Settling the Sample Complexity of Online Reinforcement Learning

Zihan Zhang* Princeton
Yuxin Chen† UPenn
Jason D. Lee* Princeton
Simon S. Du‡ Univ. of Washington

July 26, 2023

Abstract

A central issue lying at the heart of online reinforcement learning (RL) is data efficiency. While a number of recent works achieved asymptotically minimal regret in online RL, the optimality of these results is only guaranteed in a "large-sample" regime, imposing enormous burn-in cost in order for their algorithms to operate optimally. How to achieve minimax-optimal regret without incurring any burn-in cost has been an open problem in RL theory.

We settle this problem for the context of finite-horizon inhomogeneous Markov decision processes. Specifically, we prove that a modified version of Monotonic Value Propagation (MVP), a model-based algorithm proposed by Zhang et al. (2021), achieves a regret on the order of (modulo log factors)

$$\min\left\{\sqrt{SAH^3K},\, HK\right\},$$

where S is the number of states, A is the number of actions, H is the planning horizon, and K is the total number of episodes. This regret matches the minimax lower bound for the entire range of sample size $K \geq 1$, essentially eliminating any burn-in requirement. It also translates to a PAC sample complexity (i.e., the number of episodes needed to yield ε-accuracy) of $\frac{SAH^3}{\varepsilon^2}$ up to log factor, which is minimax-optimal for the full ε-range. Further, we extend our theory to unveil the influences of problem-dependent quantities like the optimal value/cost and certain variances. The key technical innovation lies in the development of a new regret decomposition strategy and a novel analysis paradigm to decouple complicated statistical dependency — a long-standing challenge facing the analysis of online RL in the sample-hungry regime.

Keywords: online RL; minimax regret; burn-in cost; optimal sample complexity; model-based algorithms

Contents

*Department of Electrical and Computer Engineering, Princeton University; email: {zz5478,jasonlee}@princeton.edu.
†Department of Statistics and Data Science, University of Pennsylvania; email: yuxinc@wharton.upenn.edu.
‡Paul G. Allen School of Computer Science and Engineering, University of Washington; email: ssdu@cs.washington.edu.

1 Introduction

In reinforcement learning (RL), an agent is often asked to learn optimal decisions (i.e., the ones that maximize cumulative reward) through real-time "trial-and-error" interactions with an unknown environment. This task is commonly dubbed as *online RL*, underscoring the critical role of adaptive online data collection and differentiating it from other RL settings that rely upon pre-collected data. A central challenge in achieving sample-efficient online RL boils down to how to optimally balance exploration and exploitation during data collection, namely, how to trade off the potential revenue of exploring unknown terrain/dynamics against the benefit of exploiting past experience. While decades-long effort has been invested towards unlocking the capability of online RL, how to *fully* characterize — and more importantly, attain — its fundamental performance limit remains largely unsettled.

In this paper, we take an important step towards settling the sample complexity limit of online RL, focusing on tabular Markov Decision Processes (MDPs) with finite horizon and finite state-action space. More concretely, imagine that one seeks to learn a near-optimal policy of a time-inhomogeneous MDP with S states, A actions, and horizon length H, and is allowed to execute the MDP of interest K times and collect K sample episodes each of length H. This canonical problem is among the most extensively studied in the RL literature, with formal theoretical pursuit dating back to more than 25 years ago (e.g., Kearns and Singh (1998b)). Numerous works have since been devoted to improving the sample efficiency and/or refining the analysis framework (Azar et al., 2017; Bai et al., 2019; Brafman and Tennenholtz, 2003; Dann et al., 2017;

Domingues et al., 2021; Jaksch et al., 2010; Jin et al., 2018; Kakade, 2003; Li et al., 2021b; Ménard et al., 2021; Zanette and Brunskill, 2019; Zhang et al., 2021, 2020). As we shall elucidate momentarily, however, information-theoretic optimality has only been achieved in the "large-sample" regime. When it comes to the most challenging sample-hungry regime, there remains a substantial gap between the state-of-the-art regret upper bound and the best-known minimax lower bound, which motivates the research of this paper.

1.1 Inadequacy of prior art: enormous burn-in cost

While past research has obtained asymptotically optimal (i.e., optimal when K approaches infinity) regret bounds in the aforementioned setting, all of these results incur an enormous burn-in cost — that is, the minimum sample size needed for an algorithm to operate sample-optimally — which we explain in the sequel.

Minimax lower bound. To provide a theoretical benchmark, we first make note of the best-known minimax regret lower bound developed by Domingues et al. (2021); Jin et al. (2018):[1]

$$\text{(minimax lower bound)} \qquad \Omega\left(\min\left\{\sqrt{SAH^3K},\, HK\right\}\right), \tag{1}$$

assuming that the immediate reward at each step falls within $[0,1]$ and imposing no restriction on K.

Given that a regret of HK can be trivially achieved (as the sum of rewards in any K episodes cannot exceed HK), we shall often drop the HK term and write

$$\text{(minimax lower bound)} \qquad \Omega\big(\sqrt{SAH^3K}\big) \qquad \text{if } K \geq SAH. \tag{2}$$

Prior upper bounds and burn-in cost. We now turn to the upper bounds developed in prior literature. For ease of presentation, we shall assume

$$K \geq SAH \tag{3}$$

in the rest of this subsection unless otherwise noted. Log factors are also ignored in the discussion below.

The first paper that achieves asymptotically optimal regret is Azar et al. (2017), which came up with a model-based algorithm called UCBVI that enjoys a regret bound $\widetilde{O}\big(\sqrt{SAH^3K} + H^3S^2A\big)$. A close inspection reveals that this regret matches the minimax lower bound (2) if and only if

$$\big(\text{burn-in cost of Azar et al. (2017)}\big) \qquad K \gtrsim S^3AH^3, \tag{4}$$

due to the presence of the lower-order term H^3S^2A in the regret bound. This burn-in cost is clearly undesirable, since the sample size available in many practical scenarios might be far below this requirement.

In light of its fundamental importance in contemporary RL applications (which often have unprecedented dimensionality and relatively limited data volume), reducing the burn-in cost without compromising sample efficiency has emerged as a central problem in recent pursuit of RL theory (Agarwal et al., 2020; Dann et al., 2019; Li et al., 2022a, 2021b,d; Ménard et al., 2021; Sidford et al., 2018b; Zanette and Brunskill, 2019; Zhang et al., 2021; Zhou et al., 2023). The state-of-the-art regret upper bounds for finite-horizon inhomogeneous MDPs can be summarized below (depending on the size of K):

$$\text{(Ménard et al., 2021)} \qquad \widetilde{O}\big(\sqrt{SAH^3K} + SAH^4\big), \tag{5a}$$

$$\text{(Zhang et al., 2021; Zhou et al., 2023)} \qquad \widetilde{O}\big(\sqrt{SAH^3K} + S^2AH^2\big), \tag{5b}$$

[1] Let $\mathcal{X} = \{S, A, H, K, \frac{1}{\delta}\}$, where $1-\delta$ is the target success rate (to be seen shortly). The standard notation $f(\mathcal{X}) = O\big(g(\mathcal{X})\big)$ (or $f(\mathcal{X}) \lesssim g(\mathcal{X})$) indicates the existence of some universal constant $c_1 > 0$ such that $f(\mathcal{X}) \leq c_1 g(\mathcal{X})$; $f(\mathcal{X}) = \Omega\big(g(\mathcal{X})\big)$ (or $f(\mathcal{X}) \gtrsim g(\mathcal{X})$) means that there exists some universal constant $c_2 > 0$ such that $f(\mathcal{X}) \geq c_2 g(\mathcal{X})$; and $f(\mathcal{X}) = \Theta\big(g(\mathcal{X})\big)$ (or $f(\mathcal{X}) \asymp g(\mathcal{X})$) means that $f(\mathcal{X}) \lesssim g(\mathcal{X})$ and $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ hold simultaneously. Moreover, $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ are defined analogously, except that all logarithmic factors in $\mathcal{X}$ are hidden.

Algorithm	Regret upper bound	Range of K that attains optimal regret	Sample complexity (or PAC bound)
MVP **(this work, Theorem 1)**	$\min\{\sqrt{SAH^3K}, HK\}$	$[1,\infty)$	$\frac{SAH^3}{\varepsilon^2}$
UCBVI (Azar et al., 2017)	$\min\{\sqrt{SAH^3K}+S^2AH^3,\ HK\}$	$[S^3AH^3,\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{S^2AH^3}{\varepsilon}$
ORLC (Dann et al., 2019)	$\min\{\sqrt{SAH^3K}+S^2AH^4,\ HK\}$	$[S^3AH^5,\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{S^2AH^4}{\varepsilon}$
EULER (Zanette and Brunskill, 2019)	$\min\{\sqrt{SAH^3K}+S^{3/2}AH^3(\sqrt{S}+\sqrt{H}),\ HK\}$	$[S^2AH^3(\sqrt{S}+\sqrt{H}),\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{S^2AH^3(\sqrt{S}+\sqrt{H})}{\varepsilon}$
UCB-Adv (Zhang et al., 2020)	$\min\{\sqrt{SAH^3K}+S^2A^{3/2}H^{33/4}K^{1/4},\ HK\}$	$[S^6A^4H^{27},\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{S^{8/3}A^2H^{11}}{\varepsilon^{4/3}}$
MVP (Zhang et al., 2021)	$\min\{\sqrt{SAH^3K}+S^2AH^2,\ HK\}$	$[S^3AH,\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{S^2AH^2}{\varepsilon}$
UCB-M-Q (Ménard et al., 2021)	$\min\{\sqrt{SAH^3K}+SAH^4,\ HK\}$	$[SAH^5,\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{SAH^4}{\varepsilon}$
Q-Earlysettled-Adv (Li et al., 2021b)	$\min\{\sqrt{SAH^3K}+SAH^6,\ HK\}$	$[SAH^9,\infty)$	$\frac{SAH^3}{\varepsilon^2}+\frac{SAH^6}{\varepsilon}$
Lower bound (Domingues et al., 2021)	$\min\{\sqrt{SAH^3K}, HK\}$	n/a	$\frac{SAH^3}{\varepsilon^2}$

Table 1: Comparisons between our result and prior works that achieve asymptotically optimal regret for finite-horizon inhomogeneous MDPs (with all log factors omitted), where S (resp. A) is the number of states (resp. actions), H is the planning horizon, and K is the number of episodes. The third column reflects the burn-in cost, and the sample complexity (or PAC bound) refers to the number of episodes needed to yield ε accuracy. The results provided here account for all $K \geq 1$ and/or all $\varepsilon \in (0, H]$. Our paper is the only one that gives regret (resp. PAC) bound matching the minimax lower bound for the entire range of K (resp. ε).

meaning that even the most advanced prior results fall short of sample optimality unless

$$\big(\text{best burn-in cost in past works}\big) \qquad K \gtrsim \min\big\{SAH^5, S^3AH\big\}. \tag{6}$$

The interested reader is referred to Table 1 for more details about existing regret upper bounds and the associated sample complexity.

In summary, no prior theory was able to achieve optimal sample complexity in the data-hungry regime

$$SAH \leq K \lesssim \min\big\{SAH^5, S^3AH\big\},$$

suffering from a significant barrier of either long horizon (as in the term SAH^5) or large state space (as in the term S^3AH). In fact, the information-theoretic limit is yet to be determined within this regime (i.e., neither the achievability results nor the lower bounds had been shown to be tight), although it has been conjectured by Ménard et al. (2021) that the lower bound (1) reflects the correct scaling for any sample size K.[2]

Comparisons with other RL settings and key challenges. In truth, the incentive to minimize the burn-in cost and improve data efficiency arises in multiple other settings beyond online RL. For instance, in an idealistic setting that assumes access to a simulator or a generative model — a model that allows the learner to query arbitrary state-action pairs to draw samples — a recent work Li et al. (2020) developed a perturbed model-based approach that is provably optimal without incurring any burn-in cost. Analogous results have been obtained in Li et al. (2021d) for offline RL — a setting that requires policy learning to be performed based on historical data — unveiling the full-range optimality of a pessimistic model-based algorithm.

Unfortunately, the algorithmic and analysis frameworks developed in the above two works fail to accommodate the online counterpart. The main hurdle stems from the complicated statistical dependency; for instance, in online RL, the estimated transition probabilities at one time step and the value function estimates of the next step are oftentimes statistically dependent (unless we waste data), while in contrast,

[2]Note that the original conjecture in Ménard et al. (2021) was $\widetilde{\Theta}\big(\sqrt{SAH^3K}+SAH^2\big)$. Combining it with the trivial upper bound HK allows one to remove the term SAH^2 (with a little algebra).

there is no such dependency in the simulator setting (as it typically assumes samples are taken completely independently). How to decouple the intricate statistical dependency without compromising data efficiency constitutes the key innovation of this work. More in-depth technical discussions will be provided in Section 4.

1.2 A peek at our main contributions

We are now positioned to summarize the main findings of this paper. Focusing on time-inhomogeneous finite-horizon MDPs, our main contributions can be divided into two parts: the first part fully settles the minimax-optimal regret and sample complexity of online RL, whereas the second part further extends and augments our theory to make apparent the impacts of certain problem-dependent quantities. Throughout this subsection, the regret metric $\mathsf{Regret}(K)$ captures the cumulative sub-optimality gap (i.e., the gap between the performance of the policy iterates and that of the optimal policy) over all K episodes, to be formally defined in (15).

1.2.1 Settling the optimal sample complexity with no burn-in cost

Our first result *fully* determines the sample complexity limit of online RL in a minimax sense, allowing one to attain the optimal regret regardless of the number K of episodes that can be collected.

Theorem 1. *For any $K \geq 1$ and any $0 < \delta < 1$, one can design an algorithm (see Algorithm 1) obeying*

$$\mathsf{Regret}(K) \leq \widetilde{O}\left(\min\left\{\sqrt{SAKH^3},\, HK\right\}\right) \tag{7}$$

with probability at least $1 - \delta$.

The optimality of our regret bound (7) can be readily seen given that it matches the minimax lower bound (1) (modulo some logarithmic factor). One can also easily translate the above regret bound into sample complexity or probably approximately correct (PAC) bounds: the proposed algorithm is able to return an ε-suboptimal policy with high probability using at most

$$\text{(sample complexity)} \qquad \widetilde{O}\left(\frac{SAH^3}{\varepsilon^2}\right) \quad \text{episodes} \tag{8}$$

(or equivalently, $\widetilde{O}\big(\frac{SAH^4}{\varepsilon^2}\big)$ sample transitions as each episode has length H). Remarkably, this result holds true for the entire ε range (i.e., any $\varepsilon \in (0, H]$), essentially eliminating the need of any burn-in cost. It is noteworthy that even in the presence of an idealistic generative model, this order of sample size is un-improvable (Azar et al., 2013; Li et al., 2020).

The algorithm proposed herein is a modified version of *Monotonic Value Propagation (MVP)*. Originally proposed by Zhang et al. (2021), the MVP method falls under the category of the model-based approach, a family of algorithms that construct explicit estimates of the probability transition kernel before value estimation and policy learning. Notably, a technical obstacle that obstructs the progress in understanding model-based algorithms arises from the exceedingly large model dimensionality: given that the dimension of the transition kernel scales proportionally with S^2, all existing analyses for model-based online RL fell short of effectiveness unless the sample size already far exceeds S^2 (Azar et al., 2017; Zhang et al., 2021). To overcome this undesirable source of burn-in cost, a crucial step is to empower the analysis framework in order to accommodate the highly sub-sampled regime (i.e., a regime where the sample size scales linearly with S), which we shall elaborate on in Section 4. The full proof of Theorem 1 will be postponed to Appendix C.

1.2.2 Extension: optimal problem-dependent regret bounds

In practice, RL algorithms often perform far more appealingly than what their worst-case performance guarantees would suggest. This motivates a recent line of works that goes beyond worst-case regret to investigate optimal performance in a more problem-dependent fashion (Dann et al., 2021; Fruit et al., 2018; Jin et al., 2020; Simchowitz and Jamieson, 2019; Talebi and Maillard, 2018; Tirinzoni et al., 2021;

Wagenmaker et al., 2022; Xu et al., 2021; Yang et al., 2021; Zanette and Brunskill, 2019; Zhao et al., 2023; Zhou et al., 2023). Encouragingly, the proposed algorithm automatically achieves optimality on a more refined problem-dependent level, without requiring prior knowledge of additional problem-specific knowledge. This results in a couple of extended theorems that take into account different problem-dependent quantities.

The first extension below investigates how the optimal value influences the regret bound.

Theorem 2 (Optimal value-dependent regret). *For any $K \geq 1$ and any $0 < \delta < 1$, Algorithm 1 satisfies*

$$\mathsf{Regret}(K) \leq \widetilde{O}\left(\min\left\{\sqrt{SAH^2Kv^*} + SAH^2,\ Kv^*\right\}\right) \tag{9}$$

with probability at least $1-\delta$, where v^* is the value of the optimal policy averaged over the initial state distribution (see (29) for the formal definition of v^*).

There is also no shortage of applications where the use of a cost function is preferred over a value function (Agarwal et al., 2017; Allen-Zhu et al., 2018; Lee et al., 2020; Wang et al., 2023). For this purpose, we provide another variation based upon the optimal cost.

Corollary 1 (Optimal cost-dependent regret). *For any $K \geq 1$ and any $0 < \delta < 1$, Algorithm 1 achieves*

$$\mathsf{Regret}(K) \leq \widetilde{O}\left(\min\left\{\sqrt{SAH^2Kc^*} + SAH^2,\ K(H-c^*)\right\}\right) \tag{10}$$

with probability exceeding $1-\delta$, where c^* denotes the cost of the optimal policy averaged over the initial state distribution (to be formally defined in (31)).

It is worth noting that: despite the apparent similarity of the statements of Theorem 2 and Corollary 1, they do not imply each other, although their proofs are very similar to each other.

Finally, we establish another regret bound that reflects the effect of certain variance quantities of interest.

Theorem 3 (Optimal variance-dependent regret). *For any $K \geq 1$ and any $0 < \delta < 1$, Algorithm 1 obeys*

$$\mathsf{Regret}(K) \leq \widetilde{O}\left(\min\left\{\sqrt{SAHK\mathrm{var}} + SAH^2,\ KH\right\}\right) \tag{11}$$

with probability at least $1-\delta$, where var *is a certain variance-type metric (to be formally defined in (35)).*

Two remarks concerning the above extensions are in order:

- In the worst-case scenarios, the quantities v^*, c^* and var can all be as large as the order of H, in which case Theorems 2-3 readily recover Theorem 1. In contrast, the advantages of Theorems 2-3 over Theorem 1 become more evident in those favorable cases (e.g., the situation where $v^* \ll H$ or $c^* \ll H$, or the case when the environment is nearly deterministic (so that var ≈ 0)).
- Interestingly, the regret bounds in Theorems 2-3 and Corollary 1 all contain a lower-order term SAH^2, and one might naturally wonder whether this term is essential. To demonstrate the unavoidable nature of this term and hence the optimality of Theorems 2-3 and Corollary 1, we will establish matching lower bounds, to be detailed in Section 5.

1.3 Related works

Let us take a moment to discuss several related theoretical works on tabular RL; there has also been an active line of research that exploits low-dimensional function approximation to further reduce sample complexity, which is beyond the scope of this paper.

Our discussion below focuses on two mainstream approaches that have received widespread adoption: the model-based approach and the model-free approach. In a nutshell, model-based algorithms decouple model estimation and policy learning, and often use the learned transition kernel to compute the value function and find a desired policy. In stark contrast, the model-free approach attempts to estimate the optimal value function and optimal policy directly without explicit estimation of the model. In general, model-free algorithms only require $O(SAH)$ memory for the purpose of storing Q-functions and value functions, while the model-based counterpart might require $O(S^2AH)$ space in order to store the estimated transition kernel.

Sample complexity for RL with a simulator. As an idealistic setting that separates the problem of exploration from that of estimation, RL with a simulator (or generative model) has been studied by numerous works, allowing the learner to query any state-action pairs and draw independent samples (Agarwal et al., 2020; Azar et al., 2013; Beck and Srikant, 2012; Chen et al., 2020; Cui and Yang, 2021; Even-Dar and Mansour, 2003; Kakade, 2003; Kearns and Singh, 1998a; Li et al., 2021a, 2022a, 2020; Pananjady and Wainwright, 2020; Shi et al., 2023; Sidford et al., 2018a,b; Wainwright, 2019a,b). While both model-based and model-free approaches are capable of achieving asymptotic sample optimality (Agarwal et al., 2020; Azar et al., 2013; Sidford et al., 2018b; Wainwright, 2019b), all model-free algorithms that enjoy asymptotically optimal sample complexity suffer from dramatic burn-in cost. Thus far, only the model-based approach has been shown to fully eliminate the burn-in cost for both discounted infinite-horizon MDPs and inhomogeneous finite-horizon MDPs (Li et al., 2020). The optimal sample complexity for time-homogeneous finite-horizon MDPs remains open.

Sample complexity for offline RL. The emergent subfield of offline RL is concerned with learning based purely on a pre-collected dataset (Levine et al., 2020). A frequently used mathematical model assumes that historical data are collected (often independently) using some behavior policy, and the key challenges (compared with RL with a simulator) come from distribution shift and incomplete data coverage. The sample complexity of offline RL has been the focus of a large strand of recent works, with asymptotically optimal sample complexity achieved by multiple algorithms (Jin et al., 2021; Li et al., 2022b, 2021c; Qu and Wierman, 2020; Rashidinejad et al., 2021; Ren et al., 2021; Shi et al., 2022; Wang et al., 2022; Xie et al., 2021; Yan et al., 2022; Yin et al., 2022). Akin to the simulator setting, the fully optimal sample complexity (without burn-in cost) has only been achieved via the model-based approach when it comes to discounted infinite-horizon and inhomogeneous finite-horizon settings (Li et al., 2022b). All model-free methods incur substantial burn-in cost. Time-homogeneous finite-horizon MDPs also remain unsettled.

Sample complexity for online RL. Obtaining optimal sample complexity (or regret bound) in online RL without incurring any burn-in cost has been one of the most fundamental open problems in RL theory. In fact, the past decades have witnessed a flurry of activity towards improving the sample efficiency of online RL, partial examples including Agrawal and Jia (2017); Bartlett and Tewari (2009); Brafman and Tennenholtz (2003); Cai et al. (2019); Dann and Brunskill (2015); Dann et al. (2017); Domingues et al. (2021); Dong et al. (2019); Efroni et al. (2019); Fruit et al. (2018); Jaksch et al. (2010); Ji and Li (2023); Jin et al. (2018); Kakade (2003); Kearns and Singh (1998b); Kolter and Ng (2009); Lattimore and Hutter (2012); Li et al. (2021b, 2023, 2021d); Ménard et al. (2021); Neu and Pike-Burke (2020); Osband et al. (2013); Pacchiano et al. (2020); Russo (2019); Strehl et al. (2006); Strehl and Littman (2008); Szita and Szepesvári (2010); Tarbouriech et al. (2021); Wang et al. (2020); Xiong et al. (2022); Zanette and Brunskill (2019); Zhang et al. (2021, 2022, 2020). Unfortunately, no work has settled this problem completely: the state-of-the-art result for model-based algorithms still incurs a burn-in that scales at least quadratically in S (Zhang et al., 2021), while the burn-in cost of the best model-free algorithms (particularly with the aid of variance reduction introduced in Zhang et al. (2020)) still suffers from sub-optimal horizon dependency (Li et al., 2021b).

