Learning rates for Q-Learning

Eyal Even-Dar and Yishay Mansour

School of Computer Science, Tel-Aviv University
evend@cs.tau.ac.il mansour@cs.tau.ac.il

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 the 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 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 that this exponential behavior is inherent for a linear learning rate.

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 Markov Decision Process (MDP), which has become the dominating approach in Reinforcement Learning [SB98,BT96].

An MDP includes states, which abstract the environment, actions, which are the available actions to the agent, and for each state and action a distribution of next states, the state 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 the rewards to a single value that the agent tries to optimize. A discounted return has a parameter $\gamma \in (0,1)$ where the reward received in 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 base, where we first reconstruct a model of the MDP, and then find an optimal policy for the approximate model. The second approach are implicit methods that updates the information after each step, and based on this have an estimate to the optimal policy. The most popular of those methods in Q-learning [Wat89].

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, [BT96,Tsi94,WD92,MS96,JJS94,BM00]. 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 at 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 require to show that the convergence is with probability one).

We use the proof technique of [BT96], which is based on convergence of stochastic processes, to derive convergence rates for Q-learning. The most interesting outcome of our investigation is the relationship between the form of the learning rates and the rate of convergence. 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. 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 model that using 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 show an exponential behavior for linear learning rate. The lower bound simply shows that if the initial value is one and all the rewards are zero, it takes $O((1/\epsilon)^{1/(1-\gamma)})$ updates, using a linear learning rate, until we reach a value of ϵ.

The convergence rate of Q-learning in a batch setting, where many samples are averaged for each update, was analyzed in [KS98]. A batch setting does not have a learning rate and it has much of the flavor of model based techniques, since each update is an average of many samples.

The convergence of Q-learning with linear learning rate was studied in [Sze97] 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 distributed i.i.d.). For this model [Sze97] shows a convergence rate, which is exponential in $1/(1-\gamma)$. In [BGS99] an exponential lower bound in the number of the states is given for undiscounted return.

2 The Model

We define a Markov Decision process (MDP) as follows

Definition 2.1. *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), $P^M_{i,j}(a)$ is the transition probability from state i to state j when performing action $a \in U(i)$ in state i, and $R_M(a, s)$ is the reward received when performing action a in state s.*

We assume that $R_M(a, s)$ is non-negative and bounded by R_{max}, i.e, $\forall s, a : 0 \leq R_M(a, s) \leq R_{max}$. For simplicity we assume that the reward $R_M(a, s)$ is deterministic, however all our results apply when $R_M(a, s)$ 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 $R_M(a, s)$), 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 $Q^\pi_M(s,a) = R_M(a,s) + \gamma \sum_{\bar{s}} P_{s,\bar{s}}(a) V^\pi_M(\bar{s})$, 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 $\beta = (1-\gamma)/3$, which will be useful for 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 [Put94].) This implies that for any policy π and any state s we have $V^{\pi^*}_M(s) \geq V^\pi_M(s)$, and $\pi^*(s) = argmax_a(R_M(a,s) + \gamma(\sum_{s'} P^M_{s,s'}(a) \max_b Q(s',b))$. The optimal policy is also the only fixed point of the operator, $(TQ)(s,a) = R_M(a,s) + \gamma \sum_{s'} P_{s,s'}(a) \max_b Q(s',b)$. 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 $\|V^*_M - V^\pi_M\|_\infty \leq \epsilon$.

The *covering time*, denoted by L, is a bound on the number of steps needed, from any start state, until all state-action pairs are performed. Note that the covering time is a function of both the MDP and the underlying strategy. Initially we assume that from any start state, within L steps all state-action pairs are performed. 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 pairs are performed.

The Parallel Sampling Model, $PS(M)$, as was introduced in [KS98]. 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 $R_M(a,s)$. The advantage of this model is that it allows ignoring the exploration and focusing 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 [Wat89] estimates the state-action value function (for discounted return) as follows:

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

where s' is the state reached from state s when performing action a at time t. Let $T^{i,u}$ be the set of times, where action u was performed at state i, then $\alpha_t(\bar{s},\bar{a}) = 0$ for $t \notin T^{\bar{s},\bar{a}}$. It is known 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 $\sum_{t=1}^\infty \alpha^2_t(s,a) < \infty$ ([BT96,Tsi94,WD92,MS96,JJS94]).

