Nearly Optimal Policy Optimization with Stable at Any Time Guarantee

Tianhao Wu [* 1] **Yunchang Yang** [* 2 3] **Han Zhong** [* 2] **Liwei Wang** [4] **Simon S. Du** [5] **Jiantao Jiao** [1]

Abstract

Policy optimization methods are one of the most widely used classes of Reinforcement Learning (RL) algorithms. However, theoretical understanding of these methods remains insufficient. Even in the episodic (time-inhomogeneous) tabular setting, the state-of-the-art theoretical result of policy-based method in Shani et al. (2020) is only $\widetilde{O}(\sqrt{S^2 A H^4 K})$ where S is the number of states, A is the number of actions, H is the horizon, and K is the number of episodes, and there is a $\sqrt{SH}$ gap compared with the information theoretic lower bound $\widetilde{\Omega}(\sqrt{SAH^3 K})$ (Jin et al., 2018). To bridge such a gap, we propose a novel algorithm Reference-based Policy Optimization with Stable at Any Time guarantee (RPO-SAT), which features the property "Stable at Any Time". We prove that our algorithm achieves $\widetilde{O}(\sqrt{SAH^3 K} + \sqrt{AH^4 K})$ regret. When $S > H$, our algorithm is minimax optimal when ignoring logarithmic factors. To our best knowledge, RPO-SAT is the first computationally efficient, nearly minimax optimal policy-based algorithm for tabular RL.

1. Introduction

Reinforcement Learning (RL) has achieved phenomenal successes in solving complex sequential decision-making problems (Silver et al., 2016; 2017; Levine et al., 2016; Gu et al., 2017). Most of these empirical successes are driven by policy-based (policy optimization) methods, such as policy gradient (Sutton et al., 1999), natural policy gradient (NPG) (Kakade, 2001), trust region policy optimization

*Equal contribution [1]University of California, Berkeley [2]Center for Data Science, Peking University [3]Peng Cheng Laboratory [4]Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University [5]University of Washington. Correspondence to: Tianhao Wu <thw@berkeley.edu>, Yunchang Yang <yangyc@pku.edu.cn>, Han Zhong <hanzhong@stu.pku.edu.cn>.

(TRPO) (Schulman et al., 2015), and proximal policy optimization (PPO) (Schulman et al., 2017). For example, Haarnoja et al. (2018) proposed a policy-based state-of-the-art reinforcement learning algorithm, soft actor-critic (SAC), which outperformed value-based methods in a variety of real world robotics tasks including manipulation and locomotion. In fact, Kalashnikov et al. (2018) observed that compared with value-based methods such as Q-learning, policy-based methods work better with dense reward. On the other hand, for sparse reward cases in robotics, value-based methods perform better.

Motivated by this, a line of recent work (Fazel et al., 2018; Bhandari & Russo, 2019; Liu et al., 2019; Wang et al., 2019; Agarwal et al., 2021) provides global convergence guarantees for these popular policy-based methods. However, to achieve this goal, they made several assumptions. Agarwal et al. (2021) assumes they have the access to either the exact population policy gradient or an estimate of it up to certain precision for all states uniformly compared with the state distribution induced by π^*, bypassing the hardness of exploration. They showed that even with this stringent assumption, the convergence rate would depend on the distribution mismatch coefficient $D_\infty = \max_s \left(\frac{d_{s_0}^{\pi^*}(s)}{\mu(s)} \right)$, where μ is the starting state distribution of the algorithm and $d_{s_0}^{\pi^*}(s)$ is the stationary state distribution of the optimal policy π^* starting from s_0. This dependency is problematic since D_∞ is small only when the initial distribution has a *good coverage* of the optimal stationary distribution, which may not happen in practice.

However, in online value-based RL, algorithm such as Azar et al. (2017) can achieve fast convergence rate (or regret) independent of the distribution mismatch coefficient, or equivalently, without the *coverage* assumption. Though value-based methods have achieved the information theoretical optimal regret in tabular (Azar et al., 2017) and linear MDPs settings (Zanette et al., 2020), it remains unclear whether policy-based methods can achieve information theoretically optimal regret in the same settings. To address this issue, Cai et al. (2020) proposed the idea of optimism in policy optimization, which seems similar to the value-based optimism but different in nature, since it encourages optimism for Q^π instead of Q^* (Section 4). With this new idea, Cai et al. (2020); Shani et al. (2020) managed to estab-

lish the regret guarantees without additional assumptions, though the regret is suboptimal.

In this work, we focus on the same setting as in Shani et al. (2020): episodic *tabular* MDPs with *unknown transitions*, *stochastic rewards/losses*, and *bandit feedback*. In this setting, the state-of-the-art result of policy-based method is $\widetilde{O}(\sqrt{S^2AH^4K})$ (Shani et al., 2020). Here S and A are the cardinality of states and actions, respectively, H is the episode horizon, and K is the number of episodes. Compared with the information theoretic limit (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018; Domingues et al., 2021), there is still a gap $\sqrt{SH}$.

1.1. Main Contributions

In this paper we present a novel provably efficient policy optimization algorithm, *Reference-based Policy Optimization with Stable at Any Time guarantee* (RPO-SAT). We establish a high probablity regret upper bound $\widetilde{O}(\sqrt{SAH^3K} + \sqrt{AH^4K})$ for our algorithm. Importantly, if $S > H$, the main term in this bound matches the information theoretic limit $\Omega(\sqrt{SAH^3K})$ (Jin et al., 2018), up to some lower order term $\widetilde{O}(\text{poly}(S, A, H)K^{1/4})$. We introduce our algorithmic innovations and analytical innovations as follows:

Algorithmic innovations:

- We introduce a novel reference V estimator in our algorithm. It is conceptually simple and easy to implement as it just updates the reference value to be the mean of empirical V values when some conditions are triggered (cf. Algorithm 1 Lines 18-20).

- We carefully incorporate the reference V estimator into our bonus term, in the way of adding a weighted absolute difference between the estimated V values and the reference V values to control the instability of the estimation process. Readers may refer to Section 4 for more details.

- Another highlight is that we modify the policy improvement phase of our algorithm to meet a novel property, which we called **Stable at Any Time (SAT)**. (For a detailed definition see Equation (10) in Section 4.) More specifically, instead of using the KL-divergence regularization term proposed in Shani et al. (2020), we use the ℓ_2 regularization term. This is crucial to ensure SAT.

Analytical innovations:

- We prove that our algorithm satisfies the SAT property. The analysis is done by two steps: First, we establish a *1*-st step regret bound $\widetilde{O}(\sqrt{S^2AH^4K})$. Second, we use a new technique "*Forward Induction*" to prove the

same for all the h-th step regret. Here the h-step regret is defined in (4). Readers may refer to Section 4 for more details of the "Forward Induction" technique.

- We show that the combination of the SAT property and the simple reference V estimator yields a precise approximation of V^*. We use this property to derive a $O(\sqrt{SAH^3K})$ upper bound for the sum of bonus terms, which leads to a $\sqrt{S}$ reduction in terms of regret.

1.2. Related Work

Our work contributes to the theoretical investigations of policy-based methods in RL (Cai et al., 2020; Shani et al., 2020; Lancewicki et al., 2020; Fei et al., 2020; He et al., 2021; Zhong et al., 2021; Luo et al., 2021; Zanette et al., 2021). The most related policy-based method is proposed by Shani et al. (2020), who also studies the episodic tabular MDPs with unknown transitions, stochastic losses, and bandit feedback. It is important to understand whether it is possible to eliminate this gap and thus achieve minimax optimality, or alternatively to show that this gap is inevitable for policy-based methods.

We also provide interesting practical insights. For example, the usage of reference estimator and ℓ_2 regularization with decreasing learning rate to stabilize the estimate of V and Q. The use of reference estimator can be traced to (Zhang et al., 2020b). They use the reference estimator to maximize the data utilization, hence reduce the estimation variance. However, our usage of reference estimator is different from theirs. The reason why they can reduce a $\sqrt{H}$ factor is because the bottleneck term is only estimated using $1/H$ fraction of data, while the usage of reference can make use of all the data, hence fully utilizing available data. However, for policy-based RL, there is no such problem of data utilization. In fact, the bottleneck of policy-based methods is the *instability* of Q estimation, therefore we use the reference estimator to stabilize the estimation process and reduce a $\sqrt{S}$ factor in the regret. Readers may turn to Section 4 for a detailed explanation for "instability".

Our work is also closely related to another line of work on value-based methods. In particular, Azar et al. (2017); Zanette & Brunskill (2019); Zhang et al. (2020a;b); Menard et al. (2021) have shown that the value-based methods can achieve $\widetilde{O}(\sqrt{SAH^3K})$ regret upper bound, which matches the information theoretic limit. Different from these works, we are the first to prove the (nearly) optimal regret bound for policy-based methods.

2. Preliminaries

A finite horizon stochastic Markov Decision Process (MDP) with time-variant transitions $\mathcal{M}$ is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, P = \{P_h\}_{h=1}^{H}, c = \{c_h\}_{h=1}^{H})$, where $\mathcal{S}$ and $\mathcal{A}$

are finite state and action spaces with cardinality S and A, respectively, and $H \in \mathbb{N}$ is the horizon of the MDP. At time step h, and state s, the agent performs an action a, transitions to the next state s' with probability $P_h(s' \mid s, a)$, and suffers a random cost $C_h(s, a) \in [0, 1]$ drawn i.i.d from a distribution with expectation $c_h(s, a)$.

A stochastic policy $\pi : \mathcal{S} \times [H] \to \Delta_A$ is a mapping from states and time-step indices to a distribution over actions, i.e., $\Delta_A = \{\pi \in \mathbb{R}^A : \sum_a \pi(a) = 1, \pi(a) \geq 0\}$. The performance of a policy π when starting from state s at time h is measured by its value function, which is defined as

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, \pi, P \right]. \quad (1)$$

The expectation is taken with respect to the randomness of the transition, the cost function and the policy. The Q-function of a policy given the state action pair (s, a) at time-step h is defined by

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a, \pi, P \right]. \quad (2)$$

By the above definitions, for any fixed policy π, we can obtain the Bellman equation

$$\begin{aligned} Q_h^\pi(s, a) &= c_h(s, a) + P_h(\cdot \mid s, a)V_{h+1}^\pi(\cdot), \\ V_h^\pi(s) &= \langle Q_h^\pi(s, \cdot), \pi_h(\cdot \mid s)\rangle. \end{aligned} \quad (3)$$

An optimal policy π^* minimizes the value for all states s and time-steps h simultaneously, and its corresponding optimal value is denoted by $V_h^*(s) = \min_\pi V_h^\pi(s)$, for all $h \in [H]$. We consider an agent that repeatedly interacts with an MDP in a sequence of K episodes such that the starting state at the k-th episode, s_1^k, is initialized by a fixed state s_1[*].

In this paper we define the notion of *h-th step regret*, Regret_h, as follows:

$$\text{Regret}_h(K) = \sum_{k=1}^{K} \left(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k)\right). \quad (4)$$

When $h = 1$, this matches the traditional definition of regret, which measures the performance of the agent starting from s_1. In this case we also use $\text{Regret}(K)$ for simplicity.