1.4 Notation

Before proceeding, let us introduce a set of notation to be used throughout. For any set $\mathcal{X}$, $\Delta(\mathcal{X})$ represents the set of probability distributions over the set $\mathcal{X}$. For any positive integer N, we denote $[N] = \{1, \ldots, N\}$. For any two vectors x, y with the same dimension, we use xy to abbreviate $x^\top y$. For any integer $S > 0$, any probability vector $p \in \Delta([S])$ and another vector $v = [v_i]_{1\le i\le S}$, we denote by $\mathbb{V}(p, v) \coloneqq pv^2 - (pv)^2$ the associated variance, where $v^2 = [v_i^2]_{1\le i\le S}$ represents element-wise square of v. Let $\mathbf{1}$ and $\mathbf{0}$ indicate respectively the all-one vector and the all-zero vector. Let $\mathbf{1}_s$ denote the vector with 1 at the s-th coordinate and 0 at other coordinates. We shall often use $\{\cdot\}_{(\lambda)}$ as shorthand for $\{\cdot\}_{\lambda\in\Lambda}$, where Λ is the set of all proper choices of the index λ; for example, $\{\cdot\}_{(s,a,h,k)}$ denotes $\{\cdot\}_{(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]}$. Without loss of generality, we assume throughout that K is a power of 2 to streamline presentation.

2 Preliminaries

In this section, we introduce the basics of tabular online RL, as well as some basic assumptions to be imposed throughout.

Basics of finite-horizon MDPs. This paper concentrates on time-inhomogeneous (or nonstatioanry) finite-horizon MDPs. Throughout the paper, we employ $\mathcal{S} = \{1, \ldots, S\}$ to denote the state space, $\mathcal{A} = \{1, \ldots, A\}$ the action space, and H the planning horizon. The notation $P = \big\{P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\big\}_{1 \le h \le H}$ denotes the probability transition kernel of the MDP; for any current state s at any step h, if action a is taken, then the state at the next step $h+1$ of the environment is randomly drawn from $P_{s,a,h} \coloneqq P_h(\cdot \,|\, s, a) \in \Delta(\mathcal{S})$. Also, the notation $R = \big\{R_{h,s,a} \in \Delta([0, H])\big\}_{1 \le h \le H, s \in \mathcal{S}, a \in \mathcal{A}}$ indicates the reward distribution; that is, while executing action a in state s at step h, the agent receives an immediate reward — which is non-negative and possibly stochastic — drawn from the distribution $R_{h,s,a}$ with mean $r_h(s, a)$. Additionally, a deterministic policy $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{1 \le h \le H}$ stands for an action selection rule, so that the action selected in state s at step h is given by $\pi_h(s)$. The readers can consult standard textbooks (e.g., Bertsekas (2019)) for more extensive descriptions.

In each episode, the learner starts from an initial state s_1 independently drawn from some fixed (but unknown) distribution $\mu \in \Delta(\mathcal{S})$. For each step $1 \le h \le H$, the learner takes action a_h, gains an immediate reward $r_h \sim R_{h,s_h,a_h}$, and the environment transits to the state s_{h+1} at step $h+1$ according to $P_{s_h,a_h,h}$. All of our results in this paper operate under the following assumption on the total reward.

Assumption 1. *In any episode, it holds that* $0 \le \sum_{h=1}^{H} r_h \le H$.

As can be easily seen, Assumption 1 is less stringent than another common choice that assumes $r_h \in [0, 1]$ for any h in any episode. In particular, Assumption 1 allows for sparse and spiky rewards along an episode; more discussions can be found in Jiang and Agarwal (2018); Wang et al. (2020).

Value function and Q-function. For any given policy π, one can define the value function $V^\pi = \{V_h^\pi : \mathcal{S} \to \mathbb{R}\}$ and the Q-function $Q^\pi = \{Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ such that

$$V_h^\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'} \,\middle|\, s_h = s \right], \qquad \forall (s, h) \in \mathcal{S} \times [H], \tag{12a}$$

$$Q_h^\pi(s, a) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'} \,\middle|\, (s_h, a_h) = (s, a) \right], \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{12b}$$

where the expectation $\mathbb{E}_\pi[\cdot]$ is taken over the randomness of the MDP under policy π, that is, the trajectory chooses $a_{h'} = \pi_h(s_h)$ for all $h \le h' \le H$ (resp. $h < h' \le H$) in the definition of V_h^π (resp. Q_h^π). Accordingly, we can define the optimal value function and the optimal Q-function respectively as follows:

$$V_h^*(s) \coloneqq \max_\pi V_h^\pi(s), \qquad \forall (s, h) \in \mathcal{S} \times [H], \tag{13a}$$

$$Q_h^*(s, a) \coloneqq \max_\pi Q_h^\pi(s, a) \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{13b}$$

Two important properties are worth mentioning: (a) the optimal value and the optimal Q-function are linked by the Bellman equation:

$$Q_h^*(s, a) = r_h(s, a) + (P_{s,a,h})^\top V_{h+1}^*, \qquad V_h^*(s) = \max_{a'} Q_h^*(s, a'), \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]; \tag{14}$$

(b) there exists a deterministic policy, denoted by π^*, that achieves optimality for all state-action-step tuples simultaneously, that is,

$$V_h^{\pi^*}(s) = V_h^*(s) \qquad \text{and} \qquad Q_h^{\pi^*}(s, a) = Q_h^*(s, a), \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Data collection process and performance metrics. During the learning process, the learner is allowed to collect K episodes of samples (using arbitrary policies it selects). More precisely, in the k-th episode, the learner is given an independently generated initial state $s_1^k \sim \mu$, and executes policy π^k (chosen based on data collected in previous episodes) to obtain a sample trajectory $\left\{(s_h^k, a_h^k, r_h^k)\right\}_{1\le h\le H}$, with s_h^k, a_h^k and r_h^k denoting the state, action and immediate reward at step h of this episode.

To evaluate the learning performance, a widely used metric is the (cumulative) regret over all K episodes:

$$\mathsf{Regret}(K) := \sum_{k=1}^{K} \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right), \tag{15}$$

and our goal is to design an online RL algorithm that minimizes $\mathsf{Regret}(K)$ regardless of the allowable sample size K. It is also well-known (see, e.g., Jin et al. (2018)) that a regret bound can often be readily translated into a PAC sample complexity result (which counts the number of episodes needed to find an ε-optimal policy $\widehat{\pi}$ in the sense that $\mathbb{E}_{s_1\sim\mu}\left[V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1)\right] \le \varepsilon$). For instance, the standard reduction argument in Jin et al. (2018) reveals that: if an algorithm achieves $\mathsf{Regret}(K) \le f(S,A,H)K^{1-\alpha}$ for some function f and some parameter $\alpha \in (0,1)$, then by randomly selecting a policy from $\{\pi^k\}_{1\le k\le K}$ as $\widehat{\pi}$ one achieves $\mathbb{E}_{s_1\sim\mu}\left[V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1)\right] \lesssim f(S,A,H)K^{-\alpha}$, thus resulting in a sample complexity bound of $\left(\frac{f(S,A,H)}{\varepsilon}\right)^{1/\alpha}$.

3 A model-based algorithm: Monotonic Value Propagation

In this section, we formally describe our algorithm: a simple variation of the model-based algorithm called *Monotonic Value Propagation* proposed by Zhang et al. (2021). We present the full procedure in Algorithm 1, and point out several key ingredients.

- *Optimistic updates using upper confidence bounds (UCB).* The algorithm implements the optimism principle in the face of uncertainty by adopting the frequently used UCB-based framework (Azar et al., 2017; Jin et al., 2018). More specifically, the learner maintains upper estimates for both the value and Q-function, by calculating the following optimistic Bellman equation backward (from $h = H, \ldots, 1$):

$$Q_h(s,a) \leftarrow \min\left\{\widehat{r}_h(s,a) + \langle \widehat{P}_{s,a,h}, V_{h+1}\rangle + b_h(s,a), H\right\}, \tag{16a}$$
$$V_h(s) \leftarrow \max_a Q_h(s,a). \tag{16b}$$

  Here, Q_h (resp. V_h) is the running estimate for the Q-function (resp. value function), $\widehat{r}_h(s,a) \in \mathbb{R}$ is an estimate of the mean reward at (s,a,h), $\widehat{P}_{s,a,h} \in \mathbb{R}^S$ indicates an estimate of the transition probability vector from (s,a,h), whereas $b_h(s,a) \ge 0$ is some suitably chosen bonus term that compensates for the uncertainty.

- *Monotonic bonus functions.* Another crucial step in order to ensure near-optimal regret lies in careful designs of the data-driven bonus terms $\{b_h(s,a)\}$ in (16a). Here, we adopt the monotonic bonus function for MVP originally proposed in Zhang et al. (2021), to be made precise in (17). Compared to the bonus function in `Euler` (Zanette and Brunskill, 2019) and `UCBVI` (Azar et al., 2017), the monotonic bonus form has a cleaner structure that effectively avoid large lower order terms. In order to enable variance-aware regret, we also need to keep track of the empirical variance of the (stochastic) immediate rewards.

- *An epoch-based procedure and a doubling trick.* A key step of our algorithm is to update the empirical transition kernel and empirical rewards in an epoch-based fashion. More concretely, the whole learning process is divided into several consecutive epochs via a simple doubling rule. That is, once the number of visits to a (s,a,h)-tuple reaches a power of 2, we end the current epoch, reconstruct the empirical transition kernel and rewards using data from this epoch (cf. lines 11 and 13 of Algorithm 1), compute the Q-function and value function using the newly updated transition kernel and rewards (cf. (18)), and then start a new epoch. In each epoch, the learned policy is induced by the optimistic Q-function estimate computed based on the empirical transition kernel of the *current* epoch.

Remark 1 (Doubling batch). *We note that a doubling update rule has also been used in the original MVP (Zhang et al., 2021). A major difference between our modified version and the original one is that: when the visitation count for some (s,a,h) reaches 2^i for some integer i, we only use the second half of the samples (i.e., the $\{2^{i-1}+j\}_{j=1}^{2^{i-1}}$-th samples) to compute the empirical model, whereas the original MVP makes use of all the 2^i samples. This step is crucial for decoupling statistical dependence.*

Algorithm 1: Monotoinic Value Propagation (MVP) (Zhang et al., 2021)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, total number of episodes K, confidence parameter δ, $c_1 = \frac{460}{9}$, $c_2 = 2\sqrt{2}$ and $c_3 = \frac{544}{9}$.

initialization: for all $(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$, set $\theta_h(s,a) \leftarrow 0$, $\kappa_h(s,a) \leftarrow 0$, $\overline{N}_h(s,a,s') \leftarrow 0$, $N_h(s,a,s') \leftarrow 0$, $n_h(s,a) \leftarrow 0$, $Q_h(s,a) \leftarrow H-h+1$, $V_h(s) \leftarrow H-h+1$.

for $k = 1, 2, \ldots$ **do**

  Set π^k such that $\pi_h^k(s) = \arg\max_a Q_h(s,a)$ for all $s \in \mathcal{S}$ and $h \in [H]$. /* policy iterate. */

  for $h = 1, 2, \ldots, H$ **do**

    Observe s_h^k, take action $a_h^k = \arg\max_a Q_h(s_h^k, a)$, receive r_h^k, observe s_{h+1}^k. /* sampling. */

    $(s,a,s') \leftarrow (s_h^k, a_h^k, s_{h+1}^k)$.

    Update $\overline{N}_h(s,a) \leftarrow \overline{N}_h(s,a)+1$, $N_h(s,a,s') \leftarrow N_h(s,a,s')+1$, $\theta_h(s,a) \leftarrow \theta_h(s,a) + r_h^k$, $\kappa_h(s,a) \leftarrow \kappa_h(s,a) + (r_h^k)^2$.

    /* perform updates using data of this epoch. */

    if $\overline{N}_h(s,a) \in \{1, 2, \ldots, 2^{\log_2 K}\}$ **then**

      $n_h(s,a) \leftarrow \sum_{\widetilde{s}} N_h(s,a,\widetilde{s})$. // number of visits to (s,a,h) in this epoch.

      $\widehat{r}_h(s,a) \leftarrow \frac{\theta_h(s,a)}{n_h(s,a)}$. // empirical rewards of this epoch.

      $\widehat{\sigma}_h(s,a) \leftarrow \frac{\kappa_h(s,a)}{n_h(s,a)}$. // empirical squared rewards of this epoch.

      $\widehat{P}_{s,a,h}(\widetilde{s}) \leftarrow \frac{N_h(s,a,\widetilde{s})}{n_h(s,a)}$ for all $\widetilde{s} \in \mathcal{S}$. // empirical transition for this epoch.

      Set TRIGGERED = TRUE, and $\theta_h(s,a) \leftarrow 0$, $\kappa_h(s,a) \leftarrow 0$, $N_h(s,a,\widetilde{s}) \leftarrow 0$ for all $\widetilde{s} \in \mathcal{S}$.

  /* optimistic Q-estimation using empirical model of this epoch. */

  if TRIGGERED= TRUE **then**

    Set TRIGGERED = FALSE, and $V_{H+1}(s) \leftarrow 0$ for all $s \in \mathcal{S}$.

    for $h = H, H-1, \ldots, 1$ **do**

      for $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**

$$b_h(s,a) \leftarrow c_1\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a,h}, V_{h+1})\log\frac{1}{\delta}}{\max\{n_h(s,a),1\}}} + c_2\sqrt{\frac{\big(\widehat{\sigma}_h(s,a) - (\widehat{r}_h(s,a))^2\big)\log\frac{1}{\delta}}{\max\{n_h(s,a),1\}}} + c_3\frac{H\log\frac{1}{\delta}}{\max\{n_h(s,a),1\}}, \tag{17}$$

$$Q_h(s,a) \leftarrow \min\big\{\widehat{r}_h(s,a) + \langle\widehat{P}_{s,a,h}, V_{h+1}\rangle + b_h(s,a), H\big\},\ V_h(s) \leftarrow \max_a Q_h(s,a). \tag{18}$$

4 Technical overview

In this section, we point out the technical hurdles the previous approach encounters when mitigating the burn-in cost, and put forward a new strategy to overcome such hurdles.

4.1 Technical barriers in prior theory

A high-level diagnosis of the technical obstacles. Let us first single out a technical challenge on a high level. In the regret analysis, one central step is to control the error term $(\widehat{P}-P)V$, where $\widehat{P}$ represents a certain empirical transition kernel (constructed based on collected data), P stands for the true transition kernel, and V is a certain value function estimate. The analytical difficult arises in that V is often statistically dependent on $\widehat{P}$. A couple of strategies have been adopted in prior works to address this issue.

- The first strategy, which has been commonly used for model-based algorithms, decomposes the error term as (Azar et al., 2017; Dann et al., 2017; Zhang et al., 2021)

$$(\widehat{P}-P)V = (\widehat{P}-P)V^* + (\widehat{P}-P)(V-V^*).$$

  Given V^* is independent of $\widehat{P}$, one can apply Bernstein-style concentration inequalities to control the first term $(\widehat{P}-P)V^*$. As for the second term $(\widehat{P}-P)(V-V^*)$, note that $V-V^*$ might become exceedingly small when K is large enough; if this were the case, then one could simply bound this term via $\|\widehat{P}-P\|_1\|V-V^*\|_\infty$, which would become a negligible lower-order term. This approach, however, becomes problematic when K is not large enough, as this crude bound leads to an extra $\widetilde{O}(\sqrt{S})$ factor in the lower-order term due to the use of $\|\widehat{P}-P\|_1$.

- We now turn to the analysis of model-free algorithms (Jin et al., 2018; Li et al., 2021b; Ménard et al., 2021; Zhang et al., 2020). One way that has been used in earlier analyses (e.g., Jin et al. (2018)) can be described as follows: the learner first computes a value estimate V, and then employs news samples to construct $\widehat{P}$, which facilitates the analysis of $(\widehat{P}-P)V$ owing to certain independence between $(\widehat{P}-P)$ and V. Nevertheless, this strategy falls short of sample efficiency (even in an asymptotic large-sample sense), given that only the samples collected after computation of V are utilized. To enable asymptotic sample optimality, Zhang et al. (2021) proposed a solution called reference-advantage decomposition (or variance reduction). This strategy maintains a reference value estimate V^{ref} (computed using a previous batch of data in a way that obeys $V \approx V^{\mathrm{ref}}$) and decomposes

$$(\widehat{P}-P)V = (\widehat{P}-P)V^{\mathrm{ref}} + (\widehat{P}-P)(V-V^{\mathrm{ref}}),$$

  where the first term can be easily controlled if $\widehat{P}$ is based on data collected after V^{ref} is determined, and the second term vanishes if $V \approx V^{\mathrm{ref}}$. Unfortunately, this strategy also fails to enable all-regime optimality, since even computing the first version of V^{ref} at the initial stage already requires a large sample size.

A closer inspection on prior analysis for UCB-based algorithms. Next, let us take a closer inspection on the regret analysis for UCB-based model-based algorithms, in order to better illuminate the part that calls for novel analysis.

In each episode $k = 1, \ldots, K$, we update our estimates for the Q-function and the value function as follows: working backward (i.e., $h = H, H-1, \ldots, 1$), we set

$$Q_h^k(s,a) = \min\Big\{\widehat{r}_h^k(s,a) + \big\langle \widehat{P}_{s,a,h}^k, V_{h+1}^k\big\rangle + b_h^k(s,a),\ H\Big\}, \qquad \forall (s,a)\in\mathcal{S}\times\mathcal{A} \tag{19a}$$

$$V_h^k(s) = \max_a Q_h^k(s,a), \qquad \forall s\in\mathcal{S}. \tag{19b}$$

Here, $\widehat{r}_h^k(s,a)$ and $\widehat{P}_{s,a,h}^k$ represent respectively the empirical reward and the empirical transition model for the k-th episode, and $b_h^k(s,a)$ stands for a bonus function properly chosen to ensure that $Q_h^k(s,a) \geq Q_h^*(s,a)$ with high probability. These are computed from the collected data. We will specify $\widehat{P}_{s,a,h}^k$ and $\widehat{r}_h^k(s,a)$ later in Section 4.2, and $b_h^k(s,a) \geq 0$ has been described in Section 3. It has been shown using standard decomposition arguments that (Azar et al., 2017; Jaksch et al., 2010; Zhang et al., 2021)

$$\mathsf{Regret}(K) \lesssim \sum_{k,h} b_h^k\big(s_h^k, a_h^k\big) + \underbrace{\left|\sum_{k,h}\Big(\widehat{P}_{s_h^k,a_h^k,h}^k - P_{s_h^k,a_h^k,h}\Big)V_{h+1}^k\right|}_{=:T_{\mathrm{err}}} + \left|\sum_{k,h}\Big(\widehat{r}_h^k\big(s_h^k,a_h^k\big) - r_h\big(s_h^k,a_h^k\big)\Big)\right|. \tag{20}$$

In order to achieve full-range optimal regret, one needs to bound the three terms on the right-hand side of (20) carefully. The first term can be bounded in a rate-optimal manner (i.e., $\widetilde{O}\big(\sqrt{\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k)/N} + H/N\big)$) if we adopt the bonus construction in Zhang et al. (2021) for the original MVP (here, we omit the bonus tailored to stochastic rewards for simplicity). In the meantime, the third term on the right-hand side of (20) can be be easily coped with via standard Bernstein-style concentration inequalities.

The term that is the most challenging to control is the second error term T_{err} on the right-hand side of (20). Given the statistical dependency between $\widehat{P}^k_{s_h^k,a_h^k,h}$ and V_{h+1}^k, it is often difficult to directly apply concentration inequalities.[3] To see this, note that the estimation of $\widehat{P}^{k-1}_{s_h^k,a_h^k,h}$ determines the policy π^k for the k-th round, which in turns affects $\{\widehat{P}^k_{s,a,h}\}_{(s,a,h)}$ and V_{h+1}^k. On the other hand, $\widehat{P}^k_{s_h^k,a_h^k,h}$ is highly correlated with $\widehat{P}^{k-1}_{s_h^k,a_h^k,h}$ for most (s,a,h,k)-tuples, thus implying that V_{h+1}^k is not independent of $\widehat{P}^k_{s_h^k,a_h^k,h}$. In most prior analysis for model-based algorithms (Azar et al., 2017; Dann et al., 2017; Zanette and Brunskill, 2019; Zhang et al., 2021), this term T_{err} is decomposed as

$$\sum_{k,h}\Big(\widehat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h}\Big)V_{h+1}^k = \sum_{k,h}\Big(\widehat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h}\Big)V_{h+1}^* + \sum_{k,h}\Big(\widehat{P}_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h}\Big)(V_{h+1}^k - V_{h+1}^*).$$

The first term above can be bounded easily since V_{h+1}^* is fixed and independent of $\widehat{P}^k_{s_h^k,a_h^k,h}$. In comparison, the second term on the right-hand side of the above equation is a lower-order term, which would vanish as V_{h+1}^* converges to V_{h+1}^* (which would happen as K becomes large enough). Such arguments, however, are loose when analyzing the initial stage of the learning process — given that $V_{h+1}^k - V_{h+1}^*$ is not sufficiently small — resulting in a potentially large lower-order term and hence large burn-in cost.

4.2 A novel approach to decouple V from $\widehat{P}$

To address the above-mentioned issue, the key lies in decoupling the statistical dependence between $\widehat{P}^k_{s_h^k,a_h^k,h}$ and V_{h+1}^k. Let us first look at the relationship between $\widehat{P}^k_{s_h^k,a_h^k,h}$ and V_{h+1}^k. Let $\{N_h^k(s,a)\}_{(s,a,h)}$ be the number of visits to a state-action-step tuple (s,a,h) before the k-th episode. Note that V_{h+1}^k is determined by the samples after the h-th step up to the k-th episode. We can find that, $\widehat{P}^k_{s_h^k,a_h^k,h}$ *at most decides the count* after the h-th step. Therefore, if we pretend that the visitation counts $\{N_h^k(s,a)\}_{s,a,h,k}$ are fixed independent of $\widehat{P}^k_{s_h^k,a_h^k,h}$, then we can obtain a desired high-probability upper bound $\widetilde{O}\big(\sqrt{\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k)/N_h^k(s_h^k,a_h^k)}\big)$ on the quantity of interest $(\widehat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h})V_{h+1}^k$. One natural strategy is then to first develop such bounds for $\{N_h^k(s,a)\}_{s,a,h,k}$, and then invoke a covering argument that applies a union bound over all possible choices of $\{N_h^k(s,a)\}_{s,a,h,k}$.

Unfortunately, there are exponentially many choices for $\{N_h^k(s,a)\}_{s,a,h,k}$, thus preventing one from invoking the uniform convergence argument. In order to perform proper compression of the set of all possible choices of $\{N_h^k(s,a)\}_{(s,a,h,k)}$, we introduce doubling batches (as described in Section 3) during estimation of the value functions and Q-functions. To facilitate analysis, we have the following definitions.

