Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regret

Yingjie Fei[1] **Zhuoran Yang**[2] **Yudong Chen**[3] **Zhaoran Wang**[1] **Qiaomin Xie**[3]
[1] Northwestern University; yf275@cornell.edu, zhaoranwang@gmail.com
[2] Princeton University; zy6@princeton.edu
[3] Cornell University; {yudong.chen, qiaomin.xie}@cornell.edu

Abstract

We study risk-sensitive reinforcement learning in episodic Markov decision processes with unknown transition kernels, where the goal is to optimize the total reward under the risk measure of exponential utility. We propose two provably efficient model-free algorithms, Risk-Sensitive Value Iteration (RSVI) and Risk-Sensitive Q-learning (RSQ). These algorithms implement a form of risk-sensitive optimism in the face of uncertainty, which adapts to both risk-seeking and risk-averse modes of exploration. We prove that RSVI attains an $\tilde{O}\big(\lambda(|\beta|H^2)\cdot\sqrt{H^3S^2AT}\big)$ regret, while RSQ attains an $\tilde{O}\big(\lambda(|\beta|H^2)\cdot\sqrt{H^4SAT}\big)$ regret, where $\lambda(u)=(e^{3u}-1)/u$ for $u>0$. In the above, β is the risk parameter of the exponential utility function, S the number of states, A the number of actions, T the total number of timesteps, and H the episode length. On the flip side, we establish a regret lower bound showing that the exponential dependence on $|\beta|$ and H is unavoidable for any algorithm with an $\tilde{O}(\sqrt{T})$ regret (even when the risk objective is on the same scale as the original reward), thus certifying the near-optimality of the proposed algorithms. Our results demonstrate that incorporating risk awareness into reinforcement learning necessitates an exponential cost in $|\beta|$ and H, which quantifies the fundamental tradeoff between risk sensitivity (related to aleatoric uncertainty) and sample efficiency (related to epistemic uncertainty). To the best of our knowledge, this is the first regret analysis of risk-sensitive reinforcement learning with the exponential utility.

1 Introduction

Risk-sensitive reinforcement learning (RL) concerns learning to act in a dynamic environment while taking into account risks that arise during the learning process. Effective management of risks in RL is critical to many real-world applications such as autonomous driving [32], real-time strategy games [56], financial investment [44], etc. In neuroscience, risk-sensitive RL has been applied to model human behaviors in decision making [46, 52].

In this paper, we consider risk-sensitive RL with the exponential utility [34] under episodic Markov decision processes (MDPs) with unknown transition kernels. Informally, the agent aims to maximize a risk-sensitive objective function of the form

$$V=\frac{1}{\beta}\log\left\{\mathbb{E}e^{\beta R}\right\}, \tag{1}$$

where R is the total reward the agent receives, and $\beta\neq 0$ is a real-valued parameter that controls risk preference of the agent; see Equation (2) for a formal definition of V. The objective V admits the Taylor expansion $V=\mathbb{E}[R]+\frac{\beta}{2}\mathrm{Var}(R)+O(\beta^2)$. It can be seen that for $\beta>0$ the agent

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

is risk-seeking (favoring high uncertainty in R), for $\beta < 0$ the agent is risk-averse (favoring low uncertainty in R), and a larger $|\beta|$ implies higher risk-sensitivity. When $\beta \to 0$, the agent tends to be risk-neutral and the objective reduces to the expected reward objective $V = \mathbb{E}[R]$ standard in RL. Therefore, the risk-sensitive objective in (1) covers the entire spectrum of risk sensitivity by varying β. In addition, the formulation (1) is closely related to RL with constraints. For example, a negative risk parameter β controls the tail of a risk distribution so as to mitigate the chance of receiving a total reward R that is excessively low. We refer to [42, Section 2.1] for an in-depth discussion of this connection.

