Model-Based RL with Confidence Bounds: Bridging Regret and Sample Complexity

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 6, 2024

Abstract

We study the sample complexity of model-based online reinforcement learning in time-inhomogeneous finite-horizon Markov Decision Processes (MDPs). Leveraging recent advances in upper confidence bound techniques, we provide minimax optimal regret bounds for both stochastic and deterministic settings. Our primary contribution lies in refining the role of problem-dependent quantities, such as variance and the reward distribution's statistical properties, which impact the fundamental limits of performance. We further establish that our results generalize across several policy iteration algorithms, showcasing tight dependence on key MDP parameters. This work advances the understanding of regret minimization in online RL, offering novel insights for both theoretical and applied settings.

Keywords: Reinforcement Learning, Sample Complexity, Finite-Horizon MDP, Regret Minimization, Upper Confidence Bound.

1 Introduction

The task of reinforcement learning (RL) has garnered considerable interest in recent years, particularly in applications involving sequential decision-making under uncertainty. In this paper, we focus on online RL within the setting of time-inhomogeneous finite-horizon Markov Decision Processes (MDPs), where both the state transitions and rewards can vary across different time steps. This setting, while more general than time-homogeneous cases, poses significant challenges for understanding the sample complexity and regret behavior of common algorithms.

Our study aims to settle key questions related to the sample complexity of online RL in finite-horizon MDPs. Specifically, we explore how various problem-dependent parameters, such as the variance of reward distributions and transition dynamics, influence the learning performance of RL algorithms. We propose a *model-based* approach, employing Monotonic Value Propagation with upper confidence bound (UCB) updates to optimize the learning process.

Main Contributions We summarize the main findings of this paper, focusing on time-inhomogeneous finite-horizon MDPs. Our contributions extend and clarify the impact of key problem-dependent quantities, such as the variance of reward distributions and the statistical properties of transition kernels, on the sample complexity and regret behavior of online RL algorithms.

- We provide minimax optimal regret bounds for time-inhomogeneous finite-horizon MDPs, highlighting the tight dependence on problem-specific factors like reward variance and transition kernel properties.

- Our results generalize across both stochastic and deterministic environments, demonstrating the robustness of our analysis in various settings.

- We propose a model-based algorithm, Monotonic Value Propagation, which uses UCB-based optimism to achieve near-optimal regret bounds, with both theoretical guarantees and practical applicability.

1.1 Organization

The paper is organized as follows: Section 2 introduces the basic problem formulation for online RL in time-inhomogeneous finite-horizon MDPs. Section 3 presents our proposed algorithm, Monotonic Value Propagation, and its theoretical analysis. Section 4 discusses the main regret bounds and sample complexity results. Finally, Section 5 concludes the paper with potential directions for future research.

2 Problem formulation

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 non-statioanry) 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 $R = \big\{R_{s,a,h} \in \Delta([0, H])\big\}_{1 \le h \le H, s \in \mathcal{S}, a \in \mathcal{A}}$ 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 $R_{s,a,h}$. We shall also denote by $r = \big\{r_h(s,a)\big\}_{1 \le h \le H, s \in \mathcal{S}, a \in \mathcal{A}}$ the mean reward function, so that $r_h(s,a) \coloneqq \mathbb{E}_{r' \sim R_{s,a,h}}[r'] \in [0, H]$ for any (s,a,h)-tuple. 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., [Ber19]) for more extensive descriptions.

In each episode, a trajectory $(s_1, a_1, r'_1, s_2, \ldots, s_H, a_H, r'_H)$ is rolled out as follows: 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 $r'_h \sim R_{s_h,a_h,h}$, 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. *For any possible trajectory $(s_1, a_1, r'_1, \ldots, s_H, a_H, r'_H)$, one always has $0 \le \sum_{h=1}^{H} r'_h \le H$.*

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 [JA18, WDYK20].

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

$$V_h^\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_{j=h}^{H} r'_j \,\middle|\, s_h = s \right], \qquad \forall (s,h) \in \mathcal{S} \times [H], \tag{1a}$$

$$Q_h^\pi(s,a) \coloneqq \mathbb{E}_\pi \left[\sum_{j=h}^{H} r'_j \,\middle|\, (s_h, a_h) = (s,a) \right], \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \tag{1b}$$

where the expectation $\mathbb{E}_\pi[\cdot]$ is taken over the randomness of an episode $\big\{(s_h, a_h, r'_h)\big\}_{1 \le h \le H}$ generated under policy π, that is, $a_j = \pi_j(s_j)$ for every $h \le j \le H$ (resp. $h < j \le H$) is chosen in the definition of V_h^π (resp. Q_h^π). Accordingly, we define the optimal value function and the optimal Q-function respectively as:

$$V_h^\star(s) \coloneqq \max_\pi V_h^\pi(s), \qquad \forall (s,h) \in \mathcal{S} \times [H], \tag{2a}$$

$$Q_h^\star(s,a) \coloneqq \max_\pi Q_h^\pi(s,a) \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{2b}$$

Throughout this paper, we shall often abuse the notation by letting both V_h^π and $V_h^\star$ (resp. Q_h^π and $Q_h^\star$) represent S-dimensional (resp. SA-dimensional) vectors containing all elements of the corresponding value functions (resp. Q-functions). Two important properties are worth mentioning: (a) the optimal value and the optimal Q-function are linked by the Bellman equation:

$$Q_h^\star(s,a) = r_h(s,a) + \big\langle P_{h,s,a}, V_{h+1}^\star \big\rangle, \qquad V_h^\star(s) = \max_{a'} Q_h^\star(s,a'), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]; \tag{3}$$

(b) there exists a deterministic policy, denoted by $\pi^\star$, that achieves optimal value functions and Q-functions for all state-action-step tuples simultaneously, that is,

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

Data collection protocol 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 $\big\{(s_h^k, a_h^k, r_h^k)\big\}_{1 \le h \le H}$, 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:

$$\mathsf{Regret}(K) \coloneqq \sum_{k=1}^{K} \Big(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) \Big) \tag{4}$$

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., [JAZBJ18]) that a regret bound can often be readily translated into a PAC sample complexity result, the latter of which counts the number of episodes needed to find an ε-optimal policy $\widehat{\pi}$ in the sense that $\mathbb{E}_{s_1 \sim \mu}\big[V_1^\star(s_1) - V_1^{\widehat{\pi}}(s_1)\big] \le \varepsilon$. For instance, the reduction argument in [JAZBJ18] 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 $\mathbb{E}_{s_1 \sim \mu}\big[V_1^\star(s_1) - V_1^{\widehat{\pi}}(s_1)\big] \lesssim f(S,A,H)K^{-\alpha}$, thus resulting in a sample complexity bound of $\big(\frac{f(S,A,H)}{\varepsilon}\big)^{1/\alpha}$.

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 [ZJD21]. 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 (see, e.g., `UCBVI` by [AOM17]). More specifically, the learner calculates the optimistic Bellman equation backward (from $h = H, \ldots, 1$): it first computes an empirical estimate $\widehat{P} = \{\widehat{P}_h \in \mathbb{R}^{SA\times S}\}_{1\le h\le H}$ of the transition probability kernel as well as an empirical estimate $\widehat{r} = \{\widehat{r}_h \in \mathbb{R}^{SA}\}_{1\le h\le H}$ of the mean reward function, and then maintains upper estimates for the associated value function and Q-function using

$$Q_h(s,a) \leftarrow \min\left\{\widehat{r}_h(s,a) + \langle \widehat{P}_{s,a,h}, V_{h+1}\rangle + b_h(s,a), H\right\} \tag{5a}$$

$$V_h(s) \leftarrow \max_a Q_h(s,a) \tag{5b}$$

  for all state-action pairs. Here, Q_h (resp. V_h) indicates the running estimate for the Q-function (resp. value function), whereas $b_h(s,a) \ge 0$ is some suitably chosen bonus term that compensates for the uncertainty. The above opportunistic Q-estimate in turn allows one to obtain a policy estimate (via a simple greedy rule), which will then be exeuted to collect new data. The fact that we first estimate the model (i.e., the transition kernel and mean rewards) makes it a model-based approach. Noteworthily, the empirical model $(\widehat{P}, \widehat{r})$ shall be updated multiple times as new samples continue to arrive, and hence the updating rule (5) will be invoked a couple of times as well.

