Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments

Runlong Zhou [1] **Zihan Zhang** **Simon S. Du** [1]

Abstract

We study variance-dependent regret bounds for Markov decision processes (MDPs). Algorithms with variance-dependent regret guarantees can automatically exploit environments with low variance (e.g., enjoying constant regret on deterministic MDPs). The existing algorithms are either variance-independent or suboptimal. We first propose two new environment norms to characterize the fine-grained variance properties of the environment. For model-based methods, we design a variant of the `MVP` algorithm (Zhang et al., 2021a). We apply new analysis techniques to demonstrate that this algorithm enjoys variance-dependent bounds with respect to the norms we propose. In particular, this bound is *simultaneously minimax optimal for both stochastic and deterministic MDPs*, the first result of its kind. We further initiate the study on model-free algorithms with variance-dependent regret bounds by designing a reference-function-based algorithm with a novel *capped-doubling reference update schedule*. Lastly, we also provide lower bounds to complement our upper bounds.

1. Introduction

We consider episodic reinforcement learning (RL) on tabular Markov Decision Processes (MDPs). Existing algorithms can be categorized into two classes: model-based methods whose space complexity scales quadratically with the number of states (Auer et al., 2008; Agrawal & Jia, 2017; Azar et al., 2017; Dann et al., 2017; 2019; Zanette & Brunskill, 2019; Zhang et al., 2021a) and model-free methods whose space complexity scales linearly with the number of states (Jin et al., 2018; Bai et al., 2019; Zhang et al., 2020b; Li et al., 2021).

The MDPs in practice often enjoy benign structures, so problem-dependent regret bounds are of great interest (Zanette & Brunskill, 2019). RL algorithms often perform far better on these MDPs than what their worst-case guarantees would suggest. Motivated by this observation, we want to systematically study algorithms with regrets depending on quantities that characterizes the hardness of MDPs. Ideally, such algorithms should *automatically exploit the MDP instance without the prior knowledge of problem-dependent quantities*. As a motivating example, for time-homogeneous MDPs with total reward bounded by 1, the minimax regret bound for deterministic MDPs is $O(SA)$ where S and A are number of states and actions, respectively and the worst-case minimax optimal regret bound for stochastic MDPs is $\widetilde{O}\left(\sqrt{SAK}\right)$ where K is the number of episodes. Many problems can be formulated as deterministc MDPs, such as shortest path (maze, real world navigation), combinatorial optimization, Atari games (Mnih et al., 2013) and many games (mountain car, lunar lander, robotics, etc.) in OpenAI Gym (Brockman et al., 2016). Deterministic systems can also approximate stochastic systems well (see Section 2 and 6 in Bertsekas (2012)). We want an algorithm designed for generic stochastic MDPs with worst-case minimax optimal regret bound while enjoying the $O(SA)$ bound when the MDP is deterministic.

Zanette & Brunskill (2019) is a pioneer work which provides a model-based algorithm whose regret scales with variance-depedent quantities. They defined a quantity, $\mathbb{Q}^\star$, named *the maximum per-step conditional variance* to characterize the randomness of the MDP instance, and showed a regret bound of $\widetilde{O}(\sqrt{H\mathbb{Q}^\star \cdot SAK} + H^{5/2}S^2A)$, where H is the planning horizon. This bound is still not satisfactory because: ① There exist MDPs with $\mathbb{Q}^\star = \Omega(1)$, so the regret reduces to $\widetilde{O}(\sqrt{HSAK})$ which does not match the minimax optimal bound $\widetilde{O}(\sqrt{SAK})$. ② For deterministic MDPs ($\mathbb{Q}^\star = 0$), the regret reduces to $\widetilde{O}(H^{5/2}S^2A)$, which does not match the optimal $O(SA)$ bound.

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. Correspondence to: Simon S. Du <ssdu@cs.washington.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1.1. Contributions

This paper makes the following contributions which significantly advance our understanding of problem-dependent bounds in reinforcement learning.

- First, We introduce *the total multi-step conditional variance*, Var_K^{Σ} and *the maximum policy-value variance*, $\mathsf{Var}^{\star}$, to provide fine-grained characterizations of the variance in the MDP (see Section 4 for the formal definitions). Importantly, regret bounds that depend on these quantities will reduce to the minimax optimal bound in the worst case whereas the existing notion $H\mathbb{Q}^{\star}$ cannot.

- Second, for model-based methods, we identify the obstacles preventing the current state-of-the-art minimax optimal algorithm, `MVP` (Zhang et al., 2021a), from being variance-dependent. We make necessary improvements and introduce a truncation method to bound the total variance. We show the regret bound of the improved algorithm, `MVP-V`, scales with $\mathsf{Var}^{\star}$ or Var_K^{Σ}. In particular, these bounds imply that, `MVP-V` is *minimax optimal for both the classes of stochastic and deterministic MDPs*. To our knowledge, this is first result of such kind. See Table 1 for comparions between model-based methods.

- Third, we initiate the study of model-free algorithms with variance-dependent regrets. We explain why existings model-free algorithms cannot be variance-dependent. We futher propose a new model-free algorithm, `UCB-Advantage-V`, which relies on a *a capped-doubling manner of updates for reference values*. We further utilize a novel analysis technique which *bounds value gaps directly from the existing uniform convergence bound* to give the first variance-dependent bound for model-free algorithms. Importantly, this bound reduces to the minimax optimal bound for the worst-case MDPs. See Table 2 for comparisons between model-free algorithms.

- Lastly, we prove minimax regret lower bounds for the class of MDPs with bounded variances. We show that the main order terms of our regret upper bounds match these lower bounds, so our proposed algorithms are *minimax optimal for the class of variance-bounded MDPs*.

1.2. Technical Overview

For model-based algorithms, existing state-of-the-art work (Zhang et al., 2021a) fails to be variance-dependent. It is hard to bound the total variance by its expectation using martingale concentration inequalities directly, while avoiding an H-dependency. This is because the total variance within an episode can be as large as $\Omega(H)$. We introduce a novel analysis technique which *truncates* the total variance of each episode to a constant and apply martingale concentration inequalities on this sequence, and show that with high probability there is no truncation. We also apply a more refined concentration inequality to the transition model to have a dependency on the maximum support instead of the size of the state space. This step is crucial in obtaining the tight bound for deterministic MDPs.

For the model-free algorithm, existing work (Zhang et al., 2020b) upper-bounds all the four bias terms in their Equation (13) by variance-independent main order terms. We identify the problem incurred by the large bias in reference values, and replace the update with *a capped-doubling manner*. Since too frequent updates discard past data very often, this method balances between the summation of gaps of value functions and the waste of data. We *integrate directly over the error* between the estimated value and the optimal value to bound the total squared gaps between them, whereas Zhang et al. (2020b) bound them with a coarse binary gap of either H or the final gap. Combined with many other finer-grained analyses throughout the proof, we can finally remove all the variance-independent main order terms except for the total variance.

1.3. Paper Overview

The paper is organized as follows. We first list basic concepts of MDPs in Section 3, then define variance quantities in Section 4. Our main results then come in three sections: Sections 5 and 6 show the algorithms, theorems, corollaries and proof sketches of our model-based and model-free methods, respectively. Section 7 shows our lower bounds for the class of variance-bounded MDPs.

2. Related Works

Minimax optimal regret bounds. Algorithms for regret minimization can be categorized into two classes: model-based and model-free. Being model-free means the space complexity is $O(HSA)$, prohibiting the estimation of the whole transition model $P_h(s'|s,a)$. For model-based methods, there are previous work (Zhang et al., 2021a; 2022; 2020a; Wang et al., 2020) achieving a property called *horizon-free*, which allows only logarithmic dependency on H for regrets. As explained in Jiang & Agarwal (2018), in many scenarios with a long planning horizon, the interesting regime is $K \ll H$. This underscores the importance of being horizon-free, because for H-dependent bounds, only when $K \gg H$ they become sub-linear in K. Being horizon-free is challenging, because it requires utilizing transition data for the same state-action pair from different steps and handling a spike in rewards. There are many works other than those we cite in Section 1 giving nearly minimax optimal bounds: Bartlett & Tewari (2012); Osband et al. (2013); Osband & Van Roy (2017); Fruit et al. (2018a); Talebi & Maillard (2018); Simchowitz & Jamieson (2019); Russo (2019); Zhang & Ji (2019); Neu & Pike-Burke (2020); Xiong et al. (2021); Pacchiano et al.

Algorithm	Regret	Variance-Dependent	Stochastic-Optimal	Deterministic-Optimal	Horizon-Free
Euler Zanette & Brunskill (2019)	$\widetilde{O}(\sqrt{H\mathbb{Q}^\star \cdot SAK} + H^{5/2}S^2A)$	**Yes**	No	No	No
	$\widetilde{O}(\sqrt{SAK} + H^{5/2}S^2A)$	No	**Yes**	No	No
MVP Zhang et al. (2021a)	$\widetilde{O}(\sqrt{SAK} + S^2A)$	No	**Yes**	No	**Yes**
MVP-V This work	$\widetilde{O}(\sqrt{\min\{\mathsf{Var}_K^\Sigma, \mathsf{Var}^\star K\}SA} + \Gamma SA)$	**Yes**	**Yes**	**Yes**	**Yes**

Table 1. Comparisons between model-based algorithms for time-homogeneous MDPs with total reward bounded by 1. $\widetilde{O}$ hides logarithm factors. S, A, Γ, H and K are number of states, actions, maximum support of the transition model, planning horizon and interaction episodes. $\mathbb{Q}^\star$, Var_K^Σ and $\mathsf{Var}^\star$ are variance notations in Section 4. $\mathbb{Q}^\star$ and Var_K^Σ are upper bounded by 1 in the worst case and become 0 when the MDP is deterministic. An "Yes" in each column means: Variance-Dependent: the regret has a main order term scaling with any variance notation. Stochastic-Optimal: the regret has a main order term of $\widetilde{O}(\sqrt{SAK})$ which matches the minimax lower bound. Deterministic-Optimal: the regret is $\widetilde{O}(SA)$ on deterministic MDPs (with variance equal to 0). Horizon-Free: the regret has only logarithmic dependency on H.

Algorithm	Regret	Variance-Dependent	Stochastic-Optimal
Q-learning (UCB-B) Jin et al. (2018)	$\widetilde{O}(\sqrt{H^4SAK} + H^{9/2}S^{3/2}A^{3/2})$	No	No
UCB-Advantage Zhang et al. (2020b)	$\widetilde{O}(\sqrt{H^3SAK} + \sqrt[4]{H^{33}S^8A^6K})$	No	**Yes**
Q-EarlySettled-Advantage Li et al. (2021)	$\widetilde{O}(\sqrt{H^3SAK} + H^6SA)$	No	**Yes**
UCB-Advantage-V This work	$\widetilde{O}(\sqrt{\min\{\mathsf{Var}_K^\Sigma, \mathsf{Var}^\star K\}HSA} + \sqrt[4]{H^{15}S^5A^3K})$	**Yes**	**Yes**

Table 2. Comparison between model-free algorithms for time-inhomogeneous MDPs with every reward bounded by 1. An "Yes" in each column means: Variance-Dependent: the bound scales with the variance term that characterizes the randomness of the environment; Stochastic-Optimal: in the worst-case, the regret's dominating term becomes $\widetilde{O}(\sqrt{H^3SAK})$ which matches the minimax lower bound.

(2020). We compare our results with the state-of-the-art in Table 1 (model-based) and Table 2 (model-free).

Variance-dependent results. Talebi & Maillard (2018) provides a problem-dependent regret bound that scales with the variance of the next step value functions under strong assumptions on ergodicity of the MDP. Namely, they define $\boldsymbol{V}^\star_{s,a}$ for each (s,a) pair and derives a regret of $\tilde{O}(\sqrt{S\sum_{s,a}\boldsymbol{V}^\star_{s,a}T})$ under the infinite horizon setting.

Simchowitz & Jamieson (2019) combines gap-dependent regret with variances. The standard notation $\mathtt{gap}(s,a)$ is the gap between the optimal value function and the optimal Q-function, and $\mathtt{gap}_{\min}$ is the minimum non-zero gap. Let $\mathtt{Var}^\star_{h,s,a}$ be the variance of optimal value function at (h,s,a) triple, their regret approximately scales as

$$\tilde{O}\left(\sum_{s,a}\frac{H\max_h \mathtt{Var}^\star_{h,s,a}}{\max\{\mathtt{gap}(s,a),\ \mathtt{gap}_{\min}\}}\log(K)\right).$$

Variance-aware bounds exist in bandits (Kim et al., 2021; Zhang et al., 2021b; Zhou et al., 2021; Zhao et al., 2023; 2022). We notice two concurrent works: Zhao et al. (2023) studies variance-dependent regret upper bounds for linear bandits and linear mixture MDPs, and Li & Sun (2023) studies linear bandits and linear MDPs. They both define the same variance as Var_K^Σ (one of the two quantities also proposed by us under the tabular setting). More recent work generalized variance-aware bound from MDPs to latent MDPs (Zhou et al., 2022).

Other problem-dependent results. Most problem-dependent results prior to Zanette & Brunskill (2019) focus on the infinite-horizon setting. Some depend on the range of value functions (Bartlett & Tewari, 2012; Fruit et al., 2018b). Maillard et al. (2014) introduces a hardness measure called *distribution norm*. There are gap-dependent results for multi-armed bandits and RL (Even-Dar et al., 2006; Auer et al., 2008; Xu et al., 2021; Yang et al., 2021).

Jin et al. (2020) shows that with a slight modification, the algorithm in Zanette & Brunskill (2019) can have a *first-order regret*, with the main order term depending on the optimal value function. Wagenmaker et al. (2022) provides a first-order regret for linear MDPs. When the total reward is bounded by 1 almost surely, for any policy its variance is

not larger than this value. This means a first-order dependency is weaker than a variance-dependency.

3. Preliminaries

Notations. For any event $\mathcal{E}$, we use $\mathbb{1}[\mathcal{E}]$ to denote the indicator function. For any random variable X, we use $\mathbb{V}(X)$ to denote its variance. For any set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the probability simplex over $\mathcal{X}$. For any positive integer n, we denote $[n] := \{1, 2, \ldots, n\}$. For any probability distribution P, we use $\mathsf{supp}(P) = \sum_x \mathbb{1}[P(x) > 0]$ to denote the size of its support. Suppose x and y are n-dimensional vectors, we denote $xy := \sum_{i=1}^n x_i y_i$ and $x^q := (x_1^q, x_2^q, \ldots, x_n^q)$ for any number q. If $x \in \Delta([n])$, we use $\mathbb{V}(x, y) = \sum_i x_i (y_i - xy)^2 = xy^2 - (xy)^2$ to denote the variance of y under x. We use $\mathbf{1}_k$ to denote a vector with all 0 but an only 1 on the k-th position.

Finite-horizon MDPs. A finite-horizon MDP can be described by a tuple $M = (\mathcal{S}, \mathcal{A}, P, R, H)$. $\mathcal{S}$ is the finite state space with size S and $\mathcal{A}$ is the finite action space with size A. For any $h \in [H]$, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function and $R_h : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ is the reward distribution with mean $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. H is the planning horizon, i.e., episode length. We denote $\Gamma := \max_{h,s,a} \mathsf{supp}(P_h(\cdot|s, a))$ as the maximum support of the transition function, and $P_{s,a,h} := P_h(\cdot|s, a)$.

Conditions for MDPs. We have two conditions more general than the ordinary setting. As explained below them, getting tight regret bounds are harder when they are met.

Condition 1. *For any policy π, the total reward in a single episode is upper-bounded by 1 almost surely.*

For those MDPs not satisfying Condition 1, we can normalize all the rewards by $1/H$. Such a conversion usually multiplies a factor of $1/H$ to the regret, but cannot take into account a spike in rewards, e.g., some $r_h(s, a) = 1$.

Condition 2. *The MDP is time-homogeneous. Namely, there exist $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, $R : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $P_h(\cdot|s, a) = P(\cdot|s, a)$, $R_h(s, a) = R(s, a)$ and $r_h(s, a) = r(s, a)$ for any $h \in [H]$.*

For simplicity, we denote $P_{s,a} := P(\cdot|s, a)$ and $P_{s,a,s'} := P(s'|s, a)$. Any time-inhomogeneous MDP can be instantiated by a time-homogeneous one to satisfy Condition 2. Let a mega state space $\mathcal{S} = \cup_{h=1}^H \mathcal{S}_h$, where each $\mathcal{S}_h$ corresponds to the state space of the time-inhomogeneous MDP. For any (h, s, a), we construct $P(s'_{h+1}|s_h, a) = P_h(s'|s, a)$ and $R(s_h, a) = R_h(s, a)$, where s_h is the corresponding state of s in $\mathcal{S}_h$. S is multiplied by H while Γ remains the same, and the regret changes accordingly. This condition underscores the algorithm's ability to use information of the same state-action pair from different steps.

We will introduce quantities in Section 4 to quantify determinism, but a fully-deterministic MDP is very worth studying because the regret lower bound is the well-known $\Omega(SA)$ (under Conditions 1 and 2). Thus, we care about whether the algorithms can have a constant regret (up to logarithm factors) under the assumption of determinism.

Assumption 3. *The MDP is deterministic. Namely, for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $R_h(s, a)$ and $P_h(\cdot|s, a)$ map to a single real number and a single state respectively.*

Policies. A history-independent deterministic policy π chooses an action based on the current state and time step. Formally, $\pi = \{\pi_h\}_{h\in[H]}$ where $\pi_h : \mathcal{S} \to \mathcal{A}$ maps a state to an action. We use Π to denote the set of all such policies.

Episodic RL on MDPs. Upon choosing action a at state s when it is the h-th step in an episode, the agent observes a reward $r \sim R_h(s, a)$ and the next state $s' \sim P_h(\cdot|s, a)$. When $h = H$, the episode ends after this observation. Thus, a policy π induces a (random) trajectory $(\{s_h, a_h, r_h\}_{h\in[H]}, s_{H+1})$ where s_1 is exogenously generated, $a_h = \pi_h(s_h), r_h \sim R_h(s_h, a_h)$ and $s_{h+1} \sim P_h(\cdot|s_h, a_h)$ for $h \in [H]$. The episodic RL on MDPs proceeds in a total of K episodes. When one episode ends, a new initial state s_1 is generated. The agent should (adaptively) choose a policy π^k for the k-th episode, put it into action and cannot change it within an episode.

Value functions and Q-functions. Given a policy π, we define its value function and Q-function as

$$V_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{t=h}^H r_t \;\middle|\; s_h = s \right],$$

$$Q_h^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{t=h}^H r_t \;\middle|\; (s_h, a_h) = (s, a) \right].$$

It is easy to verify that $Q_h^\pi(s, a) = r_h(s, a) + P_{s,a,h} V_{h+1}^\pi$.

Performance measure. The goal of episodic RL on MDPs is to find the policy which maximizes the total reward collected in an episode, regardless of the initial state. Given M, such a goal can be achieved using dynamic programming. Given this, we denote $V^\star := V^{\pi^\star}$ and $Q^\star := Q^{\pi^\star}$. We use cumulative regret as a performance measure:

$$\mathsf{Regret}(K) := \sum_{k=1}^K (V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)).$$

4. Variance Quantities for MDPs

We use the notion of variance to quantify the degree of determinism of MDPs. The first is called *the maximum per-step conditional variance* (Zanette & Brunskill, 2019), which is only relevant to the optimal value function.

Definition 4. *The maximum per-step conditional variance for a particular MDP is defined as:*

$$\mathbb{Q}^\star := \max_{h,s,a}\{\mathbb{V}(R_h(s,a)) + \mathbb{V}(P_{s,a,h}, V^\star_{h+1})\}.$$

Zanette & Brunskill (2019) directly use $H\mathbb{Q}^\star$ to upper-bound the total per-step conditional variances in an episode, a quantity which can be upper-bounded by a constant (see Lemmas 29 and 42). So when $\mathbb{Q}^\star \geqslant \Omega(H)$ (or $\Omega(1/H)$ under Condition 1), $H\mathbb{Q}^\star$ is not tight. In light of this, we define *the total multi-step conditional variance* as a better notation in place of $H\mathbb{Q}^\star$. This quantity is also proposed in concurrent works (Zhao et al., 2023; Li & Sun, 2023).