Definition 1 (Doubling batch and estimations of transitions, rewards, and squared rewards)**.** *For any (s,a,h), the i-th batch for (s,a,h) is the collection of $2^{i-2}+j$-th sample for $j = 1,2,\ldots,2^{i-2}$ for $i \geq 2$, and the first sample for $i = 1$.*[4] *We define $\widehat{P}^{(j)}_{s,a,h}$, $\widehat{r}_h^{(j)}(s,a)$ and $\widehat{\sigma}_h^{(j)}(s,a)$ to be the empirical transition probability, the empirical reward, and the empirical squared reward of the j-th batch for (s,a,h), respectively. For completeness, we define the 0-th batch for each (s,a,h) as an empty set, and set $\widehat{P}^{(0)}_{s,a,h} = \frac{1}{S}\mathbf{1}$, $\widehat{r}_h^{(0)}(s,a) = 0$ and $\widehat{\sigma}_h^{(0)}(s,a) = 0$ for the 0-th batch.*

[3] This is different from the simulator and offline RL setting for inhomogeneous MDPs (Li et al., 2022b, 2020), as $\widehat{P}^k_{s_h^k,a_h^k,h}$ and V_{h+1}^k are (or can be made) independent therein.

[4] It is possible that the total count of (s,a,h) is less than K after K episodes. In this case, we add some virtual samples to fill the $\log_2(K)+1$ batches.

Definition 2 (Profile). *Fix $k \in [K]$. Let $I^k_{s,a,h}$ be the largest integer obeying $2^{I^k_{s,a,h}-1} \le N^k_h(s,a)$ for each (s,a,h). In particular, when $N^k_h(s,a) = 0$, we set $I^k_{s,a,h} = 0$. The profile for the k-th episode is defined as*

$$\mathcal{I}^k := \left\{I^k_{s,a,h}\right\}_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}. \tag{21}$$

We further let $\mathcal{I} := \{\mathcal{I}^k\}_{k=1}^K$ be the total profile.

With regards to the online RL, we can define a natural filtration induced by the sequential learning process. The formal definition is as follows.

Definition 3 (Online filtration). *For any $(h,k) \in [H+1]\times[K]$, let $\mathcal{F}^k_h$ be the σ-algebra induced by events happening before the h-th step in the k-th episode. Then $\{\mathcal{F}^k_h\}_{(h,k)\in[H]\times[K]}$ — with proper ordering in accordance with the sequential learning process — constructs a filtration $\mathcal{F}_{\text{online}}$, which we shall refer to as "online filtration" throughout.*

In order to facilitate analysis, we find it helpful to introduce another filtration below tailored to a generative model (which can be defined in a more flexible way than the online counterpart).

Definition 4 (Generative filtration). *Consider an order over all state-action pairs in $\mathcal{S}\times\mathcal{A}$ such that $\mathcal{S}\times\mathcal{A} = \{(s^{(i)},a^{(i)})\}_{i=1}^{SA}$. Let us employ the following sampling order of the state-action-step tuples:*

$$\begin{array}{cccc} (s^{(1)},a^{(1)},H) & (s^{(2)},a^{(2)},H) & \cdots & (s^{(SA)},a^{(SA)},H) \\ (s^{(1)},a^{(1)},H-1) & (s^{(2)},a^{(2)},H-1) & \cdots & (s^{(SA)},a^{(SA)},H-1) \\ & \cdots & & \\ (s^{(1)},a^{(1)},1) & (s^{(2)},a^{(2)},1) & \cdots & (s^{(SA)},a^{(SA)},1) \end{array} \tag{22}$$

where for each (s,a,h) we draw K independent sample transitions from the generative model. For any $1 \le t \le K$, define $\overline{\mathcal{F}}_{s,a,h}(t)$ to be the σ-algebra induced by events happening after the t-th sample of (s,a,h) is collected. For any $1 \le z \le SAHK$, define $\widetilde{\mathcal{F}}(z) := \overline{\mathcal{F}}_{s^{(i)},a^{(i)},h}(t)$, where (i,h,t) is chosen be such that $z = (H-h)\cdot SA\cdot K + (i-1)\cdot K + t$. Then (a proper ordering of) $\{\widetilde{F}(z)\}_{z=1}^{SAKH}$ constructs a filtration $\mathcal{F}_{\text{gen}}$, which we shall refer to as "generative filtration" throughout.

In words, there are a total number of $SAHK$ samples to be collected (K i.i.d. samples for each (s,a,h)), and we introduce a sequential ordering of them, with $\widetilde{F}(z)$ denoting the σ-algebra after the z-th sample. For convenience, we assume that all initial states have been generated from μ before the online learning process starts. Then $\widetilde{F}(SAHK)$ could be viewed as an expansion of $\mathcal{F}^k_{H+1}$, since one could simulate the whole online learning process using the $SAHK$ independent samples in the generative filtration. In other words, for each event in the online filtration $\mathcal{F}_{\text{online}}$, it is measurable w.r.t. the generative filtration $\mathcal{F}_{\text{gen}}$. This allows one to conduct analysis based on the generative filtration (as we shall detail momentarily). We also remark that we will only use the generative filtration $\mathcal{F}_{\text{gen}}$ when necessary, given that the analysis under the online filtration $\mathcal{F}_{\text{online}}$ is more natural for an online learning problem.

Doubling batch updates. We update the value function and Q-function with the doubling batches. Namely, in the k-th episode, we choose $\widehat{P}^k_{s,a,h} = \widehat{P}^{(I^k_{s,a,h})}_{s,a,h}$, $\widehat{r}^k_h(s,a) = \widehat{r}^{(I^k_{s,a,h})}_h(s,a)$ and $\widehat{\sigma}^k_h(s,a) = \widehat{\sigma}^{(I^k_{s,a,h})}_h(s,a)$ for any (s,a,h,k). Our update rule is a slightly different from previous doubling update rules (Jaksch et al., 2010), where the algorithm keeps running a policy until the visitation count of some (s,a,h) doubles and then uses the whole dataset to compute the empirical transition model. In contrast, we divide the whole dataset into disjoint batches following Definition 1, and only use the latest batch to compute the empirical transition model. We design this update rule because the samples in different batches are not correlated, which could help us decouple the value function and empirical transition model. Crucially, our update rule preserves sample efficiency, since the latest batch always contains at least half of the samples.

4.3 Key lemmas

As discussed above, under the generative filtration $\mathcal{F}_{\text{gen}}$, the random vector $\widehat{P}^k_{s_h^k,a_h^k,h}$ is conditionally independent of V^k_{h+1} for any $(k,h) \in [K] \times [H]$ if we fix $\mathcal{I}$. Then we can view the error term

$$T_{\text{err}} = \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\widehat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h}\right)V^k_{h+1}$$

as a martingale difference and obtain a desired bound. Following this intuition, we introduce our key lemma to bound the error term T_{err} with the doubling batch updates mentioned above.

Let $\mathcal{C}$ be a set which contains all possible values of the total profiles $\mathcal{I}$. One key novelty is to obtain a tight bound on $|\mathcal{C}|$, which we will discuss later.

Now, fix any $\mathcal{J} \in \mathcal{C}$, and consider the event $\mathcal{E}(\mathcal{J},\delta)$ defined as follows:

$$\mathcal{E}(\mathcal{J},\delta) := \left\{\mathcal{I} = \mathcal{J},\, T_{\text{err}} \le \sqrt{L\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}\big(P_{s_h^k,a_h^k,h}, V^k_{h+1}\big)\left(SAH + \log\frac{1}{\delta}\right)} + LH\left(SAH + \log\frac{1}{\delta}\right)\right\}, \quad (23)$$

where L is a logarithmic term in (S,A,H,K) to be defined shortly. We claim that

$$\Pr\big(\mathcal{E}(\mathcal{J},\delta)\big) \ge 1-\delta \quad (24)$$

for any $\mathcal{J} \in \mathcal{C}$ and $\delta \in (0,1)$. Then by applying the union bound over $\mathcal{J} \in \mathcal{C}$ and rescaling δ as $\delta/|\mathcal{C}|$, we obtain that with probability at least $1-\delta$,

$$T_{\text{err}} \le L\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}\big(P_{s_h^k,a_h^k,h}, V^k_{h+1}\big)\left(SAH + \log\frac{|C|}{\delta}\right)} + LH\left(SAH + \log\frac{|C|}{\delta}\right).$$

Consequently, we are left with accomplishing the following two steps:

1. Prove that $\Pr(\mathcal{E}(\mathcal{J},\delta)) \ge 1-\delta$ (i.e., (24)).
2. Determine $\mathcal{C}$ and bound $|\mathcal{C}|$ properly.

Proof of inequality (24). Towards this end, we need the lemma below.

Lemma 4. *Fix any $\mathcal{J} = \{J^k\}_{k=1}^K$. For each $h \times [H]$, let $\mathcal{X}_{h+1}$ be a set of vectors obeying*

- $\|X\|_\infty \le H$, $\forall X \in \mathcal{X}_{h+1}$;
- $\mathcal{X}_{h+1}$ *is determined by* $\{\widehat{P}^{(J^k_{s,a,h'})}_{s,a,h'}, \widehat{r}^{(J^k_{s,a,h'})}_{h'}(s,a), \widehat{\sigma}^{(J^k_{s,a,h'})}_{h'}(s,a)\}_{h' \in [h+1,H], k\in[K],(s,a)}$ *and* $\{J^k\}_{k=1}^K$;
- $|\mathcal{X}_{h+1}| \le W$ *for any* $1 \le h \le H$ *and some* $W \in \mathbb{N}$;
- *the all-zero vector* $\boldsymbol{0} \in \mathcal{X}_{h+1}$ *for each* $h \in [H]$.

Then with probability at least $1-\delta$, it holds that

$$\begin{aligned}&\left|\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\widehat{P}^{(J^k_{s_h^k,a_h^k,h})}_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h}\right)X^k_{h+1}\right| \\ &\le \sqrt{L\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}\big(P_{s_h^k,a_h^k,h}, X^k_{h+1}\big)\left(SAH\log W + \log\frac{1}{\delta}\right)} + LH\left(SAH\log W + \log\frac{1}{\delta}\right) \quad (25)\end{aligned}$$

for any sequence $\{X^k_{h+1}\}_{(h,k)}$ such that $X^k_{h+1} \in \mathcal{X}_{h+1}, \forall (h,k) \in [H]\times[K]$, where $L = 200(\log_2(K)+1)^2$.

The proof of Lemma 4 is based on a martingale concentration inequality in the view of $\mathcal{F}_{\text{gen}}$. We first fix the choice of $X^k_{h+1} \in \mathcal{X}_{h+1}$ for each (h,k), and then verify that $\sum_{s,a}\sum_{h=1}^{H} 2^{l-2}(\widehat{P}^{(l)}_{s,a,h} - P_{s,a,h})Y_{s,a,h}$ is a martingale difference for any $l \ge 2$ and $\{Y_{s,a,h}\}_{(s,a,h)}$ as long as $Y_{s,a,h}$ is selected from $\{X^k_{h+1}\}_{k=1}^K$ according to some specific rule for any (s,a,h). See Appendix B.1 for the proof of Lemma 4.

Bounding the size of possible profiles $|\mathcal{C}|$. Next, we turn to the second problem concerning $|\mathcal{C}|$. Let us choose

$$\mathcal{C} \coloneqq \left\{ \mathcal{J} = \{J^1, J^2, \ldots, J^K\} \,\middle|\, J^\tau \le J^{\tau+1}, \forall 1 \le \tau \le K-1, J^\tau \in \left\{0 \cup [\log_2 K]\right\}^{SAH}, \forall \tau \right\}. \tag{26}$$

Given that $i^k \le i^{k+1}$ for $1 \le k \le K-1$, it is easily seen that $\mathcal{I} \in \mathcal{C}$. The lemma below serves to upper bound the size of $\mathcal{C}$.

Lemma 5. *The choice (26) obeys* $|\mathcal{C}| \le (4SAHK)^{SAH(\log_2 K+1)}$.

In proving Lemma 5, we use the increasing property that $J^\tau \le J^{\tau+1}, \forall 1 \le \tau \le K-1$ for $\mathcal{J} = \{J^1, J^2, \ldots, J^K\} \in \mathcal{C}$. The naive bound for the size of $\mathcal{C}$ is $(\log_2(K)+1)^{SAHK}$, which is too large for our purpose. By virtue of the increasing property, we are actually counting the number of increasing paths in the SAH-dimensional grid $\{[\log_2(K)] \cup 0\}^{SAH}$. For each increasing path, there are at most $SAH(\log_2(K)+1)$ steps and at most SAH directions for each step. Then the proof can be completed with some primitive combinatorial computations. The detailed proof can be found in Appendix B.2.

With Lemma 4 and Lemma 5 in mind, we can invoke a uniform convergence argument to reach the lemma below; the proof of this lemma is postponed to Appendix B.3.

Lemma 6. *Recall that* $\mathcal{I} = \{I^k\}_{k=1}^K$ *is the total profile and the fact that* $\widehat{P}^k_{s,a,h} = \widehat{P}^{(I^k_{s,a,h})}_{s,a,h}, \widehat{r}^k_h(s,a) = \widehat{r}^{(I^k_{s,a,h})}_h(s,a), \widehat{\sigma}^k_h(s,a) = \widehat{\sigma}^{(I^k_{s,a,h})}_h(s,a)$ *for any proper* (s,a,k,h). *For each* $h \times [H]$, *let* $\mathcal{X}_{h+1}$ *be a set of vectors be such that: (1)* $\|X\|_\infty \le H, \forall X \in \mathcal{X}_{h+1}$ *; (2)* $\mathcal{X}_{h+1}$ *is determined by* $\{\widehat{P}^k_{s,a,h'}, \widehat{r}^k_{h'}(s,a), \widehat{\sigma}^k_{h'}(s,a)\}_{h+1 \le h' \le H, 1 \le k \le K, (s,a)}$ *and* $\{I^k\}_{k=1}^K$*; (3)* $|\mathcal{X}_{h+1}| \le W$ *for any* $1 \le h \le H$ *and some* $W \in \mathbb{N}$*; (4) the zero vector* $\mathbf{0} \in \mathcal{X}_{h+1}$ *for each* $h \in [H]$. *Then with probability at least* $1-\delta$, *it holds that*

$$\begin{aligned} &\left| \sum_{k=1}^K \sum_{h=1}^H \left(\widehat{P}^k_{s_h^k, a_h^k, h} - P_{s_h^k, a_h^k, h} \right) X^k_{h+1} \right| \\ &\le \sqrt{L_1 \sum_{k=1}^K \sum_{h=1}^H \mathbb{V}(P_{s_h^k, a_h^k, h}, X^k_{h+1}) \left(SAH \log W + \log \frac{1}{\delta} \right)} + L_1 H \left(SAH \log W + \log \frac{1}{\delta} \right) \end{aligned} \tag{27}$$

for any sequence $\{X^k_{h+1}\}_{h,k}$ *such that* $X^k_{h+1} \in \mathcal{X}_{h+1}, \forall (h,k) \in [H] \times [K]$, *where* $L_1 = 4000 \log_2^2(K) \log(SAHK)$.

In Algorithm 1, we compute V^k_{h+1} by the following rule.

$$\begin{aligned} &V^k_{H+1}(s) = 0; \forall s \in \mathcal{S}; \\ &V^k_{h'}(s) = \min \left\{ \max_a \left(\widehat{r}^k_{h'}(s,a) + \widehat{P}^k_{s,a,h'} V^k_{h'+1} + b(\widehat{P}^k_{s,a,h'}, \widehat{r}^k_{h'}(s,a), \widehat{\sigma}^k_{h'}(s,a), n^k_h(s,a)) \right), H \right\} \end{aligned} \tag{28}$$

for all $s \in \mathcal{S}$ and $h' = H, H-1, \ldots, h+1$. Here $b(\cdot,\cdot,\cdot,\cdot)$ is some proper bonus function and $n^k_h(s,a)$ is the size of the $I^k_{s,a,h}$-th batch of (s,a,h). Then V^k_{h+1} is determined by $\{\widehat{P}^k_{s,a,h'}, \widehat{r}^k_{h'}(s,a), \widehat{\sigma}^k_{h'}(s,a)\}_{h+1 \le h' \le H, 1 \le k' \le k, (s,a)}$ and $\{I^k\}_{k=1}^K$, thus allowing us to apply Lemma 6 to bound T_{err} by choosing $\mathcal{X}_{h+1} = \{V^k_{h+1}\}_{k=1}^K$. In addition, it is worth noting that Lemma 6 is more general compared to our original target to bound T_{err}, since $\mathcal{X}_{h+1}$ can be chosen as arbitrary functions.

5 Extensions

With the refined error bound derived in Section 4.3, we can readily obtain more refined regret bounds for Algorithm 1 to reflect the role of several problem-dependent quantities. Most of the arguments in the analysis are similar to those in the previous work Zhou et al. (2023). Detailed proofs are postponed to Appendix D and Appendix F.

Value-based regret bounds. Thus far, we have not yet introduced the crucial quantity $v^\star$ in Theorem 2, which we define now. When the initial states are drawn from μ, v^{star} stands for the weighted optimal value:

$$v^* := \mathbb{E}_{s\sim\mu}\big[V_1^*(s)\big]. \tag{29}$$

Encouragingly, the value-dependent regret bound in Theorem 2 is still minimax-optimal, as asserted by the following lower bound.

Theorem 7. *Consider any $p \in [0,1]$ and $K \geq 1$. For any learning algorithm, there exists an MDP with S states, A actions and horizon H obeying $v^* \leq Hp$ and*

$$\mathbb{E}\big[\mathsf{Regret}(K)\big] \gtrsim \min\big\{\sqrt{SAH^3Kp},\, KHp\big\}. \tag{30}$$

In fact, the construction of the hard instance (as required in Theorem 7) is quite simple. Design a new branch with 0 reward and set the probability of reaching this branch to be $1-p$. Also, with probability p, we direct the learner to a hard instance with regret $\Omega(\min\{\sqrt{SAH^3Kp}, KpH\})$ and optimal value H. This guarantees that the optimal value $v^* \leq Hp$ and that the expected regret is at least $\Omega(\min\{\sqrt{SAH^3Kp}, KHp\}) \gtrsim \min\{\sqrt{SAH^2Kv^*}, Kv^*\}$. See Appendix G for more details.

Cost-based regret bounds. Next, we turn to the cost-aware regret bound as in Corollary 1. Note that all other results except for Corollary 1 are about rewards as opposed to cost. In order to facilitate discussion, let us first formally introduce the cost-based scenarios.

Suppose that the reward distributions $\{R_{h,s,a}\}_{(s,a,h)}$ are replaced with the cost distributions $\{C_{h,s,a}\}_{(s,a,h)}$, where each distribution $C_{h,s,a} \in \Delta([0,H])$ has mean $c_h(s,a)$. In the h-th step of an episode, the learner pays an immediate cost $c_h \sim C_{h,s_h,a_h}$ instead of receiving an immediate reward r_h, and the objective of the learner is instead to minimize the total cost $\sum_{h=1}^H c_h$ (in an expected sense). The optimal cost quantity c^* is then defined as

$$c^* := \min_\pi \mathbb{E}_{\pi, s_1\sim\mu}\left[\sum_{h=1}^H c_h\right]. \tag{31}$$

Similarly, we can re-define the Q-function and value function as follows:

$$\begin{aligned}
\mathtt{Q}_h^\pi(s,a) &:= \mathbb{E}_\pi\left[\sum_{h'=h}^H c_{h'} \,\middle|\, (s_h,a_h) = (s,a)\right], && \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H],\\
\mathtt{V}_h^\pi(s) &:= \mathbb{E}_\pi\left[\sum_{h'=h}^H c_{h'} \,\middle|\, s_h = s\right], && \forall (s,h) \in \mathcal{S}\times\times[H],
\end{aligned}$$

where we use different fonts to differentiate them from the original Q-function and value function. The optimal cost function is then given by $\mathtt{Q}_h^*(s,a) = \min_\pi \mathtt{Q}_h^\pi(s,a)$ and $\mathtt{V}_h^*(s) = \min_\pi \mathtt{V}_h^\pi(s)$. Given the definitions above, we overload the notation $\mathsf{Regret}(K)$ to denote the regret for the cost-based scenario as

$$\mathsf{Regret}(K) := \sum_{k=1}^K \Big(\mathtt{V}_1^{\pi^k}(s_1^k) - \mathtt{V}_1^*(s_1^k)\Big).$$

One can also simply regard the cost minimization problem as reward maximization with negative rewards by choosing $r_h = -c_h$. This way allows us to apply Algorithm 1 directly, except that (18) is replaced by

$$Q_h(s,a) \leftarrow \max\Big\{\min\Big\{\widehat{r}_h(s,a) + \widehat{P}_{s,a,h}V_{h+1} + b_h(s,a),\, 0\Big\},\, -H\Big\}. \tag{32}$$

Note that the proof of Corollary 1 closely resembles that of Theorem 2, which can be found in Appendix E.

To confirm the tightness of Corollary 1, we develop the following matching lower bound, which basically employs the same hard instance as in the proof of Theorem 7.

Corollary 2. *Consider any $p \in [0, \frac{1}{4}]$ and any $K \geq 1$. For any algorithm, one can construct an MDP with S states, A actions and horizon H obeying $c^* = \Theta(Hp)$ and*

$$\mathbb{E}\big[\mathsf{Regret}(K)\big] \gtrsim \min\big\{\sqrt{SAH^3Kp} + SAH^2,\, KH(1-p)\big\}.$$

Variance-dependent regret bound. The final regret bound presented in Theorem 3 depends on a sort of variance metrics. Towards this end, let us first make precise the variance metrics of interest:

(i) The first variance metric is defined as

$$\mathrm{var}_1 \coloneqq \max_{\pi} \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathbb{V}\big(P_{s_h,a_h,h}, V_{h+1}^*\big) + \sum_{h=1}^{H} \mathrm{Var}\big(R_h(s_h,a_h)\big) \right], \tag{33}$$

where $\{(s_h,a_h)\}_{1\le h\le H}$ represents a sample trajectory under policy π. This captures the maximal possible expected sum of variance with respect to the optimal value function $\{V_h^*\}_{h=1}^H$.

(ii) Another useful variance metric is defined as

$$\mathrm{var}_2 \coloneqq \max_{\pi,s} \mathrm{Var}_\pi \left[\sum_{h=1}^{H} r_h \,\Big|\, s_1 = s \right], \tag{34}$$

where $\{r_h\}_{1\le h\le H}$ denotes a sample sequence of immediate rewards under policy π. This indicates the maximal possible variance of the accumulative reward.

The interested reader is referred to Zhou et al. (2023) for further discussion about these two metrics. Our final variance metric is then defined as

$$\mathrm{var} \coloneqq \min\big\{\mathrm{var}_1, \mathrm{var}_2\big\}. \tag{35}$$

With the above metric var in mind, we can then revisit Theorem 3. When the transition model is fully deterministic, the regret bound in Theorem 3 simplifies to

$$\mathsf{Regret}(K) \le \widetilde{O}\big(\min\big\{SAH^2,\, HK\big\}\big)$$

for any $K \ge 1$, which is roughly the cost of visiting each state-action pair. The full proof of Theorem 3 is postponed to Appenndix F.