Notations and Definitions We denote the number of times that the agent has visited state s, state-action pair (s, a) and state-action-transition pair (s, a, s') at the h-th step by $n_h^k(s)$, $n_h^k(s, a)$ and $n_h^k(s, a, s')$ respectively. We denote by $\overline{X}_k$ the empirical average of a random variable X. All quantities are based on experience gathered until the

[*]Our subsequent analysis can be extended to the setting where the initial state is sampled from a fixed distribution.

end of the k-th episode. We denote by $\Delta_\mathcal{A}$ the probability simplex over the action space, i.e. $\Delta_\mathcal{A} = \{(p_1, ..., p_{|\mathcal{A}|}) \mid p_i \geq 0, \sum_i p_i = 1\}$.

We use $\widetilde{O}(X)$ to refer to a quantity that depends on X up to a poly-log expression of a quantity at most polynomial in S, A, K, H and δ^{-1}. Similarly, $\lesssim$ represents $\leq$ up to numerical constants or poly-log factors. We define $X \vee Y := \max\{X, Y\}$.

3. RPO-SAT: Reference-based Policy Optimization with Stable at Any Time guarantee

---

Algorithm 1 Reference-based Policy Optimization with Stable at Any Time guarantee (RPO-SAT)

1: initialize $Q_h(x, a) \leftarrow 0$, $V_h(x, a) \leftarrow 0$ and $V_h^{\text{ref}}(x, a) \leftarrow 0$, $C_0 = \sqrt{S^3 A H^3}$
2: **for** episode $k = 1, ..., K$ **do**
3:     Rollout a trajectory by acting π^k
4:     Update counters and empirical model $n^k = \{n_h^k\}_{h\in[H]}, \overline{c}^k = \{\overline{c}_h^k\}_{h\in[H]}, \overline{P}^k = \{\overline{P}_h^k\}_{h\in[H]}$
5:     **for** Step $h = H, ..., 1$ **do**
6:         **for** $\forall s, a \in \mathcal{S} \times \mathcal{A}$ **do**
7:             Calculate u_h^k as in (5)
8:             $b_h^k(s, a) = \min\{u_h^k(s, a), \sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln\frac{3SAHT}{\delta'}}{n_h^k(s,a)}}\}$
9:             $Q_h^k(s, a) = \max\{\overline{c}_h^k(s, a) + \overline{P}_h^k(\cdot \mid s, a)V_{h+1}^k(\cdot) - b_h^k(s, a), 0\}$
10:             **for** $\forall s \in \mathcal{S}$ **do**
11:                 $V_h^k(s) = \langle Q_h^k(s, \cdot), \pi_h^k(\cdot|s)\rangle$
12:             **end for**
13:         **end for**
14:     **end for**
15:     **for** $\forall h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$ **do**
16:         $\pi_h^{k+1}(\cdot|s) = \operatorname{argmin}_{\pi_h} \eta_k \langle Q_h^k(\cdot, s), \pi_h - \pi_h^k\rangle + \|\pi_h - \pi_h^k\|_2^2$
17:     **end for**
18:     **for** $\forall h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$ such that $n_h^k(s) \geq C_0\sqrt{k}$ **do**
19:         $V_h^{\text{ref}}(s) = \frac{1}{n_h^k(s)} \sum_{i=1}^{k} V_h^i(s_h^i)\mathbf{1}[s_h^i = s]$
20:     **end for**
21: **end for**

In this section, we present our algorithm RPO-SAT (Reference-based Policy Optimization with Stable at Any Time guarantee). The pseudocode is given in Algorithm 1.

We start by reviewing the optimistic policy optimization

algorithms OPPO and POMD proposed by Cai et al. (2020); Shani et al. (2020). In each update, OPPO and POMD involve a policy improvement phase and a policy evaluation phase. In policy evaluation phase, OPPO and POMD explicitly incorporate a UCB bonus function into the estimated Q-function to promote exploration. Then, in the policy improvement phase, OPPO and POMD improve the policy by Online Mirror Descent (OMD) with KL-regularization, where the estimated Q-function serves as the gradient. Compared with existing optimistic policy optimization algorithms in Shani et al. (2020); Cai et al. (2020), our algorithm has three novelties.

First, in policy evaluation phase, we introduce the reference V estimator (Lines 18-20). Specifically, for any $s \in \mathcal{S}$, if the number of visiting s satisfies the condition in Line 18, we update the reference value to be the empirical mean estimator of V (Line 19). Zhang et al. (2020b) also adopts a reference value estimator. Nevertheless their update conditions and methods are different from us, and they reduce a $\sqrt{H}$ factor while we reduce a $\sqrt{S}$ factor.

Second, we make some modifications in policy improvement phase. Specifically, the policy optimization step in the h-th step of the k-th episode is

$$\pi_h^{k+1}(\cdot|s) = \mathrm{argmin}_{\pi_h} \eta_k \langle Q_h^k(\cdot, s), \pi_h - \pi_h^k \rangle + D(\pi_h, \pi_h^k),$$

where η_k is the stepsize in the k-th episode, D is some distance measure and Q_h^k is the estimated Q-function in the h-th step of the k-th episode. Different from Shani et al. (2020); Cai et al. (2020) choosing $D(\pi_h, \pi_h^k) = \mathrm{KL}(\pi_h, \pi_h^k)$, we choose $D(\pi_h, \pi_h^k) = \|\pi_h - \pi_h^k\|_2^2$. In this case the solution of the OMD is:

$$\pi_h^{k+1}(\cdot|s) = \Pi_{\Delta_{\mathcal{A}}} \left(\pi_h^k(\cdot|s) - \eta_k Q_h^k(\cdot, s) \right)$$

where $\Pi_{\Delta_{\mathcal{A}}}$ is the Euclidean projection onto $\Delta_{\mathcal{A}}$, i.e. $\Pi_{\Delta_{\mathcal{A}}}(\boldsymbol{x}) = \mathrm{argmin}_{\boldsymbol{y} \in \Delta_{\mathcal{A}}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$. Unlike previous works, we also adopt a decreasing learning rate schedule instead of a fixed learning rate. These modifications are necessary for our analysis since they ensure the SAT property, making it possible to learn a good reference V value.

Finally, we design a novel bonus term which carefully incorporate the reference V estimator mentioned before by adding a weighted absolute difference Term (iii), which is also referred to as the *instability* term. Specifically, we define

$$u_h^k(s, a) = \mathrm{Term}(i) + \mathrm{Term}(ii) + \mathrm{Term}(iii) + \mathrm{Term}(iv) \tag{5}$$

where

$$\mathrm{Term}(i) = \sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}},$$

$$\mathrm{Term}(ii) = \sqrt{\frac{6\mathbb{V}_{Y \sim \overline{P}_h^k(\cdot|s,a)} V_{h+1}^{\mathrm{ref}}(Y) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s,a)}} +$$
$$\sqrt{\frac{4H\ln\left(\frac{2SAHK}{\delta'}\right)}{S \cdot n_h^k(s,a)}} + \frac{8\sqrt{SH^2}\ln\left(\frac{2SAHK}{\delta'}\right)}{3n_h^k(s,a)},$$

$$\mathrm{Term}(iii) = \sum_{y \in \mathcal{S}} \left(\sqrt{\frac{2\overline{P}_h^k(y|s,a)(1-\overline{P}_h^k(y|s,a))\ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s,a)-1}} \right.$$
$$\left. + \frac{7\ln\left(\frac{2SAHK}{\delta'}\right)}{3n_h^k(s,a)} \right) |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)|,$$

$$\mathrm{Term}(iv) = \sqrt{\frac{4H\ln\frac{3SAHT}{\delta'}}{n_h^k(s,a)} + \frac{7SH\ln\left(\frac{2SAHK}{\delta'}\right)}{3n_h^k(s,a)}}$$
$$+ \frac{2S^{3/2}A^{1/4}H^{7/4}K^{1/4}\sqrt{\ln\left(\frac{2SAHK}{\delta'}\right)}}{n_h^k(s,a)}.$$

where (i) is the estimation error for reward functions and (ii)-(iv) are estimation errors for transition kernels. We will provide more explanations for this seemingly complicated term in Section 4. Furthermore, we set the bonus as

$$b_h^k(s, a) = \min \left\{ u_h^k(s, a), \right.$$
$$\left. \sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln\frac{3SAHT}{\delta'}}{n_h^k(s,a)}} \right\}. \tag{6}$$

With this carefully chosen bonus function, we can also achieve optimism like previous work in optimistic policy optimization (Shani et al., 2020; Cai et al., 2020). Notably, our bonus function is smaller than that in Shani et al. (2020), thus leads to our tighter regret bound.

Now we state our main theoretical results for RPO-SAT.

Theorem 3.1 (Regret bound)**.** *Suppose in Algorithm 1, we choose* $\eta_t = O(\sqrt{1/(H^2 At)})$, *and* $C_0 = \sqrt{S^3 A H^3}$, *then for sufficiently large K, we have*

$$\mathrm{Regret}(K) \leq \widetilde{O}(\sqrt{SAH^3K} + \sqrt{AH^4K} +$$
$$S^{5/2}A^{5/4}H^{3/2}K^{1/4}).$$

We provide a proof sketch in Section 5. The full proof is in appendix B. Note that previous literature shows that the regret lower bound is $\Omega(\sqrt{SAH^3K})$ (Jin et al., 2018). Hence our result matches the information theoretic limit up to logarithmic factors when $S > H$.

4. Technique Overview

In this section, we illustrate the main steps of achieving near optimal regret bound and introduce our key techniques.

Achieving Optimism via Reference. Similar to previous works (Cai et al., 2020; Shani et al., 2020), to achieve optimism, a crucial step is to design proper bonus term to upper bound $(\overline{P}_h^k - P_h)(\cdot \mid s,a)V_{h+1}^k(\cdot)$. For example, Jaksch et al. (2010); Shani et al. (2020) bound this term in a separate way:

$$\left|(\overline{P}_h^k - P_h)(\cdot \mid s,a)V_{h+1}^k(\cdot)\right| \tag{7}$$

$$\leq \|(\overline{P}_h^k - P_h)(\cdot \mid s,a)\|_1 \cdot \|V_{h+1}^k(\cdot)\|_\infty \leq \widetilde{O}\left(\sqrt{\frac{SH^2}{n_h^k(s,a)}}\right),$$

which will leads to an additional $\sqrt{S}$ factor due to the absence of making use of the optimism of Q^k. This issue is later addressed by value-based algorithm UCBVI (Azar et al., 2017). They divide the term into two separate terms:

$$(\overline{P}_h - P_h)(\cdot \mid s,a)V_{h+1}^k(\cdot) = (\overline{P}_h^k - P_h)(\cdot \mid s,a)V_{h+1}^*(\cdot)$$

$$+ (\overline{P}_h^k - P_h)(\cdot \mid s,a)(V_{h+1}^k - V_{h+1}^*)(\cdot). \tag{8}$$