Definition 5. *The total multi-step conditional variance for a trajectory $\tau = \{s_h, a_h\}_{h\in[H]}$ is defined as:*

$$\mathsf{Var}^\Sigma_\tau := \sum_{h=1}^{H}(\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_{s_h,a_h,h}, V^\star_{h+1})).$$

During the learning process, let the trajectory of the k-th episode be τ^k, then we denote $\mathsf{Var}^\Sigma_{(k)} := \mathsf{Var}^\Sigma_{\tau^k}$, and $\mathsf{Var}^\Sigma_K := \sum_{k=1}^K \mathsf{Var}^\Sigma_{(k)}$.

We introduce another type of variance, called *the maximum policy-value variance*, which is novel in the literature.

Definition 6. *For any policy $\pi \in \Pi$, its maximum value variance is defined as $\mathsf{Var}^\pi := \max_{s\in\mathcal{S}} \mathsf{Var}^\pi_1(s)$, where*

$$\mathsf{Var}^\pi_1(s) := \mathbb{E}_\pi\left[\sum_{h=1}^{H}\left(\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_{s_h,a_h,h}, V^\pi_{h+1})\right) \,\middle|\, s_1 = s\right].$$

The maximum policy-value variance for a particular MDP is defined as:

$$\mathsf{Var}^\star := \max_{\pi\in\Pi} \mathsf{Var}^\pi.$$

$\mathsf{Var}^\pi_1(s)$ is the variance of total reward of π starting from s, and the justification can be found in Appendix B.1.

Under Condition 1, by Lemma 20 we know that $\mathsf{Var}^\pi_1(s) \leqslant V^\pi_1(s) \leqslant V^\star_1(s)$. So $\mathsf{Var}^\star \leqslant V^\star_1(s)$. This means a variance-dependent regret is better than a first-order regret.

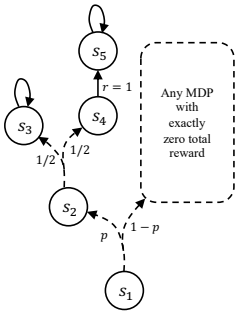

Figure 1. Example of $\mathsf{Var}^\star$ being arbitrarily smaller than $\mathsf{Var}^\Sigma_{(k)}$. Dashed arrows represent probabilistic transitions and solid arrows represent deterministic ones. The only reward comes at state s_4 and on choosing any action.

4.1. Comparing $\mathsf{Var}^\Sigma_{(k)}$ and $\mathsf{Var}^\star$

We use this subsection to demonstrate that a small $\mathsf{Var}^\Sigma_{(k)}$ does not imply a small $\mathsf{Var}^\star$, and vice versa.

Deterministic MDPs have $\mathsf{Var}^\Sigma_{(k)} = \mathsf{Var}^\star = 0$. Trivially, $\mathsf{Var}^\star = 0 \implies \mathsf{Var}^\Sigma_{(k)} = 0$, while the reverse is not ture.

When $\mathsf{Var}^\Sigma_{(k)} = 0 < \mathsf{Var}^\star$. Consider the following *time-homogeneous* MDP with horizon H: For each state s there is a good action a_1 with a *deterministic* reward $r(s,a_1) = 1/H$, and all other actions a' have a *deterministic* reward $r(s,a') = 0$. For any state-action pair (s,a), the transition is identically $P_{s,a,s'} = 1/S$.

The optimal policy always chooses a_1 at any state s, so for any s and h, $V^\star_h(s) = (H-h+1)/H$. For any (h,s,a),

$$\mathbb{V}(R(s,a)) + \mathbb{V}(P_{s,a}, V^\star_{h+1}) = 0,$$

which means $\mathsf{Var}^\Sigma_{(k)} = 0$. However, let π be a policy with $\pi_H(s_1) = a'$ for a certain state s_1, and $\pi_h(s) = a_1$ for any other h or s. Then π yields cumulative rewards of 1 and $1 - 1/H$, each with non-zero probabilities. So $\mathsf{Var}^\star > 0$.

This example shows that deterministic MDPs are not the only MDPs satisfying $\mathsf{Var}^\Sigma_{(k)} = 0$, and that $\mathsf{Var}^\Sigma_{(k)} = 0$ does not imply $\mathsf{Var}^\star = 0$.

When $\mathsf{Var}^\Sigma_{(k)} = 1/4 > \mathsf{Var}^\star$. Consider the *time-homogeneous* MDP in Figure 1: $P_{s_1,a,s_2} = p$ for any $a \in \mathcal{A}$, and the rest probability is into an MDP with *no reward at all*. s_2 is a state which we want to have a high $\mathsf{Var}^\Sigma_{(k)}$: $P_{s_2,a,s_3} = P_{s_2,a,s_4} = 1/2$, where s_3 and s_4 are states with value 0 and 1 respectively. Thus at $s_2, a, h = 3$,

$$\mathsf{Var}^\Sigma_{(k)} \geqslant \mathbb{V}(R(s_2,a)) + \mathbb{V}(P_{s_2,a}, V^\star_3) = \frac{1}{4}.$$

We also have that for any policy π, $V^\pi_1(s_1) = p/2$, so by Lemma 20, $\mathsf{Var}^\star \leqslant p/2$. Taking p arbitrarily small gives an arbitrarily large gap between $\mathsf{Var}^\Sigma_{(k)}$ and $\mathsf{Var}^\star$.

This example shows that a small $\mathsf{Var}^\star$ does not imply a small $\mathsf{Var}^\Sigma_{(k)}$, so using only $\mathsf{Var}^\Sigma_{(k)}$ is insufficient.

Algorithm 1 MVP-V

Input and initialize: Logarithm term ι; Trigger set $\mathcal{L} \leftarrow \{2^{i-1} \mid 2^i \leqslant KH, i = 1, 2, \ldots\}$.
Set all $N(s,a)$, $n(s,a)$, $\theta(s,a)$, $\phi(s,a)$, $N(s,a,s')$, $\widehat{P}_{s,a,s'}$ to be 0 and all $Q_h(s,a)$, $V_h(s)$ to be 1.
for $k = 1, 2, \ldots, K$ **do**
  Observe s_1^k.
  for $h = 1, 2, \ldots, H$ **do**
    Take action $a_h^k = \arg\max_a Q_h(s_h^k, a)$;
    Receive reward r_h^k and observe s_{h+1}^k;
    Set $(s, a, s', r) \leftarrow (s_h^k, a_h^k, s_{h+1}^k, r_h^k)$;
    Set $N(s,a) \overset{+}{\leftarrow} 1$, $\theta(s,a) \overset{+}{\leftarrow} r$, $\phi(s,a) \overset{+}{\leftarrow} r^2$, $N(s,a,s') \overset{+}{\leftarrow} 1$.
    if $N(s,a) \in \mathcal{L}$ **then**
      Set $\widehat{r}(s,a) \leftarrow \theta(s,a)/N(s,a)$;
      Set $\widehat{\mathsf{VarR}}(s,a) \leftarrow \phi(s,a)/N(s,a) - \widehat{r}(s,a)^2$;
      Set $\widehat{P}_{s,a,\widetilde{s}} \leftarrow N(s,a,\widetilde{s})/N(s,a)$ for all $\widetilde{s} \in \mathcal{S}$;
      Set $n(s,a) \leftarrow N(s,a)$;
      Set TRIGGERED = TRUE.
    end if
  end for
  if TRIGGERED **then**
    for $h = H, H-1, \ldots, 1$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      Set
$$b_h(s,a) \leftarrow 4\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V_{h+1})\iota}{\max\{n(s,a),1\}}} + 2\sqrt{\frac{\widehat{\mathsf{VarR}}(s,a)\iota}{\max\{n(s,a),1\}}} + \frac{21\iota}{\max\{n(s,a),1\}};$$
$$Q_h(s,a) \leftarrow \min\{1, \widehat{r}(s,a) + \widehat{P}_{s,a}V_{h+1} + b_h(s,a)\};$$
$$V_h(s) \leftarrow \max_a Q_h(s,a).$$
    end for
    Set TRIGGERED = FALSE.
  end if
end for

5. Results of the Model-Based Algorithm

We propose MVP-V (Algorithm 1, where "V" stands for "Variance-dependent"), a model-based algorithm with a variance-dependent regret bound. Based on MVP (Zhang et al., 2021a) which is minimax optimal under Conditions 1 and 2, we make necessary alterations to make the regret variance-dependent.

Common notations. These are notations shared with our model-free algorithm. Let s_h^k, a_h^k and r_h^k denote the state, action and reward at the h-th step of the k-th episode. Let V_h^k and Q_h^k denote V_h and Q_h at the beginning of the k-th episode. $\widetilde{O}$ hides $\mathsf{polylog}(H, S, A, K, 1/\delta)$ factors.

Algorithm description. MVP-V re-plans whenever a state-action pair's visitation is doubled. The bonus function depends on the variance of the next-step value functions. It achieves variance-dependent regret by using the empirical variances of rewards in the bonus, as opposed to using the empirical rewards themselves in MVP. MVP-V is capable of handling Conditions 1 and 2 and Assumption 3.

Theorem 7. *Assume that Conditions 1 and 2 hold. Let $\delta \in (0,1)$ be the confidence parameter and that H, S, A, K, δ be known. With probability at least $1-\delta$, the regret of MVP-V (Algorithm 1) run with*

$$\iota = 99\left(\ln\left(\frac{3000^2 H^5 S^7 A^5 K^5}{\delta^2}\right) + 1\right)$$

is bounded by

$$\mathsf{Regret}(K) \leqslant \widetilde{O}(\sqrt{\min\{\mathsf{Var}_K^{\Sigma}, \mathsf{Var}^{\star} K\} SA} + \Gamma SA).$$

When Condition 1 holds, we have $\mathsf{Var}^{\star} \leqslant 1$. Thus, our result is better than the $\widetilde{O}(\sqrt{SAK} + S^2A)$ regret of MVP, and achieves the *horizon-free* (only logarithm dependency on H) property. It is also strictly better than the $\widetilde{O}(\sqrt{H\mathbb{Q}^{\star} \cdot SAK} + H^{5/2}S^2A))$ regret in Zanette & Brunskill (2019).

Proof sketch. See Appendix B.2 for the rigorous proof. We follow the outline in Zhang et al. (2021a), realizing that the total regret is upper-bounded by $M_1 + M_2 + M_3$, where

$$M_1 \approx \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k, a_h^k} - \mathbf{1}_{s_{h+1}^k})V_{h+1}^k,$$
$$M_2 \approx \sum_{k=1}^{K}\sum_{h=1}^{H}(V_h^k(s_h^k) - r(s_h^k, a_h^k) - P_{s_h^k, a_h^k}V_{h+1}^k),$$
$$M_3 \approx \sum_{k=1}^{K}\left(\sum_{h=1}^{H} r(s_h^k, a_h^k) - V_1^{\pi^k}(s_1^k)\right).$$

We expand $r(s_h^k, a_h^k)$ by Bellman equation to derive a tighter bound for M_3. This change is necessary to remove a variance-independent $\widetilde{O}(\sqrt{K})$ term. M_1, M_2, M_3 can be then related to a crucial variance term

$$M_4 \approx \sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbb{V}(R(s_h^k, a_h^k)) + \mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^k))$$

so the regret is approximately $\widetilde{O}(\sqrt{SAM_4})$. The difference between Var_K^{Σ} and M_4 is of a lower order. To upper bound M_4 directly, we introduce *bonus-correction* terms

$$\mathsf{bc}_h^k(s,a) := V_h^k(s) - P_{s,a}V_{h+1}^k - r(s,a).$$

Let $\mathsf{BC}_h^k(s) := \mathsf{bc}_h^k(s,a) + P_{s,a}\mathsf{BC}_{h+1}^k$ with $a = \pi_h^k(s)$, then it can be proven that $\mathsf{BC}_h^k(s) = V_h^k(s) - V_h^{\pi^k}(s)$. Thus, M_4 can be bounded by the sum of

$$Z \approx \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k, a_h^k}, \mathsf{BC}_{h+1}^k)$$

and

$$W = \sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbb{V}(R(s_h^k,a_h^k)) + \mathbb{V}(P_{s_h^k,a_h^k}, V_{h+1}^{\pi^k})),$$

where Z is of a lower order and $W \leqslant \widetilde{O}(\mathsf{Var}^\star K)$. However, the bound of W cannot be derived using martingale concentration inequalities directly, because the summation of variances within an episode can be of order $\Omega(H)$, which will introduce a constant term of H, ruining the *horizon-free* property. We first prove that the total variance in an episode is bounded by $\widetilde{O}(1)$ with high probability, then the martingale concentration inequality can be applied to terms *truncated* to $\widetilde{O}(1)$. To get the Γ-dependency in the lower order term, we observe that $P_{s,a} = 0 \implies \widehat{P}_{s,a} = 0$ and put this into concentration bounds.

Corollaries. We study deterministic MDPs first.

Corollary 8. *Assume that Conditions 1 and 2 and Assumption 3 hold. Let $\delta \in (0,1)$ be the confidence parameter and that H, S, A, K, δ be known. With probability at least $1-\delta$, the regret of* `MVP-V` *(Algorithm 1) run with $\iota = 99(\ln(3000^2 H^5 S^7 A^5 K^5/\delta^2)+1)$ is bounded by* $\mathsf{Regret}(K) \leqslant \widetilde{O}(SA)$.

This is because $\mathsf{Var}^\star = 0$ and $\Gamma = 1$ when the MDP is deterministic. With a more refined analysis, we can totally eliminate the dependency on δ. Up to logarithm factors, `MVP-V` matches the lower bound of $\Omega(SA)$. So `MVP-V` is *minimax optimal for the class of deterministic MDPs*.

Another corollary arises when we remove Conditions 1 and 2. For `MVP-V` to work properly, we need to apply the conversion methods written below the conditions.

Corollary 9. *Let $\delta \in (0,1)$ be the confidence parameter and that H, S, A, K, δ be known. With probability at least $1-\delta$, the regret of* `MVP-V` *(Algorithm 1) run with $\iota = 99(\ln(3000^2 H^{12} S^7 A^5 K^5/\delta^2)+1)$ is bounded by*

$$\mathsf{Regret}(K) \leqslant \widetilde{O}(\sqrt{\min\{\mathsf{Var}_K^\Sigma, \mathsf{Var}^\star K\} HSA} + H^2\Gamma SA).$$

Readers may notice that the scaling in main order term is not standard. This is because when removing Condition 1, Var_K^Σ and $\mathsf{Var}^\star$ automatically scale by H^2.

6. Results of the Model-Free Algorithm

We propose `UCB-Advantage-V` (Algorithm 2) to initiate the study of model-free algorithms with variance-dependent regrets.

Algorithm description. In `UCB-Advantage-V`, the update of value functions is triggered by the beginning of stages for each (s,a,h) triple separately, and the stage design approximately makes use of the last $1/H$ fraction of data. Besides, the algorithm maintains *reference values* by assigning value functions to them at another frequency. The update rule using *the reference value decomposition* can be illustrated as:

$$Q_h(s,a) \leftarrow \widehat{P_{s,a,h} V_{h+1}^{\mathsf{ref}}} + P_{s,a,h}(\widehat{V_{h+1} - V_{h+1}^{\mathsf{ref}}}) + \widehat{r}_h(s,a) + b_h^k(s,a),$$

Algorithm 2 `UCB-Advantage-V`

Input and initialize: Logarithm term ι; Stage lengths $e_1 = H$, $e_{i+1} = \lfloor (1+1/H)e_i \rfloor$ and stage trigger set $\mathcal{L} \leftarrow \{\sum_{i=1}^{j} e_i \mid j = 1,2,\ldots\}$; Reference trigger set $\mathcal{R} \leftarrow \{60000 \cdot 2^{2i} SAH^3\iota \mid i = 1,2,\ldots,\iota^\star\}$.
Set all $N_h(s,a)$, $\check{N}_h(s,a)$, $\theta_h(s,a)$, $\phi_h(s,a)$, $\check{\upsilon}_h(s,a)$, $\check{\mu}_h(s,a)$, $\check{\sigma}_h(s,a)$, $\mu_h^{\mathsf{ref}}(s,a)$, $\sigma_h^{\mathsf{ref}}(s,a)$ to be 0 and all $V_h(s), Q_h(s,a), V_h^{\mathsf{ref}}(s,a)$ to be H.
for $k = 1,2,\ldots,K$ **do**
  Observe s_1^k.
  for $h = 1,2,\ldots,H$ **do**
    Take action $a_h^k = \arg\max_a Q_h(s_h^k,a)$;
    Receive reward r_h^k and observe s_{h+1}^k;
    Update accumulators:
    $n := N_h(s_h^k,a_h^k) \overset{+}{\leftarrow} 1,\ \check{n} := \check{N}_h(s_h^k,a_h^k) \overset{+}{\leftarrow} 1;$
    $\theta := \theta_h(s_h^k,a_h^k) \overset{+}{\leftarrow} r_h^k,\ \phi := \phi_h(s_h^k,a_h^k) \overset{+}{\leftarrow} (r_h^k)^2;$
    $\check{\upsilon} := \check{\upsilon}_h(s_h^k,a_h^k) \overset{+}{\leftarrow} V_{h+1}(s_{h+1}^k);$
    $\check{\mu} := \check{\mu}_h(s_h^k,a_h^k) \overset{+}{\leftarrow} V_{h+1}(s_{h+1}^k) - V_{h+1}^{\mathsf{ref}}(s_{h+1}^k);$
    $\check{\sigma} := \check{\sigma}_h(s_h^k,a_h^k) \overset{+}{\leftarrow} (V_{h+1}(s_{h+1}^k) - V_{h+1}^{\mathsf{ref}}(s_{h+1}^k))^2;$
    $\mu^{\mathsf{ref}} := \mu_h^{\mathsf{ref}}(s_h^k,a_h^k) \overset{+}{\leftarrow} V_{h+1}^{\mathsf{ref}}(s_{h+1}^k);$
    $\sigma^{\mathsf{ref}} := \sigma_h^{\mathsf{ref}}(s_h^k,a_h^k) \overset{+}{\leftarrow} (V_{h+1}^{\mathsf{ref}}(s_{h+1}^k))^2.$
    if $n \in \mathcal{L}$ **then**
      Set
      $\widehat{r} \leftarrow \dfrac{\theta}{n},\ \widehat{\mathsf{VarR}} \leftarrow \dfrac{\phi}{n} - \left(\dfrac{\theta}{n}\right)^2;$
      $\bar{b} \leftarrow 2\sqrt{\dfrac{H^2\iota}{\check{n}}};$
      $\nu^{\mathsf{ref}} \leftarrow \dfrac{\sigma^{\mathsf{ref}}}{n} - \left(\dfrac{\mu^{\mathsf{ref}}}{n}\right)^2,\ \check{\nu} = \dfrac{\check{\sigma}}{\check{n}} - \left(\dfrac{\check{\mu}}{\check{n}}\right)^2;$
      $b \leftarrow 4\sqrt{\dfrac{\nu^{\mathsf{ref}}\iota}{n}} + 4\sqrt{\dfrac{\check{\nu}\iota}{\check{n}}} + 2\sqrt{\dfrac{\widehat{\mathsf{VarR}}\iota}{n}} + \dfrac{90H\iota}{\check{n}};$
      $Q_h(s_h^k,a_h^k) \leftarrow \min\left\{\widehat{r} + \dfrac{\check{\upsilon}}{\check{n}} + \bar{b},\ \widehat{r} + \dfrac{\mu^{\mathsf{ref}}}{n} + \dfrac{\check{\mu}}{\check{n}} + b,\ Q_h(s_h^k,a_h^k)\right\};$
      $V_h(s_h^k) \leftarrow \max Q_h(s_h^k,a).$
      Set $N_h(s_h^k,a_h^k) \leftarrow 0$, $\check{\mu}_h(s_h^k,a_h^k) \leftarrow 0$, $\check{\upsilon}_h(s_h^k,a_h^k) \leftarrow 0$, $\check{\sigma}_h(s_h^k,a_h^k) \leftarrow 0$.
    end if
    if $\sum_a N_h(s_h^k,a) \in \mathcal{R}$ **then** $V_h^{\mathsf{ref}}(s_h^k) \leftarrow V_h(s_h^k)$.
  end for
end for

where $b_h^k(s,a)$ is the bonus, $\widehat{r}_h(s,a)$, $\widehat{P_{s,a,h}V_{h+1}^{\mathsf{ref}}}$ and $\widehat{P_{s,a,h}(V_{h+1}-V_{h+1}^{\mathsf{ref}})}$ are empirical estimates of $r_h(s,a)$, $P_{s,a,h}V_{h+1}^{\mathsf{ref}}$ and $P_{s,a,h}(V_{h+1}-V_{h+1}^{\mathsf{ref}})$ respectively. In addition, a very simple update rule

$$Q_h(s,a) \leftarrow \widehat{P_{s,a,h}V_{h+1}} + \widehat{r}_h(s,a) + b_h^k(s,a)$$

is also in use to provide a guarantee of uniform convergence of estimated value functions.