The challenge of risk-sensitive RL lies both in the non-linearity of the objective function and in designing a risk-aware exploration mechanism. In particular, as we elaborate in Section 2.2, the non-linear objective function (1) induces a non-linear Bellman equation. Classical RL algorithms are inappropriate in this setting, as their design crucially relies on the linearity of Bellman equations. On the other hand, effective exploration has been well known to be crucial to RL algorithm design, yet it is not clear how to design an algorithm that efficiently explores uncertain environments while at the same time adapting to the risk-sensitive objective (1) of agents with different risk parameter β.

To address these difficulties, we propose two model-free algorithms, Risk-Sensitive Value Iteration (RSVI) and Risk-Sensitive Q-learning (RSQ). Specifically, RSVI is a batch algorithm and RSQ is an online algorithm; both families of batch and online algorithms see broad applications in practice. We demonstrate in Section 3 that our proposed algorithms implement a form of risk-sensitive optimism for exploration. Importantly, the exact implementation of optimism depends on both the magnitude and the sign of the risk parameter, and therefore applies to both risk-seeking and risk-averse modes of learning. Letting $\lambda(u) = (e^{3u} - 1)/u$ for $u > 0$, we prove that RSVI attains an $\tilde{O}\big(\lambda(|\beta|H^2) \cdot \sqrt{H^3S^2AT}\big)$ regret, and RSQ achieves an $\tilde{O}\big(\lambda(|\beta|H^2) \cdot \sqrt{H^4SAT}\big)$ regret. Here, S and A are the numbers of states and actions, respectively, T is the total number of timesteps, and H is the length of each episode. These regret bounds interpolate across different regimes of risk sensitivity and subsume existing results under the risk-neutral setting. Compared with risk-neutral RL (corresponding to $\beta \to 0$), our general regret bounds feature an exponential dependency on $|\beta|$ and H, even though the risk-sensitive objective (1) is on the same scale as the total reward; see Figure 1 for a plot of the exponential factor $\lambda(|\beta|H^2)$. Complementarily, we prove a lower bound showing that such an exponential dependency is inevitable for any algorithm and thus certifies the near-optimality of the proposed algorithms. To the best of our knowledge, our work provides the first regret analysis of risk-sensitive RL with the exponential utility.

Our upper and lower bounds demonstrate the fundamental tradeoff between risk sensitivity and sample efficiency in RL.[1] Broadly speaking, risk sensitivity is associated with *aleatoric* uncertainty, which originates from the inherent randomness of state transition, actions and rewards, whereas sample efficiency is associated with *epistemic* uncertainty, which arises from imperfect knowledge of the environment/system and can be reduced by more exploration [20, 24]. These two notions of uncertainty are usually decoupled in the regret analysis of risk-neutral RL—in particular, using the expected reward as the objective effectively suppresses the aleatoric uncertainty. In risk-sensitive RL, we establish that there is a fundamental connection and tradeoff between these two forms of uncertainty: the risk-seeking and risk-averse regimes both incur an exponential cost in $|\beta|$ and H on the regret, whereas the regret is polynomial in H in the risk-neutral regime.

Our contributions. The contributions of our work can be summarized as follows:

- We consider the problem of risk-sensitive RL with the exponential utility. We propose two provably efficient model-free algorithms, namely RSVI and RSQ, that implement risk-sensitive optimism in the face of uncertainty;
- We provide regret analysis for both algorithms over the entire spectrum of risk parameter β. As $\beta \to 0$, we show that our results recover the existing regret bounds in the risk-neutral setting;
- We provide a lower bound result that certifies the near-optimality of our upper bounds and reveals a fundamental tradeoff between risk sensitivity and sample complexity.

[1] By standard arguments, regret can be translated into sample complexity bounds and vice versa; see [38].

