Nearly Minimax Optimal Reinforcement Learning for Discounted MDPs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the reinforcement learning problem for discounted Markov Decision Processes (MDPs) under the tabular setting. We propose a model-based algorithm named UCBVI-γ, which is based on the *optimism in the face of uncertainty principle* and the Bernstein-type bonus. We show that UCBVI-γ achieves an $\widetilde{O}\big(\sqrt{SAT}/(1-\gamma)^{1.5}\big)$ regret, where S is the number of states, A is the number of actions, γ is the discount factor and T is the number of steps. In addition, we construct a class of hard MDPs and show that for any algorithm, the expected regret is at least $\widetilde{\Omega}\big(\sqrt{SAT}/(1-\gamma)^{1.5}\big)$. Our upper bound matches the minimax lower bound up to logarithmic factors, which suggests that UCBVI-γ is nearly minimax optimal for discounted MDPs.

1 Introduction

The goal of reinforcement learning (RL) is designing algorithms to learn the optimal policy through interactions with the unknown dynamic environment. Markov decision process (MDPs) plays a central role in reinforcement learning due to their ability to describe the time-independent state transition property. More specifically, the discounted MDP is one of the standard MDPs in reinforcement learning to describe sequential tasks without interruption or restart. For discounted MDPs, with a *generative model* [12], several algorithms with near-optimal sample complexity have been proposed. More specifically, Azar et al. [3] proposed an Empirical QVI algorithm which achieves the optimal sample complexity to find the optimal value function. Sidford et al. [22] proposed a sublinear randomized value iteration algorithm that achieves a near-optimal sample complexity to find the optimal policy, and Sidford et al. [23] further improved it to reach the optimal sample complexity. Since generative model is a powerful oracle that allows the algorithm to query the reward function and the next state for any state-action pair (s, a), it is natural to ask whether there exist online RL algorithms (without generative model) that achieve optimality.

To measure an online RL algorithm, a widely used notion is *regret*, which is defined as the summation of sub-optimality gaps over time steps. The regret is firstly introduced for episodic and infinite-horizon average-reward MDPs and later extended to discounted MDPs by [15, 30, 35, 35]. Liu and Su [15] proposed a double Q-learning algorithm with the UCB exploration (Double Q-learning), which enjoys $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{2.5})$ regret, where S is the number of states, A is the number of actions, γ is the discount factor and T is the number of steps. While Double Q-learning enjoys a standard $\sqrt{T}$-regret, it still does not match the lower bound proved in [15] in terms of the dependence on S, A and $1/(1-\gamma)$. Recently, Zhou et al. [34] proposed a UCLK$^+$ algorithm for discounted MDPs with the linear mixture MDP assumption which achieves $\widetilde{O}\big(d\sqrt{T}/(1-\gamma)^{1.5}\big)$ regret, where d

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

is the dimension of the feature mapping. However, directly applying their algorithm to our setting would yield an $\widetilde{O}\big(S^2A\sqrt{T}/(1-\gamma)^{1.5}\big)$ regret [1], which is even worse in the dependence on S, A.

In this paper, we aim to close this gap by designing a practical algorithm with a nearly optimal regret. In particular, we propose a model-based algorithm named UCBVI-γ for discounted MDPs without using the generative model. At the core of our algorithm is to use a "refined" Bernstein-type bonus and the *law of total variance* [3, 4], which together can provide tighter upper confidence bound (UCB). Our contributions are summarized as follows:

- We propose a model-based algorithm UCBVI-γ to learn the optimal value function under the discounted MDP setting. We show that the regret of UCBVI-γ in first T steps is upper bounded by $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{1.5})$. Our regret bound strictly improves the best existing regret $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{2.5})$[2] in [15] by a factor of $(1-\gamma)^{-1}$.
- We also prove a lower bound of the regret by constructing a class of hard-to-learn discounted MDPs, which can be regarded as a *chain* of the hard MDPs considered in [15]. We show that for any algorithm, its regret in the first T steps can not be lower than $\widetilde{\Omega}(\sqrt{SAT}/(1-\gamma)^{1.5})$ on the constructed MDP. This lower bound also strictly improves the lower bound $\Omega(\sqrt{SAT}/(1-\gamma) + \sqrt{AT}/(1-\gamma)^{1.5})$ proved by [15].
- The nearly matching upper and the lower bounds together suggest that the proposed UCBVI-γ algorithm is minimax-optimal up to logarithmic factors.

We compare the regret of UCBVI-γ with previous online algorithms for learning discounted MDPs in Table 1.

Notation For any positive integer n, we denote by $[n]$ the set $\{1, \dots, n\}$. For any two numbers a and b, we denote by $a \vee b$ as the shorthand for $\max(a, b)$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C such that $a_n \le Cb_n$, and we write $a_n = \Omega(b_n)$ if there exists an absolute constant C such that $a_n \ge Cb_n$. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to further hide the logarithmic factors.