We make three major alterations to `UCB-Advantage`: ① We use empirical variances of rewards in bonuses. ② Due to a more refined analysis, we remove the $\widetilde{O}(H(n^{-3/4}+\check{n}^{-3/4}))$ term in bonuses. ③ The reference value functions are updated in *a capped-doubling manner* (cf. Line 13).

Alteration ③ is crucial to make the main order term variance-dependent, because there exist constant gaps between reference values and optimal values, whose summation contributes to the regret as the main order term in `UCB-Advantage`. By choosing an appropriate number of updates, we can *balance between the total constant gap and the deviation introduced by frequent updates*, making the total variance the only factor in the main order term.

Theorem 10. *Let $\delta \in (0,1)$ be the confidence parameter and that H, S, A, K, δ be known. With probability at least $1-\delta$, the regret of `UCB-Advantage-V` (Algorithm 2) run with*

$$\iota = 99\left(\ln\left(\frac{7000^2(HSAK)^5}{\delta^2}\right)+1\right)$$

and

$$i^\star = \left\lceil \frac{1}{2}\log_2\left(\frac{K}{H^5S^3A\iota^2}\right)\right\rceil$$

is bounded by

$$\mathsf{Regret}(K) \leqslant \widetilde{O}(\sqrt{\min\{\mathsf{Var}_K^\Sigma, \mathsf{Var}^\star K\}HSA} + \sqrt[4]{H^{15}S^5A^3K}).$$

We have $\mathsf{Var}^\star \leqslant H^2$, so our result is strictly better than the $\widetilde{O}(\sqrt{H^3SAK}+\sqrt[4]{H^{33}S^8A^6K})$ regret of `UCB-Advantage`.

Extra notations. Let $V_h^{\mathsf{ref},k}$ denote V_h^{ref} at the beginning of the k-th episode, and $V_h^{\mathsf{REF}} := V_h^{\mathsf{ref},K+1}$ denote the final reference value function. Let $\nu_h^{\mathsf{ref},k}, \check{\nu}_h^k, b_h^k$ denote $\nu^{\mathsf{ref}}, \check{\nu}, b$ for the value of $Q_h^k(s_h^k,a_h^k)$. Let $N_h^k(s)$ denote $\sum_a N_h(s,a)$ at the beginning of the k-th episode. Let n_h^k and $\check{n}_h^k$ be the total number of visits to (s_h^k,a_h^k,h) prior to the current stage and during the stage immediately before the current stage with respect to the same triple.

Proof sketch. See Appendix B.3 for the rigorous proof. From Zhang et al. (2020b), the regret is roughly $\sum_{k=1}^K\sum_{h=1}^H(\psi_{h+1}^k+\xi_{h+1}^k+\phi_{h+1}^k+b_h^k)$, where

$$\begin{aligned}
\psi_{h+1}^k &\approx V_{h+1}^{\mathsf{ref},k}(s_{h+1}^k) - V_{h+1}^{\mathsf{REF}}(s_{h+1}^k),\\
\xi_{h+1}^k &\approx (P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k})(V_{h+1}^k - V_{h+1}^\star),\\
\phi_{h+1}^k &= (P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k})(V_{h+1}^\star - V_{h+1}^{\pi^k}).
\end{aligned}$$

All these four terms are bounded loosely in Zhang et al. (2020b) such that they are all main order terms. To establish a variance-dependent regret, we prove that only the b term is the main order term *after our aforementioned alterations*. The ϕ term is a martingale and shown to be $\widetilde{O}(H)$. For the rest terms, we need the following argument:

$$N_h^k(s) \geqslant N_0(\epsilon) = \widetilde{O}\left(\frac{H^5SA}{\epsilon^2}\right) \implies 0 \leqslant V_h^k(s) - V_h^\star(s) \leqslant \epsilon.$$

Notice that the reference trigger set $\mathcal{R}$ in Algorithm 2 is composed of $N_0(\beta_i)$ for $i \in [i^\star]$ where $\beta_i := H/2^i$. There is a constant gap of at least $\beta_{i^\star}$ between $V_h^{\mathsf{REF}}(s)$ and $V_h^\star(s)$ in the worst case, because the number of updates is capped by $i^\star$. This argument branches into two corollaries. The first one is we can bound value gaps directly:

$$\sum_{k=1}^K\sum_{h=1}^H(V_h^k(s_h^k) - V_h^\star(s_h^k))^2 \leqslant \widetilde{O}(H^6SA).$$

This can be utilized to bound the ξ term. The second one is that, we define

$$B_h^{\mathsf{ref},k}(s) := \sum_{i=1}^{i^\star}\beta_{i-1}\mathbb{1}[N_0(\beta_{i-1}) \leqslant N_h^k(s) < N_0(\beta_i)],$$

then $V_h^{\mathsf{ref},k}(s) - V_h^{\mathsf{REF}}(s) \leqslant B_h^{\mathsf{ref},k}(s)$, $V_h^{\mathsf{ref},k}(s) - V_h^\star(s) \leqslant B_h^{\mathsf{ref},k}(s) + \beta_{i^\star}$ and

$$\sum_{k,h} B_h^{\mathsf{ref},k}(s_h^k) \leqslant \widetilde{O}(H^5S^2A2^{i^\star}), \sum_{k,h}(B_h^{\mathsf{ref},k}(s_h^k))^2 \leqslant \widetilde{O}(H^6S^2Ai^\star).$$

So we can directly bound the ψ term. We show that

$$\begin{aligned}
\nu_h^{\mathsf{ref},k} &\lessapprox \widetilde{O}(\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^\star) + (B_{h+1}^{\mathsf{ref},k}(s_{h+1}^k))^2 + \beta_{i^\star}^2),\\
\check{\nu}_h^k &\lessapprox O((B_{h+1}^{\mathsf{ref},k}(s_{h+1}^k))^2 + \beta_{i^\star}^2).
\end{aligned}$$

The b term is hence bounded by

$$\widetilde{O}(\sqrt{\mathsf{Var}_K^\Sigma HSA} + \sqrt{H^5SAK/2^{2i^\star}}).$$

Analogous to the proof of Theorem 7, the difference between Var_K^Σ and $\mathsf{Var}^\star K$ is of a lower order. Finally, the lower order terms are

$$\widetilde{O}(\sqrt{H^5SAK/2^{2i^\star}} + H^5S^2A2^{i^\star}).$$

We derive Theorem 10 by choosing the optimal $i^\star$.

Corollary. We study deterministic MDPs.

Corollary 11. *Assume that Assumption 3 holds. Let $\delta \in (0,1)$ be the confidence parameter and that H, S, A, K, δ be known. With probability at least $1-\delta$, the regret of* `UCB-Advantage-V` *(Algorithm 2) run with $\iota = 99(\ln(7000^2(HSAK)^5/\delta^2)+1)$ and $i^\star = \lceil 1/2 \cdot \log_2(K/H^5S^3A\iota^2)\rceil$ is bounded by*

$$\mathsf{Regret}(K) \leqslant \widetilde{O}(\sqrt[4]{H^{15}S^5A^3K}).$$

Notice that the regret under Assumption 3 is not constant which we desire, this may be due to some fundamental limit of model-free algorithms. However, since the research on model-free algorithms is still at its nascent stage and there lack thorough understanding, our result provides the first look into the potential of such algorithms.

Intuitively, for any algorithm to have a constant regret on deterministic MDPs, its value functions should also converge in a constant steps. Previous model-free algorithms all use historical data to estimate value functions. These data are biased because some of them are not up-to-date, making value functions hard to converge in a constant steps. Here we identify difficulties for existing algorithms to be variance-dependent for all K-related terms.

`Q-learning (UCB-B)` (Jin et al., 2018). In their proof of Lemma C.3, when bounding $|P_3 - P_4|$, there is a variance-independent $1/\sqrt{t}$ term in the gap between the estimations and true values. Notice that their result is possible to be variance-dependent by not loosening $\sqrt{H^7SA\iota/t} \leqslant H + H^6SA\iota/t$ above their Equation (C.10) while introducing a variance-independent $K^{1/4}$ term.

`UCB-Advantage` (Zhang et al., 2020b). There are biases in the reference value functions, because they are updated for only finite times. If the update is not capped by a threshold, readers can easily verify that the ψ term will become a variance-independent main order term.

`Q-EarlySettled-Advantage` (Li et al., 2021). There is a same issue about the constant gap between the reference value and the optimal value when bounding $\mathcal{R}_3$ defined in their Equation (39c).

7. Regret Lower Bounds

We show that for any algorithm and any variance $\mathcal{V}$, there always exists an MDP such that the regret main order terms of Theorem 7, Corollary 9 and Theorem 10 are tight. This means that `MVP-V` and `UCB-Advantage-V` are *minimax optimal for the class of variance-bounded MDPs*. The proofs for this section are deferred to Appendix B.4.

Theorem 12. *Assume $S \geqslant 6$, $A \geqslant 2$, $H \geqslant 3\lfloor \log_2(S-2)\rfloor$ and $0 < \mathcal{V} \leqslant O(1)$. For any algorithm $\boldsymbol{\pi}$, there exists an MDP $\mathcal{M}_{\boldsymbol{\pi}}$ such that:*

- *It satisfies Conditions 1 and 2;*
- *$\mathsf{Var}_\tau^\Sigma, \mathsf{Var}^\star = \Theta(\mathcal{V})$ for any possible trajectory τ;*
- *For $K \geqslant SA$, the expected regret of $\boldsymbol{\pi}$ in $\mathcal{M}_{\boldsymbol{\pi}}$ after K episodes satisfies*

$$\mathbb{E}\left[\sum_{k=1}^{K}(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)) \,\middle|\, \mathcal{M}_{\boldsymbol{\pi}}, \boldsymbol{\pi}\right] = \Omega(\sqrt{\mathcal{V}SAK}).$$

Theorem 13. *Assume $S \geqslant 6$, $A \geqslant 2$, $H \geqslant 3\lfloor \log_2(S-2)\rfloor$ and $0 < \mathcal{V} \leqslant O(H^2)$. For any algorithm $\boldsymbol{\pi}$, there exists an MDP $\mathcal{M}_{\boldsymbol{\pi}}$ such that:*

- *$\mathsf{Var}_\tau^\Sigma, \mathsf{Var}^\star = \Theta(\mathcal{V})$ for any possible trajectory τ;*
- *For $K \geqslant HSA$, the expected regret of $\boldsymbol{\pi}$ in $\mathcal{M}_{\boldsymbol{\pi}}$ after K episodes satisfies*

$$\mathbb{E}\left[\sum_{k=1}^{K}(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)) \,\middle|\, \mathcal{M}_{\boldsymbol{\pi}}, \boldsymbol{\pi}\right] = \Omega(\sqrt{\mathcal{V}HSAK}).$$

8. Conclusion

We systematically study variance-dependent regret bounds for MDPs by introducing new notions of variances, proposing model-based and model-free algorithms respectively, and providing regret lower bounds for the class of variance-bounded MDPs. Our results improve upon the previous algorithms and achieves minimax optimal regrets for the class of variance-bounded MDPs. Our model-based algorithm is minimax optimal for deterministic MDPs. Finally, we identify some possible limit of current model-free algorithms. One possible future direction is to find a new model-free algorithm with a constant regret for deterministic MDPs.

Acknowledgements

SSD acknowledges the support of NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844, NSF IIS 2229881.

References

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *ICML*, 2017.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32, 2019.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

Bertsekas, D. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. In *NeurIPS*, 2021.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *NIPS*, 2017.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In Feldman, V., Ligett, K., and Sabato, S. (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 578–598. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating markov decision processes. *Advances in Neural Information Processing Systems*, 31, 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018b.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44, 02 2016. doi: 10.1287/moor.2017.0928.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Kim, Y., Yang, I., and Jun, K.-S. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *arXiv preprint arXiv:2111.03289*, 2021.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776, 2021.

Li, X. and Sun, Q. Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *arXiv preprint arXiv:2303.05606*, 2023.

Maillard, O.-A., Mann, T. A., and Mannor, S. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 27, 2014.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K., and Roberts, S. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

Russo, D. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pp. 770–805. PMLR, 2018.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. In *Neural Information Processing Systems*, 2021.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

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

Xiong, Z., Shen, R., and Du, S. S. Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703*, 2021.

Xu, H., Ma, T., and Du, S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *ICML*, 2019.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, Z., Du, S. S., and Ji, X. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020b.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *COLT*, 2021a.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021b.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12653–12662. PMLR, 2021c.

Zhang, Z., Ji, X., and Du, S. S. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Annual Conference Computational Learning Theory*, 2022.

Zhao, H., Zhou, D., He, J., and Gu, Q. Bandit learning with general function classes: Heteroscedastic noise and variance-dependent regret bounds. *arXiv preprint arXiv:2202.13603*, 2022.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. *arXiv preprint arXiv:2302.10371*, 2023.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

Zhou, R., Wang, R., and Du, S. S. Horizon-free reinforcement learning for latent markov decision processes. *arXiv preprint arXiv:2210.11604*, 2022.

A. Technical Lemmas

Lemma 14 (Hoeffding's Inequality). *Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables with values in $[0, b]$ and let $\delta > 0$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > b\sqrt{\frac{\ln(2/\delta)}{2n}}\right] \leqslant \delta.$$

Lemma 15 (Bennett's Inequality, Theorem 3 in Maurer & Pontil (2009)). *Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables with values in $[0, b]$ and let $\delta > 0$. Define $\mathbb{V}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\mathbb{V}[Z]\ln(2/\delta)}{n}} + \frac{b\ln(2/\delta)}{n}\right] \leqslant \delta.$$

Lemma 16 (Theorem 4 in Maurer & Pontil (2009)). *Let $Z, Z_1, \ldots, Z_n$ $(n \geqslant 2)$ be i.i.d. random variables with values in $[0, b]$ and let $\delta > 0$. Define $\bar{Z} = \frac{1}{n}Z_i$ and $\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\hat{V}_n\ln(2/\delta)}{n-1}} + \frac{7b\ln(2/\delta)}{3(n-1)}\right] \leqslant \delta.$$

Lemma 17 (Lemma 11 in (Zhang et al., 2021c)). *Let $(M_n)_{n\geqslant 0}$ be a martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \leqslant c$ for some $c > 0$ and any $n \geqslant 1$. Let $\mathrm{Var}_n = \sum_{k=1}^{n}\mathbb{E}[(M_k - M_{k-1})^2|\mathcal{F}_{k-1}]$ for $n \geqslant 0$, where $\mathcal{F}_k = \sigma(M_1, \ldots, M_k)$. Then for any positive integer n and any $\epsilon, \delta > 0$, we have that*

$$\mathbb{P}\left[|M_n| \geqslant 2\sqrt{2\mathrm{Var}_n\ln(1/\delta)} + 2\sqrt{\epsilon\ln(1/\delta)} + 2c\ln(1/\delta)\right] \leqslant 2\left(\log_2\left(\frac{nc^2}{\epsilon}\right) + 1\right)\delta.$$

Lemma 18 (Lemma 10 in Zhang et al. (2022)). *Let $X_1, X_2, \ldots$ be a sequence of random variables taking values in $[0, l]$. Define $\mathcal{F}_k = \sigma(X_1, X_2, \ldots, X_{k-1})$ and $Y_k = \mathbb{E}[X_k \mid \mathcal{F}_k]$ for $k \geqslant 1$. For any $\delta > 0$, we have that*

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \geqslant 3\sum_{k=1}^{n} Y_k + l\ln(1/\delta)\right] \leqslant \delta,$$
$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geqslant 3\sum_{k=1}^{n} X_k + l\ln(1/\delta)\right] \leqslant \delta.$$

Lemma 19 (Lemma 30 in Chen et al. (2021)). *For any two random variables X, Y, we have*

$$\mathbb{V}(XY) \leqslant 2\mathbb{V}(X)(\sup|Y|)^2 + 2(\mathbb{E}[X])^2\mathbb{V}(Y).$$

Consequently, $\sup|X| \leqslant C$ *implies* $\mathbb{V}(X^2) \leqslant 4C^2\mathbb{V}(X)$.

Lemma 20 (Bhatia–Davis Inequality). *For any random variable X, $\mathbb{V}(X) \leqslant (\sup X - \mathbb{E}[X])(\mathbb{E}[X] - \inf X)$.*

B. Missing Proofs

B.1. Justification for Definition 6

Let $X_h^\pi(s)$ denote the random variable of cumulative reward starting from s as the h-th step. Clearly, $V_h^\pi(s) = \mathbb{E}[X_h^\pi(s)]$. We denote $\mathsf{Var}_h^\pi(s) := \mathbb{V}(X_h^\pi(s))$. Since $\pi \in \Pi$ is deterministic, let $a = \pi_h(s)$. Law of total variance states that $\mathbb{V}(Y) = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}(\mathbb{E}[Y|X])$, so

$$\begin{aligned}
\mathsf{Var}_h^\pi(s) &= \mathbb{E}_{r\sim R_h(s,a), s'\sim P_{s,a,h}}[\mathbb{V}(r + X_{h+1}^\pi(s'))] + \mathbb{V}_{r\sim R_h(s,a), s'\sim P_{s,a,h}}(\mathbb{E}[r + X_{h+1}^\pi(s')]) \\
&= \mathbb{E}_{r\sim R_h(s,a), s'\sim P_{s,a,h}}[\mathbb{V}(X_{h+1}^\pi(s'))] + \mathbb{V}_{r\sim R_h(s,a), s'\sim P_{s,a,h}}(r + \mathbb{E}[X_{h+1}^\pi(s')]) \\
&= \mathbb{E}_{s'\sim P_{s,a,h}}[\mathsf{Var}_{h+1}^\pi(s')] + \mathbb{V}_{r\sim R_h(s,a)}(r) + \mathbb{V}_{s'\sim P_{s,a,h}}(V_{h+1}^\pi(s'))
\end{aligned}$$

$$= P_{s,a,h}\mathsf{Var}^{\pi}_{h+1} + \mathbb{V}(R_h(s,a)) + \mathbb{V}(P_{s,a,h}, V^{\pi}_{h+1}).$$

Let $d^{\pi}_h \in \Delta(\mathcal{S})(\cdot|s)$ denote the state visitation distribution at the h-th step conditioned on the first state being s, i.e.,

$$d^{\pi}_h(s'|s) := \mathbb{P}_{\pi}[s_h = s' \mid s_1 = s].$$

By induction, we can prove that (with $a_h = \pi_h(s_h)$)

$$\begin{aligned}\mathsf{Var}^{\pi}_1(s) &= \sum_{h=1}^{H} \mathbb{E}_{s_h \sim d^{\pi}_h(\cdot|s)}[\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_{s_h,a_h,h}, V^{\pi}_{h+1})] \\ &= \mathbb{E}_{\pi}\left[\sum_{h=1}^{H}\left(\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_{s_h,a_h,h}, V^{\pi}_{h+1})\right) \,\middle|\, s_1 = s\right].\end{aligned}$$

B.2. Model-Based Algorithm: `MVP-V` (Algorithm 1)

Summary of notations. Let s^k_h, a^k_h and r^k_h denote the state, action and reward at the h-th step of the k-th episode. Let $V^k_h(s)$, $Q^k_h(s,a)$, $n^k(s,a)$ and $\widehat{P}^k_{s,a,s'}$ denote $V_h(s)$, $Q_h(s,a)$, $n(s,a)$ and $\widehat{P}_{s,a,s'}$ at the beginning of the k-th episode.

Let $\mathcal{K}$ be the set of indexes of episodes in which no update is triggered. By the update rule, it is obvious that $\left|\mathcal{K}^C\right| \leqslant SA(\log_2(KH)+1)$. Let $h_0(k)$ be the first time an update is triggered in the k-th episode if there is an update in this episode and otherwise $H+1$. Define $\mathcal{X}_0 := \{(k, h_0(k)) \mid k \in \mathcal{K}^C\}$ and $\mathcal{X} := \{(k,h) \mid k \in \mathcal{K}^C, h_0(k)+1 \leqslant h \leqslant H\}$.

Let $I(k,h) := \mathbb{1}[(k,h) \notin \mathcal{X}]$. We use the "check" notation to denote the original value timed with $I(k,h)$, e.g., $\check{V}^k_h := V^k_h I(k,h)$ and $\check{\beta}^k_h := \beta^k_h I(k,h)$.

