Nearly Minimax Optimal Reinforcement Learning for Discounted MDPs

Jiafan He [1] **Dongruo Zhou** [1] **Quanquan Gu** [1]

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* (Kakade et al., 2003), several algorithms with near-optimal sample complexity have been proposed. More specifically, Azar et al. (2013) proposed an Empirical QVI algorithm which achieves the optimal sample complexity to find the optimal value function. Sidford et al. (2018a) proposed a sublinear randomized value iteration algorithm that achieves a near-optimal sample complexity to find the optimal policy, and Sidford et al. (2018b) 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 suboptimality gaps over time steps. The regret is firstly introduced for episodic and infinite-horizon average-reward MDPs and later extended to discounted MDPs by (Liu and Su, 2020; Yang et al., 2021; Zhou et al., 2021b;b). Liu and Su (2020) 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 (Liu and Su, 2020) in terms of the dependence on S, A and $1/(1-\gamma)$. Recently, Zhou et al. (2021a) proposed a UCLK$^+$ algorithm for discounted MDPs under the linear mixture MDP assumption and achieved $\widetilde{O}\big(d\sqrt{T}/(1-\gamma)^{1.5}\big)$ regret, where d 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, which is even worse that of double Q-learning (Liu and Su, 2020) in terms of 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* (Azar et al., 2013; 2017), 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 $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{2.5})$[1] in (Liu and Su, 2020) by a factor

[1] Department of Computer Science, University of California, Los Angeles, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] The regret definition in (Liu and Su, 2020) 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.

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 (Liu and Su, 2020). 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 (Liu and Su, 2020).

- 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 $\{1,\ldots,n\}$. 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 $a_n \leq Cb_n$, and we write $a_n = \Omega(b_n)$ if there exists an absolute constant C such that $a_n \geq Cb_n$. 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 (1999) firstly proposed a phased Q-Learning which learns an ϵ-optimal policy with $\widetilde{O}(SA/((1-\gamma)^7\epsilon^2))$ sample complexity for $\epsilon \leq 1/(1-\gamma)$. Later on, Strehl et al. (2006) proposed a delay-Q-learning algorithm, which achieves $\widetilde{O}(SA/((1-\gamma)^8\epsilon^4))$ sample complexity of exploration. Wang (2017) proposed a randomized primal-dual method algorithm, which improves the sample complexity to $\widetilde{O}(SA/((1-\gamma)^4\epsilon^2))$ for $\epsilon \leq 1/(1-\gamma)$ under the ergodicity assumption. Later, Sidford et al. (2018b) proposed a sublinear randomized value iteration algorithm and achieved $\widetilde{O}(SA/((1-\gamma)^4\epsilon^2))$ sample complexity for $\epsilon \leq 1$. Sidford et al. (2018a) 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 $\epsilon \leq 1$. Wainwright (2019) 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. (2019) 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. (2020b) 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 (2020) focused on regret minimization for the infinite-horizon discounted MDP and showed the connection between regret and sample complexity of exploration. Liu and Su (2020) proposed a Double Q-Learning algorithm, which achieves $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{2.5})$ regret within T steps. Furthermore, Liu and Su (2020) 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. (2013) proposed an empirical QVI algorithm which achieves $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ sample complexity. Azar et al. (2013) 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 (2010) proposed an MoRmax algorithm, which achieves $\widetilde{O}(SA/((1-\gamma)^6\epsilon^2))$ sample complexity. Later, Lattimore and Hutter (2012) 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. (2019) proposed a refined analysis for the empirical QVI algorithm which achieves $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ sample complexity when $\epsilon \leq 1/\sqrt{1-\gamma}$.

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, $r:\mathcal{S}\times\mathcal{A}\rightarrow[0,1]$ 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 $\pi_t:\{\mathcal{S}\times\mathcal{A}\}^{t-1}\times\mathcal{S}\rightarrow\mathcal{A}$ 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 de-

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 (Liu and Su, 2020) are derived from their sample complexity results. See Appendix B.1 for more details.

	Algorithm	Sample complexity	Regret
Model-free	Delay-Q-learning (Strehl et al., 2006)	$\widetilde{O}\Big(\frac{SA}{(1-\gamma)^8\epsilon^4}\Big)$	$\widetilde{O}\Big(\frac{S^{1/5}A^{1/5}T^{4/5}}{(1-\gamma)^{9/5}}\Big)$
	Q-learning with UCB (Dong et al., 2019)	$\widetilde{O}\Big(\frac{SA}{(1-\gamma)^7\epsilon^2}\Big)$	$\widetilde{O}\Big(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{8/3}}\Big)$
	UCB-multistage (Zhang et al., 2020b)	$\widetilde{O}\Big(\frac{SA}{(1-\gamma)^{5.5}\epsilon^2}\Big)$	$\widetilde{O}\Big(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{13/6}}\Big)$
	UCB-multistage-adv (Zhang et al., 2020b)	$\widetilde{O}\Big(\frac{SA}{(1-\gamma)^3\epsilon^2}\Big)$[2]	$\widetilde{O}\Big(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{4/3}}\Big)$
	Double Q-learning (Liu and Su, 2020)	N/A	$\widetilde{O}\Big(\frac{\sqrt{SAT}}{(1-\gamma)^{2.5}}\Big)$
Model-based	R-max (Brafman and Tennenholtz, 2002)	$\widetilde{O}\Big(\frac{S^2A}{(1-\gamma)^6\epsilon^3}\Big)$	$\widetilde{O}\Big(\frac{S^{1/2}A^{1/4}T^{3/4}}{(1-\gamma)^{7/4}}\Big)$
	MoRmax (Szita and Szepesvári, 2010)	$\widetilde{O}\Big(\frac{SA}{(1-\gamma)^6\epsilon^2}\Big)$	$\widetilde{O}\Big(\frac{S^{1/3}A^{1/3}T^{2/3}}{(1-\gamma)^{7/3}}\Big)$
	UCRL (Lattimore and Hutter, 2012)	$\widetilde{O}\Big(\frac{S^2A}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{O}\Big(\frac{S^{2/3}A^{1/3}T^{2/3}}{(1-\gamma)^{4/3}}\Big)$
	UCBVI-γ **(Our work)**	N/A	$\widetilde{O}\Big(\frac{\sqrt{SAT}}{(1-\gamma)^{1.5}}\Big)$
Lower bound	N/A	$\widetilde{\Omega}\Big(\frac{SA}{(1-\gamma)^3\epsilon^2}\Big)$ (Lattimore and Hutter, 2012)	$\widetilde{\Omega}\Big(\frac{\sqrt{SAT}}{(1-\gamma)^{1.5}}\Big)$ **(Our work)**

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

fine the action-value function and value function at step t as follows:

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

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

where $a_{t+i} = \pi_{t+i}(s_{t+i})$, and $s_{t+i+1} \sim \mathbb{P}(\cdot|s_{t+i}, \pi_{t+i}(s_{t+i}))$. 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^*:

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

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. (2017), 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 (C.1). 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 (Liu and Su, 2020), our UCBVI-γ uses a Bernstein-type bonus which

Algorithm 1 Upper Confidence Value-iteration UCBVI-γ

```
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 $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$
for step $t = 1, \dots$ do
    Let $\pi_t(\cdot) \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_t(\cdot, a)$, take action $a_t \leftarrow \pi_t(s_t)$ and receive next state $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$
    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 $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$
    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 $N_t(s_t, a_t, s_{t+1}) \leftarrow N_t(s_t, a_t, s_{t+1}) + 1$
    For all $s \in \mathcal{S}, a \in \mathcal{A}$, set
        $\mathbb{P}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a) \vee 1}.$    (4.1)
    Update new value function $Q_{t+1}(s,a)$ and $V_{t+1}(s)$ by
        $Q_{t+1}(s,a) = \min\big\{Q_t(s,a), r(s,a) + \gamma[\mathbb{P}_t V_t](s,a) + \gamma \text{UCB}_t(s,a)\big\}, V_{t+1}(s) = \max_{a \in \mathcal{A}} Q_{t+1}(s,a).$    (4.2)
    where $\text{UCB}_t(s,a)$ is denoted as (C.1)
end for
```

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 (C.1) can be regarded as a standard Bernstein-type bonus on the empirical measure $\mathbb{P}_t(\cdot|s,a)$ with an additional error term.

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 Regret(T) as follow:

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

The same regret has been used in prior work (Yang et al., 2021; Zhou et al., 2021b;a) on discounted MDPs. It is related to the "sample complexity of exploration" (Kakade et al., 2003; Lattimore and Hutter, 2012; Dong et al., 2019). For more details about the connection between the regret and the sample complexity, please refer to Appendix B.

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

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

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

Theorem 4.3. Suppose $\gamma \geq 2/3, A \geq 30$ and $T \geq 100SAL/(1-\gamma)^4$, then for any algorithm, there exists an MDP such that

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

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

Remark 4.4. When T is large enough and $A = \widetilde{\Omega}(1)$, Theorem 4.3 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. 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 $T \leq \max\{S^3A/(1-\gamma)^4, SA/(1-\gamma)\}$, and we leave it as an open problem for future work.

References

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* .

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.

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.

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.

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

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

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

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* .

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

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

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

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*.

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

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

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

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

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

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

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* .

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

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

SIDFORD, A., WANG, M., WU, X. and YE, Y. (2018b). 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.

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

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*.

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.

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

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

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

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

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* .

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

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

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

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

A. Other Related Works

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) (Dann and Brunskill, 2015; Osband and Van Roy, 2016; Azar et al., 2017; Osband and Van Roy, 2017; Jin et al., 2018; Dann et al., 2019; Simchowitz and Jamieson, 2019; Russo, 2019; Zanette and Brunskill, 2019; Zhang et al., 2020a; Neu and Pike-Burke, 2020; Pacchiano et al., 2020). 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.

Comparison Algorithm 1 with other Algorithms Compared with UCBVI algorithm in Azar et al. (2017), 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 (2012), UCBVI-γ does not need to call an additional extended value iteration sub-procedure (Jaksch et al., 2010; Strehl and Littman, 2008), which is not easy to implement even with infinite computation (Lattimore and Hutter, 2012).

B. More Discussions on the Regret and Sample Complexity

B.1. Converting Sample Complexity of Exploration to Regret

In this subsection, we shows the relationship between the sample complexity of exploration and the regret.

The definition of regret in Defintion 4.1 is related to the “sample complexity of exploration” $N(\epsilon, \delta)$ (Kakade et al., 2003; Lattimore and Hutter, 2012; Dong et al., 2019), which is the upper bound on the number of steps t such that $V^*(s_t) - V_t^\pi(s_t) \geq \epsilon$ with probability at least $1-\delta$. Compared with the regret, sample complexity of exploration focuses on the sub-optimalities at all steps t, rather than the first T steps, and ignores the small sub-optimalities. Though both metrics have been used to describe the performance of an algorithm, these two metrics are not directly comparable. More specifically, algorithms with fewer but larger sub-optimalities will have a small sample complexity of exploration but a high regret. In contrast, algorithms with a lot of moderate sub-optimalities will have a high sample complexity of exploration but a low regret.

By the definition of the sample complexity exploration $N(\epsilon, \delta)$, with probability at least $1-\delta$, the number of steps t where $V^*(s_t) - V_t^\pi(s_t) \geq \epsilon$ is upper bounded by $N(\epsilon, \delta)$. Thus, for the regret within T steps, we have following inequality:

$$\begin{aligned}\text{Regret}(T) &= \sum_{t=1}^{T} \big[V^*(s_t) - V_t^\pi(s_t)\big] \\ &= \sum_{t\in[T], V^*(s_t)-V_t^\pi(s_t)\geq\epsilon} \big[V^*(s_t) - V_t^\pi(s_t)\big] + \sum_{t\in[T], V^*(s_t)-V_t^\pi(s_t)<\epsilon} \big[V^*(s_t) - V_t^\pi(s_t)\big] \\ &\leq \frac{N(\epsilon,\delta)}{1-\gamma} + T\epsilon, \end{aligned} \tag{B.1}$$

where the inequality holds due to the definition of $N(\epsilon, \delta)$. Furthermore,if an algorithm achieve sample complexity $N(\epsilon,\delta) = O(B\epsilon^{-\alpha})$, then we can choose $\epsilon = T^{-1/(\alpha+1)}(1-\gamma)^{1/(\alpha+1)}B^{-1/(\alpha+1)}$ to minimize the (B.1). Thus, we have

$$\begin{aligned}\text{Regret}(T) &\leq \frac{N(\epsilon,\delta)}{1-\gamma} + T\epsilon \\ &= O\Big(\frac{B\epsilon^{-\alpha}}{1-\gamma} + T\epsilon\Big) \\ &= O\big(B^{1/(\alpha+1)}(1-\gamma)^{-1/(\alpha+1)}T^{\alpha/(\alpha+1)}\big). \end{aligned} \tag{B.2}$$

Furthermore, the best result in sample complexity of exploration (Zhang et al., 2020b) achieves $\widetilde{O}\Big(SA/\big((1-\gamma)^3\epsilon^2\big)\Big)$ sample complexity and this result implies $\widetilde{O}(S^{1/3}A^{1/3}(1-\gamma)^{-4/3}T^{2/3})$ regret, which is worse than our result by a $T^{1/6}$ factor.