They bound the first term using straightforward application of Chernoff-Hoeffding inequality, which removes the $\sqrt{S}$ factor since V^* is deterministic. Thanks to the fact that $V_h^k \leq V_h^*$ for any $(k,h) \in [K] \times [H]$, they can bound the second term successfully (see Appendix C for more details). However, this approach is not applicable for policy-based methods which improve the policy in a conservative way instead of choosing the greedy policy (when the stepsize $\eta \to \infty$, the OMD update becomes the greedy policy, and for any $\eta < \infty$ this update is "conservatively" greedy). This key property of policy-based methods makes it only possible to ensure $V_h^k \leq V_h^{\pi^k}$, the optimism $V_h^k \leq V_h^*$ doesn't hold in general. To tackle this challenge, we notice that as long as V_h^k converges to V_h^* sufficiently fast (at least on average), $(\overline{P}_h^k - P_h)(\cdot \mid s,a)(V_{h+1}^k - V_{h+1}^*)(\cdot)$ can be bounded by a term related to the rate of convergence. This leads to the important notion called **Stable at Any Time (SAT)**. Precisely, we introduce the *reference V estimator* $V^{\text{ref}} = \{V_h^{\text{ref}}\}_{h\in[H]}$ and decompose $(\overline{P}_h^k - P_h)(\cdot \mid s,a)V_{h+1}^k(\cdot)$ as

$$\begin{aligned}
&(\overline{P}_h^k - P_h)(\cdot \mid s,a)V_{h+1}^k(\cdot) \\
&= \underbrace{(\overline{P}_h^k - P_h)(\cdot \mid s,a)V_{h+1}^*(\cdot)}_{(a)} \\
&+ \underbrace{(\overline{P}_h^k - P_h)(\cdot \mid s,a)(V_{h+1}^k - V_{h+1}^{\text{ref}})(\cdot)}_{(b)} \\
&+ \underbrace{(\overline{P}_h^k - P_h)(\cdot \mid s,a)(V_{h+1}^{\text{ref}} - V_{h+1}^*)(\cdot)}_{(c)}.
\end{aligned} \tag{9}$$

By standard concentration inequalities, we can bound the first term by the quantity depending on the variance of V_{h+1}^*. Once we have SAT, an easy implication is that $|V_h^{\text{ref}}(s) - V_h^*(s)| \leq O(\sqrt{H/S})$ for any $n_h^k(s)$ large enough (Lemma 5.3). We can replace the variance of V_{h+1}^* (unknown) by the variance of V_{h+1}^{ref} (known) and bound the second and third term in an analogous way of (7) and remove the factor $\sqrt{S}$ as desired. Specifically, we can upper bound Terms $(a),(b),(c)$ in (9) by Terms $(ii),(iii),(iv)$ in (5), respectively.

Stable at Any Time. The high-level idea is that in order to control $(\overline{P}_h^k - P_h)(\cdot \mid s,a)(V_{h+1}^k - V_{h+1}^*)(\cdot)$, term $V_{h+1}^k - V_{h+1}^*$ must satisfy some properties. For example, Zhang et al. (2020b) shows that in the senario of greedy policy, this term converges to zero as the number of visit goes to infinity. However, due to the nature of conservative policy update scheme, we cannot guarantee the convergence, unless we make the *coverage* assumption as in Agarwal et al. (2021). Fortunately, it's possible to obtain an average convergence guarantee, which we called SAT. Specifically, we say that an algorithm satisfy the property of SAT if for all K' and h,

$$\sum_{k=1}^{K'} |V_h^k(s_h^k) - V_h^*(s_h^k)| \leq O(\sqrt{K'}). \tag{10}$$

The meaning of **Stable at Any Time** can be interpreted as follows: The above inequality implies that the estimation V_h^k varies around the fixed value V_h^*, hence **"stable"**. And since we impose that the inequality holds for all h and K', hence **"at any time"**. For this reason, we also name $\sum_{y\in\mathcal{S}} \sqrt{\frac{\overline{P}_h^k(y|s,a)}{n_h^k(s,a)}}|V_{h+1}^k(y) - V_{h+1}^{\text{ref}}(y)|$ in the bonus term (5) as the **instability** term, since it's an upper bound of $(\overline{P}_h^k - P_h)(\cdot \mid s,a)(V_{h+1}^k - V_{h+1}^{\text{ref}})(\cdot)$, which measures the instability of V^k with respect to the fixed reference V^{ref}.

Another nice implication is that we can obtain a precise estimate of V^* if SAT is satisfied. Specifically, combining the update rules of reference V estimator and the fact that $n_h^k(s) \geq C_0\sqrt{k}$, we have:

$$\begin{aligned}
&|V_h^{\text{ref}}(s) - V_h^*(s)| \\
&\leq \frac{1}{n_h^k(s)} \sum_{i=1}^k |V_h^i(s_h^i) - V_h^*(s_h^i)|\mathbf{1}[s_h^i = s] \leq O(1/C_0).
\end{aligned}$$

Coarse Regret Bound. We are able to show that (10) can be implied by

$$\text{Regret}_1(K') \leq O(\sqrt{K'}). \tag{11}$$

for any $K' \in [K]$. See Lemma 5.3 for details. We point out that establishing Equation (10) from coarse regret bound is highly non-trivial. There exists two challenges:

1. Previous policy-based methods (Cai et al., 2020; Shani et al., 2020) choose a constant mirror descent stepsize depending on K, which impedes us from obtaining regret bound sublinear on K' for all $K' \in [K]$.

2. For the step $2 \leq h \leq H$, the state s_h^k is not fixed, which makes the OMD term difficult to bound.

To this end, we use the following two novel techniques:

1. Replace the unbounded KL-divergence regularization term by the bounded ℓ_2 regularization term, which allows us to choose varying stepsize which depends on the current time step instead of K.

2. *Forward Induction*, that is, deriving the regret upper bound for the *1*-st step and obtaining the regret upper bounds for h-th ($2 \leq h \leq H$) step by induction. See Section 5 for more details.

Putting these together, we can remove the additional factor $\sqrt{S}$ as desired. If we also adopt the Bernstein variance reduction technique (Azar et al., 2017), we can further improve a $\sqrt{H}$ factor in the main term.

5. Proof Sketch of Theorem 3.1

In this section, our goal is to provide a sketch of the proof of Theorem 3.1. One key ingredient in regret analysis in RL is the optimism condition, namely

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) \leq 0. \tag{12}$$

This condition guarantees that Q_h^k is an optimistic estimation of $Q_h^{\pi^k}$ (or Q_h^*), and that the regret can be bounded by the sum of bonus terms (Jaksch et al., 2010; Azar et al., 2017; Cai et al., 2020; Shani et al., 2020).

In our case, it does not appear straightforward to show such condition hold or not since the bonus term b_h^k depends on the value of V^{ref}, which we do not know in advance. Dealing with this problem needs additional effort. But let's put it aside for a while, and assumes that (12) holds temporarily. In Section 5.1, we first demonstrate the intuition of the proof under this assumption. Then in Section 5.2, we show how to remove this assumption, with additional techniques.

5.1. A warm-up: the Optimism Assumption

In this section, we first demonstrate the intuition of the proof, under the assumption that optimism condition (12) holds. We first recall the useful notion of h-th step regret in episode K' as follows:

$$\mathrm{Regret}_h(K') = \sum_{k=1}^{K'} \left(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k) \right).$$

The notion aligns with the normal regret if $h = 1$, in other words, we have $\mathrm{Regret}_1(K') = \mathrm{Regret}(K')$. We first state the following lemma, which bounds the estimation error using bonus function:

Lemma 5.1 (Bounding estimation error)**.** *Suppose that optimism holds, namely for* $\forall k \leq K, h \leq H, s \in \mathcal{S}, a \in \mathcal{A}$,

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot|s,a)V_{h+1}^k \leq 0.$$

Then for $\forall K' \leq K, h' \leq H$, *it holds with high probability that*

$$\sum_{k=1}^{K'} \left(V_h^{\pi^k}(s_h^k) - V_h^k(s_h^k) \right)$$

$$\leq \widetilde{O}\left(\sum_{k=1}^{K'} \sum_{h=h'}^{H} b_h^k(s_h^k, a_h^k) \right) + \widetilde{O}(\sqrt{H^3 K'}).$$

Proof. See §B.1 for a detailed proof. □

The above lemma is useful since it tells us that the sum of the estimation error $V_h^{\pi^k}(s_h^k) - V_h^k(s_h^k)$ can be roughly viewed as the sum of bonus function. By the choice of our bonus term, we have $b_h^k(s,a) \leq \sqrt{\frac{SH^2}{n^k(s,a)}}$. Following standard techniques from (Shani et al., 2020), we get a easy corollary $\sum_{k=1}^{K'} \left(V_h^{\pi^k}(s_h^k) - V_h^k(s_h^k) \right) \leq \widetilde{O}(\sqrt{S^2AH^4K'})$.

Next, we show that our algorithm satisfies a coarse regret bound:

Lemma 5.2 (Coarse regret bound)**.** *With the same assumptions and notations in Lemma 5.1, we have for* $\forall K' \leq K$:

$$\mathrm{Regret}(K') = \sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1^k) - V_1^*(s_1^k) \right) \leq \widetilde{O}(\sqrt{S^2AH^4K'}). \tag{13}$$

Proof. We first decompose the regret term in the following way:

$$\sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1) - V_1^*(s_1) \right)$$

$$= \underbrace{\sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1) - V_1^k(s_1) \right)}_{(i)} + \underbrace{\sum_{k=1}^{K'} \left(V_1^k(s_1) - V_1^*(s_1) \right)}_{(ii)} \tag{14}$$

Term (i) can be bounded by $\widetilde{O}(\sqrt{S^2AH^4K'})$ using the corollary of Lemma 5.1. We take a closer look at Term (ii),

by standard regret decomposition lemma (Lemma A.1):

Term $(ii) =$

$$\underbrace{\sum_{k,h} \mathbb{E}\left[\langle Q_h^k(s_h,\cdot), \pi_h^k(\cdot \mid s_h) - \pi_h^*(\cdot \mid s_h)\rangle \mid s_1, \pi^*, P\right]}_{(iii)}$$

$$+\underbrace{\sum_{k,h} \mathbb{E}\left[Q_h^k(s_h,a_h) - c_h(s_h,a_h) - P_h(\cdot \mid s_h,a_h)V_{h+1}^k \mid s_1, \pi^*, P\right]}_{(iv)}.$$

Term (iii) is also called the OMD term. Using Lemma B.6, we have

$$\text{Term } (iii) \leq \widetilde{O}(\sqrt{AH^4K'}).$$

We note that the reason why we change the Bregman term and the learning rate schedule is to ensure a Term $(iii) \leq \widetilde{O}(\sqrt{K'})$ type bound. In (Shani et al., 2020), they use the KL-divergence as the Bregman term and a fixed learning rate that is a function of K, hence their bound is Term $(iii) \leq \widetilde{O}(\sqrt{\log(A)H^4K})$. Although our choice of the ℓ_2 penalty term and decreasing learning rate leads to a larger dependence on A, but it has a better dependence on K', which is crucial to ensure that we can learn a good reference function.