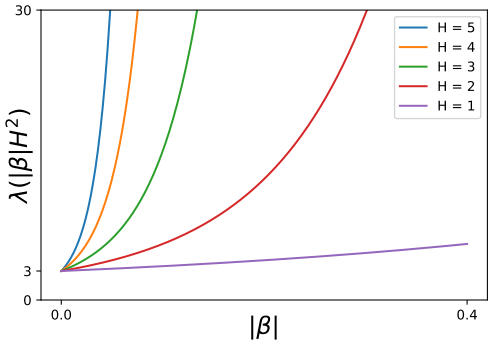

Figure 1: Scaling of $\lambda(|\beta|H^2)$ in risk sensitivity $|\beta|$ for different values of episode length H.

Related work. RL with risk-sensitive utility functions have been studied in several work. The work [45] proposes TD(0) and Q-learning-style algorithms that transform temporal differences instead of cumulative rewards, and proves their convergence. Risk-sensitive RL with a general family of utility functions is studied in [52], which also proposes a Q-learning algorithm with convergence guarantees. The work of [28] studies a risk-sensitive policy gradient algorithm, though with no theoretical guarantees. We remark that while substantial work has been devoted to designing risk-sensitive RL algorithms and proving their convergence, the issues of exploration, sample efficiency and regret bounds have rarely been studied. Our work narrows this gap in the literature by studying regret bounds of model-free algorithms for risk-sensitive RL.

The exponential utility has also been been investigated in the more classical setting of MDPs. Following the seminal work of [34], this line of work includes [7, 9–11, 14, 21, 25, 29, 30, 33, 43, 48, 51, 58, 61]. Note that these papers impose more restrictive assumptions and study different types of results than ours. Specifically, they assume known transition kernels or access to simulators, and they do not conduct finite-time or finite-sample analysis. Another related direction to ours is RL with risk/safety constraints studied by [1, 2, 16–19, 26, 27, 49, 54, 59, 62], and readers are also referred to [31] for an excellent survey on this topic. Compared to our work, that line of work focuses on constrained RL problems with different risk criteria. Other related problems include risk-sensitive games [5, 6, 8, 15, 35, 37, 40, 57], and risk-sensitive bandits [13, 22, 23, 42, 50, 53, 55, 60, 63]. Bandit problems are special cases of the RL problem that we investigate, with both the number of states and episode length being equal to one. As such, both our settings and results are more general than those obtained in bandit problems.

Notations. For a positive integer n, let $[n] \coloneqq \{1, 2, \ldots, n\}$. For two non-negative sequences $\{a_i\}$ and $\{b_i\}$, we write $a_i \lesssim b_i$ if there exists a universal constant $C > 0$ such that $a_i \leq C b_i$ for all i. We write $a_i \asymp b_i$ if $a_i \lesssim b_i$ and $b_i \lesssim a_i$. We use $\tilde{O}(\cdot)$ to denote $O(\cdot)$ while hiding logarithmic factors.

2 Problem setup

2.1 Episodic MDPs and risk-sensitive objective

We consider the setting of episodic MDPs, denoted by MDP$(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ is the set of possible actions, H is the length of each episode, and $\mathcal{P} = \{P_h\}_{h\in[H]}$ and $\mathcal{R} = \{r_h\}_{h\in[H]}$ are the sets of state transition kernels and reward functions, respectively. In particular, for each $h \in [H]$, $P_h(\cdot \,|\, s, a)$ is the distribution of the next state if action a is taken in state s at step h. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite discrete spaces, and let $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ denote their cardinalities. We assume that the agent does not have access to $\{P_h\}$ and that each $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic function.

An agent interacts with an episodic MDP as follows. At the beginning of each episode, an initial state s_1 is chosen arbitrarily by the environment. In each step $h \in [H]$, the agent observes a state $s_h \in \mathcal{S}$, chooses an action $a_h \in \mathcal{A}$, and receives a reward $r_h(s_h, a_h)$. The MDP then transitions into a new

state $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$. We use the convention that the episode terminates when a state s_{H+1} at step $H+1$ is reached, at which the agent does not take an action and receives no reward.