B.2. Comparison with the Regret in (Liu and Su, 2020)

Our definition is similar to that of Liu and Su (2020). Note that Liu and Su (2020) define the regret as $\text{Regret}^{\text{Liu}}(T) = \sum_{t=1}^{T} \Delta_t$, where $\Delta_t = (1-\gamma)V^*(s_t) - r(s_t, a_t)$. Comparing the definition in Liu and Su (2020) with our definition, we can show that $(1-\gamma)\text{Regret}(T) \approx \text{Regret}^{\text{Liu}}(T)$ since

$$(1-\gamma)\sum_{t=1}^{T} V_t^{\pi}(s_t) \approx (1-\gamma)\sum_{t=1}^{T}\sum_{i=0}^{\infty}\gamma^i r(s_{t+i}, a_{t+i}) \approx \sum_{t=1}^{T} r(s_t, a_t),$$

where the first approximate equality holds due to Azuma-Hoeffding inequality and the second approximate equality holds due to $0 \le r(s,a) \le 1$. Therefore, our regret definition is equivalent to that in (Liu and Su, 2020) up to a $1-\gamma$ factor.

C. Proof of the Main Results

In this section, we provide the proofs of Theorems 4.2 and 4.3.

C.1. Proof of Theorem 4.2

In this subsection, we prove Theorem 4.2 and we first denote the definition of bonus term $\text{UCB}_t(s,a)$ in Algorithm 1:

$$\text{UCB}_t(s,a) = \sqrt{\frac{8U\text{Var}_{s'\sim\mathbb{P}_t(\cdot,s,a)}(V_t(s'))}{N_t(s,a)\vee 1}} + \frac{8U/(1-\gamma)}{N_t(s,a)\vee 1} + \sqrt{\frac{8\sum_{s'}\mathbb{P}_t(s'|s,a)\min\big\{100B_t(s'), 1/(1-\gamma)^2\big\}}{N_t(s,a)\vee 1}}, \quad \text{(C.1)}$$

where $B_t(s') = \beta/\big[(1-\gamma)^5\big(N_t(s')\vee 1\big)\big]$. 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 C.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 $Q_t(s,a) \ge Q^*(s,a),\ V_t(s) \ge V^*(s)$.

Equipped with Lemma C.1, we can decompose the regret of UCBVI-γ as follows:

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

where the inequality holds due to Lemma C.1. Therefore, it suffices to bound $\text{Regret}'(T)$. We have

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

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

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

$$+\underbrace{\sum_{t=1}^{T}\gamma[(\mathbb{P}_{t-1}-\mathbb{P})V^*](s_t,a_t)}_{I_3}+\underbrace{\sum_{t=1}^{T}\gamma\text{UCB}_{t-1}(s_t,a_t)}_{I_4}+I_5, \tag{C.2}$$

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

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

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

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 C.4. With probability at least $1-2\delta-\delta/(1-\gamma)$, we have

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

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 C.5. With probability at least $1-4\delta-\delta/(1-\gamma)$, we have

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

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

$$I_5\le\frac{\sqrt{2T\log(1/\delta)}}{1-\gamma}. \tag{C.3}$$

Substituting the upper bounds of terms I_1 to I_5 from Lemma C.2 to Lemma C.5, as well as (C.3), into (C.2), and taking a union bound to let all the events introduced in Lemma C.2 to Lemma C.5 and (C.3) hold, we have with probability at least $1-20TU^2\delta/(1-\gamma)^2$, the following inequality holds:

$$(1-\gamma)\text{Regret}'(T)\le\frac{160S^2A^{1.5}U^{3.5}}{(1-\gamma)^{2.5}}+\frac{54U\sqrt{SAT}}{\sqrt{1-\gamma}}+\frac{2\sqrt{2TU}}{1-\gamma}+12U\sqrt{\frac{SA\text{Regret}'(T)}{1-\gamma}}. \tag{C.4}$$

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

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

This completes our proof.

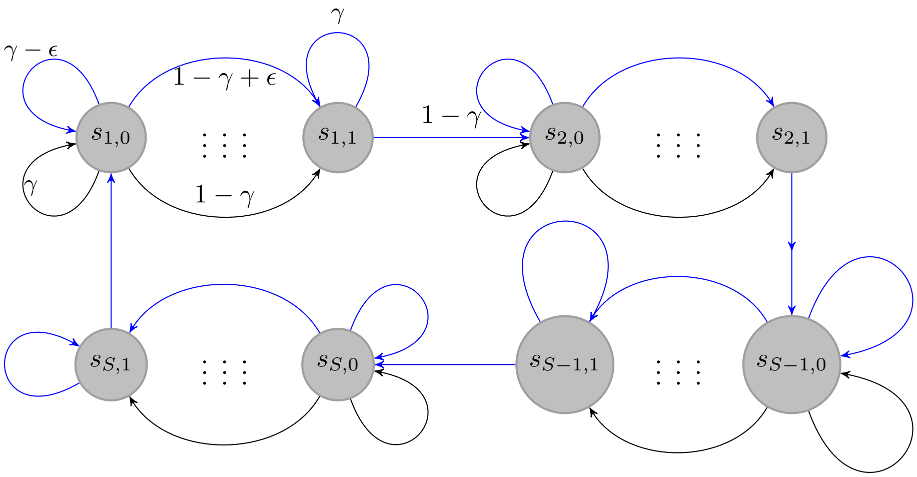

Figure 1. A class of hard-to-learn MDPs considered in Theorem 4.3. The MDP can be regarded as a combination of S two-state MDPs, each of which is an MDP illustrated on the top-left corner. In addition, the i-th two-state MDP has the a_i^*-th action as its optimal action. The blue arrows represent the optimal actions in different states. $\epsilon = \sqrt{A(1-\gamma)/K}/24$.

C.2. Proof of Theorem 4.3

In this subsection, we provide the proof of Theorem 4.3. 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 (Liu and Su, 2020). 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 (Jaksch et al., 2010), both instances use S copies of a single MDP. However, unlike the MDP in (Jaksch et al., 2010) 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 C.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 C.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 (Liu and Su, 2020). 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:

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

where

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

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 C.7. If $K \ge 10SA/(1-\gamma)^4$, then for each $j \in [S]$, we have

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

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 C.7, we have

$$I_1 = \sum_{j=1}^{S}\mathbb{E}^*\bigg[\sum_{i=1}^{K} V^*(X_{j,i}) - \frac{r(X_{j,i},A_{j,i})}{1-\gamma}\bigg] \ge \frac{\sqrt{SAT}}{2304\sqrt{10}(1-\gamma)^{1.5}} - \frac{S}{(1-\gamma)^2}. \tag{C.5}$$

To bound I_2, we need the following lemma.

Lemma C.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 C.8 gives a crude lower bound of I_2. Taking expectation over Lemma C.8 and taking summation over all states, we have

$$\begin{aligned} I_2 &\ge \sum_{j=1}^{S}\mathbb{E}^*\bigg[\bigg(-\frac{\sqrt{2T_j\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4}{(1-\gamma)^2}\bigg)\bigg| T_j > K\bigg]\mathbb{P}^*[T_j > K] - \sum_{j=1}^{S} T \cdot \frac{2ST\delta\log T}{1-\gamma} \\ &\ge \sum_{j=1}^{S}\mathbb{E}^*\bigg[-\frac{\sqrt{2T_j\log(1/\delta)\log T}}{(1-\gamma)^{1.5}}\bigg] - \frac{4S}{(1-\gamma)^2} - \frac{2S^2T^2\delta\log T}{1-\gamma} \\ &\ge \sum_{j=1}^{S} -\frac{\sqrt{2\mathbb{E}^*[T_j]\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4S}{(1-\gamma)^2} - \frac{2S^2T^2\delta\log T}{1-\gamma} \\ &\ge -\frac{\sqrt{2ST\log(1/\delta)\log T}}{(1-\gamma)^{1.5}} - \frac{4S}{(1-\gamma)^2} - \frac{2S^2T^2\delta\log T}{1-\gamma}, \end{aligned} \tag{C.6}$$

where the first inequality holds due to Lemma C.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 C.9. When $K \geq 10A\log(1/\delta)/(1-\gamma)^4$, 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:

$$I_3 \geq -\sum_{j=1}^{S} \frac{K}{1-\gamma}\mathbb{P}^*[T_j < K] \geq -\frac{ST\delta}{5(1-\gamma)}, \tag{C.7}$$

where the first inequality holds due to $0 \leq r(X_{j,i}, A_{j,i}) \leq 1$ and the second inequality holds due to Lemma C.9. Finally, setting $\delta = 1/\big(4ST^2(1-\gamma)\log T\big)$, we can verify that the requirements of K in Lemma C.7 and Lemma C.9 hold when T satisfies $T \geq 100SAL/(1-\gamma)^4$, 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 (C.6) and (C.7), and combining (C.5), (C.6), (C.7) and Lemma C.6, we have

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

which completes the proof of Theorem 4.3.

D. Proof of Lemmas in Section C.1

In this section, we prove Lemma C.1 to Lemma C.5. For simplicity, we introduce the following shorthand notations:

$$\begin{aligned}
\mathbb{V}^*(s,a) &= \text{Var}_{s'\sim\mathbb{P}(\cdot|s,a)}\big(V^*(s')\big),\\
\mathbb{V}^\pi_t(s,a) &= \text{Var}_{s'\sim\mathbb{P}(\cdot|s,a)}\big(V^\pi_{t+1}(s')\big),\\
\mathbb{V}_t(s,a) &= \text{Var}_{s'\sim\mathbb{P}_t(\cdot|s,a)}(V_t(s')),\\
\mathbb{V}^*_t(s,a) &= \text{Var}_{s'\sim\mathbb{P}_t(\cdot|s,a)}(V^*(s')).
\end{aligned}$$

We start with a list of technical lemmas that will be used to prove Lemma C.1 to Lemma C.5. We first provide the Azuma-Hoeffding and Bernstein inequalities.

Lemma D.1 (Azuma–Hoeffding inequality, Cesa-Bianchi and Lugosi 2006)**.** Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \leq M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Then for any $0 < \delta < 1$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^n x_i \leq M\sqrt{2n\log(1/\delta)}.$$

Lemma D.2 (Bernstein inequality, Cesa-Bianchi and Lugosi 2006)**.** Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \leq M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Suppose that

$$\sum_{i=1}^n \mathbb{E}(x_i^2|\mathcal{G}_i) \leq v$$

for some constant v. Then for any $\delta > 0$, with probability at least $1-\delta$,

$$\sum_{i=1}^n x_i \leq \sqrt{2v\log(1/\delta)} + \frac{2M\log(1/\delta)}{3}.$$

The following first lemma provides basic inequalities for the summations of counted numbers $N_i(s_i, a_i)$ and $N_i(s_i)$.

Lemma D.3. For all $t \in [T]$ and subset $\mathcal{C} \subseteq [T]$, we have

$$\sum_{i=1}^t \frac{1}{N_{i-1}(s_i, a_i) \vee 1} \leq SA\log(3T),$$

$$\sum_{i=1}^{t} \frac{1}{N_{i-1}(s_i) \vee 1} \leq S\log(3T),$$
$$\sum_{i \in \mathcal{C}} \frac{1}{\sqrt{N_{i-1}(s_i, a_i) \vee 1}} \leq \sqrt{SA\log(3T)|\mathcal{C}|}.$$

Next lemma upper bounds the difference between the empirical measure $\mathbb{P}_{t-1}$ and $\mathbb{P}$, with respect to the true variance of the optimal value function $\mathbb{V}^*(s, a)$.

Lemma D.4. If $0 \leq V^*(s) \leq 1/(1-\gamma)$ for all $s \in \mathcal{S}$, then with probability at least $1-\delta$, for all $t \in [T], s \in \mathcal{S}, a \in \mathcal{A}$, we have

$$\big[(\mathbb{P}_t - \mathbb{P})V^*\big](s, a) \leq \sqrt{\frac{2\mathbb{V}^*(s, a)\log(SAT/\delta)}{N_{t-1}(s, a) \vee 1}} + \frac{2\log(SAT/\delta)}{3(1-\gamma)\big(N_{t-1}(s, a) \vee 1\big)}.$$

Similar to Lemma D.4, the following lemmas also upper bounds the difference between the empirical measure $\mathbb{P}_{t-1}$ and $\mathbb{P}$, but with respect to the estimated variance.

Lemma D.5 (Theorem 4 in Maurer and Pontil 2009)**.** Let $Z, Z_1, .., Z_n$ be i.i.d random variable with value in $[0, M]$ and let $\delta > 0$, then with probability at least $1-\delta$, we have

$$\mathbb{E}Z - \frac{1}{n}\sum_{i=1}^{n} Z_i \leq \sqrt{\frac{2\mathbb{V}_n Z \log(1/\delta)}{n}} + \frac{7M\log(1/\delta)}{3n},$$

where $\mathbb{V}_n Z$ is the estimated variance $\mathbb{V}_n Z = \sum_{1 \leq i < j \leq n}(Z_i - Z_j)^2/n(n-1)$.

Lemma D.6. If $0 \leq V^*(s) \leq 1/(1-\gamma)$ for all $s \in \mathcal{S}$, then with probability at least $1-\delta$, for all $t \in [T], s \in \mathcal{S}, a \in \mathcal{A}$, we have

$$\big[(\mathbb{P} - \mathbb{P}_t)V^*\big](s, a) \leq \sqrt{\frac{2\mathbb{V}^*_{t-1}(s, a)\log(SAT/\delta)}{N_{t-1}(s, a) \vee 1}} + \frac{7\log(SAT/\delta)}{3(1-\gamma)\big(N_{t-1}(s, a) \vee 1\big)}.$$

The next lemma shows that the total variance of the nonstationary policy π can be upper bounded by $O(T/(1-\gamma))$. It is worth noting that a trivial bound which bounds $\mathbb{V}^\pi_i(s_i, a_i)$ by $1/(1-\gamma)^2$ only gives an $O(T/(1-\gamma)^2)$ bound.