Proof of Theorem 7. We first run `MVP-V` (Algorithm 1) with $\iota = 99(\ln(HSAK/\delta)+1)$ which is large enough for all the probabilistic inequalities to hold. This choice will make the success probability be $1 - \mathsf{poly}(S,A,H,K,\iota)\delta$. The lemmas are also proved assuming this choice of ι at first.

Based on Lemma 7 in Zhang et al. (2021a), by Lemmas 23 and 25 we have that

$$\mathsf{Regret}(K) \leqslant \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s^k_h,a^k_h}\check{V}^k_{h+1} - \check{V}^k_{h+1}(s^k_{h+1}))}_{=:M_1} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\check{\beta}^k_h(s^k_h,a^k_h)}_{=:M_2} + \underbrace{\sum_{k=1}^{K}\left(\sum_{h=1}^{H} r(s^k_h,a^k_h)I(k,h) - V^{\pi^k}_1(s^k_1)\right)}_{=:M_3} + \left|\mathcal{K}^C\right|.$$

We utilize Equation (39) in Zhang et al. (2021a): for any non-negative sequence $(w^k_h)_{k\in[K],h\in[H]}$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{I(k,h)}{n^k(s^k_h,a^k_h)} \leqslant O(SA\iota), \quad \sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{w^k_h I(k,h)}{n^k(s^k_h,a^k_h)}} \leqslant O\left(\sqrt{SA\iota\sum_{k=1}^{K}\sum_{h=1}^{H} w^k_h I(k,h)} + SA\iota\right).$$

Thus by Lemma 25 we have

$$M_2 \leqslant O\left(\sqrt{SA\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbb{V}(R(s^k_h,a^k_h)) + \mathbb{V}(P_{s^k_h,a^k_h}, V^k_{h+1}))I(k,h)}_{\approx:M_4}\iota} + \sqrt{\Gamma SA\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h}, V^k_{h+1} - V^{\star}_{h+1})I(k,h)}_{\approx:M_5}\iota}\right.$$
$$\left. + \sqrt{SA\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbb{V}(R(s^k_h,a^k_h)) + \mathbb{V}(P_{s^k_h,a^k_h}, V^{\star}_{h+1}))I(k,h)}_{\approx:M_6}\iota} + \Gamma SA\iota^2\right).$$

We need to substitute $I(k,h)$ with $I(k,h+1)$ to get the precise definition of M_4, M_5 and M_6. Such substitution only introduces an error of $O(|\mathcal{K}^C|)$. By $\mathbb{V}(X+Y) \leqslant 2\mathbb{V}(X) + 2\mathbb{V}(Y)$,

$$M_6 \leqslant O(M_4 + M_5), \quad \text{and} \quad M_4 \leqslant O(M_5 + M_6).$$

- If we use the former relation, $M_2 \leqslant O(\sqrt{SAM_4}\iota + \sqrt{\Gamma SAM_5}\iota + \Gamma SA\iota^2)$. Plugging in Lemma 26 and Lemma 28, we have

$$M_2 \leqslant O\left(\sqrt{\Gamma SAM_2}\iota + \sqrt{SA\sum_{k=1}^{K}\mathsf{Var}^{\pi^k}\iota} + \Gamma SA\iota^2\right).$$

  Solving the inequality gives

$$M_2 \leqslant O\left(\sqrt{SA\sum_{k=1}^{K}\mathsf{Var}^{\pi^k}\iota} + \Gamma SA\iota^2\right).$$

  By Lemma 8 of Zhang et al. (2021a), $M_1 \leqslant O(\sqrt{M_4\iota} + \iota)$. Further plugging in Lemma 27 gives

$$\mathsf{Regret}(K) \leqslant M_1 + M_2 + M_3 + |\mathcal{K}^C| \leqslant O\left(\sqrt{SA\sum_{k=1}^{K}\mathsf{Var}^{\pi^k}\iota} + \Gamma SA\iota^2\right) \leqslant O(\sqrt{\mathsf{Var}^\star SAK}\iota + \Gamma SA\iota^2).$$

- If we use the latter relation, $M_2 \leqslant O(\sqrt{SAM_6}\iota + \sqrt{\Gamma SAM_5}\iota + \Gamma SA\iota^2)$. First plug in Lemma 28, we have $M_2 \leqslant O(\sqrt{\Gamma SAM_2}\iota + \sqrt{SAM_6}\iota + \Gamma SA\iota^2)$, which implies

$$M_2 \leqslant O(\sqrt{SAM_6}\iota + \Gamma SA\iota^2).$$

  For the regret, we need $M_1 \leqslant O(\sqrt{M_4\iota} + \iota)$, Lemmas 27 and 29 and $M_4 \leqslant O(M_5 + M_6)$, so

$$\mathsf{Regret}(K) \leqslant O(\sqrt{SAM_6}\iota + \Gamma SA\iota^2) \leqslant O(\sqrt{\mathsf{Var}_K^\Sigma SA}\iota + \Gamma SA\iota^2).$$

The above results hold with probability at least $1 - 19HS^2AK\iota\delta$. To establish the final result, we need to scale δ to make the success probability be $1-\delta'$. We upper-bound ι by $100(HSAK/\sqrt{\delta}+1)$ and solve the inequality:

$$1900HS^2AK\left(\frac{HSAK}{\sqrt{\delta}} + 1\right)\delta \leqslant \delta'.$$

Take $\delta = (\delta'/3000H^2S^3A^2K^2)^2$. By $\ln(HSAK/(\delta/3000H^2S^3A^2K^2)^2) \leqslant O(\iota)$, we conclude the proof. □

Lemma 21. *Define the following events:*

$$\mathcal{E}_1 := \left\{\forall(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K],\ \left|(\widehat{P}^k_{s,a,s'} - P_{s,a,s'})V^\star_{h+1}\right| \leqslant 2\sqrt{\frac{\mathbb{V}(\widehat{P}^k_{s,a}, V^\star_{h+1})\iota}{n^k(s,a)}} + \frac{14\iota}{3n^k(s,a)}\right\}, \quad (1)$$

$$\mathcal{E}_2 := \left\{\forall(s,a,k) \in \mathcal{S}\times\mathcal{A}\times[K],\ \left|\widehat{r}^k(s,a) - r(s,a)\right| \leqslant 2\sqrt{\frac{\widehat{\mathsf{Var}\mathsf{R}}^k(s,a)\iota}{n^k(s,a)}} + \frac{14\iota}{3n^k(s,a)}\right\}, \quad (2)$$

$$\mathcal{E}_3 := \left\{\forall(s,a,s',k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[K],\ \left|\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right| \leqslant \sqrt{\frac{2P_{s,a,s'}\iota}{n^k(s,a)}} + \frac{\mathbb{1}[P_{s,a,s'}>0]\iota}{n^k(s,a)}\right\}, \quad (3)$$

$$\mathcal{E}_4 := \left\{\forall(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K],\ \left|(\widehat{P}^k_{s,a,s'} - P_{s,a,s'})V^\star_{h+1}\right| \leqslant \sqrt{\frac{2\mathbb{V}(P_{s,a}, V^\star_{h+1})\iota}{n^k(s,a)}} + \frac{\iota}{n^k(s,a)}\right\}. \quad (4)$$

We have that

$$\mathbb{P}[\mathcal{E}_1] \geqslant 1 - HSAK\iota\delta, \quad \mathbb{P}[\mathcal{E}_2] \geqslant 1 - SAK\iota\delta, \quad \mathbb{P}[\mathcal{E}_3] \geqslant 1 - S^2AK\iota\delta, \quad \mathbb{P}[\mathcal{E}_4] \geqslant 1 - HSAK\iota\delta.$$

Proof of Lemma 21. $\mathbb{P}[\mathcal{E}_1]$ and $\mathbb{P}[\mathcal{E}_2]$ are direct results by applying Lemma 16 and $\frac{1}{x-1} \leqslant \frac{2}{x}$, taking union bounds over the mentioned quantifiers and that $n^k(s,a) \in \{1, 2, \ldots, \lfloor \log_2(HK) \rfloor\}$. $\mathbb{P}[\mathcal{E}_3]$ and $\mathbb{P}[\mathcal{E}_4]$ are direct results by applying Lemma 15 and that $P_{s,a,s'} = 0 \implies \widehat{P}^k_{s,a,s'} = 0$, finally taking union bounds over the mentioned quantifiers and that $n^k(s,a) \in \{1, 2, \ldots, \lfloor \log_2(HK) \rfloor\}$. □

Lemma 22 (Adapted from Lemma 14 in Zhang et al. (2021a) and Lemma 16 in Tarbouriech et al. (2021))**.** *For any fixed dimension D, let $\Upsilon := \{v \in \mathbb{R}^D : v \geqslant 0, \|v\|_\infty \leqslant B\}$. For any two constants c_1, c_2 satisfying $c_1^2 \leqslant c_2$, let $f : \Delta([D]) \times \Upsilon \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $f(p, v, n, \iota) = pv + \max\left\{c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}, c_2\frac{B\iota}{n}\right\}$. Then for all $p \in \Delta([D]), v \in \Upsilon$ and $n, \iota > 0$,*

1. *$f(p, v, n, \iota)$ is non-decreasing in v, i.e.,*

$$\forall (v, v') \in \Upsilon^2, v \leqslant v', \text{ it holds that } f(p, v, n, \iota) \leqslant f(p, v', n, \iota);$$

2. *$f(p, v, n, \iota) \geqslant pv + \frac{c_1}{2}\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} + \frac{c_2}{2}\frac{B\iota}{n}$.*

Lemma 23. *Conditioned on the successful events of Lemma 21, we have that for any $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, $Q^k_h(s,a) \geqslant Q^\star_h(s,a)$.*

Proof of Lemma 23. Let k be fixed and omit it for simplicity. The proof is conducted by induction in the order of $h = H+1, H, \ldots, 1$. $Q_{H+1}(s,a) = 0 \geqslant 0 = Q^\star_{H+1}(s,a)$ holds trivially for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Now assume $Q_{h+1}(s,a) \geqslant Q^\star_{h+1}(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, hence $V_{h+1}(s) = \max_{a \in \mathcal{A}} Q_{h+1}(s,a) \geqslant \max_{a \in \mathcal{A}} Q^\star_{h+1}(s,a) = V^\star_{h+1}(s)$ for any $s \in \mathcal{S}$.

$$\begin{aligned}
&\widehat{r}(s,a) + \widehat{P}_{s,a}V_{h+1} + b_h(s,a) \\
&= \left(\widehat{r}(s,a) + 2\sqrt{\frac{\widehat{\mathsf{VarR}}(s,a)\iota}{n(s,a)}} + \frac{5\iota}{n(s,a)}\right) + \left(\widehat{P}_{s,a}V_{h+1} + 4\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V_{h+1})\iota}{n(s,a)}} + \frac{16\iota}{n(s,a)}\right) \\
&\overset{\text{(i)}}{\geqslant} r(s,a) + \underbrace{\widehat{P}_{s,a}V_{h+1} + \max\left\{4\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V_{h+1})\iota}{n(s,a)}}, \frac{16\iota}{n(s,a)}\right\}}_{=f(\widehat{P}_{s,a}, V_{h+1}, \iota, n(s,a))} \\
&\overset{\text{(ii)}}{\geqslant} r(s,a) + \underbrace{\widehat{P}_{s,a}V^\star_{h+1} + \max\left\{4\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V^\star_{h+1})\iota}{n(s,a)}}, \frac{16\iota}{n(s,a)}\right\}}_{=f(\widehat{P}_{s,a}, V^\star_{h+1}, \iota, n(s,a))} \\
&\overset{\text{(iii)}}{\geqslant} r(s,a) + \widehat{P}_{s,a}V^\star_{h+1} + 2\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V^\star_{h+1})\iota}{n(s,a)}} + \frac{8\iota}{n(s,a)} \\
&\overset{\text{(iv)}}{\geqslant} r(s,a) + P_{s,a}V^\star_{h+1} \\
&= Q^\star_h(s,a),
\end{aligned}$$

where (i) is by $\mathcal{E}_2$ (Equation (2)); (ii) is by recognizing the last part as the function in Lemma 22, $(c_1, c_2, B) = (4, 16, 1)$ satisfying that $c_1^2 \leqslant c_2$ and using the first property based on the induction that $V_{h+1} \geqslant V^\star_{h+1}$; (iii) is by the second property in Lemma 22; (iv) is $\mathcal{E}_1$ (Equation (1)). So $Q_h(s,a) \geqslant Q^\star_h(s,a)$. □

Lemma 24. *With probability at least $1 - 2SAK\iota\delta$, we have that for any $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times [K]$,*

$$\widehat{\mathsf{VarR}}^k(s,a) \leqslant O\left(\mathbb{V}(R(s,a)) + \frac{\iota}{n^k(s,a)}\right).$$

Proof of Lemma 24. Let s, a, k be fixed and omit k for simplicity. Assume that all the $n(s,a)$ realizations of $R(s,a)$ are $(r_{s,a}^{(i)})_{i=1}^{n(s,a)}$. We have that

$$\widehat{\mathsf{VarR}}(s,a) = \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} (r_{s,a}^{(i)})^2 - \left(\frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} r_{s,a}^{(i)} \right)^2.$$

From Lemma 15,

$$\mathbb{P}\left[\left| \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} (r_{s,a}^{(i)})^2 - \mathbb{E}[R(s,a)^2] \right| > \sqrt{\frac{2\mathbb{V}(R(s,a)^2)\iota}{n(s,a)}} + \frac{\iota}{n(s,a)} \right] \leqslant \delta,$$
$$\mathbb{P}\left[\left| \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} r_{s,a}^{(i)} - \mathbb{E}[R(s,a)] \right| > \sqrt{\frac{2\mathbb{V}(R(s,a))\iota}{n(s,a)}} + \frac{\iota}{n(s,a)} \right] \leqslant \delta.$$

By Lemma 19, $\mathbb{V}(R(s,a)^2) \leqslant 4\mathbb{V}(R(s,a))$. So

$$\begin{aligned}
&\left| \widehat{\mathsf{VarR}}(s,a) - \mathbb{V}(R(s,a)) \right| \\
&\leqslant \left| \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} (r_{s,a}^{(i)})^2 - \mathbb{E}[R(s,a)^2] \right| + \left| \left(\frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} r_{s,a}^{(i)} \right)^2 - \mathbb{E}[R(s,a)]^2 \right| \\
&\leqslant 2\sqrt{\frac{2\mathbb{V}(R(s,a))\iota}{n(s,a)}} + \frac{\iota}{n(s,a)} + \left| \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} r_{s,a}^{(i)} + \mathbb{E}[R(s,a)] \right| \left| \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} r_{s,a}^{(i)} - \mathbb{E}[R(s,a)] \right| \\
&\leqslant 2\sqrt{\frac{2\mathbb{V}(R(s,a))\iota}{n(s,a)}} + \frac{\iota}{n(s,a)} + 2\sqrt{\frac{2\mathbb{V}(R(s,a))\iota}{n(s,a)}} + \frac{2\iota}{n(s,a)} \\
&= 4\sqrt{\frac{2\mathbb{V}(R(s,a))\iota}{n(s,a)}} + \frac{3\iota}{n(s,a)}.
\end{aligned}$$

Using $2\sqrt{xy} \leqslant x + y$, we have

$$\widehat{\mathsf{VarR}}(s,a) \leqslant \mathbb{V}(R(s,a)) + 4\sqrt{\frac{2\mathbb{V}(R(s,a))\iota}{n(s,a)}} + \frac{3\iota}{n(s,a)} \leqslant 2\mathbb{V}(R(s,a)) + \frac{11\iota}{n(s,a)}.$$

This completes the proof. □

Lemma 25. *Conditioned on the successful events of Lemmas 21 and 24, we have that for any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$*

$$Q_h^k(s,a) - r(s,a) - P_{s,a} V_{h+1}^k \leqslant \beta_h^k(s,a),$$

where

$$\beta_h^k(s,a) = O\left(\sqrt{\frac{\mathbb{V}(P_{s,a}, V_{h+1}^k)\iota}{n^k(s,a)}} + \sqrt{\frac{\mathbb{V}(P_{s,a}, V_{h+1}^\star)\iota}{n^k(s,a)}} + \sqrt{\frac{\Gamma\mathbb{V}(P_{s,a}, V_{h+1}^k - V_{h+1}^\star)\iota}{n^k(s,a)}} + \sqrt{\frac{\mathbb{V}(R(s,a))\iota}{n^k(s,a)}} + \frac{\Gamma\iota}{n^k(s,a)} \right).$$

Proof of Lemma 25. For any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,

$$\begin{aligned}
&(\widehat{P}_{s,a}^k - P_{s,a})(V_{h+1}^k - V_{h+1}^\star) \\
&= \sum_{s' \in \mathcal{S}} (\widehat{P}_{s,a,s'}^k - P_{s,a,s'})[V_{h+1}^k(s') - V_{h+1}^\star(s') - P_{s,a}(V_{h+1}^k - V_{h+1}^\star)]
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leqslant} \sum_{s'\in\mathcal{S}} \sqrt{\frac{2P_{s,a,s'}\iota}{n^k(s,a)}} \left|V^k_{h+1}(s') - V^\star_{h+1}(s') - P_{s,a}(V^k_{h+1} - V^\star_{h+1})\right| + \sum_{s'\in\mathcal{S}} \frac{\mathbb{1}[P_{s,a,s'}>0]\iota}{n^k(s,a)} \\
&\leqslant \sqrt{\frac{2\iota}{n^k(s,a)}} \sum_{s'\in\mathcal{S}} \sqrt{\mathbb{1}[P_{s,a,s'}>0]P_{s,a,s'}[V^k_{h+1}(s') - V^\star_{h+1}(s') - P_{s,a}(V^k_{h+1} - V^\star_{h+1})]^2} + \frac{\Gamma\iota}{n^k(s,a)} \\
&\overset{\text{(ii)}}{\leqslant} \sqrt{\frac{2\iota}{n^k(s,a)}} \sqrt{\sum_{s'\in\mathcal{S}} \mathbb{1}[P_{s,a,s'}>0]} \sqrt{\sum_{s'\in\mathcal{S}} P_{s,a,s'}[V^k_{h+1}(s') - V^\star_{h+1}(s') - P_{s,a}(V^k_{h+1} - V^\star_{h+1})]^2} + \frac{\Gamma\iota}{n^k(s,a)} \\
&= \sqrt{\frac{2\Gamma\mathbb{V}(P_{s,a}, V^k_{h+1} - V^\star_{h+1})}{n^k(s,a)}} + \frac{\Gamma\iota}{n^k(s,a)},
\end{aligned}
$$

where (i) is by $\mathcal{E}_3$ (Equation (3)); (ii) is by Cauchy-Schwarz inequality. While retaining most other steps in Appendix C.1 of Zhang et al. (2021a) which require $\mathcal{E}_4$ (Equation (4)), we have

$$
\beta^k_h(s,a) = O\left(\sqrt{\frac{\mathbb{V}(\widehat{P}^k_{s,a}, V^k_{h+1})\iota}{n^k(s,a)}} + \sqrt{\frac{\mathbb{V}(P_{s,a}, V^\star_{h+1})\iota}{n^k(s,a)}} + \sqrt{\frac{\Gamma\mathbb{V}(P_{s,a}, V^k_{h+1} - V^\star_{h+1})\iota}{n^k(s,a)}} + \sqrt{\frac{\widehat{\mathsf{VarR}}^k(s,a)\iota}{n^k(s,a)}} + \frac{\Gamma\iota}{n^k(s,a)}\right).
$$

Similar as the steps above Equation (36) in Zhang et al. (2021a) which require $\mathcal{E}_3$ (Equation (3)), we have that

$$
\mathbb{V}(\widehat{P}^k_{s,a}, V^k_{h+1}) \leqslant O\left(\mathbb{V}(P_{s,a}, V^k_{h+1}) + \frac{\Gamma\iota}{n^k(s,a)}\right).
$$

Combined with Lemma 24 we have the desired result. □

Lemma 26. *Conditioned on the successful events of Lemma 25, with probability at least $1 - 5K\iota\delta$, we have that*

$$
M_4 \leqslant O\left(\sum_{k=1}^K \mathsf{Var}^{\pi^k} + M_2 + SA\iota^2\right) \leqslant O(\mathsf{Var}^\star K + M_2 + SA\iota^2).
$$