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:
$\forall s,a : Q_0(s,a) = 0$

$\forall s, a : Q_{t+1}(s,a) = (1-\alpha_t^\omega)Q_t(s,a) + \alpha_t^\omega(R_M(a,s) + \gamma \max_{a\in U(\bar{s})} Q_t(\bar{s},a))$, where $\bar{s}$ is the state reached from state s when performing action a. 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). We add the assumption that the underlying strategy has a covering time of L. The updates are as follows:

$$\forall s, a : Q_0(s,a) = 0$$
$$\forall s, a : Q_{t+1}(s,a) = (1-\alpha_t^\omega(s,a))Q_t(s,a) + \alpha_t^\omega(s,a)(R_M(a,s) + \gamma \max_{b\in U(\bar{s})} Q_t(\bar{s},b))$$

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 at time t we perform action a in state s 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 $1/(1-\gamma) = 1/(3\beta)$. In contrast, 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 in 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 4.1. *Let Q_T be the value of the synchronous Q-learning algorithm using polynomial learning rate at time T. Then $||Q_T - Q^*|| \le \epsilon$, given that*

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

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 $\Omega((\ln(1/\epsilon)/\epsilon^2)^{1/\omega} + (\ln(1/\epsilon))^{1/(1-\omega)})$ which is optimized when ω approaches one. Considering the dependence only on β, recall that $V_{max} = R_{max}/(3\beta)$, 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 4.2. *Let Q_T be the value of the synchronous Q-learning algorithm using linear learning rate at time T. Then $||Q_T - Q^*|| \leq \epsilon$, given that*

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

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 4.3. *Let Q_T be the value of the asynchronous Q-learning algorithm using polynomial learning rate at time T. Then $||Q_T - Q^*|| \leq \epsilon$, given that*

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

The dependence on the covering time, in the above theorem, is $\Omega(L^{2+1/\omega} + L^{1/(1-\omega)})$, which is optimized for $\omega = \sqrt{2}/2$. 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 4.4. *Let Q_T be the value of the asynchronous Q-learning algorithm using linear learning rate at time T. Then $||Q_T - Q^*|| \leq \epsilon$, given that*

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

The following Theorem shows that a linear learning rate may require an exponential dependence on $1/(3\beta) = 1/(1-\gamma)$, thus showing that the gap between linear learning rate and polynomial learning rate is real and does exist for some MDPs. (The proof is in Appendix A.)

Theorem 4.5. *There exists an MDP such that Q-Learning with linear learning rate after $T = \Omega((\frac{1}{\epsilon})^{\frac{1}{1-\gamma}})$ steps has $||Q_T - Q^*_M|| > \epsilon$.*

5 Background from Stochastic Algorithms

Before we derive our proofs we first introduce the proof given in [BT96] for the convergence of stochastic iterative algorithms, and then show that Q-Learning algorithms fall in this category. In this section we review the proof for the convergence in the limit, and in the next section we will analyze the rate at which different Q-learning algorithms converge. (We will try to keep the background as close as possible to our needs for this paper.)

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

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

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 mappings.

Definition 5.1. *An iterative stochastic algorithm is well behaved if:*

1. *The step size* $\alpha_t(i)$ *satisfy (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 following fact (proven in [BT96]) shows the connection with the Q-learning algorithms.

Fact 5.1 *Synchronous Q-learning and Asynchronous Q-learning are both well behaved stochastic iterative algorithms.*

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

Theorem 5.2. *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 in [BT96]. Without loss of generality, assume that $X^* = 0$ and $\|X_0\| \leq A$. The value of X_t is bounded since for any history F_t we have $\|w_t\| \leq A$, and since $\|X_0\| \leq A$, hence, for any t we have $\|X_t\| \leq A$.

Recall that $\beta = \frac{1-\gamma}{3}$. 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

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

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 $t \geq \tau_k$,

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

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 $Y_{t;\tau_k}$ (the proof can be found in [BT96]).

