Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP

Kefan Dong∗†, Yuanhao Wang∗†, Xiaoyu Chen§2, and Liwei Wang¶2,3

1Institute for Interdisciplinary Information Sciences, Tsinghua University, China
2Key Laboratory of Machine Perception, MOE, School of EECS, Peking University
3Center for Data Science, Peking University, Beijing Institute of Big Data Research

January 29, 2019

Abstract

A fundamental question in reinforcement learning is whether model-free algorithms are sample efficient. Recently, Jin et al. [6] proposed a Q-learning algorithm with UCB exploration policy, and proved it has nearly optimal regret bound for finite-horizon episodic MDP. In this paper, we adapt Q-learning with UCB-exploration bonus to infinite-horizon MDP with discounted rewards without accessing a generative model. We show that the sample complexity of exploration of our algorithm is bounded by \(\tilde{O}(\frac{SA}{\epsilon^2(1-\gamma)^7})\). This improves the previously best known result of \(\tilde{O}(\frac{SA}{\epsilon^4(1-\gamma)^8})\) in this setting achieved by delayed Q-learning [13], and matches the lower bound in terms of \(\epsilon\) as well as \(S\) and \(A\) except for logarithmic factors.

1 Introduction

The goal of reinforcement learning is to construct algorithms that learn and plan in sequential decision making systems when the underlying system dynamics are unknown. A typical model in RL is Markov Decision Process (MDP). At each time step, the environment is in state \(s\). The agent may take an action \(a\), obtain a reward, and then the environment may transit to another state. In reinforcement learning, the transition probability distribution is unknown. The algorithm needs to learn the transition dynamics of MDP, while aiming to maximize the cumulative reward. This causes an exploration-exploitation dilemma: whether to act to gain new information (explore) or to act consistently with past experience to maximize reward (exploit).

Theoretical analysis of reinforcement learning falls into two broad categories: those assuming a simulator (a.k.a. generative model), and those without a simulator. In the first category, the algorithm is able to query the outcome of any state action pair from an oracle. The emphasis is on the number of calls needed to estimate the \(Q\) value or to output a near-optimal policy. There has been extensive research in literature following this line of research, the majority of which focuses on

∗These two authors contributed equally
†dkf16@mails.tsinghua.edu.cn
‡yuanhao-16@mails.tsinghua.edu.cn
§cxy30@pku.edu.cn
¶wanglw@cis.pku.edu.cn

arXiv:1901.09311v1 [cs.LG] 27 Jan 2019
discounted infinite horizon MDPs \cite{1,4,12}. The current results have achieved near-optimal time and sample complexities \cite{12}.

Without a simulator, there is a dichotomy between finite-horizon and infinite-horizon settings. In finite-horizon settings, there are straightforward definitions for both regret and sample complexity; the latter is defined as the number of samples needed before the policy becomes near optimal. In this setting, extensive research in the past decade \cite{6,2,5} has achieved great progress, and established nearly-tight bounds for both regret and sample complexity.

The infinite-horizon setting is a very different matter. First of all, the performance measure we use cannot be a straightforward extension of the sample complexity defined above. Instead, the measure of sample efficiency we adopt is the so-called sample complexity of exploration \cite{7}, which is also a widely-accepted definition. This measure counts the number of times that the algorithm “makes mistakes” along the whole trajectory. See also \cite{14} for further discussions regarding this issue.

Several model based algorithms have been proposed for infinite horizon MDP, for example R-max \cite{3}, MoRmax \cite{15} and UCLR-\gamma \cite{8}. It is noteworthy that there still exists a considerable gap between the state-of-the-art algorithm and the theoretical lower bound \cite{8}, but the gap is only about \(1/(1 - \gamma)\).

Though model-based algorithms have been proved to be sample efficient in various MDP settings, most state-of-the-art RL algorithms are developed in the model-free paradigm \cite{11,10,9}. Model-free algorithms are more flexible and require less space, which have achieved remarkable performance on benchmarks such as Atari games and simulated robot control problems.

For infinite horizon MDPs without access to simulator, the best known result of sample complexity of exploration is \(\tilde{O}\left(\frac{SA}{\epsilon(1-\gamma)^2}\right)\), achieved by delayed Q-learning \cite{13}. It is also the first algorithm to achieve a \(\tilde{O}(SA)\) bound in sample complexity of exploration. The authors provide a novel strategy of argument when proving the upperbound for the sample complexity of exploration, namely identifying a sufficient condition for optimality, and then bound the number of times that this condition is violated.

However, the results of Delayed Q-learning still leave a quadratic gap in \(1/\epsilon\) from the best-known lowerbound. This is partly because, many samples collected by the algorithms actually have no direct effect on the learning process. If one wants to bridge this gap, intuitively, the algorithm might need to make use of every sample. This, as well as the success of the Q-learning with UCB algorithm \cite{6} in proving a regret bound in finite-horizon settings, motivates us to incorporate a UCB-like exploration term into our algorithm.

In this work, we propose a Q-learning algorithm with a UCB exploration policy. We show the sample complexity of exploration bound of the algorithm is \(\tilde{O}\left(\frac{SA}{\epsilon(1-\gamma)^7}\right)\). This strictly improves the previous best known result due to Delayed Q-learning. It also matches the lower bound in terms in the dependence on \(\epsilon\), \(S\) and \(A\) (logarithmic factors ignored).

Although our algorithm is quiet similar to the algorithm proposed by \cite{6}, the analysis of the sample complexity of exploration for infinite-horizon MDP is challenging so that the techniques developed in \cite{6} do not directly apply here. Please see Section 3.2 for detailed explanations.

The rest of the paper is organized as follows. After introducing the notation used in the paper in Section 2, we describe our infinite Q-learning with UCB algorithm in Section 3. We then state our main theoretical results, which are in the form of PAC sample complexity bounds. In Section 4 we present some interesting properties beyond sample complexity bound. Finally, we conclude the paper in Section 5.
2 Preliminary

We consider a Markov Decision Process defined by a five tuple \( \langle S, A, p, r, \gamma \rangle \), where \( S \) is the state space, \( A \) is the action space, \( p(s'|s,a) \) is the transition function, \( r : S \times A \rightarrow [0, 1] \) is the deterministic reward function, and \( 0 \leq \gamma < 1 \) is the discount factor for rewards. Let \( S = |S| \) and \( A = |A| \) denote the number of states and the number of actions respectively.

Starting from a state \( s_1 \), the agent interacts with the environment for infinite number of time steps. At each time step, the agent observes state \( s_t \in S \), picks action \( a_t \in A \), and receives reward \( r_t \), then the system transits to next state \( s_{t+1} \).

A policy \( \pi_t : S \rightarrow A \) refers to the non-stationary control policy of the algorithm in step \( t \). We use \( V^{\pi_t}(s) \) to denote the value function under policy \( \pi_t \), which is defined as \( V^{\pi_t}(s) = \mathbb{E}[\sum_{i=1}^{\infty} \gamma^{i-1} r(s_i, \pi_t(s_i))|s_1 = s] \). We also use \( V^*(s) = \sup_{\pi} V^\pi(s) \) to denote the value function of the optimal policy. Accordingly, we define \( Q^\pi_t(s,a) = r(s,a) + \mathbb{E}[\sum_{i=2}^{\infty} \gamma^{i-1} r(s_i, \pi_t(s_i))|s_1 = s, a_1 = a] \) as the Q function under policy \( \pi_t \); \( Q^*(s,a) \) is the Q function under optimal policy \( \pi^* \).

We use the sample complexity of exploration defined in [7] to measure the learning efficiency of our algorithm. This sample complexity definition has been widely used in previous works [13, 8, 14].

**Definition 1.** Sample complexity of Exploration of an algorithm ALG \( \mathcal{G} \) is defined as the number of time steps \( t \) such that the non-stationary policy \( \pi_t \) at time \( t \) is not \( \epsilon \)-optimal for current state \( s_t \), i.e. \( V^{\pi_t}(s_t) < V^*(s_t) - \epsilon \).

We use the following definition of PAC-MDP \([13]\).

**Definition 2.** An algorithm ALG is said to be \( \text{PAC-MDP} \) (Probably Approximately Correct in Markov Decision Processes) if, for any \( \epsilon \) and \( \delta \), the sample complexity of ALG is less than some polynomial in the relevant quantities \( (S, A, 1/\epsilon, 1/\delta, 1/(1-\gamma)) \), with probability at least \( 1 - \delta \).

Finally, recall that Bellman equation is defined as the following:

\[
\begin{align*}
V^{\pi_t}(s) &= Q^{\pi_t}(s, \pi_t(s)) \\
Q^{\pi_t}(s,a) &:= (r_t + \gamma \mathbb{E}[V^{\pi_t}](s')) \quad \forall s \in S, \\
V^*(s) &= Q^*(s, \pi^*(s)) \\
Q^*(s,a) &:= (r_t + \gamma \mathbb{E}[V^*](s')) \quad \forall s \in S,
\end{align*}
\]

which is frequently used in our analysis. Here we denote \( \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\pi_t}(s') \).