Proof of Lemma 26. Define

$$
\mathsf{bc}^k_h(s,a) := V^k_h(s) - P_{s,a}V^k_{h+1} - r(s,a) \in [-1,1], \tag{5}
$$

which stands for bonus-correction. By Lemma 25, $\mathsf{bc}^k_h(s,a) \leqslant \beta^k_h(s,a)$. However, we make the distinction here to be more precise. Let $\mathsf{BC}^k_h(s) := \mathsf{bc}^k_h(s,a) + P_{s,a}\mathsf{BC}^k_{h+1}$ with $a = \pi^k_h(s)$ and boundary condition $\mathsf{BC}^k_{H+1}(s) := 0$. We can prove by induction that

$$
\mathsf{BC}^k_h(s) = (V^k_h - V^{\pi^k}_h)(s) \in [0,1]. \tag{6}
$$

First,

$$
\mathsf{BC}^k_H(s) = \mathsf{bc}^k_H(s,a) = V^k_H(s) - r(s,a) = V^k_H(s) - V^{\pi^k}_H(s).
$$

Then assume that $\mathsf{BC}^k_{h+1} = V^k_{h+1} - V^{\pi^k}_{h+1}$, we have

$$
\begin{aligned}
\mathsf{BC}^k_h(s) &= \mathsf{bc}^k_h(s,a) + P_{s,a}(V^k_{h+1} - V^{\pi^k}_{h+1}) \\
&= V^k_h(s) - P_{s,a}V^k_{h+1} - r(s,a) + P_{s,a}(V^k_{h+1} - V^{\pi^k}_{h+1}) \\
&= V^k_h(s) - (r(s,a) + P_{s,a}V^{\pi^k}_{h+1}) \\
&= V^k_h(s) - V^{\pi^k}_h(s).
\end{aligned}
$$

Define a series of random variables and their truncated values: for any $k \in [K]$,

$$
W^k := \sum_{h=1}^H (\mathbb{V}(R(s^k_h, a^k_h)) + \mathbb{V}(P_{s^k_h, a^k_h}, V^{\pi^k}_{h+1})), \quad \overline{W^k} := \min\{W^k, 50\iota\}.
$$

Correspondingly, define the following event, which means there is no truncation:

$$\mathcal{E}_W := \{W^k = \overline{W^k},\ \forall k \in [K]\}.$$

We now calculate the probability of no truncation happens. For any fixed $1 \leqslant k \leqslant K$,

$$\begin{aligned}
W^k &\overset{\text{(i)}}{\leqslant} \sum_{h=1}^{H} [P_{s_h^k,a_h^k}(V_{h+1}^{\pi^k})^2 - (V_{h+1}^{\pi^k}(s_{h+1}^k))^2] + \sum_{h=1}^{H} [(V_h^{\pi^k}(s_h^k))^2 - (P_{s_h^k,a_h^k} V_{h+1}^{\pi^k})^2] + \sum_{h=1}^{H} r(s_h^k, a_h^k) - (V_1^{\pi^k}(s_1^k))^2 \\
&\overset{\text{(ii)}}{\leqslant} 2\sqrt{2\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k}, (V_{h+1}^{\pi^k})^2)\iota} + 6\iota + 2\sum_{h=1}^{H} (V_h^{\pi^k}(s_h^k) - P_{s_h^k,a_h^k} V_{h+1}^{\pi^k}) + \sum_{h=1}^{H} r(s_h^k, a_h^k) \\
&\overset{\text{(iii)}}{\leqslant} 4\sqrt{2\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k}, V_{h+1}^{\pi^k})\iota} + 3\sum_{h=1}^{H} r(s_h^k, a_h^k) + 6\iota \\
&\leqslant 4\sqrt{2W^k\iota} + 3 + 6\iota,
\end{aligned}$$

where (i) is by Lemma 20, $\mathbb{V}(R(s,a)) \leqslant \mathbb{E}[R(s,a)]$; (ii) is by Lemma 17 with $c = \epsilon = 1$, which happens with probability at least $1 - 2\iota\delta$, and $a^2 - b^2 \leqslant (a+b)\max\{a-b, 0\}$ when $a, b \geqslant 0$; (iii) is by Lemma 19 with $C = 1$. Solving the inequality of W^k, we have that

$$W^k \leqslant 50\iota.$$

This means $\mathbb{P}[\mathcal{E}_W] \geqslant 1 - 2K\iota\delta$.

From now on, we suppose $\mathcal{E}_W$ holds. We are ready to bound M_4:

$$\begin{aligned}
M_4 &= \sum_{k=1}^{K}\sum_{h=1}^{H} (\mathbb{V}(R(s_h^k, a_h^k)) + \mathbb{V}(P_{s_h^k,a_h^k}, V_{h+1}^k))I(k, h+1) \\
&\overset{\text{(i)}}{\leqslant} 2\sum_{k=1}^{K}\sum_{h=1}^{H} (\mathbb{V}(R(s_h^k, a_h^k)) + \mathbb{V}(P_{s_h^k,a_h^k}, V_{h+1}^{\pi^k}))I(k, h+1) + 2\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k}, \mathsf{BC}_{h+1}^k)I(k, h+1)}_{=:Z} \qquad (7) \\
&\leqslant 2\sum_{k=1}^{K} W^k + 2Z \\
&\overset{\text{(ii)}}{=} 2\sum_{k=1}^{K} \overline{W^k} + 2Z \\
&\overset{\text{(iii)}}{\leqslant} 6\sum_{k=1}^{K} \mathbb{E}[\overline{W^k} \mid \mathcal{F}_k] + 2Z + 100\iota \\
&\leqslant 6\sum_{k=1}^{K} \mathbb{E}[W^k \mid \mathcal{F}_k] + 2Z + 100\iota^2 \\
&= 6\sum_{k=1}^{K} \mathsf{Var}_1^{\pi^k}(s_1^k) + 2Z + 100\iota^2 \\
&\leqslant 6\sum_{k=1}^{K} \mathsf{Var}^{\pi^k} + 2Z + 100\iota^2 \\
&\leqslant 6\mathsf{Var}^\star K + 2Z + 100\iota^2,
\end{aligned}$$

where (i) is by Equation (6) and $\mathbb{V}(X+Y) \leqslant 2\mathbb{V}(X) + 2\mathbb{V}(Y)$; (ii) is by $\mathcal{E}_W$; (iii) is by Lemma 18 with $l = 50\iota$, which happens with probability at least $1 - \delta$.

It remains to bound the quantity Z we encountered:

$$
\begin{aligned}
Z &= \sum_{k=1}^{K}\sum_{h=1}^{H}[P_{s_h^k,a_h^k}(\mathsf{BC}_{h+1}^k)^2 - (\mathsf{BC}_{h+1}^k(s_{h+1}^k))^2]I(k,h+1) \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}[(\mathsf{BC}_h^k(s_h^k))^2 I(k,h) - (P_{s_h^k,a_h^k}\mathsf{BC}_{h+1}^k)^2 I(k,h+1)] - \sum_{k=1}^{K}(\mathsf{BC}_1^k(s_1^k))^2 \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}[P_{s_h^k,a_h^k}(\mathsf{BC}_{h+1}^k)^2 - (\mathsf{BC}_{h+1}^k(s_{h+1}^k))^2]I(k,h+1) \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}[(\mathsf{BC}_h^k(s_h^k))^2 - (P_{s_h^k,a_h^k}\mathsf{BC}_{h+1}^k)^2]I(k,h+1) + \left|\mathcal{K}^C\right| \\
&\overset{\text{(i)}}{\leqslant} 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},(\mathsf{BC}_{h+1}^k)^2)I(k,h+1)\iota} + 6\iota \\
&\quad + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{\mathsf{BC}_h^k(s_h^k) - P_{s_h^k,a_h^k}\mathsf{BC}_{h+1}^k, 0\}I(k,h+1) + \left|\mathcal{K}^C\right| \\
&\overset{\text{(ii)}}{\leqslant} 4\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},\mathsf{BC}_{h+1}^k)I(k,h+1)\iota} + 6\iota + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{\mathsf{bc}_h^k(s_h^k,a_h^k),0\}I(k,h+1) + \left|\mathcal{K}^C\right| \\
&\leqslant 4\sqrt{2Z\iota} + 6\iota + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\check{\beta}_h^k(s_h^k,a_h^k) + 2\left|\mathcal{K}^C\right| \\
&\leqslant 4\sqrt{2Z\iota} + 2M_2 + 8SA\iota,
\end{aligned}
$$

where (i) is by Lemma 17 with $c = \epsilon = 1$, which happens with probability at least $1 - 2\iota\delta$; (ii) is by Lemma 19 with $C = 1$. Solving the inequality of Z, we have that

$$
Z \leqslant 4M_2 + 48SA\iota. \tag{8}
$$

So plugging back into the bound of M_4 gives the final result. □

Lemma 27. *Conditioned on the successful events of Lemma 26, with probability at least $1 - 2\iota\delta$, we have that*

$$
M_3 \leqslant O(\sqrt{M_4\iota} + \sqrt{M_2\iota} + SA\iota).
$$

Proof of Lemma 27.

$$
\begin{aligned}
M_3 &= \sum_{k=1}^{K}\left(\sum_{h=1}^{H} r(s_h^k,a_h^k)I(k,h) - V_1^{\pi^k}(s_1^k)I(k,1)\right) \\
&= \sum_{k=1}^{K}\left(\sum_{h=1}^{H}(V_h^{\pi^k}(s_h^k) - P_{s_h^k,a_h^k}V_{h+1}^{\pi^k})I(k,h) - V_1^{\pi^k}(s_1^k)I(k,1)\right) \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}(V_{h+1}^{\pi^k}(s_{h+1}^k) - P_{s_h^k,a_h^k}V_{h+1}^{\pi^k})I(k,h+1) + \left|\mathcal{K}^C\right| \\
&\overset{\text{(i)}}{\leqslant} 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},V_{h+1}^{\pi^k})I(k,h+1)\iota} + 6\iota + SA\iota \\
&\overset{\text{(ii)}}{\leqslant} 4\sqrt{M_4\iota} + 4\sqrt{Z\iota} + 7SA\iota \\
&\overset{\text{(iii)}}{\leqslant} 4\sqrt{M_4\iota} + 8\sqrt{M_2\iota} + 35SA\iota,
\end{aligned}
$$

where (i) is by Lemma 17 with $c = \epsilon = 1$, which happens with probability at least $1 - 2\iota\delta$; (ii) is by Equation (6), $\mathbb{V}(X+Y) \leqslant 2\mathbb{V}(X) + 2\mathbb{V}(Y)$ and definition of Z (Equation (7)); (iii) is by Equation (8). □

Lemma 28. *Conditioned on the successful events of Lemma 25, with probability at least $1 - 2\iota\delta$, we have that*

$$M_5 \leqslant O(M_2 + SA\iota).$$

Proof of Lemma 28. Define $\widetilde{V}_h^k = V_h^k - V_h^\star$.

$$\begin{aligned}
M_5 &= \sum_{k=1}^{K}\sum_{h=1}^{H}[P_{s_h^k,a_h^k}(\widetilde{V}_{h+1}^k)^2 - (\widetilde{V}_{h+1}^k(s_{h+1}^k))^2]I(k,h+1) \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}[(\widetilde{V}_h^k(s_h^k))^2 I(k,h) - (P_{s_h^k,a_h^k}\widetilde{V}_{h+1}^k)^2 I(k,h+1)] - \sum_{k=1}^{K}(\widetilde{V}_1^k(s_1^k))^2 \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}[P_{s_h^k,a_h^k}(\widetilde{V}_{h+1}^k)^2 - (\widetilde{V}_{h+1}^k(s_{h+1}^k))^2]I(k,h+1) \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}[(\widetilde{V}_h^k(s_h^k))^2 - (P_{s_h^k,a_h^k}\widetilde{V}_{h+1}^k)^2]I(k,h+1) + \left|\mathcal{K}^C\right| \\
&\overset{\text{(i)}}{\leqslant} 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},(\widetilde{V}_{h+1}^k)^2)I(k,h+1)\iota + 6\iota} \\
&\quad + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{\widetilde{V}_h^k(s_h^k) - P_{s_h^k,a_h^k}\widetilde{V}_{h+1}^k, 0\}I(k,h+1) + \left|\mathcal{K}^C\right| \\
&\overset{\text{(ii)}}{\leqslant} 4\sqrt{2M_5\iota} + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\breve{\beta}_h^k(s_h^k,a_h^k) + 2\left|\mathcal{K}^C\right| \\
&\leqslant 4\sqrt{2M_5\iota} + 2M_2 + 8SA\iota,
\end{aligned}$$

where (i) is by Lemma 17 with $c = \epsilon = 1$, which happens with probability at least $1 - 2\iota\delta$; (ii) is by Lemma 19 with $C = 1$ and the following argument: by Lemma 25,

$$\widetilde{V}_h^k(s_h^k) - P_{s_h^k,a_h^k}\widetilde{V}_{h+1}^k \leqslant \widetilde{Q}_h^k(s_h^k,a_h^k) - P_{s_h^k,a_h^k}\widetilde{V}_{h+1}^k \leqslant \beta_h^k(s_h^k,a_h^k).$$

Solving the inequality of M_5, we have that

$$M_5 \leqslant 4M_2 + 48SA\iota.$$

This completes the proof. □

Lemma 29. *With probability at least $1 - 4K\iota\delta$, we have that for any $k \in [K]$,*

$$\mathsf{Var}_{(k)}^{\Sigma} \leqslant O(\iota).$$

As a result,

$$M_6 \leqslant \mathsf{Var}_K^{\Sigma} \leqslant O(K\iota).$$

Proof of Lemma 29. For any $k \in [K]$,

$$\mathsf{Var}_{(k)}^{\Sigma} \leqslant \sum_{h=1}^{H}[P_{s_h^k,a_h^k}(V_{h+1}^\star)^2 - (V_{h+1}^\star(s_{h+1}^k))^2] + \sum_{h=1}^{H}[(V_h^\star(s_h^k))^2 - (P_{s_h^k,a_h^k}V_{h+1}^\star)^2] + \sum_{h=1}^{H} r(s_h^k,a_h^k) - (V_1^\star(s_1^k))^2$$

$$\overset{\text{(i)}}{\leqslant} 2\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},(V_{h+1}^\star)^2)\iota}+6\iota+2\sum_{h=1}^{H}\max\{\underbrace{V_h^\star(s_h^k)-P_{s_h^k,a_h^k}V_{h+1}^\star}_{\geqslant Q_h^\star(s_h^k,a_h^k)-P_{s_h^k,a_h^k}V_{h+1}^\star\geqslant 0},0\}+1$$

$$\overset{\text{(ii)}}{\leqslant} 4\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},V_{h+1}^\star)\iota}+7\iota+2\sum_{h=1}^{H}(V_{h+1}^\star(s_{h+1}^k)-P_{s_h^k,a_h^k}V_{h+1}^\star)+\underbrace{2V_1^\star(s_1^k)}_{\leqslant 2}$$

$$\overset{\text{(iii)}}{\leqslant} 4\sqrt{2\mathsf{Var}_{(k)}^{\Sigma}\iota}+9\iota+4\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k},V_{h+1}^\star)\iota}+12\iota$$

$$\leqslant 8\sqrt{2\mathsf{Var}_{(k)}^{\Sigma}\iota}+21\iota,$$

where (i) is by Lemma 17 with $c=\epsilon=1$, which happens with probability at least $1-2\iota\delta$; (ii) is by Lemma 19 with $C=1$; (iii) is by Lemma 17 with $c=\epsilon=1$, which happens with probability at least $1-2\iota\delta$. Solving the inequality of $\mathsf{Var}_{(k)}^{\Sigma}$, we have that

$$\mathsf{Var}_{(k)}^{\Sigma}\leqslant 170\iota.$$

So taking a union bound over k we have the desired result. □

B.3. Model-Free Algorithm: `UCB-Advantage-V` (Algorithm 2)

Summary of notations. Let s_h^k, a_h^k and r_h^k denote the state, action and reward at the h-th step of the k-th episode. Let $V_h^k(s), Q_h^k(s,a), V_h^{\mathsf{ref},k}, N_h^k(s,a)$ and $\check{N}_h^k(s,a)$ denote $V_h(s), Q_h(s,a), N_h(s,a)$ and $\check{N}_h(s,a)$ at the beginning of the k-th episode. Let $V_h^{\mathsf{REF}}:=V_h^{\mathsf{ref},K+1}$ denote the final reference value function. Let $N_h^k(s):=\sum_{a\in\mathcal{A}}N_h^k(s,a)$. $N_h^{K+1}(s,a)$ denotes the total number of visits of (s,a,h) after all K episodes are done.

Define $e_1=H$ and $e_{i+1}=\lfloor(1+1/H)e_i\rfloor$. The definition of stages is with respect to the triple (s,a,h). For any fixed pair of k and h, we say that (k,h) falls in the j-th stage of (s,a,h) if and only if $(s,a)=(s_h^k,a_h^k)$ and the total visit number of (s_h^k,a_h^k) after the k-th episode is in $(\sum_{i=1}^{j-1}e_i,\sum_{i=1}^{j}e_i]$.

Let $\check{\upsilon}_h^k$, $\check{\mu}_h^k$, $\check{\sigma}_h^k$, $\mu_h^{\mathsf{ref},k}$, $\sigma_h^{\mathsf{ref},k}$, $\widehat{\mathsf{VarR}}_h^k$, $\bar{b}_h^k$, $\nu_h^{\mathsf{ref},k}$, $\check{\nu}_h^k$ and b_h^k denote $\check{\upsilon}$, $\check{\mu}$, $\check{\sigma}$, μ^{ref}, σ^{ref}, $\widehat{\mathsf{VarR}}$, $\bar{b}$, ν^{ref}, $\check{\nu}$ and b calculated for the value of $Q_h^k(s_h^k,a_h^k)$.

For each k and h, let n_h^k be the total number of visits to (s_h^k,a_h^k,h) prior to the current stage with respect to the same triple and let n_h^k be the number of visits to the same triple during the stage immediately before the current stage. Let $l_{h,i}^k$ and $\check{l}_{h,i}^k$ denote the index of the i-th episode among the n_h^k and $\check{n}_h^k$ episodes defined above, respectively. When h and k are clear from the context, we use l_i and $\check{l}_i$ for short.

Proof of Theorem 10. We first run `UCB-Advantage-V` (Algorithm 2) with $\iota=99(\ln(HSAK/\delta)+1)$ which is large enough for all the probabilistic inequalities to hold. This choice will make the success probability be $1-\mathsf{poly}(S,A,H,K,\iota)\delta$. The lemmas are also proved assuming this choice of ι at first.

Define

$$\psi_{h+1}^k:=\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}(V_{h+1}^{\mathsf{ref},l_i}-V_{h+1}^{\mathsf{REF}}),$$

$$\xi_{h+1}^k:=\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}[P_{s_h^k,a_h^k,h}(V_{h+1}^{\mathsf{ref},\check{l}_i}-V_{h+1}^\star)-(V_{h+1}^{\mathsf{ref},\check{l}_i}(s_{h+1}^{\check{l}_i})-V_{h+1}^\star(s_{h+1}^{\check{l}_i}))],$$

$$\phi_{h+1}^k:=P_{s_h^k,a_h^k,h}(V_{h+1}^\star-V_{h+1}^{\pi^k})-(V_{h+1}^\star(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)).$$

Combining Section 4.2 in Zhang et al. (2020b) with Lemmas 30 to 33 and 36 to 41, we have that with probability at least

$1-35HSAK\iota\delta$,

$$\begin{aligned}\mathsf{Regret}(K) &\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}(\psi_{h+1}^k+\xi_{h+1}^k+\phi_{h+1}^k+2b_h^k)\\ &\leqslant O(\sqrt{\mathsf{Var}_K^{\Sigma}HSA\iota}+\sqrt{H^5SAK\iota^2/2^{2i^\star}}+H^5S^2A2^{i^\star}\iota^2).\end{aligned}$$

Taking $i^\star=\lceil 1/2\cdot\log_2(K/H^5S^3A\iota^2)\rceil$, we have:

$$\mathsf{Regret}(K)\leqslant O(\sqrt{\mathsf{Var}_K^{\Sigma}HSA\iota}+\sqrt[4]{H^{15}S^5A^3K\iota^6}).$$

Now we apply Lemma 42. With probability at least $1-46HSAK\iota\delta$,

$$\mathsf{Regret}(K)\leqslant O\left(\sqrt{HSAK\iota\sum_{k=1}^{K}\mathsf{Var}^{\pi^k}}+\sqrt[4]{H^{15}S^5A^3K\iota^6}\right)\leqslant O(\sqrt{\mathsf{Var}^\star HSAK\iota}+\sqrt[4]{H^{15}S^5A^3K\iota^6}).$$

The final result is established by scaling δ to make the success probability be $1-\delta'$. We upper-bound ι by $100(HSAK/\sqrt{\delta}+1)$ and solve the inequality:

$$4600HSAK\left(\frac{HSAK}{\sqrt{\delta}}+1\right)\delta\leqslant\delta'.$$

Take $\delta=(\delta'/7000H^2S^2A^2K^2)^2$. By $\ln(HSAK/(\delta/7000H^2S^2A^2K^2)^2)\leqslant O(\iota)$, we conclude the proof. □

Lemma 30. *With probability at least $1-15HSAK\iota\delta$, we have that for any $(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$,*

$$Q_h^\star(s,a)\leqslant Q_h^{k+1}(s,a)\leqslant Q_h^k(s,a).$$

Proof of Lemma 30. Recall that the update rule is:

$$Q_h(s_h^k,a_h^k)\leftarrow\min\left\{\underbrace{\widehat{r}_h(s_h^k,a_h^k)+\frac{\check{v}}{\check{n}}+\bar{b}}_{①},\ \underbrace{\widehat{r}_h(s_h^k,a_h^k)+\frac{\mu^{\mathsf{ref}}}{n}+\frac{\check{\mu}}{\check{n}}+b}_{②},\ Q_h(s_h^k,a_h^k)\right\}. \tag{9}$$

We prove by induction on k. Clearly for $k=1$ the argument is true.