Combining with the fact that optimism holds, we have Term $(iv) \leq 0$, therefore the lemma is proved. $\square$

Now we have a $\text{Regret}(K') = \widetilde{O}(\sqrt{K'})$ bound for all $K' \leq K$. We show that this immediately implies the following key theorem, which we called the Average Convergence Lemma. The following lemma guarantees that the reference value is a good approximation of V^*:

Lemma 5.3 (Average convergence of V^k). *With the same assumptions and notations in Lemma 5.1, we have for all $K' \leq K, h \leq H$,*

$$\text{Regret}_h(K') \leq \widetilde{O}(\sqrt{S^2AH^4K'}). \tag{15}$$

As a consequence, SAT holds, for $\forall K' \leq K, h \leq H$,

$$\sum_{k=1}^{K'} |V_h^k(s_h^k) - V_h^*(s_h^k)| \leq \widetilde{O}(\sqrt{S^2AH^4K'}).$$

Proof. For the proof of Equation (15), we use a novel technique called *Forward Induction*, starting from the case when $h = 1$, $\text{Regret}_1(K') = \text{Regret}(K') \leq \widetilde{O}(\sqrt{S^2AH^4K'})$, which is true according to Lemma 13. Using induction we can proof Equation (15) for all h, which we will leave the details to Appendix B.3. For the second statement, we note

that

$$|V_h^k(s_h^k) - V_h^*(s_h^k)|$$
$$\leq |V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)| + |V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k)|$$
$$= \left(V_h^{\pi^k}(s_h^k) - V_h^k(s_h^k)\right) + \left(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k)\right).$$

These two terms can be handled by Lemma 5.1 and Equation (15) separately, hence finished the proof of Lemma 5.3. $\square$

As mentioned before, Theorem 5.3 is significant in the sense that it implies the reference value $V_h^{\text{ref}}(s) = \frac{1}{n_h^k(s)} \sum_{i=1}^{k} V_h^i(s_h^i)\mathbf{1}[s_h^i = s]$ is close to the optimal value if $n_h^k(s)$ is large enough, in other words, if $n_h^k(s) \geq C_0\sqrt{k} = \sqrt{S^3AH^3k}$:

$$|V_h^{\text{ref}}(s) - V_h^*(s)|$$
$$\leq \frac{1}{n_h^k(s)} \sum_{i=1}^{k} |V_h^i(s_h^i) - V_h^*(s_h^i)|\mathbf{1}[s_h^i = s]$$
$$\leq \frac{1}{n_h^k(s)} \sum_{i=1}^{k} |V_h^i(s_h^i) - V_h^*(s_h^i)|$$
$$\leq \widetilde{O}\left(\sqrt{\frac{H}{S}}\right)$$

This explains the reason why we choose $C_0 = \sqrt{S^3AH^3}$ in the algorithm, the $1/\sqrt{S}$ bound serves a key role in reducing the $\sqrt{S}$ factor in the regret.

Finally, we are ready to prove Theorem 3.1. Combining Lemma 5.1 and Equation (14),

$$\sum_{k=1}^{K}\left(V_1^{\pi^k}(s_1^k) - V_1^*(s_1^k)\right)$$
$$\leq \widetilde{O}\left(\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k)\right) + \widetilde{O}(\sqrt{AH^4K}). \tag{16}$$

The only thing left is to prove that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \leq \widetilde{O}\left(\sqrt{SAH^3K} + S^{\frac{5}{2}}A^{\frac{5}{4}}H^{\frac{11}{4}}K^{\frac{1}{4}}\right).$$

By Lemma B.7, this is true. Therefore we have finished the proof.

5.2. How to prove optimism

In this section, we show how to remove the optimism assumption in Lemma 5.1, Lemma 13 and Lemma 5.3. In fact, we will first prove the following lemma, which shows that both optimism and coarse regret bound holds for all episodes.

Lemma 5.4 (Coarse analysis of regret). *For any $K' \in [K]$, we have the following regret bound:*

$$\text{Regret}(K') = \sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1^k) - V_1^*(s_1^k)\right) \leq \widetilde{O}(\sqrt{S^2 AH^4 K'}). \tag{17}$$

and the optimism holds, namely

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) \leq 0. \tag{18}$$

The full proof is in Appendix B.2. Here we mainly explain the core idea. To prove Lemma 5.4 we use induction. For the first few episodes of the algorithm, the bonus term is dominated by $\sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln\frac{3SAHT}{\delta'}}{n_h^k(s,a)}}$, which is the same as in Shani et al. (2020). Therefore the optimism and coarse regret bound automatically holds.

In the induction step, assume that optimism and coarse regret holds for all $k \leq K'$ and h. Then by optimism we have that 1-step regret is bounded by $\widetilde{O}(\sqrt{S^2 AH^4 K'})$. Then by Lemma 5.3, we have that $\text{Regret}_h(K') \leq \widetilde{O}(\sqrt{S^2 AH^4 K'})$. This means that if in the $K' + 1$-th episode, a new reference function is calculated, then this new reference function must be $\sqrt{H/S}$ accurate, i.e. $|V^{\text{ref}}(s) - V^*(s)| \leq O(\sqrt{H/S})$. Hence, the carefully designed bonus guarantees that optimism still holds for $k = K' + 1$, which finishes the proof. The induction process is shown in Figure 1.

Finally, thanks to optimism, we can bound the regret by $\sum_{k=1}^{K} \sum_{h=1}^{H} b_h^k(s_h^k, a_h^k)$. The rest follows from the same arguments in the previous section.
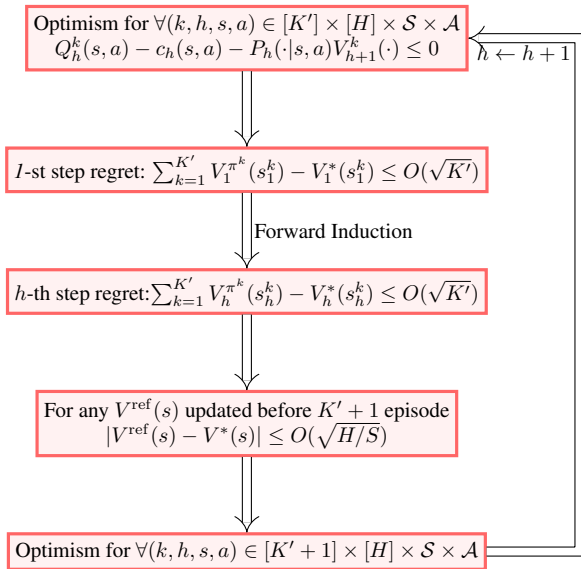

Figure 1. Proving optimism and h-th step regret by induction

6. Conclusion and Future Work

In this paper, we proposed the first optimistic policy optimization algorithm for tabular, episodic RL that can achieve regret guarantee $\widetilde{O}(\sqrt{SAH^3K} + \sqrt{AH^4K} + \text{poly}(S, A, H)K^{1/4})$. This algorithm improves upon previous results (Shani et al., 2020) and matches the information theoretic limit $\Omega(\sqrt{SAH^3K})$ when $S > H$. Our results also raise a number of promising directions for future work. Theoretically, can we design better policy-based methods that can eliminate the constraint $S > H$? Practically, can we leverage the insight of RPO-SAT to improve practical policy-based RL algorithms? Specifically, can we design different regularization terms to stabilize the V/Q estimation process to make the algorithm more sample efficient? We look forward to answering these questions in the future.

Acknowledgement

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 1467–1476. PMLR, 2018.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. Dynamic regret of policy optimization in non-stationary environments. *arXiv preprint arXiv:2007.00148*, 2020.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

He, J., Zhou, D., and Gu, Q. Nearly optimal regret for learning adversarial mdps with linear function approximation. *arXiv preprint arXiv:2102.08940*, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Learning adversarial markov decision processes with delayed feedback. *arXiv preprint arXiv:2012.14843*, 2020.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pp. 10565–10576, 2019.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *arXiv preprint arXiv:2107.08346*, 2021.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Menard, P., Domingues, O. D., Shang, X., and Valko, M. Ucb momentum q-learning: Correcting the bias without forgetting. *arXiv preprint arXiv:2103.01312*, 2021.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

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

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020b.

Zhong, H., Yang, Z., Wang, Z., and Szepesvári, C. Optimistic policy optimization is provably efficient in non-stationary mdps. *arXiv preprint arXiv:2110.08984*, 2021.

A. Regret Decomposition and Failure Events

A.1. Regret Decomposition

For any $(k, h) \in [K] \times [H]$, we define

$$
\begin{aligned}
\zeta_{k,h}^1 &= [V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)] - [Q_h^k(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k)], \\
\zeta_{k,h}^2 &= [P_h(\cdot \mid s_h^k, a_h^k)V_{h+1}^k(\cdot) - P_h(\cdot \mid s_h^k, a_h^k)V_{h+1}^{\pi^k}(\cdot)] - [V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)].
\end{aligned}
\tag{19}
$$

By the definition, we have that $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ represent the randomness of executing a stochastic policy $\pi_h^k(\cdot \mid s_h^k)$ and the randomness of observing the next state from stochastic transition kernel $P_h(\cdot \mid s_h^k, a_h^k)$, respectively. With these notations, we have the following standard regret decomposition lemma (Cai et al., 2020; Shani et al., 2020).

Lemma A.1 (Regret Decomposition). *For any $(K', h') \in [K] \times [H]$, it holds that*

$$
\begin{aligned}
\mathrm{Regret}_{h'}(K') &= \underbrace{\sum_{k=1}^{K'} \left(V_{h'}^{\pi^k}(s_{h'}^k) - V_{h'}^k(s_{h'}^k) \right)}_{(i)} + \underbrace{\sum_{k=1}^{K'} \left(V_{h'}^k(s_{h'}^k) - V_{h'}^*(s_{h'}^k) \right)}_{(ii)} \\
&= \underbrace{\sum_{k=1}^{K'} \sum_{h=h'}^{H} \left[c_h(s_h^k, a_h^k) + P_h(\cdot \mid s_h^k, a_h^k)V_{h+1}^k(\cdot) - Q_h^k(s_h^k, a_h^k) \right]}_{(i.1)} + \underbrace{\sum_{k=1}^{K'} \sum_{h=h'}^{H} (\zeta_{k,h}^1 + \zeta_{k,h}^2)}_{(i.2)} \\
&\quad + \underbrace{\sum_{k=1}^{K'} \sum_{h=h'}^{H} \mathbb{E}\left[\langle Q_h^k(s_h, \cdot), \pi_h^k(\cdot \mid s_h) - \pi_h(\cdot \mid s_h) \rangle \mid s_{h'} = s_{h'}^k, \pi, P \right]}_{(ii.1)} \\
&\quad + \underbrace{\sum_{k=1}^{K'} \sum_{h=h'}^{H} \mathbb{E}\left[Q_h^k(s_h, a_h) - c_h(s_h, a_h) - P_h(\cdot \mid s_h, a_h)V_{h+1}^k(\cdot) \mid s_{h'} = s_{h'}^k, \pi, P \right]}_{(ii.2)}
\end{aligned}
$$

Proof. See Cai et al. (2020) for a detailed proof. □

A.2. Failure Events

Definition A.2. We define the following failure events:

$$F_k^0 = \left\{ \exists h : \left| \sum_{k'=1}^{k} \sum_{h'=h}^{H} (\zeta_{k',h'}^1 + \zeta_{k',h'}^2) \right| \geq \sqrt{16H^3 K \ln \frac{2H}{\delta'}} \right\},$$

$$F_k^1 = \left\{ \exists s, a, h : \left| c_h(s,a) - \bar{c}_h^k(s,a) \right| \geq \sqrt{\frac{2 \ln \frac{2SAHT}{\delta'}}{n_h^k(s,a)}} \right\},$$

$$F_k^2 = \left\{ \exists s, a, h : \left\| P_h(\cdot \mid s,a) - \overline{P}_h^k(\cdot \mid s,a) \right\|_1 \geq \sqrt{\frac{4S \ln \frac{3SAHT}{\delta'}}{n_h^k(s,a)}} \right\},$$

$$F_k^3 = \left\{ \exists s, a, h : \left| \left(P_h(\cdot \mid s,a) - \overline{P}_h^k(\cdot \mid s,a) \right) V_h^*(\cdot) \right| \geq H\sqrt{\frac{4 \ln \frac{3SAHT}{\delta'}}{n_h^k(s,a)}} + \frac{2H \ln \frac{2SAHT}{\delta'}}{3n_h^k(s,a)} \right\},$$

$$F_k^4 = \left\{ \exists s', s, a, h : \left| \overline{P}_h^k(s' \mid s,a) - P_h(s' \mid s,a) \right| \geq \sqrt{\frac{2\overline{P}_h^k(s' \mid s,a)(1 - \overline{P}_h^k(s' \mid s,a)) \ln \left(\frac{2SAHK}{\delta'} \right)}{n_h^k(s,a) - 1}} + \frac{7 \ln \left(\frac{2SAHK}{\delta'} \right)}{3n_h^k(s,a)} \right\},$$

$$F = \bigcup_{k=1}^{K} \left(F_k^0 \cup F_k^1 \cup F_k^2 \cup F_k^3 \cup F_k^4 \right),$$

where $\delta' = \frac{\delta}{5}$.

