 demonstrates the strong relationship between discount factor, γ, and convergence rate. In this experiment, we again see similar behavior in both MDPs. When the discount factor approaches one, Q-learning using linear learning rate estimation of the Q value becomes unreliable, while Q-learning using learning rate of $\omega = 0.7$ remains stable (the error is below 0.1).

Figure 3 compares two different learning rates $\omega = 0.6$ and $\omega = 0.9$ for ten state MDPs (both random and line) and finds an interesting tradeoff. For low precision levels, the learning rate of $\omega = 0.6$ was superior, while for high precision levels the learning rate of $\omega = 0.9$ was superior. An explanation for this behavior is that the dependence in terms of ε is $\Omega((\ln(1/\varepsilon)/\varepsilon^2)^{1/\omega} + (\ln(1/\varepsilon))^{1/(1-\omega)})$, which is optimized as the learning rate approaches one.

Our last experimental result is M_0, the lower bound example from Section 10. Here the difference between the learning rates is the most significant, as shown in Figure 4.

6. Background from Stochastic Algorithms

Before we derive our proofs, we first introduce the proof given by Bertsekas and Tsitsiklis (1996) for the convergence of stochastic iterative algorithms; in Section 7 we show that Q-learning algorithms fall in this category. In this section we review the proof for convergence in the limit, and in the next sections we will analyze the rate at which different Q-learning algorithms converge. (We will try to keep the background as close as possible to the needs for this paper rather than giving the most general results.)

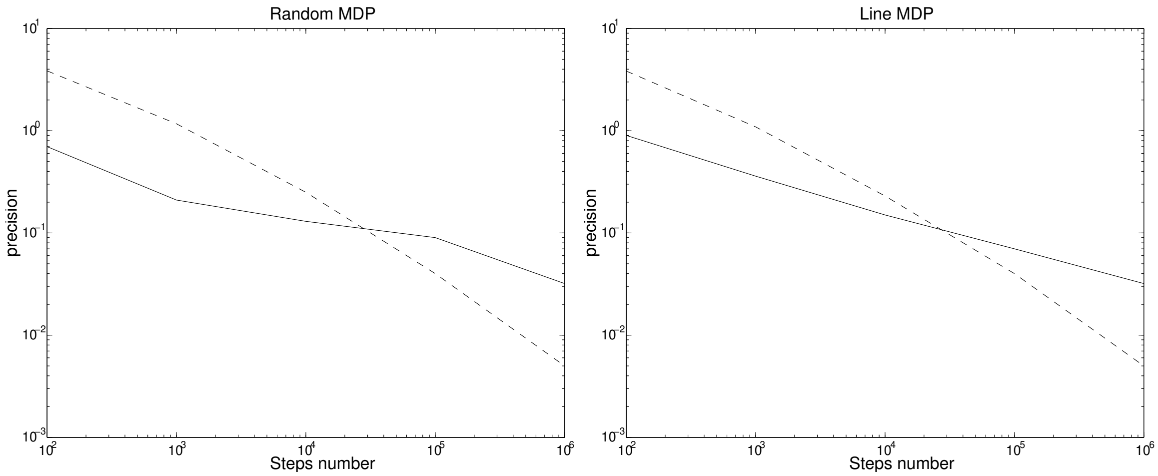

Figure 3: Random and Line MPDs (10 states each), where the discount factor is $\gamma = 0.9$. The dashed line is synchronous Q-learning using $\omega = 0.9$ and the the solid line is synchronous Q-learning using $\omega = 0.6$.

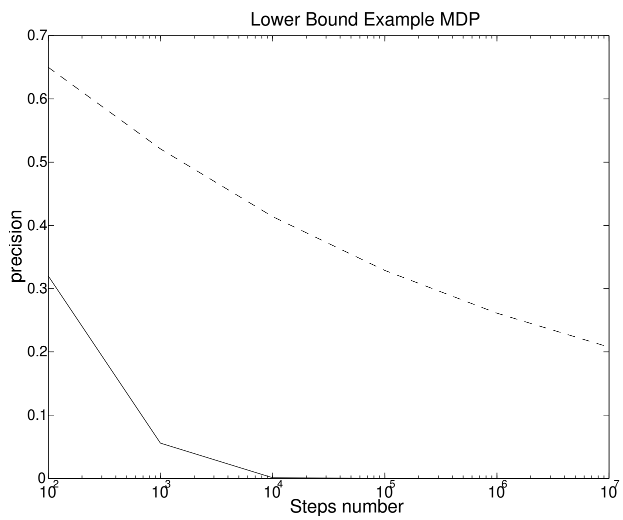

Figure 4: Lower bound example M_0, with discount factor $\gamma = 0.9$. Q-learning ran with two different learning rates, linear (dashed line) and $\omega = 0.65$ (solid line).

This section considers a general type of *iterative stochastic algorithms*, which is computed as follows:

$$X_{t+1}(i) = (1 - \alpha_t(i))X_t(i) + \alpha_t(i)((H_t X_t)(i) + w_t(i)), \qquad (2)$$

where w_t is a bounded random variable with zero expectation, and each H_t is assumed to belong to a family $\mathcal{H}$ of pseudo contraction mappings (See Bertsekas and Tsitsiklis (1996) for details).

Definition 7 *An iterative stochastic algorithm is well-behaved if:*

1. *The step size* $\alpha_t(i)$ *satisfies (1)* $\sum_{t=0}^{\infty} \alpha_t(i) = \infty$*, (2)* $\sum_{t=0}^{\infty} \alpha_t^2(i) < \infty$ *and (3)* $\alpha_t(i) \in (0,1)$.

2. *There exists a constant A that bounds* $w_t(i)$ *for any history* F_t*, i.e.,* $\forall t, i: \ |w_t(i)| \leq A$.

3. *There exists* $\gamma \in [0,1)$ *and a vector* X^* *such that for any X we have* $||H_t X - X^*|| \leq \gamma ||X - X^*||$.

The main theorem states that a well-behaved stochastic iterative algorithm converges in the limit.

Theorem 8 *[Bertsekas and Tsitsiklis (1996)] Let* X_t *be the sequence generated by a well-behaved stochastic iterative algorithm. Then* X_t *converges to* X^* *with probability* 1.

The following is an outline of the proof given by Bertsekas and Tsitsiklis (1996). Without loss of generality, assume that $X^* = 0$ and $\|X_0\| \leq A$. The value of X_t is bounded since $\|X_0\| \leq A$ and for any history F_t we have $\|w_t\| \leq A$; hence, for any t we have $\|X_t\| \leq A$.

Recall that $\beta = \frac{1-\gamma}{2}$. Let $D_1 = A$ and $D_{k+1} = (1-\beta)D_k$ for $k \geq 1$. Clearly the sequence D_k converges to zero. We prove by induction that for every k there exists some time τ_k such that for any $t \geq \tau_k$ we have $\|X_t\| \leq D_k$. Note that this will guarantee that at time $t \geq \tau_k$ for any i the value $\|X_t(i)\|$ is in the interval $[-D_k, D_k]$.

The proof is by induction. Assume that there is such a time τ_k and we show that there exists a time τ_{k+1} such that for $t \geq \tau_{k+1}$ we have $\|X_t\| \leq D_{k+1}$. Since D_k converges to zero this proves that X_t converges to zero, which equals X^*. For the proof we define for $t \geq \tau$ the quantity

$$W_{t+1;\tau}(i) = (1-\alpha_t(i))W_{t;\tau}(i) + \alpha_t(i)w_t(i),$$

where $W_{\tau;\tau}(i) = 0$. The value of $W_{t;\tau}$ bounds the contributions of $w_j(i)$, $j \in [\tau, t]$, to the value of X_t (starting from time τ). We also define for $t \geq \tau_k$,

$$Y_{t+1;\tau}(i) = (1-\alpha_t(i))Y_{t;\tau}(i) + \alpha_t(i)\gamma D_k$$

where $Y_{\tau_k;\tau_k} = D_k$. Notice that $Y_{t;\tau_k}$ is a deterministic process. The following lemma gives the motivation for the definition of $Y_{t;\tau_k}$.