For case ① in Equation (9), we have that (omit the subscripts of h and superscripts of k for simplicity)

$$\begin{aligned}Q_h^{k+1}(s,a)&=r_h(s,a)+\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}V_{h+1}^{l_i}(s_{h+1}^{l_i})+(\widehat{r}_h(s,a)-r_h(s,a))+\bar{b}\\ &\overset{\text{(i)}}{\geqslant} r_h(s,a)+\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}V_{h+1}^{\star}(s_{h+1}^{l_i})+(\widehat{r}_h(s,a)-r_h(s,a))+\bar{b}\\ &\overset{\text{(ii)}}{\geqslant} r_h(s,a)+P_{s,a,h}V_{h+1}^\star-\sqrt{\frac{H^2\iota}{2\check{n}}}+(\widehat{r}_h(s,a)-r_h(s,a))+\bar{b}\\ &\overset{\text{(iii)}}{\geqslant} r_h(s,a)+P_{s,a,h}V_{h+1}^\star-\sqrt{\frac{H^2\iota}{2\check{n}}}-\sqrt{\frac{\iota}{2n}}+\bar{b}\\ &\geqslant Q_{h+1}^\star(s,a),\end{aligned}$$

where (i) is by induction $V^u\geqslant V^\star$ for any $1\leqslant u\leqslant k$; (ii) is by Lemma 14 with $b=H$, which holds with probability at least $1-\delta$; (iii) is by Lemma 14 with $b=1$, which holds with probability at least $1-\delta$.

Define

$$\chi_1:=\frac{1}{n}\sum_{i=1}^{n}(V_{h+1}^{\mathsf{ref},l_i}(s_{h+1}^{l_i})-P_{s,a,h}V_{h+1}^{\mathsf{ref},l_i}),$$

$$\chi_2 := \frac{1}{\check{n}} \sum_{i=1}^{n} [(V_{h+1}^{l_i} - V_{h+1}^{\mathsf{ref},l_i})(s_{h+1}^{l_i}) - P_{s,a,h}(V_{h+1}^{l_i} - V_{h+1}^{\mathsf{ref},l_i})].$$

For case ② in Equation (9), we have that

$$\begin{aligned}
Q_h^{k+1}(s,a) &= \widehat{r}_h(s,a) + P_{s,a,h}\left(\frac{1}{n}\sum_{i=1}^{n} V_{h+1}^{\mathsf{ref},l_i}\right) + P_{s,a,h}\left(\frac{1}{\check{n}}\sum_{i=1}^{\check{n}} (V_{h+1}^{\check{l}_i} - V_{h+1}^{\mathsf{ref},\check{l}_i})\right) + \chi_1 + \chi_2 + b \\
&\overset{\text{(i)}}{\geqslant} r_h(s,a) + P_{s,a,h}\left(\frac{1}{\check{n}}\sum_{i=1}^{\check{n}} V_{h+1}^{\check{l}_i}\right) + \chi_1 + \chi_2 + (r_h(s,a) - \widehat{r}_h(s,a)) + b \\
&\overset{\text{(ii)}}{\geqslant} r_h(s,a) + P_{s,a,h} V_{h+1}^{\star} + \chi_1 + \chi_2 + (r_h(s,a) - \widehat{r}_h(s,a)) + b \\
&= Q_h^{\star}(s,a) + \chi_1 + \chi_2 + (r_h(s,a) - \widehat{r}_h(s,a)) + b,
\end{aligned}$$

where (i) is by that $V_{h+1}^{\mathsf{ref},u}$ is non-increasing in u; (ii) is by the induction $V^u \geqslant V^\star$ for any $1 \leqslant u \leqslant k$.

From Lemma 17 with $c = H, \epsilon = c^2$, we have that with probability at least $1 - 2\iota\delta$,

$$|\chi_1| \leqslant \frac{1}{n}\left(2\sqrt{2\underbrace{\sum_{i=1}^{n} \mathbb{V}(P_{s,a,h}, V_{h+1}^{\mathsf{ref},l_i})}_{=:X}\,\iota} + 6H\iota\right). \tag{10}$$

Define

$$\begin{aligned}
\chi_3 &:= \sum_{i=1}^{n} [P_{s,a,h}(V_{h+1}^{\mathsf{ref},l_i})^2 - (V_{h+1}^{\mathsf{ref},l_i}(s_{h+1}^{l_i}))^2], \\
\chi_4 &:= \frac{1}{n}\left(\sum_{i=1}^{n} V_{h+1}^{\mathsf{ref},l_i}(s_{h+1}^{l_i})\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} P_{s,a,h} V_{h+1}^{\mathsf{ref},l_i}\right)^2, \\
\chi_5 &:= \frac{1}{n}\left(\sum_{i=1}^{n} P_{s,a,h} V_{h+1}^{\mathsf{ref},l_i}\right)^2 - \sum_{i=1}^{n} (P_{s,a,h} V_{h+1}^{\mathsf{ref},l_i})^2,
\end{aligned}$$

then it is easy to verify that

$$X = n\nu^{\mathsf{ref}} + \chi_3 + \chi_4 + \chi_5. \tag{11}$$

By Lemma 17 with $c = H^2, \epsilon = c^2$, and Lemma 19 with $C = H$, we have that with probability at least $1 - 2\iota\delta$,

$$\chi_3 \leqslant 2\sqrt{2\sum_{i=1}^{n} \mathbb{V}(P_{s,a,h}, (V_{h+1}^{\mathsf{ref},l_i})^2)\iota} + 6H^2\iota \leqslant 4H\sqrt{2X\iota} + 6H^2\iota. \tag{12}$$

By Lemma 17 with $c = H, \epsilon = c^2$, we have that with probability at least $1 - 2\iota\delta$,

$$\chi_4 \leqslant \frac{1}{n}\left|\sum_{i=1}^{n} (V_{h+1}^{\mathsf{ref},l_i}(s_{h+1}^{l_i}) + P_{s,a,h} V_{h+1}^{\mathsf{ref},l_i})\right| \left|\sum_{i=1}^{n} (V_{h+1}^{\mathsf{ref},l_i}(s_{h+1}^{l_i}) - P_{s,a,h} V_{h+1}^{\mathsf{ref},l_i})\right| \leqslant 2H(2\sqrt{2X\iota} + 6H\iota). \tag{13}$$

By Cauchy-Schwarz inequality, $\chi_5 \leqslant 0$. Thus, $X \leqslant n\nu^{\mathsf{ref}} + 18H^2\iota + 8H\sqrt{2X\iota}$. Solving the inequality,

$$X \leqslant 2n\nu^{\mathsf{ref}} + 164H^2\iota.$$

Plugging back into Equation (10), we have

$$\chi_1 \leqslant 4\sqrt{\frac{\nu^{\mathsf{ref}}\iota}{n}} + \frac{(4\sqrt{82} + 6)H\iota}{n}.$$

By a similar reasoning, we have that with probability at least $1-6\iota\delta$,

$$\chi_2 \leqslant 4\sqrt{\frac{\breve{\nu}\iota}{\breve{n}}} + \frac{(4\sqrt{82}+6)H\iota}{\breve{n}}.$$

By Lemma 16 and $\frac{1}{x-1} \leqslant \frac{2}{x}$, we have that with probability at least $1-\delta$,

$$|r_h(s,a) - \widehat{r}_h(s,a)| \leqslant 2\sqrt{\frac{\widehat{\mathsf{VarR}}_h(s,a)\iota}{n}} + \frac{14\iota}{3n}. \tag{14}$$

Therefore, we have $b \geqslant |\chi_1| + |\chi_2| + |r_h(s,a) - \widehat{r}_h(s,a)|$, which means $Q_h^{k+1}(s,a) \geqslant Q_h^\star(s,a)$. □

Lemma 31 (Adapted from Lemma 5 and Corollary 6 in Zhang et al. (2020b), and Corollary 6 in Chen et al. (2021))**.** *Conditioned on the successful events of Lemma 30, with probability at least $1-HK\delta$, we have that for any $\epsilon \in (0,H]$ and any $h \in [H]$,*

$$\sum_{k=1}^{K} \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant \epsilon] \leqslant 60000\frac{H^5SA\iota}{\epsilon^2} =: N_0(\epsilon).$$

As a result, for every state s we have that

$$N_h^k(s) \geqslant N_0(\epsilon) \implies 0 \leqslant V_h^k(s) - V_h^\star(s) \leqslant \epsilon.$$

Proof of Lemma 31. To derive the constant 60000, we only need to solve the inequality:

$$\sum_{k=1}^{K} \mathbb{1}[\delta_h^k \geqslant \epsilon] \leqslant \frac{\sum_{k=1}^{K} \mathbb{1}[\delta_h^k \geqslant \epsilon]\delta_h^k}{\epsilon} \leqslant \frac{240H^{5/2}\sqrt{\|w\|_\infty SA\iota \sum_{k=1}^{K} \mathbb{1}[\delta_h^k \geqslant \epsilon]} + 3SAH^3\|w\|_\infty}{\epsilon}$$

which is below Equation (48) in Zhang et al. (2020b), using $x \leqslant a\sqrt{x} + b \implies x \leqslant a^2 + 2b$.

The second part can be proven in a similar way as Corollary 6 in Chen et al. (2021). □

Lemma 32. *Conditioned on the successful events of Lemma 31, we have that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(V_h^k(s_h^k) - V_h^\star(s_h^k)) \leqslant O(\sqrt{H^7SAK\iota}),$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}(V_h^k(s_h^k) - V_h^\star(s_h^k))^2 \leqslant O(H^6SA\iota^2).$$

Proof of Lemma 32. Let c be a fixed constant, then

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}(V_h^k(s_h^k) - V_h^\star(s_h^k)) &= \sum_{k=1}^{K}\sum_{h=1}^{H}(V_h^k(s_h^k) - V_h^\star(s_h^k))\mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) < c] \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}(V_h^k(s_h^k) - V_h^\star(s_h^k))\mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant c] \\
&\overset{\text{(i)}}{\leqslant} cHK + \sum_{k=1}^{K}\sum_{h=1}^{H}\int_0^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x]\mathbb{1}[(V_h^k(s_h^k) - V_h^\star(s_h^k)) \geqslant c]\,\mathrm{d}x \\
&= cHK + \sum_{k=1}^{K}\sum_{h=1}^{H}\int_c^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x]\,\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= cHK + \int_c^H \left(\sum_{k=1}^K \sum_{h=1}^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x] \right) \mathrm{d}x \\
&\overset{\text{(ii)}}{\leqslant} cHK + \int_c^H O\left(\frac{H^6 SA\iota}{x^2} \right) \mathrm{d}x \\
&\leqslant O\left(cHK + \frac{H^6 SA\iota}{c} \right),
\end{aligned}
$$

where (i) is by $n = \int_0^\infty \mathbb{1}[n \geqslant x]\,\mathrm{d}x$ for any non-negative real n and $V_h^\star \leqslant V_h^k \leqslant H$ (Lemma 30); (ii) is by Lemma 31. Taking $c = \sqrt{H^5 SA\iota/K}$ gives the first result.

Similarly,

$$
\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H (V_h^k(s_h^k) - V_h^\star(s_h^k))^2 \\
&= \sum_{k=1}^K \sum_{h=1}^H (V_h^k(s_h^k) - V_h^\star(s_h^k))^2 \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) < c] \\
&\quad + \sum_{k=1}^K \sum_{h=1}^H (V_h^k(s_h^k) - V_h^\star(s_h^k))^2 \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant c] \\
&\leqslant c^2 HK + \sum_{k=1}^K \sum_{h=1}^H \left(\int_0^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x] \mathbb{1}[(V_h^k(s_h^k) - V_h^\star(s_h^k)) \geqslant c]\,\mathrm{d}x \right)^2 \\
&= c^2 HK + \sum_{k=1}^K \sum_{h=1}^H \left(\int_c^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x]\,\mathrm{d}x \right)^2 \\
&= c^2 HK + \int_c^H \left(\int_c^H \left(\sum_{k=1}^K \sum_{h=1}^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x] \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant y] \right) \mathrm{d}x \right) \mathrm{d}y \\
&= c^2 HK + \int_c^H \left(\int_c^y \left(\sum_{k=1}^K \sum_{h=1}^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant y] \right) \mathrm{d}x + \int_y^H \left(\sum_{k=1}^K \sum_{h=1}^H \mathbb{1}[V_h^k(s_h^k) - V_h^\star(s_h^k) \geqslant x] \right) \mathrm{d}x \right) \mathrm{d}y \\
&\leqslant c^2 HK + \int_c^H \left((y-c) O\left(\frac{H^6 SA\iota}{y^2} \right) + \int_y^H O\left(\frac{H^6 SA\iota}{x^2} \right) \mathrm{d}x \right) \mathrm{d}y \\
&\leqslant c^2 HK + O\left(H^6 SA\iota \int_c^H \frac{\mathrm{d}y}{y} \right) \\
&= O\left(c^2 HK + H^6 SA\iota \ln \frac{H}{c} \right),
\end{aligned}
$$

and taking $c = 1/\sqrt{HK}$ gives the second result. □

Lemma 33. *Define $\beta_i := H/2^i$ for $i \in \{0, 1, \ldots, i^\star\}$, $N_0^0 := 0$ and $N_0^i := N_0(\beta_i) = 60000 \cdot 2^{2i} SAH^3\iota$ (defined in Lemma 31) for $i \in [i^\star]$. Define*

$$
B_h^{\mathsf{ref},k}(s) := \sum_{i=1}^{i^\star} \beta_{i-1} \mathbb{1}[N_0^{i-1} \leqslant N_h^k(s) < N_0^i].
$$

Conditioned on the successful events of Lemma 31, we have that

$$
\begin{aligned}
V_h^{\mathsf{ref},k}(s) - V_h^{\mathsf{REF}}(s) &\leqslant B_h^{\mathsf{ref},k}(s), \\
V_h^{\mathsf{ref},k}(s) - V_h^\star(s) &\leqslant B_h^{\mathsf{ref},k}(s) + \beta_{i^\star},
\end{aligned}
$$

and

$$\sum_{k=1}^{K}\sum_{h=1}^{H} B_h^{\mathsf{ref},k}(s_h^k) \leqslant O(H^5S^2A2^{i^\star}\iota),$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H} (B_h^{\mathsf{ref},k}(s_h^k))^2 \leqslant O(H^6S^2Ai^\star\iota).$$

Proof of Lemma 33. For $i \leqslant i^\star - 1$, by Lemma 31, if $N_h^k(s) \geqslant N_0^i = N_0(\beta_i)$ then $V_h^k(s) - V_h^\star(s) \leqslant \beta_i$. Let k_0 be the minimum k such that $N_h^k(s) \geqslant N_0^i$. By the updating rule in Algorithm 2 and non-increasing property of V^k (Lemma 30), it must satisfy that $V_h^k(s) \leqslant V_h^{\mathsf{ref},k}(s) \leqslant V_h^{k_0}(s)$. Since $V_h^\star(s) \leqslant V_h^{\mathsf{REF}}(s)$ (Lemma 30), we have that

$$V_h^{\mathsf{ref},k}(s) - V_h^{\mathsf{REF}}(s) \leqslant V_h^{k_0}(s) - V_h^\star(s) \leqslant \beta_i.$$

If $N_h^k(s) \geqslant N_0^{i^\star} = N_0(\beta_{i^\star})$ then $V_h^{\mathsf{ref},k}(s) = V_h^{\mathsf{REF}}(s)$ and $V_h^{\mathsf{ref},k}(s) - V_h^\star(s) \leqslant \beta_{i^\star}$, which corresponds to $B_h^{\mathsf{ref},k}(s) = 0$. Since the indicator functions are disjoint, we have the first part of results.

The remaining result is proven by:

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} B_h^{\mathsf{ref},k}(s_h^k) &= \sum_{s\in\mathcal{S}}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{i^\star} \frac{H}{2^{i-1}}\mathbb{1}[s = s_h^k, N_0^{i-1} \leqslant N_h^k(s) < N_0^i] \\
&\leqslant \sum_{s\in\mathcal{S}}\sum_{h=1}^{H}\sum_{i=1}^{i^\star} \frac{H}{2^{i-1}} N_0^i \\
&\leqslant O\left(SH\sum_{i=1}^{i^\star}\frac{H}{2^i}\cdot 2^{2i}SAH^3\iota\right) \\
&\leqslant O(H^5S^2A2^{i^\star}\iota); \\
\sum_{k=1}^{K}\sum_{h=1}^{H} (B_h^{\mathsf{ref},k}(s_h^k))^2 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{i^\star} \beta_{i-1}^2\mathbb{1}[N_0^{i-1} \leqslant N_h^k(s) < N_0^i] \\
&\leqslant O\left(SH\sum_{i=1}^{i^\star}\frac{H^2}{2^{2i}}\cdot 2^{2i}SAH^3\iota\right) \\
&\leqslant O(H^6S^2Ai^\star\iota).
\end{aligned}$$

This completes the proof. □

Lemma 34 (Lemma 11 in Zhang et al. (2020b))**.** *For any non-negative weights $(w_h(s,a))_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]}$ and $\alpha \in (0,1)$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k)}{(n_h^k)^\alpha} \leqslant \frac{2^\alpha}{1-\alpha}\sum_{s,a,h} w_h(s,a)(N_h^{K+1}(s,a))^{1-\alpha},$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k)}{(\check{n}_h^k)^\alpha} \leqslant \frac{2^{2\alpha}H^\alpha}{1-\alpha}\sum_{s,a,h} w_h(s,a)(N_h^{K+1}(s,a))^{1-\alpha}.$$

In the case $\alpha = 1$, it holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k)}{n_h^k} \leqslant 2\sum_{s,a,h} w_h(s,a)\ln N_h^{K+1}(s,a),$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k)}{\check{n}_h^k} \leqslant 4H\sum_{s,a,h} w_h(s,a)\ln N_h^{K+1}(s,a).$$

Lemma 35. *For any non-negative sequence $(X_h^k)_{k\in[K],h\in[H]}$, we have that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}X_h^{l_{h,i}^k}\leqslant 2\iota\sum_{k=1}^{K}\sum_{h=1}^{H}X_h^k,$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}X_h^{\check{l}_{h,i}^k}\leqslant\left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{h=1}^{H}X_h^k.$$

Proof of Lemma 35. Refer to Equation (58) in Zhang et al. (2020b) for the first inequality. Refer to Equation (15) and the paragraph below it in Zhang et al. (2020b) for the second inequality. □

Lemma 36. *Conditioned on the successful events of Lemma 33, with probability at least $1-\delta$, we have that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\psi_{h+1}^k\leqslant O(H^5S^2A2^{i^\star}\iota).$$