Lemma A.3. *It holds that*

$$\Pr(F) \leq \delta.$$

Proof. These standard concentration inequalities also appear in Azar et al. (2017); Cai et al. (2020); Shani et al. (2020). For completeness, we present the proof sketch here.

Note that the martingales $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ defined in (19) satisfy $|\zeta_{k,h}^1 + \zeta_{k,h}^2| \leq 4H$. By Azuma-Hoeffding inequality, we have $\Pr(F^0) \leq \delta'$.

Let $F^1 = \bigcup_{k=1}^{K} F_k^1$. By Hoeffding's inequality, we have

$$\Pr \left\{ \left| c_h(s,a) - \bar{c}_h^k(s,a) \right| \geq \sqrt{\frac{2 \ln \frac{1}{\delta'}}{n_h^k(s,a)}} \right\} \leq \delta' \qquad (20)$$

Using a union bound over all s, a and all possible values of $n_k(s,a)$ and k., we have $\Pr \{F^c\} \leq \delta'$.

Let $F^2 = \bigcup_{k=1}^{K} F_k^2$. Then $\Pr \{F^2\} \leq \delta'$, which is implied by (Weissman et al., 2003) while applying union bound on all s, a and all possible values of $n_k(s,a)$ and k.

Let $F^3 = \bigcup_{k=1}^{K} F_k^3$. According to (Azar et al., 2017), we have that with probability at least $1 - \delta'$

$$\left| \left[\left(P_h - \overline{P}_h^k \right) V_h^* \right] (s,a) \right| \leq \sqrt{\frac{2H^2 \ln \left(\frac{2HK}{\delta'} \right)}{n_h^k(s,a)}} + \frac{2H \ln \left(\frac{2HK}{\delta'} \right)}{3n_h^k(s,a)}. \qquad (21)$$

Take a union bound over s, a, k, we have $\Pr \{F^3\} \leq \delta'$.

Let $F^4 = \bigcup_{k=1}^{K} F_k^4$. The Empirical Bernstein inequality (Theorem 4 in (Maurer & Pontil, 2009)) combined with a union bound argument on $s, a, s', n_h^k(s,a)$ also implies the following bound holds with probability at least $1 - \delta'$:

$$\left| \overline{P}_h^k(s' \mid s,a) - P_h(s' \mid s,a) \right| \leq \sqrt{\frac{2\overline{P}(s' \mid s,a)(1 - \overline{P}(s' \mid s,a)) \ln \left(\frac{2SAHK}{\delta'} \right)}{n_h^k(s,a) - 1}} + \frac{7 \ln \left(\frac{2SAHK}{\delta'} \right)}{3n_h^k(s,a)} \qquad (22)$$

Therefore, $\Pr\{F^4\} \leq \delta'$

Finally, take a union bound with $\delta' = \frac{\delta}{5}$, we have $\Pr\{F\} \leq \delta$. $\qquad\square$

Below we will assume that the failure event F does not happen, which is with high probability.

B. Missing Proofs for Section 5

B.1. Proof of Lemma 5.1

Proof. By Lemma A.1, we have

$$\sum_{k=1}^{K'} \left(V_h^{\pi_k}(s_h^k) - V_h^k(s_h^k) \right)$$

$$+ \sum_{k=1}^{K'} \sum_{h=h'}^{H} \mathbb{E}\left[\langle Q_h^k(s_h, \cdot), \pi_h^k(\cdot \mid s_h) - \pi_h(\cdot \mid s_h) \rangle \mid s_{h'} = s_{h'}^k, \pi, P \right] + \sum_{k=1}^{K'} \sum_{h=h'}^{H} (\zeta_{k,h}^1 + \zeta_{k,h}^2).$$

Here $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ are martingales defined in (19) satisfying $|\zeta_{k,h}^1 + \zeta_{k,h}^2| \leq 4H$. Under Azuma-Hoeffding inequalities and the optimistic assumption that $-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot|s,a)V_{h+1}^k \leq 0$, we finish the proof of Lemma 5.1. $\qquad\square$

B.2. Full Proof of Theorem 3.1

As discussed in Section 5, we only need to prove the coarse regret in Lemma 5.4, but this time without the optimism assumption. We restate the lemma for ease of reading.

Lemma B.1 (Coarse analysis of dynamic regret, restatement of Lemma 5.4). *Conditioned on F^c, for any $K' \in [K]$, we have the following regret bound:*

$$\text{Regret}(K') = \sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1^k) - V_1^*(s_1^k) \right) \leq \widetilde{O}(\sqrt{S^2 A H^4 K'}). \tag{23}$$

and the optimism holds, namely

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) \leq 0. \tag{24}$$

Remark B.2. In our algorithm, we change the Bregman penalty term from KL-divergence $d_{KL}(\pi_h \| \pi_h^k)$ to $\|\pi_h - \pi_h^k\|_2^2$, we also let the learning rate to be of the form $\eta_t = O(\frac{1}{\sqrt{t}})$ instead of a constant depending on K. We note that although this will cause more regret in the OMD term, it is crucial to obtain a $\text{Regret}(K') = O(\sqrt{K'})$ bound instead of $\text{Regret}(K') = O(K')$.

Proof. First, by Lemma A.1, we decompose the regret in the following way.

$$\sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1) - V_1^*(s_1) \right)$$

$$= \underbrace{\sum_{k=1}^{K'} \sum_{h=1}^{H} \left[c_h(s_h^k, a_h^k) + P_h(\cdot \mid s_h^k, a_h^k) V_{h+1}^k(\cdot) - Q_h^k(s_h^k, a_h^k) \right]}_{(i)} + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} (\zeta_{k,h}^1 + \zeta_{k,h}^2)}_{(ii)}$$

$$+ \underbrace{\sum_{k=1}^{K'} \sum_{h=1}^{H} \mathbb{E} \left[\langle Q_h^k(s_h, \cdot), \pi_h^k(\cdot \mid s_h) - \pi_h(\cdot \mid s_h) \rangle \mid s_1 = s_1^k, \pi, P \right]}_{(iii)}$$

$$+ \underbrace{\sum_{k=1}^{K'} \sum_{h=1}^{H} \mathbb{E} \left[Q_h^k(s_h, a_h) - c_h(s_h, a_h) - P_h(\cdot \mid s_h, a_h) V_{h+1}^k(\cdot) \mid s_1 = s_1^k, \pi, P \right]}_{(iv)}.$$

We first prove (23) and (24) for early episodes, namely when $n_h^k < C_0 \sqrt{k}$ for all s (which is true at the beginning of the algorithm), then the regret bound holds. Note that in this case, conditioned on F^c, we have $\frac{SH\sqrt{C_0}K'^{1/4}}{n_h^k(s,a)} \geq \sqrt{\frac{H^2 S}{n_h^k(s,a)}}$, which implies $b_h^k(s,a) = \sqrt{\frac{2\ln \frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln \frac{3SAHT}{\delta'}}{n_h^k(s,a)}}$. In this case, by Lemma B.3, we have $-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) \leq 0$, which further implies that

$$\text{Term } (i) \leq O\left(\sum_{k=1}^{K'} \sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \right)$$

$$\leq O\left(\sum_{k=1}^{K'} \sum_{h=1}^{H} \sqrt{\frac{2\ln \frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln \frac{3SAHT}{\delta'}}{n_h^k(s,a)}} \right)$$

$$\leq \widetilde{O}(\sqrt{S^2 A H^4 K'}),$$

and

$$\text{Term } (iv) \leq 0.$$

Also, by Lemma B.3, we know that (24) holds in this period. Applying Lemma B.6 to Term (iii), we have

$$\text{Term } (iii) \leq \widetilde{O}(\sqrt{A H^4 K'})$$

Therefore, under event F^c, we have

$$\sum_{k=1}^{K'} \left(V_1^{\pi^k}(s_1) - V_1^*(s_1) \right) \leq \widetilde{O}(\sqrt{S^2 A H^4 K'})$$

Next, we prove (23) and (24) for the remaining episodes. We prove this claim by induction. In fact, we will prove the following claim: for each episode, (23) and (24) hold. We have shown that this claim holds for the first episodes.

Assume that for $k = K' - 1$, we have $\sum_{k=1}^{K'-1}(V_1^{\pi^k}(s_1) - V_1^*(s_1)) \leq \widetilde{O}(\sqrt{S^2 A H^4(K'-1)})$, we want to prove that for $k = K'$, we have $\sum_{k=1}^{K'}(V_1^{\pi^k}(s_1) - V_1^*(s_1)) \leq \widetilde{O}(\sqrt{S^2 A H^4 K'})$.

If there exists (h, s, a) such that $n_h^{K'}(s) \geq C_0 \sqrt{K'}$ (i.e. we construct $V_h^{\text{ref}}(s)$ in this episode), then since (23) and (24)

holds for previous episodes, using Lemma 5.3 we have

$$\left|V_h^{\mathrm{ref}}(s) - V_h^*(s)\right| = \frac{1}{n_h^{K'}(s)}\left|\sum_{i=1}^{K'}\left[V_h^i(s_h^i) - V_h^*(s)\right]\mathbf{1}[s_h^i = s]\right|$$

$$\leq \frac{1}{n_h^{K'}(s)}\sum_{i=1}^{K'}\left|V_h^i(s_h^i) - V_h^*(s)\right|\mathbf{1}[s_h^i = s]$$

$$\leq \frac{1}{C_0\sqrt{K'}}\sum_{i=1}^{K'}\left|V_h^i(s) - V_h^*(s)\right|$$

$$\leq \frac{1}{C_0\sqrt{K'}}\sqrt{S^2AH^4K'} = \frac{\sqrt{S^2AH^4}}{C_0}$$

Therefore, by Lemma B.4 we know that Q_h^k is an optimistic estimation of Q_h^*. Together with Lemmas A.1 and B.6, we have

$$\sum_{k=1}^{K'}\left(V_1^{\pi^k}(s_1) - V_1^*(s_1)\right)$$

$$\leq O\Big(\sum_{k=1}^{K'}\sum_{h=1}^{H}b_h^k(s_h^k, a_h^k)\Big) + O(\sqrt{AH^4K'})$$

$$\leq O\Big(\sum_{k=1}^{K'}\sum_{h=1}^{H}\sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln\frac{3SAHT}{\delta'}}{n_h^k(s,a)}}\Big) + O(\sqrt{AH^4K'})$$

$$\leq \widetilde{O}(\sqrt{S^2AH^4K'}),$$

which concludes our proof.

