Q-learning with Logarithmic Regret

Kunhe Yang
ykh17@mails.tsinghua.edu.cn
Tsinghua University

Lin F. Yang
linyang@ee.ucla.edu
University of California, Los Angeles

Simon S. Du
ssdu@cs.washington.edu
University of Washington

Abstract

This paper presents the first non-asymptotic result showing a model-free algorithm can achieve logarithmic cumulative regret for episodic tabular reinforcement learning if there exists a strictly positive sub-optimality gap. We prove that the optimistic Q-learning studied in [Jin et al. 2018] enjoys a $\mathcal{O}\Big(\frac{SA\cdot\mathrm{poly}(H)}{\Delta_{\min}}\log(SAT)\Big)$ cumulative regret bound where S is the number of states, A is the number of actions, H is the planning horizon, T is the total number of steps, and $\Delta_{\min}$ is the minimum sub-optimality gap of the optimal Q-function. This bound matches the lower bound in terms of S, A, T up to a $\log(SA)$ factor. We further extend our analysis to the discounted setting and obtain a similar logarithmic cumulative regret bound.

1 Introduction

Q-learning (Watkins and Dayan, 1992) is one of the most popular classes of methods for solving reinforcement learning (RL) problems. Q-learning tries to estimate the optimal state-action value function (Q-function). With a Q-function, at every state, one can just greedily choose the action with the largest Q value to interact with the RL environment. Compared to another popular class of methods, model-based learning, Q-learning algorithms (or more generally, model-free algorithms) often enjoy better memory and time efficiency[1]. These are the main reasons why Q-learning is applied in solving a wide range of RL problems (Mnih et al., 2015).

While model-free methods are widely applied in practice, most theoretical works study model-based RL. In one of the most fundamental RL frameworks, tabular RL, which is the focus of this paper, the majority of works study model-based algorithms (Kearns and Singh, 1999; Kakade, 2003; Singh and Yee, 1994; Azar et al., 2013, 2017; Dann and Brunskill, 2015; Dann et al., 2017; Agarwal et al., 2019; Simchowitz and Jamieson, 2019) with a few exceptions (Strehl et al., 2006; Jin et al., 2018; Wang et al., 2019; Zhang et al., 2020). From a regret minimization point of view, the state-of-the-art analysis demonstrates that one can achieve a $\sqrt{T}$-type regret bound where T is the number of episodes. Although these bounds are sharp in the worst-case scenario, they do not reveal the favorable structures of the environment, which can significantly decrease the regret.

One such structure is the existence of a strictly positive sub-optimality gap, i.e., for every state, there is a strictly positive value gap between the optimal action(s) and the rest (cf. Definition 2.1). In practice, arguably, nearly all environments with finite action sets satisfy some sub-optimality gap conditions. In Atari-games, e.g., Freeway, the optimal action has a value that is usually very distinctive from the rest of actions. In many other environments with finite number of actions, e.g. those control environments in OpenAI gym (Brockman et al., 2016), the gap condition usually holds. Similar gap conditions can be observed in other environments (see e.g. Kakade (2003)).

Theoretically, the sub-optimality gap is extensively investigated in the bandit problems, which can be viewed as RL problems with the planning horizon being 1. With this structure, one can drastically decrease the $\sqrt{T}$-type regret to $\log T$-type regret (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2018; Slivkins et al., 2019). For RL, most existing works that can leverage this structure require additional assumptions about the environment, such as finite hitting time and ergodicity (Jaksch et al., 2010; Tewari, 2007; Ok et al., 2018) or access to a generator (Zanette et al., 2019).[2]

[1] See Section 2 for the precise definitions of model-free and model-based algorithms in the tabular setting.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Corresponds to Lin F. Yang and Simon S. Du.

[2] The simulator allows the user to query any state-action

Recently, Simchowitz and Jamieson (2019) presented a systematic study of episodic tabular RL with the gap structure. They presented a novel algorithm which achieves the near-optimal $\sqrt{T}$-type regret in the worst scenario and $\log T$-type regret if there exists a strictly positive sub-optimality gap. Furthermore, they also provided instance-dependent lower bounds for a class of reasonable algorithms. See Section 1.1 for more detailed discussions.

However, to our knowledge, all existing works that obtain $\log T$-type regret bounds are about model-based algorithms. It remains open whether model-free algorithms such as Q-learning can achieve $\log T$-type regret bounds. Indeed, this is a challenging task. As discussed in Simchowitz and Jamieson (2019), their analysis framework cannot be applied to model-free algorithms directly. Later in this section, we also provide some technical explanations on why their approach is difficult to adopt.

Our Contributions We answer the aforementioned open problem by proving that the optimistic Q-learning algorithm studied in Jin et al. (2018) enjoys $\mathcal{O}\Big(\frac{SAH^6}{\Delta_{\min}}\log(SAT)\Big)$ cumulative regret where S is the number states, A is the number of actions, H is the planning horizon and $\Delta_{\min}$ is the minimum sub-optimality gap. To our knowledge, this is the first result showing model-free algorithms can achieve $\log T$-type regret. Furthermore, our bound matches the lower bound by Simchowitz and Jamieson (2019) in terms of S, A and T up to a $\log(SA)$ factor. Importantly, the algorithm does not need to know $\Delta_{\min}$.

Second, we extend our analysis to the infinite-horizon discounted setting with the regret defined in Liu and Su (2020), for which we show the optimistic Q-learning achieves $\mathcal{O}\Big(\frac{SA}{\Delta_{\min}(1-\gamma)^6}\log\Big(\frac{SAT}{\Delta_{\min}(1-\gamma)}\Big)\Big)$ regret where $0<\gamma<1$ is the discount factor.

Main Challenges Here we explain the main challenges of using existing analyses and give an overview of our main techniques at a high level. The existing proof in Jin et al. (2018) bounds the regret in terms of a weighted sum of the estimation error of Q-function. Note the estimation error scales $1/\sqrt{T}$ which in turn gives a $\sqrt{T}$-type regret, but cannot give a $\log T$-type regret bound.