Proof of Lemma 36. Since ψ_h^k is non-negative and $(1+1/H)^{h-1}\leqslant 3$ when $h\leqslant H$, we have that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\psi_{h+1}^k&\leqslant O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\psi_{h+1}^k\right)\\&=O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}(V_{h+1}^{\mathsf{ref},l_{h,i}^k}-V_{h+1}^{\mathsf{REF}})\right)\\&\overset{\text{(i)}}{\leqslant}O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}B_{h+1}^{\mathsf{ref},l_{h,i}^k}\right)\\&\overset{\text{(ii)}}{\leqslant}O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}P_{s_h^k,a_h^k,h}B_{h+1}^{\mathsf{ref},k}\right)\\&\overset{\text{(iii)}}{\leqslant}O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}B_h^{\mathsf{ref},k}(s_h^k)+H\iota\right)\\&\overset{\text{(iv)}}{\leqslant}O(H^5S^2A2^{i^\star}\iota),\end{aligned}$$

where (i) is by Lemma 33; (ii) is by Lemma 35; (iii) is by Lemma 18 with $l=H$, which holds with probability at least $1-\delta$; (iv) is by Lemma 33. □

Lemma 37. *Conditioned on the successful events of Lemma 32, with probability at least $1-5HSA\iota\delta$, we have that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\xi_{h+1}^k\leqslant O(H^{7/2}SA\iota^{3/2}).$$

Proof of Lemma 37. We borrow the beginning part of proof of Lemma 15 in Zhang et al. (2020b), and perform more fine-grained analyses on the remaining part. Let x_h^k be the number of elements in current stage with respect to (s_h^k,a_h^k,h). Define

$$\theta_{h+1}^j:=\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbb{1}[\check{l}_{h,i}^k=j],\quad\widetilde{\theta}_{h+1}^j:=\left(1+\frac{1}{H}\right)^{h-1}\frac{\left\lfloor(1+1/H)x_h^j\right\rfloor}{x_h^j}\leqslant 3,$$

and

$$\mathcal{K}:=\{(k,h)\mid\theta_{h+1}^k=\widetilde{\theta}_{h+1}^k\},\quad\mathcal{K}_h^\perp(s,a):=\{k\mid(s_h^k,a_h^k)=(s,a),k\text{ is in the second last stage of }(s,a,h)\}.$$

Let $\theta_{h+1}(s,a)$ and $\widetilde{\theta}_{h+1}(s,a)$ denote θ^k_{h+1} and $\widetilde{\theta}^k_{h+1}$ respectively for some $k \in \mathcal{K}^{\perp}_h(s,a)$. By Equation (61) in Zhang et al. (2020b),

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\xi^k_{h+1} \leqslant \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\widetilde{\theta}^k_{h+1}[P_{s^k_h,a^k_h,h}(V^k_{h+1}-V^\star_{h+1})-(V^k_{h+1}(s^k_{h+1})-V^\star_{h+1}(s^k_{h+1}))]}_{=:①}$$
$$+\underbrace{\sum_{(k,h)\in\overline{\mathcal{K}}}(\theta^k_{h+1}-\widetilde{\theta}^k_{h+1})[P_{s^k_h,a^k_h,h}(V^k_{h+1}-V^\star_{h+1})-(V^k_{h+1}(s^k_{h+1})-V^\star_{h+1}(s^k_{h+1}))]}_{=:②}.$$

We now bound both terms:

$$① \overset{\text{(i)}}{\leqslant} O\left(\sqrt{\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h},(V^k_{h+1}-V^\star_{h+1}))}_{=:Z}\,\iota}+H\iota\right),$$

$$② \overset{\text{(ii)}}{=} \sum_{s,a,h}(\theta_{h+1}(s,a)-\widetilde{\theta}_{h+1}(s,a))\sum_{k\in\mathcal{K}^{\perp}_h(s,a)}[P_{s^k_h,a^k_h,h}(V^k_{h+1}-V^\star_{h+1})-(V^k_{h+1}(s^k_{h+1})-V^\star_{h+1}(s^k_{h+1}))]$$
$$\leqslant \sum_{s,a,h}\left|\theta_{h+1}(s,a)-\widetilde{\theta}_{h+1}(s,a)\right|\left|\sum_{k\in\mathcal{K}^{\perp}_h(s,a)}[P_{s^k_h,a^k_h,h}(V^k_{h+1}-V^\star_{h+1})-(V^k_{h+1}(s^k_{h+1})-V^\star_{h+1}(s^k_{h+1}))]\right|$$
$$\overset{\text{(iii)}}{\leqslant} O\left(\sum_{s,a,h}\left(\sqrt{\sum_{k\in\mathcal{K}^{\perp}_h(s,a)}\mathbb{V}(P_{s^k_h,a^k_h,h},V^k_{h+1}-V^\star_{h+1})\iota}+H\iota\right)\right)$$
$$\overset{\text{(iv)}}{\leqslant} O\left(\sqrt{HSA\iota\sum_{s,a,h}\sum_{k\in\mathcal{K}^{\perp}_h(s,a)}\mathbb{V}(P_{s^k_h,a^k_h,h},V^k_{h+1}-V^\star_{h+1})}+H^2SA\iota\right)$$
$$\overset{\text{(v)}}{\leqslant} O(\sqrt{HSAZ\iota}+H^2SA\iota),$$

where (i) is by Lemma 17 with $c=3H, \epsilon=c^2$, which holds with probability at least $1-2\iota\delta$; (ii) is by the step above Equation (63) in Zhang et al. (2020b); (iii) is by $\left|\theta_{h+1}(s,a)-\widetilde{\theta}_{h+1}(s,a)\right| \leqslant 3$ and Lemma 17 with $c=H, \epsilon=c^2$, which holds with probability at least $1-2HSA\iota\delta$; (iv) is by Cauchy-Schwarz inequality; (v) is by the following argument: for any non-negative sequence $(X^k_h)_{k\in[K],h\in[H]}$,

$$\begin{aligned}\sum_{s,a,h}\sum_{k\in\mathcal{K}^{\perp}_h(s,a)}X^k_h &= \sum_{k=1}^{K}\sum_{h=1}^{H}X^k_h\sum_{s,a,h'}\sum_{k'\in\mathcal{K}^{\perp}_{h'}(s,a)}\mathbb{1}[(k',h')=(k,h)]\\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}X^k_h\sum_{s,a}\mathbb{1}[k\in\mathcal{K}^{\perp}_h(s,a)]\\ &\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}X^k_h\sum_{s,a}\mathbb{1}[(s^k_h,a^k_h)=(s,a)]\\ &\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}X^k_h.\end{aligned}$$

It remains to bound Z:

$$Z \leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}P_{s^k_h,a^k_h,h}(V^k_{h+1}-V^\star_{h+1})^2$$

$$\overset{\text{(i)}}{\leqslant} O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}(V_{h+1}^{k}(s_{h+1}^{k})-V_{h+1}^{\star}(s_{h+1}^{k}))^2+H^2\iota\right)$$
$$\overset{\text{(ii)}}{\leqslant} O(H^6SA\iota^2),$$

where (i) is by Lemma 18 with $l=H^2$, which holds with probability at least $1-\delta$; (ii) is by Lemma 32. □

Lemma 38. *With probability at least $1-6\iota\delta$, we have that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\phi_{h+1}^{k}\leqslant O(H\iota).$$

Proof of Lemma 38. Since $(1+1/H)^{h-1}\leqslant 3$ when $h\leqslant H$, we have that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\phi_{h+1}^{k}=\sum_{k=1}^{K}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}[P_{s_h^k,a_h^k,h}(V_{h+1}^{\star}-V_{h+1}^{\pi^k})-(V_{h+1}^{\star}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k))]$$
$$\overset{\text{(i)}}{\leqslant} O\left(\sqrt{\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\star}-V_{h+1}^{\pi^k})\,\iota}_{=:Y}}+H\iota\right),$$

where (i) is by Lemma 17 with $c=3H,\epsilon=c^2$, which happens with probability at least $1-2\iota\delta$. Next we bound Y:

$$\begin{aligned}
Y&=\sum_{k=1}^{K}\sum_{h=1}^{H}[P_{s_h^k,a_h^k,h}(V_{h+1}^{\star}-V_{h+1}^{\pi^k})^2-(V_{h+1}^{\star}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k))^2]\\
&\quad+\sum_{k=1}^{K}\sum_{h=1}^{H}\{(V_h^{\star}(s_h^k)-V_h^{\pi^k}(s_h^k))^2-[P_{s_h^k,a_h^k,h}(V_{h+1}^{\star}-V_{h+1}^{\pi^k})]^2\}-(V_1^{\star}(s_1^k)-V_1^{\pi^k}(s_1^k))^2\\
&\overset{\text{(i)}}{\leqslant} 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},(V_{h+1}^{\star}-V_{h+1}^{\pi^k})^2)\iota}+6H^2\iota\\
&\quad+2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{\underbrace{V_h^{\star}(s_h^k)-V_h^{\pi^k}(s_h^k)-P_{s_h^k,a_h^k,h}(V_{h+1}^{\star}-V_{h+1}^{\pi^k})}_{\geqslant Q_h^{\star}(s_h^k,a_h^k)-r_h(s_h^k,a_h^k)-P_{s_h^k,a_h^k,h}V_{h+1}^{\star}=0},0\}\\
&\overset{\text{(ii)}}{\leqslant} 4H\sqrt{2Y\iota}+6H^2\iota+2H\sum_{k=1}^{K}\sum_{h=1}^{H}[V_{h+1}^{\star}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)-P_{s_h^k,a_h^k,h}(V_{h+1}^{\star}-V_{h+1}^{\pi^k})]\\
&\quad+\underbrace{2H(V_1^{\star}(s_1^k)-V_1^{\pi^k}(s_1^k))}_{\leqslant 2H^2}\\
&\overset{\text{(iii)}}{\leqslant} 4H\sqrt{2Y\iota}+8H^2\iota+4H\sqrt{2Y\iota}+12H^2\iota\\
&\leqslant 8H\sqrt{2Y\iota}+20H^2\iota,
\end{aligned}$$

where (i) is by Lemma 17 with $c=H^2,\epsilon=c^2$, which happens with probability at least $1-2\iota\delta$, and $a^2-b^2\leqslant(a+b)\max\{a-b,0\}$ when $a,b\geqslant 0$; (ii) is by Lemma 19 with $C=H$; (iii) is by by Lemma 17 with $c=H,\epsilon=c^2$, which happens with probability at least $1-2\iota\delta$. Solving the inequality of Y, we have that

$$Y\leqslant 168H^2\iota.$$

Plugging Y back gives the desired result. □

Lemma 39. *Conditioned on the successful events of Lemma 33, with probability at least* $1-4\iota\delta$*, we have that for any* $(k,h)\in[K]\times[H]$

$$\nu_h^{\mathsf{ref},k} \leqslant O\left(\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^\star)+\frac{H^2\iota+\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}(B_{h+1}^{\mathsf{ref},l_i})^2}{n_h^k}+\beta_{i^\star}^2\right).$$

Proof of Lemma 39. Let (k,h) be fixed. We prove by first bounding $\nu_n^{\mathsf{ref},k}-\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\mathsf{ref},l_i})$. By Equation (11),

$$\nu_n^{\mathsf{ref},k}-\frac{1}{n_h^k}\underbrace{\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\mathsf{ref},l_i})}_{=X\text{(abusing notation)}}=-\frac{\chi_3+\chi_4+\chi_5}{n_h^k}.$$

Since we can use Equations (12) and (13), we only need to bound $-\chi_5$.

$$\begin{aligned}
-\chi_5 &= \sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{ref},l_i})^2-\frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{ref},l_i}\right)^2\\
&\overset{\text{(i)}}{\leqslant} \sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{ref},l_i})^2-\frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{REF}}\right)^2\\
&= \sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{ref},l_i}+P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{REF}})(P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{ref},l_i}-P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{REF}})\\
&\leqslant 2H\sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{ref},l_i}-P_{s_h^k,a_h^k,h}V_{h+1}^{\mathsf{REF}})\\
&\overset{\text{(ii)}}{\leqslant} 2H\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}B_{h+1}^{\mathsf{ref},l_i},
\end{aligned}$$

where (i) is by $V_{h+1}^{\mathsf{ref},l_i}\geqslant V_{h+1}^{\mathsf{REF}}$ (Lemma 30); (ii) is by Lemma 33. Combining bounds of χ_3 (Equation (12)) and χ_4 (Equation (13)), we have:

$$\nu_n^{\mathsf{ref},k}-\frac{X}{n_h^k}\leqslant\frac{8H\sqrt{2X\iota}+18H^2\iota+2H\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}B_{h+1}^{\mathsf{ref},l_i}}{n_h^k}.$$

Since $8H\sqrt{2X\iota}\leqslant X+32H^2\iota$, we have:

$$\nu_n^{\mathsf{ref},k}-\frac{2X}{n_h^k}\leqslant O\left(\frac{H^2\iota+H\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}B_{h+1}^{\mathsf{ref},l_i}}{n_h^k}\right).$$

For the desired result, we finally bound:

$$\begin{aligned}
\frac{X}{n_h^k}-2\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^\star) &= \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}(\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\mathsf{ref},l_i})-2\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^\star))\\
&\overset{\text{(i)}}{\leqslant} \frac{2}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\mathsf{ref},l_i}-V_{h+1}^\star)\\
&\leqslant \frac{2}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}(V_{h+1}^{\mathsf{ref},l_i}-V_{h+1}^\star)^2
\end{aligned}$$

$$\overset{\text{(ii)}}{\leqslant} \frac{2}{n_h^k} \sum_{i=1}^{n_h^k} P_{s_h^k, a_h^k, h}(B_{h+1}^{\mathsf{ref}, l_i} + \beta_{i^\star})^2$$
$$\leqslant \frac{4}{n_h^k} \sum_{i=1}^{n_h^k} P_{s_h^k, a_h^k, h}(B_{h+1}^{\mathsf{ref}, l_i})^2 + 4\beta_{i^\star}^2,$$

where (i) is by $\mathbb{V}(X+Y) \leqslant 2\mathbb{V}(X) + 2\mathbb{V}(Y)$; (ii) is by Lemma 33. So by $HP_{s_h^k, a_h^k, h} B_{h+1}^{\mathsf{ref}, l_i} \leqslant O(H^2 + P_{s_h^k, a_h^k, h}(B_{h+1}^{\mathsf{ref}, l_i})^2)$ we have the result. □

Lemma 40 (Analogous to Lemma 24)**.** *With probability at least $1 - 2HSAK\iota\delta$, we have that for any $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,*

$$\widehat{\mathsf{VarR}}_h^k(s, a) \leqslant O\left(\mathbb{V}(R_h(s, a)) + \frac{\iota}{n_h^k(s, a)}\right).$$

Lemma 41. *Conditioned on the successful events of Lemmas 39 and 40, with probability at least $1 - \delta$, we have that*

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^{h-1} b_h^k \leqslant O(\sqrt{\mathsf{Var}_K^\Sigma HSA\iota} + \sqrt{H^5 SAK\iota^2 / 2^{2i^\star}} + H^4 S^{3/2} A i^{\star 1/2} \iota^2).$$

Proof of Lemma 41. Since b_h^k is non-negative and $(1 + 1/H)^{h-1} \leqslant 3$ when $h \leqslant H$, we have that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^{h-1} b_h^k \leqslant O\left(\sum_{k=1}^{K} \sum_{h=1}^{H} \left(\sqrt{\frac{\nu_h^{\mathsf{ref}, k} \iota}{n_h^k}} + \sqrt{\frac{\breve{\nu}_h^k \iota}{\breve{n}_h^k}} + \sqrt{\frac{\widehat{\mathsf{VarR}}_h^k \iota}{n_h^k}} + \frac{H\iota}{\breve{n}_h^k}\right)\right)$$
$$\leqslant O\left(\sum_{k=1}^{K} \sum_{h=1}^{H} \left(\sqrt{\frac{\nu_h^{\mathsf{ref}, k} \iota}{n_h^k}} + \sqrt{\frac{\breve{\nu}_h^k \iota}{\breve{n}_h^k}} + \sqrt{\frac{\mathbb{V}(R_h(s_h^k, a_h^k))\iota}{n_h^k}} + \frac{H\iota}{\breve{n}_h^k}\right)\right),$$

where the last step is by Lemma 40. Using Lemma 34, we have that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H\iota}{\breve{n}_h^k} \leqslant O(H^3 SA\iota^2).$$

Next, we bound the terms of ν^{ref} and $\breve{\nu}$ separately.

Plugging in Lemma 39, we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left(\sqrt{\frac{\nu_h^{\mathsf{ref}, k} \iota}{n_h^k}} + \sqrt{\frac{\mathbb{V}(R_h(s_h^k, a_h^k))\iota}{n_h^k}}\right)$$
$$\leqslant O\left(\sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\frac{(\mathbb{V}(R_h(s_h^k, a_h^k)) + \mathbb{V}(P_{s_h^k, a_h^k, h}, V_{h+1}^\star))\iota}{n_h^k}} + \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H\iota}{n_h^k}\right.$$
$$\left. + \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{n_h^k} \sqrt{\sum_{i=1}^{n_h^k} P_{s_h^k, a_h^k, h}(B_{h+1}^{\mathsf{ref}, l_{h,i}^k})^2 \iota} + \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\frac{\beta_{i^\star}^2 \iota}{n_h^k}}\right)$$
$$\overset{\text{(i)}}{\leqslant} O\left(\sum_{s,a,h} \sqrt{N_h^{K+1}(s, a)(\mathbb{V}(R_h(s_h^k, a_h^k)) + \mathbb{V}(P_{s,a,h}, V_{h+1}^\star))\iota} + H^2 SA\iota^2\right.$$

$$+\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\iota}{n_h^k}}\sqrt{\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}(B_{h+1}^{\mathsf{ref},l_{h,i}^k})^2}+\sqrt{\beta_{i^\star}^2\iota}\sum_{s,a,h}\sqrt{N_h^{K+1}(s,a)}\Bigg)$$

$$\overset{\text{(ii)}}{\leqslant} O\Bigg(\sqrt{HSA\iota\sum_{s,a,h}N_h^{K+1}(s,a)(\mathbb{V}(R_h(s,a))+\mathbb{V}(P_{s,a,h},V_{h+1}^\star))}+H^2SA\iota^2+$$

$$+\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\iota}{n_h^k}}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,h}(B_{h+1}^{\mathsf{ref},l_{h,i}^k})^2}+\sqrt{HSA\beta_{i^\star}^2\iota\sum_{s,a,h}N_h^{K+1}(s,a)}\Bigg)$$

$$\overset{\text{(iii)}}{\leqslant} O\Bigg(\sqrt{HSA\iota\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbb{V}(R_h(s_h^k,a_h^k))+\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^\star))}_{=\mathsf{Var}_K^\Sigma}}$$

$$+H^2SA\iota^2+\sqrt{H^2SA\beta_{i^\star}^2K\iota}+\sqrt{HSA\iota^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}P_{s_h^k,a_h^k,h}(B_{h+1}^{\mathsf{ref},k})^2}\Bigg)$$

$$\overset{\text{(iv)}}{\leqslant} O\left(\sqrt{\mathsf{Var}_K^\Sigma HSA\iota}+H^2SA\iota^2+\sqrt{H^2SA\beta_{i^\star}^2K\iota}+\sqrt{HSA\iota^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}(B_h^{\mathsf{ref},k}(s_h^k))^2+H\iota}\right)$$

$$\overset{\text{(v)}}{\leqslant} O(\sqrt{\mathsf{Var}_K^\Sigma HSA\iota}+\sqrt{H^4SAK\iota/2^{2i^\star}}+H^{7/2}S^{3/2}Ai^{\star 1/2}\iota^{3/2}).$$

where (i) is by Lemma 34; (ii) is by Cauchy-Schwarz inequality; (iii) is by Lemmas 34 and 35; (iv) is by Lemma 18 with $l=H$, which happens with probability at least $1-\delta$; (v) is by Lemma 33.