Lemma 9 *[Bertsekas and Tsitsiklis (1996)] For every i, we have*

$$-Y_{t;\tau_k}(i) + W_{t;\tau_k}(i) \leq X_t(i) \leq Y_{t;\tau_k}(i) + W_{t;\tau_k}(i)$$

Next we use Lemma 9 to complete the proof of Theorem 8. From the definition of $Y_{t;\tau}$ and the assumption that $\sum_{t=0}^{\infty} \alpha_t = \infty$, it follows that $Y_{t;\tau}$ converges to γD_k as t goes to infinity. In addition $W_{t;\tau_k}$ converges to zero as t goes to infinity. Therefore there exists a time τ_{k+1} such that $Y_{t;\tau} \leq (\gamma + \frac{\beta}{2})D_k$, and $|W_{t;\tau_k}| \leq \beta D_k/2$. This fact, together with Lemma. 9, yields that for $t \geq \tau_{k+1}$,

$$||X_t|| \leq (\gamma + \beta)D_k = D_{k+1},$$

which completes the proof of Theorem 8.

7. Applying the Stochastic Theorem to Q-learning

In this section we show that both synchronous and asynchronous Q-learning are well-behaved iterative stochastic algorithms. The proof is similar in spirit to the proof given by Bertsekas and Tsitsiklis (1996) At the beginning we deal with synchronous Q-learning. First define operator H as

$$(HQ)(i,a) = \sum_{j=0}^{n} P_{ij}(a)(R(i,a) + \gamma \max_{b\in U(j)} Q(j,b))$$

Rewriting Q-learning with H, we get

$$Q_{t+1}(i,a) = (1-\alpha_t(i,a))Q_t(i,a) + \alpha_t(i,a)((HQ_t)(i,a) + w_t(i,a)).$$

Let $\bar{i}$ is the state reached by performing at time t action a in state i and $r(i,s)$ is the reward observed at time i; then

$$w_t(i,s) = r(i,s) + \gamma \max_{b\in U(\bar{i})} Q_t(\bar{i},b) - \sum_{j=0}^{n} P_{ij}(a)\left(R(i,a) + \gamma \max_{b\in U(j)} Q_t(j,b)\right)$$

In synchronous Q-learning, H is computed simultaneously on all states actions pairs.

Lemma 10 *Synchronous Q-learning is a well-behaved iterative stochastic algorithm.*

Proof We know that for any history F_t $E[w_t(i,a)|F_t] = 0$ and $|w_t(i,a)| \le V_{max}$. We also know that for $\frac{1}{2} < \omega \le 1$ we have that $\sum \alpha_t(s,a) = \infty$, $\sum \alpha_t^2(s,a) < \infty$ and $\alpha_t(s,a) \in (0,1)$.
We need only show that H satisfies the contraction property.

$$\begin{aligned}
|(HQ)(i,a) - (H\bar{Q})(i,a)| &\le \sum_{j=0}^{n} P_{ij}(a)|\gamma \max_{b\in U(j)} Q(j,b) - \gamma \max_{b\in U(j)} \bar{Q}(j,b)| \\
&= \sum_{j=0}^{n} P_{ij}(a)\gamma|\max_{b\in U(j)} Q(j,b) - \max_{b\in U(j)} \bar{Q}(j,b)| \\
&\le \sum_{j=0}^{n} P_{ij}(a)\gamma \max_{b\in U(j)} |Q(j,b) - \bar{Q}(j,b)| \\
&\le \gamma \sum_{j=0}^{n} P_{ij}(a)\|Q-\bar{Q}\| \le \gamma\|Q-\bar{Q}\|
\end{aligned}$$

Since we update all (i,a) pairs simultaneously, synchronous Q-learning is well-behaved stochastic iterative algorithm. ∎

We next show that Theorem 8 can be applied also to asynchronous Q-learning.

Lemma 11 *Asynchronous Q-learning, where the input sequence has a finite covering time L, is a well-behaved iterative stochastic algorithm.*

Proof We define HQ for every start state i and start time t_1 of a phase (beginning of the covering time) until the end of the phase (completing the covering time) at time t_2, during which all state-action pairs are updated. Since a state action can be performed more than once, $HQ(i,a)$ can be performed more than once. We consider time t, in which the policy performs action a at state i and Q_t is the vector. We have that

$$\begin{aligned}
|(HQ_t)(i,a)-(HQ^*)(i,a)| &\leq \sum_{j=0}^{n} P_{ij}(a)|\gamma \max_{b\in U(j)} Q_t(j,b) - \gamma \max_{b\in U(j)} Q^*(j,b)| \\
&= \sum_{j=0}^{n} P_{ij}(a)\gamma| \max_{b\in U(j)} Q_t(j,b) - \max_{b\in U(j)} Q^*(j,b)| \\
&\leq \sum_{j=0}^{n} P_{ij}(a)\gamma \max_{b\in U(j)} |Q_t(j,b) - Q^*(j,b)| \\
&\leq \sum_{j\in A} P_{ij}(a)\gamma \max_{b\in U(j)} |Q_t(j,b) - Q^*(j,b)| \\
&\quad + \sum_{j\in B} P_{ij}(a)\gamma \max_{b\in U(j)} |Q_t(j,b) - Q^*(j,b)| \\
&\leq \gamma\|Q_t - Q^*\|,
\end{aligned}$$

where A includes the states for which during (t_1,t) all the actions in $U(i)$ were performed, and $B=S-A$. We conclude that $\|Q_{t'}-Q^*\| \leq \|Q_{t'-1}-Q^*\|$, since we only change at each time a single state-action pair, which satisfies $|HQ_t(i,a)-Q^*(i,a)| \leq \gamma|Q_t-Q^*|$. We look at the operator H after performing all state-action pairs, $\|HQ-Q^*\| \leq max_{i,a}|HQ_t(i,a)-Q^*(i,a)| \leq \gamma\|Q_t-Q^*\| \leq \gamma\|Q-Q^*\|$. ■

8. Synchronous Q-learning

In this section we give the proof of Theorems 2 and 3. Our main focus will be the value of $r_t = \|Q_t - Q^*\|$, and our aim is to bound the time until $r_t \leq \varepsilon$. We use a sequence of values D_i, such that $D_{k+1} = (1-\beta)D_k$ and $D_1 = V_{max}$. As in Section 6, we will consider times τ_k such that for any $t \geq \tau_k$ we have $r_t \leq D_k$. We call the time between τ_k and τ_{k+1} the kth iteration. (Note the distinction between a step of the algorithm and an iteration, which is a sequence of many steps.)

Our proof has two parts. The first (and simple) part is bounding the number of iterations until $D_i \leq \varepsilon$. The bound is derived in the following Lemma.

Lemma 12 *For $m \geq \frac{1}{\beta}\ln(V_{max}/\varepsilon)$ we have $D_m \leq \varepsilon$.*

Proof We have that $D_1 = V_{max}$ and $D_i = (1-\beta)D_{i-1}$. We want to find the m that satisfies $D_m = V_{max}(1-\beta)^m \leq \varepsilon$. By taking a logarithm over both sides of the inequality we get $m \geq \frac{1}{\beta}\ln(V_{max}/\varepsilon)$. ■

The second (and much more involved) part is to bound the number of steps in an iteration. We use the following quantities introduced in Section 6. Let $W_{t+1,\tau}(s,a) = (1-\alpha_t^{\omega}(s,a))W_{t,\tau}(s,a) + \alpha_t^{\omega}(s,a)w_t(s,a)$, where $W_{\tau;\tau}(s,a)=0$ and

$$w_t(s,a) = R(s,a) + \gamma \max_{b\in U(s')} Q_t(s',b) - \sum_{j=1}^{|S|} P_{s,j}(a)\left(R(s,a) + \gamma \max_{b\in U(j)} Q_t(j,b)\right),$$

where s' is the state reached after performing action a at state s. Let

$$Y_{t+1;\tau_k}(s,a) = (1-\alpha_t^\omega(s,a))Y_{t;\tau_k}(s,a) + \alpha_t^\omega(s,a)\gamma D_k,$$

where $Y_{\tau_k;\tau_k}(s,a) = D_k$. Our first step is to rephrase Lemma 9 for our setting.

Lemma 13 *For every state s action a and time τ_k, we have*

$$-Y_{t;\tau_k}(s,a) + W_{t;\tau_k}(s,a) \le Q^*(s,a) - Q_t(s,a) \le Y_{t;\tau_k}(s,a) + W_{t;\tau_k}(s,a)$$

The above lemma suggests (once again) that in order to bound the error r_t one can bound $Y_{t;\tau_k}$ and $W_{t;\tau_k}$ separately, and the two bounds imply a bound on r_t. We first bound the Y_t term, which is deterministic process, and then we bound the term, $W_{t;\tau}$, which is stochastic.