Lemma D.7. With probability at least $1 - \delta/(1-\gamma)$, we have

$$\gamma^2 \sum_{t=1}^{T} \mathbb{V}^\pi_t(s_t, a_t) \leq \frac{5T}{1-\gamma} + \frac{25\log(1/\delta)}{3(1-\gamma)^3}.$$

Based on previous concentration Lemma, we define the following high probability events and our proof of Lemma C.2 to Lemma C.5 relies on these high probability events. Let $\mathcal{E}$ denote the event when the conclusion of Lemma C.1 holds. Then by Lemma C.1, we have $\Pr(\mathcal{E}) \geq 1 - 64T\delta\log^2 T/(1-\gamma)^2$. We also define the following event:

$$\mathcal{E}_1 = \bigg\{\big[(\mathbb{P}_t - \mathbb{P})V^*\big](s, a) \leq \sqrt{\frac{2\mathbb{V}^*(s, a)\log(SAT/\delta)}{N_{t-1}(s, a) \vee 1}} + \frac{2\log(SAT/\delta)}{3(1-\gamma)\big(N_{t-1}(s, a) \vee 1\big)}, \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]\bigg\},$$

$$\mathcal{E}_2 = \bigg\{\big[(\mathbb{P} - \mathbb{P}_t)V^*\big](s, a) \leq \sqrt{\frac{2\mathbb{V}^*_{t-1}(s, a)\log(SAT/\delta)}{N_{t-1}(s, a) \vee 1}} + \frac{7\log(SAT/\delta)}{3(1-\gamma)\big(N_{t-1}(s, a) \vee 1\big)}, \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]\bigg\},$$

$$\begin{aligned}
\mathcal{E}_3 &= \bigg\{\mathbb{P}_{t-1}(s'|s_t,a_t) - \mathbb{P}(s'|s_t,a_t) \le \sqrt{\frac{2\mathbb{P}(s'|s_t,a_t)(1-\mathbb{P}(s'|s_t,a_t))\mathrm{log}(ST/\delta)}{N_{t-1}(s_t,a_t)\vee 1}}, \\
&\qquad + \frac{2\log(ST/\delta)}{3\big(N_{t-1}(s_t,a_t)\vee 1\big)} \forall s\in\mathcal{S}, a\in\mathcal{A}, t\in[T]\bigg\}, \\
\mathcal{E}_4 &= \bigg\{\sum_{t=1}^{T}\mathbb{P}(s'|s_t,a_t)\big(V_{t-1}(s') - V^*(s')\big) \le \sum_{t=1}^{T}\big(V_{t-1}(s_{t+1}) - V^*(s_{t+1})\big) + \frac{\sqrt{2T\log(1/\delta)}}{1-\gamma}\bigg\}, \\
\mathcal{E}_5 &= \{\gamma^2\sum_{t=1}^{T}\mathbb{V}_t^{\pi}(s_t,a_t) \le \frac{5T}{1-\gamma} + \frac{25\log(1/\delta)}{3(1-\gamma)^3}\}, \\
\mathcal{E}_6 &= \bigg\{\sum_{t=1}^{T}\big[\mathbb{P}(V_{t-1} - V_{t+1}^{\pi})\big](s_t,a_t) - \sum_{t=1}^{T}\big[V_{t-1}(s_{t+1}) - V_{t+1}^{\pi}(s_{t+1})\big] \le \frac{\sqrt{2T\log(1/\delta)}}{1-\gamma}\bigg\}, \\
\mathcal{E}_7 &= \bigg\{\sum_{t=1}^{T}\big[\mathbb{P}(V^* - V_{t+1}^{\pi})\big](s_t,a_t) - \sum_{t=1}^{T}\big[V^*(s_{t+1}) - V_{t+1}^{\pi}(s_{t+1})\big] \le \frac{\sqrt{2T\log(1/\delta)}}{1-\gamma}\bigg\}, \\
\mathcal{E}_8 &= \bigg\{\big\|\mathbb{P}_{t-1}(\cdot|s,a) - \mathbb{P}(\cdot|s,a)\big\|_1 \le \frac{\sqrt{2S\mathrm{log}(T/\delta)}}{\sqrt{N_{t-1}(s,a)\vee 1}}, \forall s\in\mathcal{S}, a\in\mathcal{A}, t\in[T]\bigg\}, \\
\mathcal{E}_9 &= \bigg\{\sum_{t=1}^{T}\sum_{s'}\mathbb{P}(s'|s_t,a_t)\min\Big\{\frac{100S^2A^2U^5}{(1-\gamma)^5\big(N_{t-1}(s')\vee 1\big)}, \frac{1}{(1-\gamma)^2}\Big\} \\
&\qquad \le \sum_{t=1}^{T}\min\Big\{\frac{100S^2A^2U^5}{(1-\gamma)^5\big(N_{t-1}(s_{t+1})\vee 1\big)}, \frac{1}{(1-\gamma)^2}\Big\} + \frac{\sqrt{2TU}}{(1-\gamma)^2}\bigg\},
\end{aligned}$$

where $U = \log(40SAT^3\log^2 T/(\delta(1-\gamma)^2))$. For these high probability events, according to the Lemma D.1, we have $\Pr(\mathcal{E}_4) \ge 1-\delta, \Pr(\mathcal{E}_6) \ge 1-\delta, \Pr(\mathcal{E}_7) \ge 1-\delta, \Pr(\mathcal{E}_8) \ge 1-\delta, \Pr(\mathcal{E}_9) \ge 1-\delta$. According to the Lemma D.2, we have $\Pr(\mathcal{E}_3) \ge 1-\delta$. According to the Lemma D.4, we have $\Pr(\mathcal{E}_1) \ge 1-\delta$. According to the Lemma D.6, we have $\Pr(\mathcal{E}_2) \ge 1-\delta$. According to the Lemma D.7, we have $\Pr(\mathcal{E}_5) \ge 1-\delta/(1-\gamma)$.

The next lemma shows that the total difference between the optimal variance and the variance induced by π can be bounded in terms of $\mathrm{Regret}'(T)$.

Lemma D.8. On the event $\mathcal{E}_7$, we have

$$\sum_{i=1}^{T}\big(\mathbb{V}^*(s_i,a_i) - \mathbb{V}_i^{\pi}(s_i,a_i)\big) \le \frac{2\mathrm{Regret}'(T)}{1-\gamma} + \frac{2+\sqrt{2T\mathrm{log}(1/\delta)}}{(1-\gamma)^2}.$$

Similar to Lemma D.8, the next lemma shows that the total difference between the estimated variance and the variance induced by π can be upper-bounded in terms of $\mathrm{Regret}'(T)$.

Lemma D.9. On the event $\mathcal{E}_6 \cap \mathcal{E}_8$, we have

$$\sum_{i=1}^{T}\big(\mathbb{V}_{i-1}(s_i,a_i) - \mathbb{V}_i^{\pi}(s_i,a_i)\big) \le \frac{2\mathrm{Regret}'(T)}{1-\gamma} + \frac{9S\sqrt{2AT\mathrm{log}(T/\delta)\log(3T)}}{(1-\gamma)^2}.$$

D.1. Proof of Lemma C.1

For simplicity, we denote $U = \log(SAT^2/\delta)$ and $H = \lfloor 2\log T/(1-\gamma)\rfloor + 1$ and for $h\in[H]$, we define

$$\mathrm{Regret}'(t,s,h) = \sum_{1\le i\le t, s_i = s}\gamma^h\big[V_{i+h}(s_{i+h}) - V_{i+h}^{\pi}(s_{i+h})\big].$$

Then we have the following lemma.

Lemma D.10. For each $t \in [T]$, with probability at least $1 - 4H^2\delta$, for all $s \in \mathcal{S}, h \in [H]$, we have

$$\text{Regret}'(t, s, h) \le \frac{16SAU^2\sqrt{N_t(s)}}{(1-\gamma)^{2.5}} + \frac{4S^2A^{1.5}U^3}{(1-\gamma)^{3.5}}.$$

In addition, if $N_t(s) > 0$, we have

$$V_t(s) - V^*(s) \le \frac{20SAU^2}{(1-\gamma)^{2.5}\sqrt{N_t(s)}}.$$

Now, we start the proof of Lemma C.1,

Proof of Lemma C.1. We prove this lemma by induction. At the first step $t = 1$, for all $s \in \mathcal{S}$, we have $V_1(s) = 1/(1-\gamma) \ge V^*(s)$. When Lemma C.1 holds for the first t steps, we consider for each $s \in \mathcal{S}, a \in \mathcal{A}$, then by the update rule (4.2), we have

$$Q_{t+1}(s,a) = \min\Big\{Q_t(s,a), r(s,a) + \gamma[\mathbb{P}_t V_t](s,a) + \gamma\text{UCB}_t(s,a)\Big\}.$$

If $Q_{t+1}(s,a) = Q_t(s,a)$, then by induction, we have

$$Q_{t+1}(s,a) \ge r(s,a) + \frac{8\gamma U}{1-\gamma} \ge r(s,a) + \gamma[\mathbb{P}V^*](s,a) = Q^*(s,a),$$

where the first inequality holds due to (4.2) in Algorithm 1 and the second inequality holds due to $0 \le V^*(s) \le 1/(1-\gamma)$. Otherwise, if $N_t(s,a) = 0$, then we have

$$Q_{t+1}(s,a) = Q_t(s,a) \ge Q^*(s,a).$$

When $N_t(s,a) > 0$, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&Q_{t+1}(s,a) - Q^*(s,a)\\
&= \gamma[\mathbb{P}_t V_t](s,a) + \gamma\text{UCB}_t(s,a) - \gamma[\mathbb{P}V^*](s,a)\\
&= \gamma\text{UCB}_t(s,a) + \gamma[(\mathbb{P}_t - \mathbb{P})V^*](s,a) + \gamma[\mathbb{P}_t(V_t - V^*)](s,a)\\
&\ge \gamma\text{UCB}_t(s,a) + \gamma[(\mathbb{P}_t - \mathbb{P})V^*](s,a)\\
&\ge \gamma\text{UCB}_t(s,a) - \gamma\sqrt{\frac{4\mathbb{V}_t^*(s,a)U}{N_t(s,a)\vee 1}} - \frac{8U\gamma}{(1-\gamma)\big(N_t(s,a)\vee 1\big)}\\
&\ge \gamma\sqrt{\frac{8\mathbb{V}_t(s,a)U}{N_t(s,a)\vee 1}} - \gamma\sqrt{\frac{4\mathbb{V}_t^*(s,a)U}{N_t(s,a)\vee 1}} + \gamma\sqrt{\frac{8\sum_{s'}\mathbb{P}_t(s'|s,a)\min\big\{100B_t(s'), 1/(1-\gamma)^2\big\}}{N_t(s,a)\vee 1}},
\end{aligned} \tag{D.1}$$

where the first inequality holds due to $V_t(s) \ge V^*(s)$, the second inequality holds due to Lemma D.6 and the third inequality holds due to the definition of UCB_t in (C.1). For the term $\mathbb{V}_t^*(s,a)$, we have

$$\begin{aligned}
\mathbb{V}_t^*(s,a) &= \mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\Big[\big(V^*(s') - \mathbb{E}[V^*(s')]\big)^2\Big]\\
&= \mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\Big[\big(V^*(s') - V_t(s') - \mathbb{E}[V^*(s') - V_t(s')] + V_t(s') - \mathbb{E}[V_t(s')]\big)^2\Big]\\
&\le 2\mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\Big[\big(V_t(s') - \mathbb{E}[V_t(s')]\big)^2\Big]\\
&\qquad + 2\mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\Big[\big(V^*(s') - V_t(s') - \mathbb{E}[V^*(s') - V_t(s')]\big)^2\Big]\\
&\le 2\mathbb{V}_t(s,a) + 2\mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\Big[\big(V^*(s') - V_t(s')\big)^2\Big],
\end{aligned} \tag{D.2}$$

where the first inequality holds due to $(x+y)^2 \le 2x^2+2y^2$ and the second inequality holds due to $\mathbb{E}\big[(X-\mathbb{E}[X])^2\big] \le \mathbb{E}[X^2]$. Substituting (D.2) into (D.1), with probability at least $1-4(t+1)H^2\delta$, we have

$$
\begin{aligned}
Q_{t+1}(s,a) - Q^*(s,a) &\ge \gamma\sqrt{\frac{8\mathbb{V}_t(s,a)U}{N_t(s,a)\vee 1}} + \gamma\sqrt{\frac{8\sum_{s'}\mathbb{P}_t(s'|s,a)\min\big\{100B_t(s'),1/(1-\gamma)^2\big\}}{N_t(s,a)\vee 1}} \\
&\quad - \gamma\sqrt{\frac{8\mathbb{V}_t(s,a)U + 8U\mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\big(V^*(s')-V_t(s')\big)^2}{N_t(s,a)\vee 1}} \\
&\ge \gamma\sqrt{\frac{8\sum_{s'}\mathbb{P}_t(s'|s,a)\min\big\{100B_t(s'),1/(1-\gamma)^2\big\}}{N_t(s,a)\vee 1}} \\
&\quad - \gamma\sqrt{\frac{8U\mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}\big(V^*(s')-V_t(s')\big)^2}{N_t(s,a)\vee 1}} \\
&\ge 0,
\end{aligned}
$$

where the first inequality holds due to (D.1), the second inequality holds due to (D.2), the third inequality holds due to $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, the last inequality holds due to Lemma D.10 with probability at least $1-4H^2\delta$ and induction hypothesis with probability at least $1-4tH^2\delta$. In addition, for all $s \in \mathcal{S}$, we have

$$
V_{t+1}(s) = \max_{a\in\mathcal{A}} Q_{t+1}(s,a) \ge \max_{a\in\mathcal{A}} Q^*(s,a) = V^*(s).
$$