To finish up, let us develop a matching lower bound to corroborate the tightness and optimality of Theorem 3.

Theorem 8. *Consider any $p \in [0,1]$ and any $K \ge 1$. For any algorithm, one can find an MDP with S states, A actions, and horizon H satisfying $\max\{\frac{\mathrm{var}_1}{H^2}, \frac{\mathrm{var}_2}{H^2}\} \le p$ and*

$$\mathbb{E}\big[\mathsf{Regret}(K)\big] \gtrsim \min\big\{\sqrt{SAH^3Kp} + SAH^2,\, KH\big\}.$$

The proof of Theorem 8 resembles that of Theorem 7, except that we need to construct a hard instance when $K \le SAH/p$. For this purpose, we construct a fully deterministic MDP (i.e., all of its transitions are deterministic and all rewards are fixed), and show that the learner has to visit about half of the state-action-layer tuples in order to learn a near-optimal policy. The proof details are deferred to Appendix G.

6 Discussion

Focusing on tabular online RL in time-inhomogeneous finite-horizon MDPs, this paper has established the minimax-optimal regret (resp. sample complexity) — up to log factors — for the entire range of sample size $K \ge 1$ (resp. target accuracy level $\varepsilon \in (0, H]$), thereby fully settling an open problem at the core of recent RL theory. The MVP algorithm studied herein is model-based in nature. Remarkably, the model-based approach remains the only family of algorithms that is capable of obtaining minimax optimality without burn-ins, regardless of the data collection mechanism in use (e.g., online RL, offline RL, and the simulator setting). We have further unlocked the optimality of this algorithm in a more refined manner, making apparent the effect of several problem-dependent quantities (e.g., optimal value/cost, variance statistics) upon the fundamental

performance limits. The new analysis and algorithmic techniques put forward herein might shed important light on how to understand other RL settings as well.

Moving forward, there are multiple directions that anticipate further theoretical pursuit. To begin with, is it possible to develop a model-free algorithm — which often exhibits more favorable memory complexity compared to the model-based counterpart — that achieves full-range minimax optimality? As alluded to previously, existing paradigms that rely on reference-advantage decomposition (or variance reduction) seem to incur a high burn-in cost (Li et al., 2021b; Zhang et al., 2020), thus calling for new ideas to overcome this barrier. Additionally, multiple other tabular settings (e.g., time-homogeneous finite-horizon MDPs, discounted infinite-horizon MDPs) have also suffered from similar issues regarding burn-in requirements (Ji and Li, 2023; Zhang et al., 2021). Take time-homogeneous finite-horizon MDPs for example: in order to achieve optimal sample efficiency, one needs to carefully deal with the statistical dependency incurred by aggregating data from across different time steps to estimate the same transition matrix (due to the homogeneous nature of P), which results in more intricate issues than the time-homogeneous counterpart. We believe that resolving these two open problems will greatly enhance our theoretical understanding about online RL and beyond.

Acknowledgement

We thank for Qiwen Cui for helpful discussions. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-19-1-0030 and FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994. SSD acknowledges the support of NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844, NSF IIS 2229881.

A Technical lemmas

Lemma 9. *Let $(M_n)_{n\geq 0}$ be a martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ for some $c > 0$ and any $n \geq 1$. Let $\mathrm{Var}_n = \sum_{k=1}^{n} \mathbb{E}\left[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}\right]$ for $n \geq 0$, where $\mathcal{F}_k = \sigma(M_1, ..., M_k)$. Then for any positive integer n, and any $\epsilon, \delta > 0$, one has*

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

Lemma 10 (Lemma 30 in Chen et al. (2021))**.** *Let X be a random variable and $\|X\|_\infty$ denotes the largest possible value of X. Let $\mathrm{Var}(X)$ denote the variance of X. Then $\mathrm{Var}(X^2) \leq 4\|X\|_\infty^2 \mathrm{Var}(X)$.*

Lemma 11 (Lemma 10 in Zhang et al. (2022))**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables taking value in $[0, l]$. Define $\mathcal{F}_k = \sigma(X_1, X_2, \ldots, X_{k-1})$ and $Y_k = \mathbb{E}[X_k|\mathcal{F}_k]$ for $k \geq 1$. For any $\delta > 0$, we have*

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \geq 3\sum_{k=1}^{n} Y_k + l\ln\frac{1}{\delta}\right] \leq \delta$$
$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geq 3\sum_{k=1}^{n} X_k + l\ln\frac{1}{\delta}\right] \leq \delta.$$

Lemma 12 (Bennet's inequality)**.** *Let $Z, Z_1, ..., Z_n$ be i.i.d. random variables with values in $[0, 1]$ and let $\delta > 0$. Define $\mathbb{V}Z = \mathbb{E}\left[(Z - \mathbb{E}Z)^2\right]$. Then one has*

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

Lemma 13 (Theorem 4 in Maurer and Pontil (2009))**.** *Let $Z, Z_1, ..., Z_n$ ($n \geq 2$) be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. Define $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\widehat{V}_n = \frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z})^2$. Then we have*

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

Lemma 14. *Recall the definition of $N_h^k(s_h^k, a_h^k)$ in Algorithm 1. It holds that*

$$\sum_{k=1}^K \sum_{h=1}^H \frac{1}{\max\{N_h^k(s_h^k, a_h^k), 1\}} \leq 2SAH \log_2 K \tag{36}$$

Proof. By definition, for any fixed (s,a,h) we have

$$\sum_{k=1}^K \frac{1}{N_h^k(s_h^k, a_h^k)} \mathbb{I}\Big[(s,a) = \big(s_h^k, a_h^k\big)\Big] \leq \log_2 K + 1. \tag{37}$$

Summing over all (s,a,h) completes the proof. □

B Missing proofs in Section 4

B.1 Proof of Lemma 4

Since $\mathcal{X}_{h+1}$ has at most W elements, we can write $\mathcal{X}_{h+1} = \{x_{h+1}(w)\}_{w=1}^W$, where each $x_{h+1}(w)$ could be regarded as a function of $\{\widehat{P}_{s,a,h'}^{(J_{s,a,h}^k)}, \widehat{r}^{(J_{s,a,h'}^k)_{h'}(s,a)}, \widehat{\sigma}_{h'}^{(J_{s,a,h'}^k)}(s,a)\}_{h+1\leq h'\leq H, 1\leq k\leq K,(s,a)}$ and $\{J^k\}_{k=1}^K$. Now we fix a group of indices $\{w_{s,a,h}\}_{(s,a,h)}$ where $1 \leq w_{s,a,h} \leq W$ for each (s,a,h).

Fix $2 \leq l \leq \log_2(K) + 1$. We consider to bound the term

$$T(l, \{w_{s,a,h}\}_{(s,a,h)}) \coloneqq 2^{l-2} \sum_{s,a,h} (\widehat{P}_{s,a,h}^{(l)} - P_{s,a,h}) x_{h+1}(w_{s,a,h}).$$

Recall that $\widehat{P}_{s,a,h}^{(l)}$ is the empirical transition of the l-th batch of (s,a,h), i.e., the empirical transition of the $2^{l-2}+1$-th to 2^{l-1}-th samples of (s,a,h). Also recall the definition of $\mathcal{F}_{\text{gen}} = \{\widetilde{F}(z)\}_{z=1}^{SAHK}$ in Definition 4. Conditioned on $\widetilde{\mathcal{F}}((H-h)\cdot SAK)$, $\{x_{h+1}(w_{s,a,h})\}_{(s,a)}$ is fixed, and $\{2^{l-2}\widehat{P}_{s,a,h}^{(l)}\}_{(s,a)}$ are mutually independent multinomial random variables. Let $v_{s,a,h}(t,l)$ be the next state of the t-th sample of the l-th batch of (s,a,h). By writing

$$T(l, \{w_{s,a,h}\}_{(s,a,h)}) = \sum_{s,a,h} \sum_{\tau=1}^{2^{l-2}} (\mathbf{1}_{v_{s,a,h}(\tau,l)} - P_{s,a,h}) \cdot x_{h+1}(w_{s,a,h}), \tag{38}$$

using Lemma 9, with probability at least $1 - 10SAH^2K^2\delta'$,

$$T(l, \{w_{s,a,h}\}_{(s,a,h)}) \leq 2\sqrt{2} \cdot \sqrt{2^{l-2} \sum_{s,a,h} \mathbb{V}(P_{s,a,h}, x_{h+1}(w_{s,a,h})) \log\frac{1}{\delta'}} + 3H \log\frac{1}{\delta'}. \tag{39}$$

Note that $\{w_{s,a,h}\}_{(s,a,h)}$ has at most $(W)^{SAH}$ choices. Applying the union bound and rescaling δ' to $\frac{\delta'}{|W|^{SAH}}$, we see that: with probability at least $1 - 10SAH^2K^2\delta'$,

$$\begin{aligned}
&T(l, \{w_{s,a,h}\}_{(s,a,h)}) \\
&= 2^{l-2} \sum_{s,a,h} (\widehat{P}_{s,a,h}^{(l)} - P_{s,a,h}) x_{h+1}(w_{s,a,h})
\end{aligned}$$

$$\leq 2\sqrt{2}\cdot\sqrt{2^{l-2}\sum_{s,a,h}\mathbb{V}(P_{s,a,h},x_{h+1}(w_{s,a,h}))\left(2SAH\log W+\log\frac{1}{\delta'}\right)}+6H\left(SAH\log K+\log\frac{1}{\delta'}\right) \tag{40}$$

holds for any $\{w_{s,a,h}\}_{(s,a,h)}$ such that $1\leq w_{s,a,h}\leq W,\forall(s,a,h)$.

For $l=1$, we have that $\sum_{s,a,h}(\widehat{P}^{(l)}_{s,a,h}-P_{s,a,h})x_{h+1}(w_{s,a,h})\leq SAH^2$ trivially.

Now we rewrite

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\widehat{P}^{(J^k_{s^k_h,a^k_h,h})}_{s^k_h,a^k_h,h}-P_{s^k_h,a^k_h,h}\right)X^k_{h+1}\\
&=\sum_{l=0}^{\log_2(K)}\sum_{s,a,h}(\widehat{P}^{(l)}_{s,a,h}-P_{s,a,h})\sum_{k=1}^{K}\mathbb{I}[(s^k_h,a^k_h)=(s,a),I^k_{s,a,h}=l]X^k_{h+1}\\
&\leq\sum_{l=1}^{\log_2(K)}\sum_{o=1}^{2^{l-1}}\sum_{s,a,h}(\widehat{P}^{(l)}_{s,a,h}-P_{s,a,h})\sum_{k=1}^{K}\mathbb{I}[(s^k_h,a^k_h)=(s,a),I^k_{s,a,h}=l,\overline{N}^k_{s^k_h,a^k_h,h}=2^{l-1}+o]X^k_{h+1}+SAH^2\\
&=\sum_{l=1}^{\log_2(K)}\sum_{o=1}^{2^{l-1}}\sum_{s,a,h}(\widehat{P}^{(l)}_{s,a,h}-P_{s,a,h})X^{k_{l,o,s,a,h}}_{h+1}+SAH^2,
\end{aligned}\tag{41}$$

where $k_{l,o,s,a,h}$ denotes the index of the $(2^{l-1}+o)$-th sample of (s,a,h) in the online learning process. Recall that $\overline{N}^{K+1}_h(s,a)$ is the total visit count of (s,a,h) in K episodes. If $\overline{N}^{K+1}_h(s,a)<2^{l-1}+o$, we set $k_{l,o,s,a,h}=\infty$ and $X^{\infty}_{h+1}=0$. Fix $2\leq l\leq\log_2(K)+1$ and $1\leq o\leq 2^{l-1}$, we can find $\{w^{(l,o)}_{s,a,h}\}_{(s,a,h)}$ be such that $x_{h+1}(w^{(l,o)}_{s,a,h})=X^{k_{l,o,s,a,h}}_{h+1}$ for any proper (s,a,h). Using (40) for (l,o) such that $2\leq l\leq\log_2(K)+1$ and $1\leq o\leq 2^{l-1}$, with probability exceeding $1-2K\cdot 10SAH^2K^2\delta'$,

$$\begin{aligned}
&\sum_{l=1}^{\log_2(K)}\sum_{o=1}^{2^{l-1}}\sum_{s,a,h}(\widehat{P}^{(l)}_{s,a,h}-P_{s,a,h})X^{k_{l,o,s,a,h}}_{h+1}\\
&\leq\sum_{l=1}^{\log_2(K)}\sum_{o=1}^{2^{l-1}}\frac{1}{2^{l-2}}\left(\sqrt{16\cdot 2^{l-2}\sum_{s,a,h}\mathbb{V}(P_{s,a,h},X^{k_{l,o,s,a,h}}_{h+1})\cdot\left(SAH\log W+\log\frac{1}{\delta'}\right)}\right.\\
&\qquad\left.+6H\left(SAH\log W+\log\frac{1}{\delta'}\right)\right)\\
&\leq\sum_{l=1}^{\log_2(K)}\sqrt{32\cdot\sum_{s,a,h}\sum_{o=1}^{2^{l-1}}\mathbb{V}(P_{s,a,h},X^{k_{l,o,s,a,h}}_{h+1})\cdot\left(SAH\log W+\log\frac{1}{\delta'}\right)}\\
&\qquad+\sum_{l=1}^{\log_2(K)}12H\left(SAH\log W+\log\frac{1}{\delta'}\right)\\
&\leq\sqrt{64\log_2(K)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h},X^k_{h+1})\cdot(SAH\log(W)+\log(\frac{1}{\delta'}))}\\
&\qquad+12(\log_2 K)H\left(SAH\log W+\log\frac{1}{\delta'}\right).
\end{aligned}\tag{42}$$

In the last inequality, we have applied Cauchy's inequality and the fact that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h},X^k_{h+1})$$

$$= \sum_{l=1}^{\log_2(K)} \sum_{s,a,h} \sum_{o=1}^{2^{l-1}} \mathbb{V}(P_{s,a,h}, X_{h+1}^{k_{l,o,s,a,h}}) + \sum_{s,a,h} \sum_{k=1}^{K} \mathbb{I}\Big[(s_h^k, a_h^k) = (s,a), \overline{N}_h^k(s_h^k, a_h^k) = 0\Big] \mathbb{V}(P_{s,a,h}, X_{h+1}^k).$$

Using (41) and (42), and replacing δ with $\delta/(20SAH^2K^3)$, we finish the proof.

B.2 Proof of Lemma 5

Let $M = \log_2(K)$ and $N = SAH$. Let $\check{\mathcal{C}}(l) := \{\mathcal{J} = \{J^1, J^2, \ldots, J^l\} | J^\tau < J^{\tau+1}, \forall 1 \le \tau \le l-1, J^\tau \in \{0 \cup [M]\}^N, \forall \tau\}$ and $\check{\mathcal{C}} = \cup_{l \ge 1} \check{\mathcal{C}}(l)$. In words, $\check{\mathcal{C}}(l)$ is the set of *strict* increasing path in $\{0 \cup [M]\}^N$ with length l and $\check{\mathcal{C}}$ is the set of all *strict* increasing path.

We define $\text{Proj} : \mathcal{C} \to \check{\mathcal{C}}$ by mapping $\mathcal{J} \in \mathcal{C}$ to $\check{\mathcal{J}} \in \check{\mathcal{C}}$, where $\check{\mathcal{J}}$ is the set of all different elements in $\mathcal{J}$. Let $\mathcal{F}(\check{J}) := \{\mathcal{J} \in \mathcal{C} | \text{Proj}(\mathcal{J}) = \check{\mathcal{J}}\}$ for each $\check{\mathcal{J}} \in \check{\mathcal{C}}$. Because $\check{\mathcal{J}}$ is a *strict* increasing path, there are at most $MN+1$ elements in $\check{\mathcal{J}}$. As a result, the size of $\mathcal{F}(\mathcal{J})$ is at most the solution of the equation below

$$\sum_{i=1}^{MN+1} x_i = K, x_i \in \mathbb{N}, \forall 1 \le i \le MN+1$$

which is $\begin{pmatrix} K+MN \\ MN \end{pmatrix} = \frac{(K+MN)!}{(MN)!K!} \le (K+MN)^{MN} \le (2K)^{MN}$. It then holds that $|\mathcal{C}| \le |\check{\mathcal{C}}| \cdot (2K)^{MN}$.

We further consider the set $\check{\mathcal{C}}(MN+1)$. For $\check{\mathcal{J}} = \{J^1, J^2, \ldots, J^{MN+1}\} \in \check{\mathcal{C}}(MN+1)$, with pigeonhole principle, we have that $J^1 = [0, 0, \ldots, 0]^\top$ and $J^{MN+1} = [M, M, \ldots, M]^\top$. Moreover, for each $1 \le \tau \le MN$, J^τ and $J^{\tau+1}$ differ only at one dimension with distance 1. In this way, we can view $\check{\mathcal{J}}$ as an MN-step increasing path from $[0, 0, \ldots, 0]^\top$ to $[M, M, \ldots, M]^\top$. In each step, we have at most N directions. As a result, there are at most N^{MN} such paths, which implies that $|\check{\mathcal{C}}(MN+1)| \le N^{MN}$. Finally, noting that for any $\check{\mathcal{J}} \in \check{\mathcal{C}}$, there exists some $\check{\mathcal{J}}' \in \check{\mathcal{C}}(MN+1)$ such that $\check{\mathcal{J}} \subset \check{\mathcal{J}}'$, we conclude that $|\mathcal{C}| \le (2K)^{MN} |\check{\mathcal{C}}| \le (2K)^{MN} 2^{MN+1} |\check{\mathcal{C}}(MN+1)| \le (4KN)^{MN+1}$.

B.3 Proof of Lemma 6

Without loss of generality, we write $\widehat{P}^{(I^k)} = \widehat{P}^k = \{\widehat{P}^k_{s,a,h}\}_{(s,a,h)} = \{\widehat{P}^{(I^k_{s,a,h})}_{s,a,h}\}_{(s,a,h)}$. Then we can regard $\mathcal{X}_{h+1}$ as a function of $\{\widehat{P}^{(I^k)}\}_{k=1}^K$, i.e., $\mathcal{X}_{h+1} = \mathcal{X}^k_{h+1}(\{\widehat{P}^{(I^k)}\}_{k=1}^K, \{I^k\}_{k=1}^K)$. Let $\widetilde{\mathcal{E}}$ be the event where there exists $X^k_{h+1} \in \mathcal{X}_{h+1}(\{\widehat{P}^{(I^k)}\}_{k=1}^K, \{I^k\}_{k=1}^K), \forall (h,k) \in [H] \times [K]$ such that (27) does not hold. So it suffices to prove that $\Pr(\widetilde{\mathcal{E}}) \le \delta$. Fix $\mathcal{J} = \{J^k\}_{k=1}^K \in \mathcal{C}$. We consider the event $\mathcal{E}(\mathcal{J})$ where there exists a sequence $X^k_{h+1} \in \mathcal{X}_{h+1}(\{\widehat{P}^{(J^k)}\}_{k=1}^K, \{J^k\}_{k=1}^K), \forall (h,k) \in [H] \times [K]$ such that

$$\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H (\widehat{P}^{(J^k_{s,a,h})}_{s_h^k, a_h^k, h} - P_{s_h^k, a_h^k, h}) X^k_{h+1} \\
&\le \sqrt{L \sum_{k=1}^K \sum_{h=1}^H \mathbb{V}(P_{s_h^k, a_h^k, h}, X^k_{h+1}) \left(SAH \log(W) + \log(|\mathcal{C}|) + \log(1/\delta)\right)} \\
&\qquad + LH(SAH \log(W) + \log(|\mathcal{C}|) + \log(1/\delta)),
\end{aligned} \tag{43}$$

does not hold, where $L = 200(\log_2(K)+1)^2$. Let $\widetilde{\mathcal{E}}(\mathcal{J})$ be the event where there exists a sequence $X^k_{h+1} \in \mathcal{X}_{h+1}(\{\widehat{P}^{(J^k)}\}_{k=1}^K), \forall (h,k) \in [H] \times [K]$ such that (27) does not holds and $\mathcal{I} = \mathcal{J}$. Because $|\mathcal{C}| \le (4SAHK))^{SAH \log_2(K)+1}$, we have $\widetilde{\mathcal{E}}(\mathcal{J}) \subset \mathcal{E}(\mathcal{J})$. With Lemma 4, we learn that $\Pr(\widetilde{\mathcal{E}}(\mathcal{J})) \le \Pr(\mathcal{E}(\mathcal{J})) \le \frac{\delta}{|\mathcal{C}|}$. As a result, we have $\Pr(\cup_{\mathcal{J} \in \mathcal{C}} \widetilde{\mathcal{E}}(\mathcal{J})) \le \delta$. By noting that $\widetilde{\mathcal{E}} \subset \cup_{\mathcal{J} \in \mathcal{C}} \widetilde{\mathcal{E}}(\mathcal{J})$, we obtain that $\Pr(\widetilde{\mathcal{E}}) \le \delta$. The proof is completed.

C Regret analysis (proof of Theorem 1)

This section is devoted to the proof of Theorem 1. Below we assume that $K \geq BSAH$, where $B = 4000 \log_2^3(K) \log(3SAH) \log(\frac{1}{\delta})$. Let π^k be the policy in the k-th episode. Let $\bar{N}_h^k(s,a)$ be the count of (s,a,h) before the k-th episode and $N_h^k(s,a)$ be the count of the doubling batch used to compute the value function in the k-th episode. In particular, when $\bar{N}_h^k(s,a) = 0$, we define $N_h^k(s,a) = 1$. Let V_h^k and Q_h^k be respectively the value of V_h and Q_h before the k-th episode for all proper (s,a,k,h). Recall that $\hat{P}_{s,a,h}^k$ is the value of $\hat{P}_{s,a,h}$ before the k-th episode. Let $\hat{r}_h^k(s,a)$ be the empirical reward function before the k-th episode of (s,a). Let $\hat{\sigma}_h^k(s,a)$ be the empirical variance before the k-th episode for the state-action pair (s,a), i.e., the value of $\hat{\sigma}_h(s,a)$ before the k-th episode.

C.1 Optimism

Lemma 15. *With probability* $1-4SAHK\delta$, $Q_h^k(s,a) \geq Q_h^*(s,a)$ *and* $V_h^k(s) \geq V_h^*(s)$ *for any proper* (s,a,h,k).

Proof. For $p \in \Delta^S$, $v \in \mathbb{R}^S$, $\|v\|_\infty \leq H$ and $n \in N^+$, let

$$f(p,v,n) = pv + \max\left\{\frac{20}{3}\sqrt{\frac{\mathbb{V}(p,v)\log(\frac{1}{\delta})}{n}}, \frac{400}{9}\cdot\frac{H\log(\frac{1}{\delta})}{n}\right\}.$$

Direct computation shows that $f(p,v,n)$ is non-decreasing in each dimension of v as below.