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\check{\nu}_h^k &\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(V_{h+1}^{\mathsf{ref},\check{l}_{h,i}^k}(\check{s}_{h+1}^{\check{l}_{h,i}^k})-V_{h+1}^{\check{l}_{h,i}^k}(\check{s}_{h+1}^{\check{l}_{h,i}^k}))^2\\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(V_{h+1}^{\mathsf{ref},\check{l}_{h,i}^k}(\check{s}_{h+1}^{\check{l}_{h,i}^k})-V_{h+1}^\star(\check{s}_{h+1}^{\check{l}_{h,i}^k}))^2\\
&\overset{\text{(i)}}{\leqslant} \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(B_{h+1}^{\mathsf{ref},\check{l}_{h,i}^k}(\check{s}_{h+1}^{\check{l}_{h,i}^k})+\beta_{i^\star})^2\\
&\overset{\text{(ii)}}{\leqslant} O\left(\sum_{k=1}^{K}\sum_{h=1}^{H}(B_h^{\mathsf{ref},k}(s_h^k))^2+\beta_{i^\star}^2HK\right)\\
&\overset{\text{(iii)}}{\leqslant} O(H^6S^2Ai^\star\iota+H^3K/2^{2i^\star}).
\end{aligned}$$

where (i) is by Lemma 33; (ii) is by Lemma 35; (iii) is by Lemma 33. So

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\check{\nu}_h^k\iota}{\check{n}_h^k}}\leqslant\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\iota}{\check{n}_h^k}}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\check{\nu}_h^k}\overset{\text{(i)}}{\leqslant} O(H^4S^{3/2}Ai^{\star 1/2}\iota^{3/2}+\sqrt{H^5SAK\iota^2/2^{2i^\star}}).$$

where (i) is by Lemma 34. □

Lemma 42. *With probability at least $1-11K\iota\delta$, the following results hold: For any $k\in[K]$,*

$$\mathsf{Var}_{(k)}^\Sigma\leqslant O(H^2\iota),$$

hence

$$\mathsf{Var}_K^{\Sigma} \leqslant O(H^2 K\iota).$$

Alternatively,

$$\mathsf{Var}_K^{\Sigma} \leqslant O\left(\sum_{k=1}^{K} \mathsf{Var}^{\pi^k} + H^2\iota^2\right) \leqslant O(\mathsf{Var}^{\star} K + H^2\iota^2).$$

Proof of Lemma 42. We first prove the result depending on Var_K^{Σ} similar to Lemma 29. For any $k \in [K]$,

$$\begin{aligned}
\mathsf{Var}_{(k)}^{\Sigma} &\overset{\text{(i)}}{\leqslant} \sum_{h=1}^{H}[P_{s_h^k,a_h^k,h}(V_{h+1}^{\star})^2 - (V_{h+1}^{\star}(s_{h+1}^k))^2] + \sum_{h=1}^{H}[(V_h^{\star}(s_h^k))^2 - (P_{s_h^k,a_h^k,h}V_{h+1}^{\star})^2] + \sum_{h=1}^{H} r_h(s_h^k,a_h^k) - (V_1^{\star}(s_1^k))^2 \\
&\overset{\text{(ii)}}{\leqslant} 2\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},(V_{h+1}^{\star})^2)\iota} + 6H^2\iota + 2H\sum_{h=1}^{H}\max\{\underbrace{V_h^{\star}(s_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^{\star}}_{\geqslant Q_h^{\star}(s_h^k,a_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^{\star} \geqslant 0}, 0\} + H \\
&\overset{\text{(iii)}}{\leqslant} 4H\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\star})\iota} + 7H^2\iota + 2H\sum_{h=1}^{H}(V_{h+1}^{\star}(s_{h+1}^k) - P_{s_h^k,a_h^k,h}V_{h+1}^{\star}) + \underbrace{2HV_1^{\star}(s_1^k)}_{\leqslant 2H^2} \\
&\overset{\text{(iv)}}{\leqslant} 4H\sqrt{2\mathsf{Var}_{(k)}^{\Sigma}\iota} + 9H^2\iota + 4H\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\star})\iota} + 12H\iota \\
&\leqslant 8H\sqrt{2\mathsf{Var}_{(k)}^{\Sigma}\iota} + 21H^2\iota,
\end{aligned}$$

where (i) is by by Lemma 20, $\mathbb{V}(R_h(s,a)) \leqslant \mathbb{E}[R_h(s,a)]$; (ii) is by Lemma 17 with $c = H^2, \epsilon = c^2$, which happens with probability at least $1 - 2\iota\delta$; (iii) is by Lemma 19 with $C = H$; (iv) is by Lemma 17 with $c = H, \epsilon = c^2$, which happens with probability at least $1 - 2\iota\delta$. Solving the inequality of $\mathsf{Var}_{(k)}^{\Sigma}$, we have that

$$\mathsf{Var}_{(k)}^{\Sigma} \leqslant 170H^2\iota.$$

So taking a union bound over k we have the first result.

Next we prove the result depending on $\mathsf{Var}^{\star}$. This is similar to the proof of Lemma 26. Define a series of random variables and their truncated values: for any $k \in [K]$,

$$W^k := \sum_{h=1}^{H}(\mathbb{V}(R_h(s_h^k,a_h^k)) + \mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})), \quad \overline{W^k} := \min\{W^k, 50H^2\iota\}.$$

By $\mathbb{V}(P_{s,a,h},V_{h+1}^{\star}) \leqslant 2\mathbb{V}(P_{s,a,h},V_{h+1}^{\pi^k}) + 2\mathbb{V}(P_{s,a,h},V_{h+1}^{\star} - V_{h+1}^{\pi^k})$, we know that

$$\mathsf{Var}_K^{\Sigma} \leqslant 2\sum_{k=1}^{K} W^k + 2Y,$$

where $Y \leqslant O(H^2\iota)$ (with probability at least $1 - 4\iota\delta$) is defined in the proof of Lemma 38. Correspondingly, define the following event, which means there is no truncation:

$$\mathcal{E}_W := \{W^k = \overline{W^k},\ \forall k \in [K]\}.$$

For any fixed $1 \leqslant k \leqslant K$,

$$W^k \leqslant \sum_{h=1}^{H}[P_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})^2 - (V_{h+1}^{\pi^k}(s_{h+1}^k))^2] + \sum_{h=1}^{H}[(V_h^{\pi^k}(s_h^k))^2 - (P_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k})^2] + \sum_{h=1}^{H} r_h(s_h^k,a_h^k) - (V_1^{\pi^k}(s_1^k))^2$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leqslant} 2\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},(V_{h+1}^{\pi^k})^2)\iota} + 6H^2\iota + 2H\sum_{h=1}^{H}(V_h^{\pi^k}(s_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k}) + H \\
&\overset{\text{(ii)}}{\leqslant} 4H\sqrt{2\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})\iota} + 7H^2\iota + 2H\sum_{h=1}^{H} r_h(s_h^k,a_h^k) \\
&\leqslant 4H\sqrt{2W^k\iota} + 9H^2\iota,
\end{aligned}
$$

where (i) is by Lemma 17 with $c = H^2, \epsilon = c^2$, which happens with probability at least $1 - 2\iota\delta$; (ii) is by Lemma 19 with $C = H$. Solving the inequality, $W^k \leqslant 50H^2\iota$. This means, $\mathbb{P}[\mathcal{E}_W] \geqslant 1 - 2K\iota\delta$. Now on suppose $\mathcal{E}_W$ holds, then

$$
\begin{aligned}
\sum_{k=1}^{K} W^k &= \sum_{k=1}^{K}\overline{W^k} \\
&\overset{\text{(i)}}{\leqslant} 3\sum_{k=1}^{K}\mathbb{E}[\overline{W^k} \mid \mathcal{F}_k] + 50H^2\iota^2 \\
&\leqslant 3\sum_{k=1}^{K}\mathbb{E}[W^k \mid \mathcal{F}_k] + 50H^2\iota^2 \\
&= 3\sum_{k=1}^{K}\mathsf{Var}_1^{\pi^k}(s_1^k) + 50H^2\iota^2 \\
&\leqslant 3\sum_{k=1}^{K}\mathsf{Var}^{\pi^k} + 50H^2\iota^2 \\
&\leqslant 3\mathsf{Var}^{\star}K + 50H^2\iota^2,
\end{aligned}
$$

where (i) is by Lemma 18 with $l = 50H^2\iota$, which happens with probability at least $1 - \delta$. □

B.4. Proof of Lower Bounds

We modify Theorem 9 in Domingues et al. (2021) for a bounded-reward, time-homogeneous lower bound (Theorem 12). Theorem 13 is much more straightforward. To this end, we borrow necessary notations from Domingues et al. (2021), adapted to time-homogeneous setting.

A policy π interacting with an MDP $\mathcal{M}$ defines a stochastic process denote by $((S_h^k, A_h^k, R_h^k)_{h\in[H]})_{k\geqslant 1}$, where S_h^k, A_h^k and R_h^k are the random variables representing the state, the action and the reward at time h of episode k. As explained by Lattimore & Szepesvári (2020), the Ionescu-Tulcea theorem ensures the existence of probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\mathcal{M}})$ such that

$$
\mathbb{P}_{\mathcal{M}}[S_{h+1}^k = s | A_h^k, I_h^k] = P(s|S_h^k, A_h^k), \quad \text{and} \quad \mathbb{P}_{\mathcal{M}}[A_h^k = a | I_h^k] = \pi_h^k(a|I_h^k),
$$

where $\boldsymbol{\pi} = (\pi_h^k)_{k\in[K],h\in[H]}$ and

$$
I_h^k := (S_1^1, A_1^1, R_1^1, \ldots, S_H^1, A_H^1, R_H^1, S_1^2, A_1^2, R_1^2, \ldots, S_H^{k-1}, A_H^{k-1}, R_H^{k-1}, S_1^k, A_1^k, R_1^k, \ldots, S_h^k)
$$

is the random vector containing all state-action pairs observed up to step h of episode k, but not including A_h^k. Here we assume the rewards are *deterministic* as in Domingues et al. (2021). Next, we denote by $\mathbb{P}_{\mathcal{M}}^{I_H^K}$ the pushforward measure of I_H^K under $\mathbb{P}_{\mathcal{M}}$,

$$
\mathbb{P}_{\mathcal{M}}^{I_H^K}[i_H^K] := \mathbb{P}_{\mathcal{M}}[I_H^K = i_H^K] = \prod_{k=1}^{K}\prod_{h=1}^{H}\pi_h^k(a_h^k|i_h^k)P(s_{h+1}^k|s_t^k, a_t^k), \tag{15}
$$

where i_H^K is a realization of I_H^K.

Definition 43. *The Kullback-Leibler divergence between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on a measurable space $(\Omega, \mathcal{G})$ is defined as*

$$\mathsf{KL}(\mathbb{P}_1, \mathbb{P}_2) := \int_\Omega \ln\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}(\omega)\right) \mathrm{d}\mathbb{P}_1(\omega),$$

if $\mathbb{P}_1 \ll \mathbb{P}_2$ and $+\infty$ otherwise. For Bernoulli distributions, we define $\forall (p,q) \in [0,1]^2$,

$$\mathsf{kl}(p,q) := \mathsf{KL}(\mathcal{B}(p), \mathcal{B}(q)) = p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right).$$

Lemma 44 (Adapted from Lemma 5 in Domingues et al. (2021))**.** *Let $\mathcal{M}$ and $\mathcal{M}'$ be two MDPs that are identical except for their transition probabilities, denoted by P and P', respectively. Assume that we have $\forall (s,a), P(\cdot|s,a) \ll P'(\cdot|s,a)$. Then for any K,*

$$\mathsf{KL}\left(\mathbb{P}_{\mathcal{M}}^{I_H^K}, \mathbb{P}_{\mathcal{M}'}^{I_H^K}\right) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}_{\mathcal{M}}\left[N_{s,a}^K\right] \mathsf{KL}(P(\cdot|s,a), P'(\cdot|s,a)),$$

where $N_{s,a}^K := \sum_{k=1}^K \sum_{h=1}^H \mathbb{1}[(S_h^k, A_h^k) = (s,a)]$.

Lemma 45 (Lemma 1 in Garivier et al. (2016))**.** *Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0,1]$, we have*

$$\mathsf{KL}(\mathbb{P}_1, \mathbb{P}_2) \geqslant \mathsf{kl}(\mathbb{E}_1[Z], \mathbb{E}_2[Z]),$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively.

Proof of Theorem 12. We retain most of the proof of Theorem 9 and Appendix C in Domingues et al. (2021), while incorporating the hard instance design in Section 5.5.1 in Zhou et al. (2022). Namely, we change the A-ary tree in Domingues et al. (2021) with the binary tree in Zhou et al. (2022). This change does not affect the proof, while circumventing the requirement of $S = 3 + (A^d - 1)/(A-1)$ where d is the tree height. We can find $S' = 2 + 2^{\lfloor \log_2(S-2) \rfloor} = \Omega(S)$ and replace S with S'. We still use $d = \lfloor \log_2(S-2) \rfloor$ to denote the tree height.

We change the transition at s_g: $P(s_b|s_g, a) = 1$ for any $a \in \mathcal{A}$. This means for any trajectory, the agent can only get reward once, then loops at s_b.

Another important change in design is to scale the reward at s_g by $t \leqslant 1$, with t depending on $\mathcal{V}$ the variance we desire. So $r(s_g, a) = t$.

To be precise, let $\mathbb{E}_0$ and $\mathbb{E}_{(\ell^\star, a^\star)}$ be the expectation taken with respect to the reference MDP (with no special leaf-action pair) and $\mathcal{M}_{(\ell^\star, a^\star)}$. We have that

$$\mathcal{R}_K(\boldsymbol{\pi}, \mathcal{M}_{(\ell^\star,a^\star)}) \geqslant tK\varepsilon\left(1 - \frac{1}{K}\mathbb{E}_{(\ell^\star,a^\star)}\left[N_{(\ell^\star,a^\star)}^K\right]\right),$$

where $N_{(\ell^\star,a^\star)}^K = \sum_{k=1}^K \mathbb{1}[(S_{d+1}^k, A_{d+1}^k) = (s,a)]$. Hence,

$$\max_{(\ell^\star,a^\star)} \mathcal{R}_K(\boldsymbol{\pi}, \mathcal{M}_{(\ell^\star,a^\star)}) \geqslant tK\varepsilon\left(1 - \frac{1}{LAK}\sum_{(\ell^\star,a^\star)} \mathbb{E}_{(\ell^\star,a^\star)}\left[N_{(\ell^\star,a^\star)}^K\right]\right). \tag{16}$$

Since $N_{(\ell^\star,a^\star)}^K/K \in [0,1]$, by Lemma 45,

$$\mathsf{kl}\left(\frac{1}{K}\mathbb{E}_0\left[N_{(\ell^\star,a^\star)}^K\right], \frac{1}{K}\mathbb{E}_{(\ell^\star,a^\star)}\left[N_{(\ell^\star,a^\star)}^K\right]\right) \leqslant \mathsf{KL}(\mathbb{P}_0^{I_H^K}, \mathbb{P}_{(\ell^\star,a^\star)}^{I_H^K}).$$

By Lemma 44,

$$\mathsf{KL}\left(\mathbb{P}_0^{I_H^K}, \mathbb{P}_{(\ell^\star,a^\star)}^{I_H^K}\right) = \mathbb{E}_0\left[N_{(\ell^\star,a^\star)}^K\right] \mathsf{kl}\left(\frac{1}{2}, \frac{1}{2} + \varepsilon\right).$$

Assume that $\varepsilon \leqslant 1/4$, then $\mathsf{kl}(1/2, 1/2+\varepsilon) \leqslant 4\varepsilon^2$. By Pinsker's inequality, $(p-q)^2 \leqslant \mathsf{kl}(p,q)/2$, it implies

$$\frac{1}{K}\mathbb{E}_{(\ell^\star,a^\star)}\left[N^K_{(\ell^\star,a^\star)}\right] \leqslant \frac{1}{K}\mathbb{E}_0\left[N^K_{(\ell^\star,a^\star)}\right] + \sqrt{2}\varepsilon\sqrt{\mathbb{E}_0\left[N^K_{(\ell^\star,a^\star)}\right]}.$$

Since $\sum_{(h^\star,a^\star)} N^K_{(\ell^\star,a^\star)} = K$, by Cauchy-Schwarz inequality we have

$$\frac{1}{K}\sum_{(\ell^\star,a^\star)} \mathbb{E}_{(\ell^\star,a^\star)}\left[N^K_{(\ell^\star,a^\star)}\right] \leqslant 1 + \sqrt{2}\varepsilon\sqrt{LAK}.$$

Plugging this back to Equation (16), and taking $\varepsilon = (1-1/LA)\sqrt{LA/8K}$, we have

$$\max_{(\ell^\star,a^\star)} \mathcal{R}_K(\boldsymbol{\pi}, \mathcal{M}_{(\ell^\star,a^\star)}) \geqslant \Omega(t\sqrt{SAK}).$$

To ensure that $\varepsilon \leqslant 1/4$, we need $K \geqslant SA$.

Now we calculate the variances. We know that $V^\star_{d+2}(s_b) = 0$ and $V^\star_{d+2}(s_g) = t$. For any trajectory τ, we look at step $h = d+1$. If $(s_h, a_h) \neq (\ell^\star, a^\star)$, then

$$\mathsf{Var}^\Sigma_\tau \geqslant \mathbb{V}((1/2,1/2),(0,t)) = \Omega(t^2).$$

If $(s_h, a_h) = (\ell^\star, a^\star)$, then

$$\mathsf{Var}^\Sigma_\tau \geqslant \mathbb{V}((1/2-\varepsilon,1/2+\varepsilon),(0,t)) = \left(\frac{1}{4}-\varepsilon^2\right)\Omega(t^2).$$

Notice that $\varepsilon \leqslant 1/4$, so $\mathsf{Var}^\Sigma_\tau \geqslant \Omega(t^2)$ for any τ, and

$$\mathsf{Var}^\star \geqslant \mathsf{Var}^{\pi^\star} \geqslant \min_\tau \mathsf{Var}^\Sigma_\tau \geqslant \Omega(t^2).$$

Since the total reward in each episode is upper-bounded by t, we know that $\mathsf{Var}^\Sigma_\tau, \mathsf{Var}^\star \leqslant O(t^2)$. Thus,

$$\mathsf{Var}^\Sigma_\tau, \mathsf{Var}^\star = \Theta(t^2).$$

For the desired result, we set $t = \Theta(\sqrt{\mathcal{V}})$. □

Proof of Theorem 13. We retain most of the proof of Theorem 9 in Domingues et al. (2021), while incorporating the hard instance design in Section 5.5.1 in Zhou et al. (2022). Namely, we change the A-ary tree in Domingues et al. (2021) with the binary tree in Zhou et al. (2022). This change does not affect the proof, while circumventing the requirement of $S = 3 + (A^d - 1)/(A-1)$ where d is the tree height. We can find $S' = 2 + 2^{\lfloor \log_2(S-2) \rfloor} = \Omega(S)$ and replace S with S'. We still use $d = \lfloor \log_2(S-2) \rfloor$ to denote the tree height.

Another important change in design is to scale the reward at s_g by $t \leqslant 1$, with t depending on $\mathcal{V}$ the variance we desire. So $r_h(s_g, a) = t\mathbb{1}[h \geqslant \overline{H} + d + 1]$. This modification does not affect the choice of ε and $\overline{H}$ in Domingues et al. (2021), only scales the optimal value and regret linearly, so we have that

$$\max_{(h^\star,\ell^\star,a^\star)} \mathcal{R}_K(\boldsymbol{\pi}, \mathcal{M}_{(h^\star,\ell^\star,a^\star)}) \geqslant \Omega(t\sqrt{H^3SAK}).$$

Now we calculate the variances. We know that $V^\star_{\overline{H}+d+1}(s_b) = 0$ and $V^\star_{\overline{H}+d+1}(s_g) = t(H - \overline{H} - d) = \Omega(tH)$. For any trajectory τ, we look at step $h = \overline{H} + d$. If $(s_h, a_h) \neq (\ell^\star, a^\star)$, then

$$\mathsf{Var}^\Sigma_\tau \geqslant \mathbb{V}((1/2,1/2),(0,\Omega(tH))) = \Omega(t^2H^2).$$

If $(s_h, a_h) = (\ell^\star, a^\star)$, then

$$\mathsf{Var}^\Sigma_\tau \geqslant \mathbb{V}((1/2-\varepsilon,1/2+\varepsilon),(0,\Omega(tH))) = \left(\frac{1}{4}-\varepsilon^2\right)\Omega(t^2H^2).$$

Notice that $\varepsilon \leqslant 1/4$ in Domingues et al. (2021), so $\mathsf{Var}_\tau^\Sigma \geqslant \Omega(t^2H^2)$ for any τ, and

$$\mathsf{Var}^\star \geqslant \mathsf{Var}^{\pi^\star} \geqslant \min_\tau \mathsf{Var}_\tau^\Sigma \geqslant \Omega(t^2H^2).$$

Since the total reward in each episode is upper-bounded by $O(tH)$, we know that $\mathsf{Var}_\tau^\Sigma, \mathsf{Var}^\star \leqslant O(t^2H^2)$. Thus,

$$\mathsf{Var}_\tau^\Sigma, \mathsf{Var}^\star = \Theta(t^2H^2).$$

For the desired result, we set $t = \Theta(\sqrt{\mathcal{V}}/H)$. □