2 Related Work

Model-free Algorithms for Discounted MDPs. A large amount of reinforcement learning algorithms like Q-learning can be regarded as model-free algorithms. These algorithms directly learn the action-value function by updating the values of each state-action pair. Kearns and Singh [13] firstly proposed a phased Q-Learning which learns an ϵ-optimal policy with $\widetilde{O}(SA/((1-\gamma)^7\epsilon^2))$ sample complexity for $\epsilon \le 1/(1-\gamma)$. Later on, Strehl et al. [25] proposed a delay-Q-learning algorithm, which achieves $\widetilde{O}(SA/((1-\gamma)^8\epsilon^4))$ sample complexity of exploration. Wang [29] proposed a randomized primal-dual method algorithm, which improves the sample complexity to $\widetilde{O}(SA/((1-\gamma)^4\epsilon^2))$ for $\epsilon \le 1/(1-\gamma)$ under the ergodicity assumption. Later, Sidford et al. [23] proposed a sublinear randomized value iteration algorithm and achieved $\widetilde{O}(SA/((1-\gamma)^4\epsilon^2))$ sample complexity for $\epsilon \le 1$. Sidford et al. [22] further improved the empirical QVI algorithm and proposed a variance-reduced QVI algorithm, which improves the sample complexity to $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ for $\epsilon \le 1$. Wainwright [28] proposed a variance-reduced Q-learning algorithm, which is an extension of the Q-learning algorithm and achieves $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ sample complexity. In addition, Dong et al. [9] proposed an infinite Q-learning with UCB and improved the sample complexity of exploration to $\widetilde{O}(SA/((1-\gamma)^7\epsilon^2))$. Zhang et al. [33] proposed a UCB-multistage algorithm which attains the $\widetilde{O}(SA/((1-\gamma)^{5.5}\epsilon^2))$ sample complexity of exploration, and proposed a UCB-multistage-adv algorithm which attains a better sample complexity $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ in the high accuracy regime. Recently, Liu and Su [15] focused on regret minimization for the infinite-horizon discounted MDP and showed the connection between regret and sample complexity of exploration. Liu and Su [15]

[1] Linear mixture MDP assumes that there exists a feature mapping $\boldsymbol{\phi}(s'|s,a) \in \mathbb{R}^d$ and a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta} \rangle$. It can be verified that any MDP is automatically a linear mixture MDP with a S^A-dimensional feature mapping [2, 35].

[2] The regret definition in [15] differs from our definition by a factor of $(1-\gamma)^{-1}$. Here we translate their regret from their definition to our definition for a fair comparison. A detailed comparison can be found in Appendix.

Table 1: Comparison of RL algorithms for discounted MDPs in terms of sample complexity and regret. Note that the regret bounds for all the compared algorithms except Double Q-learning [15] are derived from their sample complexity results. See Appendix A.1 for more details.

	Algorithm	Sample complexity	Regret
Model-free	Delay-Q-learning [25]	$\widetilde{O}\left(\frac{SA}{(1-\gamma)^8\epsilon^4}\right)$	$\widetilde{O}\left(\frac{S^{1/5}A^{1/5}T^{4/5}}{(1-\gamma)^{9/5}}\right)$
	Q-learning with UCB [9]	$\widetilde{O}\left(\frac{SA}{(1-\gamma)^7\epsilon^2}\right)$	$\widetilde{O}\left(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{8/3}}\right)$
	UCB-multistage [33]	$\widetilde{O}\left(\frac{SA}{(1-\gamma)^{5.5}\epsilon^2}\right)$	$\widetilde{O}\left(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{13/6}}\right)$
	UCB-multistage-adv [33]	$\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$[3]	$\widetilde{O}\left(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{4/3}}\right)$
	Double Q-learning [15]	N/A	$\widetilde{O}\left(\frac{\sqrt{SAT}}{(1-\gamma)^{2.5}}\right)$
Model-based	R-max [5]	$\widetilde{O}\left(\frac{S^2A}{(1-\gamma)^6\epsilon^3}\right)$	$\widetilde{O}\left(\frac{S^{1/2}A^{1/4}T^{3/4}}{(1-\gamma)^{7/4}}\right)$
	MoRmax [27]	$\widetilde{O}\left(\frac{SA}{(1-\gamma)^6\epsilon^2}\right)$	$\widetilde{O}\left(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{7/3}}\right)$
	UCRL [14]	$\widetilde{O}\left(\frac{S^2A}{(1-\gamma)^3\epsilon^2}\right)$	$\widetilde{O}\left(\frac{S^{2/3}A^{1/3}T^{2/3}}{(1-\gamma)^{4/3}}\right)$
	UCBVI-γ **(Our work)**	N/A	$\widetilde{O}\left(\frac{\sqrt{SAT}}{(1-\gamma)^{1.5}}\right)$
Lower bound	N/A	$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$ [14]	$\widetilde{\Omega}\left(\frac{\sqrt{SAT}}{(1-\gamma)^{1.5}}\right)$ **(Our work)**

2. It holds when $\epsilon \leq 1/\text{poly}(S, A, 1/(1-\gamma))$.

proposed a Double Q-Learning algorithm, which achieves $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{2.5})$ regret within T steps. Furthermore, Liu and Su [15] constructed a series of hard MDPs and showed that the expected regret for any algorithm is lower bounder by $\widetilde{\Omega}\big(\sqrt{SAT}/(1-\gamma) + \sqrt{AT}/(1-\gamma)^{1.5}\big)$. There still exists a $1/(1-\gamma)$-gap between the upper and lower regret bounds. In contrast to the aforementioned model-free algorithms, our proposed algorithm is model-based.

Model-based Algorithms for Discounted MDP. Our UCBVI-γ falls into the category of model-based reinforcement learning algorithms. Model-based algorithms maintain a model of the environment and update it based on the observed data. They will form the policy based on the learnt model. More specifically, to learn the ϵ-optimal value function, Azar et al. [3] proposed an empirical QVI algorithm which achieves $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ sample complexity. Azar et al. [3] proposed an empirical QVI algorithm which improves the sample complexity to $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ for $\epsilon \leq 1/\sqrt{(1-\gamma)S}$. Szita and Szepesvári [27] proposed an MoRmax algorithm, which achieves $\widetilde{O}(SA/((1-\gamma)^6\epsilon^2))$ sample complexity. Later, Lattimore and Hutter [14] proposed a UCRL algorithm, which achieves $\widetilde{O}(S^2A/((1-\gamma)^3\epsilon^2))$ sample complexity in general and $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ sample complexity with a strong assumption on the state transition. Recently, Agarwal et al. [1] proposed a refined analysis for the empirical QVI algorithm which achieves $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ sample complexity when $\epsilon \leq 1/\sqrt{1-\gamma}$.