8.1 Synchronous Q-learning using a Polynomial Learning Rate

We start with Q-learning using a polynomial learning rate and show that the duration of iteration k, which starts at time τ_k and ends at time τ_{k+1}, is bounded by τ_k^ω. For synchronous Q-learning with a polynomial learning rate we define $\tau_{k+1} = \tau_k + \tau_k^\omega$, where τ_1 will be specified latter.

Lemma 14 *Consider synchronous Q-learning with a polynomial learning rate and assume that for any $t \ge \tau_k$ we have $Y_{t;\tau_k}(s,a) \le D_k$. Then for any $t \ge \tau_k + \tau_k^\omega = \tau_{k+1}$ we have $Y_{t;\tau_k}(s,a) \le D_k(\gamma + \frac{2}{e}\beta)$.*

Proof Let $Y_{\tau_k;\tau_k}(s,a) = \gamma D_k + \rho_{\tau_k}$, where $\rho_{\tau_k} = (1-\gamma)D_k$. We can now write

$$Y_{t+1;\tau_k}(s,a) = (1-\alpha_t^\omega)Y_{t;\tau_k}(s,a) + \alpha_t^\omega \gamma D_k = \gamma D_k + (1-\alpha_t^\omega)\rho_t,$$

where $\rho_{t+1} = \rho_t(1-\alpha_t^\omega)$. We would like to show that after time $\tau_{k+1} = \tau_k + \tau_k^\omega$ for any $t \ge \tau_{k+1}$ we have $\rho_t \le \frac{2}{e}\beta D_k$. By definition we can rewrite ρ_t as

$$\rho_t = (1-\gamma)D_k \prod_{l=1}^{t-\tau_k}(1-\alpha_{l+\tau_k}^\omega) = 2\beta D_k \prod_{l=1}^{t-\tau_k}(1-\alpha_{l+\tau_k}^\omega) = 2\beta D_k \prod_{l=1}^{t-\tau_k}(1-\frac{1}{(l+\tau_k)^\omega}),$$

where the last identity follows from the fact that $\alpha_t^\omega = 1/t^\omega$. Since the α_t^ω's are monotonically decreasing

$$\rho_t \le 2\beta D_k (1-\frac{1}{\tau_k^\omega})^{t-\tau_k}.$$

For $t \ge \tau_k + \tau_k^\omega$ we have

$$\rho_t \le 2\beta D_k (1-\frac{1}{\tau_k^\omega})^{\tau_k^\omega} \le \frac{2}{e}\beta D_k.$$

Hence, $Y_{t;\tau_k}(s,a) \le (\gamma + \frac{2}{e}\beta)D_k$. ■

Next we bound the term $W_{t;\tau_k}$ by $(1-\frac{2}{e})\beta D_k$. The sum of the bounds for $W_{t;\tau_k}(s,a)$ and $Y_{t;\tau_k}(s,a)$ would be $(\gamma+\beta)D_k = (1-\beta)D_k = D_{k+1}$, as desired.

Definition 15 *Let $W_{t;\tau_k}(s,a) = (1-\alpha_t^\omega(s,a))W_{t-1;\tau_k}(s,a) + \alpha_t^\omega(s,a)w_t(s,a)$ $= \sum_{i=\tau_k+1}^{t} \eta_i^{k,t}(s,a)w_i(s,a)$, where $\eta_i^{k,t}(s,a) = \alpha_{i+\tau_k}^\omega(s,a)\prod_{j=\tau_k+i+1}^{t}(1-\alpha_j^\omega(s,a))$ and let $W_{t;\tau_k}^l(s,a) = \sum_{i=\tau_k+1}^{\tau_k+l} \eta_i^{k,t}(s,a)w_i(s,a)$.*

Note that in the synchronous model $\alpha^{\omega}_t(s,a)$ and $\eta^{k,t}_i(s,a)$ are identical for every state action pair. We also note that $W^{t-\tau_k+1}_{t;\tau_k}(s,a) = W_{t;\tau_k}(s,a)$. We have bounded the term $Y_{t;\tau_k}$, for $t = \tau_{k+1}$. This bound holds for any $t \geq \tau_{k+1}$, since the sequence $Y_{t;\tau_k}$ is monotonically decreasing. In contrast, the term $W_{t;\tau_k}$ is stochastic. Therefore it is not sufficient to bound $W_{\tau_{k+1};\tau_k}$, but we need to bound $W_{t;\tau_k}$ for $t \geq \tau_{k+1}$. However, it is sufficient to consider $t \in [\tau_{k+1}, \tau_{k+2}]$. The following lemma bounds the coefficients in that interval.

Lemma 16 *For any $t \in [\tau_{k+1}, \tau_{k+2}]$ and $i \in [\tau_k, t]$, we have $\eta^{k,t}_i = \Theta(\frac{1}{\tau_k{}^{\omega}})$,*

Proof Since $\eta^{k,t}_i = \alpha^{\omega}_{i+\tau_k}\prod_{j=\tau_k+i+1}^{t}(1-\alpha^{\omega}_j)$,we can divide $\eta^{k,t}_t$ into two parts, the first one $\alpha^{\omega}_{i+\tau_k}$ and the second one $\mu = \prod_{j=\tau_k+i+1}^{t}(1-\alpha^{\omega}_j)$. We show that the first one is $\Theta(\frac{1}{\tau^{\omega}_k})$ and the second is constant.
Since $\alpha^{\omega}_{i+\tau_k}$ are monotonically decreasing we have for every $i \in [\tau_k, \tau_{k+2}]$ we have $\alpha^{\omega}_{\tau_k} \leq \alpha^{\omega}_i \leq \alpha^{\omega}_{\tau_{k+2}}$, thus $\frac{1}{\tau^{\omega}_k} \leq \alpha^{\omega}_t \leq \frac{1}{(\tau_{k+1}+\tau^{\omega}_{k+1})^{\omega}} \leq \frac{1}{(3\tau^{\omega}_k+\tau_k)^{\omega}} < \frac{1}{4\tau^{\omega}_k}$. Next we bound μ. Clearly μ is bounded from above by 1. Also $\mu \geq \prod_{j=\tau_k}^{\tau_{k+2}}(1-\alpha_j) \geq (1-\frac{1}{\tau_k{}^{\omega}})^{3\tau_k{}^{\omega}} \geq \frac{1}{e^3}$. Therefore, we have that for every $t \in [\tau_k, \tau_{k+2}]$, $\eta^{k,t}_i = \Theta(\frac{1}{\tau_k{}^{\omega}})$. ■

We introduce Azuma's inequality, which bounds the deviations of a martingale. The use of Azuma's inequality is mainly needed for the asynchronous case.

Lemma 17 (Azuma 1967) *Let $X_0, X_1, \ldots, X_n$ be a martingale sequence such that for each $1 \leq k \leq n$,*

$$|X_k - X_{k-1}| \leq c_k,$$

where the constant c_k may depend on k. Then for all $n \geq 1$ and any $\varepsilon > 0$

$$Pr\left[|X_n - X_0| > \varepsilon\right] \leq 2e^{-\frac{\varepsilon}{2\sum_{k=1}^{n}c_k^2}}$$

Next we show that Azuma's inequality can be applied to $W^l_{t;\tau_k}$.

Lemma 18 *For any $t \in [\tau_{k+1}, \tau_{k+2}]$ and $1 \leq l \leq t$ we have that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, which satisfies*

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| \leq \Theta(\frac{V_{max}}{\tau^{\omega}_k})$$

Proof We first note that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, since

$$\begin{aligned} E[W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)|F_{\tau_k+l-1}] &= E[\eta^{k,t}_{\tau_k+l}(s,a)w_{\tau_k+l}(s,a)|F_{\tau_k+l-1}] \\ &= \eta^{k,t}_{\tau_k+l}E[w_{\tau_k+l}(s,a)|F_{\tau_k+l-1}] = 0. \end{aligned}$$

By Lemma 16 we have that $\eta^{k,t}_l(s,a) = \Theta(1/\tau^{\omega}_k)$, thus

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| = \eta^{k,t}_{\tau_k+l}(s,a)|w_{\tau_k+l}(s,a)| \leq \Theta(\frac{V_{max}}{\tau^{\omega}_k}).$$

■

The following lemma provides a bound for the stochastic error caused by the term $W_{t;\tau_k}$ by using Azuma's inequality.