$\square$

B.3. Proof of Lemma 5.3

Proof. We have

$$\sum_{k=1}^{K'}|V_h^k(s_h^k) - V_h^*(s_h^k)|$$

$$\leq \sum_{k=1}^{K'}|V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)| + \sum_{k=1}^{K'}|V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k)|$$

$$= \sum_{k=1}^{K'}(V_h^{\pi^k}(s_h^k) - V_h^k(s_h^k)) + \sum_{k=1}^{K'}(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k))$$

where in the last step we use optimism and definition of V^*. By Lemma 5.1, the first term $\sum_{k=1}^{K'}(V_h^{\pi^k}(s_h^k) - V_h^k(s_h^k))$ can be bounded as

$$\sum_{k=1}^{K'}\left(V_h^{\pi^k}(s_h) - V_h^k(s_h)\right)$$

$$\leq O\Big(\sum_{k=1}^{K'}\sum_{h'=h}^{H}b_{h'}^k(s_{h'}^k, a_{h'}^k)\Big) + O(\sqrt{H^3K'}) \tag{25}$$

$$\leq O\Big(\sum_{k=1}^{K'}\sum_{h'=h}^{H}\sqrt{\frac{H^2S}{n_{h'}^k(s_{h'}^k, a_{h'}^k)}}\Big) + O(\sqrt{H^3K'})$$

$$\leq \widetilde{O}(\sqrt{S^2AH^4K'}).$$

For the second term $\sum_{k=1}^{K'}(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k))$, we use a forward induction trick:

First we have $\sum_{k=1}^{K'}(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k)) \leq O(\sqrt{S^2AH^4K'})$ holds for $h = 1$. By induction, if the claim holds up to h, then we have

$$\sum_{k=1}^{K'} \left(V_h^{\pi^k}(s_h^k) - V_h^*(s_h^k) \right)$$

$$= \sum_{k=1}^{K'} \left(\langle Q_h^{\pi^k}(s_h^k, \cdot), \pi_h^k(\cdot|s_h^k) \rangle - \langle Q_h^*(s_h^k, \cdot), \pi_h^*(\cdot|s_h^k) \rangle \right)$$

$$\geq \sum_{k=1}^{K'} \left(\langle Q_h^{\pi^k}(s_h^k, \cdot), \pi_h^k(\cdot|s_h^k) \rangle - \langle Q_h^*(s_h^k, \cdot), \pi_h^k(\cdot|s_h^k) \rangle \right)$$

$$= \sum_{k=1}^{K'} \langle P_h(s_h^k, \cdot)(V_{h+1}^{\pi^k} - V_{h+1}^*), \pi_h^k(\cdot|s_h^k) \rangle$$

$$= \sum_{k=1}^{K'} \left(V_{h+1}^{\pi^k}(s_{h+1}^k) - V_{h+1}^*(s_{h+1}^k) \right) + \underbrace{P_h(\cdot|s_h^k, \pi_h^k(s_h^k))(V_{h+1}^{\pi^k} - V_{h+1}^*)(\cdot) - (V_{h+1}^{\pi^k} - V_{h+1}^*)(s_{h+1}^k)}_{(a)},$$

where Term (a) is a martingale. Using Azuma-Hoeffding inequality, we have Term $(a) \leq O(\sqrt{H^2K'})$. Therefore by induction, for all $h \in [H]$ we have

$$\sum_{k=1}^{K'} V_{h+1}^{\pi^k}(s_{h+1}^k) - V_{h+1}^*(s_{h+1}^k) \leq \sum_{k=1}^{K'} V_1^{\pi^k}(s_1^k) - V_1^*(s_1^k) + O(h\sqrt{H^2K'}) \leq O(\sqrt{S^2AH^4K'}),$$

which concludes the proof of Lemma 5.3. $\qquad\square$

B.4. Useful Lemmas

B.4.1. OPTIMISM

Lemma B.3 (Optimism at the beginning). *Conditioned on F^c, when $b_h^k(s,a) = \sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}} + H\sqrt{\frac{4S\ln\frac{3SAHT}{\delta'}}{n_h^k(s,a)}}$, we have that*

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) \leq 0. \tag{26}$$

Proof. Recall that Q_h^k takes form

$$Q_h^k(s,a) = \max\{\bar{c}_h^k(s,a) + \overline{P}_h^k(\cdot \mid s,a)V_{h+1}^k(\cdot) - b_h^k(s,a), 0\}. \tag{27}$$

We have

$$c_h(s,a) + P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) - Q_h^k(s,a)$$
$$\leq c_h(s,a) - \bar{c}_h^k(s,a) + P_h(\cdot \mid s,a)V_{h+1}^k(\cdot) - \overline{P}_h^k(\cdot \mid s,a)V_{h+1}^k(\cdot) + b_h^k(s,a). \tag{28}$$

Under the event F^c, we have

$$c_h(s,a) - \bar{c}_h^k(s,a) \leq \sqrt{\frac{2\ln\frac{2SAHT}{\delta'}}{n_h^k(s,a)}}, \tag{29}$$

and

$$
\left| P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) - \overline{P}_h^k(\cdot \mid s, a) V_{h+1}^k(\cdot) \right|
$$

$$
\leq \left\| \overline{P}_h^k(\cdot \mid s_h, a_h) - P_h(\cdot \mid s_h, a_h) \right\|_1 \left\| V_{h+1}^k(\cdot) \right\|_\infty
$$

$$
\leq H \cdot \left\| \overline{P}_h^k(\cdot \mid s_h, a_h) - P_h(\cdot \mid s_h, a_h) \right\|_1
$$

$$
\leq H \sqrt{\frac{4S \ln \frac{3SAHT}{\delta'}}{n_h^k(s,a)}}. \tag{30}
$$

Here the first inequality follows from Cauchy-Schwartz inequality, the second inequality uses the fact that $\|V_h^k(\cdot)\|_\infty \leq H$ for any $(k,h) \in [K] \times [H]$, and the last inequality holds conditioned on the event F^c. Plugging (29) and (30) into (28), together with the definition of b_h^k, we obtain

$$
c_h(s,a) + P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) - Q_h^k(s,a) \leq 2b_h^k. \tag{31}
$$

Similarly, by the definition of Q_h^k in (27), we have

$$
Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s, a) V_{h+1}^k(\cdot)
$$

$$
\leq \max \left\{ \left(\overline{c}_h^k(s,a) - c_h(s,a) \right) + \left(\overline{P}_h^k(\cdot \mid s, a) V_{h+1}^k(\cdot) - P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) \right) - b_h^k(s,a), 0 \right\}
$$

$$
\leq \max\{ b_h^k - b_h^k, 0 \} = 0, \tag{32}
$$

where the last inequality follows from (29) and (30). Combining (31) and (32), we finish the proof. $\square$

Lemma B.4. *Conditioned on the event F^c, when the reference function V^{ref} satisfies $|V^{\mathrm{ref}}(s) - V^*(s)| \leq \frac{\sqrt{S^2 AH^4}}{C_0}$ where $C_0 = \sqrt{S^3 AH^3}$, we have*

$$
-2b_h^k(s,a) \leq Q_h^k(s,a) - c_h(s,a) - P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) \leq 0.
$$

Proof. When F does not happen, we have

$$
\left| c_h(s,a) - \overline{c}_h^k(s,a) \right| \leq \sqrt{\frac{2 \ln \frac{2SAHT}{\delta'}}{n_h^k(s,a)}}
$$

and

$$
\left| P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) - \overline{P}_h^k(\cdot \mid s, a) V_{h+1}^k(\cdot) \right| \leq H \sqrt{\frac{4S \ln \frac{3SAHT}{\delta'}}{n_h^k(s,a)}}.
$$

Meanwhile, we have

$$
\left| P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) - \overline{P}_h^k(\cdot \mid s, a) V_{h+1}^k(\cdot) \right|
$$

$$
\leq \left| \left(\overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) V_{h+1}^*(\cdot) \right| + \left| \left(\overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) (V_{h+1}^k - V_{h+1}^{\mathrm{ref}})(\cdot) \right|
$$

$$
+ \left| \left(\overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) (V_{h+1}^{\mathrm{ref}} - V_{h+1}^*)(\cdot) \right| \tag{33}
$$

For the first term, we have

$$
\left| \left(\overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) V_{h+1}^* \right|
$$

$$
\leq \sqrt{\frac{2 \mathbb{V}_{Y \sim P_h(\cdot \mid s, a)} V_{h+1}^*(Y) \ln \left(\frac{2SAHK}{\delta'} \right)}{n_h^k(s,a)}} + \frac{7 \ln \left(\frac{2SAHK}{\delta'} \right)}{3 n_h^k(s,a)}
$$

$$
\leq \sqrt{\frac{4 \mathbb{V}_{Y \sim P_h(\cdot \mid s, a)} V_{h+1}^{\mathrm{ref}}(Y) \ln \left(\frac{2SAHK}{\delta'} \right)}{n_h^k(s,a)}} + \sqrt{\frac{4H \ln \left(\frac{2SAHK}{\delta'} \right)}{S \cdot n_h^k(s,a)}} + \frac{7 \ln \left(\frac{2SAHK}{\delta'} \right)}{3 n_h^k(s,a)}, \tag{34}
$$

where the last inequality follows from the fact that $\sqrt{2\mathbb{V}(X)} \le 2\sqrt{\mathbb{V}(Y) + \mathbb{V}(Y - X)} \le 2\sqrt{\mathbb{V}(Y)} + 2\sqrt{\mathbb{V}(Y - X)}$ and $|V_{h+1}^* - V_{h+1}^{\text{ref}}| \le \sqrt{H/S}$. Under the event F^c, we have

$$\left| \overline{P}_h^k(y \mid s, a) - P_h(y \mid s, a) \right| \le \sqrt{\frac{2\overline{P}_h^k(y \mid s, a)(1 - \overline{P}_h^k(y \mid s, a)) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a) - 1}} + \frac{7 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)},$$

By AM-GM inequality, we have

$$\frac{1}{2}\overline{P}_h^k(y \mid s, a) + \frac{\ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)} \ge \sqrt{\frac{2\overline{P}_h^k(y \mid s, a) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)}}$$

$$\ge \sqrt{\frac{2\overline{P}_h^k(y \mid s, a)\left(1 - \overline{P}_h^k(y \mid s, a)\right) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a) - 1}}$$

which further implies that

$$P_h(y \mid s, a) - \frac{3}{2} \cdot \overline{P}_h^k(y \mid s, a) \le \frac{10 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}.$$

Then, we have

$$\mathbb{V}_{Y \sim P_h(\cdot|s,a)} V_{h+1}^{\text{ref}}(Y) = \sum_{y \in \mathcal{S}} P_h(y \mid s, a) \left(V_{h+1}^{\text{ref}}(y) - P_h^k(\cdot \mid s, a) V_{h+1}^{\text{ref}}(\cdot) \right)^2$$

$$\le \sum_{y \in \mathcal{S}} P_h(y \mid s, a) \left(V_{h+1}^{\text{ref}}(y) - \overline{P}_h^k(\cdot \mid s, a) V_{h+1}^{\text{ref}}(\cdot) \right)^2$$

$$\le \sum_{y \in \mathcal{S}} \left(\frac{3}{2}\overline{P}_h^k(y \mid s, a) + \frac{10 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)} \right) \cdot \left(V_{h+1}^{\text{ref}}(y) - \overline{P}_h^k(\cdot \mid s, a) V_{h+1}^{\text{ref}}(\cdot) \right)^2$$

$$\le \frac{3}{2}\mathbb{V}_{Y \sim \overline{P}_h^k(\cdot|s,a)} V_{h+1}^{\text{ref}}(Y) + \frac{10 SH^2 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}. \tag{35}$$