A policy $\pi = \{\pi_h\}_{h\in[H]}$ of an agent is a sequence of functions $\pi_h : \mathcal{S} \to \mathcal{A}$, where $\pi_h(s)$ is the action that the agent takes in state s at step h of an episode. For each $h \in [H]$, we define the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π as the expected value of cumulative rewards the agent receives under a risk measure of exponential utility by executing policy π starting from an arbitrary state at step h. Specifically, we have

$$V_h^\pi(s) \coloneqq \frac{1}{\beta} \log\left\{ \mathbb{E}\left[\exp\left(\beta \sum_{h'=h}^{H} r_{h'}(s_{h'}, \pi_{h'}(s_{h'})) \right) \,\middle|\, s_h = s \right] \right\}, \tag{2}$$

for each $(h, s) \in [H] \times \mathcal{S}$. Here $\beta \neq 0$ is the risk parameter of the exponential utility: $\beta > 0$ corresponds to a risk-seeking value function, $\beta < 0$ corresponds to a risk-averse value function, and as $\beta \to 0$ the agent tends to be risk-neutral and we recover the classical value function $V_h^\pi(s) = \mathbb{E}[\sum_{h=1}^{H} r_h(s_h, \pi_h(s_h)) \mid s_h = s]$ in RL. The goal of the agent is to find a policy π such that $V_1^\pi(s)$ is maximized for all state $s \in \mathcal{S}$. Note the logarithm and rescaling by $1/\beta$ in the above definition, which puts the objective $V_1^\pi(s)$ on the same scale as the total reward; this scaling property is made formal in Lemma 1 below.

2.2 Bellman equations and regret

We further define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which gives the expected value of the risk measured by the exponential utility when the agent starts from an arbitrary state-action pair at step h and follows policy π afterwards; that is,

$$Q_h^\pi(s, a) \coloneqq \frac{1}{\beta} \log\left\{ \exp(\beta \cdot r_h(s, a)) \mathbb{E}\left[\exp\left(\beta \sum_{h'=h+1}^{H} r_{h'}(s_{h'}, a_{h'}) \right) \,\middle|\, s_h = s, a_h = a \right] \right\},$$

for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. The Bellman equation associated with policy π is given by

$$\begin{aligned} Q_h^\pi(s, a) &= r_h(s, a) + \frac{1}{\beta} \log\left\{ \mathbb{E}_{s' \sim P_h(\cdot \mid s, a)} \left[\exp\left(\beta \cdot V_{h+1}^\pi(s') \right) \right] \right\}, \\ V_h^\pi(s) &= Q_h^\pi(s, \pi_h(s)), \qquad V_{H+1}^\pi(s) = 0, \end{aligned} \tag{3}$$

which holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Under some mild regularity conditions, there always exists an optimal policy π^* which gives the optimal value $V_h^*(s) = \sup_\pi V_h^\pi(s)$ for all $(h, s) \in [H] \times \mathcal{S}$ [7]. The Bellman optimality equation is given by

$$\begin{aligned} Q_h^*(s, a) &= r_h(s, a) + \frac{1}{\beta} \log\left\{ \mathbb{E}_{s' \sim P_h(\cdot \mid s, a)} \left[\exp\left(\beta \cdot V_{h+1}^*(s') \right) \right] \right\}, \\ V_h^*(s) &= \max_{a \in \mathcal{A}} Q_h^*(s, a), \qquad V_{H+1}^*(s) = 0. \end{aligned} \tag{4}$$

This equation implies that the optimal policy π^* is the greedy policy with respect to the optimal action-value function $\{Q_h^*\}_{h\in[H]}$. Hence, to find the optimal policy π^*, it suffices to estimate the optimal action-value function. We note that both Bellman equations (3) and (4) are non-linear in the value and action-value functions due to non-linearity of the exponential utility. This is in contrast with their linear risk-neutral counterparts.