- *An epoch-based procedure and a doubling trick.* Compared to the original `UCBVI` [AOM17], one distinguishing feature of `MVP` is to update the empirical transition kernel and empirical rewards in an epoch-based fashion, as motivated by a doubling update framework adopted in [JOA10]. More concretely, the whole learning process is divided into consecutive epochs via a simple doubling rule; namely, whenever there exits a (s,a,h)-tuple whose visitation count reaches a power of 2, we end the current epoch, reconstruct the empirical model (cf. lines 13 and 15 of Algorithm 1), compute the Q-function and value function using the newly updated transition kernel and rewards (cf. (7)), and then start a new epoch with an updated sampling policy. This stands in stark contrast with the original `UCBVI`, which computes new estimates for the transition model, Q-function and value function in every episode. With this doubling rule in place, the estimated transition probability vector for each (s,a,h)-tuple will be updated by no more than $\log_2 K$ times, a feature that plays a pivotal role in significantly reducing some sort of covering number needed in our covering-based analysis (as we shall elaborate on shortly in Section C). 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 — which will then be employed to collect samples in *all* episodes of the next epoch. More technical explanations of the doubling update rule will be provided in Section C.2.

- *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 (5a). Here, we adopt the monotonic Bernstein-style bonus function for `MVP` originally proposed in [ZJD21], to be made precise in (6). Compared to the bonus function in `Euler` [ZB19] and `UCBVI` [AOM17], the

monotonic bonus form has a cleaner structure that effectively avoids large lower-order terms. Note that in order to enable variance-aware regret, we also need to keep track of the empirical variance of the (stochastic) immediate rewards.

Remark. We note that a doubling update rule has also been used in the original `MVP` [ZJD21]. A subtle difference between our modified version and the original one lies in that: when the visitation count for some (s, a, h) reaches 2^i for some integer $i \geq 1$, we only use the second half of the samples (i.e., the $\{2^{i-1}+l\}_{l=1}^{2^{i-1}}$-th samples) to compute the empirical model, whereas the original `MVP` makes use of all the 2^i samples. This modified step turns out to be helpful in our analysis, while still preserving sample efficiency in an orderwise sense (since the latest batch always contains at least half of the samples).

4 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 investigate optimal performance in a more problem-dependent fashion [TM18, SJ19, ZB19, ZZD23, FPLO18, XMD21, YYD21, JKSY20, WCS+22, ZHZ+23, DMMZ21, TPL21]. 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 1 (Optimal value-dependent regret)**.** *For any $K \geq 1$, Algorithm 1 satisfies*

$$\mathbb{E}\big[\mathsf{Regret}(K)\big] \lesssim \min\big\{\sqrt{SAH^2Kv^\star}, Kv^\star\big\}\log^5(SAHK) \tag{8}$$

where $v^\star$ is the value of the optimal policy averaged over the initial state distribution (to be formally defined in (11)*).*

Moreover, there is also no shortage of applications where the use of a cost function is preferred over a value function [AKL+17, AZBL18, LLWZ20, WZW+23]. For this purpose, we provide another variation based upon the optimal cost.

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

$$\mathsf{Regret}(K) \leq \widetilde{O}\left(\min\big\{\sqrt{SAH^2Kc^\star} + SAH^2,\ K(H-c^\star)\big\}\right) \tag{9}$$

with probability exceeding $1-\delta$, where $c^\star$ denotes the cost of the optimal policy averaged over the initial state distribution (to be formally defined in (13)*).*

It is worth noting that: despite the apparent similarity between the statements of Theorem 1 and Theorem 2, 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.

Algorithm 1: Monotoinic Value Propagation (`MVP`) [ZJD21]

Input: number of states S, number of actions A, horizon H, total number of episodes K, confidence parameter δ

Initialize: $\mathcal{L} = \{1, 2, \ldots, 2^{\log_2(K)}\}$; $c_1 = \frac{460}{9}, c_2 = 2\sqrt{2}, c_3 = \frac{544}{9}$

Initialization: $\theta_h(s,a) \leftarrow 0, \kappa_h(s,a) \leftarrow 0, N_h^{\mathsf{all}}(s,a,s') \leftarrow 0, N_h(s,a,s') \leftarrow 0, n_h(s,a) \leftarrow 0, Q_h(s,a) \leftarrow H-h+1, V_h(s) \leftarrow H-h+1, \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$;

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

  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 reward r_h^k and observe s_{h+1}^k

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

    Set $N_h^{\mathsf{all}}(s,a) \leftarrow N_h^{\mathsf{all}}(s,a) + 1$, $N_h(s,a,s') \leftarrow N_h(s,a,s') + 1$, $\theta_h(s,a) \leftarrow \theta_h(s,a) + r_h^k$, $\kappa_h(s,a) \leftarrow \kappa_h(s,a) + (r_h^k)^2$

    \\ *Update empirical rewards and transition probability*

    if $N_h^{\mathsf{all}}(s,a) \in \mathcal{L}$ **then**

      Set $\widehat{r}_h(s,a) \leftarrow \frac{\theta_h(s,a)}{\sum_{\widetilde{s}'} N_h(s,a,\widetilde{s}')}$ `// empirical rewards of this epoch`

      Set $\widehat{\sigma}_h(s,a) \leftarrow \frac{\kappa_h(s,a)}{\sum_{\widetilde{s}'} N_h(s,a,\widetilde{s}')} - (\widehat{r}_h(s,a))^2$

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

      Set $n_h(s,a) \leftarrow \sum_{\widetilde{s}'} N_h(s,a,\widetilde{s}')$

      Set TRIGGERED = TRUE

      $\theta_h(s,a) \leftarrow 0, \kappa_h(s,a) \leftarrow 0, N_h(s,a,\widetilde{s}) \leftarrow 0, \forall \widetilde{s} \in \mathcal{S}$;

    end if

  end for

  \\ *Update Q-function*

  if TRIGGERED **then**

    $V_{H+1}(s) \leftarrow 0, \forall s \in \mathcal{S}$;

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

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

        Set

$$b_h(s,a) \leftarrow c_1 \sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a,h}, V_{h+1}) \log(\frac{1}{\delta'})}{\max\{n_h(s,a), 1\}}} + c_2 \sqrt{\frac{(\widehat{\sigma}_h(s,a) - (\widehat{r}_h(s,a))^2) \log(\frac{1}{\delta'})}{\max\{n_h(s,a), 1\}}} + c_3 \frac{H \log(\frac{1}{\delta'})}{\max\{n_h(s,a), 1\}} \tag{6}$$

$$Q_h(s,a) \leftarrow \min\{\widehat{r}_h(s,a) + \widehat{P}_{s,a,h} V_{h+1} + b_h(s,a), H\} \tag{7}$$

$$V_h(s) \leftarrow \max_a Q_h(s,a)$$

      end for

    end for

    Set TRIGGERED = FALSE

  end if

end for

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

$$\mathsf{Regret}(K) \leq \widetilde{O}\left(\min\left\{\sqrt{SAHK\mathrm{var}} + SAH^2,\, KH\right\}\right) \tag{10}$$

with probability at least $1-\delta$, where var *is a certain variance-type metric (to be formally defined in* (17)*).*

Two remarks concerning the above extensions are in order:

- In the worst-case scenarios, the quantities $v^\star$, $c^\star$ and var can all be as large as the order of H, in which case Theorems 1-3 readily recover Theorem 7. In contrast, the advantages of Theorems 1-3 over Theorem 7 become more evident in those favorable cases (e.g., the situation where $v^\star \ll H$ or $c^\star \ll H$, or the case when the environment is nearly deterministic (so that var ≈ 0)).
- Interestingly, the regret bounds in Theorems 1-3 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 1-3, we will provide matching lower bounds.