Despite two possible points such that $\frac{20}{3}\sqrt{\frac{\mathbb{V}(p,v)\log(\frac{1}{\delta})}{n}} = \frac{400}{9}\cdot\frac{H\log(\frac{1}{\delta})}{n}$,

$$\begin{aligned}\frac{\partial f}{\partial v(s)} &= p(s) + \frac{20}{3}\mathbb{I}\left[\frac{20}{3}\sqrt{\frac{\mathbb{V}(p,v)\log(\frac{1}{\delta})}{n}} \geq \frac{400}{9}\frac{H\log(\frac{1}{\delta})}{n}\right]\frac{p(s)(v(s)-pv)\log(\frac{1}{\delta})}{\sqrt{n\mathbb{V}(p,v)\log(\frac{1}{\delta})}} \\ &\geq \min\{p(s) + p(s)\frac{(v(s)-pv)}{H}, p(s)\} \\ &= 0. \end{aligned} \tag{44}$$

Fix h,k. In the case $N_h^k(s,a) \leq 2$, $Q_h^k(s,a) = H-h+1 \geq Q_h^*(s,a)$ and $_h^k(s) = H-h+1 \geq V_h^*(s)$ for any proper (s,a,h). Assume $N_h^k(s,a) > 2$ and $Q_{h+1}^k \geq Q_{h+1}^*$. It then follows that $V_{h+1}^k \geq V_{h+1}^*$. According to the update rule in (17), we either have that $Q_h^k(s,a) = H-h+1$ or

$$\begin{aligned}&Q_h^k(s,a) \\ &= \hat{r}_h^k(s,a) + \hat{P}_{s,a,h}^k V_{h+1}^k + c_1\sqrt{\frac{\mathbb{V}(\hat{P}_{s,a,h}^k, V_{h+1}^k)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + c_2\sqrt{\frac{\left(\hat{\sigma}_h^k(s,a) - (\hat{r}_h^k(s,a))^2\right)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + c_3\frac{H\log(\frac{1}{\delta})}{N_h^k(s,a)} \\ &\geq \hat{r}_h^k(s,a) + 2\sqrt{2}\sqrt{\frac{\left(\hat{\sigma}_h^k(s,a) - (\hat{r}_h^k(s,a))^2\right)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + \frac{28H\log(\frac{1}{\delta})}{3N_h^k(s,a)} + f(\hat{P}_{s,a,h}^k, V_{h+1}^k, N_h^k(s,a)) \\ &\geq \hat{r}_h^k(s,a) + 2\sqrt{2}\sqrt{\frac{\left(\hat{\sigma}_h^k(s,a) - (\hat{r}_h^k(s,a))^2\right)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + \frac{28H\log(\frac{1}{\delta})}{3N_h^k(s,a)} + f(\hat{P}_{s,a,h}^k, V_{h+1}^*, N_h^k(s,a)). \end{aligned} \tag{45}$$

for any (s,a).

By Lemma 13, and recalling the definition of $\sigma_h^k(s,a)$, we have that,

$$\mathbb{P}\left[|(\hat{P}_{s,a,h}^k - P_{s,a,h})V_{h+1}^*| > 2\sqrt{\frac{\mathbb{V}(\hat{P}_{s,a,h}^k, V_{h+1}^*)\iota}{N_h^k(s,a)}} + \frac{14\iota}{3n^k(s,a)}\right]$$

$$\leq \mathbb{P}\left[|(\hat{P}^k_{s,a,h} - P_{s,a,h})V^*_{h+1}| > \sqrt{\frac{2\mathbb{V}(\hat{P}^k_{s,a,h}, V^*_{h+1})\iota}{N^k_h(s,a)-1}} + \frac{7\iota}{3N^k_h(s,a)-1}\right]$$
$$\leq 2\delta \tag{46}$$

and

$$\mathbb{P}\left[|\hat{r}^k_h(s,a) - r(s,a)| > 2\sqrt{\frac{\left(\hat{\sigma}^k_h(s,a) - (\hat{r}^k_h(s,a))^2\right)\iota}{N^k_h(s,a)}} + \frac{28H\iota}{3N^k_h(s,a)}\right]$$
$$\leq 2\delta, \tag{47}$$

Note that (46) implies that $f(\hat{P}^k_{s,a,h}, V^*_{h+1}, N^k_h(s,a)) \geq P_{s,a,h}V^*_{h+1}$.

Therefore, with probability $1-4\delta$, $Q^k_h(s,a) \geq r_h(s,a) + P_{s,a,h}V^*_{h+1} = Q^*_h(s,a)$. By induction, we learn that with probability $1-4SAHK\delta$, $Q^k_h(s,a) \geq Q^*_h(s,a)$ for any proper (s,a,h,k). It then follows $V^k_h = \max_a Q^k_h(s,a) \geq \max_a Q^*_h(s,a) \geq V^*_h(s)$. The proof is completed.

□

C.2 Regret decomposition

Recall the definition that $\pi^k_h(s) = \arg\max_a Q^k_h(s,a)$. With probability $1-2SAHK\delta$,

$$\begin{aligned}
\text{Regret}(K) :&= \sum_{k=1}^K (V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1)) \\
&\leq \sum_{k=1}^K \left(V^k_1(s^k_1) - V^{\pi^k}_1(s^k_1)\right) \\
&\leq \sum_{k=1}^K \sum_{h=1}^H \left((\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})V^k_{h+1} + b^k_h(s^k_h,a^k_h)\right) + \sum_{k=1}^K \sum_{h=1}^H (P_{s^k_h,a^k_h,h} - \mathbf{1}_{s^k_{h+1}})V^k_{h+1} \\
&\qquad + \sum_{k=1}^K \left(\sum_{h=1}^H \hat{r}^k_h(s^k_h,a^k_h) - V^{\pi^k}_1(s^k_1)\right),
\end{aligned} \tag{48}$$

where $b^k_h(s^k_h,a^k_h) := c_1\sqrt{\frac{\mathbb{V}(\hat{P}^k_{s^k_h,a^k_h,h}, V^k_{h+1})\log(\frac{1}{\delta})}{N^k_h(s^k_h,a^k_h)}} + c_2\sqrt{\frac{\left(\hat{\sigma}^k_h(s,a) - (\hat{r}^k_h(s,a))^2\right)\log(\frac{1}{\delta})}{N^k_h(s^k_h,a^k_h)}} + c_3\frac{H\log(\frac{1}{\delta})}{N^k_h(s^k_h,a^k_h)}$. Here the first inequality is by Lemma 15, and the second inequality is by the Lemma below:

Lemma 16. *For each $k \in [K]$,*

$$V^k_1(s^k_1) \leq \sum_{h=1}^H \left((\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})V^k_{h+1} + b^k_h(s^k_h,a^k_h) + r_h(s^k_h,a^k_h) + (P_{s^k_h,a^k_h,h} - \boldsymbol{1}_{s^k_{h+1}})V^k_{h+1}\right).$$

Proof. By definition, for each $h \in [H]$

$$\begin{aligned}
V^k_h(s^k_h) &\leq r_h(s^k_h,a^k_h) + \hat{P}^k_{s^k_h,a^k_h,h}V^k_{h+1} + b^k_h(s^k_h,a^k_h) \\
&= (\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})V^k_{h+1} + b^k_h(s^k_h,a^k_h) + r_h(s^k_h,a^k_h) + (P_{s^k_h,a^k_h,h} - \mathbf{1}_{s^k_{h+1}})V^k_{h+1} + V^k_{h+1}(s^k_{h+1})
\end{aligned}$$

Taking sum over $h \in [H]$, we have that

$V^k_1(s^k_1)$

$$\leq \sum_{h=1}^{H} \left((\hat{P}^k_{s^k_h, a^k_h, h} - P_{s^k_h, a^k_h, h}) V^k_{h+1} + b^k_h(s^k_h, a^k_h) + r_h(s^k_h, a^k_h) + (P_{s^k_h, a^k_h, h} - \mathbf{1}_{s^k_{h+1}}) V^k_{h+1} \right) + V^k_{H+1}(s^k_{H+1}). \quad (49)$$

The proof is completed because $V^k_{H+1} = \mathbf{0}$. □

Define $T_1 = \sum_{k=1}^{K} \sum_{h=1}^{H} \left((\hat{P}^k_{s^k_h, a^k_h, h} - P_{s^k_h, a^k_h, h}) V^k_{h+1} \right)$, $T_2 = \sum_{k=1}^{K} \sum_{h=1}^{H} b^k_h(s^k_h, a^k_h)$, $T_3 = \sum_{k=1}^{K} \sum_{h=1}^{H} (P_{s^k_h, a^k_h, h} - \mathbf{1}_{s^k_{h+1}}) V^k_{h+1}$ and $T_4 = \sum_{k=1}^{K} \left(\sum_{h=1}^{H} \hat{r}^k_h(s^k_h, a^k_h) - V^{\pi^k}_1(s^k_1) \right)$.

We can easily bound T_2, T_3 and T_4 as in the following.

Bound of T_2 By definition, we can write

$$\begin{aligned} T_2 &= \sum_{k=1}^{K} \sum_{h=1}^{H} b^k_h(s^k_h, a^k_h) \\ &= \frac{460}{9} \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\frac{\mathbb{V}(\hat{P}^k_{s^k_h, a^k_h, h}, V^k_{h+1}) \log(\frac{1}{\delta})}{N^k_h(s^k_h, a^k_h)}} + 2\sqrt{2} \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\frac{\left(\hat{\sigma}^k_h(s^k_h, a^k_h) - (\hat{r}^k_h(s^k_h, a^k_h))^2 \right) \log(\frac{1}{\delta})}{N^k_h(s^k_h, a^k_h)}} \\ &\qquad + \frac{544}{9} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H \log(\frac{1}{\delta})}{N^k_h(s^k_h, a^k_h)}. \end{aligned} \quad (50)$$

Using Cauchy's inequality and Lemma 14, we obtain

$$\begin{aligned} T_2 &\leq \frac{460}{9} \sqrt{2SAH \log_2(K) \log(\frac{1}{\delta}) \sum_{k,h} \mathbb{V}(\hat{P}^k_{s^k_h, a^k_h, h}, V^k_{h+1})} \\ &\qquad + 4\sqrt{SAH \log_2(K) \log(\frac{1}{\delta})} \sqrt{\sum_{k,h} \left(\hat{\sigma}^k_h(s^k_h, a^k_h) - (\hat{r}^k_h(s^k_h, a^k_h))^2 \right)} + \frac{1088}{9} SAH^2 \log_2(K) \log(\frac{1}{\delta}) \\ &= \frac{460}{9} \sqrt{2SAH \log_2(K) \log(\frac{1}{\delta}) T_5} \\ &\qquad + 4\sqrt{SAH \log_2(K) \log(\frac{1}{\delta})} \sqrt{\sum_{k,h} \left(\hat{\sigma}^k_h(s^k_h, a^k_h) - (\hat{r}^k_h(s^k_h, a^k_h))^2 \right)} + \frac{1088}{9} SAH^2 \log_2(K) \log(\frac{1}{\delta}) \\ &\leq \frac{460}{9} \sqrt{2SAH \log_2(K) \log(\frac{1}{\delta}) T_5} \\ &\qquad + 4\sqrt{SAH^2 \log_2(K) \log(\frac{1}{\delta})} \sqrt{\sum_{k,h} \hat{r}^k_h(s^k_h, a^k_h)} + \frac{1088}{9} SAH^2 \log_2(K) \log(\frac{1}{\delta}), \end{aligned} \quad (51)$$

where we define $T_5 = \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{V}(\hat{P}^k_{s^k_h, a^k_h, h}, V^k_{h+1})$, and the last inequality is by the fact that $\left(\hat{\sigma}^k_h(s^k_h, a^k_h) - (\hat{r}^k_h(s^k_h, a^k_h))^2 \right) \leq H \hat{r}^k_h(s, a)$ for any proper (s, a, h, k).

We bound $\sum_{k,h} \hat{r}^k_h(s^k_h, a^k_h)$ by means of the lemma below.

Lemma 17. *With probability $1 - 2SAHK\delta$, it holds that*

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{h=1}^{H} \left| \hat{r}^k_h(s^k_h, a^k_h) - r_h(s^k_h, a^k_h) \right| \\ &\leq 4SAH^2 + 4\sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H \log(\frac{1}{\delta})}{N^k_h(s^k_h, a^k_h)}} \cdot \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} r_h(s^k_h, a^k_h)} + 24 \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H \log(\frac{1}{\delta})}{N^k_h(s^k_h, a^k_h)}. \end{aligned}$$

Proof of Lemma 17. In view of Lemma 13, with probability $1-2SAHK\delta$ we have

$$\hat{r}_h^k(s,a) - r_h(s,a) \le 2\sqrt{2}\sqrt{\frac{\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + \frac{28H\log(\frac{1}{\delta})}{3N_h^k(s,a)} \tag{52}$$

$$\le 2\sqrt{2}\sqrt{\frac{H\hat{r}_h^k(s,a)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + \frac{28H\log(\frac{1}{\delta})}{3N_h^k(s,a)} \tag{53}$$

for any proper (s,a,h,k) be such that $N_h^k(s,a) > 2$. Solve the inequality above, we can obtain

$$|\hat{r}_h^k(s,a) - r_h(s,a)| \le 4\sqrt{\frac{Hr_h(s,a)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + 24\frac{H\log(\frac{1}{\delta})}{N_h^k(s,a)}. \tag{54}$$

It is then seen that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right| \\
&\le 4SAH^2 + \sum_{k=1}^{K}\sum_{h=1}^{H}\left(4\sqrt{\frac{Hr_h(s_h^k,a_h^k)\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}} + 24\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}\right) \\
&\le 4SAH^2 + 4\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}r_h(s_h^k,a_h^k)} + 24\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}.
\end{aligned}$$

Here, the second inequality is due to Cauchy's inequality and the third inequality is a consequence of Lemma 14. We have thus completed the proof. □

With Lemma 17 and Lemma 14 in place, and noting that $\sum_{k=1}^{K}\sum_{h=1}^{H} r_h(s_h^k,a_h^k) \le KH$, with probability $1-2SAHK\delta$,

$$T_2 \le 61\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 145SAH^2\log_2(K)\log(\frac{1}{\delta}) \tag{55}$$

Bound of T_3 By virtue of Lemma 9, we see that with probability $1-10SAH^2K^2\delta$,

$$T_3 \le 2\sqrt{2}\cdot\sqrt{T_6\log(\frac{1}{\delta})} + \log(\frac{1}{\delta}) + 2H\log(\frac{1}{\delta}) \le 2\sqrt{2}\cdot\sqrt{T_6\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}), \tag{56}$$

where $T_6 = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k)$.

Bound of T_4 Note that

$$T_4 = \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right) + \sum_{k=1}^{K}\left(\sum_{h=1}^{H}r_h(s_h^k,a_h^k) - \mathbb{V}_1^{\pi^k}(s_1^k)\right). \tag{57}$$

For the first term, using Lemma 17 and noting that $\sum_{k=1}^{K}\sum_{h=1}^{H} r_h(s_h^k,a_h^k) \le KH$, with probability $1-2SAHK\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}(\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)) \\
&\le 4SAH^2 + 4\sqrt{2\log_2(K)SAH^2}\cdot\sqrt{KH} + 24\log_2(K)SAH^2\log(\frac{1}{\delta})
\end{aligned} \tag{58}$$

$$\leq 4\sqrt{2SAH^3K\log_2(K)\log(\frac{1}{\delta})} + 28SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{59}$$

Noting that $E_k := \sum_{h=1}^{H} r_h(s_h^k, a_h^k) - V_1^{\pi^k}(s_1^k)$ is a zero-mean random variable bounded by H, by Lemma 9, with probability $1-2\delta$, it holds that

$$\sum_{k=1}^{K} E_k \leq 2\sqrt{2}\cdot\sqrt{\sum_{k=1}^{K}\mathrm{Var}(E_k)\log(\frac{1}{\delta})} + 3H^2\log(\frac{SAHK}{\delta}) \leq 2\sqrt{2KH^2\log(\frac{1}{\delta})} + 3H^2\log(\frac{SAHK}{\delta}), \tag{60}$$

where $\mathrm{Var}(E_k)$ denoting teh variance of E_k conditioned on the $\mathcal{F}_1^k$.

Putting (59) and (60) together, we obtain that with probability $1-4SAHK\delta$,

$$T_4 \leq 6\sqrt{2SAH^3K\log_2(K)\log(\frac{1}{\delta})} + 31SAH^2\log_2(K)\log(\frac{SAHK}{\delta}). \tag{61}$$

Bound of T_5 and T_6 Direct computation gives that

$$\begin{aligned}
T_5 :&= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(\hat{P}^k_{s_h^k,a_h^k,h}, V_{h+1}^k) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left((\hat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h})(V_{h+1}^k)^2 + P_{s_h^k,a_h^k,h}(V_{h+1}^k)^2 - (\hat{P}^k_{s_h^k,a_h^k,h}V_{h+1}^k)^2\right) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}(\hat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h})(V_{h+1}^k)^2 + \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k})(V_{h+1}^k)^2 \\
&\qquad + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{V_h^k(s_h^k) - \hat{P}^k_{s_h^k,a_h^k,h}V_{h+1}^k, 0\} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}(\hat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h})(V_{h+1}^k)^2 + \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k})(V_{h+1}^k)^2 \\
&\qquad + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}b_h^k(s_h^k,a_h^k) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}r_h(s_h^k,a_h^k).
\end{aligned} \tag{62}$$

Let $T_7 = \sum_{k=1}^{K}\sum_{h=1}^{H}(\hat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h})(V_{h+1}^k)^2$ and $T_8 = \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k})(V_{h+1}^k)^2$.

Using Freedman's inequality and the fact that $\mathrm{Var}(X^2) \leq 4H^2\mathrm{Var}(X)$ for any random variable X with support on $[-H, H]$, we have that

$$T_8 \leq 2\sqrt{2}\sqrt{4H^2T_6\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}). \tag{63}$$

In a similar way, we can bound T_6 as

$$\begin{aligned}
T_6 :&= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k}\right)(V_{h+1}^k)^2 + \sum_{k=1}^{K}\sum_{h=1}^{H}(V_{h+1}^k)^2(s_{h+1}^k) - \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k,h}V_{h+1}^k)^2 \\
&\leq 2\sqrt{2}\sqrt{4H^2T_6\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{V_h^k(s_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^k, 0\}
\end{aligned}$$

$$\leq 2\sqrt{2}\sqrt{4H^2T_6\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}) + 2HT_2 + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})V^k_{h+1}, 0\} + 2KH^2. \tag{64}$$

Let $T_9 = \sum_{k=1}^{K}\sum_{h=1}^{H}\max\{(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})V^k_{h+1}, 0\}$.

C.3 Bounds of the error terms

Now we deal with T_1, T_7 and T_9. We first recall the definition.

$$T_1 = \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)V^k_{h+1};$$
$$T_7 = \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)(V^k_{h+1})^2;$$
$$T_9 = \sum_{k=1}^{K}\sum_{h=1}^{H}\max\{(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})V^k_{h+1}, 0\}.$$

Recall $B = 4000\log_3^2(K)\log(3SAH)\log(\frac{1}{\delta})$. By the update rule (18), V^k_{h+1} is determined by the $\{\hat{P}^k_{s,a,h'}\}_{(h+1\leq h'\leq H,s,a)}$ and $\{I^k_{s,a,h'}\}_{h+1\leq h'\leq H,s,a}$. Using Lemma 6 with $\mathcal{X}_{h+1} = \{V^k_{h+1}\}_{k=1}^K \cup \{0\}, \{(V^k_{h+1})^2/H\}_{k=1}^K \cup \{0\}$ and $\{V^k_{h+1}\}_{k=1}^K \cup \{0\}$, and noting that $\mathrm{Var}(X^2) \leq 4\|X\|_\infty^2\mathrm{Var}(X)$, we have that with probability $1 - 30SAH^2K^2\delta$,

$$T_1 \leq \sqrt{BSAH\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1}) + BSAH^2} = \sqrt{BSAHT_6} + BSAH^2; \tag{65}$$
$$T_7 \leq H\sqrt{4BSAH\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1}) + 4BSAH^3} = H\sqrt{4BSAHT_6} + 4BSAH^3; \tag{66}$$
$$T_9 \leq \sqrt{BSAH\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1}) + BSAH^2} = \sqrt{BSAHT_6} + BSAH^2. \tag{67}$$

C.4 Putting all pieces together

Rewrite (55),(56),(61),(62),(64),(63),(65),(66) and (67) as below. With probability $1 - 100SAH^2K\delta$,

$$T_2 \leq 61\sqrt{BSAHT_5} + 145BSAH^2; \tag{68}$$
$$T_3 \leq \sqrt{8BT_6} + 3HB; \tag{69}$$
$$T_4 \leq 9\sqrt{BSAH^3K} + 31BSAH^2; \tag{70}$$
$$T_5 \leq T_7 + T_8 + 2HT_2 + 2KH^2; \tag{71}$$
$$T_6 \leq \sqrt{32BH^2T_6} + 2HT_2 + 2HT_9 + 3H^2B + 2KH^2; \tag{72}$$
$$T_8 \leq \sqrt{32BH^2T_6} + 3BH^2; \tag{73}$$
$$T_1 \leq \sqrt{BSAHT_6} + BSAH^2; \tag{74}$$
$$T_7 \leq H\sqrt{4BSAHT_6} + 4BSAH^3; \tag{75}$$

$$T_9 \le \sqrt{BSAHT_6} + BSAH^2. \tag{76}$$

To solve the inequalities (68) to (76), we use the fact that $a \le \sqrt{bc} + d$ implies that $a \le \epsilon b + \frac{1}{2\epsilon}c + d$ for any $b, c \ge 0, a, d \in \mathbb{R}$ and any $\epsilon > 0$. Using this arguments, we have

$$\begin{aligned}
HT_2 &\le \epsilon T_5 + (\frac{1}{2\epsilon} + 1)61BSAH^3 + 145BSAH^3; \\
T_6 &\le \epsilon T_6 + 2HT_2 + 2HT_9 + (3 + \frac{32}{\epsilon})BH^2 + 2KH^2; \\
HT_9 &\le \epsilon T_6 + \left(\frac{1}{2\epsilon} + 1\right) BSAH^3; \\
T_8 &\le \epsilon T_6 + \left(\frac{16}{\epsilon} + 1\right) BH^2; \\
T_7 &\le \epsilon T_6 + \left(\frac{2}{\epsilon} + 4\right) BSAH^3.
\end{aligned}$$

Then we have that

$$\begin{aligned}
T_5 &\le T_7 + T_8 + 2HT_2 + 2KH^2 \le 2\epsilon T_5 + 2\epsilon T_6 + \left(\frac{100}{\epsilon} + 300\right) BSAH^3 + 2KH^2; \\
T_6 &\le 3\epsilon T_6 + 2\epsilon T_5 + \left(\frac{200}{\epsilon} + 300\right) BSAH^3 + 2KH^2.
\end{aligned} \tag{77}$$

Choosing $\epsilon = \frac{1}{20}$, we learn that $T_5 + T_6 \le O(BSAH^3 + KH^2)$. Recall that $K \ge SAHB$. We then have that $T_5, T_6 = O(KH^2 + BSAH^3) = O(KH^2)$. As a result, we have that $T_1 = O(\sqrt{BSAH^3K})$, $T_7, T_8 = O(\sqrt{BSAH^5K})$, $T_2 = O(\sqrt{BSAH^3K})$ and $T_3 = O(\sqrt{BKH^2})$. We then conclude that the total regret is bounded by $\tilde{O}(\sqrt{SAH^3K\log(\frac{1}{\delta})})$. The proof of Theorem 1 is completed by replacing δ with $\frac{\delta}{100SAH^2K}$.