Plugging (35) into (34), we obtain

$$\left| \left(\overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) V_{h+1}^* \right|$$

$$\le \sqrt{\frac{6 \mathbb{V}_{Y \sim \overline{P}_h^k(\cdot|s,a)} V_{h+1}^{\text{ref}}(Y) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)}} + \sqrt{\frac{4H \ln\left(\frac{2SAHK}{\delta'}\right)}{S \cdot n_h^k(s, a)}} + \frac{8\sqrt{SH^2} \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}. \tag{36}$$

Meanwhile, it holds that

$$\left| \left(\overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) \left(V_{h+1}^k - V_{h+1}^{\text{ref}} \right) \right|$$

$$\le \sum_{y \in \mathcal{S}} \left| \overline{P}_h^k(y \mid s, a) - P_h(y \mid s, a) \right| \cdot |V_{h+1}^k(y) - V_{h+1}^{\text{ref}}(y)|$$

$$\le \sum_{y \in \mathcal{S}} \left(\sqrt{\frac{2\overline{P}_h(y \mid s, a)(1 - \overline{P}_h(y \mid s, a)) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a) - 1}} + \frac{7 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)} \right) |V_{h+1}^k(y) - V_{h+1}^{\text{ref}}(y)|, \tag{37}$$

where the last inequality follows from the definition of event F^c. Plugging (36) and (37) into (33), we have

$$
\left| P_h(\cdot \mid s, a) V_{h+1}^k(\cdot) - \overline{P}_h^k(\cdot \mid s, a) V_{h+1}^k(\cdot) \right|
$$

$$
\leq \sqrt{\frac{6 \mathbb{V}_{Y \sim \overline{P}_h^k(\cdot|s,a)} V_{h+1}^{\text{ref}}(Y) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)}} + \sqrt{\frac{4H \ln\left(\frac{2SAHK}{\delta'}\right)}{S \cdot n_h^k(s, a)}} + \frac{8\sqrt{SH^2} \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}
$$

$$
+ \sum_{y \in \mathcal{S}} \left(\sqrt{\frac{2\overline{P}_h(y \mid s, a)(1 - \overline{P}_h(y \mid s, a)) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a) - 1}} + \frac{7 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)} \right) |V_{h+1}^k(y) - V_{h+1}^{\text{ref}}(y)|
$$

$$
+ \underbrace{\sum_{y \in \mathcal{S}} \left| \overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right| |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)|}_{(b)}
$$

For Term (b), we divide $\mathcal{S}$ into two sets: $\mathcal{S}_0 = \{y \in \mathcal{S} : n_h^k(y) \geq C_0 \sqrt{k}\}$ and $\mathcal{S}_0^c$. Since $|V_{h+1}^{\text{ref}}(s) - V_{h+1}^*(s)| \leq \frac{\sqrt{S^2 AH^4}}{C_0}$, we have

$$
\sum_{y \in \mathcal{S}_0} \left| \overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s_h, a_h) \right| |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)|
$$

$$
\leq \sqrt{\frac{4S \ln \frac{3SAHT}{\delta'}}{n_h^k(s, a)}} \frac{\sqrt{S^2 AH^4}}{C_0}
$$

For $y \in \mathcal{S}_0^c$, we have

$$
\sum_{y \in \mathcal{S}_0^c} \left| \overline{P}_h^k(\cdot \mid s, a) - P_h(\cdot \mid s_h, a_h) \right| |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)|
$$

$$
\leq \sum_{y \in \mathcal{S}_0^c} \left(\sqrt{\frac{2\overline{P}_h^k(s' \mid s, a)(1 - \overline{P}_h^k(s' \mid s, a)) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a) - 1}} + \frac{7 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)} \right) |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)|
$$

$$
\leq \sum_{y \in \mathcal{S}_0^c} \left(\sqrt{\frac{4\overline{P}_h^k(s' \mid s, a) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)}} \right) |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)| + \frac{7 \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)} |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)|
$$

$$
= \sum_{y \in \mathcal{S}_0^c} \sqrt{\frac{4 n_h^k(s, a, y) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)^2}} |V_{h+1}^{\text{ref}}(y) - V_{h+1}^*(y)| + \frac{7SH \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}
$$

$$
\leq \sum_{y \in \mathcal{S}_0^c} \sqrt{\frac{4 n_h^k(y) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)^2}} H + \frac{7SH \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}
$$

$$
\leq \sum_{y \in \mathcal{S}_0^c} \sqrt{\frac{4 C_0 \sqrt{K} \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)^2}} H + \frac{7SH \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}
$$

$$
\leq \frac{2S\sqrt{C_0} K^{1/4} H \sqrt{\ln\left(\frac{2SAHK}{\delta'}\right)}}{n_h^k(s, a)} + \frac{7SH \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}
$$

Therefore, we have

$$
\begin{aligned}
&\left| \left(\overline{P}_h^k \left(\cdot \mid s_h, a_h \right) - P_h \left(\cdot \mid s_h, a_h \right) \right) V_{h+1}^k \right| \\
&\leq \sqrt{ \frac{ 6 \mathbb{V}_{Y \sim \overline{P}_h^k (\cdot | s,a)} V_{h+1}^{\mathrm{ref}}(Y) \ln \left(\frac{2SAHK}{\delta'} \right) }{ n_h^k(s,a) } } + \sqrt{ \frac{ 4H \ln \left(\frac{2SAHK}{\delta'} \right) }{ S \cdot n_h^k(s,a) } } + \frac{ 8 \sqrt{SH^2} \ln \left(\frac{2SAHK}{\delta'} \right) }{ 3 n_h^k(s,a) } \\
&\quad + \sum_{y \in \mathcal{S}} \left(\sqrt{ \frac{ 2 \overline{P}_h^k(s' \mid s,a)(1 - \overline{P}_h^k(s' \mid s,a)) \ln \left(\frac{2SAHK}{\delta'} \right) }{ n_h^k(s,a) - 1 } } + \frac{ 7 \ln \left(\frac{2SAHK}{\delta'} \right) }{ 3 n_h^k(s,a) } \right) |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)| \\
&\quad + \sqrt{ \frac{ 4S \ln \frac{3SAHT}{\delta'} }{ n_h^k(s,a) } } \frac{ \sqrt{S^2 A H^4} }{ C_0 } + \frac{ 2S \sqrt{C_0} K^{1/4} H \sqrt{ \ln \left(\frac{2SAHK}{\delta'} \right) } }{ n_h^k(s,a) } + \frac{ 7SH \ln \left(\frac{2SAHK}{\delta'} \right) }{ 3 n_h^k(s,a) }
\end{aligned}
$$

By definition, $Q_h^k(s,a) = \max\{ \overline{c}_h^k(s,a) + \overline{P}_h^k(\cdot \mid s,a) V_{h+1}^k(\cdot) - b_h^k(s,a), 0 \}$. Then we can prove that (31) and (32) still hold, which finishes the proof. $\qquad\square$

B.4.2. MIRROR DESCENT

The mirror descent (MD) algorithm (Beck & Teboulle, 2003) is a proximal convex optimization method that minimizes a linear approximation of the objective together with a proximity term, defined in terms of a Bregman divergence between the old and new solution estimates. In our analysis we choose the Bregman divergence to be the l_2 norm. If $\{f_k\}_{k=1}^K$ is a sequence of convex functions $f_k : \mathbb{R}^d \to \mathbb{R}$ and C is a constraints set, the k-th iterate of MD is the following:

$$
x_{k+1} \in \arg\min_{x \in C} \left\{ \eta_k \left\langle g_k \left(x_k \right), x - x_k \right\rangle + \| x - x_k \|_2^2 \right\}, \tag{38}
$$

where η_k is the stepsize. The MD algorithm ensures $\mathrm{Regret}\,(K') = \sum_{k=1}^{K'} f(x_k) - \min_x f(x) \leq O(\sqrt{K'})$ for all $K' \in [K]$.

The following lemma (Theorem 6.8 in (Orabona, 2019)) is a fundamental inequality for analysis of OMD regret, which will be used in our analysis.

Lemma B.5 (OMD Regret, Theorem 6.8 in (Orabona, 2019)). *Assume $\eta_{k+1} \leq \eta_k, k = 1, \ldots, K$. Then, using OMD with the l_2 norm, learning rate $\{\eta_k\}$ and uniform initialization $x_1 = [1/d, \ldots, 1/d]$, the following regret bounds hold*

$$
\sum_{t=1}^T \left\langle g_t, x_t - u \right\rangle \leq \max_{1 \leq t \leq T} \frac{ B_\psi \left(u; x_t \right) }{ \eta_T } + \frac{1}{2} \sum_{t=1}^T \eta_t \| g_t \|_2^2. \tag{39}
$$

In our analysis, by adapting the above lemma to our notation, we get the following lemma.

Lemma B.6 (OMD in Policy Optimization). *Assume $\eta_{k+1} \leq \eta_k, k = 1, \ldots, K$. Then, using OMD with the l_2 norm, learning rate $\{\eta_k\}$ and uniform initialization $\pi_h^1(\cdot \mid s) = [1/A, \ldots, 1/A]$, the following regret bounds hold*

$$
\sum_{k=1}^K \left\langle Q_h^k(s, \cdot), \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s) \right\rangle \leq \frac{2}{\eta_K} + \frac{1}{2} \sum_{k=1}^K \eta_k \sum_a \left(Q_h^k(s,a) \right)^2. \tag{40}
$$

Moreover, if we choose $\eta_k = 1/\sqrt{AH^2 k}$, we have

$$
\sum_{k=1}^K \sum_{h=1}^H \left\langle Q_h^k(s, \cdot), \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s) \right\rangle \leq 3\sqrt{AH^4 K}. \tag{41}
$$

Proof. Fix $h \in [H]$. Replace g_t with $Q_h^k(s, \cdot)$ and x_t with $\pi_h^k(\cdot \mid s)$ in (39). Since $\| \pi_h^* - \pi_h^k \|_2^2 \leq 2$ and $Q_h^k(s,a) \leq H$, we

have

$$\sum_{k=1}^{K} \left\langle Q_h^k(s, \cdot), \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s) \right\rangle \leq \frac{2}{\eta_K} + \frac{1}{2} \sum_{k=1}^{K} \eta_k \sum_a \left(Q_h^k(s, a) \right)^2$$

$$\leq \frac{2}{\eta_K} + \frac{1}{2} \sum_{k=1}^{K} \eta_k A H^2.$$

Let $\eta_k = 1/\sqrt{AH^2 k}$, we have $\sum_{k=1}^{K} \eta_k \leq \sqrt{\frac{4K}{AH^2}}$, which further implies that

$$\sum_{k=1}^{K} \left\langle Q_h^k(s, \cdot), \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s) \right\rangle \leq 3\sqrt{AH^2 K}$$

for any $h \in [H]$. Taking summation over $h \in [H]$ concludes our proof. $\qquad\square$

B.5. Sum of Bonus

Lemma B.7. *It holds that*

$$\sum_{k=1}^{K} \sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \leq \widetilde{O}(\sqrt{SAH^3 K} + S^{5/2} A^{5/4} H^{11/4} K^{1/4}). \tag{42}$$