3 Main Results

In this section, we present the UCB Q-learning algorithm and the sample complexity bound.

3.1 Algorithm

Our UCB Q-learning algorithm (Algorithm 1) starts from an optimistic estimation of action value function \( Q(s,a) \) and its historical minimum value \( \hat{Q}(s,a) \). The target value \( U \) at time step \( t \) consists of three parts: the immediate reward \( r(s_t,a_t) \), the discounted value function of the next state \( \gamma \hat{V}(s_{t+1}) \) and the UCB term \( b_k \). Here \( \hat{V}(s) \leftarrow \max_{a \in A} Q(s,a) \) is the value function maintained by our algorithm.

The learning rate is defined as

\[
\alpha_k = \frac{H + 1}{H + k}.
\]
the number of time steps when arriving at not considered as a good policy under the sample complexity defined in infinite horizon MDPs. Therefore exploration, if behaves poorly at , it will contribute to the sample complexity every time is encountered, which may happen infinite times since the trajectory has infinite length. Therefore is not considered as a good policy under the sample complexity defined in infinite horizon MDPs.

3.2 Sample Complexity of Exploration

Our main result is the following sample complexity of exploration bound.

**Theorem 1.** For any , , with probability , the sample complexity of exploration (i.e., the number of time steps such that is not -optimal at ) of Algorithm 1 is at most

\[
\hat{O}\left(\frac{SA \ln 1/\delta}{\epsilon^2 (1 - \gamma)^4}\right),
\]

where \(\hat{O}\) hides log factors of \(1/\epsilon, 1/(1 - \gamma)\) and \(SA\).

We first point out the obstacles for proving the theorem and why the techniques in do not directly apply here. We then give a high level description of the ideas of our approach.

One important issue is caused by the difference in sample complexity for finite and infinite horizon MDP. For finite horizon settings, sample complexity only characterizes the performance (i.e., \(V^*(s) - V^*(s)\)) of a policy at the starting state of episodes . On the contrary, for infinite horizon settings, sample complexity of exploration characterizes the performance for all actions that the policy takes. For example, consider an MDP where there exists a state such that, for any policy , can only be reached with exponentially small probability from the starting state . The performance of policy at state , denoted as \(\delta(s) = V^*(s) - V^*(s)\), only contributes an exponentially small part (i.e., \(p\delta(s)\)) to the performance at , which is the only consideration of sample complexity defined finite horizon. More concretely, even if policy takes the worst actions at state , it can still be an -optimal policy for . However, for sample complexity of exploration, if behaves poorly at , it will contribute to the sample complexity every time is encountered, which may happen infinite times since the trajectory has infinite length. Therefore is not considered as a good policy under the sample complexity defined in infinite horizon MDPs.

---

**Algorithm 1 Infinite Q-learning with UCB**

| Parameters: | \(\epsilon_1, H, \tau, \gamma\) |
|-------------|----------------------------------|
| Initialize | \(Q(s, a), \hat{Q}(s, a) \leftarrow \frac{1}{1 - \gamma}, N(s, a) \leftarrow 0\) |
| for \(t = 1, 2, \ldots\) do |
| 5: Take action \(a_t \leftarrow \arg\max_a \hat{Q}(s_t, a')\) |
| 6: Receive reward \(r_t\) and transit to \(s_{t+1}\) |
| 7: \(N(s_t, a_t) \leftarrow N(s_t, a_t) + 1\) |
| 8: \(k \leftarrow N(s_t, a_t), b_k \leftarrow \frac{c_2}{1 - \gamma} \sqrt{\frac{H_\tau(k)}{k}}\) |
| 9: \(U = \left[r(s_t, a_t) + b_k + \gamma \hat{V}(s_{t+1})\right]\) |
| 10: \(Q(s_t, a_t) \leftarrow (1 - \alpha_k)Q(s_t, a_t) + \alpha_k U\) |
| end for |

\(\iota(k) = \ln(SA(k + 1)(t + 2)/\delta)\) is an log factor. \(\epsilon_1\) is a parameter whose value will be determined in the proof of Theorem 1. \(H\) is chosen as \(\frac{\ln 1/((1 - \gamma)\epsilon_1)}{\ln 1/\gamma} \leq \frac{\ln 1/((1 - \gamma)\epsilon_1)}{1 - \gamma}\).
The major reason that the techniques in [6] do not apply to our problem is the following. The key idea in [6] is a clever design of the learning rate so that at episode \( k \) and time \( h \), the learning error can be decomposed as the errors from a set of consecutive episodes before \( k \) at time \( h + 1 \). This nice property allows one to control the regret in a recursive manner. However the property heavily relies on the fact that the setting is episodic finite-horizon MDP.

For the setting of infinite-horizon MDP, the above property does not exist anymore. For our setting, suppose at time \( t \) the agent is at state \( s_t \) and takes action \( a_t \). Then the learning error at \( t \) only depends on those previous time steps such that the agent encountered the same state as \( s_t \) and took the same action as \( a_t \). Thus the learning error at time \( t \) cannot be decomposed as errors from a set of consecutive time steps before \( t \), but errors from a set of non-consecutive time steps without any structure. (Please see Fig. 1 for an illustration.) Therefore, we have to control the sum of learning errors over an unstructured set of time steps. This makes the analysis challenging.

![Figure 1: An illustration of how error is propagated for Q-learning algorithms in (a) finite horizon settings; (b) infinite horizon settings. At blocks with the same color the same state-action pair is experienced. The arrows indicate how one error contributes to future errors. The dashed frames indicate errors of interest (terms in the summation).](image)

Now we give a very brief description of the proof of Theorem 1. The basic idea is to establish a sufficient condition so that \( \pi_t \) learned at step \( t \) is \( \epsilon \)-optimal for state \( s_t \), i.e. \( V^*(s_t) - V^\pi_t(s_t) \leq \epsilon \). At a high level this follows the approach of [13]. It turns out that a sufficient condition for \( V^*(s_t) - V^\pi_t(s_t) \leq \epsilon \) is that \( V^*(s_{t'}) - Q^*(s_{t'}, a_{t'}) \) is small for a few time steps \( t' \) within an interval \( [t, t + R] \) for a carefully chosen \( R \). We then bound the total number of bad time steps on which \( V^*(s_t) - Q^*(s_t, a_t) \) is large for the whole MDP. This in turn relies on a key technical lemma (Lemma 2), which controls the weighted sum of learning errors.

We now present the formal proof of Theorem 1.

**Proof.** (Proof for Theorem 1) For a fixed \( s_t \), let \( \text{TRAJ}(R) \) be the set of length-\( R \) trajectories starting from \( s_t \). Our goal is to give a sufficient condition so that \( \pi_t \), the policy learned at step \( t \), is \( \epsilon \)-optimal. For any \( \epsilon_2 > 0 \), define \( R := \ln \frac{1}{\epsilon_2 (1 - \gamma) / (1 - \gamma)} \). Denote \( V^*(s_t) - Q^*(s_t, a_t) \) by \( \Delta_t \). We have

\[
V^*(s_t) - V^\pi_t(s_t)
\]
\[ V^*(s_t) - Q^*(s_t, a_t) + Q^*(s_t, a_t) - V^{\pi_t}(s_t) \]
\[ = V^*(s_t) - Q^*(s_t, a_t) + \gamma \mathbb{P} \left( V^* - V^{\pi + 1} \right)(s_t, \pi_t(s_t)) \]
\[ = V^*(s_t) - Q^*(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1}|s_t, \pi_t(s_t)) \cdot [V^*(s_{t+1}) - Q^*(s_{t+1}, a_{t+1})] + \gamma \sum_{s_{t+1}, s_{t+2}} p(s_{t+2}|s_{t+1}, \pi_{t+1}(s_{t+1})) \cdot p(s_{t+1}|s_t, \pi_t(s_t)) \cdot [V^*(s_{t+2}) - Q^*(s_{t+2}, a_{t+2})] \]
\[ \cdots \]
\[ \leq \epsilon_2 + \sum_{\text{traj} \in \text{TRAJ}(R)} p(\text{traj}) \cdot \left[ \sum_{j=0}^{R-1} \gamma^j \Delta_{t+j} \right], \quad (1) \]
where the last inequality holds because \( \frac{R}{1-\gamma} \leq \epsilon_2 \), which follows from the definition of \( R \).