For model-based algorithms, Simchowitz and Jamieson (2019) introduced a novel notion, *optimistic surplus* (cf. Equation (35)), which can be bounded by the estimation error of the transition probability. The logarithmic regret bound can be proved via a clipping trick on top of the optimistic surplus.

pair.

Unfortunately, as acknowledged by Simchowitz and Jamieson (2019), their analysis is highly tailored to model-based algorithms. First, model-free algorithms do not estimate the probability transition, so we cannot bound the optimistic surplus via this approach. Secondly, although we can also obtain a formula for the optimistic surplus in each episode using the update rules of the Q-learning algorithm, the formula depends on the estimation error of Q-function in previous episodes. This dependency makes it difficult to bound the optimistic surplus. See Section 8 for more technical details.

Technique Overview In this paper, we adopt an entirely different *counting* approach. We first write the total regret as expected sum over sub-optimality gaps appearing in the whole learning process, then use the estimation error of Q-function and the definition of sub-optimality gap to upper bound the number of times the algorithm takes suboptimal actions.

To obtain a sharp dependency on $\Delta_{\min}$, we divide the interval $[\Delta_{\min}, H]$ (the range of all gaps) into multiple subintervals. We then bound the sum of learning error in each subinterval by its maximum value times the number of steps falling into this subinterval. The number of steps in each layer is bounded through computing the weighted sum of learning error across all the episodes $k \in [K]$. See detailed discussion in Lemma 4.3 and Lemma 4.2.

Organization This paper is organized as follows. In Section 1.1 we discuss related works. In Section 2, we introduce necessary definitions and backgrounds. In Section 3, we present our main results and discussions. In Section 4, we give the proof of our theorem on the episodic setting. We conclude in Section 5 and leave remaining proofs to the appendix.

1.1 Related Work

Gap-independent Finite-horizon and Infinite-horizon Discounted RL [3] There is a long list of results about regret or sample complexity of tabular RL, dating back to Singh and Yee (1994). One line of works require access to a simulator where the agent can query samples freely from any state-action pair of the environment and therefore the agent does not need to design a strategy to explore the environment. (Kearns and Singh, 1999; Kakade, 2003; Singh and Yee, 1994; Azar et al., 2013; Sidford et al., 2018b,a; Agarwal et al., 2019; Zanette et al., 2019; Li et al., 2020).

[3] There is another line of works on gap-independent infinite-horizon average-reward setting. This setting is beyond the scope of this paper.

Another line of works drop the simulator assumption and thus the agent needs to use advanced techniques, such as upper confidence bound (UCB) to explore the state space (Azar et al., 2017; Dann and Brunskill, 2015; Dann et al., 2017, 2019; Jin et al., 2018; Strehl et al., 2006; Zhang et al., 2020; Simchowitz and Jamieson, 2019; Zanette and Brunskill, 2019; Wang et al., 2019). In terms of the regret, the state-of-art result shows one can achieve $\widetilde{\mathcal{O}}\Big(\sqrt{SAH^2T} + \text{poly}\,(S, A, H)\Big)$ regret for which the first term nearly match the $\Omega\left(\sqrt{SAH^2T}\right)$ up to logarithmic factors (Dann and Brunskill, 2015; Osband and Roy, 2016).[4] Among these results, only a few are for model-free algorithms (Strehl et al., 2006; Jin et al., 2018; Wang et al., 2019; Zhang et al., 2020) and only very recently, Jin et al. (2018); Zhang et al. (2020) showed Q-learning can achieve $\sqrt{T}$-type regret bounds.

Sub-optimality Gap The results about gap-dependent regret bounds for MDP algorithms can be categorized into asymptotic bounds and non-asymptotic bounds. Asymptotic bounds are only valid when the total number of steps T is large enough. These bounds often suffer from the worst-case dependency on some problem-specific quantities, such as diameter and worst-case hitting time. Under the infinite-horizon average-reward setting, Auer and Ortner (2007) provided a logarithmic regret algorithm for irreducible MDPs. Besides dependency on hitting times, their regret also depends inversely on Δ_*^2, the squared distance between optimal and second-optimal policy. Along this direction and improving over previous algorithm of Burnetas and Katehakis (1997), Tewari and Bartlett (2008) proposed an algorithm called Optimistic Linear Programming (OLP). OLP is proved to have $C(P)\log T$ regret asymptotically in T, where $C(P)$ depends on some diameter-related quantity as well as the sum over reciprocals of gaps for (x, a) inside a critical set.

For non-asymptotic bounds, Jaksch et al. (2010) introduced UCRL2 algorithm, which enjoys $\widetilde{\mathcal{O}}\Big(\frac{D^2S^2A}{\Delta_*}\log T\Big)$ regret where D is the diameter. More recently, Ok et al. (2018) derived problem-specific lower bounds for both structured and unstructured MDPs. Their lower bound scales $SA\log T$ for unstructured MDP and $c\log T$ for structured MDP, where this c depends on both the minimal action sub-optimality gap and the span of bias function, which can be bounded by diameter D. For non-asymptotic bounds, Simchowitz and Jamieson (2019) proved that model-based optimistic algorithm StrongEuler has gap-dependent regret bound that holds uniformly over T. Moreover, their bounds depend only on H and not on any term such as hitting time or diameter. In Section 3, we compare our result with the one in Simchowitz and Jamieson (2019) in more detail.

[4] In this paper, we study the same setting as in Jin et al. (2018) where the reward at each level is in $[0, 1]$, and the transition probabilities at each level can be different. In another setting, the total reward is bounded by 1 and the transition probabilities at each level are the same. The latter setting is more challenging to analyze and the worst-case sample complexity is still open (Jiang and Agarwal, 2018; Wang et al., 2020).

2 Preliminaries

Episodic MDP An episodic Markov decision process (MDP) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, P, r)$, where $\mathcal{S}$ is the finite state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the finite action space with $|\mathcal{A}| = A$, $H \in \mathbb{Z}_+$ is the planning horizon, $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator at step h that takes a state-action pair and returns a distribution over states, and $r_h : \mathcal{S}\times\mathcal{A} \to [0, 1]$ is the deterministic reward function at step h. Each episode starts at an initial state $x_1 \in \mathcal{S}$ picked by an adversary.