Upper and Lower Bounds for Episodic MDPs. There is a line of work which aims at proving sample complexity or regret for episodic MDPs (MDPs which consist of restarting episodes) [7, 18, 4, 19, 11, 8, 24, 21, 31, 32, 17, 20]. Compared with the episodic MDP, discounted MDPs involve only one infinite-horizon sample trajectory, suggesting that any two states or actions on the trajectory are dependent. Such a dependence makes the learning of discounted MDPs more challenging.

3 Preliminaries

We consider infinite-horizon discounted Markov Decision Processes (MDP) which are defined by a tuple $(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$. Here $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$,

$\gamma \in (0,1)$ is the discount factor, $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, $\mathbb{P}(s'|s,a)$ is the transition probability function, which denotes the probability that state s transfers to state s' with action a. For simplicity, we assume the reward function is *deterministic and known*. A *non-stationary policies* π is a collection of function $\{\pi_t\}_{t=1}^{\infty}$, where each function $\pi_t : \{\mathcal{S} \times \mathcal{A}\}^{t-1} \times \mathcal{S} \to \mathcal{A}$ maps history $\{s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t = s\}$ to an action. For any non-stationary policy π, we denote $\pi_t(s) = \pi_t(s; s_1, a_1, ..., s_{t-1}, a_{t-1})$ for simplicity. We define the action-value function and value function at step t as follows:

$$Q_t^{\pi}(s,a) = \mathbb{E}\Big[\sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i}) \Big| s_1, ..., s_t = s, a_t = a\Big],$$

$$V_t^{\pi}(s) = \mathbb{E}\Big[\sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i}) \Big| s_1, ..., s_t = s\Big],$$

where $a_{t+i} = \pi_{t+i}(s_{t+i})$, and $s_{t+i+1} \sim \mathbb{P}\big(\cdot\,|s_{t+i}, \pi_{t+i}(s_{t+i})\big)$. In addition, we denote the optimal action-value function and the optimal value function as $Q^*(s,a) = \sup_{\pi} Q_1^{\pi}(s,a)$ and $V^*(s) = \sup_{\pi} V_1^{\pi}(s)$ respectively. Note that the optimal action-value function and the optimal value function are independent of the step t. For simplicity, for any function $V : \mathcal{S} \to R$, we denote $[\mathbb{P}V](s,a) = \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} V(s')$. According to the definition of the value function, we have the following non-stationary Bellman equation and Bellman optimality equation for non-stationary policy π and optimal policy π^*:

$$Q_t^{\pi}(s,a) = r(s,a) + \gamma[\mathbb{P}V_{t+1}^{\pi}](s,a), \; Q^*(s,a) = r(s,a) + \gamma[\mathbb{P}V^*](s,a). \tag{3.1}$$

4 Main Results

4.1 Algorithm

In this subsection, we propose the Upper Confidence Bound Value Iteration-γ (UCBVI-γ) algorithm, which is illustrated in Algorithm 1. The algorithm framework of UCBVI-γ follows the UCBVI algorithm proposed in Azar et al. [4], which can be regarded as the counterpart of UCBVI-γ in the episodic MDP setting.

UCBVI-γ is a model-based algorithm that maintains an empirical measure $\mathbb{P}_t$ at each step t. At the beginning of the t-th iteration, UCBVI-γ takes action a_t based on the greedy policy induced by $Q_t(s_t, a)$ and transits to the next state s_{t+1}. After receiving the next state s_{t+1}, UCBVI-γ computes the empirical transition probability function $\mathbb{P}_t(s'|s,a)$ in (4.1). Based on empirical transition probability function $\mathbb{P}_t(s'|s,a)$, UCBVI-γ updates $Q_{t+1}(s,a)$ by performing one-step value iteration on $Q_t(s,a)$ with an additional upper confidence bound (UCB) term $\text{UCB}_t(s,a)$ defined in (4.3). Here the UCB bonus term is used to measure the uncertainty of the expectation of the value function $V_t(s)$. Unlike previous work, which adapts a Hoeffding-type bonus [15], our UCBVI-γ uses a Bernstein-type bonus which brings a tighter upper bound by accessing the variance of $V_t(s)$, denoted by $\text{Var}_{s' \sim \mathbb{P}(\cdot|,s,a)} V_t(s')$. However, since the probability transition $\mathbb{P}(\cdot|s,a)$ is unknown, it is impossible to calculate the exact variance of V_t. Instead, UCBVI-γ estimates the variance by considering the variance of V_t over the empirical probability transition function $\mathbb{P}_t(\cdot|s,a)$ defined in (4.1). Therefore, the final UCB bonus term in (4.3) can be regarded as a standard Bernstein-type bonus on the empirical measure $\mathbb{P}_t(\cdot|s,a)$ with an additional error term.

Compared with UCBVI algorithm in Azar et al. [4], the action-value function $Q_t(s,a)$ in UCBVI-γ is updated in a forward way from step 1 to step T with the initial value $Q_1(s,a) = 1/(1-\gamma)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, while UCBVI updates its action-value function in a backward way from $Q_{t,H}$ to $Q_{t,1}$ with initial value $Q_{t,H}(s,a) = 0$. Compared with UCRL in Lattimore and Hutter [14], UCBVI-γ does not need to call an additional extended value iteration sub-procedure [10, 26], which is not easy to implement even with infinite computation [14].

4.2 Regret Analysis

In this subsection, we provide the regret bound of UCBVI-γ. We first give the formal definition of the regret for the discounted MDP setting.