D Proof of the value-based regret bound (proof of Theorem 2)

We recall the definition of T_1-T_9 listed as below. We will continue by proving tighter bounds for some of these terms with respect to v^*.

$$\begin{aligned}
T_1 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)V^k_{h+1}; \\
T_2 &= \sum_{k=1}^{K}\sum_{h=1}^{H} b^k_h(s^k_h, a^k_h); \\
T_3 &= \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s^k_h,a^k_h,h} - \mathbf{1}_{s^k_{h+1}})V^k_{h+1}; \\
T_4 &= \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}^k_h(s^k_h, a^k_h) - V_1^{\pi^k}(s^k_1)\right); \\
T_5 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(\hat{P}^k_{s^k_h,a^k_h,h}, V^k_{h+1}); \\
T_6 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1});
\end{aligned}$$

$$T_7 = \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h}\right)(V^k_{h+1})^2;$$
$$T_8 = \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k})(V^k_{h+1})^2;$$
$$T_9 = \sum_{k=1}^{K}\sum_{h=1}^{H}\max\left\{(\hat{P}^k_{s_h^k,a_h^k,h} - P_{s_h^k,a_h^k,h})V^k_{h+1}, 0\right\}.$$

Also recall (68)-(76). We will provide refined analysis for the bound of T_2, T_4, T_5 and T_6, and leave other bounds invariant.

Bound of T_2 Recall that in (51), we show that

$$T_2 \le \frac{460}{9}\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 4\sqrt{SAH^2\log_2(K)\log(\frac{1}{\delta})}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\hat{r}^k_h(s_h^k,a_h^k)} + \frac{1088}{9}SAH^2\log_2(K)\log(\frac{1}{\delta}).$$

By definition of T_4, and the fact that $\sum_{k=1}^{K}V_1^*(s_1^k) \le 3Kv^* + H\log(\frac{1}{\delta})$ with probability $1-\delta$ (Lemma 11), it then holds that

$$T_2 \le \frac{460}{9}\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 4\sqrt{SAH^2\log_2(K)\log(\frac{1}{\delta})}\sqrt{T_4+3Kv^*} + 130SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{78}$$

Bound of T_4 Note that

$$\begin{aligned}T_4 &= \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}^k_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right)\\ &= \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}^k_h(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right) + \sum_{k=1}^{K}\left(\sum_{h=1}^{H}r_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right).\end{aligned} \tag{79}$$

Let $\check{T}_1 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}^k_h(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right)$ and $\check{T}_2 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H}r_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right)$ By Lemma 17 and Lemma 14, with probability $1-2SAHK\delta$,

$$\begin{aligned}\check{T}_1 &\le 4\sqrt{2SAH^2\log_2(K)}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}r_h(s_h^k,a_h^k)} + 52SAH^2\log_2(K)\log(\frac{1}{\delta})\\ &\le 4\sqrt{2SAH^2\log_2(K)}\cdot\sqrt{\check{T}_2+3Kv^*} + 60SAH^2\log_2(K)\log(\frac{1}{\delta}).\end{aligned} \tag{80}$$

On the other hand, by Lemma 9, with probability $1-2SAHK\delta$,

$$\check{T}_2 \le 2\sqrt{2\sum_{k=1}^{K}\mathbb{E}_{\pi^k,s_1\sim\mu_1}\left[\left(\sum_{h=1}^{H}r_h(s_h,a_h)\right)^2\right]\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta})$$
$$2\sqrt{2H\sum_{k=1}^{K}\mathbb{E}_{\pi^k,s_1\sim\mu_1}\left[\sum_{h=1}^{H}r_h(s_h,a_h)\right]\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta})$$

$$\leq 2\sqrt{2KHv^* \log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}) \tag{81}$$

$$\leq 2Kv^* + 5H\log(\frac{1}{\delta}). \tag{82}$$

Combining (80), (81) with (82), with probability $1 - 4SAHK\delta$,

$$\check{T}_1 \leq 12\sqrt{SAH^2Kv^* \log_2(K)\log(\frac{1}{\delta})} + 80SAH^2\log_2(K)\log(\frac{1}{\delta})$$

$$\check{T}_2 \leq 2\sqrt{2KHv^* \log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}).$$

As a result, we have that

$$T_4 \leq 14sqrtSAH^2Kv^* \log_2(K)\log(\frac{1}{\delta}) + 83SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{83}$$

Bound of T_5 Recall that in (62), we show that

$$T_5 \leq \sum_{k=1}^{K}\sum_{h=1}^{H}(\hat{P}^k_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h})(V^k_{h+1})^2 + \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s^k_h,a^k_h,h} - \mathbf{1}_{s^k_{h+1}})(V^k_{h+1})^2$$
$$+ 2H\sum_{k=1}^{K}\sum_{h=1}^{H} b^k_h(s^k_h,a^k_h) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H} r_h(s^k_h,a^k_h).$$

With (82), with probability $1 - 4SAHK\delta$,

$$T_5 \leq T_7 + T_8 + 2HT_2 + 4HKv^* + 10H^2\log(\frac{1}{\delta}). \tag{84}$$

Bound of T_6 Recall (64)

$$T_6 \leq 2\sqrt{8T_6\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{V^k_h(s^k_h) - P_{s^k_h,a^k_h,h}V^k_{h+1}, 0\}$$

$$\leq 2\sqrt{8T_6\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}) + 2HT_2 + 2HT_9 + 2H\sum_{k=1}^{K}\sum_{h=1}^{H} r_h(s^k_h,a^k_h)$$

$$\leq\leq 2\sqrt{8T_6\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{V^k_h(s^k_h) - P_{s^k_h,a^k_h,h}V^k_{h+1}, 0\}$$

$$\leq 2\sqrt{8T_6\log(\frac{1}{\delta})} + 2HT_9 + 4HKv^* + 6H^2\log(\frac{1}{\delta}) + 2HT_2. \tag{85}$$

Putting all together Assume that $K \geq \frac{BSAH^2}{v^*}$. Then $\sqrt{BSAH^2Kv^*} \geq BSAH^2$. Solving (78),(56),(83),(84),(137),(63),(65 and (67), we have that, with probability $1-100SAH^2K\delta$, $T_6 = O(BHKv^* + BSAH^3)$, $T_1 = O(\sqrt{BSAH^2Kv^*})$, $T_7, T_8 = O(\sqrt{BSAH^4Kv^*})$, $T_5 = O(BHKv^*)$, $T_2 = O(\sqrt{BSAH^2Kv^*})$ and $T_3 = O(\sqrt{BHKv^*})$. We then conclude that the total regret is bounded by $O(\sqrt{BSAH^2Kv^*})$. In the case $K \leq \frac{BSAH^2}{v^*}$, the regret bound is trivially $O(Kv^*)$. The proof is completed by replacing δ with $\frac{\delta}{100SAH^2K}$.

E Proof of Corollary 1

In this section, we will use r to denote the negative reward, that is, $r = -c$. Recall (32):

$$Q_h(s,a) \leftarrow \max\{\min\left\{\hat{r}_h(s,a) + \hat{P}_{s,a,h}V_{h+1} + b_h(s,a), 0\right\}, -H\}.$$

Recall the definition of T_1-T_9. We note that the analysis of T_1, T_3, T_7, T_8 and T_9 in Appendix D applies for the case the reward function is negative. So it suffices to provide bounds for T_2, T_4, T_5 and T_6 with respect to c^*.

Bound of T_2 Recall that

$$\begin{aligned}
T_2 &= \sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{460}{9}\sqrt{\frac{\mathbb{V}(\hat{P}^k_{s_h^k,a_h^k,h}, V_{h+1}^k)\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}}\right. \\
&\qquad \left. + 2\sqrt{2}\sqrt{\frac{\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right)\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}} + \frac{544}{9}\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}\right). \qquad (86)
\end{aligned}$$

For the first and third term in right hand side of (86), we can use Cauchy's inequality to obtain that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}(\hat{P}^k_{s_h^k,a_h^k,h}, V_{h+1}^k)\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}} \\
&\le \sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})\sum_{k,h}\mathbb{V}(\hat{P}^k_{s_h^k,a_h^k,h}, V_{h+1}^k)} \\
&= \sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} \qquad (87)
\end{aligned}$$

and

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)} \le 2SAH^2\log_2(K)\log(\frac{1}{\delta}). \qquad (88)$$

For the second term, noting that

$$\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k)^2)\right) \le -H\hat{r}_h^k(s_h^k,a_h^k),$$

we have

$$\begin{aligned}
&\sqrt{\frac{\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right)\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}} \\
&\le \sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})}\sqrt{H\sum_{k,h}-\hat{r}_h^k(s_h^k,a_h^k)} \\
&\le \sqrt{2SAH^2\log_2(K)\log(\frac{1}{\delta})}\sqrt{T_4 + 3Kc^* + \sum_{k=1}^{K}(-V_1^{\pi^k}(s_1^k) + V_1^*(s_1^k)) + \sum_{k=1}^{K}(-V_1^*(s_1^k) - 3c^*)}. \qquad (89)
\end{aligned}$$

By Lemma 11, with probability $1-\delta$,

$$\sum_{k=1}^{K}-V_1^*(s_1^k) \le 3Kc^* + H\log(\frac{1}{\delta}).$$

On the other hand, we note that

$$\sum_{k=1}^{K}(-V_1^{\pi^k}(s_1^k) + V_1^*(s_1^k)) = \text{Regret}(K) = T_1 + T_2 + T_3 + T_4. \qquad (90)$$

Putting all together, we obtain that, with probability $1-\delta$,

$$T_2 \le 90\sqrt{SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 4\sqrt{SAH^2\log_2(K)\log(\frac{1}{\delta})}\sqrt{T_1+T_2+T_3+2T_4+3Kv^*} + 130SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{91}$$

Bound of T_4 Recall that

$$\begin{aligned} T_4 &= \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}_h^k(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right) \\ &= \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right) + \sum_{k=1}^{K}\left(\sum_{h=1}^{H} r_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right). \end{aligned} \tag{92}$$

Also recall that $\check{T}_1 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right)$ and $\check{T}_2 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H} r_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right)$. We continue with a lemma to bound the empirical reward for negative reward function.

Lemma 18. *With probability $1-2SAHK\delta$, it holds that*

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right| \\ &\le 4SAH^2 + 4\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}-r_h(s_h^k,a_h^k)} + 24\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}. \end{aligned}$$

The proof of Lemma 18 is basically the same as that of Lemma 17, except for that r is replaced with $-r$.

By Lemma 18 and Lemma 14, with probability $1-3SAHK\delta$,

$$\begin{aligned} |\check{T}_1| &\le 4\sqrt{2SAH^2\log_2(K)}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}-r_h(s_h^k,a_h^k)} + 52SAH^2\log_2(K)\log(\frac{1}{\delta}) \\ &\le 4\sqrt{2SAH^2\log_2(K)}\cdot\sqrt{\check{T}_2 + 3Kc^* + \sum_{k=1}^{K}(-V_1^*(s_1^k) - 3c^*)} + 52SAH^2\log_2(K)\log(\frac{1}{\delta}) \\ &\le 4\sqrt{2SAH^2\log_2(K)}\cdot\sqrt{\check{T}_2 + 3Kc^*} + 60SAH^2\log_2(K)\log(\frac{1}{\delta}), \end{aligned} \tag{93}$$

where in the last line we use the fact

$$\sum_{k=1}^{K} -V_1^*(s_1^k) \le 3Kc^* + H\log(\frac{1}{\delta}) \tag{94}$$

with probability $1-\delta$ (Lemma 11).

On the other hand, by Lemma 9 and (94), with probability $1-3SAHK\delta$,

$$\begin{aligned} |\check{T}_2| &\le 2\sqrt{2\sum_{k=1}^{K}\mathbb{E}_{\pi^k,}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h)\right)^2 | s_1 = s_1^k\right]\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}) \\ &2\sqrt{2H\sum_{k=1}^{K}\mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H} -r_h(s_h,a_h)|s_1 = s_1^k\right]\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}) \end{aligned}$$

$$\leq 2\sqrt{2H\left(\sum_{k=1}^{K}\left(-V_1^{\pi^k}(s_1^k)+V_1^*(s_1^k)\right)+\sum_{k=1}^{K}\left(-V_1^*(s_1^k)-3c^*\right)+3Kc^*\right)\log(\frac{1}{\delta})}+3H\log(\frac{1}{\delta}) \tag{95}$$

$$\leq 3Kc^*+T_1+T_2+T_3+T_4+9H\log(\frac{1}{\delta}). \tag{96}$$

Combining (93), (95) with (96), with probability $1-4SAHK\delta$,

$$|\check{T}_1|\leq 16\sqrt{SAH^2(Kc^*+T_1+T_2+T_3+T_4)\log_2(K)\log(\frac{1}{\delta})}+200SAH^2\log_2(K)\log(\frac{1}{\delta})$$

$$|\check{T}_2|\leq 2\sqrt{2H(3Kc^*+T_1+T_2+T_3+T_4)\log(\frac{1}{\delta})}+9H\log(\frac{1}{\delta}).$$

As a result, we have that

$$|T_4|\leq 22\sqrt{SAH^2(Kc^*+T_1+T_2+T_3+T_4)\log_2(K)\log(\frac{1}{\delta})}+209SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{97}$$

Bound of T_5 Using the arguments in (62), and noting the update rule (32), we have

$$T_5\leq\sum_{k=1}^{K}\sum_{h=1}^{H}(\hat{P}^k_{s_h^k,a_h^k,h}-P_{s_h^k,a_h^k,h})(V_{h+1}^k)^2+\sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k)^2+2H\sum_{k=1}^{K}\sum_{h=1}^{H}-r_h(s_h^k,a_h^k).$$

Recall that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}-r_h(s_h^k,a_h^k)=-\check{T}_2-\sum_{k=1}^{K}V_1^{\pi^k}(s_1)\leq-\check{T}_2+\sum_{k=1}^{K}V_1^*(s_1^k). \tag{98}$$

By (94), with probability $1-5SAHK\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}-r_h(s_h^k,a_h^k)\leq 2\sqrt{2H(3Kc^*+T_1+T_2+T_3+T_4)\log(\frac{1}{\delta})}+3Kc^*+10H\log(\frac{1}{\delta}). \tag{99}$$

As a result, we have that

$$T_5\leq T_7+T_8+2HT_2+4\sqrt{2H^3(3Kc^*+T_1+T_2+T_3+T_4)\log(\frac{1}{\delta})}+6HKc^*+20H^2\log(\frac{1}{\delta}). \tag{100}$$

with probability $1-5SAHK\delta$.

Bound of T_6 Using the arguments in (62), (94) and (98), and noting the update rule (32), with probability $1-3SAHK\delta$

$$\begin{aligned}T_6&\leq 2\sqrt{8T_6\log(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta})+2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{P_{s_h^k,a_h^k,h}V_{h+1}^k-V_h^k(s_h^k),0\}\\&\leq 2\sqrt{8T_6\log(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta})+2HT_9+2H\sum_{k=1}^{K}\sum_{h=1}^{H}-r_h(s_h^k,a_h^k)\\&\leq 2\sqrt{8T_6\log(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta})+2HT_9\\&\qquad+2H\left(2\sqrt{2H(3Kc^*+T_1+T_2+T_3+T_4)\log(\frac{1}{\delta})}+3Kc^*+10H\log(\frac{1}{\delta})\right).\end{aligned} \tag{101}$$

Putting all together Solving (91),(56),(97),(100),(101),(63),(65),(66) and (67), we have that, with probability $1-100SAH^2K\delta$, $T_6 = O(HKc^*+BSAH^3)$, $T_1 = O(\sqrt{BSAH^2Kc^*}+BSAH^2)$, $T_7, T_8 = O(\sqrt{BSAH^4Kc^*}+BSAH^3)$, $T_5 = O(HKc^* + BSAH^2)$, $T_2 = O(\sqrt{BSAH^2Kc^*} + BSAH^2)$ and $T_3 = O(\sqrt{BHKc^*} + BSAH^2)$. We then conclude that the total regret is bounded by $O(\sqrt{BSAH^2Kc^*} + BSAH^2)$. On the other hand, the regret bound is trivially bounded by $O(K(H - c^*))$. The proof is completed by replacing δ with $\frac{\delta}{100SAH^2K}$.

F Proof of the variance-dependent regret bounds

F.1 Proof of Theorem 3

In this section, we will present the proof of Theorem 3. The proof contains two parts, where we respectively prove regret bounds of $\widetilde{O}\left(\min\{\sqrt{SAHK\mathrm{var}_1} + SAH^2, KH\}\right)$ and $\widetilde{O}\left(\min\{\sqrt{SAHK\mathrm{var}_2} + SAH^2, KH\}\right)$. Formally we have the following lemmas.

Lemma 19. *With probability exceeding $1-\delta$, the regret of Algorithm 1 is at most $\widetilde{O}(\min\{\sqrt{SAHK\mathrm{var}_1} + SAH^2, KH\})$*

Lemma 20. *With probability at least $1-\delta$, the regret of Algorithm 1 is at most $\widetilde{O}(\min\{\sqrt{SAHK\mathrm{var}_2} + SAH^2, KH\})$*

Putting the two regret bounds together and rescaling δ to $\delta/2$, we conclude the proof.

F.2 Proof of Lemma 19

Recall that

$$T_4 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\hat{r}_h^k(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right);$$
$$T_5 = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(\hat{P}_{s_h^k,a_h^k,h}, V_{h+1}^k);$$
$$T_6 = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k).$$

Recall that $B = 4000\log_3^2(K)\log(3SAH)\log(\frac{1}{\delta})$.

F.2.1 Bound of T_2

Recall in (55), we show that

$$T_2 \le \frac{460}{9}\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 4\sqrt{SAH\log_2(K)\log(\frac{1}{\delta})}\sqrt{\sum_{k,h}\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right)} + \frac{1088}{9}SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{102}$$

Define the variance of $R_h(s,a)$ as $v_h(s,a)$. We then have the following lemma.

Lemma 21. *With probability $1-4SAHK\delta$,*

$$\sum_{k,h}\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right) \le 6K\mathrm{var}_1 + 242SAH^3\log_2(K)\log(\frac{1}{\delta}). \tag{103}$$

Proof. We first control each $\hat{\sigma}_h^k(s_h^k,a_h^k)-(\hat{r}_h^k(s_h^k,a_h^k))^2$ with $v_h(s,a)$. Fix (s,a,h,k). Using Lemma 11, with probability $1-2\delta$,

$$N_h^k(s,a)\left(\hat{\sigma}_h^k(s_h^k,a_h^k)-(\hat{r}_h^k(s_h^k,a_h^k))^2\right)\le 3N_h^k v_h(s,a)+H^2\log(\frac{1}{\delta}). \tag{104}$$

Then we have that, with probability $1-2SAHK\delta$,

$$\begin{aligned}&\sum_{k,h}\left(\hat{\sigma}_h^k(s_h^k,a_h^k)-(\hat{r}_h^k(s_h^k,a_h^k))^2\right)\\ &\le 3\sum_{k,h}v_h(s_h^k,a_h^k)+\sum_{k,h}\frac{H^2\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}\le 3\sum_{k,h}v_h(s_h^k,a_h^k)+2SAH^3\log_2(K)\log(\frac{1}{\delta}).\end{aligned} \tag{105}$$

Now it suffices to control $\sum_{k,h}v_h(s_h^k,a_h^k)$. Let $\tilde{V}_h^k(s):=\mathbb{E}_{\pi^k}[\sum_{h'=h}^H v_{h'}(s_{h'},a_{h'})|s_h=s]$ be the value function with reward as $\{v_h(s,a)\}$ and policy π^k. Then $\tilde{V}_h^k(s,a)\le H^2$.

Then, by Lemma 9, with probability $1-2SAHK\delta$,

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H v_h(s_h^k,a_h^k)-\sum_{k=1}^K\tilde{V}_1^k(s_1^k)&=\sum_{k=1}^K\left(\sum_{h=1}^H\left(\mathbf{1}_{s_{h+1}^k}-P_{s_h^k,a_h^k,h}\right)\tilde{V}_{h+1}^k\right)\\ &\le 2\sqrt{2\sum_{k=1}^K\sum_{h=1}^H\mathbb{V}(P_{s_h^k,a_h^k,h},\tilde{V}_{h+1}^k)\log(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta}).\end{aligned} \tag{106}$$

On the other hand, using Lemma 9 again, we obtain that with probability $1-2SAHK\delta$

$$\begin{aligned}&\sum_{k=1}^K\sum_{h=1}^H\mathbb{V}(P_{s_h^k,a_h^k,h},\tilde{V}_{h+1}^k)\\ &=\sum_{k=1}^K\sum_{h=1}^H\left(P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k}\right)(\tilde{V}_{h+1}^k)^2\\ &\qquad+\sum_{k=1}^H\sum_{h=1}^H\left((\tilde{V}_{h+1}^k(s_{h+1}^k))^2-(\tilde{V}_h^k(s_h^k))^2\right)+\sum_{k=1}^K\sum_{h=1}^H\left((\tilde{V}_h^k(s_h^k))^2-(P_{s_h^k,a_h^k,h}\tilde{V}_{h+1}^k)^2\right)\\ &\le 2\sqrt{8H^4\sum_{k=1}^K\sum_{h=1}^H\mathbb{V}(P_{s_h^k,a_h^k,h},\tilde{V}_{h+1}^k)\log(\frac{1}{\delta})}+2H^2\sum_{k=1}^K\sum_{h=1}^H v_h(s_h^k,a_h^k)+3H^4\log(\frac{1}{\delta})\\ &\le 4H^2\sum_{k=1}^K\sum_{h=1}^H v_h(s_h^k,a_h^k)+42H^4\log(\frac{1}{\delta}).\end{aligned} \tag{107}$$

By (106) and (107), we learn that, with probability $1-4SAHK\delta$,

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^K v_h(s_h^k,a_h^k)&\le\sum_{k=1}^K\tilde{V}_1^k(s_1^k)+2\sqrt{8H^2\sum_{k=1}^K\sum_{h=1}^H v_h(s_h^k,a_h^k)\log\frac{1}{\delta}+84H^4\log^2(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta})\\ &\le 2\sum_{k=1}^K\tilde{V}_1^k(s_1^k)+80H^2\log(\frac{1}{\delta})\\ &\le 2K\mathrm{var}_1+80H^2\log(\frac{1}{\delta}).\end{aligned} \tag{108}$$

□

With Lemma 21 and (102), with probability $1-4SAHK\delta$,

$$T_2 \le \frac{460}{9}\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 12\sqrt{SAH\log_2(K)\log(\frac{1}{\delta})}\sqrt{2K\mathrm{var}_1} + 157SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{109}$$

F.2.2 Bound of T_4

Recall that $T_4 = \check{T}_1 + \check{T}_2$ where $\check{T}_1 = \sum_{k=1}^K \sum_{h=1}^H \left(\hat{r}_h^k(s_h^k, a_h^k) - r_h(s_h^k, a_h^k)\right)$ and $\check{T}_2 = \sum_{k=1}^K \left(\sum_{h=1}^H r_h(s_h^k, a_h^k) - V_1^{\pi^k}(s_1^k)\right)$.