Thus, by induction, we complete the proof of Lemma C.1. □

D.2. Proof of Lemma C.2

Proof of Lemma C.2. We have

$$
\begin{aligned}
&\sum_{t=1}^{T}\gamma\big(V_{t-1}(s_{t+1}) - V^{\pi}_{t+1}(s_{t+1})\big) \\
&\quad = \underbrace{\gamma\sum_{t=1}^{T}\big(V_{t-1}(s_{t+1}) - V_{t+1}(s_{t+1})\big)}_{I_1} + \underbrace{\gamma\sum_{t=1}^{T}\big(V_{t+1}(s_{t+1}) - V^{\pi}_{t+1}(s_{t+1})\big)}_{I_2}.
\end{aligned}
$$

For the term I_1, we have

$$
\begin{aligned}
\sum_{t=1}^{T}\gamma\big(V_{t-1}(s_{t+1}) - V_{t+1}(s_{t+1})\big) &\le \gamma\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\big[V_{t-1}(s) - V_{t+1}(s)\big] \\
&= \gamma\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\big[V_{t-1}(s) - V_{t+1}(s)\big] \\
&= \gamma\sum_{s\in\mathcal{S}}\big(V_0(s) + V_1(s) - V_T(s) - V_{T+1}(s)\big) \\
&\le \frac{2S\gamma}{1-\gamma},
\end{aligned}
\tag{D.3}
$$

where the first inequality holds due to $V_{t-1}(s) \ge V_{t+1}(s)$ by (4.2) in Algorithm 1, and the second inequality holds due to $0 \le V_t(s) \le 1/(1-\gamma)$. For the term I_2, we have

$$
\begin{aligned}
I_2 &= \gamma\sum_{t=2}^{T+1}\big(V_t(s_t) - V^{\pi}_t(s_t)\big) \\
&= \gamma\mathrm{Regret}'(T) + \gamma\big(V_{T+1}(s_{T+1}) - V^{\pi}_{T+1}(s_{T+1})\big) - \gamma\big(V_1(s_1) - V^{\pi}_1(s_1)\big)
\end{aligned}
$$

$$\leq \gamma \mathrm{Regret}'(T) + \frac{2\gamma}{1-\gamma}, \tag{D.4}$$

where the inequality holds due to $0 \leq V_t(s), V_t^{\pi}(s) \leq 1/(1-\gamma)$. Combining (D.3) and (D.4), we complete the proof of Lemma C.2. □

D.3. Proof of Lemma C.3

Proof of Lemma C.3. On the event $\mathcal{E}$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \gamma\big[(\mathbb{P}_{t-1} - \mathbb{P})(V_{t-1} - V^*)\big](s_t, a_t) \\
&= \gamma \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}} \big(\mathbb{P}_{t-1}(s'|s_t, a_t) - \mathbb{P}(s'|s_t, a_t)\big)\big(V_{t-1}(s') - V^*(s')\big) \\
&\leq \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}} \left[\sqrt{\frac{2\mathbb{P}(s'|s_t, a_t)(1 - \mathbb{P}(s'|s_t, a_t))\log(2ST/\delta)}{N_{t-1}(s_t, a_t) \vee 1}} + \frac{2\log(ST/\delta)}{3\big(N_{t-1}(s_t, a_t) \vee 1\big)}\right]\big(V_{t-1}(s') - V^*(s')\big) \\
&\leq \underbrace{\sum_{t=1}^{T} \sum_{s' \in \mathcal{S}} \sqrt{2\log(ST/\delta)}\sqrt{\frac{\mathbb{P}(s'|s_t, a_t)}{N_{t-1}(s_t, a_t) \vee 1}}\big(V_{t-1}(s') - V^*(s')\big)}_{I_1} + \underbrace{\sum_{t=1}^{T} \frac{2S\log(ST/\delta)}{3(1-\gamma)\big(N_{t-1}(s_t, a_t) \vee 1\big)}}_{I_2},
\end{aligned} \tag{D.5}$$

where first inequality holds due to the definition of $\mathcal{E}_2$ and the second inequality holds due to $0 \leq V_{t+1}(s') - V^*(s') \leq 1/(1-\gamma)$. To bound term I_1, we separate $\mathcal{S}$ into two subsets $\mathcal{S}_t^1 \cup \mathcal{S}_t^2$, where

$$\mathcal{S}_t^1 = \left\{s \in \mathcal{S} : \mathbb{P}(s|s_t, a_t)\big(N_{t-1}(s_t, a_t) \vee 1\big) \geq \frac{8\log(ST/\delta)}{(1-\gamma)^2}\right\},\ \mathcal{S}_t^2 = \mathcal{S}/\mathcal{S}_t^1.$$

Then on the event $\mathcal{E}_4$, we have

$$\begin{aligned}
I_1 &= \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t^1} \mathbb{P}(s'|s_t, a_t)\sqrt{2\log(ST/\delta)}\sqrt{\frac{1}{\mathbb{P}(s'|s_t, a_t)\big(N_{t-1}(s_t, a_t) \vee 1\big)}}\big(V_{t-1}(s') - V^*(s')\big) \\
&\quad + \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t^2} \frac{\sqrt{2\log(ST/\delta)\mathbb{P}(s'|s_t, a_t)\big(N_{t-1}(s_t, a_t) \vee 1\big)}}{N_{t-1}(s_t, a_t) \vee 1}\big(V_{t-1}(s') - V^*(s')\big) \\
&\leq \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t^1} (1-\gamma)\mathbb{P}(s'|s_t, a_t)\big(V_{t-1}(s') - V^*(s')\big)/2 + \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t^2} \frac{4\log(ST/\delta)}{3(1-\gamma)^2\big(N_{t-1}(s_t, a_t) \vee 1\big)} \\
&\leq \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t^1} (1-\gamma)\mathbb{P}(s'|s_t, a_t)\big(V_{t-1}(s') - V^*(s')\big)/2 + \frac{4S^2A\log(ST/\delta)\log(3T)}{3(1-\gamma)^2} \\
&\leq \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}} (1-\gamma)\mathbb{P}(s'|s_t, a_t)\big(V_{t-1}(s') - V^*(s')\big)/2 + \frac{4S^2A\log(ST/\delta)\log(3T)}{3(1-\gamma)^2} \\
&\leq (1-\gamma)/2 \cdot \left[\sum_{t=1}^{T} \big(V_{t-1}(s_{t+1}) - V^*(s_{t+1})\big) + \frac{\sqrt{2T\log(1/\delta)}}{1-\gamma}\right] + \frac{4S^2A\log(ST/\delta)\log(3T)}{3(1-\gamma)^2} \\
&\leq (1-\gamma)/2 \cdot \sum_{t=1}^{T} \big(V_{t-1}(s_{t+1}) - V_{t+1}^{\pi}(s_{t+1})\big) + \sqrt{2T\log(1/\delta)} + \frac{4S^2A\log(ST/\delta)\log(3T)}{3(1-\gamma)^2} \\
&\leq (1-\gamma)/2 \cdot \left[\mathrm{Regret}'(T) + \frac{(2S+2)}{1-\gamma}\right] + \sqrt{2T\log(1/\delta)} + \frac{4S^2A\log(ST/\delta)\log(3T)}{3(1-\gamma)^2},
\end{aligned} \tag{D.6}$$

where the first inequality holds due to separate condition of $\mathbb{P}(s')$, the second inequality holds due to Lemma D.3, the third inequality holds due to $V_{t-1}(s') \geq V^*(s')$, the fourth inequality holds due to the definition of event $\mathcal{E}_4$, the fifth inequality holds due to $V^* \geq V_{t+1}^{\pi}$, and the last inequality holds due to Lemma C.2. For the term I_2, according to Lemma D.3, we have

$$I_2 \leq \frac{2S^2 A\log(ST/\delta)\log(3T)}{3(1-\gamma)}. \tag{D.7}$$

Substituting (D.6),(D.7) into (D.5), we complete the proof of Lemma C.3. □

D.4. Proof of Lemma C.4

Proof of Lemma C.4. On the event $\mathcal{E}_1 \cap \mathcal{E}_5 \cap \mathcal{E}_7$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \gamma[(\mathbb{P}_{t-1} - \mathbb{P})V^*](s_t, a_t) \\
&\quad \leq \sum_{t=1}^{T} \gamma\sqrt{\frac{2\mathbb{V}^*(s_t,a_t)\log(SAT/\delta)}{N_{t-1}(s_t,a_t) \vee 1}} + \frac{2\log(SAT/\delta)\gamma}{(1-\gamma)\big(N_{t-1}(s_t,a_t)\vee 1\big)} \\
&\quad \leq \gamma\sqrt{2\log(SAT/\delta)}\sqrt{\sum_{t=1}^{T}\mathbb{V}^*(s_t,a_t)}\sqrt{\sum_{t=1}^{T}\frac{1}{N_{t-1}(s_t,a_t)\vee 1}} + \sum_{t=1}^{T}\frac{2\gamma\log(SAT/\delta)}{(1-\gamma)\big(N_{t-1}(s_t,a_t)\vee 1\big)} \\
&\quad \leq \gamma U\sqrt{2SA}\sqrt{\sum_{t=1}^{T}\mathbb{V}^*(s_t,a_t)} + \frac{2\gamma SAU^2}{1-\gamma} \\
&\quad = \gamma U\sqrt{2SA}\sqrt{\sum_{t=1}^{T}\mathbb{V}_t^{\pi}(s_t,a_t) + \sum_{t=1}^{T}\mathbb{V}^*(s_t,a_t) - \sum_{t=1}^{T}\mathbb{V}_t^{\pi}(s_t,a_t)} + \frac{2\gamma SAU^2}{1-\gamma} \\
&\quad \leq U\sqrt{2SA}\sqrt{\frac{5T}{1-\gamma} + \frac{29U}{3(1-\gamma)^3} + \frac{2\text{Regret}'(T)}{1-\gamma} + \frac{\sqrt{2TU}}{(1-\gamma)^2}} + \frac{2SAU^2}{1-\gamma},
\end{aligned} \tag{D.8}$$

where the first inequality holds due to the definition of event $\mathcal{E}_1$, the second inequality holds due to Cauchy-Schwarz inequality, the third inequality holds due to Lemma D.3 and the definition of U, and the last inequality holds due to Lemma D.8 and the definition of event $\mathcal{E}_5$. Thus, we complete the proof of Lemma C.4. □

D.5. Proof of Lemma C.5

Proof of Lemma C.5. For the term $\text{UCB}_{t-1}(s_t, a_t)$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\gamma\text{UCB}_{t-1}(s_t,a_t) &\leq \underbrace{\sum_{t=1}^{T}\gamma\sqrt{\frac{8U\mathbb{V}_{t-1}(s_t,a_t)}{N_{t-1}(s_t,a_t)\vee 1}}}_{I_1} + \underbrace{\sum_{t=1}^{T}\gamma\frac{8U}{(1-\gamma)\big(N_{t-1}(s_t,a_t)\vee 1\big)}}_{I_2} \\
&\quad + \underbrace{\sum_{t=1}^{T}\gamma\sqrt{\frac{8\sum_{s'}\mathbb{P}_t(s'|s_t,a_t)\min\big\{100B_t(s'), 1/(1-\gamma)^2\big\}}{N_{t-1}(s_t,a_t)\vee 1}}}_{I_3}.
\end{aligned} \tag{D.9}$$

For the term I_1, on the event $\mathcal{E}_5 \cap \mathcal{E}_6 \cap \mathcal{E}_8$, we have

$$\begin{aligned}
I_1 &\leq \gamma\sqrt{8U\sum_{t=1}^{T}\mathbb{V}_{t-1}(s_t,a_t)}\sqrt{\sum_{t=1}^{T}\frac{1}{N_{t-1}(s_t,a_t)\vee 1}} \\
&\leq \gamma U\sqrt{8SA}\sqrt{\sum_{t=1}^{T}\mathbb{V}_{t-1}(s_t,a_t)}
\end{aligned}$$

$$
\begin{aligned}
&= \gamma U\sqrt{8SA}\sqrt{\sum_{i=1}^{T} \mathbb{V}_t^{\pi}(s_t,a_t) + \sum_{t=1}^{T} \mathbb{V}_{t-1}(s_t,a_t) - \sum_{i=1}^{T} \mathbb{V}_t^{\pi}(s_t,a_t)} \\
&\le U\sqrt{8SA}\sqrt{\frac{5T}{1-\gamma} + \frac{29U}{3(1-\gamma)^3} + \frac{2\text{Regret}'(T)}{1-\gamma} + \frac{9SU\sqrt{AT}}{(1-\gamma)^2}},
\end{aligned}
\tag{D.10}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to Lemma D.3, the last inequality holds due to the definition of event $\mathcal{E}_5$ and Lemma D.9. For the term I_2, by Lemma D.3, we have

$$
I_2 = \sum_{t=1}^{T} \frac{8U}{(1-\gamma)\big(N_{t-1}(s_t,a_t) \vee 1\big)} \le \frac{8SAU^2}{1-\gamma}. \tag{D.11}
$$