Lemma 5.1. *For every i, we have*

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

Proof. 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 $\gamma D_k + \beta D_k$ 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 $Y_{t;\tau} \leq (\gamma+\beta)D_k + \beta D_k/2$, and $|W_{t;\tau_k}| \leq \beta D_k/2$. This fact, together with Lemma. 5.1, yields that for $t \geq \tau_{k+1}$,

$$||X_t|| \leq (\gamma + 2\beta)D_k = D_{k+1},$$

which completes the proof. □

6 Synchronous Q-Learning

In this section we sketch the proof of Theorem 4.1. The proof of Theorem 4.2 is similar in spirit and can be found in Appendix B. Our main focus will be the value of $r_t = \|Q_t - Q^*\|$, and our aim is to bound the time until $r_t \leq \epsilon$. 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 5, 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 $D_i \leq \epsilon$. The bound is derived in the following fact.

Fact 6.1 *For $m \geq \frac{1}{\beta}\ln(V_{max}/\epsilon)$ we have $D_m \leq \epsilon$*

The second (and much more involve) part is to bound the number of steps in an iteration. We use the following quantities introduced in Section 5. Let $W_{t+1,\tau}(s,a) = (1-\alpha_t^\omega)W_{t,\tau}(s,a) + \alpha_t^\omega w_t(s,a)$, where $W_{\tau;\tau}(s,a) = 0$ and

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

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

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

where $Y_{\tau_k;\tau_k}(s,a) = D_k$. Our first step is to rephrase Lemma 5.1 for our setting.

Lemma 6.1. *For every state s action a and time τ_k, we have*

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

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. 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 polynomial learning rate we define $\tau_{k+1} = \tau_k + \tau_k^\omega$, where τ_1 will be specified latter.

Lemma 6.2. *Consider synchronous 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 $t \geq \tau_k + \tau_k^\omega = \tau_{k+1}$ we have $Y_{t;\tau_k}(s,a) \leq D_k(\gamma + \beta + \frac{2}{e}\beta)$*

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

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

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

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

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

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

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

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

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

Next we bound the term $W_{t;\tau_k}(s,a)$ 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 $(\gamma+2\beta)D_k = (1-\beta)D_k = D_{k+1}$, as desired.

Definition 6.1. *Let* $W_{t;\tau_k}(s,a) = (1-\alpha_t^\omega)W_{t-1;\tau_k}(s,a) + \alpha_t^\omega w_t(s,a)$ $= \sum_{i=\tau+1}^{t} \eta_i^{k,t} w_{\tau_k+i}(s,a)$ *, where* $\eta_i^{k,t} = \alpha_{i+\tau_k}^\omega \prod_{j=\tau_k+i+1}^{t}(1-\alpha_j^\omega)$.

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

The following lemma provides a bound for the stochastic error caused by the term $W_{t;\tau_k}$.

Lemma 6.4. *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.*

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

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. Note that by definition $w_{\tau_k+i}(s,a)$ has zero mean and is bounded by V_{max} for any history. Since $\eta_i^{k,t} = \Theta(1/\tau_k^\omega)$ and $t-\tau_k = \Theta(\tau_k^\omega)$, we can define a random variables $\tilde{w}_{i+\tau_k}(s,a) = (t-\tau_k)\eta_i^{k,t} w_{i+\tau_k}(s,a)$, which have zero mean and are bounded by $\Theta(V_{max})$. Using a Chernoff bound for every state-action pair we can derive that,

$$\begin{aligned} Pr\left[W_{t;\tau_k}(s,a) \geq \tilde{\epsilon}\right] &= Pr\left[\sum_{i=\tau_k}^{t} w_i(s,a)\eta_{i-\tau_k}^{k,t} \geq \tilde{\epsilon}\right] \\ &= Pr\left[\frac{1}{t-\tau_k}\sum_{i=\tau_k}^{t} \tilde{w}_i(s,a) \geq \tilde{\epsilon}\right] \leq e^{-c\tau_k^\omega \tilde{\epsilon}^2/V_{max}^2}, \end{aligned}$$

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

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

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,

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

Setting $\tilde{\epsilon} = (1-2/e)\beta D_k$ 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 6.5. *Consider synchronous 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 $W_{t;\tau_k} \leq (1-\frac{2}{e})\beta D_k$, 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 $\tau_0 = \Theta((\frac{V_{max}^2 \ln(V_{max}|S|\ |A|/(\delta\beta\epsilon))}{\beta^2\epsilon^2})^{1/\omega})$

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 $\tau_{k+1} = \tau_k + \tau_k^\omega$, which bounds the total time required. (The proof follows from Lemma 7.3.)