In the rest of this paper, we develop more refined regret bounds for Algorithm 1 in order to reflect the role of several problem-dependent quantities. Detailed proofs are postponed to §6.1 and §6.3.

5 Minimax lower bounds

5.1 Value-based regret bounds

Thus far, we have not yet introduced the crucial quantity $v^\star$ in Theorem 1, which we define now. When the initial states are drawn from μ, we define $v^\star$ to be the weighted optimal value:

$$v^\star := \mathbb{E}_{s\sim\mu}\big[V_1^\star(s)\big] \tag{11}$$

Encouragingly, the value-dependent regret bound we develop in Theorem 1 is still minimax-optimal, as asserted by the following lower bound.

Theorem 4. *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 $v^\star \leq Hp$ and*

$$\mathbb{E}\big[\mathsf{Regret}(K)\big] \gtrsim \min\left\{\sqrt{SAH^3Kp},\, KHp\right\}. \tag{12}$$

In fact, the construction of the hard instance (as in the proof of Theorem 4) 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 obeys $v^\star \leq Hp$ and that the expected regret is at least

$$\Omega\Big(\min\left\{\sqrt{SAH^3Kp}, KHp\right\}\Big) \gtrsim \min\left\{\sqrt{SAH^2Kv^\star}, Kv^\star\right\}.$$

See §7 for more details.

5.2 Cost-based regret bounds

Next, we turn to the cost-aware regret bound as in Theorem 2. Note that all other results except for Theorem 2 (and a lower bound in this subsection) are about rewards as opposed to cost. In order to facilitate discussion, let us first formally formulate 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 $c^\star$ is then defined as

$$c^\star := \min_\pi \mathbb{E}_{\pi, s_1\sim\mu}\left[\sum_{h=1}^H c_h\right] \tag{13}$$

In this cost-based setting, we find it convenient to 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 adopt different fonts to differentiate them from the original Q-function and value function. The optimal cost function is then define by

$$\mathtt{Q}_h^\star(s,a) = \min_\pi \mathtt{Q}_h^\pi(s,a) \qquad \text{and} \qquad \mathtt{V}_h^\star(s) = \min_\pi \mathtt{V}_h^\pi(s).$$

Given the definitions above, we overload the notation $\mathsf{Regret}(K)$ to denote the regret for the cost-based scenario as

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

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 (7) is replaced by

$$Q_h(s,a) \leftarrow \max\Big\{\min\Big\{\widehat{r}_h(s,a) + \widehat{P}_{s,a,h}V_{h+1} + b_h(s,a),\, 0\Big\},\, -H\Big\} \tag{14}$$

Note that the proof of Theorem 2 closely resembles that of Theorem 1, which can be found in §6.2.

To confirm the tightness of Theorem 2, we develop the following matching lower bound, which resorts to a similar hard instance as in the proof of Theorem 4.

Theorem 5. *Consider any $p \in [0,1/4]$ and any $K \geq 1$. For any algorithm, one can construct an MDP with S states, A actions and horizon H obeying $c^\star \asymp Hp$ and*

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

The proof of this lower bound can be found in §7.2.

5.3 Variance-dependent regret bound

The final regret bound presented in Theorem 3 depends on some 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

$$\mathrm{var}_1 \coloneqq \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \mathbb{V}\big(P_{s_h,a_h,h}, V_{h+1}^{\star}\big) + \sum_{h=1}^{H} \mathrm{Var}\big(R_h(s_h,a_h)\big)\right] \tag{15}$$

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 $\{V_h^{\star}\}_{h=1}^{H}$.

(ii) Another useful variance metric is defined as

$$\mathrm{var}_2 \coloneqq \max_{\pi,s} \mathrm{Var}_{\pi}\left[\sum_{h=1}^{H} r_h \,\Big|\, s_1 = s\right] \tag{16}$$

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 [ZZD23] for further discussion about these two metrics. Our final variance metric is then defined as

$$\mathrm{var} \coloneqq \min\big\{\mathrm{var}_1, \mathrm{var}_2\big\} \tag{17}$$

With the above variance metrics in mind, we can then revisit Theorem 3. As a special case, 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 6.3.

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

Theorem 6. *Consider any $p \in [0,1]$ and any $K \ge 1$. For any algorithm, one can find an MDP instance with S states, A actions, and horizon H satisfying $\max\{\mathrm{var}_1, \mathrm{var}_2\} \le H^2 p$ and*

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

The proof of Theorem 6 resembles that of Theorem 4, 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 §7.

6 Proofs for optimal problem-dependent regret bounds

6.1 Proof of the value-based regret bound (proof of Theorem 1)

Recall that

$$B = 4000(\log_2 K)^3 \log(3SAH) \log \frac{1}{\delta'} \qquad \text{with } \delta' = \frac{\delta}{200SAH^2K^2} \tag{18}$$

Consider first the scenario where $K \leq \frac{BSAH^2}{v^\star}$: the regret bound can be upper bounded by

$$\begin{aligned}\mathbb{E}\big[\mathsf{Regret}(K)\big] &= \mathbb{E}\left[\sum_{k=1}^{K}\Big(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\Big)\right] \leq \mathbb{E}\left[\sum_{k=1}^{K} V_1^\star(s_1^k)\right] = K\mathbb{E}_{s_1\sim\mu}\big[V_1^\star(s_1)\big] \\ &= Kv^\star = \min\Big\{\sqrt{BSAH^2Kv^\star}, Kv^\star\Big\} \end{aligned} \tag{19}$$

As a result, the remainder of the proof is dedicated to the the case with

$$K \geq \frac{BSAH^2}{v^\star} \tag{20}$$

To begin with, recall that the proof of Theorem 7 in Section **??** consists of bounding the quantities $T_1, \ldots, T_9$ (see (**??**), (**??**) and (**??**)) and recall that $\delta' = \frac{\delta}{200SAH^2K^2}$. In order to establish Theorem 1, we need to develop tighter bounds on some of these quantities (i.e., T_2, T_4, T_5 and T_6) to reflect their dependency on $v^\star$ (cf. (11)).

Bounding T_2. Recall that we have shown in (**??**) that

$$\begin{aligned} T_2 \leq \frac{460}{9}\sqrt{2SAH(\log_2 K)\Big(\log\frac{1}{\delta'}\Big)T_5} \\ + 4\sqrt{SAH^2(\log_2 K)\log\frac{1}{\delta'}}\sqrt{\sum_{k,h}\widehat{r}_h^k(s_h^k,a_h^k)} + \frac{1088}{9}SAH^2(\log_2 K)\log\frac{1}{\delta'}. \end{aligned}$$

In view of the definition of T_4 (cf. (**??**)) as well as the fact that $\sum_{k=1}^{K} V_1^\star(s_1^k) \leq 3Kv^\star + H\log\frac{1}{\delta'}$ holds with probability at least $1-\delta'$ (see Lemma 8), we arrive at

$$\sum_{k,h}\widehat{r}_h^k(s_h^k,a_h^k) \leq T_4 + \sum_k V_1^{\pi_k}(s_1^k) \leq T_4 + \sum_k V_1^\star(s_1^k) \leq T_4 + 3Kv^\star + H\log\frac{1}{\delta'} \tag{21}$$

which in turn gives

$$\begin{aligned} T_2 \leq \frac{460}{9}\sqrt{2SAH(\log_2 K)\Big(\log\frac{1}{\delta'}\Big)T_5} \\ + 4\sqrt{SAH^2(\log_2 K)\log\frac{1}{\delta'}}\sqrt{T_4 + 3Kv^\star} + 130SAH^2(\log_2 K)\log\frac{1}{\delta'} \end{aligned} \tag{22}$$