Definition 4.1. For a given non-stationary policy π, we define the regret $\text{Regret}(T)$ as follow:

$$\text{Regret}(T) = \sum_{t=1}^{T} \big[V^*(s_t) - V_t^{\pi}(s_t)\big].$$

Algorithm 1 Upper Confidence Value-iteration UCBVI-γ

1: Receive state s_1 and set initial value function $Q_1(s,a) \leftarrow 1/(1-\gamma)$, $N_0(s,a) = N_0(s,a,s') = N_0(s) \leftarrow 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$

2: **for** step $t = 1, \ldots$ **do**

3: Let $\pi_t(\cdot) \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_t(\cdot, a)$, take action $a_t \leftarrow \pi_t(s_t)$ and receive next state $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$

4: Set $N_t(s) \leftarrow N_{t-1}(s)$, $N_t(s,a) \leftarrow N_{t-1}(s,a)$ and $N_t(s,a,s') \leftarrow N_{t-1}(s,a,s')$ for all $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$

5: Update $N_t(s_t) \leftarrow N_t(s_t) + 1, N_t(s_t, a_t) \leftarrow N_t(s_t, a_t) + 1$ and $N_t(s_t, a_t, s_{t+1}) \leftarrow N_t(s_t, a_t, s_{t+1}) + 1$

6: For all $s \in \mathcal{S}, a \in \mathcal{A}$, set

$$\mathbb{P}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a) \vee 1}. \tag{4.1}$$

7: Update new value function $Q_{t+1}(s,a)$ and $V_{t+1}(s)$ by

$$\begin{aligned} Q_{t+1}(s,a) &= \min\left\{Q_t(s,a), r(s,a) + \gamma[\mathbb{P}_t V_t](s,a) + \gamma \mathrm{UCB}_t(s,a)\right\}, \\ V_{t+1}(s) &= \max_{a\in\mathcal{A}} Q_{t+1}(s,a). \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \mathrm{UCB}_t(s,a) = &\sqrt{\frac{8U\mathrm{Var}_{s'\sim\mathbb{P}_t(\cdot,s,a)}(V_t(s'))}{N_t(s,a)\vee 1}} + \frac{8U/(1-\gamma)}{N_t(s,a)\vee 1} \\ &+ \sqrt{\frac{8\sum_{s'}\mathbb{P}_t(s'|s,a)\min\left\{100B_t(s'), 1/(1-\gamma)^2\right\}}{N_t(s,a)\vee 1}}, \end{aligned} \tag{4.3}$$

and $B_t(s') = \beta/\left[(1-\gamma)^5\left(N_t(s')\vee 1\right)\right]$.

8: **end for**

The same regret has been used in prior work [30, 35, 34] on discounted MDPs. It is related to the "sample complexity of exploration" [12, 14, 9]. For more details about the connection between the regret and the sample complexity, please refer to Appendix A.

With Definition 4.1, we introduce our main theorem, which gives an upper bound on the regret for UCBVI-γ.

Theorem 4.2. Let $U = \log(40SAT^3\log^2 T/(\delta(1-\gamma)^2))$. If we set $\beta = S^2A^2U^5$ in UCBVI-γ, then with probability at least $1-\delta$, the regret of UCBVI-γ in Algorithm 1 is bounded by

$$\mathrm{Regret}(T) \le \frac{752S^2A^{1.5}U^{3.5}}{(1-\gamma)^{3.5}} + \frac{60U\sqrt{SAT}}{(1-\gamma)^{1.5}} + \frac{4\sqrt{TU}}{(1-\gamma)^2}.$$

Remark 4.3. Notice that when $T = \widetilde{\Omega}(S^3A^2/(1-\gamma)^4)$ and $SA = \Omega(1/(1-\gamma))$, the regret is bounded by $\widetilde{O}\big(\sqrt{SAT}/(1-\gamma)^{1.5}\big)$. In addition, since $\mathrm{Regret}(T) \le T/(1-\gamma)$ holds for any T, we have $\mathbb{E}[\mathrm{Regret}(T)] = \widetilde{O}\big(\sqrt{SAT}/(1-\gamma)^{1.5} + T\delta/1-\gamma\big)$. When choosing $\delta = 1/T$, we have $\mathbb{E}[\mathrm{Regret}(T)] = \widetilde{O}\big(\sqrt{SAT}/(1-\gamma)^{1.5}\big)$.

We also provide a regret lower bound, which suggests that our UCBVI-γ is nearly minimax optimal.

Theorem 4.4. Suppose $\gamma \ge 2/3$, $A \ge 30$ and $T \ge 100SAL/(1-\gamma)^4$, then for any algorithm, there exists an MDP such that

$$\mathbb{E}[\mathrm{Regret}(T)] \ge \frac{\sqrt{SAT}}{10000(1-\gamma)^{1.5}} - \frac{4\sqrt{STL}}{(1-\gamma)^{1.5}} - \frac{8S}{(1-\gamma)^2},$$

where $L = \log\left(300S^4T^2/(1-\gamma)\right)\log(10ST)$.

Remark 4.5. When T is large enough and $A = \widetilde{\Omega}(1)$, Theorem 4.4 suggests that the lower bound of regret is $\widetilde{\Omega}(\sqrt{SAT}/(1-\gamma)^{1.5})$. It can be seen that the regret of UCBVI-γ in Theorem 4.2 matches this lower bound up to logarithmic factors. Therefore, UCBVI-γ is nearly minimax optimal.

5 Proof of the Main Results

In this section, we provide the proofs of Theorems 4.2 and 4.4. The missing proofs are deferred to the appendix.

5.1 Proof of Theorem 4.2

In this subsection, we prove Theorem 4.2. For simplicity, let $\delta' = (1-\gamma)^2\delta/(80T\log^2 T)$, then $U = \log(SAT^2/\delta')$. We first present the following key lemma, which shows that the optimal value functions V^* and Q^* can be upper bounded by the estimated functions V_t and Q_t with high probability:

Lemma 5.1. With probability at least $1 - 64T\delta\log^2 T/(1-\gamma)^2$, for all $t \in [T], s \in \mathcal{S}, a \in \mathcal{A}$, we have $Q_t(s,a) \geq Q^*(s,a),\ V_t(s) \geq V^*(s)$.