For any fixed trajectory of length \( R \) starting from \( s_t \), let \( \Delta_{t'} = V^*(s_{t'}) - Q^*(s_{t'}, a_{t'}) \). Consider the sequence \((\Delta_{t'})_{t' \leq t+R} \). Let \( X_t^{(i)} \) be the \( i \)-th largest item of \((\Delta_{t'})_{t' \leq t+R} \). Rearranging Eq. (1), we obtain
\[ V^*(s_t) - V^{\pi_t}(s_t) \leq \epsilon_2 + \mathbb{E}_{\text{traj}} \left[ \sum_{i=1}^{R} \gamma^{i-1} X_t^{(i)} \right]. \]

Although the right-hand-side of the above inequality consists of the sum of \( R \) items \( X_t^{(1)}, \ldots, X_t^{(R)} \), we claim that there are \( \lceil \log_2 R \rceil \) items, if they are small then \( \pi_t \) is good. Specifically, let \( \xi_i = \frac{1}{2^{\gamma^{i+1}}} \epsilon_2 \left( \frac{1}{1-\gamma} \right)^{i+1} \). If for all \( 0 \leq i \leq \lceil \log_2 R \rceil \),
\[ \mathbb{E}[X_t^{(2^i)}] \leq \xi_i, \quad (2) \]
then \( V^*(s_t) - V^{\pi_t}(s_t) \leq 2\epsilon_2 \).

To see why this is true, note that \( X_t^{(i)} \) is monotonically decreasing with respect to \( i \). Eq. (2) implies that for \( 1/2 < \gamma < 1 \),
\[ \mathbb{E} \left[ \sum_{i=1}^{R} \gamma^{i-1} X_t^{(i)} \right] = \sum_{i=1}^{R} \gamma^{i-1} \mathbb{E}[X_t^{(i)}] \]
\[ \leq \sum_{i=1}^{R} \gamma^{i-1} 2^{-\lceil \log_2 i \rceil} \epsilon_2 \left( \frac{1}{1-\gamma} \right)^{i-1} \]
\[ \leq \sum_{i=1}^{R} \frac{\gamma^{i-1}}{i} \epsilon_2 \left( \frac{1}{1-\gamma} \right)^{i-1} \leq \epsilon_2, \]
where the last inequality follows from the fact that \( \sum_{i=1}^{\infty} \frac{\gamma^{i-1}}{i} = \frac{1}{2} \ln \frac{1}{1-\gamma} \).

For fixed \( i \), let \( M = 2 \log_2 \frac{1}{\xi_R(1-\gamma)} \), and \( \eta_j = \frac{\xi_i}{M+1} \cdot 2^{j-1} \). The reason behind the choice of \( M \) is
to ensure that \( \eta_M > 1/(1 - \gamma) \). It follows that, for \( 1 \leq j \leq M \),

\[
E[X_t^{(2^i)}] = \int_0^{1/(1 - \gamma)} \Pr[X_t^{(2^i)} > x] \, dx \\
\leq \eta_t + \sum_{j=1}^M \eta_j \Pr[X_t^{(2^i)} > \eta_{j-1}].
\]

Thus, a sufficient condition for \( E[X_t^{(2^i)}] \leq \xi_i \) is,

\[
\eta_j \Pr[X_t^{(2^i)} > \eta_{j-1}] \leq \frac{\xi_i}{M + 1}, \quad \forall 1 \leq j \leq M.
\] (3)

Eq. (3) implies that if a time step \( t \) is not \((2\epsilon_2)\)-optimal, there exists \( 0 \leq i < \lfloor \log_2 R \rfloor \) and \( 1 \leq j \leq M \) such that

\[
\eta_j \Pr[X_t^{(2^i)} > \eta_{j-1}] > \xi_i.
\]

Now, the sample complexity can be bounded by the number of \((i, j)\) pairs so that the above inequality holds.

We first fix \( i \) and consider how many \( j \) there are. The following lemma helps to provide an estimation.

**Lemma 1.** For fixed \( t \) and \( \eta > 0 \), let \( B^{(t)}_\eta \) be the event that \( V^*(s_t) - Q^*(s_t, a_t) > \frac{\eta}{1 - \gamma} \) in step \( t \). If \( \eta > 2\epsilon_1 \), then with probability at least \( 1 - \delta/2 \),

\[
\sum_{t=1}^{\infty} I[B^{(t)}_\eta] \leq \frac{SA \ln S \ln 1/\delta}{\eta^2(1 - \gamma)^3} \cdot \text{polylog}\left( \frac{1}{\epsilon_1}, \frac{1}{1 - \gamma} \right),
\] (4)

where \( I[\cdot] \) is the indicator function.

By lemma 1 for any \( 1 \leq j \leq M \),

\[
\sum_{t=1}^{\infty} I[V^*(s_t) - Q^*(s_t, a_t) > \eta_{j-1}] \leq C,
\]

where

\[
C = \frac{SA \ln S \ln 1/\delta}{\eta_j^2(1 - \gamma)^5} \cdot \tilde{P}.
\] (5)

Here \( \tilde{P} \) is a shorthand for \( \text{polylog}\left( \frac{1}{\epsilon_1}, \frac{1}{1 - \gamma} \right) \).

Let \( A_t = I[X_t^{(2^i)} \geq \eta_j] \) be a Bernoulli random variable, and \( \{\mathcal{F}_t\}_{t \geq 1} \) be the filtration generated by random variables \( \{(s_\tau, a_\tau) : 1 \leq \tau \leq t\} \). Since \( A_t \) is \( F_{t+R} \)-measurable, for any \( 0 \leq k < R \), \( \{A_{k+tR} - E[A_{k+tR} | \mathcal{F}_{k+tR}]\}_{t \geq 0} \) is a martingale difference sequence. For any \( 0 \leq k < R \), by Azuma-Hoeffding inequality, after \( T = O\left( \frac{C}{2\epsilon} \cdot \frac{\eta_j(M+1)}{\xi_i} \cdot \ln(RML) \right) \) time steps (if it happens that many times) with

\[
\Pr\left[ X_{k+tR}^{(2^i)} \geq \eta_j \right] = E[A_{k+tR}] > \frac{\xi_i}{\eta_j(M + 1)},
\] (6)

\footnote{We assume that \( \xi_R < 1/12 \), which is true when \( \epsilon < 1/3 \).}
we have 
\[ \sum_t A_{k+tR} \geq C/2^i \]
with probability at least \(1 - \delta/(2MRL)\). Here the \(\delta\) is the same as that in (5). The reason that \(\delta\) appears here is that \(T\) contains a \(\log 1/\delta\) factor.

On the other hand, if \(A_{j+tR}\) happens, within \([k+tR, k+tR+R-1]\), there must be at least \(2^i\) time steps at which \(V^*(s_t) - Q^*(s_t,a_t) > \eta_j\). The latter event happens at most \(C\) times, and \([k+tR, k+tR+R-1]\) are disjoint. Therefore,
\[ \sum_{t=0}^{\infty} A_{j+tR} \leq C/2^i. \]

This suggests that the event described by (6) happens at most \(T\) times for fixed \(i\) and \(j\). Via a union bound on \(0 \leq k < R\), we can show that with probability \(1 - \delta/(2ML)\), there are at most \(RT\) time steps where
\[ \Pr \left[ X_i^{(2^i)} \geq \epsilon_2 \cdot 2^{i-1} \right] > \frac{\xi_i}{\eta_j(M+1)}. \]

Then, we sum over \(j\) to upper bound the number of time steps that violate condition (3) for a fixed \(i\):
\[ \sum_{j=1}^{M} \frac{SA(M+1)R \ln 1/\delta \ln SA}{\eta_j \xi_i \cdot 2^i(1-\gamma)^6} \bar{P} = \frac{SA \ln SA \ln 1/\delta}{\epsilon_2(1-\gamma)^6} \bar{P}. \]

Finally, we sum over \(i\) to obtain an upper bound of the number of time steps when (3) is violated with some \(i\) and \(j\) (with probability \(1 - \delta/2\)):
\[ \sum_{i=0}^{\lceil \log_2 R \rceil} \frac{SA \ln SA \ln 1/\delta}{\epsilon_2(1-\gamma)^6} \bar{P} \leq \sum_{i=0}^{\lceil \log_2 R \rceil} \frac{SA \cdot 2^{i+6} \ln SA \ln 1/\delta}{\epsilon_2(1-\gamma)^6} \bar{P} \]
\[ \leq \frac{SAR \ln SA \ln 1/\delta}{\epsilon_2(1-\gamma)^6} \bar{P} \leq \frac{SA \ln SA \ln 1/\delta}{\epsilon_2(1-\gamma)^6} \bar{P}. \]

It should be stressed that throughout the lines, \(\bar{P}\) is a shorthand for an asymptotic expression, instead of an exact value. Note that our final choice of \(\epsilon_2\) and \(\epsilon_1\) is
\[ \epsilon_2 = \frac{2^{i}}{\epsilon}, \quad \epsilon_1 = \frac{\epsilon}{16R(M+1) \ln \frac{1}{1-\gamma}}. \]

The only consideration of \(\epsilon_1\) is to meet the requirements of lemma \(\bar{P}\) throughout the proof. That is, we require
\[ 2\epsilon_1 < \frac{\epsilon_2}{16R(M+1) \left( \ln \frac{1}{1-\gamma} \right)^{-1}}. \]

It is not hard to see that \(\ln \frac{1}{\epsilon_1} = \text{poly}(\ln \frac{1}{\epsilon}, \ln \frac{1}{1-\gamma})\). This immediately implies that the number of time steps such that \((V^* - V^\pi)(s_t) > \epsilon\) is
\[ \tilde{O} \left( \frac{SA \ln 1/\delta}{\epsilon^2(1-\gamma)^6} \right), \]
where hidden factors are \(\text{poly}(\ln \frac{1}{\epsilon}, \ln \frac{1}{1-\gamma}, \ln SA)\). With probability \(1 - \delta/2\), the result of lemma \(\bar{P}\) holds; together with the \(1 - \delta/2\) probability above, we can see that this sample complexity bound holds for \(1 - \delta\) probability. 