Lemma 6.6. *Let*

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

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

The proof of Theorem 4.1 follows from Lemma 6.6, Lemma 6.4 and Lemma 6.2.

7 Asynchronous Q-learning

The major difference between synchronous and asynchronous Q-learning is that in the asynchronous Q-learning updates only one state action pair at each time while the synchronous Q-learning updates all the state action pairs 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, is that each state action pair should occur enough times, so that the update can progress. To overcome this we introduce the notion of covering time, denoted by L. We first extend the analysis of the synchronous Q-learning to asynchronous Q-learning, in which each run has covering time L with probability one, this implies that for any start state in L steps all state action pairs are performed. Latter we relax the probability such that the condition holds 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 with their reward and next state.

Definition 7.1. *Let $n(s,a,t_1,t_2)$ be the number of times that (s,a) was performed in the time interval $[t_1,t_2]$.*

We first give the results for asynchronous Q-learning using polynomial learning rate. (A similar proof to linear learning rate can be found in Appendix D). We use in this section the same notations as in Section 6 for D_k, τ_k, $Y_{t;\tau_k}$ and $W_{t;\tau}$, with a different set of values for τ_k. 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 latter. As in Section 6 we first bound the value of $Y_{t;\tau_k}$ (proof omitted).

Lemma 7.1. *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 $Y_{t;\tau_k}(s,a) \leq D(\gamma + \beta + \frac{2}{e}\beta)$*

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

Lemma 7.2. *Consider asynchronous Q-learning with a polynomial learning rate. With probability at least $1-\frac{\delta}{m}$ we have for every state-action pair $|W_{t;\tau_k}(s,a)| \leq (1-\frac{2}{e})\beta D_k$ 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 $\tau_k = \Theta((\frac{L^{1+2\omega}V_{max}^2 \ln(V_{max}|S|\ |A|m/(\delta\beta D_k))}{\beta^2 D_k^2})^{1/\omega})$

Proof. We sketch the proof. Note that by definition w_{τ_k+i} has zero mean and is bounded by V_{max} for any history. 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, $\eta_i^{k,t}(s,a) \leq (L/\tau_k)^\omega$. Let $\ell = n(s,a,\tau_k,t)$ by definition $\ell = \Theta(L\tau_k^\omega)$, we can define a random variable $\tilde{w}_{i+\tau_k}(s,a) = \ell\eta_i^{k,t}(s,a)w_{i+\tau_k}(s,a)$, which has zero mean and is bounded by $O(V_{max}\ell(L/\tau_k)^\omega)$. Using a Chernoff bound for every state action pair we can derive that,

$$\begin{aligned} Pr\left[W_{t;\tau_k}(s,a) \geq \tilde{\epsilon}\right] &= Pr\left[\sum_{i=\tau_k}^{t} w_i(s,a)\eta_{i-\tau_k}^{k,t} \geq \tilde{\epsilon}\right] \\ &= Pr\left[\frac{1}{n(s,a,\tau_k,t)}\sum_{i=\tau_k}^{t} \tilde{w}_i(s,a) \geq \tilde{\epsilon}\right] \\ &\leq e^{-c\frac{\ell\tilde{\epsilon}^2\tau_k^{2\omega}}{\ell^2 L^{2\omega}V_{max}^2}} \leq e^{-c\frac{\tilde{\epsilon}^2\tau_k^{\omega}}{L^{1+2\omega}V_{max}^2}}, \end{aligned}$$

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

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

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

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

Setting $\tilde{\epsilon} = (1-2/e)\beta D_k$ give the desire bound. □

The following lemma solves the recurrence $\sum_{i=0}^m L\tau^\omega + \tau_0$ and derives the time complexity. (The proof can be found is Appendix C.)

Lemma 7.3. *Let*

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

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

The proof of Theorem 4.3 follows from Lemmas 7.1, 7.2 and 7.3. In the following lemma we relax the condition of the covering time.

Lemma 7.4. *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* $L\log(\frac{T}{L}\frac{1}{\delta})$ *steps, for a run of length* T*.*

References

[BGS99] F. Beleznay, T. Grobler, and Cs. Szepesvari. Comparing value-function estimation algorithms in undiscounted problems. Technical Report TR-99-02, Mindmaker Ltd, 1999.