Equipped with Lemma 5.1, we can decompose the regret of UCBVI-γ as follows:

$$\text{Regret}(T) \leq \sum_{t=1}^{T}\big[V_t(s_t) - V_t^{\pi}(s_t)\big] = \underbrace{\sum_{t=1}^{T}\big[Q_t(s_t,a_t) - Q_t^{\pi}(s_t,a_t)\big]}_{\text{Regret}'(T)},$$

where the inequality holds due to Lemma 5.1. Therefore, it suffices to bound $\text{Regret}'(T)$. We have

$$\begin{aligned}\text{Regret}'(T) &\leq \sum_{t=1}^{T}\Big(r(s_t,a_t) + \gamma[\mathbb{P}_{t-1}V_{t-1}](s_t,a_t) + \gamma\text{UCB}_{t-1}(s_t,a_t) - r(s_t,a_t) - \gamma[\mathbb{P}V_{t+1}^{\pi}](s_t,a_t)\Big)\\ &= \sum_{t=1}^{T}\Big(\gamma[\mathbb{P}_{t-1}V_{t-1}](s_t,a_t) + \gamma\text{UCB}_{t-1}(s_t,a_t) - \gamma[\mathbb{P}V_{t+1}^{\pi}](s_t,a_t)\Big),\end{aligned}$$

where the inequality holds due to the update rule (4.2) and the Bellman equation $Q_t^{\pi}(s_t,a_t) = r(s_t,a_t) + \gamma[\mathbb{P}V_{t+1}^{\pi}](s_t,a_t)$. We further have

$$\begin{aligned}&\sum_{t=1}^{T}\Big(\gamma[\mathbb{P}_{t-1}V_{t-1}](s_t,a_t) + \gamma\text{UCB}_{t-1}(s_t,a_t) - \gamma[\mathbb{P}V_{t+1}^{\pi}](s_t,a_t)\Big)\\ &= \underbrace{\sum_{t=1}^{T}\gamma(V_{t-1}(s_{t+1}) - V_{t+1}^{\pi}(s_{t+1}))}_{I_1} + \underbrace{\sum_{t=1}^{T}\gamma\big[(\mathbb{P}_{t-1}-\mathbb{P})(V_{t-1}-V^*)\big](s_t,a_t)}_{I_2}\\ &\quad + \underbrace{\sum_{t=1}^{T}\gamma[(\mathbb{P}_{t-1}-\mathbb{P})V^*](s_t,a_t)}_{I_3} + \underbrace{\sum_{t=1}^{T}\gamma\text{UCB}_{t-1}(s_t,a_t)}_{I_4} + I_5,\end{aligned} \tag{5.1}$$

where $I_5 = \sum_{t=1}^{T}\gamma\big[\mathbb{P}(V_{t-1} - V_{t+1}^{\pi})\big](s_t,a_t) - \gamma\big[V_{t-1}(s_{t+1}) - V_{t+1}^{\pi}(s_{t+1})\big]$. In the remaining of the proof, it suffices to bound terms I_1 to I_5 separately.

First, I_1 can be regarded as the difference between the estimated V_{t-1} and the value function V_{t+1}^{π} of policy π, and it can be bounded by the following lemma.

Lemma 5.2. For the term I_1, We have $I_1 \leq \gamma\text{Regret}'(T) + (2S+2)\gamma/1-\gamma$

Next, I_2 can be regarded as the "correction" term between the estimated V_{t-1} and the optimal value function V^*. It can be bounded by the following lemma.

Lemma 5.3. With probability at least $1 - 64T\delta\log^2 T/(1-\gamma)^2 - 3\delta$, we have

$$I_2 \leq (1-\gamma)\text{Regret}'(T)/2 + \sqrt{2T\log(1/\delta)} + \frac{5S^2A\log(ST/\delta)\log(3T)}{(1-\gamma)^2}.$$

In addition, I_3 can be regarded as the error between the empirical probability distribution $\mathbb{P}_{t-1}$ and the true transition probability $\mathbb{P}$. Note that V^* is a fixed value function that does not have any randomness. Therefore, I_3 can be bounded through the standard concentration inequalities, and its upper bound is presented in the following lemma.

Lemma 5.4. With probability at least $1 - 2\delta - \delta/(1-\gamma)$, we have

$$I_3 \leq \frac{2SAU^2}{1-\gamma} + U\sqrt{2SA}\sqrt{\frac{5T}{1-\gamma} + \frac{29U}{3(1-\gamma)^3} + \frac{2\text{Regret}'(T)}{1-\gamma} + \frac{\sqrt{2TU}}{(1-\gamma)^2}}.$$

Furthermore, I_4 can be regarded as the summation of the UCB terms, which is also the dominating term of the total regret. It can be bounded by the following lemma.

Lemma 5.5. With probability at least $1 - 4\delta - \delta/(1-\gamma)$, we have

$$I_4 \leq \frac{37S^2A^{1.5}U^{3.5}}{(1-\gamma)^{2.5}} + U\sqrt{8SA}\sqrt{\frac{5T}{1-\gamma} + \frac{29U}{3(1-\gamma)^3} + \frac{2\text{Regret}'(T)}{1-\gamma} + \frac{12SU\sqrt{AT}}{(1-\gamma)^2}}.$$