In this paper, we focus on deterministic policies. A deterministic policy π is a sequence of mappings $\pi_h : \mathcal{S} \to \mathcal{A}$ for $h = 1, \ldots, H$. Given a policy π, for a state $x \in \mathcal{S}$, the value function of state $x \in \mathcal{S}$ at the h-step is defined as

$$V_h^\pi(x) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, \pi_{h'}(x_{h'}))\middle| x_h = x\right],$$

and the associated Q-function of a state-action pair $(x, a) \in \mathcal{S}\times\mathcal{A}$ at the h-step is

$$Q_h^\pi(x, a) := r_h(x, a) + \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}(x_{h'}, \pi_{h'}(x_{h'}))\middle| \begin{matrix} x_h = x \\ a_h = a \end{matrix}\right].$$

We let π^* be the optimal policy such that $V^{\pi^*}(x) = V^*(x) = \text{argmax}_\pi V^\pi(x)$ and $Q^{\pi^*}(x, a) = Q^*(x, a) = \text{argmax}_\pi Q^\pi(x, a)$ for every (x, a). For episodic MDP, the agent interacts with the MDP for $K \in \mathbb{Z}^+$ episodes. For each episode $k = 1, \ldots, K$, the learning algorithm Alg specifies a policy π^k, plays π^k for H steps and observes trajectory $(x_1, a_1), \cdots, (x_H, a_H)$. The total number of steps is $T = KH$, and the total regret of an execution instance of Alg is then

$$\text{Regret}(K) = \sum_{k=1}^{K}\left(V_1^* - V_1^{\pi^k}\right)(x_1^k).$$

In this paper we focus on bounding the expected regret $\mathbb{E}[\text{Regret}(K)]$ where the expectation is over the randomness from the environment.

Algorithm 1 Q-learning with UCB-Hoeffding

Initialize: $Q_h(x,a) \leftarrow H$ and $N_h(x,a) \leftarrow 0$ for all $(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
Define $\alpha_t = \frac{H+1}{H+t}$, $\iota \leftarrow \log\left(SAT^2\right)$.
for episode $k \in [K]$ **do**
  receive x_1.
  for step $h \in [H]$ **do**
    Take action $a_h \leftarrow \operatorname{argmax}_{a' \in \mathcal{A}} Q_h(x_h, a')$, observe x_{h+1}.
    $t = N_h(x_h, a_h) \leftarrow N_h(x_h, a_h) + 1$,
    $b_t \leftarrow c\sqrt{H^3\iota/t}$, ▷ c is a constant that can be set to 4.
    $Q_h(x_h, a_h) \leftarrow (1-\alpha_t)\, Q_h(x_h, a_h) + \alpha_t \left[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t\right]$,
    $V_h(x_h) \leftarrow \min\left\{H, \max_{a' \in \mathcal{A}} Q_h(x_h, a')\right\}$.

Model-free Algorithm V.S. Model-based Algorithm In this paper we focus on *model-free* Q-learning algorithms. Formally, by model-free algorithms, we mean the space complexity of the algorithm scales at most *linearly* in S in contrast to the model-based algorithms whose space complexity often scales *quadratically* with S (Strehl et al., 2006; Sutton and Barto, 1998; Jin et al., 2018). For episodic MDP, we will analyze the Q-learning with UCB-Hoeffding algorithm studied in Jin et al. (2018) (cf. Algorithm 1). At a high level, this algorithm maintains an upper bound of Q^* for every (s,a) pair and choose the action greedily at every episode. The algorithm uses a carefully designed step size sequence $\{\alpha_k\}$ to update the upper bound based on the observed data. Jin et al. (2020) proved that Algorithm 1 enjoys $\left(\sqrt{H^4SAT\log\left(SAT\right)}\right)$ regret, which is the first $\sqrt{T}$-type bound for model-free algorithms.

Sub-optimality Gap Our paper investigates what structures of the MDP enable us to improve the $\sqrt{T}$-type bound. In this paper we focus on the positive sub-optimality gap condition (Simchowitz and Jamieson, 2019; Du et al., 2019c, 2020).

Definition 2.1 (Sub-optimality Gap)**.** *Given $h \in [H]$, $(x,a) \in \mathcal{S} \times \mathcal{A}$, the suboptimality gap of (x,a) at level h is defined as $\Delta_h(x,a) := V_h^*(x) - Q_h^*(x,a)$.*

Definition 2.2. *Minimum Sub-optimality Gap] Denote by $\Delta_{\min}$ the minimum non-zero gap: $\Delta_{\min} := \min_{h,x,a}\left\{\Delta_h(x,a) : \Delta_h(x,a) \neq 0\right\}$.*

Note that if $\{\Delta_h(x,a) : \Delta_h(x,a) \neq 0\} = \emptyset$, then all the states are the same, and the MDP degenerates. Otherwise we always have $\Delta_{\min} > 0$. For the rest of the paper, we focus on the case when $\Delta_{\min} > 0$. In Section 1 we have discussed why many MDPs admit this structure. Our main result is a logarithmic regret bound of Algorithm 1.

Infinite-horizon Discounted MDP In this paper we also study infinite-horizon discounted MDP, which is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \gamma, P, r)$, where every step shares the same transition operator P and reward function r. Here γ denotes the discount factor, and there is no restart during the entire process. Let $\mathcal{C} = \{\mathcal{S} \times \mathcal{A} \times [0,1]\}^* \times \mathcal{S}$ be the set of all possible trajectories of any length. A non-stationary deterministic policy $\pi : \mathcal{C} \to \mathcal{A}$ is a mapping from paths to actions. The V function and Q function are defined as below $(c_i := (x_1, a_1, r_1, \cdots, x_i))$.

$$V^\pi(x) := \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} r(x_i, \pi(c_i)) \middle| x_1 = x\right],$$

$$Q^\pi(x,a) := r(x,a) + \mathbb{E}\left[\sum_{i=2}^{\infty} \gamma^{i-1} r(x_i, \pi(c_i)) \middle| \begin{matrix} x_1 = x \\ a_1 = a \end{matrix}\right].$$

Let $V^*(s)$ and $Q^*(s,a)$ denote respectively the value function and Q function of the optimal policy π^*.