3.3 Key Lemmas

In this subsection, we state a key lemma (Lemma 2). This lemma provides the main technical tool for controlling the sum of errors. Lemma 1 in subsection 3.2 also follows from Lemma 2 and some elementary calculations.

Definition 3. A sequence $(w_t)_{t \geq 1}$ is said to be a $(C, w)$-sequence for $C, w > 0$, if $0 \leq w_t \leq w$ for all $t \geq 1$, and $\sum_{t \geq 1} w_t \leq C$.

Lemma 2. For every $(C, w)$-sequence $(w_t)_{t \geq 1}$, with probability $1 - \delta/2$, the following holds:

$$\sum_{t \geq 1} w_t (\hat{Q}_t - Q^*)(s_t, a_t) \leq \frac{C\epsilon_1}{1 - \gamma} + O\left(\frac{\sqrt{wSAHc(C)}}{(1 - \gamma)^{2.5}} + \frac{wSA\ln C}{(1 - \gamma)^3} \ln \frac{1}{(1 - \gamma)\epsilon_1}\right).$$

where $\ell(C) = \gamma(C) \ln \frac{1}{(1 - \gamma)\epsilon_1}$ is a log-factor.

Here we give a sketch of the proof. The rigorous proof can be found in supplementary material.

Proof sketch: Let $a^*_i = a_i \prod_{j=i+1}^t (1 - \alpha_j)$, $a^0_i = I[t = 0]$. By expanding the algorithm’s update rule and applying a concentration inequality, it is not hard to show that with probability $1 - \delta/2$,

$$0 \leq (\hat{Q}_t - Q^*)(s, a) \leq \frac{\alpha^*_i}{1 - \gamma} + \beta_t + \sum_{i=1}^{t} \gamma a^*_i (\hat{V}_{t_i} - V^*)(s_{t_i+1}),$$

where $t = N_p(s, a)$, $t_i = \tau(s, a, i)$ and $\beta_t = c_3\sqrt{H_t(t)/((1 - \gamma)^2)}$. Note that to use a union bound over infinite number of events, we partition probability non-uniformly among timesteps. That is, we assign probability $\delta/\{SA(k+1)(k+2)\}$ for the concentration when a state-action pair is visited the $k$-th time. By doing so, these probabilities sum to $\delta/2$.

Next, we consider the weighted sum $\sum_{t \geq 1} w_t (\hat{Q}_t - Q^*)(s_t, a_t)$ by expanding it with the inequality above. Let $n_t = N_t(s_t, a_t)$ for simplicity. The result will be

$$\sum_{t \geq 1} w_t (\hat{Q}_t - Q^*)(s_t, a_t) \leq \sum_{t \geq 1} w_t \frac{\alpha^*_i}{1 - \gamma} + \sum_{t \geq 1} w_t \beta_{n_t} + \gamma \sum_{t \geq 1} w_t \sum_{i=1}^{n_t} \alpha^*_i (\hat{V}_{\tau(s_t, a_t, i)} - V^*)(s_{\tau(s_t, a_t, i)+1}).$$

The result has three parts. The first part is easily controlled by $\frac{SA}{1 - \gamma}$. The second and trickiest part consists of a weighted sum of $\beta_i$s. The key observation is that $\beta_i$ is convex and decreasing with respect to $t$. Using the rearrangement inequality and Jensen’s inequality, we can show that the second term is $O\left(\left((1 - \gamma)^{-1}\sqrt{wSAHc(C)}\right)\right)$, where $w$ is the supreme of $\{w_t\}$. The $\ln C$ term in the result comes from technical reasons and is dominated by $\sqrt{C}$ term when $C$ is large enough.

The third part is $\gamma$ multiplied to another weighted sum with total weight no larger than $C$. By carefully chosen learning rate $\alpha$, the supreme of $\{w_t\}$ (cf. Section 4 of [4]) will only be magnified by a factor of $1 + 1/H$. That is, the term is a weighted sum whose weights are a $(C, (1 + 1/H)w)$-sequence. In fact, it can be shown that

$$\sum_{t \geq 1} w_t (\hat{Q}_t - Q^*)(s_t, a_t) \leq c_3\frac{wSAHc(C)}{1 - \gamma} + O\left(\frac{wSAH}{1 - \gamma} \ln C\right) + \gamma \sum_{t \geq 1} w_{t+1}' (\hat{Q}_{t+1} - Q^*)(s_{t+1}, a_{t+1}),$$

where $w_{t+1}'$ is the $(C, (1 + 1/H)w)$-sequence mentioned above.

Note that the last term of right hand size of this inequality has the same form of the left hand size. Therefore we can repeat this unrolling process $R = H$ times until the remaining weighted
summation is bound by $O(\epsilon)$ due to discount. Putting things together will result in the bound above.

Finally, We explain how to prove Lemma 1 with Lemma 2 (Full proof can be found in supplementary materials.) Note that since $\hat{Q}_t \geq Q^*$,

$$V^*(s_t) - Q^*(s_t, a_t) \leq \hat{Q}_t(s_t, a_t) - Q^*(s_t, a_t).$$

We now consider a set $J = \{ t : V^*(s_t) - Q^*(s_t, a_t) > \eta(1 - \gamma)^{-1} \}$, and consider the $(|J|, 1)$-weight sequence defined by $w_t = I[t \in J]$. We can now apply this lemma to weighted sum $\sum_{t \geq 1} w_t [V^*(s_t) - Q^*(s_t, a_t)]$. On the one hand, this quantity is obviously at least $|J|\eta(1 - \gamma)^{-1}$.

On the other hand, by lemma 2, it is upper bounded by the weighted sum of $(\hat{Q} - Q^*)(s_t, a_t)$, which is in turn bounded by lemma 2. Thus we get

$$|J|\eta(1 - \gamma)^{-1} \leq \frac{C \epsilon_1}{1 - \gamma} + O \left( \frac{\sqrt{SA|J|\ell(|J|)}}{(1 - \gamma)^{2.5}} + \frac{wSA\ln|J|}{(1 - \gamma)^3} \ln \frac{1}{(1 - \gamma)\epsilon_1} \right).$$

Now focus on the dependence on $|J|$. The left-hand-side has linear dependence on $|J|$, whereas the left-hand-side has a $\tilde{O}\left(\sqrt{|J|}\right)$ dependence. This allows us to solve out an upper bound on $|J|$.

4 Discussion

In this section, we discuss the implication of our results, and present some interesting properties of our algorithm beyond its sample complexity bound.

4.1 Comparison with other results

Previously, the best sample complexity bound for a model-free algorithm is $\tilde{O}\left(\frac{SA}{c^4(1 - \gamma)^{2.5}}\right)$ (hiding all logarithmic terms). To the best of our knowledge, the current best minimax lower bound for sample complexity is

$$\Omega\left(\frac{SA}{c^2(1 - \gamma)^2} \ln \frac{1}{\delta}\right)$$

due to [8]. There was a quadratic gap between this lower bound and Delayed Q-learning’s result in the dependence on $1/\epsilon$, which our result closes. The gap between our results and this lower bound lies only in the dependence on $1/(1 - \gamma)$ and logarithmic terms of $SA$, $1/1 - \gamma$ and $1/\epsilon$.

In model-based algorithms, better sample complexity results in infinite horizon settings have been shown [15, 8]. To the best of our knowledge, the best published result without further restrictions on MDPs is $\tilde{O}\left(\frac{SA}{c^2(1 - \gamma)^2}\right)$ due to [15], which is $(1 - \gamma)$ smaller than our upper bound. From a practical point of view, there is a clear distinction between model-based and model-free approaches. Our claim that we improved the best model-free result is based on such a rough classification. In this sense, we can also claim that, for infinite-horizon reinforcement learning, model-free approach can be nearly as sample efficient as the best model-based ones.

If we take a theoretical point of view, however, until this date there is no clear and definite classification criterion between model-free and model-based algorithm. One candidate criterion is based on space complexity [13]. In this sense, our algorithm is indeed much more memory-efficient. Our algorithm stores $O(SA)$ values, whereas the algorithms of [15] needs $\Omega(S^2A)$ memory even to store the transition model.
4.2 Monotonicity

In our algorithm, the state-action value based on which the algorithm acts is $\hat{Q}$. For any $(s, a)$, $\hat{Q}(s, a)$ has the interesting property of being decreasing over time. Since we know that $\hat{Q}(s, a) \geq Q^*(s, a)$, this means that the $Q$-error $\left(\hat{Q} - Q^*\right)(s, a)$ also decreases over $t$. Although this property is not used in our current proof, it comes at no cost.