Finally, I_5 is the summation of a martingale difference sequence. By Azuma-Hoeffding inequality, with probability at least $1-\delta$, we have

$$I_5 \leq \frac{\sqrt{2T\log(1/\delta)}}{1-\gamma}. \tag{5.2}$$

Substituting the upper bounds of terms I_1 to I_5 from Lemma 5.2 to Lemma 5.5, as well as (5.2), into (5.1), and taking a union bound to let all the events introduced in Lemma 5.2 to Lemma 5.5 and (5.2) hold, we have with probability at least $1 - 20TU^2\delta/(1-\gamma)^2$, the following inequality holds:

$$(1-\gamma)\text{Regret}'(T) \leq \frac{160S^2A^{1.5}U^{3.5}}{(1-\gamma)^{2.5}} + \frac{54U\sqrt{SAT}}{\sqrt{1-\gamma}} + \frac{2\sqrt{2TU}}{1-\gamma} + 12U\sqrt{\frac{SA\text{Regret}'(T)}{1-\gamma}}. \tag{5.3}$$

Using the fact that $x \leq a + b\sqrt{x} \Rightarrow x \leq 1.1a + 4b^2$, (5.3) can be further bounded as follows

$$\begin{aligned}\text{Regret}(T) &\leq \text{Regret}'(T)\\ &\leq \frac{752S^2A^{1.5}U^{3.5}}{(1-\gamma)^{3.5}} + \frac{60U\sqrt{SAT}}{(1-\gamma)^{1.5}} + \frac{4\sqrt{TU}}{(1-\gamma)^2}.\end{aligned}$$

This completes our proof.

5.2 Proof of Theorem 4.4

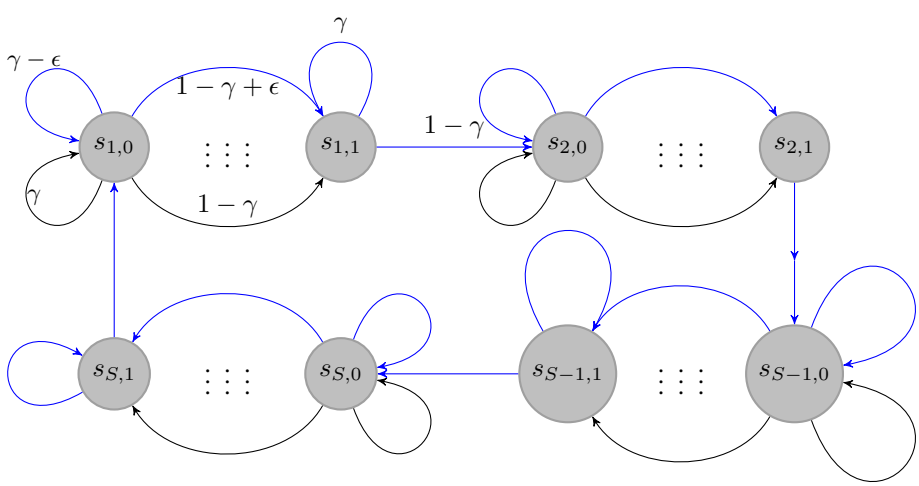

Figure 1: A class of hard-to-learn MDPs considered in Theorem 4.4. The MDP can be regarded as a combination of S two-state MDPs, each of which is an MDP illustrated on the top-left corner. In addition, the i-th two-state MDP has the a_i^*-th action as its optimal action. The blue arrows represent the optimal actions in different states. $\epsilon = \sqrt{A(1-\gamma)/K}/24$.

In this subsection, we provide the proof of Theorem 4.4. The proof of the lower bound is based on constructing a class of hard MDPs. Specifically, the state space $\mathcal{S}$ consists of $2S$ states $\{s_{i,0}, s_{i,1}\}_{i\in[S]}$

and the action space $\mathcal{A}$ contains A actions. The reward function r satisfies that $r(s_{i,0}, a) = 0$ and $r(s_{i,1}, a) = 1$ for any $a \in \mathcal{A}, i \in [S]$. The probability transition function $\mathbb{P}$ is defined as follows.

$$\mathbb{P}(s_{i,1}|s_{i,0}, a) = 1 - \gamma + \mathbb{1}_{a=a_i^*} \frac{1}{24}\sqrt{\frac{A(1-\gamma)}{K}}, \mathbb{P}(s_{i,1}|s_{i,1}, a) = \gamma,$$
$$\mathbb{P}(s_{i,0}|s_{i,0}, a) = \gamma - \mathbb{1}_{a=a_i^*} \frac{1}{24}\sqrt{\frac{A(1-\gamma)}{K}}, \mathbb{P}(s_{i+1,0}|s_{i,1}, a) = 1 - \gamma,$$

where we assume $s_{S+1,0} = s_{1,0}$ for simplicity and a_i^* is the optimal action for state $s_{i,0}$. The MDP is illustrated in Figure 1, which can be regarded as S copies of the "single" two-state MDP arranged in a circle. The two-state MDP is the same as that proposed in [15]. Each of the two-state MDP has two states and one "optimal" action a_i^* satisfied $\mathbb{P}(s_{i,1}|s_{i,0}, a_i^*) = 1 - \gamma + \epsilon$. Compared with the MDP instance in [10], both instances use S copies of a single MDP. However, unlike the MDP in [10] which only has one "optimal" action among all SA actions, our MDP which has in total S "optimal" actions, which makes it harder to analyze.

Now we begin to prove our lower bound. Let $\mathbb{E}_{\mathbf{a}^*}[\cdot]$ denote the expectation conditioned on one fixed selection of $\mathbf{a}^* = (a_1^*, \ldots, a_S^*)$. We introduce a shorthand notation $\mathbb{E}^*$ to denote $\mathbb{E}^*[\cdot] = 1/A^S \cdot \sum_{\mathbf{a}^* \in \mathcal{A}^S} \mathbb{E}_{\mathbf{a}^*}[\cdot]$. Here $\mathbb{E}^*$ is the average value of expectation over the randomness from MDP defined by different optimal actions. From now on, we aim to lower bound $\mathbb{E}^*[\text{Regret}(T)]$, since once $\mathbb{E}^*[\text{Regret}(T)]$ is lower bounded, $\mathbb{E}[\text{Regret}(T)]$ can be lower bounded by selecting $a_1^*, \ldots, a_S^*$ which maximizes $\mathbb{E}[\text{Regret}(T)]$. We set $T = 10SK$ in the following proof. Based on the definition of $\mathbb{E}^*$, we have the following lemma.