Bounding T_4. When it comes to the quantity T_4 (cf. (**??**)), we make the observation that

$$T_4 = \underbrace{\sum_{k=1}^{K}\left(\sum_{h=1}^{H}\widehat{r}_h^k(s_h^k,a_h^k) - r_h(s_h^k,a_h^k)\right)}_{=:\breve{T}_1} + \underbrace{\sum_{k=1}^{K}\left(\sum_{h=1}^{H} r_h(s_h^k,a_h^k) - V_1^{\pi^k}(s_1^k)\right)}_{=:\breve{T}_2} \tag{23}$$

Repeating the arguments for (21) yields

$$\sum_{k,h} r_h(s_h^k,a_h^k) \le \breve{T}_2 + \sum_k V_1^{\pi_k}(s_1^k) \le \breve{T}_2 + \sum_k V_1^{\star}(s_1^k) \le \breve{T}_2 + 3Kv^\star + H\log\frac{1}{\delta'} \tag{24}$$

with probability at least $1-\delta'$. Combining this with Lemma 12, we see that

$$\begin{aligned}\breve{T}_1 &\le 4\sqrt{2SAH^2\log_2 K\log\frac{1}{\delta'}}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} r_h(s_h^k,a_h^k)} + 52SAH^2(\log_2 K)\log\frac{1}{\delta'} \\ &\le 4\sqrt{2SAH^2\log_2 K\log\frac{1}{\delta'}}\sqrt{\breve{T}_2 + 3Kv^\star} + 60SAH^2(\log_2 K)\log\frac{1}{\delta'} \end{aligned} \tag{25}$$

with probability exceeding $1-3SAHK\delta'$. In addition, Lemma 6 tells us that

$$\begin{aligned}\breve{T}_2 &\le 2\sqrt{2\sum_{k=1}^{K}\mathbb{E}_{\pi^k,s_1\sim\mu}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h)\right)^2\right]\log\frac{1}{\delta'}} + 3H^2\log\frac{1}{\delta'} \\ &\le 2\sqrt{2H\sum_{k=1}^{K}\mathbb{E}_{\pi^k,s_1\sim\mu}\left[\sum_{h=1}^{H} r_h(s_h,a_h)\right]\log\frac{1}{\delta'}} + 3H\log\frac{1}{\delta'} \\ &\le 2\sqrt{2KHv^\star\log\frac{1}{\delta'}} + 3H\log\frac{1}{\delta'} \qquad (26) \\ &\le 2Kv^\star + 5H\log\frac{1}{\delta'} \qquad (27)\end{aligned}$$

with probability at least $1-2SAHK\delta'$, where the expectation operator $\mathbb{E}_{\pi^k,s_1\sim\mu}[\cdot]$ is taken over the randomness of a trajectory $\{(s_h,a_h)\}$ generated under policy π^k and initial state $s_1\sim\mu$, the last line arises from the AM-GM inequality, and the penultimate line makes use of Assumption 1 and the fact that

$$\mathbb{E}_{\pi^k,s_1\sim\mu}\left[\sum_{h=1}^{H} r_h(s_h,a_h)\right] = \mathbb{E}_{s_1\sim\mu}\left[V_1^{\pi^k}(s_1)\right] \le \mathbb{E}_{s_1\sim\mu}\left[V_1^{\star}(s_1)\right] = v^\star.$$

Taking (25), (26) and (27) together, we can demonstrate that with probability exceeding $1-5SAHK\delta'$,

$$\breve{T}_1 \le 13\sqrt{SAH^2Kv^\star(\log_2 K)\log\frac{1}{\delta'}} + 80SAH^2(\log_2 K)\log\frac{1}{\delta'}, \tag{28a}$$

$$\breve{T}_2 \le 2\sqrt{2KHv^\star\log\frac{1}{\delta'}} + 3H\log\frac{1}{\delta'}. \tag{28b}$$

Substitution into (23) reveals that: with probability exceeding $1-5SAHK\delta'$,

$$T_4 \le 15\sqrt{SAH^2Kv^\star(\log_2 K)\log\frac{1}{\delta'}} + 83SAH^2(\log_2 K)\log\frac{1}{\delta'}. \tag{29}$$

Bounding T_5. Recall that we have proven in (**??**) that

$$T_5 \le T_7 + T_8 + 2HT_2 + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{r}_h^k(s_h^k, a_h^k). \tag{30}$$

With (24) and (27) in place, we can deduce that, with probability at least $1 - 3SAHK\delta'$,

$$\sum_{k,h} r_h(s_h^k, a_h^k) \le \check{T}_2 + 3Kv^\star + H\log\frac{1}{\delta'} \le 5Kv^\star + 6H\log\frac{1}{\delta'} \tag{31}$$

Moreover, under the assumption (20), we can further bound (28a) as

$$\check{T}_1 \le \sqrt{BSAH^2Kv^\star} + BSAH^2 \le 2Kv^\star$$

with probability exceeding $1 - 3SAHK\delta'$, which combined with (31) and the assumption (20) results in

$$\sum_{k,h}\widehat{r}_h^k(s_h^k, a_h^k) = \sum_{k,h} r_h(s_h^k, a_h^k) + \check{T}_1 \le 7Kv^\star + 6H\log\frac{1}{\delta'} \le 8Kv^\star \tag{32}$$

Substitution into (30) indicates that: with probability exceeding $1 - 6SAHK\delta'$,

$$T_5 \le T_7 + T_8 + 2HT_2 + 16HKv^\star. \tag{33}$$

Bounding T_6. Making use of our bounds (**??**), (**??**) and (32), we can readily derive

$$\begin{aligned} T_6 &\le T_8 + 2HT_2 + 2HT_9 + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{r}_h(s_h^k, a_h^k) \\ &\le \sqrt{32T_6\log\frac{1}{\delta'}} + 2HT_9 + 16HKv^\star + 3H^2\log\frac{1}{\delta'} + 2HT_2 \end{aligned} \tag{34}$$

with probability at least $1 - 16SAH^2K^2\delta'$.

Putting all pieces together. Recalling our choice of B (cf. (18)), we can see from (22), (**??**), (29), (33), (34), (**??**), (**??**) and (**??**) that

$$T_2 \le \sqrt{BSAHT_5} + \sqrt{BSAH^2(T_4 + 3Kv^\star)} + BSAH^2, \tag{35a}$$
$$T_3 \le \sqrt{BT_6} + BH, \tag{35b}$$
$$T_4 \le \sqrt{BSAH^2Kv^\star} + BSAH^2, \tag{35c}$$
$$T_5 \le T_7 + T_8 + 2HT_2 + 16HKv^\star, \tag{35d}$$
$$T_6 \le \sqrt{BT_6} + 2HT_9 + 16HKv^\star + BH^2 + 2HT_2, \tag{35e}$$
$$T_8 \le \sqrt{BH^2T_6} + BH^2, \tag{35f}$$
$$T_1 \le T_9 \le \sqrt{BSAHT_6} + BSAH^2, \tag{35g}$$
$$T_7 \le H\sqrt{BSAHT_6} + BSAH^3. \tag{35h}$$