4.3 Application to finite horizon tasks

In a finite horizon setting, it is shown that when ignoring the dependence on $H, S, A$, an $\tilde{O}(T^\alpha)$ regret bound can be translated into an $\tilde{O}(\epsilon^{-1/\alpha})$ PAC sample complexity bound. On the other hand, an $\tilde{O}(\epsilon^{-\beta})$ PAC sample complexity bound can be translated into an $\tilde{O}(T^\beta/(1+\beta))$ regret bound [6]. For example, the algorithm with $\tilde{O}\left(\sqrt{T}\right)$ regret has PAC sample complexity $\tilde{O}(\epsilon^{-2})$. And the algorithm with $\tilde{O}(\epsilon^{-2})$ PAC sample complexity has a regret in the order of $\tilde{O}(T^{2/3})$.

Note that the PAC sample complexity defined for finite horizon MDP is different from that defined for infinite horizon MDP. However by applying the analysis of our algorithm on the sample complexity of exploration, it can be shown that the algorithm has a regret bound $\tilde{O}(T^{1/2})$ and a PAC sample complexity $\tilde{O}(\epsilon^{-2})$ when running on finite horizon MDPs.

First we define a mapping from a finite horizon MDP to an infinite horizon MDP so that our algorithm can be applied. For an arbitrary finite horizon MDP $\mathcal{M} = (S, \bar{A}, H, r_{1h}(s, a), p_h(s' | s, a))$ where $H$ is the length of episode, the corresponding infinite horizon MDP $\bar{\mathcal{M}} = (\bar{S}, \bar{A}, \bar{\gamma}, \bar{\nu}(\bar{s}, a), \bar{p}(\bar{s}' | \bar{s}, a))$ is defined as,

- $\bar{S} = S \times H, \bar{A} = A$;
- $\bar{\gamma} = (1 - 1/H)$;
- for a state $s$ at step $h$, let $\bar{s}_{s,h}$ be the corresponding state. For any action $a$ and next state $s'$, define $\bar{\nu}(\bar{s}_{s,h}, a) = \gamma^{H-h+1}r_{1h}(s, a)$ and $\bar{p}(\bar{s}'_{s,h+1} | \bar{s}_{s,h}, a) = p_h(s, h)$. And for $h = H$, set $\bar{\nu}(\bar{s}_{s,h}, a) = 0$ and $\bar{p}(\bar{s}'_{s,1} | \bar{s}_{s,h}, a) = I[s' = s_1]$ for a fixed starting state $s_1$.

Let $V_1^*$ be the value function in $\bar{\mathcal{M}}$ at time $t$ and $V_1^k$ the value function in $\mathcal{M}$ at episode $k$, step $h$.

It follows that $V_1^*(\bar{s}_{s,1,1}) = \gamma^{H-1}V_1^*(s_1)$. And the policy mapping is defined as $\pi_h(s) = \bar{\pi}(\bar{s}_{s,h})$ for policy $\bar{\pi}$ in $\bar{\mathcal{M}}$. Value functions in MDP $\mathcal{M}$ and $\bar{\mathcal{M}}$ are closely related in a sense that, $V_1^*(\bar{s}_{s,1,1}) = \gamma^{H-1}V_1^*(s_1)$, and any $\epsilon$-optimal policy $\bar{\pi}$ of $\bar{\mathcal{M}}$ corresponding to an $(\epsilon/\gamma^H)$-optimal policy $\pi$ in $\mathcal{M}$ (see supplementary material for proof). Note that here $\gamma^H = (1 - 1/H)^H = O(1)$ is an constant.

For any $\epsilon > 0$, by running our algorithm on $\bar{\mathcal{M}}$ for $\tilde{O}(\tfrac{3SH^2}{\epsilon^2})$ time steps, the starting state $s_1$ is visited at least $\tilde{O}(\tfrac{3SAH^2}{\epsilon^2})$ times, and at most 1/3 of them are not $\epsilon$-optimal. If we select the policy uniformly randomly from the policy $\pi^{tH+1}$ for $0 \leq t < T/H$, with probability at least 2/3 we can get an $\epsilon$-optimal policy. Therefore the PAC sample complexity is $\tilde{O}(\epsilon^{-2})$ after hiding $S, A, H$ terms.

On the other hand, we want to show that for any $K$ episodes,

$$\text{Regret}(T) = \sum_{k=1}^{T/H} \left[ V_1^*(s_1) - V_1^k(s_1) \right] \propto T^{1/2}. $$

The reason why our algorithm can have a better reduction from regret to PAC is that, after choosing $\epsilon_1$, it follows from the argument of theorem [1] that for all $\epsilon_2 > \tilde{O}(\epsilon_1/(1 - \gamma))$, the number
of $\epsilon_2$-suboptimal steps is bounded by
\[
O \left( \frac{SA \ln S \ln 1/\delta}{\epsilon_2^2(1-\gamma)^2} \text{polylog} \left( \frac{1}{\epsilon_1}, \frac{1}{1-\gamma} \right) \right)
\]
with probability $1 - \delta$. In contrast, delayed Q-learning [13] can only give an upper bound on $\epsilon_1$-suboptimal steps after setting parameter $\epsilon_1$. In other words, after setting $\epsilon_1$, we can give a $O\left(\epsilon_2^{-2}\right)$ uniform upper bound for the curve of the number of $\epsilon_2$-suboptimal steps versus $\epsilon_2$, as long as $\epsilon_2 > \tilde{O}(\epsilon_1/(1-\gamma))$.

Formally, let $X_k = V^*(s_1) - V_k^k(s_1)$ be the regret of $k$-th episode. For any $T$, set $\epsilon = \sqrt{SA/T}$ and $\epsilon_2 = \tilde{O}(\epsilon_1/(1-\gamma))$. Let $M = \lceil \log_2 \frac{1}{\epsilon_2(1-\gamma)} \rceil$. It follows that,
\[
\text{Regret}(T) \leq T\epsilon_2 + \sum_{i=1}^{M} \left( |k : \{X_k \geq \epsilon_2 \cdot 2^{i-1}\} | \right) \epsilon_2 \cdot 2^i
\]
\[
\leq \tilde{O} \left( T\epsilon_2 + \sum_{i=1}^{M} \frac{SA \ln 1/\delta}{\epsilon_2 \cdot 2^{i-2}} \right)
\]
\[
\leq \tilde{O} \left( \sqrt{SAT \ln 1/\delta} \right)
\]
with probability $1 - M\delta$. Note that the $\tilde{O}$ notation hides the polylog $(1/\epsilon_1, 1/(1-\gamma))$ which is, by our reduction, polylog $(H, T, S, A)$.

### 4.4 Future Work

There is still a quartic gap in the dependence on $1/(1-\gamma)$ between our result and the best lower bound. Future work may close this gap, either through refined analysis or through more sophisticated algorithms.

### 5 Conclusion

Infinite-horizon MDP with discounted reward is a setting that is arguably more difficult than other popular settings, such as finite-horizon MDP. Previously, the best sample complexity bound achieved by model-free reinforcement learning algorithms in this setting is $\tilde{O}\left(\frac{SA}{\epsilon^2(1-\gamma)^2}\right)$, due to Delayed Q-learning [13]. In this paper, we propose a variant of Q-learning that incorporates upper confidence bound, and show that it has a sample complexity of $\tilde{O}\left(\frac{SA}{\epsilon^2(1-\gamma)^2}\right)$. This matches the best lower bound except in dependence on $1/(1-\gamma)$ and logarithmic factors.

### 6 Acknowledgments

We thank Chongjie Zhang for helpful discussions.
References

[1] Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Speedy q-learning. In Advances in neural information processing systems, 2011.

[2] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. arXiv preprint arXiv:1703.05449, 2017.

[3] Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. J. Mach. Learn. Res., 3:213–231, March 2003.

[4] Eyal Even-Dar and Yishay Mansour. Learning rates for q-learning. Journal of Machine Learning Research, 5(Dec):1–25, 2003.

[5] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(Apr):1563–1600, 2010.

[6] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In Advances in Neural Information Processing Systems, pages 4864–4874, 2018.

[7] Sham Machandranath Kakade et al. On the sample complexity of reinforcement learning. PhD thesis, University of London London, England, 2003.

[8] Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In International Conference on Algorithmic Learning Theory, pages 320–334. Springer, 2012.

[9] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, pages 1928–1937, 2016.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[11] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International Conference on Machine Learning, pages 1889–1897, 2015.

[12] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 770–787. Society for Industrial and Applied Mathematics, 2018.

[13] Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In Proceedings of the 23rd international conference on Machine learning, pages 881–888. ACM, 2006.

[14] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. Journal of Computer and System Sciences, 74(8):1309–1331, 2008.