Definition 2.3 (Sub-optimality Gap)**.** *Given $(x,a) \in \mathcal{S} \times \mathcal{A}$, the suboptimality gap of (x,a) is defined as $\Delta(x,a) := V^*(x) - Q^*(x,a)$.*

Definition 2.4 (Minimum Sub-optimality Gap)**.** *Denote by $\Delta_{\min}$ the minimum non-zero gap: $\Delta_{\min} := \min_{x,a}\left\{\Delta(x,a) : \Delta(x,a) \neq 0\right\}$.*

Again, if $\{\Delta(x,a) : \Delta(x,a) \neq 0\} = \emptyset$, all the states are the same, and the MDP degenerates. Otherwise, we have $\Delta_{\min} > 0$.

Consider a game that starts at state x_1. A learning algorithm Alg specifies an initial non-stationary policy π_1. At each time step t, the player takes action $\pi_t(x_t)$, observes r_t and x_{t+1}, and updates π_t to π_{t+1}. The total regret of Alg for the first T steps is thus defined as $\text{Regret}(T) = \sum_{t=1}^{T}\left(V^* - V^{\pi_t}\right)(x_t)$. This definition was studied in Liu and Su (2020), which follows the sample complexity definition in Kakade (2003). For this setting, we study Algorithm 2. This is a simple adaptation of Algorithm 1 that takes γ into account, so we defer it to the appendix. We prove Algorithm 2 also enjoys a logarithmic regret bound.

3 Main Theoretical Results

Now we present our main results.

Main Result for Episodic MDP The following theorem characterizes the performance of Algorithm 1 for episodic MDP. To our knowledge, this is the first theoretical result showing a model-free algorithm can achieve logarithmic regret of tabular RL.

Theorem 3.1 (Logarithmic Regret Bound of Q-learning for Episodic MDP)**.** *The expected regret of*

Algorithm 1 for episodic tabular MDP is upper bounded by $\mathbb{E}[\text{Regret}(K)] \le \mathcal{O}\left(\frac{H^6SA}{\Delta_{\min}}\log(SAT)\right)$.

An interesting advantage of our theorem is adaptivity. Note the algorithm we analyze is exactly the same algorithm studied in Jin et al. (2020), which has been shown to achieve the worst-case $\sqrt{T}$-type regret bound. Theorem 3.1 suggests that one does not need to modify the algorithm to exploit the strictly positive minimum sub-optimality gap structure, Algorithm 1 *automatically adapts* to this benign structure. Importantly, Algorithm 1 does not need to know $\Delta_{\min}$.

Proposition 2.2 in Simchowitz and Jamieson (2019) suggested that any algorithm with sub-linear regret in the worst case, suffer an $\Omega\left(\sum_{(x,a),\Delta_1(x,a)>0}\frac{H^2}{\Delta_1(x,a)}\log T\right)$ expected regret. Therefore, the dependencies on S, A and T are nearly tight in Theorem 3.1.

One may wonder whether it is possible to obtain a regret bound that only depends the sum of positive gaps, e.g., $O\left(\sum_{(x,a),\Delta_1(x,a)>0}\frac{H^2}{\Delta_1(x,a)}\log T\right)$, unlike ours, which is a multiple of $1/\Delta_{\min}$. Unfortunately, Simchowitz and Jamieson (2019) showed, all existing algorithms, including Algorithm 1 and their algorithm, suffer an $\Omega\left(\frac{S}{\Delta_{\min}}\right)$ regret, and new algorithmic ideas are needed in order to circumvent this lower bound.

We compare Theorem 3.1 with the regret bound for model-based algorithm in Simchowitz and Jamieson (2019) (in big-$\mathcal{O}$ form):

$$\left(\sum_{\substack{(x,a):\\ \exists h\in[H],\Delta_h(x,a)>0}}\frac{H^3}{\min_h\Delta_h(x,a)}+\frac{SH^3}{\Delta_{\min}} + H^4SA\max(S,H)\log\left(\frac{SAH}{\Delta_{\min}}\right)\right)\log(SAHT)$$

First recall our bound is for a model-free algorithm which is more space-efficient and time-efficient than the model-based algorithm in Simchowitz and Jamieson (2019). In terms of the regret bound, Theorem 3.1's dependency on H is worse than that in their bound. We remark that simple model-free algorithms may have a worse dependency on H compared to model-based algorithms (e.g., see Jin et al. (2018)).

Now let us consider an environment where there are $\sim SA$ state-action pairs whose gap is $\Delta_{\min}$. Then the bound in Simchowitz and Jamieson (2019) becomes

$$\left(\frac{H^3SA}{\Delta_{\min}}+H^4SA\max\{S,H\}\log\left(\frac{SAH}{\Delta_{\min}}\right)\right)\log(SAHT).$$

In this regime, both Theorem 3.1 and their bound have an $\frac{SA}{\Delta_{\min}}$ term. Their bound also has an additional $H^4SA\max(H,S)\log\left(\frac{SAH}{\Delta_{\min}}\right)$ burn-in term which our bound does not have. When S is large compared to H and $\Delta_{\min}$, this term scales S^2 and can dominate other terms, so our bound is better. The technical reason behind this phenomenon is that Algorithm 1 uses the Hoeffding bound for constructing bonus on Q-value, which does not need burn-in.

Main Result for Infinite-horizon Discounted MDP Algorithm 1 can be easily generalized to the discounted MDP. See Algorithm 2 in the appendix. We also obtain a logarithmic regret bound for infinite-horizon discounted MDP.