Solving (35) under the assumption $K \geq \frac{BSAH^2}{v^\star}$ allows us to demonstrate that

$$T_6 \lesssim BHKv^\star \tag{36a}$$
$$T_1 \leq T_9 \lesssim \sqrt{B^2SAH^2Kv^\star} \tag{36b}$$
$$T_7 + T_8 \lesssim \sqrt{B^2SAH^4Kv^\star} \tag{36c}$$
$$T_5 \lesssim BHKv^\star \tag{36d}$$
$$T_2 \lesssim \sqrt{B^2SAH^2Kv^\star} \tag{36e}$$
$$T_3 \lesssim \sqrt{B^2HKv^\star} \tag{36f}$$
$$T_4 \lesssim \sqrt{BSAH^2Kv^\star} \tag{36g}$$

with probability exceeding $1 - 200SAH^2K^2\delta'$. Putting these bounds together with (**??**), we arrive at

$$\mathsf{Regret}(K) \leq T_1 + T_2 + T_3 + T_4 \lesssim B\sqrt{SAH^2Kv^\star}$$

with probability exceeding $1 - 200SAH^2K^2\delta'$. Replacing δ' with $\frac{\delta}{200SAH^2K^2}$ and taking $\delta = \frac{1}{2KH}$ give

$$\begin{aligned}\mathbb{E}\big[\mathsf{Regret}(K)\big] &\lesssim (1-\delta)B\sqrt{SAH^2Kv^\star} + \delta Kv^\star \lesssim B\sqrt{SAH^2Kv^\star} + 1 \asymp B\sqrt{SAH^2Kv^\star} \\ &\asymp \min\left\{B\sqrt{SAH^2Kv^\star}, BKv^\star\right\} \asymp \min\left\{\sqrt{SAH^2Kv^\star}, Kv^\star\right\}\log^5(SAHK),\end{aligned}$$

provided that $K \geq \frac{BSAH^2}{v^\star}$. Taking this collectively with (19) concludes the proof.

6.2 Proof of the cost-based regret bound (proof of Theorem 2)

We now turn to the proof of Theorem 2. For notational convenience, we shall use r to denote the negative cost (namely, $r_h = -c_h$, $\widehat{r}_h = -\widehat{c}_h$, and so on) throughout this section. We shall also use the following notation (and similar quantities like Q_h^k, V_h^k, ...)

$$\begin{aligned}Q_h(s,a) &\leftarrow \max\left\{\min\left\{\widehat{r}_h(s,a) + \widehat{P}_{s,a,h}V_{h+1} + b_h(s,a), 0\right\}, -H\right\}, \\ V_h(s) &\leftarrow \max_a Q_h(s,a),\end{aligned}$$

in order to be consistent with the reward-based setting.

Akin to the proof of Theorem 1, we need to bound the quantities $T_1, \ldots, T_9$ introduced previously (see (**??**), (**??**) and (**??**)). We note that the analysis for T_1, T_3, T_7, T_8 and T_9 in §6.1 readily applies to the negative reward case herein. Thus, it suffices to develop bounds on T_2, T_4, T_5 and T_6 to capture their dependency on $c^\star$, which forms the main content of the remainder of this section.

Bounding T_2. Recall from (**??**) that

$$\begin{aligned}T_2 = \frac{460}{9}\sum_{k,h}\sqrt{\frac{\mathbb{V}\big(\widehat{P}^k_{s_h^k,a_h^k,h}, V_{h+1}^k\big)\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}} + \\ 2\sqrt{2}\sum_{k,h}\sqrt{\frac{\Big(\widehat{\sigma}_h^k(s_h^k,a_h^k) - \big(\widehat{r}_h^k(s_h^k,a_h^k)\big)^2\Big)\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}} + \frac{544}{9}\sum_{k,h}\frac{H\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}\end{aligned} \tag{37}$$

In what follows, let us bound the three terms on the right-hand side of (37) separately.

- For the first and the third terms on the right-hand side of (37), invoking the Cauchy-Schwarz inequality and Lemma 11 gives

$$\sum_{k,h}\sqrt{\frac{\mathbb{V}\big(\widehat{P}^k_{s_h^k,a_h^k,h},V^k_{h+1}\big)\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}}\le\sqrt{2SAH(\log_2 K)\Big(\log\frac{1}{\delta'}\Big)\sum_{k,h}\mathbb{V}\big(\widehat{P}^k_{s_h^k,a_h^k,h},V^k_{h+1}\big)}$$
$$=\sqrt{2SAH(\log_2 K)\Big(\log\frac{1}{\delta'}\Big)T_5} \tag{38}$$

with T_5 defined in (**??**), and in addition,

$$\sum_{k,h}\frac{H\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}\le 2SAH^2(\log_2 K)\log\frac{1}{\delta'}. \tag{39}$$

- Let us turn to the second term on the right-hand side of (37). Observing the basic fact that

$$\widehat{\sigma}_h^k(s_h^k,a_h^k)-\big(\widehat{r}_h^k(s_h^k,a_h^k)\big)^2\le -H\widehat{r}_h^k(s_h^k,a_h^k),$$

we can combine it with Lemma 11 to derive

$$\sqrt{\frac{\Big(\widehat{\sigma}_h^k(s_h^k,a_h^k)-\big(\widehat{r}_h^k(s_h^k,a_h^k)\big)^2\Big)\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}}\le\sqrt{2SAH(\log_2 K)\log\frac{1}{\delta'}}\sqrt{H\sum_{k,h}-\widehat{r}_h^k(s_h^k,a_h^k)}$$
$$\le\sqrt{2SAH^2(\log_2 K)\log\frac{1}{\delta'}}\sqrt{-T_4+3Kc^\star+\sum_{k=1}^K\Big(-V_1^{\pi^k}(s_1^k)+V_1^\star(s_1^k)\Big)+\sum_{k=1}^K\Big(-V_1^\star(s_1^k)-3c^\star\Big)}, \tag{40}$$

where the last inequality invokes the definition of T_4 (see (**??**)). By virtue of Lemma 8 and the definition (13) of $c^\star$, one can show that

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

with probability exceeding $1-\delta'$. In addition, we note that

$$\sum_{k=1}^K\Big(-V_1^{\pi^k}(s_1^k)+V_1^\star(s_1^k)\Big)=\mathsf{Regret}(K)=T_1+T_2+T_3+T_4. \tag{41}$$

Taking these properties together with (40) yields

$$\sqrt{\frac{\Big(\widehat{\sigma}_h^k(s_h^k,a_h^k)-\big(\widehat{r}_h^k(s_h^k,a_h^k)\big)^2\Big)\log\frac{1}{\delta'}}{N_h^k(s_h^k,a_h^k)}}$$
$$\le\sqrt{2SAH^2(\log_2 K)\log\frac{1}{\delta'}}\sqrt{T_1+T_2+T_3+2|T_4|+3Kc^\star+H\log\frac{1}{\delta'}}$$

Putting the above results together, we can deduce that, with probability exceeding $1-\delta'$,

$$T_2 \le 90\sqrt{SAH(\log_2 K)\Big(\log\frac{1}{\delta'}\Big)T_5} + 4\sqrt{SAH^2(\log_2 K)\log\frac{1}{\delta'}}\sqrt{T_1+T_2+T_3+2|T_4|+3Kc^\star+H\log\frac{1}{\delta'}} + 130SAH^2(\log_2 K)\log\frac{1}{\delta'}. \tag{42}$$

Bounding T_4. When it comes to the quantity T_4, we recall that

$$T_4 = \underbrace{\sum_{k=1}^{K}\left(\sum_{h=1}^{H}\widehat{r}_h^k(s_h^k,a_h^k)-r_h(s_h^k,a_h^k)\right)}_{=:\breve{T}_1} + \underbrace{\sum_{k=1}^{K}\left(\sum_{h=1}^{H}r_h(s_h^k,a_h^k)-V_1^{\pi^k}(s_1^k)\right)}_{=:\breve{T}_2} \tag{43}$$

To control T_4, we first make note of the following result that bounds the empirical reward (for the case with negative rewards), which assists in bounding the term $\breve{T}_1$.