For the term I_3, on the event $\mathcal{E}_8 \cap \mathcal{E}_9$, we have

$$
\begin{aligned}
I_3 &\le \sqrt{8\sum_{t=1}^{T} \frac{1}{N_{t-1}(s_t,a_t) \vee 1}}\sqrt{\sum_{t=1}^{T}\sum_{s'} \mathbb{P}_t(s'|s_t,a_t)\min\left\{\frac{100S^2A^2U^5}{(1-\gamma)^5 N_{t-1}(s')}, \frac{1}{(1-\gamma)^2}\right\}} \\
&\le \sqrt{8SAU}\sqrt{\sum_{t=1}^{T}\sum_{s'} \mathbb{P}_t(s'|s_t,a_t)\min\left\{\frac{100S^2A^2U^5}{(1-\gamma)^5\big(N_{t-1}(s') \vee 1\big)}, \frac{1}{(1-\gamma)^2}\right\}} \\
&\le \sqrt{8SAU} \\
&\quad \cdot \sqrt{\sum_{i=1}^{T} \frac{\sqrt{2SU}}{(1-\gamma)^2\sqrt{N_t(s_t,a_t) \vee 1}} + \sum_{t=1}^{T}\sum_{s'} \mathbb{P}(s'|s_t,a_t)\min\left\{\frac{100S^2A^2U^5}{(1-\gamma)^5\big(N_{t-1}(s') \vee 1\big)}, \frac{1}{(1-\gamma)^2}\right\}} \\
&\le \sqrt{8SAU}\sqrt{\frac{SU\sqrt{2AT}}{(1-\gamma)^2} + \frac{\sqrt{2TU}}{(1-\gamma)^2} + \sum_{t=1}^{T}\min\left\{\frac{100S^2A^2U^5}{(1-\gamma)^5\big(N_{t-1}(s_{t+1}) \vee 1\big)}, \frac{1}{(1-\gamma)^2}\right\}} \\
&\le \sqrt{8SAU}\sqrt{\frac{SU\sqrt{2AT}}{(1-\gamma)^2} + \frac{\sqrt{2TU}}{(1-\gamma)^2} + \frac{100S^3A^2U^6}{(1-\gamma)^5}},
\end{aligned}
\tag{D.12}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to Lemma D.3, the third inequality holds due to the definition of event $\mathcal{E}_8$, the forth inequality holds due to the definition of event $\mathcal{E}_9$ and the last inequality holds due to Lemma D.3. Substituting (D.10), (D.11) and (D.12) into (D.9), we complete the proof of Lemma C.5. □

E. Proof of Lemmas in Section C.2

E.1. Proof of Lemma C.6

Proof of Lemma C.6. We have

$$
\begin{aligned}
\mathbb{E}^*\left[\sum_{t=1}^{T} V^*(s_t) - V_t^{\pi}(s_t)\right] &= \mathbb{E}^*\left[\sum_{t=1}^{T} V^*(s_t) - \sum_{k=0}^{\infty}\gamma^k r(s_{t+k},a_{t+k})\right] \\
&= \mathbb{E}^*\left[\sum_{t=1}^{T}\Big(V^*(s_t) - \sum_{k=0}^{t}\gamma^k r(s_t,a_t)\Big) - \sum_{t=T+1}^{\infty}\sum_{k=0}^{T}\gamma^{t-k} r(s_t,a_t)\right] \\
&\ge \mathbb{E}^*\left[\sum_{t=1}^{T} V^*(s_t) - \frac{r(s_t,a_t)}{1-\gamma}\right] - \sum_{t=T+1}^{\infty}\sum_{k=0}^{T}\gamma^{t-k} \\
&\ge \mathbb{E}^*\left[\sum_{t=1}^{T} V^*(s_t) - \frac{r(s_t,a_t)}{1-\gamma}\right] - \frac{4}{(1-\gamma)^2}.
\end{aligned}
\tag{E.1}
$$

where the first inequality holds due to $0 \le r(s_t, a_t) \le 1$ and the last inequality holds due to $\sum_{k=0}^{\infty} \gamma^k = 1/(1-\gamma)$. Thus, we finish the proof of Lemma C.6. □

E.2. Proof of Lemma C.7

Proof of Lemma C.7. In this proof, we follow the proof technique in (Liu and Su, 2020) and (Jaksch et al., 2010). For simplicity, we denote $\epsilon = \sqrt{A(1-\gamma)/K}/24$ and we first determine the optimal policy in these hard-to-learn MDPs. According to (3.1), for optimal policy π^*, we have

$$Q^*(s,a) = r(s,a) + \gamma[\mathbb{P}V^*](s,a),$$

For each $j \in [S]$ and state $s = s_{j,1}$, the choice of action a will not effect the reward $r(s,a)$ and the probability transition function $\mathbb{P}(\cdot|s,a)$. For optimal action a^* at state $s = s_{j,0}$, we have

$$\begin{aligned} V^*(s_{j,0}) &= r(s,a) + \gamma[\mathbb{P}V^*](s,a^*) \\ &= 0 + \gamma\mathbb{P}(s_{j,0}|s_{j,0},a^*)V^*(s_{j,0}) + \gamma\mathbb{P}(s_{j,1}|s_{j,0},a^*)V^*(s_{j,1}). \end{aligned}$$

Since $\mathbb{P}(s_{j,0}|s_{j,0},a^*) + \mathbb{P}(s_{j,1}|s_{j,0},a^*) = 1$, we have

$$(1-\gamma)V^*(s_{j,0}) = \gamma\big(V^*(s_{j,1}) - V^*(s_{j,0})\big),$$

and it implies that $V^*(s_{j,1}) \ge V^*(s_{j,0})$. Therefore, for all action $a \ne a_j^*$, we have $Q^*(s_{j,0}, a_j^*) \ge Q^*(s_{j,0}, a)$ and it further implies that the optimal action at state $s = s_{j,0}$ is a_j^*. Thus, according to the optimal bellman equation 3.1, for each $j \in [S]$, we have

$$\begin{aligned} V^*(s_{j,0}) &= \gamma(1-\gamma+\epsilon)V^*(s_{j,1}) + \gamma(\gamma-\epsilon)V^*(s_{j,0}), \\ V^*(s_{j,1}) &= 1 + \gamma(1-\gamma)V^*(s_{j+1,1}) + \gamma^2 V^*(s_{j,1}), \end{aligned}$$

and it implies that the optimal value function V^* is

$$\begin{aligned} V^*(s_{j,0}) &= \frac{\gamma - \gamma^2 + \gamma\epsilon}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}, \\ V^*(s_{j,1}) &= \frac{1 - \gamma^2 + \gamma\epsilon}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}. \end{aligned}$$

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 $A_{j,i}$. For each $j \in [S]$, by the definition of $X_{j,i}$, we have

$$\begin{aligned} &\mathbb{P}(X_{j,i} = s_{j,1}|X_{j,i-1} = s_{j,0}, A_{j,i-1}) = 1 - \gamma + \mathbb{1}_{A_{j,i}=a_j^*}\,\epsilon, \\ &\mathbb{P}(X_{j,i} = s_{j,0}|X_{j,i-1} = s_{j,0}, A_{j,i-1}) = \gamma - \mathbb{1}_{a=a_j^*}\,\epsilon, \\ &\mathbb{P}(X_{j,i} = s_{j,0}|X_{j,i-1} = s_{j,0}, A_{j,i-1}) = 1 - \gamma, \\ &\mathbb{P}(X_{j,i} = s_{j,1}|X_{j,i-1} = s_{j,1}, A_{j,i-1}) = \gamma, \end{aligned}$$

where the third equality holds because when $X_{j,i-1}$ leave state $s_{j,0}, s_{j,1}$, the next state in $s_{j,0}, s_{j,1}$ must be $s_{j,0}$. Similar to the proof of Theorem 5 in (Jaksch et al., 2010), we focus on the first K visits to the state set $\{s_{j,0}, s_{j,1}\}$ and let random variable N_0, N_1 and N_0^* denote the total number of visit state $s_{j,0}$, the total number of visit state $s_{j,1}$ and the total number of visit state $s_{j,0}$ with action a_j^*. By the same argument as the proof of Theorem 5 in (Jaksch et al., 2010), for the random variable N_1 and N_0^*, we have following property:

$$\mathbb{E}[N_1] \le \frac{K}{2} + \frac{1}{2(1-\gamma)} + \frac{\epsilon\mathbb{E}[N_0^*]}{1-\gamma}, \tag{E.2}$$

and

$$\mathbb{E}[N_0^*] \le \frac{K}{2A} + \frac{1}{2A(1-\gamma)} + \frac{\epsilon K}{2}\sqrt{\frac{K}{A(1-\gamma)}} + \frac{\epsilon K}{2\sqrt{A}(1-\gamma)}. \tag{E.3}$$

Therefore, the regret can be upper bounded by

$$\begin{aligned}
&\mathbb{E}^*\bigg[\sum_{i=1}^{K} V^*(X_{j,i}) - \frac{r(X_{j,i},A_{j,i})}{1-\gamma}\bigg]\\
&= \mathbb{E}[N_0]\big(V^*(s_{j,0}) - 0\big) + \mathbb{E}[N_1]\bigg(V^*(s_{j,1}) - \frac{1}{1-\gamma}\bigg)\\
&= \frac{(\gamma-\gamma^2+\gamma\epsilon)\big(K-\mathbb{E}[N_1]\big) - (\gamma-\gamma^2)\mathbb{E}[N_1]}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}\\
&\geq \frac{\frac{K\gamma\epsilon}{2} - \gamma - \frac{\gamma\epsilon}{2(1-\gamma)} - \frac{\mathbb{E}[N_0^*]\epsilon(2\gamma-2\gamma^2+\gamma\epsilon)}{1-\gamma}}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}\\
&\geq \frac{\frac{K\gamma\epsilon}{2} - \gamma - \frac{\gamma\epsilon}{2(1-\gamma)} - \Big(\frac{K}{2A} + \frac{1}{2A(1-\gamma)} + \frac{\epsilon K}{2}\sqrt{\frac{K}{A(1-\gamma)}} + \frac{\epsilon K}{2\sqrt{A}(1-\gamma)}\Big)\frac{\epsilon(2\gamma-2\gamma^2+\gamma\epsilon)}{1-\gamma}}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}. \qquad \text{(E.4)}
\end{aligned}$$

where the second inequality holds due to the fact that $\mathbb{E}[N_0] + \mathbb{E}[N_1] = K$, the third inequality holds due to (E.2) and the last inequality holds due to (E.3). Since $K \geq 10SA/(1-\gamma)^4$, $\gamma > 2/3$ and $A \geq 30$, (E.4) can be further bounded by

$$\begin{aligned}
&\mathbb{E}^*\bigg[\sum_{i=1}^{K} V^*(X_{j,i}) - \frac{r(X_{j,i},A_{j,i})}{1-\gamma}\bigg]\\
&\geq \frac{\frac{K\gamma\epsilon}{2} - \gamma - \frac{\gamma\epsilon}{2(1-\gamma)} - \Big(\frac{K}{2A} + \frac{1}{2A(1-\gamma)} + \frac{\epsilon K}{2}\sqrt{\frac{K}{A(1-\gamma)}} + \frac{\epsilon K}{2\sqrt{A}(1-\gamma)}\Big)\frac{\epsilon(2\gamma-2\gamma^2+\gamma\epsilon)}{1-\gamma}}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}\\
&\geq \gamma \times \frac{\frac{K\epsilon}{4} - 1 - 3\epsilon\Big(\frac{5K}{8A} + \frac{\epsilon K}{2}\sqrt{\frac{K}{A(1-\gamma)}} + \frac{\epsilon K}{2\sqrt{A}(1-\gamma)}\Big)}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}\\
&\geq \gamma \times \frac{\frac{\sqrt{AK(1-\gamma)}}{576} - 1}{(1-\gamma)(1-2\gamma^2+\gamma+\gamma\epsilon)}\\
&\geq \frac{\sqrt{AK}}{2304(1-\gamma)^{1.5}} - \frac{1}{(1-\gamma)^2}, \qquad \text{(E.5)}
\end{aligned}$$

where the second inequality holds to $\epsilon = \sqrt{A(1-\gamma)/K}/24 \leq 1-\gamma$ with $K \geq 10SA/(1-\gamma)^4$, the third inequality holds due to $\epsilon = \sqrt{A(1-\gamma)/K}/24$ with $A \geq 30$ and the last inequality holds due to $\gamma \geq 2/3$ and $\epsilon = \sqrt{A(1-\gamma)/K}/24 \leq 1-\gamma$. Therefore, we finish the proof of Lemma C.7.

□

E.3. Proof of Lemma C.8

Proof of Lemma C.8. For each $j \in [S]$ and $t \in [T]$, we denote $H = \lfloor \log T/(1-\gamma) \rfloor + 1$, random variable

$$Y_{j,i} = \sum_{k=0}^{H} \gamma^k r(X_{j,i+k}, A_{j,i+k}),$$

and filtration $\mathcal{F}_{j,i}$ contain all random variable before $X_{j,i+H}$. For simplicity, we ignore the subscript j and only focus on the subscript i.