[15] István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 1031–1038, 2010.
A Appendix

A.1 Proof of Lemma 1

Proof. Let $I = \{i : V^*(s_i) - Q^*(s_i, a_i) > \frac{\eta i}{1 - \gamma}\}$. By lemma 2 with probability $1 - \delta$,

$$\frac{\eta |I|}{1 - \gamma} \leq \sum_{i \in I} (V^*(s_i) - Q^*(s_i, a_i)) \leq \sum_{i \in I} \left[ (\hat{Q}_i - Q^*) (s_i, a_i) \right]$$

$$\leq \frac{|I|\epsilon_1}{1 - \gamma} + O \left( \frac{1}{(1 - \gamma)^{5/2}} \sqrt{SA|I|\epsilon(|I|)} + \frac{SA}{(1 - \gamma)^{3}} \ln |I| \ln \frac{1}{\epsilon_1(1 - \gamma)} \right)$$

$$\leq \frac{|I|\epsilon_1}{1 - \gamma} + O \left( \ln \frac{1}{\epsilon_1(1 - \gamma)} \left( \sqrt{SA|I|\ln (SA|I|)} \right) \frac{SA\ln |I|}{(1 - \gamma)^{5/2}} + \frac{SA\ln |I|}{(1 - \gamma)^{3}} \right)$$

Suppose that $|I| = \frac{SAk^2}{2(\eta - \epsilon_1)} \ln SA$, for some $k > 1$. Then it follows that for some constant $C_1$,

$$\frac{\eta |I|}{1 - \gamma} = \frac{k^2SA\ln SA}{(1 - \gamma)^{4\eta}} \leq \frac{2(\eta - \epsilon_1)|I|}{1 - \gamma}$$

$$\leq C_1 \sqrt{\ln \frac{1}{\delta}} \ln \frac{1}{\epsilon_1(1 - \gamma)} \left( \sqrt{SA|I|\ln (SA|I|)} \right) \frac{SA\ln |I|}{(1 - \gamma)^{5/2}} + \frac{SA\ln |I|}{(1 - \gamma)^{3}}$$

Therefore

$$k^2 \ln (SA) \leq C_1 \sqrt{\ln \frac{1}{\delta}} \ln \frac{1}{\epsilon_1(1 - \gamma)} \left( k \ln (SA + \ln |I|) \right) \left( \eta(1 - \gamma) \ln |I| \right)$$

$$\leq kC_1 \sqrt{\ln \frac{1}{\delta}} \ln \frac{1}{\epsilon_1(1 - \gamma)} \cdot \ln (SA + 2 \ln |I|)$$

$$\leq kC_1 \sqrt{\ln \frac{1}{\delta}} \ln \frac{1}{\epsilon_1(1 - \gamma)} \cdot \left( 3 \ln SA + 4 \ln k + 6 \ln \frac{1}{\eta(1 - \gamma)} \right)$$

$$\leq 6kC_1 \sqrt{\ln \frac{1}{\delta}} \ln \frac{1}{\epsilon_1(1 - \gamma)} \cdot \ln (SA + \ln ek).$$

Let $C' = 6C_1 \sqrt{\ln \frac{1}{\delta}} \ln \frac{1}{\epsilon_1(1 - \gamma)}$.

$$k \leq C'(2 + \ln k).$$

(7)

If $k \geq 10C' \ln C'$, then

$$k - C' \left( 2 + \ln \frac{k}{\delta} \right) \geq 8C' \ln C' - (2 + \ln 10)C' \geq 4C' (2 \ln C' - 4) \geq 0,$$
which means violation of (7). (We assume $C' \geq 2$.) Therefore

$$k \leq 10C' \ln C' \leq 360C_1^2 \ln^4 \frac{1}{\epsilon_1 (1 - \gamma)}.$$  

(8)

It immediately follows that

$$|I| = \frac{SAk^2}{\eta^2 (1 - \gamma)^3} \ln SA$$

$$\leq \frac{SA \ln SA}{\eta^2 (1 - \gamma)^5} \cdot \ln \frac{1}{\delta} \cdot O \left( \ln^8 \frac{1}{\epsilon_1 (1 - \gamma)} \right).$$  

(9)

A.2 Proof of Lemma 2

Fact 1. (1) The following statement holds throughout the algorithm,

$$\hat{Q}_{p+1}(s, a) \leq Q_{p+1}(s, a).$$

(2) If at time $p$, $\hat{Q}(s, a)$ is updated at line 10, then after this update is finished,

$$\hat{Q}_{p+1}(s, a) \geq Q_{p+1}(s, a).$$

Proof. Simply refer to the algorithm.

Before proving lemma 2, we will prove two auxiliary lemmas.

Lemma 3. The following properties hold for $\alpha_i^t$:

1. $\sqrt{\frac{1}{t}} \leq \sum_{i=1}^{t} \alpha_i^t \sqrt{\frac{1}{t}} \leq 2 \sqrt{\frac{1}{t}}$ for every $t \geq 1, c > 0$.

2. $\max_{i \in [t]} \alpha_i^t \leq \frac{2H}{t}$ and $\sum_{i=1}^{t} (\alpha_i^t)^2 \leq \frac{2H}{t}$ for every $t \geq 1$.

3. $\sum_{t=1}^{\infty} \alpha_i^t = 1 + 1/H$, for every $i \geq 1$.

4. $\sqrt{\frac{\ell(t)}{t}} \leq \sum_{i=1}^{t} \alpha_i^t \sqrt{\frac{\ell(i)}{t}} \leq 2 \sqrt{\frac{\ell(t)}{t}}$ where $\ell(t) = \ln(c(t+1)(t+2))$, for every $t \geq 1, c \geq 1$.

Proof. Recall that

$$\alpha_t = \frac{H + 1}{H + t}, \quad \alpha_0^t = \prod_{j=1}^{t} (1 - \alpha_j), \quad \alpha_i^t = \alpha_t \prod_{j=i+1}^{t} (1 - \alpha_j).$$

Properties 1-3 are proven by [6]. Now we prove the last property.

On the one hand,

$$\sum_{i=1}^{t} \alpha_i^t \sqrt{\frac{\ell(i)}{i}} \leq \sum_{i=1}^{t} \alpha_i^t \sqrt{\frac{\ell(t)}{i}} \leq 2 \sqrt{\frac{\ell(t)}{t}},$$

where the last inequality follows from property 3.

The left-hand side is proven by induction on $t$. For the base case, when $t = 1$, $\alpha_i^t = 1$. For $t \geq 2$, we have $\alpha_i^t = (1 - \alpha_t)\alpha_{i-1}^t$ for $1 \leq i \leq t - 1$. It follows that

$$\sum_{i=1}^{t} \alpha_i^t \sqrt{\frac{\ell(i)}{i}} = \alpha_t \sqrt{\frac{\ell(t)}{t}} + (1 - \alpha_t) \sum_{i=1}^{t-1} \alpha_i^{t-1} \sqrt{\frac{\ell(i)}{i}} \geq \alpha_t \sqrt{\frac{\ell(t)}{t}} + (1 - \alpha_t) \sqrt{\frac{\ell(t-1)}{t-1}}.$$
Since function $f(t) = \frac{u(t)}{t}$ is monotonically decreasing for $t \geq 1, c \geq 1$, we have
\[
\alpha_t \sqrt{\frac{u(t)}{t}} + (1 - \alpha_t) \sqrt{\frac{u(t - 1)}{t - 1}} \geq \alpha_t \sqrt{\frac{u(t)}{t}} + (1 - \alpha_t) \sqrt{\frac{u(t)}{t}} \geq \sqrt{\frac{u(t)}{t}}.
\]

\[\square\]

**Lemma 4.** With probability at least $1 - \delta/2$, for all $p \geq 0$ and $(s, a)$-pair,
\[
0 \leq (Q_p - Q^*)(s, a) \leq \frac{\alpha_0^p}{1 - \gamma} + \sum_{i=1}^{t} \gamma \alpha_i^p \beta_i (\hat{V}_i - V^*)(s_{i+1}) + \beta_t,
\]
\[
0 \leq (\hat{Q}_p - Q^*)(s, a),
\]
where $t = N_p(s, a), t_i = \tau(s, a, i)$ and $\beta_t = c_3\sqrt{H_\tau(t)/((1 - \gamma)^2)}$.