We first bound $\check{T}_1$. By Lemma 12 and a union bound over all proper (s,a,h,k), with probability $1-2SAHK\delta$,

$$\hat{r}_h^k(s,a) - r_h(s,a) \le \sqrt{\frac{2v_h(s,a)\log(\frac{1}{\delta})}{N_h^k(s,a)}} + \frac{H\log(\frac{1}{\delta})}{N_h^k(s,a)}. \tag{110}$$

As a result, we have that

$$\begin{aligned}|\check{T}_1| &\le \sum_{k=1}^K\sum_{h=1}^H \left(\sqrt{\frac{2v_h(s_h^k,a_h^k)\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}} + \frac{H\log(\frac{1}{\delta})}{N_h^k(s_h^k,a_h^k)}\right) \\ &\le \sqrt{4SAH\log_2(K)\log(\frac{1}{\delta})}\cdot\sqrt{\sum_{k=1}^K\sum_{h=1}^H v_h(s_h^k,a_h^k)} + 2SAH^2\log_2(K)\log(\frac{1}{\delta}).\end{aligned} \tag{111}$$

By (108), with probability $1-4SAHK\delta$,

$$\sum_{k=1}^K\sum_{h=1}^H v_h(s_h^k,a_h^k) \le 2K\mathrm{var}_1 + 80H^2\log(\frac{1}{\delta}). \tag{112}$$

Then

$$|\check{T}_1| \le \sqrt{8SAHK\mathrm{var}_1\log_2(K)\log(\frac{1}{\delta})} + 20SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{113}$$

On the other hand, to bound $\check{T}_2$, we have that

$$\check{T}_2 = \sum_{k=1}^K\sum_{h=1}^H \left(\mathbf{1}_{s_{h+1}^k} - P_{s_h^k,a_h^k,h}\right)V_{h+1}^{\pi^k}. \tag{114}$$

Using Lemma 9, with probability $1-2SAHK\delta$,

$$\begin{aligned}|\check{T}_2| &\le 2\sqrt{2\sum_{k=1}^K\sum_{h=1}^H \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^{\pi^k})\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}) \\ &\le 2\sqrt{4\sum_{k=1}^K\sum_{h=1}^H (\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*) + \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^* - V_{h+1}^{\pi^k}))\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}).\end{aligned} \tag{115}$$

Continue the computation,

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

$$
\begin{aligned}
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(P_{s_h^k,a_h^k,h}(V_{h+1}^* - V_{h+1}^{\pi^k})^2 - (P_{s_h^k,a_h^k,h}(V_{h+1}^* - V_{h+1}^{\pi^k}))^2\right) \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k}\right)(V_{h+1}^* - V_{h+1}^{\pi^k})^2 \\
&\quad + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\left\{\left(V_h^*(s_h^k) - r_h(s_h^k,a_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^*\right) - \left(V_h^{\pi^k}(s_h^k) - r_h(s_h^k,a_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k}\right), 0\right\} \\
&\le 2\sqrt{8H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^* - V_{h+1}^{\pi^k})\log(\frac{1}{\delta})} \\
&\qquad\qquad + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\left(V_h^*(s_h^k) - r_h(s_h^k,a_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^*\right) + 3H^2\log(\frac{1}{\delta}) \qquad (116)
\end{aligned}
$$

Here (116) holds with probability $1 - 2SAHK\delta$ because of Lemma 9 and Lemma 10.

Then we consider to bound

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(V_h^*(s_h^k) - r_h(s_h^k,a_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^*\right) \\
&= \sum_{k=1}^{K}\left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right) + \sum_{k=1}^{K}(V_1^{\pi^k}(s_1^k) - \sum_{h=1}^{H}r_h(s_h^k,a_h^k)) + \sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbf{1}_{s_{h+1}^k} - P_{s_h^k,a_h^k,h})V_{h+1}^*. \qquad (117)
\end{aligned}
$$

The first term in the right hand side (117) is exactly $\text{Regret}(K) = T_1 + T_2 + T_3 + T_4$, the second term is $-T_4$, and the third term is bounded by

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}(\mathbf{1}_{s_{h+1}^k} - P_{s_h^k,a_h^k,h})V_{h+1}^* \le 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*)\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}) \qquad (118)
$$

with probability $1 - 2SAHk\delta$.

It then follows that with probability $1 - 8SAHK\delta$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(V_h^*(s_h^k) - r_h(s_h^k,a_h^k) - P_{s_h^k,a_h^k,h}V_{h+1}^*\right) \\
&\le T_1 + T_2 + T_3 + 2|T_4| + 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*)\log(\frac{1}{\delta})} + 55H\log(\frac{1}{\delta}). \qquad (119)
\end{aligned}
$$

With (116), we further obtain that, with probability $1 - 8SAHK\delta$

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^* - V_{h+1}^{\pi^k}) \\
&\le 4H(T_1 + T_2 + T_3 + 2|T_4| + 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*)\log(\frac{1}{\delta})}) + 262H^2\log(\frac{1}{\delta}). \qquad (120)
\end{aligned}
$$

Define $T_{10} = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*)$. Plugging (120) into (115), with probability $1 - 10SAHK\delta$,

$$
|\check{T}_2| \le 2\sqrt{8K\text{var}_1\log(\frac{1}{\delta})} + 8\sqrt{H(T_1 + T_2 + T_3 + 2|T_4| + 2\sqrt{2T_{10}\log(\frac{1}{\delta})})\log(\frac{1}{\delta})} + 107H\log(\frac{1}{\delta})
$$

$$\leq 11\sqrt{T_{10}\log(\frac{1}{\delta})} + 16\sqrt{H(T_1+T_2+T_3+2|T_4|)\log(\frac{1}{\delta})} + 115H\log(\frac{1}{\delta}). \tag{121}$$

Recalling (113), with probability $1-10SAHK\delta$

$$\begin{aligned}|T_4| &\leq 18\sqrt{SAHT_{10}\log_2(K)\log(\frac{1}{\delta})} + 16\sqrt{H(T_1+T_2+T_3+2|T_4|)\log(\frac{1}{\delta})} + 135SAH^2\log_2(K)\log(\frac{1}{\delta}) \\ &\leq 36\sqrt{SAHT_{10}\log_2(K)\log(\frac{1}{\delta})} + 32\sqrt{H(T_1+T_2+T_3)\log(\frac{1}{\delta})} + 306SAH^2\log_2(K)\log(\frac{1}{\delta}).\end{aligned} \tag{122}$$

F.2.3 Bound of T_5 and T_6

We start with the following lemma

Lemma 22. *With probability* $1-2SAHK\delta$,

$$T_5 \leq 5T_6 + 8BSAH^3. \tag{123}$$

Proof of Lemma 22. Direct computation gives that

$$\begin{aligned}&\sum_{k,h}\mathbb{V}(\hat{P}^k_{s^k_h,a^k_h,h}, V^k_{h+1}) \\ &= \sum_{k,h}\left(\hat{P}^k_{s^k_h,a^k_h,h}(V^k_{h+1})^2 - (\hat{P}_{s^k_h,a^k_h,h}V^k_{h+1})^2\right) \\ &\leq \sum_{k,h}\left(P^k_{s^k_h,a^k_h,h}(V^k_{h+1})^2 - (P_{s^k_h,a^k_h,h}V^k_{h+1})^2\right) + \sum_{k,h}\left(\hat{P}_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)(V^k_{h+1})^2 + 2H\sum_{k,h}\left(\hat{P}_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)V^k_{h+1} \\ &\leq \sum_{k,h}\mathbb{V}(P^k_{s^k_h,a^k_h,h}, V^k_{h+1}) + \sum_{k,h}\left(\hat{P}_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)(V^k_{h+1})^2 + 2H\sum_{k,h}\left(\hat{P}_{s^k_h,a^k_h,h} - P_{s^k_h,a^k_h,h}\right)V^k_{h+1} \\ &= T_5 + T_7 + 2HT_1.\end{aligned} \tag{124}$$

Using Lemma 6 to bound T_7 and T_1, with probability $1-2SAHK\delta$, it holds that

$$\begin{aligned}\sum_{k,h}\mathbb{V}(\hat{P}^k_{s^k_h,a^k_h,h}, V^k_{h+1}) &\leq \sum_{k,h}\mathbb{V}(P^k_{s^k_h,a^k_h,h}, V^k_{h+1}) + 6\sqrt{\sum_{k,h}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1})BSAH^3} + 3BSAH^3 \\ &\leq 5\sum_{k,h}\mathbb{V}(P^k_{s^k_h,a^k_h,h}, V^k_{h+1}) + 8BSAH^3.\end{aligned} \tag{125}$$

□

By Lemma 22, it suffices to bound $T_5 = \sum_{k,h}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1})$.

Because $\mathrm{Var}(X+Y) \leq 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$ for any two random variable X, Y with finite variance, we have that

$$\begin{aligned}\sum_{k,h}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1}) &\leq 2\sum_{k,h}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^*_{h+1}) + 2\sum_{k,h}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1} - V^*_{h+1}) \\ &\leq 3K\mathrm{var}_1 + \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^*_{h+1}) - 3\mathrm{var}_1\right) + 2\sum_{k,h}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^k_{h+1} - V^*_{h+1}).\end{aligned} \tag{126}$$

Lemma 23. *With probability* $1-4SAHK\delta$, *it holds that*

$$T_{10} - 2K\mathrm{var}_1 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\mathbb{V}(P_{s^k_h,a^k_h,h}, V^*_{h+1}) - 2\mathrm{var}_1\right) \leq 80H^2\log(\frac{1}{\delta}). \tag{127}$$

Proof. Let $\bar{R}_h^*(s,a) = \mathbb{V}(P_{s,a,h}, V_{h+1}^*)$. Define

$$\bar{V}_h^k(s) = \mathbb{E}\left[\sum_{h'=h}^{H} \bar{R}_{h'}(s_{h'}, a_{h'}) | s_h = s\right].$$

Then $\bar{V}_h^k(s) \le \mathrm{var}_1 \le H^2$.

We then have that

$$\begin{aligned}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*) - \mathrm{var}_1 &= \sum_{h=1}^{H} \bar{R}_h^*(s_h^k, a_h^k) - \mathrm{var}_1 \\ &\le \sum_{h=1}^{H} \bar{R}_h^*(s_h^k, a_h^k) - \bar{V}_1^k(s_1^k) \\ &= \sum_{h=1}^{H} \left(\mathbf{1}_{s_{h+1}^k} - P_{s_h^k,a_h^k,h}\right) \bar{V}_{h+1}^k. \end{aligned} \tag{128}$$

Note that $\bar{V}^k$ only depends on π^k, which is determined before the k-th episode start. With Lemma 9, with probability $1 - 2SAHK\delta$,

$$\begin{aligned}&\sum_{k=1}^{K} \left(\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*) - \bar{V}_1^k(s_1^k)\right) \\ &\le 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, \bar{V}_{h+1}^k)\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}).\end{aligned} \tag{129}$$

We further bound

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, \bar{V}_{h+1}^k) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H} \left(P_{s_h^k,a_h^k,h}(\bar{V}_{h+1}^k)^2 - (P_{s_h^k,a_h^k,h}\bar{V}_{h+1}^k)^2\right) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H} \left(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k}\right)(\bar{V}_{h+1}^k)^2 \\ &\qquad + \sum_{k=1}^{H}\sum_{h=1}^{H} \left((\bar{V}_{h+1}^k(s_{h+1}^k))^2 - (\bar{V}_h^k(s_h^k))^2\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \left((\bar{V}_h^k(s_h^k))^2 - (P_{s_h^k,a_h^k,h}\bar{V}_{h+1}^k)^2\right) \\ &\le 2\sqrt{8H^4\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, \bar{V}_{h+1}^k)\log(\frac{1}{\delta})} + 2H^2\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{R}_h(s_h^k, a_h^k) + 3H^4\log(\frac{1}{\delta}).\end{aligned} \tag{130}$$

Here the last inequality is by Lemma 9 and Lemma 10 (with probability $1 - 2SAHK\delta$) and the fact that $\bar{V}_h^k(s_h^k) = \bar{R}_h(s_h^k, a_h^k) + P_{s_h^k,a_h^k,h}\bar{V}_{h+1}^k$.

It then follows that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, \bar{V}_{h+1}^k) \le 4H^2\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{R}_h(s_h^k, a_h^k) + 42H^4\log(\frac{1}{\delta}). \tag{131}$$

By (129) and (131), we learn that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*) \le \sum_{k=1}^{H} \bar{V}_1^k(s_1^k) + 2\sqrt{8H^2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*)\log(\frac{1}{\delta})} + 21H^2\log(\frac{1}{\delta}),$$

which further implies that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^*)\le 2\sum_{k=1}^{K}\bar{V}_1^k(s_1^k)+84H^2\log(\frac{1}{\delta})\le 2K\mathrm{var}_1+84H^2\log(\frac{1}{\delta}).$$

The proof is completed.

□

For the left term $\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^*)$, we have the lemma below.

Lemma 24. *With probability $1-2\delta$, it holds that*

$$\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^*)\le 4\sqrt{BH^2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)}+4H\sum_{k,h}b_h^k(s_h^k,a_h^k)+3BSAH^3.$$

Proof of Lemma 24. Direct computation gives that

$$\begin{aligned}
&\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^*)\\
&=\sum_{k,h}\left(P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^*)^2-(P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^*))^2\right) &(132)\\
&=\sum_{k,h}\left((P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k-V_{h+1}^*)^2\right)\\
&\qquad+\sum_{k,h}\left((V_{h+1}^k(s_{h+1}^k)-V_{h+1}^*(s_{h+1}^k))^2-((P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^*))^2)\right)\\
&=\sum_{k,h}\left((P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k-V_{h+1}^*)^2\right)+\sum_{k,h}\left((V_h^k(s_h^k)-V_h^*(s_h^k))^2-((P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^*))^2)\right). &(133)
\end{aligned}$$

By Lemma 9 and Lemma 10, with probability $1-\delta$, it holds that

$$\sum_{k,h}\left((P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k-V_{h+1}^*)^2\right)\le 2\sqrt{2}\sqrt{4H^2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^*)\log(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta}). \tag{134}$$

On the other hand, with probability $1-\delta$,

$$\begin{aligned}
&\sum_{k,h}\left((V_h^k(s_h^k)-V_h^*(s_h^k))^2-((P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^*))^2)\right)\\
&\le 2H\sum_{k,h}\max\{V_h^k(s_h^k)-P_{s_h^k,a_h^k,h}V_{h+1}^k-(V_h^*(s_h^k)-P_h^kV_{h+1}^*),0\}\\
&\le 2H\sum_{k,h}\max\{V_h^k(s_h^k)-P_{s_h^k,a_h^k,h}V_{h+1}^k-r_h(s_h^k,a_h^k),0\}\\
&\le 2H\sum_{k,h}\max\{(\hat{P}_{s_h^k,a_h^k,h}-P_{s_h^k,a_h^k,h})V_{h+1}^k,0\}+2H\sum_{k,h}b_h^k\\
&\le 2\sqrt{BSAH^3\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)}+2H\sum_{k,h}b_h^k(s_h^k,a_h^k)+BSAH^3. &(135)
\end{aligned}$$

It then follows that, with probability $1-2\delta$,

$$\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^*)\le 4\sqrt{BSAH^3\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)}+4H\sum_{k,h}b_h^k(s_h^k,a_h^k)+3BSAH^3. \tag{136}$$

The proof is completed.

□

By Lemma 23 and Lemma 24, we have that with probability $1-6SAHK\delta$,

$$\begin{aligned}
T_6 :=& \sum_{k,h} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k) \\
&\le 2\sum_{k,h} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^*) + 2\sum_{k,h} \mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k - V_{h+1}^*) \\
&\le 4K\mathrm{var}_1 + 8\sqrt{BSAH^3T_6} + 8HT_2 + 7BSAH^3 \\
&\le 8K\mathrm{var}_1 + 16HT_2 + 78BSAH^3. \qquad (137)
\end{aligned}$$

By Lemma 22 and (137), with probability $1-8SAHK\delta$, it holds that

$$T_5 := \sum_{k,h} \mathbb{V}(\hat{P}_{s_h^k,a_h^k,h}, V_{h+1}^k) \le 40K\mathrm{var}_1 + 80HT_2 + 398BSAH^3. \qquad (138)$$

F.2.4 Putting All Together

We rewrite the inequalities (74) – (73) as follows with (68), (70), (71) and (72) replaced by (109), (122) (138) and (137). Recall $B = 4000\log_2^3(K)\log(3SA)\log(\frac{1}{\delta})$.

$$\begin{aligned}
&T_1 \le \sqrt{128BSAHT_6} + 24BSAH^2; \\
&T_7 \le H\sqrt{512BSAHT_6} + 24BSAH^3; \\
&T_9 \le \sqrt{128BSAHT_6} + 24BSAH^2; \\
&T_2 \le 100\sqrt{BSAHT_5} + 140BSAH^2; \\
&T_3 \le \sqrt{8BT_6} + 3H\log(\frac{1}{\delta}); \\
&T_4 \le \sqrt{BSAHT_{10}} + 32\sqrt{BH(T_1+T_2+T_3)} + BSAH^2; \\
&T_5 \le 40K\mathrm{var}_1 + 80HT_2 + 398BSAH^3; \\
&T_6 \le 8K\mathrm{var}_1 + 16HT_2 + 78BSAH^3; \\
&T_8 \le \sqrt{32BH^2T_6} + 3BH^2.
\end{aligned}$$

On the other hand, by Lemma 23, we have

$$T_{10} \le 2K\mathrm{var}_1 + 80BH^2.$$

Solving the inequalities above, we obtain that, with probability $1-200SAH^2K^2\delta$,

$$\mathrm{Regret}(K) = T_1+T_2+T_3+T_4 \le O\left(\sqrt{BSAHK\mathrm{var}_1} + BSAH^2\right). \qquad (139)$$

The proof is completed by replacing δ with $\frac{\delta}{200SAH^2K^2}$.

F.3 Proof of Lemma 20

Following the arguments in the proof of Lemma19, we now bound T_2, T_4, T_5 and T_6 with respect to var_2.

F.3.1 Bound of T_2

Recall in (55), we show that

$$T_2 \le \frac{460}{9}\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 4\sqrt{SAH\log_2(K)\log(\frac{1}{\delta})}\sqrt{\sum_{k,h}\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right)} + \frac{1088}{9}SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{140}$$

Lemma 25. *With probability* $1-4SAHK\delta$,

$$\sum_{k,h}\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right) \le 6K\mathrm{var}_2 + 242H^2\log_2(K)\log(\frac{1}{\delta}). \tag{141}$$

Proof. Recall in Lemma 21, we show that with probability $1-4SAHK\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{\sigma}_h^k(s_h^k,a_h^k) - (\hat{r}_h^k(s_h^k,a_h^k))^2\right) \le 3\sum_{k=1}^{K}\tilde{V}_1^k(s_1^k) + 2SAH^3\log_2(K)\log(\frac{1}{\delta}). \tag{142}$$

We then complete the proof by noting that

$$\tilde{V}_1^k(s_1^k) \le \tilde{V}_1^k(s_1^k) + \mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},V_{h+1}^{\pi^k})|s_1=s_1^k\right] = \mathrm{Var}_{\pi^k}\left[\sum_{h=1}^{H}r_h(s_h,a_h)|s_1=s_1^k\right] \le \mathrm{var}_2. \tag{143}$$

□

By Lemma 25, with probability $1-4SAHK\delta$,

$$T_2 \le \frac{460}{9}\sqrt{2SAH\log_2(K)\log(\frac{1}{\delta})T_5} + 12\sqrt{SAH\log_2(K)\log(\frac{1}{\delta})}\sqrt{2K\mathrm{var}_1} + 157SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{144}$$

F.3.2 Bound of T_4

Recall that $T_4 = \check{T}_1 + \check{T}_2$ where $\check{T}_1 = \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right)$ and $\check{T}_2 = \sum_{k=1}^{K}\left(\sum_{h=1}^{H}r_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right)$.

Following the arguments in Lemma 21 and (111), with probability $1-6SAHK\delta$,

$$\begin{aligned}|\check{T}_1| &\le \sqrt{4SAH\log_2(K)\log(\frac{1}{\delta})}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}v_h(s_h^k,a_h^k)} + 2SAH^2\log_2(K)\log(\frac{1}{\delta}) \\ &\le \sqrt{8SAHK\mathrm{var}_1\log_2(K)\log(\frac{1}{\delta})} + 20SAH^2\log_2(K)\log(\frac{1}{\delta}).\end{aligned}$$

On the other hand, by Lemma 9 and the definition of var_2, with probability $1-2SAHK\delta$

$$|\check{T}_2| \le 2\sqrt{2K\mathrm{var}_2\log(\frac{1}{\delta})} + 3H\log(\frac{1}{\delta}). \tag{145}$$

Therefore, with probability $1-8SAHK\delta$,

$$T_4 \le 4\sqrt{2SAHK\mathrm{var}_2\log_2(K)\log(\frac{1}{\delta})} + 23SAH^2\log_2(K)\log(\frac{1}{\delta}). \tag{146}$$

F.3.3 Bounds of T_5 and T_6

Recall Lemma 22 states that with probability $1-2\delta$, $T_5 \le 5T_6 + 8BSAH^3$. So it suffices to bound T_5.