Since Y_i is $\mathcal{F}_i$-measurable and $0 \leq Y_i \leq 1/(1-\gamma)$, for each $k \in [H]$, with probability at least $1-\delta$, we have

$$\sum_{i=\lfloor \frac{K}{H} \rfloor + 1}^{\lfloor \frac{t}{H} \rfloor + 1} Y_{iH+k} \leq \sum_{i=\lfloor \frac{K}{H} \rfloor + 1}^{\lfloor \frac{t}{H} \rfloor + 1} \mathbb{E}\Big[Y_{iH+k} | \mathcal{F}_{(i-1)H+k}\Big] + \sqrt{\frac{2t}{1-\gamma}\log\frac{1}{\delta}}$$

$$
\begin{aligned}
&= \sum_{i=\lfloor \frac{K}{H} \rfloor+1}^{\lfloor \frac{t}{H} \rfloor+1} V^{\pi}_{iH+k}(X_{iH+k}) + \sqrt{\frac{2t}{1-\gamma}\log\frac{1}{\delta}} \\
&\le \sum_{i=\lfloor \frac{K}{H} \rfloor+1}^{\lfloor \frac{t}{H} \rfloor+1} V^{*}(X_{iH+k}) + \sqrt{\frac{2t}{1-\gamma}\log\frac{1}{\delta}},
\end{aligned} \tag{E.6}
$$

where the first inequality holds due to Lemma D.1 and the second inequality holds due to the definition of optimal value function V^*. Taking summation of (E.6), for all $k \in [H]$, with probability at least $1 - H\delta$, we have

$$
\begin{aligned}
\sum_{i=K+1}^{t} V^*(X_i) + \frac{\sqrt{2t\log\frac{1}{\delta}\log T}}{(1-\gamma)^{1.5}} &\ge \sum_{i=K+1}^{t} Y_i \\
&= \sum_{i=K+1}^{t}\sum_{k=0}^{H} \gamma^k r(X_{i+k}, A_{i+k}) \\
&\ge \sum_{i=K+1}^{t} r(X_i, A_i) \sum_{k=0}^{\min(H, i-K-1)} \gamma^i \\
&\ge \sum_{i=K+1}^{t} \frac{r(X_i, A_i)}{1-\gamma} - \frac{4}{(1-\gamma)^2},
\end{aligned}
$$

where the second inequality holds due to $0 \le r(s,a) \le 1$. Finally, taking union for all $j \in [S]$ and $t \in [T]$, we complete the proof. □

E.4. Proof of Lemma C.9

Proof of Lemma C.9. Let $Y_{j,i}$ be an indicator random variables which denote whether the agent at state $X_{j,i}$ with action $A_{j,i}$ goes to the different state. $Y_{j,i} = 1$ if the agent goes to the different state and $Y_{j,i} = 0$ if the agent stay at the same state. Let filtration $\mathcal{F}_{j,i}$ contain all random variables before $X_{j,i}$. Then, for each $j \in [S]$, with probability at least $1 - \delta$, we have

$$
\sum_{i=1}^{K} Y_{j,i} \le \sum_{i=1}^{K} \mathbb{E}\big[Y_{j,i}|\mathcal{F}_{j,i-1}\big] + \sqrt{2K\log\frac{1}{\delta}} \le (1-\gamma+\epsilon)K + \sqrt{2K\log\frac{1}{\delta}} \le 3(1-\gamma)K, \tag{E.7}
$$

where the first inequality holds due to Lemma D.1, the second inequality holds due to the definition of our MDPs and the last one holds due to the selection of K. Similarly, with probability at least $1 - \delta$, we have

$$
\sum_{i=1}^{5K} Y_{j.i} \ge \sum_{i=1}^{2K} \mathbb{E}\big[Y_{j,i}|\mathcal{F}_{j,i-1}\big] - \sqrt{10K\log\frac{1}{\delta}} \ge 5K(1-\gamma) - \sqrt{10K\log\frac{1}{\delta}} \ge 4(1-\gamma)K, \tag{E.8}
$$

where the first inequality holds due to Lemma D.1, the second inequality holds due to the definition of our MDPs and the last one holds due to the selection of K. Taking a union bound (E.7) and (E.8) for all $j \in [S]$, then we have (E.7) and (E.8) hold with probability at least $1 - 2S\delta$. Let $Z_{j,i}$ be the number of times for the agent to start from state $s_{j,i}$ and travel the next different state in the first T steps. By definition, we have

$$
Z_{j,0} + Z_{j,1} = \sum_{i=1}^{T_j} Y_{j,i}. \tag{E.9}
$$

By Pigeonhole principle, there exist a j^* such that $T_{j^*} \ge T/S = 10K > 5K$. Therefore, we have

$$
Z_{j^*,0} + Z_{j^*,1} = \sum_{i=1}^{T_{j^*}} Y_{j^*,i} \ge \sum_{i=1}^{5K} Y_{j^*,i} \ge 4(1-\gamma)K. \tag{E.10}
$$

Furthermore, after leaving the state $s_{j^*,0}$, the agent will visit all other states before arrive the state $s_{j^*,0}$ again. Thus, for any $k \in [S]$, the difference between $Z_{j^*,0}$ and $Z_{k,0}$ is at most 1, so do $Z_{j^*,1}$ and $Z_{k,1}$. Therefore, for any $k \in [S]$, we have

$$Z_{k,0} + Z_{k,1} \geq Z_{j^*,0} + Z_{j^*,1} - 2 \geq 4(1-\gamma)K - 2 > 3(1-\gamma)K \geq \sum_{i=1}^{K} Y_{k,i}, \tag{E.11}$$

where the second inequality holds due to (E.10), the third inequality holds since $K > 2/(1-\gamma)$ and the last one holds due to (E.7). Finally, by (E.9) we have $Z_{k,0} + Z_{k,1} = \sum_{i=1}^{T_k} Y_{k,i}$. Combining it with (E.11), we have $\sum_{i=1}^{T_k} Y_{k,i} > \sum_{i=1}^{K} Y_{k,i}$, which suggests that $T_k > k$. Thus, we complete the proof.

□

F. Proof of Lemmas in Appendix D

F.1. Proof of Lemma D.3

Proof of Lemma D.3. We have

$$\sum_{i=1}^{t} \frac{1}{N_{i-1}(s_i, a_i) \vee 1} = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} 1 + \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \sum_{i=1}^{N_{t-1}(s,a)} \frac{1}{i} \leq SA + \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \sum_{i=1}^{t} \frac{1}{i} \leq SA\log(3T). \tag{F.1}$$

We also have

$$\sum_{i=1}^{t} \frac{1}{N_{i-1}(s_i) \vee 1} = \sum_{s\in\mathcal{S}} 1 + \sum_{i=1}^{N_t(s)} \frac{1}{i} \leq S + \sum_{s\in\mathcal{S}} \sum_{i=1}^{t} \frac{1}{i} \leq S\log(3T).$$

According to (F.1), for a subset $\mathcal{C} \subseteq [T]$, we have

$$\sum_{i\in\mathcal{C}} \frac{1}{\sqrt{N_{i-1}(s_i, a_i) \vee 1}} \leq \sqrt{|\mathcal{C}| \sum_{i\in\mathcal{C}} \frac{1}{N_{i-1}(s_i, a_i) \vee 1}} \leq \sqrt{SA\log(3T)|\mathcal{C}|},$$

where the first inequality holds due to Cauchy-Schwarz inequality and the second inequality holds due to (F.1). Thus, we complete the proof. □

F.2. Proof of Lemma D.4

Proof of Lemma D.4. For each $s \in \mathcal{S}, a \in \mathcal{A}$, we denote $t_0 = 0$ and

$$t_i = \min\big\{t | t > t_{i-1}, (s_t, a_t) = (s, a)\big\}. \tag{F.2}$$

Here, t_i is the time which state-action pair (s, a) appear for the ith time and the random variable t_i is a stopping time. Beside, the random variable $V^*(s_{t_i+1})(i = 1, 2., ,)$ are random variable with value in $\big[0, 1/(1-\gamma)\big]$ and variance $\mathbb{V}^*(s, a)$. By Lemma D.2 and a union bound, with probability at least $1 - \delta$, for all $s \in \mathcal{S}, a \in \mathcal{A}, \tau \in [T]$, we have

$$\sum_{i=1}^{\tau} V^*(s_{t_i+1}) - \sum_{i=1}^{\tau} \mathbb{P}V^*(s, a) \leq \sqrt{2\tau\mathbb{V}^*(s, a)\log(SAT/\delta)} + \frac{2\log(SAT/\delta)}{3(1-\gamma)}.$$

Thus, for all $\tau \in [T]$, we have

$$\begin{aligned}
\big[(\mathbb{P}_{t_\tau+1} - \mathbb{P})V^*\big](s, a) &= \frac{1}{\tau}\sum_{i=1}^{\tau} V^*(s_{t_i+1}) - \frac{1}{\tau}\sum_{i=1}^{\tau} \mathbb{P}V^*(s, a) \\
&\leq \sqrt{\frac{2\mathbb{V}^*(s, a)\log(SAT/\delta)}{\tau}} + \frac{2\log(SAT/\delta)}{3(1-\gamma)\tau} \\
&= \sqrt{\frac{2\mathbb{V}^*(s, a)\log(SAT/\delta)}{N_{t_\tau}(s, a)}} + \frac{2\log(SAT/\delta)}{3(1-\gamma)N_{t_\tau}(s, a)}.
\end{aligned} \tag{F.3}$$

In addition, for $\tau = 0$, we have

$$\big[(\mathbb{P}_{t_\tau+1} - \mathbb{P})V^*\big](s,a) \le \frac{1}{1-\gamma} \le \frac{2\log(SAT/\delta)}{3(1-\gamma)\big(N_{t_\tau}(s,a)\vee 1\big)}, \tag{F.4}$$

where the first inequality holds due to $0 \le V^*(s) \le 1/(1-\gamma)$ and the second inequality holds due to $N_{t_\tau}(s,a) = 0$. Since $\mathbb{P}_t$ and $N_{t-1}(s,a)$ changed only when $t = t_\tau + 1$, we complete the proof by combining (F.3) and (F.4). □

F.3. Proof of Lemma D.6

Proof of Lemma D.6. For each $s \in \mathcal{S}, a \in \mathcal{A}$, we denote $t_0 = 0$ and denote

$$t_i = \min\big\{t | t > t_{i-1}, (s_t, a_t) = (s,a)\big\}. \tag{F.5}$$

Here, t_i is the time which state-action pair (s,a) appear for the ith time and the random variable t_i is a stopping time. Beside, the random variable $V^*(s_{t_i+1})(i = 1, 2.,,)$ are random variable with value in $\big[0, 1/(1-\gamma)\big]$ and variance $\mathbb{V}^*(s,a)$. By Lemma D.5 and a union bound, with probability at least $1-\delta$, for all $s \in \mathcal{S}, a \in \mathcal{A}, \tau \in [T]$, we have

$$\sum_{i=1}^{\tau} \mathbb{P}V_t^*(s,a) - \sum_{i=1}^{\tau} V^*(s_{t_i+1}) \le \sqrt{2\tau \mathbb{V}_{t_\tau}^*(s,a)\log(SAT/\delta)} + \frac{7\log(SAT/\delta)}{3(1-\gamma)}.$$

Thus, for all $\tau \in [T]$, we have

$$\begin{aligned}\big[(\mathbb{P} - \mathbb{P}_{t_\tau+1})V^*\big](s,a) &= \frac{1}{\tau}\Big|\sum_{i=1}^{\tau} V^*(s_{t_i+1}) - \sum_{i=1}^{\tau} \mathbb{P}V^*(s,a)\Big| \\ &\le \sqrt{\frac{2\mathbb{V}_{t_\tau}^*(s,a)\log(SAT/\delta)}{\tau}} + \frac{7\log(SAT/\delta)}{3(1-\gamma)\tau} \\ &= \sqrt{\frac{2\mathbb{V}_{t_\tau}^*(s,a)\log(SAT/\delta)}{N_{t_\tau}(s,a)}} + \frac{7\log(SAT/\delta)}{3(1-\gamma)N_{t_\tau}(s,a)}.\end{aligned} \tag{F.6}$$

In addition, for $\tau = 0$, we have

$$\big[(\mathbb{P} - \mathbb{P}_{t_\tau+1})V^*\big](s,a) \le \frac{1}{1-\gamma} \le \frac{7\log(SAT/\delta)}{3(1-\gamma)\big(N_{t_\tau}(s,a)\vee 1\big)}, \tag{F.7}$$

where the first inequality holds due to $0 \le V^*(s) \le 1/(1-\gamma)$ and the second inequality holds due to $N_{t_\tau}(s,a) = 0$. Since $\mathbb{P}_t, \mathbb{V}_{t-1}^*$ and $N_{t-1}(s,a)$ changed only when $t = t_\tau + 1$, we complete the proof by combining (F.6) and (F.7). □

F.4. Proof of Lemma D.7

Proof of Lemma D.7. For simplicity, we denote $H = \lfloor 1/(1-\gamma)\rfloor + 1, T' = \lfloor T/H\rfloor + 1$ and filtration $\mathcal{F}_t$ contained all random variables before first $t + H$ steps. Then for every $t \in [T]$, we have

$$\begin{aligned}\frac{1}{(1-\gamma)^2} &\ge \mathbb{E}\Big[\big(\sum_{i=0}^{\infty}\gamma^i r(s_{t+i}, a_{t+i})\big) - V_t^\pi(s_t)|\mathcal{F}_{t-H}\Big]^2 \\ &= \mathbb{E}\Big[\sum_{i=0}^{\infty}\gamma^i\big(r(s_{t+i}, a_{t+i}) + \gamma V_{t+i+1}^\pi(s_{t+i+1}) - V_{t+i}^\pi(s_{t+i})\big)|\mathcal{F}_{t-H}\Big]^2 \\ &= \mathbb{E}\Big[\sum_{i=0}^{\infty}\gamma^{2i}\Big[r(s_{t+i}, a_{t+i}) + \gamma V_{t+i+1}^\pi(s_{t+i+1}) - V_{t+i}^\pi(s_{t+i})\Big]^2|\mathcal{F}_{t-H}\Big] \\ &= \mathbb{E}\Big[\sum_{i=0}^{\infty}\gamma^{2i+2}\mathbb{V}_{t+i}^\pi(s_{t+i}, a_{t+i})|\mathcal{F}_{t-H}\Big]\end{aligned}$$

$$\geq \mathbb{E}\Big[\underbrace{\sum_{i=0}^{H}\gamma^{2i+2}\mathbb{V}^{\pi}_{t+i}(s_{t+i},a_{t+i})}_{X_t}|\mathcal{F}_{t-H}\Big], \tag{F.8}$$

where the first inequality holds due to $0 \leq r(s,a) \leq 1, 0 \leq V_t^{\pi}(s) \leq 1/(1-\gamma)$ and the second inequality holds due to $\mathbb{V}^{\pi}_{t+i}(s_{t+i},a_{t+i}) \geq 0$. For the random variable X_t, we have

$$|X_t| \leq \sum_{i=0}^{H}\frac{\gamma^{2i+2}}{(1-\gamma)^2} \leq \frac{1}{(1-\gamma)^3},\ \mathrm{Var}\big[|X_t||\mathcal{F}_{t-H}\big] \leq (\max|X_t|)\mathbb{E}[X_t|\mathcal{F}_{t-H}] \leq \frac{1}{(1-\gamma)^5},$$