Lemma 19 *Consider synchronous Q-learning with a polynomial learning rate. With probability at least* $1-\frac{\delta}{m}$ *we have* $|W_{t;\tau_k}| \leq (1-\frac{2}{e})\beta D_k$ *for any* $t \in [\tau_{k+1}, \tau_{k+2}]$*, i.e.,*

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]: \ |W_{t;\tau_k}| \leq (1-\frac{2}{e})\beta D_k\right] \geq 1-\frac{\delta}{m}$$

given that $\tau_k = \Theta((\frac{V_{max}^2 \ln(V_{max}|S|\,|A|m/(\delta\beta D_k))}{\beta^2 D_k^2})^{1/\omega})$.

Proof By Lemma 18 we can apply Azuma's inequality to $W_{t;\tau_k}^{t-\tau_k+1}$ with $c_i = \Theta(\frac{V_{max}}{\tau_k^\omega})$ (note that $W_{t;\tau_k}^{t-\tau_k+1} = W_{t;\tau_k}$) . Therefore, we derive that

$$Pr\left[|W_{t;\tau_k}(s,a)| \geq \tilde{\varepsilon} \mid t \in [\tau_{k+1}, \tau_{k+2}]\right] \ \leq \ 2e^{\frac{-2\tilde{\varepsilon}^2}{\sum_{i=\tau_k}^{t} c_i^2}} \leq 2e^{-c\tau_k^\omega \tilde{\varepsilon}^2/V_{max}^2},$$

for some constant $c > 0$. Set $\tilde{\delta}_k = 2e^{-c\tau_k^\omega\tilde{\varepsilon}^2/V_{max}^2}$, which holds for $\tau_k^\omega = \Theta(\ln(1/\tilde{\delta})V_{max}^2/\tilde{\varepsilon^2})$. Using the union bound we have,

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]: \ W_{t;\tau_k}(s,a) \leq \tilde{\varepsilon}\right] \leq \sum_{t=\tau_{k+1}}^{\tau_{k+2}} Pr\left[W_{t;\tau_k}(s,a) \leq \tilde{\varepsilon}\right],$$

thus taking $\tilde{\delta}_k = \frac{\delta}{m(\tau_{k+2}-\tau_{k+1})|S||A|}$ assures that with probability at least $1-\frac{\delta}{m}$ the statement hold at every state-action pair and time $t \in [\tau_{k+1}, \tau_{k+2}]$. As a result we have,

$$\tau_k = \Theta\left((\frac{V_{max}^2 \ln(|S|\,|A|\,m\tau_k^\omega/\delta)}{\tilde{\varepsilon}^2})^{1/\omega}\right) = \Theta\left((\frac{V_{max}^2 \ln(|S|\,|A|\,mV_{max}/\delta\tilde{\varepsilon})}{\tilde{\varepsilon}^2})^{1/\omega}\right)$$

Setting $\tilde{\varepsilon} = (1-2/e)\beta D_k$ gives the desire bound. ■

We have bounded for each iteration the time needed to achieve the desired precision level with probability $1-\frac{\delta}{m}$. The following lemma provides a bound for the error in all the iterations.

Lemma 20 *Consider synchronous Q-learning using a polynomial learning rate. With probability at least* $1-\delta$*, for every iteration* $k \in [1,m]$ *and time* $t \in [\tau_{k+1}, \tau_{k+2}]$ *we have* $W_{t;\tau_k} \leq (1-\frac{2}{e})\beta D_k$*, i.e.,*

$$Pr\left[\forall k \in [1,m], \ \forall t \in [\tau_{k+1}, \tau_{k+2}]: \ |W_{t;\tau_k}| \leq (1-\frac{2}{e})\beta D_k\right] \geq 1-\delta,$$

given that $\tau_0 = \Theta((\frac{V_{max}^2 \ln(V_{max}|S|\,|A|/(\delta\beta\varepsilon))}{\beta^2\varepsilon^2})^{1/\omega})$.

Proof From Lemma 19 we know that

$$\begin{aligned} &Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]: \ |W_{t;\tau_k}| \geq (1-\frac{2}{e})\beta D_k\right] \\ &\quad = Pr\left[\forall s,a \ \forall t \in [\tau_{k+1}, \tau_{k+2}]: \ \sum_{i=\tau_k}^{t} w_i(s,a)\eta_i^{k,t} \geq \tilde{\varepsilon}\right] \leq \frac{\delta}{m} \end{aligned}$$

Using the union bound we have,

$$Pr\left[\forall s,a\,\forall k\leq m,\forall t\in[\tau_{k+1},\tau_{k+2}]\quad \sum_{i=\tau_k}^{t} w_i(s,a)\eta_i^{k,t}\geq\tilde{\varepsilon}\right]$$
$$\leq \sum_{k=1}^{m} Pr\left[\forall s,a\,\forall t\in[\tau_{k+1},\tau_{k+2}]\quad \sum_{i=\tau_k}^{t} w_i(s,a)\eta_i^{k,t}\geq\tilde{\varepsilon}\right]\leq\delta$$

where $\tilde{\varepsilon}=(1-\frac{2}{e})\beta D_k$. ■

We have bounded both the size of each iteration, as a function of its starting time, and the number of the iterations needed. The following lemma solves the recurrence $\tau_{k+1}=\tau_k+\tau_k^{\omega}$ and bounds the total time required (which is a special case of Lemma 32).

Lemma 21 *Let*

$$a_{k+1}=a_k+a_k^{\omega}=a_0+\sum_{i=0}^{k} a_i^{\omega}.$$

For any constant $\omega\in(0,1)$, $a_k=O((a_0^{1-\omega}+k)^{\frac{1}{1-\omega}})=O(a_o+k^{\frac{1}{1-\omega}})$

The proof of Theorem 2 follows from Lemma 21, Lemma 20, Lemma 12 and Lemma 14.

8.2 Synchronous Q-learning using a Linear Learning Rate

In this subsection, we derive results for Q-learning with a linear learning rate. The proof is very similar in spirit to the proof of Theorem 2 and we give here analogous lemmas to the ones in Subsection 8.1. First, the number of iterations required for synchronous Q-learning with a linear learning rate is the same as that for a polynomial learning rate. Therefore, we only need to analyze the number of steps in an iteration.

Lemma 22 *Consider synchronous Q-learning with a linear learning rate and assume that for any* $t\geq\tau_k$ *we have* $Y_{t;\tau_k}(s,a)\leq D_k$. *Then for any* $t\geq(2+\psi)\tau_k=\tau_{k+1}$ *we have* $Y_{t;\tau_k}(s,a)\leq D_k(\gamma+\frac{2}{2+\psi}\beta)$

Proof Let $Y_{\tau_k;\tau_k}(s,a)=\gamma D_k+\rho_{\tau_k}$, where $\rho_{\tau_k}=(1-\gamma)D_k$. We can now write

$$Y_{t+1;\tau_k}(s,a)=(1-\alpha_t)Y_{t;\tau_k}(s,a)+\alpha_t\gamma D_k=\gamma D_k+(1-\alpha_t)\rho_t,$$

where $\rho_{t+1}=\rho_t(1-\alpha_t)$. We would like show that after time $(2+\psi)\tau_k=\tau_{k+1}$ for any $t\geq\tau_{k+1}$ we have $\rho_t\leq\beta D_k$. By definition we can rewrite ρ_t as,

$$\rho_t=(1-\gamma)D_k\prod_{l=1}^{t-\tau_k}(1-\alpha_{l+\tau_k})=2\beta D_k\prod_{l=1}^{t-\tau_k}(1-\alpha_{l+\tau_k})=2\beta D_k\prod_{l=1}^{t-\tau_k}(1-\frac{1}{l+\tau_k}),$$

where the last identity follows from the fact that $\alpha_t=1/t$. Simplifying the expression, and setting $t=(2+\psi)\tau_k$, we have

$$\rho_t\leq 2D_k\beta\frac{\tau_k}{t}=\frac{2D_k\beta}{2+\psi}$$

Hence, $Y_{t;\tau_k}(s,a) \leq (\gamma + \frac{2}{2+\psi}\beta)D_k$. ■

The following lemma enables the use of Azuma's inequality.