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

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\widehat{r}_h^k(s_h^k,a_h^k)-r_h(s_h^k,a_h^k)\right| \le 4\sqrt{2SAH^2(\log_2 K)\log\frac{1}{\delta'}}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\big(-r_h(s_h^k,a_h^k)\big)} + 52SAH^2(\log_2 K)\log\frac{1}{\delta'}.$$

Lemma 1 tells us that with probability at least $1-3SAHK\delta'$,

$$\begin{aligned}|\breve{T}_1| &\le 4\sqrt{2SAH^2(\log_2 K)\log\frac{1}{\delta'}}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\big(-r_h(s_h^k,a_h^k)\big)} + 52SAH^2(\log_2 K)\log\frac{1}{\delta'}\\ &\le 4\sqrt{2SAH^2(\log_2 K)}\cdot\sqrt{-\breve{T}_2+3Kc^\star+\sum_{k=1}^{K}\big(-V_1^\star(s_1^k)-3c^\star\big)} + 52SAH^2(\log_2 K)\log\frac{1}{\delta'}\\ &\le 4\sqrt{2SAH^2(\log_2 K)}\cdot\sqrt{\breve{T}_2+3Kc^\star} + 60SAH^2(\log_2 K)\log\frac{1}{\delta'}.\end{aligned} \tag{44}$$

Here, the last line uses the fact (see Lemma 8) that, with probability exceeding $1-\delta'$,

$$\sum_{k=1}^{K}\big(-V_1^\star(s_1^k)\big) \le 3Kc^\star + H\log\frac{1}{\delta'}. \tag{45}$$

In addition, the Freedman inequality in Lemma 6 combined with (45) implies that, with probability at least $1-3SAHK\delta$,

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

$$\leq 2\sqrt{2H\sum_{k=1}^{K}\mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H}-r_h(s_h,a_h)\,\middle|\, s_1=s_1^k\right]\log\frac{1}{\delta}}+3H\log\frac{1}{\delta}$$

$$=2\sqrt{2H\left(\sum_{k=1}^{K}\left(-V_1^{\pi^k}(s_1^k)+V_1^\star(s_1^k)\right)+\sum_{k=1}^{K}\left(-V_1^\star(s_1^k)-3c^\star\right)+3Kc^\star\right)\log\frac{1}{\delta}}+3H\log\frac{1}{\delta} \quad (46)$$

$$\leq 3Kc^\star+T_1+T_2+T_3+T_4+9H\log\frac{1}{\delta} \quad (47)$$

Combining (44), (46) with (47) reveals that, with probability at least $1-4SAHK\delta$,

$$|\check{T}_1|\leq 16\sqrt{SAH^2(Kc^\star+T_1+T_2+T_3+T_4)(\log_2 K)\log\frac{1}{\delta}}+200SAH^2(\log_2 K)\log\frac{1}{\delta}$$
$$|\check{T}_2|\leq 2\sqrt{2H(3Kc^\star+T_1+T_2+T_3+T_4)\log\frac{1}{\delta}}+9H\log\frac{1}{\delta}.$$

As a result, substitution into (43) leads to

$$|T_4|\leq 22\sqrt{SAH^2(Kc^\star+T_1+T_2+T_3+T_4)(\log_2 K)\log\frac{1}{\delta}}+209SAH^2(\log_2 K)\log\frac{1}{\delta}. \quad (48)$$

Bounding T_5. Invoking the arguments in (**??**) and recalling the update rule (14), we obtain

$$T_5\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle\widehat{P}^k_{s_h^k,a_h^k,h}-P_{s_h^k,a_h^k,h},\,\left(V_{h+1}^k\right)^2\right\rangle+\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle P_{s_h^k,a_h^k,h}-e_{s_{h+1}^k},\,\left(V_{h+1}^k\right)^2\right\rangle \quad (49)$$
$$+2H\sum_{k=1}^{K}\sum_{h=1}^{H}\big[-r_h(s_h^k,a_h^k)\big].$$

Moreover, we recall that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big[-r_h(s_h^k,a_h^k)\big]=-\check{T}_2-\sum_{k=1}^{K}V_1^{\pi^k}(s_1)\leq-\check{T}_2+\sum_{k=1}^{K}V_1^\star(s_1^k). \quad (50)$$

By virtue of (45), one sees that with probability at least $1-5SAHK\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big[-r_h(s_h^k,a_h^k)\big]\leq 2\sqrt{2H(3Kc^\star+T_1+T_2+T_3+T_4)\log\frac{1}{\delta}}+3Kc^\star+10H\log\frac{1}{\delta}. \quad (51)$$

Consequently, we arrive at

$$T_5\leq T_7+T_8+2HT_2+4\sqrt{2H^3(3Kc^\star+T_1+T_2+T_3+T_4)\log\frac{1}{\delta}}+6HKc^\star+20H^2\log\frac{1}{\delta} \quad (52)$$

with probability exceeding $1-5SAHK\delta$.

Bounding T_6. Invoking the arguments in (**??**), (45) and (50), and recalling the update rule (14), we can demonstrate that

$$\begin{aligned}
T_6 &\le 2\sqrt{8T_6\log\frac{1}{\delta}} + 3H^2\log\frac{1}{\delta} + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\max\left\{\langle P_{s_h^k,a_h^k,h}, V_{h+1}^k\rangle - V_h^k(s_h^k), 0\right\} \\
&\le 2\sqrt{8T_6\log\frac{1}{\delta}} + 3H^2\log\frac{1}{\delta} + 2HT_9 + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\left[-r_h(s_h^k,a_h^k)\right] \\
&\le 2\sqrt{8T_6\log\frac{1}{\delta}} + 3H^2\log\frac{1}{\delta} + 2HT_9 \\
&\qquad + 2H\left(2\sqrt{2H(3Kc^\star + T_1 + T_2 + T_3 + T_4)\log\frac{1}{\delta}} + 3Kc^\star + 10H\log\frac{1}{\delta}\right) \qquad (53)
\end{aligned}$$

with probability at least $1 - 3SAHK\delta$.

Putting all this together. Armed with the preceding bounds, we are ready to establish the claimed regret bound. By solving (42),(**??**),(48),(52),(53),(**??**),(**??**) and (**??**), we can show that, with probability exceeding $1 - 100SAH^2K\delta$,

$$\begin{aligned}
T_6 &\lesssim HKc^\star + BSAH^3, \\
T_1 &\lesssim \sqrt{BSAH^2Kc^\star} + BSAH^2, \\
T_7 + T_8 &\lesssim \sqrt{BSAH^4Kc^\star} + BSAH^3, \\
T_5 &\lesssim HKc^\star + BSAH^2, \\
T_2 &\lesssim \sqrt{BSAH^2Kc^\star} + BSAH^2, \\
T_3 &\lesssim \sqrt{BHKc^\star} + BSAH^2.
\end{aligned}$$

We then readily conclude that the total regret is bounded by

$$O\big(\sqrt{BSAH^2Kc^\star} + BSAH^2\big).$$

In addition, the regret bound is trivially upper bounded by $O\big(K(H - c^\star)\big)$. The proof is thus completed by combining these two regret bounds and replacing δ' with $\frac{\delta}{100SAH^2K}$.

6.3 Proof of the variance-dependent regret bounds (proof of Theorem 3)

In this section, we turn to establishing Theorem 3. The proof primarily contains two parts, as summarized in the following lemmas.