Because $\mathrm{Var}(X+Y) \le 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$ for any two random variable X, Y with finite variance, we have that

$$\begin{aligned}\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k) &\le 2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^{\pi^k}) + 2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k - V_{h+1}^{\pi^k}) \\ &\le 3K\mathrm{var}_1 + \sum_{k=1}^{K}\left(\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^{\pi^k}) - 3\mathrm{var}_1\right) + 2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^k - V_{h+1}^{\pi^k}).\end{aligned} \tag{147}$$

Lemma 26. *With probability $1-4SAHK\delta$, it holds that*

$$\sum_{k=1}^{K}\left(\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^{\pi^k}) - 2\mathrm{var}_2\right) \le 80H^2\log(\frac{1}{\delta}). \tag{148}$$

Proof. Let $\check{R}_h^k(s,a) = \mathbb{V}(P_{s,a,h}, V_{h+1}^{\pi^k})$. Define

$$\check{V}_h^k(s) = \mathbb{E}\left[\sum_{h'=h}^{H}\check{R}_{h'}^k(s_{h'},a_{h'})|s_h = s\right].$$

Then $\check{V}_h^k(s) \le \mathrm{var}_2 \le H^2$. We have that

$$\begin{aligned}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^{\pi^k}) - \mathrm{var}_2 &= \sum_{h=1}^{H}\check{R}_h^k(s_h^k,a_h^k) - \mathrm{var}_2 \\ &\le \sum_{h=1}^{H}\check{R}_h^k(s_h^k,a_h^k) - \check{V}_1^k(s_1^k) \\ &= \sum_{h=1}^{H}\left(\mathbf{1}_{s_{h+1}^k} - P_{s_h^k,a_h^k,h}\right)\check{V}_{h+1}^k.\end{aligned} \tag{149}$$

Note that $\check{V}^k$ only depends on π^k, which is determined before the k-th episode start. With Lemma 9, with probability $1-2SAHK\delta$,

$$\begin{aligned}&\sum_{k=1}^{K}\left(\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, V_{h+1}^{\pi^k}) - \check{V}_1^k(s_1^k)\right) \\ &\le 2\sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, \check{V}_{h+1}^k)\log(\frac{1}{\delta})} + 3H^2\log(\frac{1}{\delta}).\end{aligned} \tag{150}$$

We further bound

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h}, \check{V}_{h+1}^k) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(P_{s_h^k,a_h^k,h}(\check{V}_{h+1}^k)^2 - (P_{s_h^k,a_h^k,h}\check{V}_{h+1}^k)^2\right) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(P_{s_h^k,a_h^k,h} - \mathbf{1}_{s_{h+1}^k}\right)(\check{V}_{h+1}^k)^2\end{aligned}$$

$$+\sum_{k=1}^{H}\sum_{h=1}^{H}\left((\check{V}_{h+1}^{k}(s_{h+1}^{k}))^2-(\check{V}_{h}^{k}(s_{h}^{k}))^2\right)+\sum_{k=1}^{K}\sum_{h=1}^{H}\left((\check{V}_{h}^{k}(s_{h}^{k}))^2-(P_{s_h^k,a_h^k,h}\check{V}_{h+1}^{k})^2\right)$$

$$\leq 2\sqrt{8H^4\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},\check{V}_{h+1}^{k})\log(\frac{1}{\delta})}+2H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\check{R}_h(s_h^k,a_h^k)+3H^4\log(\frac{1}{\delta}). \tag{151}$$

Here the last inequality is by Lemma 9 and Lemma 10 (with probability $1-2SAHK\delta$) and the fact that $\check{V}_h^k(s_h^k)=\check{R}_h(s_h^k,a_h^k)+P_{s_h^k,a_h^k,h}\check{V}_{h+1}^k$.

It then follows that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},\check{V}_{h+1}^{k})\leq 4H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\check{R}_h(s_h^k,a_h^k)+42H^4\log(\frac{1}{\delta}). \tag{152}$$

By (150) and (152), we learn that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})\leq\sum_{k=1}^{H}\check{V}_1^k(s_1^k)+2\sqrt{8H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})\log(\frac{1}{\delta})}+21H^2\log(\frac{1}{\delta}),$$

which further implies that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})\leq 2\sum_{k=1}^{K}\check{V}_1^k(s_1^k)+84H^2\log(\frac{1}{\delta})\leq 2K\mathrm{var}_2+84H^2\log(\frac{1}{\delta}).$$

The proof is finished.

□

For the left term $\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})$, we have the lemma below.

Lemma 27. *With probability $1-4SAKH\delta$, it holds that*

$$\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\leq 4\sqrt{BH^2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)}+4H\sum_{k,h}b_h^k(s_h^k,a_h^k)+3BSAH^3.$$

Proof of Lemma 24. Direct computation gives that

$$\begin{aligned}
&\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\\
&=\sum_{k,h}\left(P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^{\pi^k})^2-(P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^{\pi^k}))^2\right) \qquad (153)\\
&=\sum_{k,h}\left((P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k-V_{h+1}^{\pi^k})^2\right)\\
&\qquad+\sum_{k,h}\left((V_{h+1}^k(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k))^2-((P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^{\pi^k}))^2)\right)\\
&=\sum_{k,h}\left((P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k-V_{h+1}^{\pi^k})^2\right)+\sum_{k,h}\left((V_h^k(s_h^k)-V_h^{\pi^k}(s_h^k))^2-((P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^{\pi^k}))^2)\right). \qquad (154)
\end{aligned}$$

By Lemma 9 and Lemma 10, with probability $1-2SAKH\delta$, it holds that

$$\sum_{k,h}\left((P_{s_h^k,a_h^k,h}-\mathbf{1}_{s_{h+1}^k})(V_{h+1}^k-V_{h+1}^{\pi^k})^2\right)\leq 2\sqrt{2}\sqrt{4H^2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\log(\frac{1}{\delta})}+3H^2\log(\frac{1}{\delta}). \tag{155}$$

On the other hand, with probability $1-2SAKH\delta$,

$$\begin{aligned}
&\sum_{k,h}\left((V_h^k(s_h^k)-V_h^{\pi^k}(s_h^k))^2-((P_{s_h^k,a_h^k,h}(V_{h+1}^k-V_{h+1}^{\pi^k}))^2)\right)\\
&\le 2H\sum_{k,h}\max\{V_h^k(s_h^k)-P_{s_h^k,a_h^k,h}V_{h+1}^k-(V_h^{\pi^k}(s_h^k)-P_h^kV_{h+1}^{\pi^k}),0\}\\
&=2H\sum_{k,h}\max\{V_h^k(s_h^k)-P_{s_h^k,a_h^k,h}V_{h+1}^k-r_h(s_h^k,a_h^k),0\}\\
&\le 2H\sum_{k,h}\max\{(\hat{P}_{s_h^k,a_h^k,h}-P_{s_h^k,a_h^k,h})V_{h+1}^k,0\}+2H\sum_{k,h}b_h^k(s_h^k,a_h^k)\\
&\le 2\sqrt{BSAH^3\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)}+2H\sum_{k,h}b_h^k(s_h^k,a_h^k)+BSAH^3. \qquad (156)
\end{aligned}$$

It then follows that, with probability $1-4SAKH\delta$,

$$\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\le 4\sqrt{BSAH^3\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)}+4H\sum_{k,h}b_h^k(s_h^k,a_h^k)+3BSAH^3. \qquad (157)$$

The proof is completed.

□

By Lemma 26 and Lemma 27, we have that with probability $1-6SAHK\delta$,

$$\begin{aligned}
T_6:=&\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)\\
&\le 2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})+2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\\
&\le 4K\mathrm{var}_2+8\sqrt{BSAH^3T_6}+8HT_2+7BSAH^3\\
&\le 8K\mathrm{var}_2+16HT_2+78BSAH^3. \qquad (158)
\end{aligned}$$

By Lemma 22 and (158), with probability $1-8SAHK\delta$, it holds that

$$T_5:=\sum_{k,h}\mathbb{V}(\hat{P}_{s_h^k,a_h^k,h},V_{h+1}^k)\le 40K\mathrm{var}_2+80HT_2+398BSAH^3. \qquad (159)$$

Then we have

$$\begin{aligned}
\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k)&\le 2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})+2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\\
&\le 6\sum_{k=1}^K\check{\mathrm{var}}^k+\sum_{k=1}^K\left(\sum_{h=1}^H\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})-3\check{\mathrm{var}}^k\right)+2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k})\\
&\le 6K\mathrm{var}_2+\sum_{k=1}^K\left(\sum_{h=1}^H\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^{\pi^k})-3\check{\mathrm{var}}^k\right)+2\sum_{k,h}\mathbb{V}(P_{s_h^k,a_h^k,h},V_{h+1}^k-V_{h+1}^{\pi^k}). \qquad (160)
\end{aligned}$$

By Lemma 26, 22 and Lemma 27, with probability $1-18SAHK\delta$, it holds that

$$T_5\le O\left(K\mathrm{var}_2+H\sqrt{T_6\left(SAH+\log(\frac{1}{\delta})\right)}+T_2+H^2(SAH+\log(\frac{1}{\delta}))\right); \qquad (161)$$

$$T_6\le O\left(K\mathrm{var}_2+H\sqrt{T_6\left(SAH+\log(\frac{1}{\delta})\right)}+T_2+H^2(SAH+\log(\frac{1}{\delta}))\right). \qquad (162)$$

F.3.4 Putting All Together

Recall $B = 4000 \log_2^3(K) \log(3SA) \log(\frac{1}{\delta})$. We rewrite the inequalities (74) – (73) as follows with (68), (70), (71) and (72) replaced by (144),(146) (159) and (158) respectively. With probability $1 - 200SAH^2K^2\delta$, it holds that

$$
\begin{aligned}
&T_1 \le \sqrt{128BSAHT_6} + 24BSAH^2; \\
&T_7 \le H\sqrt{512BSAHT_6} + 24BSAH^3; \\
&T_9 \le \sqrt{128BSAHT_6} + 24BSAH^2; \\
&T_2 \le 100\sqrt{BSAHT_5} + 140BSAH^2; \\
&T_3 \le \sqrt{8BT_6} + 3H\log(\frac{1}{\delta}); \\
&T_4 \le \sqrt{BSAHK\mathrm{var}_2} + BSAH^2; \\
&T_5 \le 40K\mathrm{var}_2 + 80HT_2 + 398BSAH^3; \\
&T_6 \le 8K\mathrm{var}_2 + 16HT_2 + 78BSAH^3; \\
&T_8 \le \sqrt{32BH^2T_6} + 3BH^2.
\end{aligned}
$$

Solving the inequalities above, we obtain that

$$
\mathrm{Regret}(K) = T_1 + T_2 + T_3 + T_4 \le O\left(\sqrt{BSAHK\mathrm{var}_2} + BSAH^2\right). \tag{163}
$$

The proof is completed by replacing δ with $\frac{\delta}{200SAH^2K^2}$.

G Minimax lower bounds

In this section we focus on the proof of the lower bounds

G.1 Proof of Theorem 7

Fix (S, A, H). We start with the following lemma.

Lemma 28. *For any $K' \ge 1$, for any algorithm, there exists an MDP with S states, A actions and horizon H, such that the regret in K' episodes is at least*

$$
\mathrm{Regret}(K') = \Omega(f(K')) := \Omega\left(\min\left\{\sqrt{SAH^3K'}, K'H\right\}\right).
$$

Proof of Lemma 28. The hard instance is based on the hard instance JAO-MDP (Jaksch et al., 2010; Jin et al., 2018). In Appendix.D (Jin et al., 2018), the authors show that when $K \ge C_0SAH$ for some constant C_0, the minimax regret lower bound is $\Omega(\sqrt{SAH^3K})$. Now we focus on the regime $K \le C_0SAH$. Without loss of generality, we assume $S = A = 2$, and the generalization to arbitrary (S, A) is routine. Recall the definition of JAO-MDP (Jaksch et al., 2010). Let the two state be x and y, and the two actions be a and b. The reward is always x at state 1 and always $\frac{1}{2}$ at state y. The transition model is give by $P_{x,a} = P_{x,b} = [1-\delta, \delta]^\top$, $P_{y,a} = [1-\delta, \delta]^\top, P_{y,b} = [1-\delta-\epsilon, \delta+\epsilon]$. Here we choose $\delta = \frac{C_1}{H}$ and $\epsilon = \frac{1}{H}$. Then the mixture time of the MDP is roughly $O(H)$. By choosing C_1 large enough, we can ensure that the MDP is C_3-mixing after the first half horizons for some proper constant $C_3 \in (0, \frac{1}{2})$.

It is then easy to show that action b is the optimal action for state y. Moreover, each time action a is chosen at state y, the learner needs to pay regret $\Omega(\epsilon H) = \Omega(1)$. On the other hand, to discriminate action a from action b at state y with probability $1 - \frac{1}{10}$, the learner needs at least $\Omega(\frac{\epsilon}{\delta^2}) = \Omega(H)$ rounds, saying C_4H rounds for some proper constant $C_4 > 0$. As a result, in the case $K \le C_4H$, the minimax regret is at least $\Omega(KH^2\epsilon) = \Omega(KH)$. When $C_4H \le K \le C_0SAH = 4C_0H$, the minimax regret is at least $\Omega(C_4H^2) = \Omega(KH)$. The proof is completed.

□

Let $\mathcal{M}$ be the hard instance for $K' = \max\left\{\frac{1}{10}Kp, 1\right\}$. We consider an MDP $\mathcal{M}'$ as below. In the first layer, for any state s, with probability p, the leaner transits to a copy of $\mathcal{M}$, and with probability $1-p$, the learner transits to a dumb state with 0 reward. Then we have $v^* \le pH$. Let $X = X_1 + X_2 + \ldots + X_k$, where $\{X_i\}_{i=1}^K$ are i.i.d. Bernoulli random variables with mean p. Let $g(X, K')$ denote the minimax regret on the hard instance $\mathcal{M}$ in X episodes. Clearly $g(X, K')$ is non-decreasing in X. Then $\mathrm{Regret}(K) \ge \mathbb{E}[g(X, K')]$. In the case $Kp \ge 10$, by Lemma 11, with probability $1/2$, $X \ge \frac{1}{10}Kp = K'$. Then it holds that $\mathbb{E}[g(X, K')] \ge \frac{1}{2}\cdot g(K', K') = \frac{1}{2}f(K') = \frac{1}{2}\cdot\Omega\left(\min\left\{\sqrt{SAH^3K'}, K'H\right\}\right) = \Omega(\sqrt{SAH^3Kp}, KHp)$. In the case $Kp < 10$, with probability $1-(1-p)^K \ge (1-e^{-Kp}) \ge \frac{Kp}{30}$, $X \ge 1$. Then $\mathbb{E}[g(X, K')] \ge \frac{Kp}{30}\cdot g(1, K') = \frac{Kp}{30}\cdot g(1,1) = \Omega(KHp)$. The proof is completed.

G.2 Proof of Corollary 2

Without loss of generality, we assume $S = A = 2$. Note that $p \le 1/4$. We consider a hard instance where the learner needs to identify the correct action for each layer. Let $\mathcal{S} = \{s_1, s_2\}$. For any action a and h, we set $P_{s_2,a,h} = \mathbf{1}_{s_2}$ and $r_h(s_2, a) = 0$. For any action $a \ne a^*$ and h, we also set $P_{s_1,a,h} = \mathbf{1}_{s_2}$ and $c_h(s_2, a) = 1$. At last, we set $P_{s_1,a^*,h} = \mathbf{1}_{s_1}$ and $r_h(s_1, a^*) = p$. Let the initial state be s_1. It is then clear that $c^* = Hp$ by choosing a^* for each layer. To identify the correct action a^* for at least half of the H layers, we need $\Omega(H)$ episodes, which implies that, there exists $C_5 > 0$ such that in the first $K \le C_5 H$ episodes, the cost of the learner is at least $\frac{H(1-p)}{2}$. Then the minimax regret is at least $\Omega(K(H - c^*)) = \Omega(KH^2(1-p))$ for $K \le C_5 H$.

In the case $C_5 H \le K \le \frac{100H}{p}$, the minimax regret is at least $\Omega(H(H - c^*)) = \Omega(H^2(1-p))$.

For $K \ge \frac{100H}{p}$, we let $\mathcal{M}$ be the hard instance with the same transition as that in Lemma 28, and set the cost function as $\frac{1}{2}$ for state x and 1 for state y with respect to $K' = Kp/10 \ge 10H$. Let $\mathcal{M}'$ be the MDP such that, in the first layer, with probability p, the learner transits to a copy of $\mathcal{M}$, and with probability $1-p$, the learner transits to a dumb state with 0 cost. Then $c^* = \Theta(Hp)$. Using Lemma 11, with probability $\frac{1}{2}$, $X \ge \frac{1}{3}Kp - \log(2) \ge \frac{1}{6}Kp$. Then $\mathrm{Regret}(K)$ is at least $\frac{1}{2}\cdot\Omega\left(\min\{\sqrt{H^3K'}, K'H\}\right) = \Omega\left(\sqrt{H^3Kp}\right)$.

The proof is completed by combining the minimax regret lower bounds for the three regimes $K \in [1, C_5H], (C_5H, \frac{100H}{p}], (\frac{100H}{p}, \infty]$.

G.3 Proof of Theorem 8

For $K \ge SAH/p$, the lower bound in Theorem 7 applies because the regret is at least $\Omega(\sqrt{SAH^3Kp})$ and the variance var is at most pH^2. On the other hand, for $1 \le K \le SAH$, by Lemma 29, the minimax regret is at least $\Omega(KH)$. For $SAH \le K \le SAH/p$, the regret is at least $\Omega(SAH^2) = \Omega(\min\{\sqrt{SAH^3Kp} + SAH^2, KH\})$. The proof is completed.

Lemma 29. *Fix $1 \le K \le SAH$. There exists an MDP with S states, A actions, horizon H, and $\mathrm{var}_1 = \mathrm{var}_2 = 0$, such that the regret is at least $\Omega(KH)$.*

Proof. We consider an MDP with deterministic transition. That is, for each (s, a, h), there is some s' such that $P_{s,a,h,s'} = 1$ and $P_{s,a,h,s''} = 0$ for any $s'' \ne s'$. The reward function is also deterministic. In this case, it is easy to verify that $\mathrm{var}_1 = \mathrm{var}_2 = 0$.

We first assume $S = 2$. For any action a and horizon h, we set $P_{s_2,a,h} = \mathbf{1}_{s_2}$ and $r_h(s_2, a) = 0$. For any action $a \ne a^*$ and h, we also set $P_{s_1,a,h} = \mathbf{1}_{s_2}$ and $r_h(s_2, a) = 0$. At last, we set $P_{s_1,a^*,h} = \mathbf{1}_{s_1}$ and $r_h(s_1, a^*) = 1$. In other words, there are a dumb state and a normal state in each horizon. The learner hopes to find the correct action to avoid the dumb state. Obviously, $V_1^*(s_1) = H$. To find a $\frac{H}{2}$-optimal policy, the learner needs to identify a^* for the first $\frac{H}{2}$ horizons, which needs at least $\Omega(HA)$ rounds in expectation. As a result, the minimax regret is at least $\Omega(KH)$ for $K \le cHA$ with some proper constant c.

We name the hard instance above as a *hard chain*. For general S, we construct $d := \frac{S}{2}$ hard chains. Let the two states in the i-th be $(s_1(i), s_2(i))$. We set the initial distribution to be the uniform distribution

over $\{s_1(i)\}_{i=1}^d$. Then $V_1^*(s_1(i)) = H$ for any $1 \le i \le d$. Let $\text{Regret}_i(K)$ be the expected regret due to the i-th hard chain. When $K \ge 100S$, by Lemma 11, with probability $\frac{1}{2}$, $s_1(i)$ is visited for at least $\frac{K}{10S} \ge 10$ times. As a result, we have that $\text{Regret}_i(K) \ge \frac{1}{2} \cdot \Omega\left(\frac{KH}{S}\right)$. Taking sum over i, we learn that the total regret is at least $\sum_{i=1}^d \text{Regret}_i(K) = \Omega(KH)$. When $K < 100S$, with probability $1 - (1 - \frac{1}{S})^K \ge 0.0001\frac{K}{S}$, $s_1(i)$ is visited for at least one time. Therefore, $\text{Regret}_i(K) \ge \Omega(\frac{KH}{S})$. Taking sum over i, we obtain that $\text{Regret}(K) = \sum_{i=1}^K \text{Regret}_i(K) = \Omega(KH)$.

□

References

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR.

Agarwal, A., Krishnamurthy, A., Langford, J., Luo, H., et al. (2017). Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, pages 4–7. PMLR.

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1184–1194. Curran Associates, Inc.

Allen-Zhu, Z., Bubeck, S., and Li, Y. (2018). Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, pages 186–194. PMLR.

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

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pages 8004–8013.

Bartlett, P. L. and Tewari, A. (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208.

Bertsekas, D. (2019). *Reinforcement learning and optimal control*. Athena Scientific.

Brafman, R. I. and Tennenholtz, M. (2003). R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3(Oct):213–231.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. (2021). Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234.

Cui, Q. and Yang, L. F. (2021). Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR.

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

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30.

Dann, C., Li, L., Wei, W., and Brunskill, E. (2019). Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1507–1516.

Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. (2021). Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12.

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

Dong, K., Wang, Y., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. (2019). Tight regret bounds for model-based reinforcement learning with greedy policies. *Advances in Neural Information Processing Systems*, 32.

Even-Dar, E. and Mansour, Y. (2003). Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *ICML 2018-The 35th International Conference on Machine Learning*, volume 80, pages 1578–1586.

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

Ji, X. and Li, G. (2023). Regret-optimal model-free reinforcement learning for discounted MDPs with short burn-in time. *arXiv preprint arXiv:2305.15546*.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020). Reward-free exploration for reinforcement learning. *International Conference on Machine Learning*.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

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

Kearns, M. and Singh, S. (1998a). Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, 11.

Kearns, M. J. and Singh, S. P. (1998b). Near-optimal reinforcement learning in polynominal time. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 260–268.

Kolter, J. Z. and Ng, A. Y. (2009). Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520.

Lattimore, T. and Hutter, M. (2012). PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer.

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. (2020). Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2021a). Is Q-learning minimax optimal? a tight sample complexity analysis. *accepted to Operations Research*.

Li, G., Chi, Y., Wei, Y., and Chen, Y. (2022a). Minimax-optimal multi-agent RL in Markov games with a generative model. *Advances in Neural Information Processing Systems*, 35:15353–15367.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022b). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

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

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *accepted to Operations Research*.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021c). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473.

Li, G., Yan, Y., Chen, Y., and Fan, J. (2023). Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*.

Li, Y., Wang, R., and Yang, L. F. (2021d). Settling the horizon-dependence of sample complexity in reinforcement learning. In *IEEE Symposium on Foundations of Computer Science*.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*.

Ménard, P., Domingues, O. D., Shang, X., and Valko, M. (2021). UCB momentum Q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618. PMLR.

Neu, G. and Pike-Burke, C. (2020). A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

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

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of Thirty Third Conference on Learning Theory (COLT)*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716.

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. (2021). Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34:15621–15634.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14433–14443.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025. PMLR.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018a). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196.

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018b). Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. Society for Industrial and Applied Mathematics.

Simchowitz, M. and Jamieson, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1153–1162.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. (2006). PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. ACM.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331.

Szita, I. and Szepesvári, C. (2010). Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*.

Talebi, M. S. and Maillard, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *arXiv preprint arXiv:1803.01626*.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. (2021). Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34.

Tirinzoni, A., Pirotta, M., and Lazaric, A. (2021). A fully problem-dependent regret lower bound for finite-horizon MDPs. *arXiv preprint arXiv:2106.13013*.

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

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*.

Wainwright, M. J. (2019b). Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

Wang, K., Zhou, K., Wu, R., Kallus, N., and Sun, W. (2023). The benefits of being distributional: Small-loss bounds for reinforcement learning. *arXiv preprint arXiv:2305.15703*.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. (2020). Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? In *Advances in Neural Information Processing Systems*.

Wang, X., Cui, Q., and Du, S. S. (2022). On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14865–14877.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407.

Xiong, Z., Shen, R., Cui, Q., Fazel, M., and Du, S. S. (2022). Near-optimal randomized exploration for tabular markov decision processes. *Advances in Neural Information Processing Systems*, 35:6358–6371.

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

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022). The efficacy of pessimism in asynchronous Q-learning. *accepted to IEEE Transactions on Information Theory*.

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

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. (2022). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*.

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

Zhang, Z., Ji, X., and Du, S. (2021). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR.

Zhang, Z., Ji, X., and Du, S. (2022). Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*.

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

Zhou, R., Zhang, Z., and Du, S. S. (2023). Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. *arXiv preprint arXiv:2301.13446*.