Lemma 23 *For any $t \geq \tau_k$ and $1 \leq l \leq t$ we have that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, which satisfies*

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| \leq \frac{V_{max}}{t}$$

Proof We first note that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, since

$$\begin{aligned} E[W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)|F_{\tau_k+l-1}] &= E[\eta^{k,t}_{\tau_k+l}(s,a)w_{\tau_k+l}(s,a)|F_{\tau_k+l-1}] \\ &= \eta^{k,t}_{\tau_k+l}E[w_{\tau_k+l}(s,a)|F_{\tau_k+l-1}] = 0. \end{aligned}$$

For linear learning rate we have that $\eta^{k,t}_{l+\tau_k}(s,a) = 1/t$, thus

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| = \frac{w_{\tau_k+l}(s,a)}{t} \leq \frac{V_{max}}{t}.$$

■

The following Lemma provides a bound for the stochastic term $W_{t;\tau_k}$.

Lemma 24 *Consider synchronous Q-learning with a linear learning rate. With probability at least $1-\frac{\delta}{m}$, we have $|W_{t;\tau_k}| \leq \frac{\psi}{2+\psi}\beta D_k$ for any $t \in [\tau_{k+1}, \tau_{k+2}]$ and any positive constant ψ, i.e.*

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]: \ W_{t;\tau_k} \leq \frac{\psi}{2+\psi}\beta D_k\right] \geq 1 - \frac{\delta}{m}$$

given that $\tau_k = \Theta(\frac{V^2_{max}\ln(V_{max}|S|\ |A|m/(\psi\delta\beta D_k))}{\psi^2\beta^2 D_k^2})$

Proof By Lemma 23 for any $t \geq \tau_{k+1}$ we can apply Azuma's inequality to $W^{t-\tau_k+1}_{t;\tau_k}$ with $c_i = \frac{V_{max}}{i+\tau_k}$ (note that $W^{t-\tau_k+1}_{t;\tau_k} = W_{t;\tau_k}$). Therefore, we derive that

$$Pr[|W_{t;\tau_k}| \geq \tilde{\varepsilon} \mid t \geq \tau_{k+1}] \leq 2e^{\frac{-2\tilde{\varepsilon}^2}{\sum_{i=\tau_k}^t c_i^2}} = 2e^{-c\frac{t^2\tilde{\varepsilon}^2}{(t-\tau_k)V^2_{max}}} \leq 2e^{-c\frac{t\tilde{\varepsilon}^2}{V^2_{max}}}$$

for some positive constant c. Using the union bound we get

$$\begin{aligned} Pr[\forall t \in [\tau_{k+1}, \tau_{k+2}]: \ |W_{t;\tau_k}| \geq \tilde{\varepsilon}] &\leq Pr[\forall t \geq (2+\psi)\tau_k: \ |W_{t;\tau_k}| \geq \tilde{\varepsilon}] \\ &\leq \sum_{t=(2+\psi)\tau_k}^{\infty} Pr[|W_{t;\tau_k}| \geq \tilde{\varepsilon}] \\ &\leq \sum_{t=(2+\psi)\tau_k}^{\infty} 2e^{-c\frac{t\tilde{\varepsilon}^2}{V^2_{max}}} = 2e^{-c\frac{((2+\psi)\tau_k)\tilde{\varepsilon}^2}{V^2_{max}}}\sum_{t=0}^{\infty} e^{-\frac{t\tilde{\varepsilon}}{V^2_{max}}} \\ &= \frac{2e^{-c\frac{(2+\psi)\tau_k\tilde{\varepsilon}^2}{V^2_{max}}}}{1-e^{\frac{-\tilde{\varepsilon}^2}{V^2_{max}}}} = \Theta(\frac{e^{-\frac{c'\tau_k\tilde{\varepsilon}^2}{V^2_{max}}}V^2_{max}}{\tilde{\varepsilon}^2}), \end{aligned}$$

where the last equality is due to Taylor expansion and c' is some positive constant. By setting $\frac{\delta}{m\,|S|\,|A|} = \Theta(\frac{e^{-\frac{c'\tau_k\tilde{\varepsilon}^2}{V_{max}^2}} V_{max}^2}{\tilde{\varepsilon}^2})$, which holds for $\tau_k = \Theta(\frac{V_{max}^2 \ln(V_{max}|S|\,|A|m/(\delta\tilde{\varepsilon}))}{\tilde{\varepsilon}^2})$, and $\tilde{\varepsilon} = \frac{\psi}{2+\psi}\beta D_k$ assures us that for every $t \geq \tau_{k+1}$ (and as a result for any $t \in [\tau_{k+1}, \tau_{k+2}]$) with probability at least $1-\frac{\delta}{m}$ the statement holds at every state-action pair. ■

We have bounded for each iteration the time needed to achieve the desired precision level with probability $1-\frac{\delta}{m}$. The following lemma provides a bound for the error in all the iterations.

Lemma 25 *Consider synchronous Q-learning using a linear learning rate. With probability* $1-\delta$*, for every iteration* $k \in [1,m]$*, time* $t \in [\tau_{k+1}, \tau_{k+2}]$ *and any constant* $\psi > 0$ *we have* $|W_{t;\tau_k}| \leq \frac{\psi\beta D_k}{2+\psi}$*), i.e.,*

$$Pr\left[\forall k \in [1,m],\ \forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \leq \frac{\psi\beta D_k}{2+\psi}\right] \geq 1-\delta,$$

given that $\tau_0 = \Theta(\frac{V_{max}^2 \ln(V_{max}|S|\,|A|m/(\psi\delta\beta\varepsilon))}{\psi^2\beta^2\varepsilon^2})$.

Proof From Lemma 19 we know that

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \geq \frac{\psi\beta D_k}{2+\psi}\right] \leq \frac{\delta}{m}$$

Using the union bound we have that,

$$\begin{aligned} & Pr\left[\forall k \leq m, \forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \geq \frac{\psi\beta D_k}{2+\psi}\right] \\ & \leq \sum_{k=1}^{m} Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \geq \frac{\psi\beta D_k}{2+\psi}\right] \leq \delta \end{aligned}$$

■

The proof of Theorem 5 follows from Lemmas 25 and 22, 12 and the fact that $a_{k+1} = (2+\psi)a_k = (2+\psi)^k a_1$.

9. Asynchronous Q-learning

The major difference between synchronous and asynchronous Q-learning is that asynchronous Q-learning updates only one state action pair at each time while synchronous Q-learning updates all state-action pairs at each time unit. This causes two difficulties: the first is that different updates use different values of the Q function in their update. This problem is fairly easy to handle given the machinery introduced. The second, and more basic problem is that each state-action pair should occur enough times for the update to progress. To ensure this, we introduce the notion of covering time, denoted by L. We first extend the analysis of synchronous Q-learning to asynchronous Q-learning, in which each run always has covering time L, which implies that from any start state, in L steps all state-action pairs are performed. Later we relax the requirement such that the condition holds only with probability $1/2$, and show that with high probability we have a covering time of $L\log T$ for a run of length T. Note that our notion of covering time does not assume a stationary distribution of the exploration strategy; it may be the case that at some periods of time certain state-action pairs are more frequent while in other periods different state-action pairs are more frequent. In fact, we do not even assume that the sequence of state-action pairs is generated by a strategy—it can be an arbitrary sequence of state-action pairs, along with their reward and next state.

Definition 26 *Let $n(s,a,t_1,t_2)$ be the number of times that the state action pair (s,a) was performed in the time interval $[t_1,t_2]$.*

In this section, we use the same notations as in Section 8 for D_k, τ_k, $Y_{t;\tau_k}$ and $W_{t;\tau_k}$, with a different set of values for τ_k. We first give the results for asynchronous Q-learning using polynomial learning rate (Subsection 9.1); we give a similar proof for linear learning rates in Subsection 9.2.