Lemma 2. *With probability exceeding $1 - \delta/2$, Algorithm 1 obeys*

$$\mathsf{Regret}(K) \le \widetilde{O}\Big(\min\big\{\sqrt{SAHK\mathrm{var}_1} + SAH^2, KH\big\}\Big).$$

Lemma 3. *With probability at least $1 - \delta/2$, Algorithm 1 satisfies*

$$\mathsf{Regret}(K) \le \widetilde{O}\Big(\min\big\{\sqrt{SAHK\mathrm{var}_2} + SAH^2, KH\big\}\Big).$$

Putting these two regret bounds together and rescaling δ to $\delta/2$, we immediately conclude the proof of Theorem 3. The remainder of this section is thus devoted to establishing Lemma 2 and Lemma 3.

7 Proofs for minimax lower bounds

In this section, we establish the lower bounds advertised in this paper.

7.1 Proof of Theorem 4

Consider any given (S, A, H). We start by establishing the following lemma.

Lemma 4. *Consider any $K' \geq 1$. For any algorithm, there exists an MDP instance with S states, A actions, and horizon H, such that the regret in K' episodes is at least*

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

Proof of Lemma 4. Our construction of the hard instance is based on the hard instance JAO-MDP constructed in [JOA10, JAZBJ18]. In [JAZBJ18, Appendix.D], the authors already showed that when $K' \geq C_0SAH$ for some constant $C_0 > 0$, the minimax regret lower bound is $\Omega(\sqrt{SAH^3K'})$. Hence, it suffices for us to focus on the regime where $K' \leq C_0SAH$. Without loss of generality, we assume $S = A = 2$, and the argument to generalize it to arbitrary (S, A) is standard and henc omitted for brevity.

Recall the construction of JAO-MDP in [JOA10]. Let the two states be x and y, and the two actions be a and b. The reward is always equal to x in state 1 and $1/2$ in state y. The probability transition kernel is given by

$$P_{x,a} = P_{x,b} = [1-\delta, \delta], \quad P_{y,a} = [1-\delta, \delta], \quad P_{y,b} = [1-\delta-\epsilon, \delta+\epsilon],$$

where we choose $\delta = C_1/H$ and $\epsilon = 1/H$. Then the mixing 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 of the horizons for some proper constant $C_3 \in (0, 1/2)$.

It is then easy to show that action b is the optimal action for state y. Moreover, whenever action a is chosen in state y, the learner needs to pay regret $\Omega(\epsilon H) = \Omega(1)$. In addition, to differentiate action a from action b in state y with probability at least $1 - \frac{1}{10}$, the learner needs at least $\Omega(\frac{\epsilon}{\delta^2}) = \Omega(H)$ rounds — let us call it C_4H rounds for some proper constant $C_4 > 0$. As a result, in the case where $K' \leq C_4H$, the minimax regret is at least $\Omega(K'H^2\epsilon) = \Omega(K'H)$. When $C_4H \leq K' \leq C_0SAH = 4C_0H$, the minimax regret is at least $\Omega(C_4H^2) = \Omega(K'H)$. This concludes the proof. □

With Lemma 4, we are ready to prove Theorem 4. Let $\mathcal{M}$ be the hard instance for $K' = \max\left\{\frac{1}{10}Kp, 1\right\}$ constructed in the proof of Lemma 4. We construct an MDP $\mathcal{M}'$ as below.

- In the first step, for any state s, with probability p, the leaner transitions to a copy of $\mathcal{M}$, and with probability $1-p$, the learner transitions to a dumb state with 0 reward.

It can be easily verified that $v^\star \leq pH$. Let $X = X_1 + X_2 + \cdots + X_k$, 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. Given that $g(X, K')$ is non-decreasing in X, one sees that

$$\mathsf{Regret}(K) \geq \mathbb{E}\big[g(X, K')\big].$$

- In the case where $Kp \geq 10$, Lemma 8 tells us that with probability at least $1/2$, $X \geq \frac{1}{10}Kp = K'$, and then it holds that

$$\mathbb{E}\big[g(X,K')\big] \geq \frac{1}{2}g(K',K') = \frac{1}{2}f(K') = \frac{1}{2}\Omega\left(\min\left\{\sqrt{SAH^3K'}, K'H\right\}\right) = \Omega(\sqrt{SAH^3Kp}, KHp).$$

- In the case where $Kp < 10$, with probability exceeding $1-(1-p)^K \geq (1-e^{-Kp}) \geq \frac{Kp}{30}$, one has $X \geq 1$. Then one has

$$\mathbb{E}\big[g(X,K')\big] \geq \frac{Kp}{30} \cdot g(1,K') = \frac{Kp}{30} \cdot g(1,1) = \Omega(KHp).$$

The preceding bounds taken together complete the proof.

7.2 Proof of Theorem 5

Without loss of generality, assume that $S = A = 2$ (as in the proof of Theorem 4). Note that $p \leq 1/4$. We would like to construct a hard instance for which the learner needs to identify the correct action for each step. Let $\mathcal{S} = \{s_1, s_2\}$, and take the initial state to be s_1. The transition kernel and cost are chosen as follows.

- For any action a and h, set $P_{s_2,a,h} = e_{s_2}$ and $c_h(s_2,a) = 0$.
- For any action $a \neq a^\star$ and h, set $P_{s_1,a,h} = e_{s_2}$ and $c_h(s_2,a) = 1$.
- Set $P_{s_1,a^\star,h} = e_{s_1}$ and $c_h(s_1,a^\star) = p$.

It can be easily checked that $c^\star = Hp$ by choosing $a^\star$ for each step. To identify the correct action $a^\star$ for at least half of the H steps, we need $\Omega(H)$ episodes, which implies that, there exists a constant $C_5 > 0$ such that in the first $K \leq C_5 H$ episodes, the cost of the learner is at least $\frac{H(1-p)}{2}$. Then the minimax regret is at least

$$\Omega\big(K(H-c^\star)\big) = \Omega\big(KH^2(1-p)\big)$$

when $K \leq C_5 H$. In the case where $C_5 H \leq K \leq \frac{100H}{p}$, the minimax regret is at least

$$\Omega\big(H(H-c^\star)\big) = \Omega\big(H^2(1-p)\big).$$

For $K \geq \frac{100H}{p}$, we let $\mathcal{M}$ be the hard instance with the same transition as that in the proof of Lemma 4, and set the cost as $1'2$ for state x and 1 for state y with respect to $K' = Kp/10 \geq 10H$. Let $\mathcal{M}'$ be the MDP such that: in the first step, with probability p, the learner transitions to a copy of $\mathcal{M}$, and with probability $1-p$, the learner transitions to a dumb state with 0 cost. Then we have $c^\star = \Theta(Hp)$. It follows from Lemma 8 that, with probability exceeding $1/2$, one has $X \geq \frac{1}{3}Kp - \log 2 \geq \frac{1}{6}Kp$. Then one has

$$\mathsf{Regret}(K) \geq \frac{1}{2}\Omega\left(\min\left\{\sqrt{H^3K'}, K'H\right\}\right) = \Omega\big(\sqrt{H^3Kp}\big).$$

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

7.3 Proof of Theorem 6

When $K \geq SAH/p$, the lower bound in Theorem 4 readily applies because the regret is at least $\Omega(\sqrt{SAH^3Kp})$ and the variance var is at most pH^2. When $SAH \leq K \leq SAH/p$, the regret is at least $\Omega(SAH^2) = \Omega(\min\{\sqrt{SAH^3Kp} + SAH^2, KH\})$. As a result, it suffices to focus on the case where $1 \leq K \leq SAH$, Towards this end, we only need the following lemma, which suffices for us to complete the proof.