Since X_t is $\mathcal{F}_t$-measurable and $\mathbb{E}[X_t|\mathcal{F}_{t-H}] \leq 1/(1-\gamma)^2$, for each $i \in [H]$, by Lemma D.2, with probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{j=0}^{T'}X_{jH+i} &\leq \sum_{j=0}^{T'}\mathbb{E}[X_{jH+i}|\mathcal{F}_{(j-1)H+i}] + \sqrt{\frac{2T'\log(1/\delta)}{(1-\gamma)^5}} + \frac{2\log(1/\delta)}{3(1-\gamma)^3}\\ &\leq \frac{T'}{(1-\gamma)^2} + \sqrt{\frac{2T'\log(1/\delta)}{(1-\gamma)^5}} + \frac{2\log(1/\delta)}{3(1-\gamma)^3}.\end{aligned} \tag{F.9}$$

Taking summation for (F.9) with all $i \in [H]$, with probability at least $1 - H\delta$, we have

$$\begin{aligned}\sum_{t=1}^{T}X_t &= \sum_{i=1}^{H}\sum_{j=0}^{T'}X_{jH+i}\\ &\leq \sum_{i=1}^{H}\bigg(\frac{T'}{(1-\gamma)^2} + \sqrt{\frac{2T'\log(1/\delta)}{(1-\gamma)^5}} + \frac{2\log(1/\delta)}{3(1-\gamma)^3}\bigg)\\ &\leq \frac{T}{(1-\gamma)^2} + \sqrt{\frac{4T\log(1/\delta)}{(1-\gamma)^6}} + \frac{4\log(1/\delta)}{3(1-\gamma)^4}\\ &\leq \frac{2T}{(1-\gamma)^2} + \frac{7\log(1/\delta)}{3(1-\gamma)^4},\end{aligned} \tag{F.10}$$

where the first inequality holds due to (F.9), the second inequality holds due to $T' = \lfloor T/H \rfloor + 1$ and the third inequality holds due to $x^2 + y^2 \geq 2xy$. By the definition of X_t, we have

$$\begin{aligned}\sum_{t=1}^{T}X_t &= \sum_{t=1}^{T}\sum_{i=0}^{H}\gamma^{2i+2}\mathbb{V}^{\pi}_{t+i}(s_{t+i},a_{t+i})\\ &\geq \sum_{t=1}^{T}\mathbb{V}^{\pi}_{t}(s_t,a_t)\sum_{i=0}^{\min\{H,t-1\}}\gamma^{2i+2}\\ &= \sum_{i=0}^{H}\gamma^{2i+2}\sum_{t=1}^{T}\mathbb{V}^{\pi}_{t}(s_t,a_t) - \sum_{t=1}^{H}\mathbb{V}^{\pi}_{t}(s_t,a_t)\sum_{i=t}^{H}\gamma^{2i+2}\\ &\geq \frac{\gamma^2-\gamma^{2H+4}}{1-\gamma^2}\sum_{t=1}^{T}\mathbb{V}^{\pi}_{t}(s_t,a_t) - \frac{1}{(1-\gamma)^2}\sum_{t=1}^{H}\sum_{i=t}^{H}\gamma^{2i+2},\end{aligned} \tag{F.11}$$

where the first inequality holds due to $\mathbb{V}^{\pi}_{t}(s_t,a_t) \geq 0$ and the second inequality holds due to $\mathbb{V}^{\pi}_{t}(s_t,a_t) \leq 1/(1-\gamma)^2$. To further bound (F.11), we have

$$\frac{\gamma^2-\gamma^{2H+4}}{1-\gamma^2} = \frac{\gamma^2}{1-\gamma^2}(1-\gamma^{2H+2}) \geq \frac{\gamma^2}{1-\gamma^2}(1-\gamma^{2/(1-\gamma)}) \geq \frac{4\cdot\gamma^2}{5(1-\gamma^2)} \geq \frac{2\gamma^2}{5(1-\gamma)}, \tag{F.12}$$

where the first inequality holds since $2H+2 = 2\lfloor 1/(1-\gamma)\rfloor + 2 \geq 2/(1-\gamma)$, the second inequality holds since $0 \leq \gamma^{1/(1-\gamma)} \leq 0.4$ when $0 \leq \gamma \leq 1$, the last one holds since $1+\gamma \leq 2$. We also have

$$\sum_{t=1}^{H}\sum_{i=t}^{H}\gamma^{2i+2} \leq \sum_{t=1}^{H}\frac{\gamma^{2t+2}}{1-\gamma^2} \leq \frac{\gamma^4}{(1-\gamma^2)^2} \leq \frac{\gamma^4}{(1-\gamma)^2}. \tag{F.13}$$

Substituting (F.12) and (F.13) into (F.11), we have

$$\sum_{t=1}^{T} X_t \geq \frac{2\gamma^2}{5(1-\gamma)}\sum_{t=1}^{T}\mathbb{V}_t^{\pi}(s_t,a_t) - \frac{\gamma^4}{(1-\gamma)^4}. \tag{F.14}$$

Finally, substituting (F.14) into (F.10), we have

$$\gamma^2\sum_{t=1}^{T}\mathbb{V}_t^{\pi}(s_t,a_t) \leq \frac{5T}{1-\gamma} + \frac{35\log(1/\delta)}{6(1-\gamma)^3} + \frac{5\gamma^4}{2(1-\gamma)^3} \leq \frac{5T}{1-\gamma} + \frac{25\log(1/\delta)}{3(1-\gamma)^3}.$$

Thus, we complete the proof. □

F.5. Proof of Lemma D.8

Proof of Lemma D.8. On the event $\mathcal{E}_7$, we have

$$\begin{aligned}
\sum_{i=1}^{T}(\mathbb{V}^*(s_i,a_i) - \mathbb{V}_i^{\pi}(s_i,a_i)) &\leq \sum_{i=1}^{t}\Big[\mathbb{P}\big((V^*)^2 - (V_{i+1}^{\pi})^2\big)\Big](s_i,a_i)\\
&= \sum_{i=1}^{T}\big[\mathbb{P}(V^* - V_{i+1}^{\pi})(V^* + V_{i+1}^{\pi})\big](s,a)\\
&\leq \frac{2}{1-\gamma}\sum_{i=1}^{T}\Big[\mathbb{P}(V^* - V_{i+1}^{\pi})\Big](s_i,a_i)\\
&\leq \frac{2}{1-\gamma}\sum_{i=1}^{T}(V^*(s_{i+1}) - V_{i+1}^{\pi}(s_{i+1})) + \frac{\sqrt{2T\log(1/\delta)}}{(1-\gamma)^2}\\
&\leq \frac{2}{1-\gamma}\text{Regret}'(T) + \frac{\sqrt{2T\log(1/\delta)}}{1-\gamma} + \frac{2}{(1-\gamma)^2},
\end{aligned}$$

where the first inequality holds because of Lemma C.1, the second inequality holds due to $0 \leq V^*(s), V_{i+1}^{\pi}(s) \leq \frac{1}{1-\gamma}$, the third inequality holds due to the definition of $\mathcal{E}_7$ and the last inequality holds due to $0 \leq V^*(s) \leq V_i(s) \leq 1/1-\gamma$. Thus, we complete the proof. □

F.6. Proof of Lemma D.9

Proof of Lemma D.9.

$$\begin{aligned}
\sum_{i=1}^{T}(\mathbb{V}_{i-1}(s_i,a_i) - \mathbb{V}_i^{\pi}(s_i,a_i)) &= \sum_{i=1}^{T}\mathbb{E}_{s'\sim\mathbb{P}_{i-1}(\cdot|s_i,a_i)}[V_{i-1}^2(s')] - \mathbb{E}_{s'\sim\mathbb{P}_{i-1}(\cdot|s_i,a_i)}[V_{i-1}(s')]^2\\
&\quad - \sum_{i=1}^{T}\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_i,a_i)}[V_{i+1}^{\pi}(s')^2] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_i,a_i)}[V_{i+1}^{\pi}(s')]^2\\
&\leq \underbrace{\sum_{i=1}^{T}\mathbb{E}_{s'\sim\mathbb{P}_{i-1}(\cdot|s_i,a_i)}[V_{i-1}^2(s')] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_i,a_i)}[V_{i-1}^2(s')]}_{I_1}
\end{aligned}$$

$$+\underbrace{\sum_{i=1}^{T} \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_i,a_i)}[V_{i-1}^2(s')] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_i,a_i)}[V_{i+1}^{\pi}(s')^2]}_{I_2}$$

$$+\underbrace{\sum_{i=1}^{T} \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_i,a_i)}[V^*(s')]^2 - \mathbb{E}_{s'\sim\mathbb{P}_{i-1}(\cdot|s_i,a_i)}[V^*(s')]^2}_{I_3},$$

where the inequality holds due to $V_{i-1}(s') \ge V^*(s') \ge V_{i+1}^{\pi}(s')$.

By the definition of event $\mathcal{E}_8$, we have

$$\big\|\mathbb{P}_{i-1}(\cdot|s,a) - \mathbb{P}(\cdot|s,a)\big\|_1 \le \frac{\sqrt{2S\log(T/\delta)}}{\sqrt{N_{i-1}(s,a)\vee 1}}. \tag{F.15}$$

Thus, for the term I_1, since $0 \le V_{i-1}^2(s') \le 1/(1-\gamma)^2$, we have

$$I_1 \le \sum_{i=1}^{T} \frac{\sqrt{2S\log(T/\delta)}}{(1-\gamma)^2\sqrt{N_{i-1}(s_i,a_i)\vee 1}} \le \frac{S\sqrt{2AT\log(3T)\log(T/\delta)}}{(1-\gamma)^2}, \tag{F.16}$$

where the first inequality holds due to (F.15) and the second inequality holds due to Lemma D.3. For the term I_2, on the event $\mathcal{E}_6$, we have

$$\begin{aligned}
I_2 &\le \sum_{i=1}^{T} \Big[\mathbb{P}\big((V_{i-1})^2 - (V_{i+1}^{\pi})^2\big)\Big](s_i,a_i)\\
&= \sum_{i=1}^{T} \big[\mathbb{P}(V_{i-1} - V_{i+1}^{\pi})(V_{i-1} + V_{i+1}^{\pi})\big](s,a)\\
&\le \frac{2}{1-\gamma}\sum_{i=1}^{T} \Big[\mathbb{P}(V_{i-1} - V_{i+1}^{\pi})\Big](s_i,a_i)\\
&\le \frac{2}{1-\gamma}\sum_{i=1}^{T}(V_{i-1}(s_{i+1}) - V_{i+1}^{\pi}(s_{i+1})) + \frac{\sqrt{2T\log(2/\delta)}}{1-\gamma}\\
&\le \frac{4S}{1-\gamma} + \frac{2}{1-\gamma}\sum_{i=1}^{T}(V_{i+1}(s_{i+1}) - V_{i+1}^{\pi}(s_{i+1})) + \frac{\sqrt{2\log(T/\delta)}}{(1-\gamma)^2}\\
&\le \frac{2}{1-\gamma}\text{Regret}'(T) + \frac{\sqrt{2T\log(1/\delta)}}{(1-\gamma)^2} + \frac{4S+2}{(1-\gamma)^2},
\end{aligned} \tag{F.17}$$

where the first inequality holds due to $V_{i-1}(s') \ge V^*(s') \ge V_{i+1}^{\pi}(s')$, the second inequality holds due to $0 \le V_{i-1}(s'), V_{i+1}^{\pi}(s') \le 1/(1-\gamma)$, the third inequality holds due to the definition of event $\mathcal{E}_6$ and the forth inequality holds due to $V_{i-1}(s') \ge V_{i+1}(s')$.