9.1 Asynchronous Q-learning using a Polynomial Learning Rate

Our main goal is to show that the size of the kth iteration is $L\tau_k^\omega$. The covering time property guarantees that in $L\tau_k^\omega$ steps each pair of state action is performed at least τ_k^ω times. For this reason we define for asynchronous Q-learning with polynomial learning rate the sequence $\tau_{k+1} = \tau_k + L\tau_k^\omega$, where τ_1 will be specified later. As in Subsection 8.1 we first bound the value of $Y_{t;\tau_k}$

Lemma 27 *Consider asynchronous Q-learning with a polynomial learning rate and assume that for any $t \geq \tau_k$ we have $Y_{t;\tau_k}(s,a) \leq D_k$. Then for any $t \geq \tau_k + L\tau_k^\omega = \tau_{k+1}$ we have $Y_{t;\tau_k}(s,a) \leq D(\gamma + \frac{2}{e}\beta)$*

Proof For each state-action pair (s,a) we are assured that $n(s,a,\tau_k,\tau_{k+1}) \geq \tau_k^\omega$, since the covering time is L and the underlying policy has made $L\tau_k^\omega$ steps. Using the fact that the $Y_{t;\tau_k}(s,a)$ are monotonically decreasing and deterministic, we can apply the same argument as in the proof of Lemma 14. ■

The next Lemma bounds the influence of each sample $w_t(s,a)$ on $W_{t;\tau_k}(s,a)$.

Lemma 28 *Let $\tilde{w}^t_{i+\tau_k}(s,a) = \eta_i^{k,t}(s,a) w_{i+\tau_k}(s,a)$ then for any $t \in [\tau_{k+1},\tau_{k+2}]$ the random variable $\tilde{w}^t_{i+\tau_k}(s,a)$ has zero mean and bounded by $(L/\tau_k)^\omega V_{max}$.*

Proof Note that by definition $w_{\tau_k+i}(s,a)$ has zero mean and is bounded by V_{max} for any history and state-action pair. In a time interval of length τ, by definition of the covering time, each state-action pair is performed at least τ/L times; therefore, $\eta_i^{k,t}(s,a) \leq (L/\tau_k)^\omega$. Looking at the expectation of $\tilde{w}_{i+\tau_k}(s,a)$ we observe that

$$E[\tilde{w}^t_{i+\tau_k}(s,a)] = E[\eta_i^{k,t}(s,a) w_{i+\tau_k}(s,a)] = \eta_i^{k,t}(s,a) E[w_{i+\tau_k}(s,a)] = 0$$

Next we prove that it is bounded as well:

$$\begin{aligned} |\tilde{w}^t_{i+\tau_k}(s,a)| &= |\eta_i^{k,t}(s,a) w_{i+\tau_k}(s,a)| \\ &\leq |\eta_i^{k,t}(s,a)| V_{max} \\ &\leq (L/\tau_k)^\omega V_{max} \end{aligned}$$

■

Next we define $W^l_{t;\tau_k}(s,a) = \sum_{i=1}^{l} \tilde{w}^t_{i+\tau_k}(s,a)$ and prove that it is martingale sequence with bounded differences.

Lemma 29 *For any $t \in [\tau_{k+1}, \tau_{k+2}]$ and $1 \le l \le t$ we have that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, which satisfies*

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| \le (L/\tau_k)^{\omega} V_{max}$$

Proof We first note that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, since

$$E[W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a) | F_{\tau_k+l-1}] \quad = \quad E[\tilde{w}^t_{l+\tau_k}(s,a) | F_{\tau_k+l-1}] = 0.$$

By Lemma 28 we have that $\tilde{w}^t_{l+\tau_k}(s,a)$ is bounded by $(L/\tau_k)^{\omega} V_{max}$, thus

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| = \tilde{w}^t_{l+\tau_k}(s,a) \le (L/\tau_k)^{\omega} V_{max}.$$

■

The following Lemma bounds the value of the term $W_{t;\tau_k}$.

Lemma 30 *Consider asynchronous Q-learning with a polynomial learning rate. With probability at least $1-\frac{\delta}{m}$ we have for every state-action pair $|W_{t;\tau_k}(s,a)| \le (1-\frac{2}{e})\beta D_k$ for any $t \in [\tau_{k+1}, \tau_{k+2}]$, i.e.*

$$Pr\left[\forall s,a\ \forall t \in [\tau_{k+1}, \tau_{k+2}]:\ |W_{t;\tau_k}(s,a)| \le (1-\frac{2}{e})\beta D_k\right] \ge 1 - \frac{\delta}{m}$$

given that $\tau_k = \Theta((\frac{L^{1+3\omega} V_{max}^2 \ln(V_{max}|S|\,|A|m/(\delta\beta D_k))}{\beta^2 D_k^2})^{1/\omega})$.

Proof For each state-action pair we look on $W^l_{t;\tau_k}(s,a)$ and note that $W_{t;\tau_k}(s,a) = W^{t-\tau_k+1}_{t;\tau_k}(s,a)$. Let $\ell = n(s,a,\tau_k,t)$, then for any $t \in [\tau_{k+1}, \tau_{k+2}]$ we have that $\ell \le \tau_{k+2} - \tau_k \le \Theta(L^{1+\omega}\tau_k^{\omega})$. By Lemma 29 we can apply Azuma's inequality to $W^{t-\tau_k+1}_{t;\tau_k}(s,a)$ with $c_i = (L/\tau_k)^{\omega} V_{max}$. Therefore, we derive that

$$\begin{aligned} Pr[|W_{t;\tau_k}(s,a)| \ge \tilde{\varepsilon} \mid t \in [\tau_{k+1}, \tau_{k+2}]] \quad &\le \quad 2e^{\frac{-\tilde{\varepsilon}^2}{2\sum_{i=\tau_k+1, i\in T^{s,a}}^{t} c_i^2}} \le 2e^{-c\frac{\tilde{\varepsilon}^2 \tau_k^{2\omega}}{\ell V_{max}^2 L^{2\omega}}} \\ &\le \quad 2e^{-c\frac{\tilde{\varepsilon}^2 \tau_k^{\omega}}{L^{1+3\omega} V_{max}^2}}, \end{aligned}$$

for some constant $c > 0$. We can set $\tilde{\delta}_k = 2e^{-c\tau_k^{\omega}\tilde{\varepsilon}^2/(L^{1+3\omega}V_{max}^2)}$, which holds for $\tau_k^{\omega} = \Theta(\ln(1/\tilde{\delta}_k) L^{1+3\omega} V_{max}^2/\tilde{\varepsilon}^2)$. Using the union bound we have

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]:\ W_{t;\tau_k}(s,a) \le \tilde{\varepsilon}\right] \le \sum_{t=\tau_{k+1}}^{\tau_{k+2}} Pr\left[W_{t;\tau_k}(s,a) \le \tilde{\varepsilon}\right],$$

thus taking $\tilde{\delta}_k = \frac{\delta}{m(\tau_{k+2}-\tau_{k+1})|S||A|}$ assures a certainty level of $1-\frac{\delta}{m}$ for each state-action pair. As a result we have

$$\tau_k^{\omega} = \Theta(\frac{L^{1+3\omega} V_{max}^2 \ln(|S|\,|A|\,m\tau_k^{\omega}/\delta)}{\tilde{\varepsilon}^2}) = \Theta(\frac{L^{1+3\omega} V_{max}^2 \ln(|S|\,|A|\,mV_{max}/(\delta\tilde{\varepsilon}))}{\tilde{\varepsilon}^2})$$

Setting $\tilde{\varepsilon} = (1-2/e)\beta D_k$ give the desired bound. ■

We have bounded for each iteration the time needed to achieve the desired precision level with probability $1-\frac{\delta}{m}$. The following lemma provides a bound for the error in all the iterations.

Lemma 31 *Consider asynchronous Q-learning using a polynomial learning rate. With probability $1-\delta$, for every iteration $k \in [1,m]$ and time $t \in [\tau_{k+1}, \tau_{k+2}]$ we have $|W_{t;\tau_k}(s,a)| \leq (1-\frac{2}{e})\beta D_k$, i.e.,*

$$Pr\left[\forall k \in [1,m],\ \forall t \in [\tau_{k+1}, \tau_{k+2}],\ \forall s,a:\ \ |W_{t;\tau_k}(s,a)| \leq (1-\frac{2}{e})\beta D_k\right] \geq 1-\delta,$$

given that $\tau_0 = \Theta((\frac{L^{1+3\omega} V_{max}^2 \ln(V_{max}|S|\ |A|m/(\delta\beta\epsilon))}{\beta^2\epsilon^2})^{1/\omega})$.