Theorem 3.2 (Logarithmic Regret Bound of Q-learning for Infinite-horizon Discounted MDP)**.** *The expected regret of Algorithm 2 for infinite-horizon discounted MDP is upper bounded by* $\mathbb{E}[\text{Regret}(T)] \le \mathcal{O}\left(\frac{SA}{\Delta_{\min}(1-\gamma)^6}\log\frac{SAT}{\Delta_{\min}(1-\gamma)}\right)$.

Theorem 3.2 suggests that model-free algorithms can achieve logarithmic regret even in the infinite-horizon discounted MDP setting. The main difference from Theorem 3.1 is that H is replaced by $\frac{1}{1-\gamma}$. By analogy, we believe the dependencies on S, A, T and $\Delta_{\min}$ are nearly tight and the dependency $\frac{1}{1-\gamma}$ can be improved. The proof of Theorem 3.2 is deferred to Appendix.

4 Proof of Theorem 3.1

In this section, we prove Theorem 3.1.

Notations Let $Q_h^k(x,a), V_h^k(x), N_h^k(x,a)$ denote the value of $Q_h(x,a), V_h(x)$,and $N_h(x,a)$ right before the k-th episode, respectively. Let $\mathbb{I}[\cdot]$ denote the indicator function. Let $\tau_h(x,a,i) := \max\left\{k : N_h^k(x,a) = i-1\right\}$ be the episode k at which $(x_h^k, a_h^k) = (x,a)$ for the i-th time. We will abbreviate $N_h^k(x_h^k,a_h^k)$ for n_h^k when no confusion can arise. α_t^i is defined by the following: $\alpha_t = \frac{H+1}{H+t}$, $\alpha_t^0 = \prod_{j=1}^t(1-\alpha_j)$ and $\alpha_t^i = \alpha_i\prod_{j=i+1}^t(1-\alpha_j)$ $(i>0)$. Let $\beta_0 = 0$ and $\beta_t = 4c\sqrt{\frac{H^3\iota}{t}}$ for $t \ge 1$.

Proof of Theorem 3.1 Our proof starts with the observation that the regret of each episode can be rewritten as the expected sum of sub-optimality gaps for each action:

$$\begin{aligned}
&\left(V_1^* - V_1^{\pi^k}\right)(x_1^k)\\
&= V_1^*(x_1^k) - Q_1^*(x_1^k,a_1^k) + \left(Q_1^* - Q_1^{\pi^k}\right)(x_1^k,a_1^k)\\
&= \Delta_1(x_1^k,a_1^k) + \mathbb{E}_{s'\sim P_1(\cdot|x_1^k,a_1^k)}\left[\left(V_2^* - V_2^{\pi^k}\right)(s')\right]\\
&= \cdots = \mathbb{E}\left[\sum_{h=1}^H \Delta_h(x_h^k,a_h^k)\middle| a_h^k = \pi_h^k(x_h^k)\right]. \qquad (1)
\end{aligned}$$

In order to bound $\Delta_h(x_h^k, a_h^k)$ by learning error $(Q_h^k - Q_h^*)(x_h^k, a_h^k)$, we define the following concentration event.

Definition 4.1 (Concentration of Learning Errors)**.**

$$\mathcal{E}_{\text{conc}} := \Bigg\{ \forall (x,a,h,k) \colon 0 \le \left(Q_h^k - Q_h^*\right)(x,a) \le \alpha_{n_h^k}^0 H + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(V_{h+1}^{\tau_h(x,a,i)} - V^* \right)\left(x_{h+1}^{\tau_h(x,a,i)}\right) + \beta_{n_h^k} \Bigg\}.$$

Intuitively, $\mathcal{E}_{\text{conc}}$ is the event in which all the learning errors of the value function is both bounded below (by zero) and bounded above.

We now refer to Jin et al. (2018) for the following lemma that shows $\mathcal{E}_{\text{conc}}$ happens with high probability via a concentration argument.

Lemma 4.1 (Concentration)**.** *Event $\mathcal{E}_{\text{conc}}$ occurs w.p. at least $1 - 1/T$.*

Lemma 4.1 suggests that Algorithm 1 is optimistic on $\mathcal{E}_{\text{conc}}$. Combining with the greedy choice of actions yields

$$V_h^*\left(x_h^k\right) = Q_h^*\left(x_h^k, a^*\right) \le Q_h^k\left(x_h^k, a^*\right) \le Q_h^k\left(x_h^k, a_h^k\right). \quad (2)$$

To bound $\Delta_h(x_h^k, a_h^k)$, the following notion introduced in Simchowitz and Jamieson (2019) is convenient. If we define $\text{clip}\left[x \middle| \delta\right] := x \cdot \mathbb{I}[x \ge \delta]$, then Ineq (2) suggests that $\Delta_h(x_h^k, a_h^k)$ can be bounded by clipped estimation error:

$$\begin{aligned} \Delta_h(x_h^k, a_h^k) &= \text{clip}\left[V_h^*\left(x_h^k\right) - Q_h^*(x_h^k, a_h^k) \middle| \Delta_{\min} \right] \\ &\le \text{clip}\left[\left(Q_h^k - Q_h^*\right)(x_h^k, a_h^k) \middle| \Delta_{\min} \right]. \end{aligned} \quad (3)$$

Our main technique to get $1/\Delta_{\min}$ instead of $1/\Delta_{\min}^2$ regret bound is to classify gaps of state-action pairs into different intervals and count them separately. Note the gap can range from $\Delta_{\min}$ to H. Thus, we divide the interval $[\Delta_{\min}, H]$ into N disjoint intervals: $[\Delta_{\min}, 2\Delta_{\min}), \cdots, \left[2^{N-1}\Delta_{\min}, 2^N \Delta_{\min}\right]$, where $N = \lceil \log_2 (H/\Delta_{\min}) \rceil$.

Lemma 4.2 below is our main technical lemma which upper bounds the number of steps Algorithm 1 chooses a sub-optimal action whose suboptimality is in a certain interval.