[BM00] 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.

[BT96] Dimitri P. Bertsekas and Jhon N. Tsitsklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

[JJS94] T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation, 6*, 1994.

[KS98] Michael Kearns and Stinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Neural Information Processing Systems 10*, 1998.

[MS96] Littman M. and Cs. Szepesvari. A generalized reinforcement learning model: convergence and applications. In *In International Conference on Machine Learning*, 1996.

[Put94] M.L Puterman. *Markov Decision Processes - Discrete Stochastic Dynamic Programming.* Jhon Wiley & Sons. Inc., New York, NY, 1994.

[SB98] Richard S. Sutton and Andrew G. Bato. *Reinforcement Learning.* Mit press, 1998.

[Sze97] Cs. Szepesvari. The asymptotic convergence-rate of q-learning. In *Neural Information Processing Systems 10*, pages 1064 – 1070, 1997.

[Tsi94] Jhon N. Tsitsklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16:185–202, 1994.

[Wat89] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.

[WD92] C. Watking and P. Dyan. Q-learning. *Machine Learning*, 8(3/4):279 –292, 1992.

A Lower bound for Q-Learning using linear learning rate

In this section we show a lower bound for Q-Learning with linear learning rate, which is $O((\frac{1}{\epsilon})^{\frac{1}{1-\gamma}})$. 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 $Q_t(s)\leq\epsilon$.

Lemma A.1. *Consider running synchronous Q-learning with linear learning rate on MDP M_0, when initializing $Q_0(s)=1$. Then for $t=O((\frac{1}{\epsilon})^{\frac{1}{1-\gamma}})$ we have $Q_t\geq\epsilon$.*

Proof. First we prove by induction on t that

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

For $t=2$ 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 $t=c(\frac{1}{\epsilon})^{\frac{1}{1-\gamma}}$, for some constant $c>0$, such that $Q_t(s)\geq\epsilon$. $\square$

B Synchronous Q-learning with linear learning rate

In this section we derive the results for A-learning with linear learning rate. The proof is very similar in spirit and we give here the analogue lemmas to the ones in Section 6. First, the number of iterations required for synchronous Q-learning with linear learning rate is the same as for polynomial learning rate. Therefore, we need to analyze the number of steps in an iteration.

Lemma B.1. *Consider synchronous 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 $t\geq 4\tau_k=\tau_{k+1}$ we have $Y_{t;\tau_k}(s,a)\leq D_k(\gamma+\beta+\frac{1}{2}\beta)$*

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

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

Therefore, $\rho_{t+1} = \rho_t(1-\alpha_t^1)$. We would like show that after time $4\tau_k = \tau_{k+1}$ for any $t \geq \tau_{k+1}$ we have $\rho_t \leq \beta D_k$. By definition we can rewrite ρ_t as,

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

where the last identity follows from the fact that $\alpha_t = 1/t$. Simplifying the expression, and setting $t = 4\tau_k$, we have,

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

Hence, $Y_{t;\tau_k}(s,a) \leq (\gamma+\beta+\frac{1}{2}\beta)D_k$. □

In a similar way to Lemma 6.4 we show,

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

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

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

Theorem 4.4 follows from Lemmas B.2 and B.1, and the fact that $a_{k+1} = 4a_k = 4^k a_1$.

C Proof of Lemma 7.3

Proof of Lemma 7.3 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 $k-1$ 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 $k=0$ we have,

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

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

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

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 $O((a_0^{1-\omega} + Lk)^{\frac{1}{1-\omega}})$. □

D Asynchronous Q-learning using Linear Learning Rate

In this section we consider asynchronous Q-learning with a linear learning rate. The sequence of times in this case is $\tau_{k+1} = \tau_k + 3L\tau_k$. Due to lack of space we only state the lemmas that bound $Y_{t;\tau}$ and then bound $W_{t;\tau}$.

Lemma D.1. *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 $t \geq \tau_k + 3L\tau_k = \tau_{k+1}$ we have $Y_{t;\tau_k}(s,a) \leq (\gamma + \beta + \frac{1}{2}\beta)D_k$*

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

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

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

Theorem 4.4 follows from Lemmas D.1, D.2 and the fact that $a_{k+1} = a_k + 3La_k = (3L+1)^k$.