Proof From Lemma 30 we know that

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \geq (1-\frac{2}{e})\beta D_k\right] \leq \frac{\delta}{m}$$

Using the union bound we have that

$$Pr\left[\forall k \leq m, \forall t \in [\tau_{k+1}, \tau_{k+2}]\ \ |W_{t;\tau_k}| \geq \tilde{\epsilon}\right] \quad \leq \quad \sum_{k=1}^{m} Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]\ \ |W_{t;\tau_k}| \geq \tilde{\epsilon}\right] \leq \delta,$$

where $\tilde{\epsilon} = (1-\frac{2}{e})\beta D_k$ ■

The following lemma solves the recurrence $\sum_{i=0}^{m-1} L\tau_i^{\omega} + \tau_0$ and derives the time complexity.

Lemma 32 *Let*

$$a_{k+1} = a_k + La_k^{\omega} = a_0 + \sum_{i=0}^{k} La_i^{\omega}$$

Then for any constant $\omega \in (0,1)$, $a_k = O((a_0^{1-\omega} + Lk)^{\frac{1}{1-\omega}}) = O(a_0 + (LK)^{\frac{1}{1-\omega}})$.

Proof We define the following series

$$b_{k+1} = \sum_{i=0}^{k} Lb_i^{\omega} + b_0$$

with an initial condition

$$b_0 = L^{\frac{1}{1-\omega}}.$$

We show by induction that $b_k \leq (L(k+1))^{\frac{1}{1-\omega}}$ for $k \geq 1$. For $k = 0$

$$b_0 = L^{\frac{1}{1-\omega}}(0+1)^{\frac{1}{1-\omega}} \leq L^{\frac{1}{1-\omega}}$$

We assume that the induction hypothesis holds for $k-1$ and prove it for k,

$$b_k = b_{k-1} + Lb_{k-1}^{\omega} \leq (Lk)^{\frac{1}{1-\omega}} + L(Lk)^{\frac{\omega}{1-\omega}} \leq L^{\frac{1}{1-\omega}} k^{\frac{\omega}{1-\omega}}(k+1) \leq (L(k+1))^{\frac{1}{1-\omega}}$$

and the claim is proved.

Now we lower bound b_k by $(L(k+1)/2)^{1/(1-\omega)}$. For $k = 0$

$$b_0 = L^{\frac{1}{1-\omega}} \geq (\frac{L}{2})^{\frac{1}{1-\omega}}$$

Assume that the induction hypothesis holds for $k-1$ and prove for k,

$$\begin{aligned} b_k &= b_{k-1} + Lb_{k-1}^{\omega} = (Lk/2)^{\frac{1}{1-\omega}} + L(Lk/2)^{\frac{\omega}{1-\omega}} = L^{\frac{1}{1-\omega}}((k/2)^{\frac{1}{1-\omega}} + (k/2)^{\frac{\omega}{1-\omega}}) \\ &\geq L^{\frac{1}{1-\omega}}((k+1)/2)^{\frac{1}{1-\omega}}. \end{aligned}$$

For $a_0 > L^{\frac{1}{1-\omega}}$ we can view the series as starting at $b_k = a_0$. From the lower bound we know that the start point has moved $\Theta(a_0^{1-\omega}/L)$. Therefore we have a total complexity of $O((a_0^{1-\omega} + Lk)^{\frac{1}{1-\omega}}) = O(a_0 + (LK)^{\frac{1}{1-\omega}}))$.

■

The proof of Theorem 4 follows from Lemmas 27, 31,12 and 32. In the following lemma we relax the condition of the covering time.

Lemma 33 *Assume that from any start state with probability* $1/2$ *in* L *steps we perform all state action pairs. Then with probability* $1-\delta$*, from any start state we perform all state action pairs in* $L\log_2(1/\delta)$ *steps, for a run of length* $[L\log_2(1/\delta)]$.

Proof The proof follows from the fact that after k intervals of length L (where k is a natural number), the probability of not visiting all state action pairs is 2^{-k}. Since we have $k = [\log_2(1/\delta)]$ we get that the probability of failing is δ. ■

Corollary 34 *Assume that from any start state with probability* $1/2$ *in* L *steps we perform all state action pairs. Then with probability* $1-\delta$*, from any start state we perform all state action pairs in* $L\log(T/\delta)$ *steps, for a run of length* T.

9.2 Asynchronous Q-learning using a Linear Learning Rate

In this section we consider asynchronous Q-learning with a linear learning rate. Our main goal is to show that the size of the kth iteration is $L(1+\psi)\tau_k$, for any constant $\psi > 0$. The covering time property guarantees that in $(1+\psi)L\tau_k$ steps each pair of state action is performed at least $(1+\psi)\tau_k$ times. The sequence of times in this case is $\tau_{k+1} = \tau_k + (1+\psi)L\tau_k$, where the τ_0 will be defined latter. We first bound $Y_{t;\tau_k}$ and then bound the stochastic term $W_{t;\tau_k}$.

Lemma 35 *Consider asynchronous Q-learning with a polynomial learning rate and assume that for any* $t \geq \tau_k$ *we have* $Y_{t;\tau_k}(s,a) \leq D_k$*. Then for any* $t \geq \tau_k + (1+\psi)L\tau_k = \tau_{k+1}$ *we have* $Y_{t;\tau_k}(s,a) \leq (\gamma + \frac{2}{2+\psi}\beta)D_k$

Proof For each state-action pair (s,a) we are assured that $n(s,a,\tau_k,\tau_{k+1}) \geq (1+\psi)\tau_k$, since in an interval of $(1+\psi)L\tau_k$ steps each state-action pair is visited at least $(1+\psi)\tau_k$ times by the definition of the covering time. Using the fact that the $Y_{t;\tau_k}(s,a)$ are monotonically decreasing and deterministic (thus independent), we can apply the same argument as in the proof of Lemma 22. ■

The following Lemma enables the use of Azuma's inequality.

Lemma 36 *For any $t \geq \tau_k$ and $1 \leq l \leq t$ we have that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, which satisfies*

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| \leq \frac{V_{max}}{n(s,a,0,t)}$$

Proof We first note that $W^l_{t;\tau_k}(s,a)$ is a martingale sequence, since

$$\begin{aligned} E[W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)|F_{\tau_k+l-1}] &= E[\eta^{k,t}_{\tau_k+l}(s,a)w_{\tau_k+l}(s,a)|F_{\tau_k+l-1}] \\ &= \eta^{k,t}_{\tau_k+l}E[w_{\tau_k+l}(s,a)|F_{\tau_k+l-1}] = 0. \end{aligned}$$

For a linear learning rate we have that $\eta^{k,t}_{\tau_k+l}(s,a) = 1/n(s,a,0,t)$, thus

$$|W^l_{t;\tau_k}(s,a) - W^{l-1}_{t;\tau_k}(s,a)| = \eta^{k,t}_{\tau_k+l}(s,a)|w_{\tau_k+l}(s,a)| \leq \frac{V_{max}}{n(s,a,0,t)}.$$

■

The following lemma bounds the value of the term $W_{t;\tau_k}$.

Lemma 37 *Consider asynchronous Q-learning with a linear learning rate. With probability at least $1-\frac{\delta}{m}$ we have for every state-action pair $|W_{t;\tau_k}(s,a)| \leq \frac{\psi}{2+\psi}\beta D_k$ for any $t \geq \tau_{k+1}$ and any positive constant ψ, i.e.*

$$Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+1}]: \ W_{t;\tau_k}(s,a) \leq \frac{\psi}{2+\psi}\beta D_k\right] \geq 1 - \frac{\delta}{m}$$

given that $\tau_k \geq \Theta((\frac{V^2_{max}\ln(V_{max}|S|\,|A|m/(\delta\beta D_k\psi))}{\psi^2\beta^2 D_k^2}))$.