Lemma 4.2 (Bounded Number of Steps in Each Interval)**.** *Under $\mathcal{E}_{\text{conc}}$, we have for every $n \in [N]$,*

$$C^{(n)} := \left| \left\{ (k,h) : \begin{array}{c} \left(Q_h^k - Q_h^*\right)(x_h^k, a_h^k) \in \\ [2^{n-1}\Delta_{\min}, 2^n \Delta_{\min}) \end{array} \right\} \right| \le \mathcal{O}\left(\frac{H^6 SA\iota}{4^n \Delta_{\min}^2} \right), \quad \text{where } \iota = \log\left(SAT^2\right).$$

Before we give the proof for Lemma 4.2, we first show how to use Lemma 4.2 to prove Theorem 3.1.

Proof of Theorem 3.1 Since the trajectories inside $\mathcal{E}_{\text{conc}}$ have bounded empirical regret, and complementary event $\overline{\mathcal{E}_{\text{conc}}}$ happens with sufficiently low probability,

$$\begin{aligned} &\mathbb{E}[\text{Regret}(K)] = \mathbb{E}\left[\sum_{k=1}^{K} \sum_{h=1}^{H} \Delta_h\left(x_h^k, a_h^k\right) \right] \\ &= \sum_{\text{traj}} \mathbb{P}(\text{traj}) \cdot \sum_{k,h} \Delta_h\left(x_h^k, a_h^k \middle| \text{traj}\right) \quad (4) \\ &\le \sum_{\text{traj} \in \mathcal{E}_{\text{conc}}} \mathbb{P}(\text{traj}) \cdot \sum_{k,h} \text{clip}\left[\left(Q_h^k - Q_h^*\right)\left(x_h^k, a_h^k \middle| \text{traj}\right) \middle| \Delta_{\min} \right] \\ &\qquad + \sum_{\text{traj} \in \overline{\mathcal{E}_{\text{conc}}}} \mathbb{P}(\text{traj}) \cdot TH \quad (5) \\ &\le \mathbb{P}(\mathcal{E}_{\text{conc}}) \sum_{n=1}^{N} 2^n \Delta_{\min} C^{(n)} + \mathbb{P}\left(\overline{\mathcal{E}_{\text{conc}}}\right) \cdot TH \quad (6) \\ &\le \sum_{n=1}^{N} \mathcal{O}\left(\frac{H^6 SA\iota}{2^n \Delta_{\min}} \right) + H \quad (7) \\ &\le \mathcal{O}\left(\frac{H^6 SA}{\Delta_{\min}} \log(SAT) \right). \end{aligned}$$

Above, (4) follows from the definition of expectation, (5) is because Ineq (3) suggests that for trajectories inside $\mathcal{E}_{\text{conc}}$, gaps can be bounded by clipped learning errors; whereas for trajectories outside of $\mathcal{E}_{\text{conc}}$, sub-optimality gaps never exceed H. (6) follows from adding an outer summation for state-action pairs over the N disjoint subintervals, then bounding the estimation error in each subinterval by its maximum value times the number of steps it contains. (7) comes from a sum of numbers in a geometric progression generated by Lemma 4.2, and the fact that $\mathbb{P}\left(\overline{\mathcal{E}_{\text{conc}}}\right) \le 1/T$ from concentration Lemma 4.1. In the final step, we notice that $\iota = \log(SAT^2) = \mathcal{O}(\log(SAT))$. □

Proof of Lemma 4.2 The proof of Lemma 4.2 relies on a general lemma (Lemma 4.3) characterizing a weighted sum of the estimation errors of Q-function.Then we choose a particular sequence of weights to prove Lemma 4.2. We remark that this general idea has appeared in Jin et al. (2018); Wang et al. (2019); Zhang et al. (2020).

Formally, we use the following definition.

Definition 4.2 ((C,w)-Sequence (Definition 3 in Wang et al. (2019)))**.** *A sequence $\{w_k\}_{k \ge 1}$ is called a (C,w)-sequence if $0 \le w_k \le w$ for all k and $\sum_k w_k \le C$.*

Lemma 4.3 (Weighted Sum of Learning Errors)**.** *On event $\mathcal{E}_{\text{conc}}$, for every $h \in [H]$, if $\{w_k\}_{k \in [K]}$ is a (C,w)-*

sequence, then:

$$\sum_{k=1}^{K} w_k \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \le ewSAH^2 + 10c\sqrt{ewSACH^5\iota}.$$

Before presenting the proof of Lemma 4.3, we refer the readers to Jin et al. (2018) for Lemma 4.4 below, which summarizes the properties of α_t^i that will be useful in our proof.

Lemma 4.4 (Properties of α_t^i). *Let $\alpha_t = \frac{H+1}{H+t}$, $\alpha_t^0 = \prod_{j=1}^t (1-\alpha_j)$ and $\alpha_t^i = \alpha_i \prod_{j=i+1}^t (1-\alpha_j)$ for $0 < i \le t$.*

(i) $\sum_{i=1}^t \alpha_t^i = 1$ and $\alpha_t^0 = 0$ for every $t \ge 1$, $\sum_{i=1}^t \alpha_t^i = 0$ and $\alpha_t^0 = 1$ for $t = 0$.

(ii) $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$ for every $i \ge 1$.

Proof of Lemma 4.3 We will recursively bound the weighted sum of step h by its next step $(h+1)$, and unroll $(H-h+1)$ times for the desired bound. As suggested by Lemma 4.1, upper bounds of learning error holds under $\mathcal{E}_{\text{conc}}$. Thus we have

$$\begin{aligned}
&\sum_{k=1}^{K} w_k \left(Q_h^k - Q_h^*\right)(x_h^k, a_h^k) \\
\le &\sum_{k=1}^{K} w_k \left(H\alpha_{n_h^k}^0 + \beta_{n_h^k} + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(V_{h+1}^{\tau_h(s,a,i)} - V_{h+1}^*\right)(x_{h+1}^{\tau_h(s,a,i)})\right) \\
= &\sum_{k=1}^{K} w_k H \alpha_{n_h^k}^0 + \sum_{k=1}^{K} w_k \beta_{n_h^k} \\
&+ \sum_{k=1}^{K} w_k \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(V_{h+1}^{\tau_h(x_h^k, a_h^k, i)} - V_{h+1}^*\right)(x_{h+1}^{\tau_h(x_h^k, a_h^k, i)}). \quad (8)
\end{aligned}$$