**Proof.** Recall that
\[
\alpha_0^p = \prod_{j=1}^{t} (1 - \alpha_j), \quad \alpha_i^p = \alpha_i \prod_{j=i+1}^{t} (1 - \alpha_j).
\]

It is not hard to see that our algorithm maintains the following $Q(s, a)$:
\[
Q_p(s, a) = \alpha_0^p \frac{1}{1 - \gamma} + \sum_{i=1}^{t} \alpha_i^p \left[ r(s, a) + b_i + \gamma \hat{V}_i(s_{i+1}) \right].
\]
Bellman optimality equation gives:
\[
Q^*(s, a) = r(s, a) + \gamma \mathbb{P}V^*(s, a) = \alpha_0^p Q^*(s, a) + \sum_{i=1}^{t} \alpha_i^p \left[ r(s, a) + \gamma \mathbb{P}V^*(s, a) \right].
\]

Subtracting the two equations:
\[
(Q_p - Q^*)(s, a) = \alpha_0^p (\frac{1}{1 - \gamma} - Q^*(s, a)) + \sum_{i=1}^{t} \alpha_i^p \left[ b_i + \gamma (V_i - V^*) (s_{i+1}) + \gamma (V^*(s_{i+1}) - \mathbb{P}V^*(s, a)) \right].
\]

The identity above holds for arbitrary $p, s$ and $a$. Now fix $s \in S, a \in A$ and $p \in \mathbb{N}$. Let $t = N_p(s, a), t_i = \tau(s, a, i)$. The $t = 0$ case is trivial; we assume $t \geq 1$ below. Now consider an arbitrary fixed $k$. Define
\[
\Delta_i = \left( \alpha_k^i \cdot I[t_i < \infty] \cdot \left( \mathbb{P}V^* - \hat{\mathbb{P}}_t V^* \right)(s, a) \right)
\]
Let $F_i$ be the $\sigma$-Field generated by random transitions from step 1 to $t_i$. Clearly $\mathbb{E}[\Delta_i|F_i] = 0$, while $\Delta_i$ is measurable in $F_{i+1}$. Also, clearly $|\Delta_i| \leq \frac{2}{1 - \gamma}$. Therefore, $\Delta_i$ is a martingale difference sequence; by the Azuma-Hoeffding inequality,
\[
\Pr \left[ \left| \sum_{i=1}^{k} \Delta_i \right| > \eta \right] \leq 2 \exp \left\{ - \frac{\eta^2}{8 (1 - \gamma)^2 \sum_{i=1}^{k} (\alpha_k^i)^2} \right\}.
\]

By choosing $\eta$, we can show that with probability $\delta/\lfloor SA(k+1)(k+2)\rfloor$
\[
\left| \sum_{i=1}^{k} \Delta_i \right| \leq \frac{c_1}{1 - \gamma} \cdot \sqrt{\sum_{i=1}^{k} (\alpha_k^i)^2} \cdot \ln \frac{2(k+1)(k+2)SA}{\delta} \leq \frac{c_2}{1 - \gamma} \sqrt{\frac{H_\tau(k)}{k}}.
\]
Here $c_1 = 2\sqrt{2}$ and $c_2 = 8$ will do. $\tau(k) = \ln \frac{(k+1)(k+2)SA}{\delta}$. By union bound, this holds for arbitrary $k$ simultaneously with probability $1 - \delta/(2SA)$; it holds for arbitrary $s', a'$ with probability $1 - \delta/2$. Therefore it holds for the random $t = N(p(s, a))$ for that probability as well ($p$ can be arbitrary).

**Proof of the right hand side of (11):** We also know that $(b_k = \frac{c_2}{1 - \gamma} \sqrt{\frac{H_t}{k}})$

\[
\frac{c_2}{1 - \gamma} \sqrt{\frac{H_t}{k}} \leq \sum_{i=1}^{k} \alpha^i b_i \leq \frac{2c_2}{1 - \gamma} \sqrt{\frac{H_t}{k}}.
\]

Therefore,

\[
(Q_p - Q^*)(s, a) \leq \gamma \left| \sum_{i=1}^{t} \Delta_i \right| + \sum_{i=1}^{t} \alpha^i \left[ \gamma(\hat{V}_t - V^*)(x_{t+1}) + b_i \right]
\]

\[
\leq \frac{3c_2}{1 - \gamma} \sqrt{\frac{H_t}{t}} + \sum_{i=1}^{t} \gamma \alpha^i (\hat{V}_t - V^*)(x_{t+1})
\]

\[
\leq \frac{\alpha_i^0}{1 - \gamma} + \sum_{i=1}^{t} \gamma \alpha^i (\hat{V}_t - V^*)(x_{t+1}) + \beta_t.
\]

Note that $\beta_t = c_3(1 - \gamma)^{-1} \sqrt{H_t/t}$; $c_3 = 3c_2$ will be enough.

**Proof of the left hand side of (11):** Now, we state a proposition that $Q_p \geq Q^*$ for all $(s, a)$ and $p \leq p'$. This proposition is obviously true when $p' = 0$. Assume we live in the $1 - \delta/2$ probability. Then

\[
(Q_p - Q^*)(s, a) \geq -\gamma \left| \sum_{i=1}^{t} \Delta_i \right| + \sum_{i=1}^{t} \alpha^i \left[ \gamma(\hat{V}_t - V^*)(x_{t+1}) + b_i \right]
\]

\[
\geq \sum_{i=1}^{t} \alpha^i b_i - \gamma \sum_{i=1}^{t} \Delta_i \geq 0.
\]

Therefore the proposition holds for $p' + 1$ as well. By induction, it holds for all $p$. We now see that (11) holds for probability $1 - \delta/2$ for all $p, s, a$. Since $Q_p(s, a)$ is always greater than $Q_{p'}(s, a)$ for some $p' \leq p$, we know that $\hat{Q}_p(s, a) \geq Q_{p'}(s, a) \geq Q^*(s, a)$, thus proving (12).

We now give a proof for lemma 2. Recall the definition for a $(C, w)$-sequence. A sequence $(w_t)_{t \geq 1}$ is said to be a $(C, w)$-sequence for $C, w > 0$, if $0 \leq w_t \leq w$ for all $t \geq 1$, and $\sum_{t \geq 1} w_t \leq C$.

**Proof.** Let $n_t = N_t(s_t, a_t)$ for simplicity, we have

\[
\sum_{t \geq 1} w_t \left( \hat{Q}_t - Q^* \right)(s_t, a_t)
\]

\[
\leq \sum_{t \geq 1} w_t \left( e_1 + (Q_t - Q^*)(s_t, a_t) \right)
\]

\[
\leq \sum_{t \geq 1} w_t (Q_t - Q^*)(s_t, a_t)
\]

\[
\leq \sum_{t \geq 1} w_t \left[ \frac{\alpha^0}{1 - \gamma} + \beta n_t + \gamma \sum_{i=1}^{n_t} \alpha^i (\hat{V}_{t(s_t, a_t, i)} - V^*) (s_{t(s_t, a_t, i)+1}) \right]
\]
The last inequality is due to lemma 4. Note that $\alpha^0_{n_t} = \mathbb{I}[n_t = 0]$; the first term in the summation can be bounded by,

$$\sum_{t \geq 1} w_t \frac{\alpha^0_{n_t}}{1 - \gamma} \leq \frac{SAw}{1 - \gamma}. \quad (14)$$

For the second term, define $u(s, a) = \sup_t N_t(s, a)$.

It follows that,

$$\sum_{t \geq 1} w_t \beta_{n_t} = \sum_{s,a} u(s,a) \sum_{i=0} w_{\tau(s,a,i)} \beta_i \leq \sum_{s,a} (1 - \gamma)^{-1} c_3 w \sqrt{\nu(C)HC_{s,a}/w} \leq c_3 (1 - \gamma)^{-1} wSAHC_{l(C)}. \quad (15)$$

Where $C_{s,a} = \sum_{t \geq 1, (s_t,a_t) = (s,a)} w_t$. Inequality (14) comes from rearrangement inequality, since $\nu(x)/x$ is monotonically decreasing. And inequality (15) comes from Jensen’s inequality.

For the third term of the summation, we have

$$\sum_{t \geq 1} w_t \sum_{i=1}^{n_t} \alpha^i_t \left( \hat{V}_{\tau(s_t,a_t,i)} - V_s^* \right) \left( s_{\tau(s_t,a_t,i)+1} \right) \leq \sum_{t' \geq 1} \left( \hat{V}_{t'} - V_s^* \right) \left( s_{t'+1} \right) \left( \sum_{t > t'}^{\infty} \frac{n_{t'}}{\alpha_{n_t}} w_t \right).$$

Define

$$w'_{t+1} = \left( \sum_{t > t'}^{\infty} \frac{n_{t'}}{\alpha_{n_t}} w_t \right).$$

We claim that $w'_t$ is a $(C, (1 + \frac{1}{H})w)$-sequence. We now prove this claim. By lemma 3, for any $t' \geq 0$,

$$w'_{t'+1} \leq w \sum_{j=n_{t'}}^{\infty} \alpha_{n_{t'}} = (1 + 1/H)w.$$

---

$^2$ $u(s, a)$ could be infinity when $(s, a)$ is visited for infinite number of times.
And by \( \sum_{j=0}^{i} \alpha_i^j = 1 \), we have \( \sum_{t' \geq 1} w_{t'+1} \leq \sum_{t \geq 1} w_t \leq C \). It follows that

\[
\sum_{t \geq 1} w_{t+1} \left( \tilde{V}_t - V^* \right) (s_{t+1}) = \sum_{t \geq 1} w_{t+1} \left( \tilde{V}_t - V^* \right) (s_{t+1}) + \sum_{t \geq 1} w_{t+1} \left( \tilde{V}_t - \tilde{V}_{t+1} \right) (s_{t+1}) 
\]