Proof By Lemma 36 we can apply Azuma's inequality on $W^{t-\tau_k+1}_{t;\tau_k}$ (We note that $W^{t-\tau_k+1}_{t;\tau_k} = W_{t;\tau_k}$) with $c_i = \Theta(\frac{V_{max}}{n(s,a,0,t})$ for any $t \geq \tau_{k+1}$. Therefore, we derive that

$$Pr[|W_{t;\tau_k}| \geq \tilde{\varepsilon}] \quad \leq \quad 2e^{\frac{-2\tilde{\varepsilon}^2}{\sum^t_{i=\tau_k, i\in T^{s,a}} c_i^2}} \leq 2e^{-c\frac{n(s,a,\tau_k,t)\tilde{\varepsilon}^2}{V^2_{max}}},$$

for some positive constant c. Let us define the following variable

$$\zeta_t(s,a) = \begin{cases} 1, & \alpha_t(s,a) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

Using the union bound and the fact in an interval of length $(1+\psi)L\tau_k$ each state-action pair is visited at least $(1+\psi)\tau_k$ times, we get

$$\begin{aligned} Pr[\forall t \in [\tau_{k+1}, \tau_{k+2}]: \ |W_{t;\tau_k}(s,a)| \geq \tilde{\varepsilon}] &\leq Pr[\forall t \geq ((1+\psi)L+1)\tau_k: \ |W_{t;\tau_k}(s,a)| \geq \tilde{\varepsilon}] \\ &\leq \sum_{t=((1+\psi)L+1)\tau_k}^{\infty} Pr[|W_{t;\tau_k}(s,a)| \geq \tilde{\varepsilon}] \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=((1+\psi)L+1)\tau_k}^{\infty} \zeta_t(s,a) 2e^{-c\frac{n(s,a,0,t)\tilde{\varepsilon}^2}{V_{max}^2}} \\
&\leq 2e^{-c\frac{((1+\psi)\tau_k)\tilde{\varepsilon}^2}{V_{max}^2}} \sum_{t=0}^{\infty} e^{-\frac{t\tilde{\varepsilon}^2}{2V_{max}^2}} \\
&= \frac{2e^{-c\frac{(1+\psi)\tau_k\tilde{\varepsilon}^2}{V_{max}^2}}}{1-e^{\frac{-\tilde{\varepsilon}^2}{V_{max}^2}}} = \Theta\left(\frac{e^{-\frac{c'\tau_k\tilde{\varepsilon}^2}{V_{max}^2}} V_{max}^2}{\tilde{\varepsilon}^2}\right),
\end{aligned}
$$

for some positive constant c'. Setting $\frac{\delta}{m\,|S|\,|A|} = \Theta(\frac{e^{-\frac{c'\tau_k\tilde{\varepsilon}^2}{V_{max}^2}} V_{max}^2}{\tilde{\varepsilon}^2})$, which hold for $\tau_k = \Theta(\frac{V_{max}^2 \ln(V_{max}|S|\,|A|m/(\delta\tilde{\varepsilon}))}{\tilde{\varepsilon}^2})$, and $\tilde{\varepsilon} = \frac{\psi}{2+\psi}\beta D_k$ assures us that for every $t \geq \tau_{k+1}$ (and as a result for any $t \in [\tau_{k+1}, \tau_{k+2}]$) with probability at least $1-\frac{\delta}{m}$ the statement holds at every state-action pair.

■

We have bounded for each iteration the time needed to achieve the desired precision level with probability $1-\frac{\delta}{m}$. The following lemma provides a bound for the error in all the iterations.

Lemma 38 *Consider synchronous Q-learning using a linear learning rate. With probability* $1-\delta$*, for every iteration* $k \in [1,m]$*, time* $t \in [\tau_{k+1}, \tau_{k+2}]$ *and any constant* $\psi > 0$ *we have* $|W_{t;\tau_k}| \leq \frac{\psi\beta D_k}{2+\psi}$*, i.e.,*

$$
Pr\left[\forall k \in [1,m],\ \forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \leq \frac{\psi\beta D_k}{2+\psi}\right] \geq 1-\delta,
$$

given that $\tau_0 = \Theta(\frac{V_{max}^2 \ln(V_{max}|S|\,|A|m/(\delta\beta\varepsilon\psi))}{\psi^2\beta^2\varepsilon^2})$.

Proof From Lemma 37 we know that

$$
Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]:\ \ |W_{t;\tau_k}| \geq \frac{\psi\beta D_k}{2+\psi}\right] \leq \frac{\delta}{m}
$$

Using the union bound we have that,

$$
\begin{aligned}
&Pr\left[\forall k \leq m, \forall t \in [\tau_{k+1}, \tau_{k+2}]\ \ |W_{t;\tau_k}| \geq \frac{\psi\beta D_k}{2+\psi}\right] \\
&\quad \leq \sum_{k=1}^{m} Pr\left[\forall t \in [\tau_{k+1}, \tau_{k+2}]\ \ |W_{t;\tau_k}| \geq \frac{\psi\beta D_k}{2+\psi}\right] \leq \delta
\end{aligned}
$$

■

Theorem 5 follows from Lemmas 35, 38, 12 and the fact that $a_{k+1} = a_k + (1+\psi)La_k = a_0((1+\psi)L+1)^k$.

10. Lower Bound for Q-learning using a Linear Learning Rate

In this section we show a lower bound for Q-learning with a linear learning rate, which is $O((\frac{1}{\varepsilon})^{\frac{1}{1-\gamma}})$. We consider the following MDP, denoted M_0, that has a single state s, a single action a, and a deterministic reward $R_{M_0}(s,a) = 0$. Since there is only one action in the MDP we denote $Q_t(s,a)$ as $Q_t(s)$. We initialize $Q_0(s) = 1$ and observe the time until $Q_t(s) \leq \varepsilon$.

Lemma 39 *Consider running synchronous Q-learning with linear learning rate on MDP M_0, when initializing $Q_0(s) = 1$. Then there is a time $t = c(\frac{1}{\varepsilon})^{\frac{1}{1-\gamma}}$ for some constant $c > 0$, such that $Q_t \geq \varepsilon$.*

Proof First we prove by induction on t that

$$Q_t(s) = \prod_{i=1}^{t-1} \frac{i+\gamma}{i+1}.$$

For $t = 1$ we have $Q_1(s) = (1-1/2)Q_0(s) + (1/2)\gamma Q_0(s) = (1+\gamma)/2$. Assume the hypothesis holds for $t-1$ and prove it for t. By definition,

$$Q_t(s) = (1-\frac{1}{t})Q_{t-1}(s) + \frac{1}{t}\gamma Q_{t-1}(s) = \frac{t-1+\gamma}{t}Q_{t-1}(s).$$

In order to help us estimate this quantity we use the Γ function. Let

$$\Gamma(x+1,k) = \frac{1\cdot 2\cdots k}{(x+1)\cdot(x+2)\cdot\cdots\cdot(x+k)}k^x$$

The limit of $\Gamma(1+x,k)$, as k goes to infinity, is constant for any x. We can rewrite $Q_t(s)$ as

$$Q_t(s) = \frac{1}{\Gamma(\gamma+1,t)}\frac{t^\gamma}{t+1} = \Theta(t^{\gamma-1})$$

Therefore, there is a time $t = c(\frac{1}{\varepsilon})^{\frac{1}{1-\gamma}}$, for some constant $c > 0$, such that $Q_t(s) \geq \varepsilon$. ∎

Acknowledgments

This research was supported in part by a grant from the Israel Science Foundation. Eyal Even-Dar was partially supported by the Deutsch Institute.

References

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

F. Beleznay, T. Grobler, and C. Szepesvari. Comparing value-function estimation algorithms in undiscounted problems. Technical Report TR-99-02, Mindmaker Ltd, 1999.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

V.S. Borkar and S.P. Meyn. The O.D.E method for convergence of stochstic approximation and reinforcement learning. *Siam J. Control*, 38 (2):447–69, 2000.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation, 6*, 1994.

Michael Kearns and Satinder P. Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In M.J. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 996–1002, 1999.

Michael L. Littman and Gaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In L. Saitta, editor, *Proceedings of the 13th International Conference on Machine Learning (ICML-96)*, pages 310–318, Bari, Italy, 1996. Morgan Kaufmann. URL `citeseer.nj.nec.com/littman96generalized.html`.

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

R. Sutton and A. Barto. *Reinforcement Learning*. MIT Press., Cambridge, MA., 1998.

C. Szepesvari. The asymptotic convergence-rate of Q-learning. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 1064–1070, 1998.

J. Tsitsiklis. Asynchronous stochastic approximation and Q-learning, 1994.

C. Watkins and P. Dyan. Q-learning. *Machine Learning*, 8(3/4):279–292, 1992.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge, England, 1989.