For the first term of (8), $n_h^k = 0$ at most once for every state-action pair, and we always have $w_k \le w$. Thus,

$$\sum_{k=1}^{K} w_k H \alpha_{n_h^k}^0 = \sum_{k=1}^{K} w_k H \mathbb{I}\big[n_h^k = 0\big] \le wSAH. \quad (9)$$

The second term of (8) can be bounded by the following inequalities with respective reasons listed below:

$$\begin{aligned}
\sum_{k=1}^{K} w_k \beta_{n_h^k} &= \sum_{s,a} \sum_{\substack{k=1 \\ (s_h^k, a_h^k) = (s,a)}}^{K} w_k \beta_{n_h^k} \\
&= 4c\sqrt{H^3\iota} \sum_{s,a} \sum_{i=2}^{N_h^K(s,a)} \frac{w_{\tau(s,a,i)}}{\sqrt{i-1}} \quad (10) \\
&\le 4c\sqrt{H^3\iota} \sum_{s,a} \sum_{i=1}^{\lceil C_{s,a}/w \rceil} \frac{w}{\sqrt{i}} \quad (11) \\
&\le 10c\sqrt{H^3\iota} \sum_{s,a} \sqrt{C_{s,a} w} \quad (12) \\
&\le 10c\sqrt{SACwH^3\iota}. \quad (13)
\end{aligned}$$

Above, (10) comes from prior definition $\beta_t = 4c\sqrt{\frac{H^3\iota}{t}}$ when $t \ge 1$ and $\beta_0 = 0$. Note that $\tau_h(x,a,i)$ is the episode where (x,a) is visited for the i-th time, so we always have $n_h^{\tau_h(x,a,i)} = i - 1$. (11) follows from a rearrangement inequality with $C_{s,a}$ defined as $C_{s,a} := \sum_{i=1}^{n_h^K(s,a)} w_{\tau(s,a,i)}$, where we always keep in mind that $0 < w_{\tau(s,a,i)} \le w$. (12) follows from the integral conversion of $\sum_i 1/\sqrt{i}$, and (13) is true because of Cauchy-Schwartz inequality where $\sum_{s,a} C_{s,a} = \sum_{k=1}^K w_k \le C$.

For the third term in Ineq (8), we notice that $V_h^k(x_h^k) = Q_h^k(x_h^k, a_h^k)$ due to greedy choice of actions and $V_h^*(x_h^k) \ge Q_{h+1}^*(x_{h+1}^k, a_{h+1}^k)$ by definition. Therefore $\left(V_h^k - V_h^*\right)(x_h^k) \le \left(Q_h^k - Q_h^*\right)(x_h^k, a_h^k)$. Note that $\forall k \in [K]$, the third term takes into account all the prior episodes $l < k$ where $(x_h^k, a_h^k) = (x_h^l, a_h^l)$, indicating that the learning error at step l is only counted by subsequent episodes $k > l$ when the same (s,a) is visited. Thus, we exchange the order of summation and obtain

$$\begin{aligned}
&\sum_{k=1}^{K} w_k \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(V_{h+1}^{\tau_h(x_h^k, a_h^k, i)} - V_{h+1}^*\right)(x_{h+1}^{\tau_h(x_h^k, a_h^k, i)}) \\
= &\sum_{l=1}^{K} \left(V_{h+1}^l - V_{h+1}^*\right)(x_{h+1}^l) \sum_{j=n_h^l+1}^{N_h^K(x_h^l, a_h^l)} w_{\tau_h(x_h^l, a_h^l, j)} \alpha_j^{n_h^l+1} \\
\le &\sum_{l=1}^{K} \left(Q_{h+1}^l - Q_{h+1}^*\right)(x_{h+1}^l, a_{h+1}^l) \sum_{j=n_h^l+1}^{N_h^K(x_h^l, a_h^l)} w_{\tau_h(x_h^l, a_h^l, j)} \alpha_j^{n_h^l+1}.
\end{aligned}$$

Then for $l \in [K]$ we let $\widetilde{w}_l = \sum_{j=n_h^l+1}^{N_h^K(x_h^l, a_h^l)} w_{\tau_h(x_h^l, a_h^l, j)} \alpha_j^{n_h^l+1}$ and further simplify the above equation to be

$$\begin{aligned}
&\sum_{k=1}^{K} w_k \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(V_{h+1}^{\tau_h(s,a,i)} - V_{h+1}^*\right)(x_{h+1}^{\tau_h(s,a,i)}) \\
\le &\sum_{l=1}^{K} \widetilde{w}_l \left(Q_{h+1}^l - Q_{h+1}^*\right)(x_{h+1}^l, a_{h+1}^l). \quad (14)
\end{aligned}$$

Next, we use Lemma 4.4 to verify that $\{\widetilde{w}_l\}_{l \in [K]}$ is a $\big(C, \big(1 + \frac{1}{H}\big)w\big)$-sequence:

$$\widetilde{w}_l \le w \sum_{j=n_h^l+1}^{N_h^K(x_h^l, a_h^l)} \alpha_j^{n_h^l+1} \le w \sum_{j \ge n_h^l+1} \alpha_j^{n_h^l+1} \le \left(1 + \frac{1}{H}\right) w,$$

$$\sum_{l=1}^{K} \widetilde{w}_l = \sum_{l=1}^{K} \sum_{j=n_h^l+1}^{N_h^K(x_h^l, a_h^l)} w_{\tau_h(x_h^l, a_h^l, j)} \alpha_j^{n_h^l+1}$$

$$= \sum_{k=1}^{K} w_k \sum_{t=1}^{n_h^k} \alpha_{n_h^k}^t = \sum_{k=1}^{K} w_k \le C. \tag{15}$$