Lemma 5.6. The expectated regret $\mathbb{E}^*[\text{Regret}(T)]$ can be lower bounded as follows:

$$\mathbb{E}^*[\text{Regret}(T)] \geq \mathbb{E}^*\left[\sum_{t=1}^{T} V^*(s_t) - \frac{r(s_t, a_t)}{1-\gamma}\right] - \frac{4}{(1-\gamma)^2}.$$

By Lemma 5.6, it suffices to lower bound $\sum_{t=1}^{T}[V^*(s_t) - r(s_t, a_t)/(1-\gamma)]$, which is $\text{Regret}^{\text{Liu}}(T)$ defined in [15]. When an agent visits the state set $\{s_{j,0}, s_{j,1}\}$ for the i-th time, we denote the state in $\{s_{j,0}, s_{j,1}\}$ it visited as $X_{j,i}$, and the following action selected by the agent as $A_{j,i}$. Let T_j be the number of steps for the agent staying in $\{s_{j,0}, s_{j,1}\}$ in the total T steps. Then the regret can be further decomposed as follows:

$$\mathbb{E}^*\left[\sum_{t=1}^{T} V^*(s_t) - \frac{r(s_t, a_t)}{1-\gamma}\right] = \sum_{j=1}^{S} \mathbb{E}^*\left[\sum_{i=1}^{T_j} V^*(X_{j,i}) - \frac{r(X_{j,i}, A_{j,i})}{1-\gamma}\right] = I_1 + I_2 + I_3,$$

where

$$I_1 = \sum_{j=1}^{S} \mathbb{E}^*\left[\sum_{i=1}^{K} V^*(X_{j,i}) - \frac{r(X_{j,i}, A_{j,i})}{1-\gamma}\right],$$
$$I_2 = \sum_{j=1}^{S} \mathbb{E}^*\left[\sum_{i=K+1}^{T_j} V^*(X_{j,i}) - \frac{r(X_{j,i}, A_{j,i})}{1-\gamma}\middle| T_j > K\right] \cdot \mathbb{P}^*[T_j > K],$$
$$I_3 = -\sum_{j=1}^{S} \mathbb{E}^*\left[\sum_{i=T_j+1}^{K} V^*(X_{j,i}) - \frac{r(X_{j,i}, A_{j,i})}{1-\gamma}\middle| T_j < K\right] \cdot \mathbb{P}^*[T_j < K].$$

Note that I_1 essentially represents the regret over S two-state MDPs in their first K steps, and it can be lower bounded through the following lemma.

Lemma 5.7. If $K \geq 10SA/(1-\gamma)^4$, then for each $j \in [S]$, we have

$$\mathbb{E}^*\left[\sum_{i=1}^{K}(1-\gamma)V^*(X_{j,i}) - r(X_{j,i}, A_{j,i})\right] \geq \frac{\sqrt{AK}}{2304\sqrt{1-\gamma}} - \frac{1}{1-\gamma}.$$

This lemma shows that the expected regret of first K steps on states $s_{j,0}$ and $s_{j,1}$ is at least $\widetilde{\Omega}\big(\sqrt{AK}/(1-\gamma)^{0.5} - 1/(1-\gamma)\big)$. Therefore by Lemma 5.7, we have

$$I_1 = \sum_{j=1}^{S} \mathbb{E}^*\left[\sum_{i=1}^{K} V^*(X_{j,i}) - \frac{r(X_{j,i}, A_{j,i})}{1-\gamma}\right] \geq \frac{\sqrt{SAT}}{2304\sqrt{10}(1-\gamma)^{1.5}} - \frac{S}{(1-\gamma)^2}. \tag{5.4}$$

To bound I_2, we need the following lemma.

Lemma 5.8. With probability at least $1-2ST\delta\log T/(1-\gamma)$, for each $j\in[S]$ and $K+1\le t\le T$, we have

$$\sum_{i=K+1}^{t} V^*(X_{j,i}) - \frac{r(X_{j,i},A_{j,i})}{1-\gamma} \ge -\frac{\sqrt{2t\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4}{(1-\gamma)^2}.$$

Lemma 5.8 gives a crude lower bound of I_2. Taking expectation over Lemma 5.8 and taking summation over all states, we have

$$\begin{aligned}
I_2 &\ge \sum_{j=1}^{S}\mathbb{E}^*\bigg[\bigg(-\frac{\sqrt{2T_j\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4}{(1-\gamma)^2}\bigg)\bigg|T_j>K\bigg]\mathbb{P}^*[T_j>K] - \sum_{j=1}^{S} T\cdot\frac{2ST\delta\log T}{1-\gamma}\\
&\ge \sum_{j=1}^{S}\mathbb{E}^*\bigg[-\frac{\sqrt{2T_j\log(1/\delta)\log T}}{(1-\gamma)^{1.5}}\bigg] - \frac{4S}{(1-\gamma)^2} - \frac{2S^2T^2\delta\log T}{1-\gamma}\\
&\ge \sum_{j=1}^{S} -\frac{\sqrt{2\mathbb{E}^*[T_j]\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4S}{(1-\gamma)^2} - \frac{2S^2T^2\delta\log T}{1-\gamma}\\
&\ge -\frac{\sqrt{2ST\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4S}{(1-\gamma)^2} - \frac{2S^2T^2\delta\log T}{1-\gamma},
\end{aligned} \tag{5.5}$$

where the first inequality holds due to Lemma 5.8, the second inequality holds since $1-2ST\delta\log T/(1-\gamma)\le 1$ and $\mathbb{E}[-X|Y]\mathbb{P}(Y)\ge\mathbb{E}[-X]$ when $X\ge 0$, the third inequality holds due to Jensen's inequality and the fact that $\sqrt{x}$ is a concave function, and the last inequality holds due to Jensen's inequality and the fact that $\sum_{j=1}^{S}\mathbb{E}^*[T_j]=T$. To bound I_3, we need the following lemma, which suggests that when K is large enough, $T_i>K$ happens with high probability:

Lemma 5.9. When $K\ge 10A\log(1/\delta)/(1-\gamma)^4$, with probability at least $1-2S\delta$, for all $i\in[S]$, we have $T_i>K$.