Lemma 5. *Consider any $1 \leq K \leq SAH$. There exists an MDP instance 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. Let us construct an MDP with deterministic transition; more precisely, 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 $s'' \neq s'$. The reward function is also chosen to be 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 $P_{s_2,a,h} = e_{s_2}$ and $r_h(s_2, a) = 0$. For any action $a \neq a^\star$ and h, we also set $P_{s_1,a,h} = e_{s_2}$ and $r_h(s_2, a) = 0$. At last, we set $P_{s_1,a^\star,h} = e_{s_1}$ and $r_h(s_1, a^\star) = 1$. In other words, there are a dumb state and a normal state in each step. The learner would naturally hope to find the correct action to avoid the dumb state. Obviously, $V_1^\star(s_1) = H$. To find an $\frac{H}{2}$-optimal policy, the learner needs to identify $a^\star$ for the first $\frac{H}{2}$ steps, requiring at least $\Omega(HA)$ rounds in expectation. As a result, the minimax regret is at least $\Omega(KH)$ when $K \leq cHA$ for some proper constant $c > 0$.

Let us refer to the hard instance above as a *hard chain*. For general S, we can construct $d \coloneqq \frac{S}{2}$ hard chains. Let the two states in the i-th hard chain 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 $V_1^\star(s_1(i)) = H$ holds for any $1 \leq i \leq d$. Let $\mathsf{Regret}_i(K)$ be the expected regret resulting from the i-th hard chain. When $K \geq 100S$, Lemma 8 tells us that with probability at least $\frac{1}{2}$, $s_1(i)$ is visited for at least $\frac{K}{10S} \geq 10$ times. As a result, we have

$$\mathsf{Regret}_i(K) \geq \frac{1}{2} \cdot \Omega\left(\frac{KH}{S}\right).$$

Summing over i, we see that the total regret is at least $\sum_{i=1}^d \mathsf{Regret}_i(K) = \Omega(KH)$. When $K < 100S$, with probability at least $1 - (1 - \frac{1}{S})^K \geq 0.0001\frac{K}{S}$, we know that $s_1(i)$ is visited for at least one time. Therefore, it holds that $\mathsf{Regret}_i(K) \geq \Omega(\frac{KH}{S})$. Summing over i, we obtain

$$\mathsf{Regret}(K) = \sum_{i=1}^K \mathsf{Regret}_i(K) = \Omega(KH)$$

as claimed. □

8 Conclusion

A central issue in online reinforcement learning (RL) is data efficiency. While several recent works have achieved asymptotically minimal regret in this domain, optimality is typically guaranteed only in a "large-sample" regime, resulting in significant burn-in costs for algorithms to function effectively. Addressing the challenge of achieving minimax-optimal regret without incurring these costs has remained an open problem in RL theory.

In this paper, we investigate the sample complexity and regret behavior of online RL in time-inhomogeneous finite-horizon Markov Decision Processes (MDPs). We settle the aforementioned problem by proving that a modified version of the Monotonic Value Propagation (`MVP`) algorithm achieves regret bounds of the order

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

where S is the number of states, A is the number of actions, H is the horizon length, and K is the total number of episodes. This result matches the minimax lower bound across the entire range of sample sizes $K \geq 1$, effectively eliminating the need for a burn-in period.

Additionally, we extend our analysis to highlight the influence of various problem-dependent quantities, such as optimal value/cost and variance statistics, on fundamental performance limits. The key technical innovation lies in a novel analysis paradigm, utilizing a new concept called "profiles," which helps to decouple complicated statistical dependencies—a long-standing challenge in the analysis of online RL in sample-hungry regimes.

By providing deeper insights into regret minimization in online RL and introducing robust algorithmic techniques, our findings contribute to advancing both the theory and practical applications of reinforcement learning.

References

[AJ17] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1184–1194. Curran Associates, Inc., 2017.

[AKL+17] Alekh Agarwal, Akshay Krishnamurthy, John Langford, Haipeng Luo, et al. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, pages 4–7, 2017.

[AKY20] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83, 2020.

[AMK13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272, 2017.

[AZBL18] Zeyuan Allen-Zhu, Sébastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, pages 186–194, 2018.

[Ber19] Dimitri Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

[BS12] Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208, 2012.

[BT03] Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3(Oct):213–231, March 2003.

[BT09] Peter L Bartlett and Ambuj Tewari. 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))*, 2009.

[CJJJL21] Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021.

[CMSS20] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020.

[CY21] Qiwen Cui and Lin F Yang. Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504, 2021.

[CYJW19] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

[DB15] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

[DLB17] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[DMKV21] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598, 2021.

[DMMZ21] Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

[DWCW19] Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*, 2019.

[EM03] Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

[EMGM19] Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. *Advances in Neural Information Processing Systems*, 32, 2019.

[FPLO18] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *ICML 2018-The 35th International Conference on Machine Learning*, volume 80, pages 1578–1586, 2018.

[Fre75] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118, 1975.

[JA18] Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

[JAZBJ18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[JKSY20] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. *International Conference on Machine Learning*, 2020.

[JL23] Xiang Ji and Gen Li. Regret-optimal model-free reinforcement learning for discounted MDPs with short burn-in time. *Advances in neural information processing systems*, 2023.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[JYW21] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096, 2021.

[Kak03] Sham M Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

[KN09] J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

[KS98a] Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

[KS98b] Michael J Kearns and Satinder P Singh. Near-optimal reinforcement learning in polynominal time. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 260–268, 1998.

[LCC+24] Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024.

[LCWC22] Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in Markov games with a generative model. *Advances in Neural Information Processing Systems*, 35:15353–15367, 2022.

[LH12] Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

[LKTF20] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[LLWZ20] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533, 2020.

[LSC+21] Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[LSC+23] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *accepted to the Annals of Statistics*, 2023.

[LWC+21] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2021.

[LWCC24] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

[LWY21] Yuanzhi Li, Ruosong Wang, and Lin F Yang. Settling the horizon-dependence of sample complexity in reinforcement learning. In *IEEE Symposium on Foundations of Computer Science*, 2021.

[LYCF24] Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning. *Conference on Learning Theory (COLT)*, 2024.

[MDSV21] Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. UCB momentum Q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618, 2021.

[MP09] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

[NPB20] Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*, 2020.

[ORVR13] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[PBPH+20] Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

[PW20] Ashwin Pananjady and Martin J Wainwright. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

[QW20] Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of Thirty Third Conference on Learning Theory (COLT)*, 2020.

[RLD+21] Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34:15621–15634, 2021.

[Rus19] Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14433–14443, 2019.

[RZM+21] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[SJ19] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

[SL08] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[SLW+06] Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. ACM, 2006.

[SLW+22] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025, 2022.

[SLW+23] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 2023.

[SS10] István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, 2010.

[SWW+18] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. 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, 2018.

[SWWY18] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. 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, 2018.

[TM18] Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *arXiv preprint arXiv:1803.01626*, 2018.

[TPL21] Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon MDPs. *arXiv preprint arXiv:2106.13013*, 2021.

[TZD+21] Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34, 2021.

[Wai19a] Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

[Wai19b] Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

[WCD22] Xinqi Wang, Qiwen Cui, and Simon S Du. On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14865–14877, 2022.

[WCS+22] Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429, 2022.

[WDYK20] Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? In *Advances in Neural Information Processing Systems*, 2020.

[WZW+23] Kaiwen Wang, Kevin Zhou, Runzhe Wu, Nathan Kallus, and Wen Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. *arXiv preprint arXiv:2305.15703*, 2023.

[XJW+21] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[XMD21] Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular MDPs via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472, 2021.

[XSC+22] Zhihan Xiong, Ruoqi Shen, Qiwen Cui, Maryam Fazel, and Simon S Du. Near-optimal randomized exploration for tabular markov decision processes. *Advances in Neural Information Processing Systems*, 35:6358–6371, 2022.

[YDWW22] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

[YLCF23] Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219, 2023.

[YYD21] Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584, 2021.

[ZB19] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

[ZHZ+23] Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. *arXiv preprint arXiv:2302.10371*, 2023.

[ZJD21] Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531, 2021.

[ZJD22] Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904, 2022.

[ZZD23] Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pages 42878–42914, 2023.

[ZZJ20] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, 2020.

[ZZJ21] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662, 2021.