For the term I_3, since $0 \le V^*(s')^2 \le 1/(1-\gamma)^2$, on the event $\mathcal{E}_8$, we have

$$I_3 \le \sum_{i=1}^{T} \frac{\sqrt{2S\log(T/\delta)}}{(1-\gamma)^2\sqrt{N_{i-1}(s_i,a_i)}} \le \frac{S\sqrt{2AT\log(T/\delta)\log(3T)}}{(1-\gamma)^2}, \tag{F.18}$$

where the first inequality holds due to (F.15) and the second inequality holds due to Lemma D.3. Taking an union bound for (F.16), (F.17) and (F.18), with probability at least $1-3\delta$, we have

$$\sum_{i=1}^{t}(\mathbb{V}_{i-1}(s_i,a_i) - \mathbb{V}_i^{\pi}(s_i,a_i)) \le \frac{2\text{Regret}'(T)}{1-\gamma} + \frac{9S\sqrt{2AT\log(T/\delta)\log(3T)}}{(1-\gamma)^2}.$$

□

F.7. Proof of Lemma D.10

Proof of Lemma D.10. For each $i \in [H]$,$s \in \mathcal{S}$ and $t \in [T]$, if $N_t(s) = 0$, the we have

$$\text{Regret}'(t, s, h) = 0 \le \frac{16SAU^2\sqrt{N_t(s)}}{(1-\gamma)^{2.5}} + \frac{20S^2A^{1.5}U^{4.5}}{(1-\gamma)^{3.5}}.$$

Otherwise, we have

$$\begin{aligned}
\text{Regret}'(t, s, h) &= \sum_{1\le i\le t, s_i = s} \gamma^h \big[V_{i+h}(s_{i+h}) - V^\pi_{i+h}(s_{i+h})\big] \\
&= \sum_{1\le i\le t, s_i = s} \gamma^h \big[Q_t(s_{i+h}, a_{i+h}) - V^\pi_{i+h}(s_{i+h})\big] \\
&\le \sum_{1\le i\le t, s_i = s} \gamma^{h+1}[\mathbb{P}_{i+h-1}V_{i+h-1}](s_{i+h}, a_{i+h}) + \gamma^h \text{UCB}_{i+h-1}(s_{i+h}, a_{i+h}) \\
&\quad - \gamma^{h+1}\mathbb{P}V^\pi_{i+h+1}(s_{i+h}, a_{i+h}) \\
&= I_1 + I_2 + I_3 + \gamma^h I_4 + \text{Regret}'(t, s, h+1), \qquad\qquad \text{(F.19)}
\end{aligned}$$

where the first inequality holds due to definition update rule (4.2). $I_1, \ldots, I_4$ are defined as follows.

$$\begin{aligned}
I_1 &= \sum_{1\le i\le t, s_i = s} \gamma^{h+1}(V_{i+h-1}(s_{i+h+1}) - V_{i+h+1}(s_{i+h+1})), \\
I_2 &= \sum_{1\le i\le t, s_i = s} \gamma^{h+1}[(\mathbb{P}_{i+h-1} - \mathbb{P})V_{i+h-1}](s_{i+h}, a_{i+h}), \\
I_3 &= \sum_{1\le i\le t, s_i = s} \gamma^{h+1}\big[\mathbb{P}(V_{i+h-1} - V^\pi_{i+h+1})\big](s_{i+h}, a_{i+h}), \\
&\quad - \gamma^{h+1}\big[V_{i+h-1}(s_{i+h+1}) - V^\pi_{i+h+1}(s_{i+h+1})\big], \\
I_4 &= \sum_{1\le i\le t, s_i = s} \text{UCB}_{i+h-1}(s_{i+h}, a_{i+h}).
\end{aligned}$$

For the term I_1, we have

$$\sum_{1\le i\le t, s_i = s} \gamma^{h+1}(V_{i+h-1}(s_{i+h+1}) - V_{i+h+1}(s_{i+h+1})) \le \sum_{i=1}^{t}\sum_{s'\in\mathcal{S}} V_{i+h-1}(s') - V_{i+h+1}(s') \le \frac{2S}{1-\gamma}, \qquad \text{(F.20)}$$

where the first inequality holds due to $V_{i+h-1}(s') \ge V_{i+h+1}(s')$ and the second inequality holds due to $0 \le V_t(s) \le 1/(1-\gamma)$.

For the term I_2, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{1\le i\le t, s_i = s} \gamma^{h+1}[(\mathbb{P}_{i+h-1} - \mathbb{P})V_{i+h-1}](s_{i+h}, a_{i+h}) \\
&\le \sum_{1\le i\le t, s_i = s} \frac{\gamma^{h+1}\sqrt{2SU}}{(1-\gamma)\sqrt{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}} \\
&\le \frac{\gamma^{h+1}\sqrt{2SU}}{(1-\gamma)}\sqrt{N_t(s)\sum_{1\le i\le t, s_i = s}\frac{1}{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}} \\
&\le \frac{\sqrt{2SU}}{1-\gamma}\sqrt{N_t(s)SAU} \\
&= \frac{SU\sqrt{2N_t(s)A}}{1-\gamma}, \qquad\qquad \text{(F.21)}
\end{aligned}$$

where the first inequality holds due to Lemma D.1 and the definition of U, the second inequality holds due to Cauchy-Schwarz inequality and the third inequality holds due to Lemma D.3.

For the term I_3, Since the random process $s_{i+h+1} \sim \mathbb{P}(\cdot|s_{i+h}, a_{i+h})$ is dependent with whether $s_{i+1}, .., s_{i+h+1} = s$, we cannot directly use Lemma D.1 to bound this term. However, we can use the same technique in the proof of Lemme D.7, which divide the time horizon into H sub-horizon and use Lemma D.1 for each sub-horizon. Compared with the upper bound of I_3 in proof of Theorem 4.3, this technique will lead to a gap of $\sqrt{H}$ and we have

$$\begin{aligned}
&\sum_{1\le i\le t, s_i=s} \gamma^{h+1}\big[\mathbb{P}(V_{i+h-1} - V^{\pi}_{i+h+1})\big](s_{i+h}, a_{i+h}) - \gamma^{h+1}\big[V_{i+h-1}(s_{i+h+1}) - V^{\pi}_{i+h+1}(s_{i+h+1})\big] \\
&\quad \le \frac{\sqrt{2N_t(s)U}}{(1-\gamma)}\sqrt{H} \\
&\quad \le \frac{2U\sqrt{N_t(s)}}{(1-\gamma)^{1.5}},
\end{aligned} \tag{F.22}$$

where the second inequality holds due to the definition of U. For the term I_4, we have

$$\begin{aligned}
&\sum_{1\le i\le t, s_i=s} \mathrm{UCB}_{i+h-1}(s_{i+h}, a_{i+h}) \\
&\quad \le \underbrace{\sum_{1\le i\le t, s_i=s} \sqrt{\frac{8U\mathbb{V}_{i+h-1}(s_{i+h}, a_{i+h})}{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}}}_{I_{41}} + \underbrace{\sum_{1\le i\le t, s_i=s} \frac{8U}{(1-\gamma)\big(N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1\big)}}_{I_{42}} \\
&\quad + \underbrace{\sum_{1\le i\le t, s_i=s} \sqrt{\frac{8\sum_{s'} \mathbb{P}_{i+h}(s'|s_{i+h}, a_{i+h}) \min\big\{100B_{i+h}(s'), 1/(1-\gamma)^2\big\}}{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}}}_{I_{43}}.
\end{aligned} \tag{F.23}$$

For the term I_{41}, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{1\le i\le t, s_i=s} \sqrt{\frac{8U\mathbb{V}_{i+h-1}(s_{i+h}, a_{i+h})}{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}} \\
&\quad \le \sqrt{8U}\sqrt{\sum_{1\le i\le t, s_i=s} \mathbb{V}_{i+h-1}(s_{i+h}, a_{i+h})}\sqrt{\sum_{1\le i\le t, s_i=s} \frac{1}{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}} \\
&\quad \le U\sqrt{8SA}\sqrt{\sum_{1\le i\le t, s_i=s} \mathbb{V}_{i+h-1}(s_{i+h}, a_{i+h})} \\
&\quad \le U\sqrt{8SA}\sqrt{\frac{2N_t(s)}{(1-\gamma)^2}},
\end{aligned} \tag{F.24}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to Lemma D.3, the last inequality holds due to $0 \le \mathbb{V}_{i+h-1}(s_{i+h}, a_{i+h}) \le 1/(1-\gamma)^2$.

For the term I_{42}, by Lemma D.3, we have

$$\sum_{1\le i\le t, s_i=s} \frac{8U}{(1-\gamma)\big(N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1\big)} \le \frac{8SAU^2}{1-\gamma}. \tag{F.25}$$

For the term I_{43}, with probability at least $1-2\delta$, we have

$$\sum_{1\le i\le t, s_i=s} \sqrt{\frac{8\sum_{s'} \mathbb{P}_{i+h}(s'|s_{i+h}, a_{i+h}) \min\big\{100B_{i+h}(s'), 1/(1-\gamma)^2\big\}}{N_{i+h-1}(s_{i+h}, a_{i+h}) \vee 1}}$$

$$
\begin{aligned}
&\le \sqrt{8\sum_{1\le i\le t, s_i=s}\frac{1}{N_{i+h-1}(s_{i+h},a_{i+h})\vee 1}}\\
&\quad\cdot\sqrt{\sum_{1\le i\le t, s_i=s}\sum_{s'}\mathbb{P}_{i+h}(s'|s_{i+h},a_{i+h})\min\left\{100B_{i+h}(s'),\frac{1}{(1-\gamma)^2}\right\}}\\
&\le \sqrt{8SAU}\sqrt{\sum_{1\le i\le t, s_i=s}\sum_{s'}\mathbb{P}_{i+h}(s'|s_{i+h},a_{i+h})\min\left\{100B_{i+h}(s'),\frac{1}{(1-\gamma)^2}\right\}}\\
&\le \sqrt{8SAU}\Bigg[\sum_{1\le i\le t, s_i=s}\Bigg(\frac{\sqrt{SU}}{(1-\gamma)^2\sqrt{N_{i+h-1}(s_{i+h},a_{i+h})\vee 1}}\\
&\quad+\sum_{s'}\mathbb{P}(s'|s,a)\min\left\{100B_{i+h}(s'),\frac{1}{(1-\gamma)^2}\right\}\Bigg)\Bigg]^{1/2}\\
&\le \sqrt{8SAU}\Bigg[\frac{SU\sqrt{AN_t(s)}}{(1-\gamma)^2}+\frac{\sqrt{2N_t(s)U}}{(1-\gamma)^2}\\
&\quad+\sum_{1\le i\le t, s_i=s}\min\left\{\frac{100S^2A^2U^5}{(1-\gamma)^5\big(N_{i+h-1}(s_{i+h+1})\vee 1\big)},\frac{1}{(1-\gamma)^2}\right\}\Bigg]^{1/2}\\
&\le \sqrt{8SAU}\sqrt{\frac{SU\sqrt{AN_t(s)}}{(1-\gamma)^2}+\frac{\sqrt{2N_t(s)U}}{(1-\gamma)^2}+\frac{100S^3A^2U^6}{(1-\gamma)^5}}, \qquad \text{(F.26)}
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to Lemma D.3, the third inequality holds due to Lemma D.1, the forth inequality holds due to Lemma D.1 and the last inequality holds due to Lemma D.3. Substituting (F.20), (F.21), (F.22), (F.23) into (F.19), with probability at least $1-4H\delta$, we have

$$
\mathrm{Regret}'(t,s,h)\le\mathrm{Regret}'(t,s,h+1)+\frac{16SAU\sqrt{N_t(s)}}{(1-\gamma)^{1.5}}+\frac{20S^2A^{1.5}U^{3.5}}{(1-\gamma)^{2.5}}. \qquad \text{(F.27)}
$$

Notice that

$$
\begin{aligned}
\mathrm{Regret}'(t,s,H)&=\sum_{1\le i\le t, s_i=s}\gamma^H\big[V_{i+H}(s_{i+H})-V^\pi_{i+H}(s_{i+H})\big]\\
&\le\sum_{1\le i\le t, s_i=s}\frac{\gamma^H}{1-\gamma}\\
&\le\sum_{1\le i\le t, s_i=s}\frac{1}{T}\\
&\le 1,
\end{aligned}
$$

where the first inequality holds due to $V_{i+H}(s_{i+H})-V^\pi_{i+H}(s_{i+H})\le 1/(1-\gamma)$ and the second inequality holds due to definition of H. Thus, taking summation of (F.27) with all $h\in[H]$, with probability at least $1-H^2\delta$, we have

$$
\mathrm{Regret}'(t,s,0)\le\frac{16SAU^2\sqrt{N_t(s)}}{(1-\gamma)^{2.5}}+\frac{20S^2A^{1.5}U^{4.5}}{(1-\gamma)^{3.5}}. \qquad \text{(F.28)}
$$

In addition, if $N_t(s)>0$, we have

$$
\begin{aligned}
V_t(s)-V^*(s)&\le\frac{1}{N_t(s)}\sum_{1\le i\le t, s_i=s}V_i(s)-V^*(s)\\
&\le\frac{1}{N_t(s)}\sum_{1\le i\le t, s_i=s}\big[V_i(s)-V_i^\pi(s)\big]
\end{aligned}
$$

$$\leq \frac{16SAU^2}{(1-\gamma)^{2.5}\sqrt{N_t(s)}} + \frac{20S^2A^{1.5}U^{4.5}}{(1-\gamma)^{3.5}N_t(s)},$$

where the first inequality holds due to $V_i(s)$ is decreasing, the second inequality holds due to $V^*(s) \geq V_i^\pi(s)$ and the third inequality holds due to (F.28). Notice that when $N_t(s) \geq S^2AU^3/(1-\gamma)^2$, we have

$$V_t(s) - V^*(s) \leq \frac{16SAU^2}{(1-\gamma)^{2.5}\sqrt{N_t(s)}} + \frac{20S^2A^{1.5}U^{4.5}}{(1-\gamma)^{3.5}N_t(s)} \leq \frac{36SAU^2}{(1-\gamma)^{2.5}\sqrt{N_t(s)}}.$$

Otherwise, we have

$$V_t(s) - V^*(s) \leq \frac{1}{1-\gamma} \leq \frac{36SAU^2}{(1-\gamma)^{2.5}\sqrt{N_t(s)}}.$$

Thus, we complete the proof of Lemma D.10. □