(16)

\[
\leq \sum_{t \geq 1} w_{t+1} \left( \tilde{V}_t - V^* \right) (s_{t+1}) + \sum_{t \geq 1} w_{t+1} \left( 2\alpha_{nt} \frac{1}{1 - \gamma} \right) 
\]

(17)

\[
\leq \sum_{t \geq 1} w_{t+1} \left( \tilde{V}_t - V^* \right) (s_{t+1}) + \mathcal{O} \left( \frac{wSAH}{1 - \gamma} \ln C \right) 
\]

(18)

\[
\leq \sum_{t \geq 1} w_{t+1} \left( \tilde{Q}_{t+1} - Q^* \right) (s_{t+1}, a_{t+1}) \geq \mathcal{O} \left( \frac{wSAH}{1 - \gamma} \ln C \right) 
\]

(19)

\[
\leq \sum_{t \geq 1} w_{t+1} \left( \tilde{Q}_{t+1} - Q^* \right) (s_{t+1}, a_{t+1}) + \mathcal{O} \left( \frac{wSAH}{1 - \gamma} \ln C \right) 
\]

(20)

Inequality \[(18)\] comes from the update rule of our algorithm. Inequality \[(19)\] comes from the fact that \( \alpha_t = (H + 1)/(H + t) \leq H/t \) and Jensen’s Inequality.

Putting all terms together, we have,

\[
\sum_{t \geq 1} w_t (\tilde{Q}_t - Q^*)(s_t, a_t) \leq c_3 \frac{\sqrt{wSAH C_t(C)}}{1 - \gamma} + \mathcal{O} \left( \frac{wSAH}{1 - \gamma} \ln C \right) + \gamma \sum_{t \geq 1} w_{t+1} \left( \tilde{Q}_{t+1} - Q^* \right) (s_{t+1}, a_{t+1}). 
\]

(21)

Observe that the forth term is another weighted sum with the same form. Then we can recursively unroll the summation term for index set \( I' \) and so on. Suppose that our original weight sequence is also denoted by \( \{w_t^{(0)}\}_{t \geq 1} \), while \( \{w_t^{(k)}\}_{t \geq 1} \) denotes the weight sequence after unrolling for \( k \) times. Let \( w^{(k)} \) be \( w \cdot (1 + 1/H)^k \). Then we can see that \( \{w_t^{(k)}\}_{t \geq 1} \) is a \((C, w^{(k)})\)-sequence. Suppose that we unroll for \( R \) times. Then

\[
\sum_{t \geq 1} w_t (\tilde{Q}_t - Q^*)(s_t, a_t) \leq c_3 \frac{\sqrt{w^{(R)}SAH C_t(C)}}{(1 - \gamma)^2} + \mathcal{O} \left( \frac{w^{(R)}SAH}{(1 - \gamma)^2} \ln C \right) + \gamma R \sum_{t \geq 1} w_{t+1}^{(R)} \left( \tilde{Q}_{t+1} - Q^* \right) (s_t, a_t) 
\]

\[
\leq c_3 \frac{\sqrt{w^{(R)}SAH C_t(C)}}{(1 - \gamma)^2} + \mathcal{O} \left( \frac{w^{(R)}SAH}{(1 - \gamma)^2} \ln C \right) + \gamma R \frac{C}{1 - \gamma}. 
\]

We set \( H = R = \frac{\ln 1/(1-\gamma)\epsilon_1}{\ln 1/\gamma} \leq \frac{\ln 1/(1-\gamma)\epsilon_1}{1-\gamma} \). It follows that \( w^{(R)} = (1 + 1/H)^R w^{(0)} \leq \epsilon w^{(0)} \), and that \( \gamma R \frac{C}{1 - \gamma} \leq C \epsilon_1 \). Therefore,

\[
\sum_{t \geq 1} w_t (\tilde{Q}_t - Q^*)(s_t, a_t) \leq \frac{C \epsilon_1}{1 - \gamma} + \mathcal{O} \left( \frac{\sqrt{wSAH C_t(C)}}{(1 - \gamma)^{2.5}} + \frac{wSA}{(1 - \gamma)^3} \ln C \ln \frac{1}{(1 - \gamma)\epsilon_1} \right). 
\]

(23)
A.3 MDP mapping

Recall that our MDP mapping from $\mathcal{M} = (S, A, H, r_h(s, a), p_h(s' | s, a))$ to $\hat{\mathcal{M}} = (\bar{S}, \bar{\bar{A}}, \bar{\bar{\gamma}}, \bar{\bar{\bar{r}}}(\bar{s}, \bar{a}), \bar{\bar{p}}(s' | \bar{s}, \bar{a}))$ is defined as,

- $\bar{S} = S \times H, \bar{\bar{A}} = A$;
- $\bar{\gamma} = (1 - 1/H);$  
- for a state $s$ at step $h$, let $\bar{s}_{s,h}$ be the corresponding state. For any action $a$ and next state $s'$, define $\bar{r}(\bar{s}_{s,h}, a) = \gamma^{H-h+1}r_h(s, a)$ and $\bar{\bar{p}}(\bar{s}_{s,h}, a) = p_h(s, h)$. And for $h = H$, set $\bar{r}(\bar{s}_{s,h}, a) = 0$ and $\bar{\bar{p}}(\bar{s}_{s,h}, a) = I[s' = s_1]$ for a fixed starting state $s_1$.

For a trajectory $\{(\bar{s}_{s,1}, \bar{a}_1), (\bar{s}_{s,2}, \bar{a}_2), \cdots \}$ in $\hat{\mathcal{M}}$, let $\{(s_1, a_1), (s_2, a_2), \cdots \}$ be the corresponding trajectory in $\mathcal{M}$. Note that $\mathcal{M}$ has a unique fixed starting state $s_1$, which means that $s_{1,H+1} = s_1$ for all $t \geq 0$. Denote the corresponding policy of $\bar{\pi}^t$ as $\pi^t$ (may be non-stationary), then we have

$$
\bar{V}^{\pi^t}(\bar{s}_{s,1}) = \mathbb{E} [\bar{r}(\bar{s}_{s,1}, \bar{a}_1) + \gamma^{H}r(\bar{s}_{s,2}, \bar{a}_2) + \cdots + \gamma^{H}r(\bar{s}_{s,H-1}, \bar{a}_{H-1}) + \gamma^{H}V^{\pi_{t+H-1}}(\bar{s}_{s,H+1,1})]
$$

$$
= \gamma^{H}\mathbb{E} [r_1(s_1, a_1) + r_2(s_2, a_2) + \cdots + r_H(s_{H-1}, a_{H-1}) + V^{\pi_{t+H}}(\bar{s}_{s,H+1,1})]
$$

$$
= \gamma^{H}V^{\pi^t}(s_1) + \gamma^{H}V^{\pi_{t+H}}(\bar{s}_{s,1}).
$$

Then for a stationary policy $\bar{\pi}$, we can conclude $\bar{V}^\pi(\bar{s}_{s,1}) = \frac{\gamma^{H}}{1-\gamma^{H}}V^\pi(s_1)$. Since the optimal policy $\bar{\pi}^*$ is stationary, we have $\bar{V}^*(\bar{s}_{s,1}) = \frac{\gamma^{H}}{1-\gamma^{H}}V^*(s_1)$.

By definition, $\bar{\pi}$ is $\epsilon$-optimal at time step $t$ means that

$$
\bar{V}^{\pi^t}(\bar{s}_{s,1}) \geq \bar{V}^*(\bar{s}_{s,1}) - \epsilon.
$$

It follows that

$$
\gamma^{H}V^{\pi^t}(s_1) + \gamma^{H}V^{\pi_{t+H}}(\bar{s}_{s,1}) = \bar{V}^{\pi^t}(\bar{s}_{s,1}) \geq \bar{V}^*(\bar{s}_{s,1}) - \epsilon,
$$

hence

$$
\gamma^{H}V^{\pi^t}(s_1) \geq (1 - \gamma^{H})\bar{V}^*(\bar{s}_{s,1}) + \gamma^{H}(\bar{V}^*(\bar{s}_{s,1}) - V^{\pi_{t+H}}(\bar{s}_{s,1})) - \epsilon \geq (1 - \gamma^{H})\bar{V}^*(\bar{s}_{s,1}) - \epsilon.
$$

Therefore we have

$$
V^{\pi^t}(s_1) \geq \frac{1 - \gamma^{H}}{\gamma^{H}}\bar{V}^*(\bar{s}_{s,1}) - \epsilon/\gamma^{H} = V^*(s_1) - \epsilon/\gamma^{H},
$$

which means that $\pi^t$ is an ($\epsilon/\gamma^{H}$)-optimal policy.