Notice that the difference of transition probability between the optimal action and suboptimal actions is $\sqrt{A(1-\gamma)}/24K$. In this case, when T is large enough, T_i is close to $T/S=10K$. Thus I_3 can be lower bounded as follows:

$$I_3 \ge -\sum_{j=1}^{S}\frac{K}{1-\gamma}\mathbb{P}^*[T_j<K] \ge -\frac{ST\delta}{5(1-\gamma)}, \tag{5.6}$$

where the first inequality holds due to $0\le r(X_{j,i},A_{j,i})\le 1$ and the second inequality holds due to Lemma 5.9. Finally, setting $\delta=1/\big(4ST^2(1-\gamma)\log T\big)$, we can verify that the requirements of K in Lemma 5.7 and Lemma 5.9 hold when T satisfies $T\ge 100SAL/(1-\gamma)^4$, and $L=\log\big(300S^4T^2/((1-\gamma)\delta)\big)\log T$. Therefore, substituting $\delta=1/\big(4ST^2(1-\gamma)\log T\big)$ into (5.5) and (5.6), and combining (5.4), (5.5), (5.6) and Lemma 5.6, we have

$$\mathbb{E}[\text{Regret}(T)] \ge \frac{\sqrt{SAT}}{10000(1-\gamma)^{1.5}} - \frac{4\sqrt{STL}}{(1-\gamma)^{1.5}} - \frac{8S}{(1-\gamma)^2},$$

which completes the proof of Theorem 4.4.

6 Conclusions and Future Work

We proposed UCBVI-γ, an online RL algorithm for discounted tabular MDPs. We show that the regret of UCBVI-γ can be upper bounded by $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{1.5})$ and we prove a matching lower bound on the expected regret $\widetilde{\Omega}(\sqrt{SAT}/(1-\gamma)^{1.5})$. There is still a gap between the upper and lower bounds when $T\le\max\{S^3A/(1-\gamma)^4, SA/(1-\gamma)\}$, and we leave it as an open problem for future work.

References

[1] Agarwal, A., Kakade, S. and Yang, L. F. (2019). Model-based reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:1906.03804* .

[2] AYOUB, A., JIA, Z., SZEPESVARI, C., WANG, M. and YANG, L. F. (2020). Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107* .

[3] AZAR, M. G., MUNOS, R. and KAPPEN, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning* **91** 325–349.

[4] AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

[5] BRAFMAN, R. I. and TENNENHOLTZ, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* **3** 213–231.

[6] CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[7] DANN, C. and BRUNSKILL, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*.

[8] DANN, C., LI, L., WEI, W. and BRUNSKILL, E. (2019). Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*. PMLR.

[9] DONG, K., WANG, Y., CHEN, X. and WANG, L. (2019). Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311* .

[10] JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11** 1563–1600.

[11] JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*.

[12] KAKADE, S. M. ET AL. (2003). *On the sample complexity of reinforcement learning*. Ph.D. thesis, University of London London, England.

[13] KEARNS, M. J. and SINGH, S. P. (1999). Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in neural information processing systems*.

[14] LATTIMORE, T. and HUTTER, M. (2012). Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*. Springer.

[15] LIU, S. and SU, H. (2020). Regret bounds for discounted mdps.

[16] MAURER, A. and PONTIL, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* .

[17] NEU, G. and PIKE-BURKE, C. (2020). A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891* .

[18] OSBAND, I. and VAN ROY, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732* .

[19] OSBAND, I. and VAN ROY, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*.

[20] PACCHIANO, A., BALL, P., PARKER-HOLDER, J., CHOROMANSKI, K. and ROBERTS, S. (2020). On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911* .

[21] RUSSO, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*.

[22] SIDFORD, A., WANG, M., WU, X., YANG, L. F. and YE, Y. (2018). Near-optimal time and sample complexities for for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492* .

[23] SIDFORD, A., WANG, M., WU, X. and YE, Y. (2018). 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*. SIAM.

[24] SIMCHOWITZ, M. and JAMIESON, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*.

[25] STREHL, A. L., LI, L., WIEWIORA, E., LANGFORD, J. and LITTMAN, M. L. (2006). Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*.

[26] STREHL, A. L. and LITTMAN, M. L. (2008). An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences* **74** 1309–1331.

[27] SZITA, I. and SZEPESVÁRI, C. (2010). Model-based reinforcement learning with nearly tight exploration complexity bounds .

[28] WAINWRIGHT, M. J. (2019). Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697* .

[29] WANG, M. (2017). Randomized linear programming solves the discounted markov decision problem in nearly-linear running time. *arXiv preprint arXiv:1704.01869* .

[30] YANG, K., YANG, L. and DU, S. (2021). Q-learning with logarithmic regret 1576–1584.

[31] ZANETTE, A. and BRUNSKILL, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210* .

[32] ZHANG, Z., ZHOU, Y. and JI, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019* .

[33] ZHANG, Z., ZHOU, Y. and JI, X. (2020). Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864* .

[34] ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *COLT*.

[35] ZHOU, D., HE, J. and GU, Q. (2021). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss the limitations and potential future works in Section 6

(c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work provides a theoretical analysis of discounted MDPs, and there is no potential negative social impact.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]