Plugging the upper bounds of three separate terms in (9), (13) and (14) back into Ineq (8) gives us

$$\sum_{k=1}^{K} w_k \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \le wSAH + 10c\sqrt{SACwH^3\iota} + \sum_{l=1}^{K} \widetilde{w}_l \big(Q_{h+1}^l - Q_{h+1}^*\big)(x_{h+1}^l, a_{h+1}^l), \tag{16}$$

where the third term is a weighted sum of learning errors of the same format, but taken at level $h+1$. In addition, it has weights $\{\widetilde{w}_l\}_{l\in[K]}$ being a $(C, (1+1/H)w)$-sequence. Therefore, the above analysis will also yield

$$\sum_{l=1}^{K} \widetilde{w}_l \big(Q_{h+1}^l - Q_{h+1}^*\big)(x_h^l, a_h^l) \le \left(1+\frac{1}{H}\right) wSAH + 10c\sqrt{SAC\left(1+\frac{1}{H}\right)wH^3\iota} + \big[\text{weighted sum at } (h+2)\big].$$

Recursing this argument for $h+1, h+2, \cdots, H$ gives us

$$\begin{aligned} &\sum_{k=1}^{K} w_{k,h} \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \\ &\le \sum_{h'=0}^{H-h} \left(SAH\,(1+1/H)^{h'} w + 10c\sqrt{SAC\,(1+1/H)^{h'} wH^3\iota} \right) \\ &\le H\left(SAHew + 10c\sqrt{SACewH^3\iota} \right). \end{aligned} \tag{17}$$

which is the desired conclusion.

With Lemma 4.3, we can easily prove Lemma 4.2 by choosing a particular (C, w)-sequence.

Proof of Lemma 4.2 For every $n \in [N]$, $h \in [H]$, let

$$\begin{aligned} w_k^{(n,h)} &:= \mathbb{I}\big[\big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \in \big[2^{n-1}\Delta_{\min}, 2^n \Delta_{\min}\big)\big], \\ C^{(n,h)} &:= \sum_{k=1}^{K} w_k^{(n,h)} \\ &= \left| \left\{ k : \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \in \big[2^{n-1}\Delta_{\min}, 2^n \Delta_{\min}\big) \right\} \right|. \end{aligned}$$

By definition, $\forall h \in [H]$ and $n \in [N]$, $\left\{w_k^{(n,h)}\right\}_{k\in[K]}$ is a $(C^{(n,h)}, 1)$-sequence. Now we consider bounding $\sum_{k=1}^{K} w_k^{(n,h)} \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k)$ from both sides. On the one hand, by Lemma 4.3,

$$\sum_{k=1}^{K} w_k^{(n,h)} \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \le eSAH^2 + 10c\sqrt{eSAC^{(n,h)}H^5\iota}.$$

On the other hand, according to the definition of $w_k^{(n,h)}$,

$$\sum_{k=1}^{K} w_k^{(n,h)} \big(Q_h^k - Q_h^*\big)(x_h^k, a_h^k) \ge \big(2^{n-1}\Delta_{\min}\big) \cdot C^{(n,h)}.$$

Combining these two sides, we obtain the following inequality of $C^{(n,h)}$:

$$\begin{aligned} &\big(2^{n-1}\Delta_{\min}\big) C^{(n,h)} \le eSAH^2 + 10c\sqrt{eSAC^{(n,h)}H^5\iota} \\ &\Rightarrow C^{(n,h)} \le \mathcal{O}\left(\frac{H^5 SA\iota}{4^n \Delta_{\min}^2}\right). \end{aligned}$$

Finally, we observe that

$$C^{(n)} = \sum_{h=1}^{H} C^{(n,h)} \le \mathcal{O}\left(\frac{H^6 SA\iota}{4^n \Delta_{\min}^2}\right),$$

which is exactly the statement of Lemma 4.2. □

5 Conclusion and Future Directions

This paper gives the first logarithmic regret bounds for Q-learning in both finite-horizon and discounted tabular MDPs. Below we list some future directions that we believe are worth exploring.

H dependence The dependency on H in our regret bound for episodic RL is H^6, which we believe is sub-optimal. As discussed in Simchowitz and Jamieson (2019), improving the H dependence is often a challenging task. Recently, Zhang et al. (2020) showed a model-free algorithm can achieve near-optimal regret in the worst case using the idea of reference value function. It would be interesting to apply this idea to improve the H dependence in our logarithmic regret bound.

Function Approximation Lastly, we note that recently researchers found the sub-optimality gap assumption is crucial for dealing with large state-space RL problems where function approximation is needed. Du et al. (2019c) presented an algorithm that enjoys polynomial sample complexity if there is a sub-optimality gap and the environment satisfies a low-variance assumption. Du et al. (2019b, 2020) further showed this assumption is necessary in certain settings. There is another line of works putting certain low-rank assumptions on MDPs (Krishnamurthy et al., 2016; Jiang et al., 2017; Dann et al., 2018; Du et al., 2019a; Sun et al., 2019; Misra et al., 2020). It would be interesting to extend our analysis to these settings and obtain logarithmic regret bounds.

Acknowledgments

LY acknowledges the support from the Simons Institute for the Theory of Computing at UC Berkeley (Theory of Reinforcement Learning).

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. On the optimality of sparse model-based planning for Markov decision processes. *arXiv preprint arXiv:1906.03804*, 2019.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. In *Advances in Neural Information Processing Systems*, pages 1422–1432, 2018.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1507–1516, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019a.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019b.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8058–8068, 2019c.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic Q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. *arXiv preprint arXiv:2002.07125*, 2020.

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

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Sham M Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for Q-learning and indirect algo-

rithms. In *Advances in neural information processing systems*, pages 996–1002, 1999.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

Gen Li, Yuing Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*, 2020.

Shuang Liu and Hao Su. Regret bounds for discounted MDPs. *arXiv preprint arXiv:2002.05138*, 2020.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8874–8882, 2018.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *ArXiv*, abs/1608.02732, 2016.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196, 2018a.

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, 2018b.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1151–1160, 2019.

Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

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.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Ambuj Tewari. *Reinforcement learning in large or unknown MDPs*. University of California, Berkeley, 2007.

Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*, 2019.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, pages 5626–5635, 2019.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition, 2020.