Proof. For simplicity we let $n_h^k = n_h^k(s_h^k, a_h^k), \overline{P}_h^k(\cdot) = \overline{P}_h^k(\cdot \mid s_h^k, a_h^k), P_h(\cdot) = P_h^k(\cdot \mid s_h^k, a_h^k)$. By the same arguments in (34), (35) and (36), we have

$$\sqrt{\frac{6 \mathbb{V}_{Y \sim \overline{P}_h^k(\cdot | s,a)} V_{h+1}^{\mathrm{ref}}(Y) \ln\left(\frac{2SAHK}{\delta'}\right)}{n_h^k(s, a)}} + \sqrt{\frac{4H \ln\left(\frac{2SAHK}{\delta'}\right)}{S \cdot n_h^k(s, a)}} + \frac{8\sqrt{SH^2} \ln\left(\frac{2SAHK}{\delta'}\right)}{3 n_h^k(s, a)}$$

$$\leq \widetilde{O}\left(\sqrt{\frac{\mathbb{V}_{Y \sim P_h^k(\cdot | s,a)} V_{h+1}^*(Y)}{n_h^k(s, a)}} + \sqrt{\frac{H}{S \cdot n_h^k(s, a)}} + \frac{\sqrt{SH^2}}{3 n_h^k(s, a)} \right).$$

Then, by the definition of b_h^k,

$$\sum_{k=1}^{K} \sum_{h=1}^{H} b_h^k(s_h^k, a_h^k)$$

$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \widetilde{O}\left(\sqrt{\frac{\mathbb{V}_{Y \sim P_h^k(\cdot | s,a)} V_{h+1}^*(Y)}{n_h^k(s, a)}} + \sum_{y \in \mathcal{S}} \sqrt{\frac{\overline{P}_h^k(y)}{n_h^k}} |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)| + \frac{S^{3/2} A^{1/4} H^{7/4} K^{1/4}}{n_h^k} \right)$$

$$\leq \widetilde{O}(\sqrt{SAH^3 K}) + \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{y \in \mathcal{S}} \widetilde{O}\left(\sqrt{\frac{\overline{P}_h^k(y)}{n_h^k}} |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)| \right) + \widetilde{O}(S^{5/2} A^{5/4} H^{11/4} K^{1/4})$$

$$\leq \widetilde{O}(\sqrt{SAH^3 K}) + \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{y \in \mathcal{S}} \widetilde{O}\left(\sqrt{\frac{P_h^k(y) + O(\sqrt{1/n_h^k})}{n_h^k}} |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)| \right) + \widetilde{O}(S^{5/2} A^{5/4} H^{11/4} K^{1/4})$$

$$\leq \widetilde{O}(\sqrt{SAH^3 K}) + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{y \in \mathcal{S}} \widetilde{O}\left(\sqrt{\frac{P_h^k(y)}{n_h^k}} |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)| \right)}_{(v)} + \widetilde{O}(S^{5/2} A^{5/4} H^{11/4} K^{1/4}) \tag{43}$$

The second inequality follows from Lemma 19 in Zhang et al. (2020b) and standard techniques (*e.g.* (Shani et al., 2020) or (Zhang et al., 2020b)). Using Lemma B.8, we can further bound Term (v) as:

$$\text{Term } (v) \leq \widetilde{O}(\sqrt{SAH^3K}) + \widetilde{O}(S^{5/2}A^{5/4}H^{3/2}K^{1/4})$$

Therefore, combining Equation (16) and (43), we complete the proof of Theorem 3.1. $\qquad\square$

Lemma B.8.

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{y\in\mathcal{S}}\left(\sqrt{\frac{P_h^k(y)}{n_h^k}}|V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)|\right) \leq \widetilde{O}(\sqrt{SAH^3K}) + \widetilde{O}(S^{5/2}A^{5/4}H^{3/2}K^{1/4})$$

Proof. We have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{y\in\mathcal{S}}\left(\sqrt{\frac{P_h^k(y)}{n_h^k}}|V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)|\right)\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sum_{y\in\mathcal{S}}P_h^k(y)\sqrt{\frac{1}{P_h^k(y)n_h^k}}|V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)| - \sqrt{\frac{1}{P_h^k(s_{h+1}^k)n_h^k}}|V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mathrm{ref}}(s_{h+1}^k)|\right)\\
&+\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{P_h^k(s_{h+1}^k)n_h^k}}|V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mathrm{ref}}(s_{h+1}^k)|\\
\leq&\widetilde{O}(\sqrt{H^3K}) + \sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{P_h^k(s_{h+1}^k)n_h^k}}|V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mathrm{ref}}(s_{h+1}^k)|\\
=&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sqrt{\frac{1}{P_h^k(s_{h+1}^k)n_h^k}}\underbrace{|V_{h+1}^k(s_{h+1}^k) - V_{h+1}^*(s_{h+1}^k)|}_{\text{the sum }\leq\widetilde{O}(\sqrt{H^4S^2AK})} + \sqrt{\frac{1}{P_h^k(s_{h+1}^k)n_h^k}}\underbrace{|V_{h+1}^*(s_{h+1}^k) - V_{h+1}^{\mathrm{ref}}(s_{h+1}^k)|}_{\text{almost }\leq\widetilde{O}(1/\sqrt{S})}\right)\\
&+\widetilde{O}(\sqrt{H^3K})\\
\leq&\widetilde{O}(S^2AH\sqrt{\sqrt{H^2S^2AK}}) + \widetilde{O}(\sqrt{S^2AH^2K}/\sqrt{S}) + \widetilde{O}(\sqrt{H^3K})\\
=&\widetilde{O}(\sqrt{SAH^3K}) + \widetilde{O}(S^{5/2}A^{5/4}H^{3/2}K^{1/4}) \tag{44}
\end{aligned}
$$

The second line is the sum of martingale differences bounded by H, directly applying Azuma-Hoeffding inequality yields the second inequality. The third inequality makes use of Lemma 11 in (Zhang et al., 2020b).

The third inequality in (44) This is a technique used in (Azar et al. 2017). We explain the main idea here.

First we define the typical state-actions pairs as

$$[y]_{k,h} \overset{\text{def}}{=} \{y : y \in \mathcal{S}, n_h^k(s_h^k, a_h^k)P(y \mid s_h^k, a_h^k) \geq 2H^2SL\}$$

which means these state-action pairs are visited frequently enough. Define $\widetilde{\Delta}_{h+1}^k(y) = |V_{h+1}^k(y) - V_{h+1}^{\mathrm{ref}}(y)|$. We have

$$
\begin{aligned}
&\sum_{y\in\mathcal{S}}\sqrt{\frac{P_h^k(y)}{n_h^k}}\widetilde{\Delta}_{h+1}^k(y)\\
=&\sum_{y\in[y]_{k,h}}\sqrt{\frac{P_h^k(y)}{n_h^k}}\widetilde{\Delta}_{h+1}^k(y) + \sum_{y\notin[y]_{k,h}}\sqrt{\frac{P_h^k(y)}{n_h^k}}\widetilde{\Delta}_{h+1}^k(y).
\end{aligned}
$$

The second term can be bounded by

$$\sum_{y\notin[y]_{k,h}}\sqrt{\frac{P_h^k(y)n_h^k}{n_h^{k^2}}}\widetilde{\Delta}_{h+1}^k(y) \leq \frac{SH\sqrt{4LH^2S}}{n_h^k}$$

So we only have to deal with the first term, in which $P_h^k(y)n_h^k$ is large. Thus the martingale can be bounded.

The last inequality in (44) As above, we only need to consider the case where $P_h^k(s_{h+1}^k)n_h^k \geq 2H^2SL$ (n_h^k is the shorthand for $n_h^k(s_h^k, a_h^k)$). Then we know the first term is bounded by $\widetilde{O}(\sqrt{SAH^3K})$. For the second term, we need to use the Multiplicative Chernoff bound (See the Wiki for Chernoff bound): $\Pr(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2\mu}{2}}$, where $0 \leq \delta \leq 1$ and $X = \sum_{i=1}^n X_i$, X_i are independent Bernoulli r.v., $\mathbb{E}[X] = \mu$. Set $X = n_h^k(s_h^k, a_h^k, s_{h+1}^k)$, $\mu = P_h^k(s_{h+1}^k)n_h^k$. Taking union bound over all $h, k, \mathcal{S}$ with $P_h^k(s_{h+1}^k)n_h^k \geq 2H^2SL$, we have that with high probability, $P_h^k(s_{h+1}^k)n_h^k \geq \frac{1}{2}n_h^k(s_h^k, a_h^k, s_{h+1}^k)$. With this we can now apply Lemma 11 in (Zhang et al. 2020). $\qquad\square$

C. Explanation of How UCBVI Uses Optimism

In Azar et al. (2017), they need to bound the term $(\overline{P}_h^k - P_h)(\cdot \mid s, a)(V_{h+1}^k - V_{h+1}^*)(\cdot)$ using optimism, as mentioned in Section 4.

Define $\Delta_h^k \overset{\text{def}}{=} V_h^* - V_h^{\pi^k}$, $\widetilde{\Delta}_h^k \overset{\text{def}}{=} V_h^k - V_h^{\pi^k}$, and $\widetilde{\delta}_h^k \overset{\text{def}}{=} \widetilde{\Delta}_h^k(s_h^k)$. We denote by $\square$ a numerical constant which can vary from line to line. We also use L to represent the logarithmic term $L = \ln(\square HSAT/\delta)$.

Using Bernstein's inequality, this term is bounded by

$$\sum_y P^{\pi^k}(y \mid s_h^k) \sqrt{\frac{\square L}{P^{\pi^k}(y \mid s_h^k) n_h^k}} \Delta_{h+1}^k(y) + \frac{\square SHL}{n_h^k}.$$

where $n_h^k \overset{\text{def}}{=} n_k(s_h^k, \pi^k(s_h^k))$. Now considering only the y such that $P^{\pi^k}(y \mid s_h^k) n_h^k \geq \square H^2L$, and since $0 \leq \Delta_{h+1}^k \leq \widetilde{\Delta}_{h+1}^k$ **by optimism**, then $\left(\widehat{P}_k^{\pi^k} - P^{\pi_k}\right)\Delta_{k,h+1}(s_h^k)$ is bounded by

$$\overline{\epsilon}_h^k + \sqrt{\frac{\square L}{P^{\pi^k}(s_{h+1}^k \mid s_h^k) n_h^k}} \widetilde{\delta}_{k,h+1} + \frac{\square SHL}{n_h^k} \leq \overline{\epsilon}_h^k + \frac{1}{H}\widetilde{\delta}_{h+1}^k + \frac{\square SHL}{n_h^k}.$$

where $\overline{\epsilon}_h^k \overset{\text{def}}{=} \sqrt{\frac{\square L}{n_h^k}}\left(\sum_y P^{\pi^k}(y \mid s_h^k) \frac{\widetilde{\Delta}_{h+1}^k(y)}{\sqrt{P^{\pi^k}(y \mid s_h^k)}} - \frac{\widetilde{\delta}_{h+1}^k}{\sqrt{P^{\pi_k}(s_{h+1}^k \mid s_h^k)}}\right)$. The sum over the neglected y such that $P^{\pi^k}(y \mid s_h^k) n_h^k < \square H^2L$ contributes to an additional term

$$\sum_y \sqrt{\frac{\square P^{\pi^k}(y \mid s_h^k) n_h^k L}{n_h^{k2}}} \Delta_{h+1}^k(y) \leq \frac{\square SH^2L}{n_h^k}.$$

Then they prove that the sum of $\overline{\epsilon}_h^k$ is of order $\widetilde{O}(\sqrt{T})$ and the sum of $\frac{1}{n_h^k}$ is a constant order term.