Under the episodic MDP setting, the agent aims to learn the optimal policy by interacting with the environment throughout a set of episodes. For each $k \geq 1$, let us denote by s_1^k the initial state chosen by the environment and π^k the policy chosen simultaneously by the agent at the beginning of episode k. The difference in values between $V_1^{\pi^k}(s_1^k)$ and $V_1^*(s_1^k)$ measures the expected regret or the sub-optimality of the agent in episode k. After K episodes, the total regret for the agent is

$$\text{Regret}(K) \coloneqq \sum_{k\in[K]} \left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right]. \tag{5}$$

We record the following simple worst-case upper bounds on the value functions and regret.

Lemma 1. *For any $(h, s, a) \in \mathcal{S} \times \mathcal{A} \times [H]$, policy π and risk parameter $\beta \neq 0$, we have*

$$0 \leq V_h^\pi(s) \leq H \quad \text{and} \quad 0 \leq Q_h^\pi(s, a) \leq H. \tag{6}$$

Consequently, for each $K \geq 1$, all policy sequences $\pi^1, \ldots, \pi^K$ and any $\beta \neq 0$, we have

$$0 \leq \text{Regret}(K) \leq KH. \tag{7}$$

Proof. Recall the assumption that the reward functions $\{r_h\}$ are bounded in $[0, 1]$. The lower bounds are immediate by definition. For the upper bound, we have $V_h^\pi(s) \leq \frac{1}{\beta} \log \{\mathbb{E}[\exp(\beta H)]\} = H$. Upper bounds for Q_h^π and the regret follow similarly. □

While straightforward, the above lemma highlights an important point: the risk and regret are on the same scale as the reward. In particular, the upper bounds above are *independent* of β and *linear* in the horizon length H—the same as in the standard MDP setting—because the $\log$ and $\exp$ functions in the definition of the objective function (2) cancel with each other in the worst case. Therefore, the exponential dependence of the regret on $|\beta|$ and H, which we establish below in Section 4, is not merely a consequence of scaling but rather is inherent in the risk-sensitive setting.

3 Algorithms

The non-linearity of the Bellman equations, discussed in Section 2.2, creates challenges in algorithmic design. In particular, standard model-free algorithms such as least-squares value iteration (LSVI) and Q-learning are no longer appropriate since they specialize to the risk-neutral setting with linear Bellman equations. In this section, we present risk-sensitive LSVI and Q-learning algorithms that adapt to both the non-linear Bellman equations and any valid risk parameter β.

3.1 Risk-Sensitive Value Iteration

We first present Risk-Sensitive Value Iteration (RSVI) in Algorithm 1. Algorithm 1 is inspired by LSVI-UCB of [39], which is in turn motivated by the idea of LSVI [12, 47] and the classical value-iteration algorithm. Like LSVI-UCB, Algorithm 1 applies the Upper Confidence Bound (UCB) by incorporating a bonus term to value estimates of state-action pairs, which therefore implements the principle of Optimism in the Face of Uncertainty (OFU) [36].

Mechanism of Algorithm 1. The algorithm mainly consists of the value estimation step (Line 6–13) and the policy execution step (Line 14–18). In Line 7, the algorithm computes the intermediate value w_h by a least-squares update

$$w_h \leftarrow \operatorname*{argmin}_{w \in \mathbb{R}^{SA}} \sum_{\tau \in [k-1]} \left[e^{\beta[r_h(s_h^\tau, a_h^\tau) + V_{h+1}(s_{h+1}^\tau)]} - w^\top \phi(s_h^\tau, a_h^\tau) \right]^2. \tag{8}$$

Here, $\{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}_{\tau \in [k-1]}$ are accessed from the dataset $\mathcal{D}_h$ for each $h \in [H]$, and $\phi(\cdot, \cdot)$ denotes the canonical basis in $\mathbb{R}^{SA}$. Line 7 can be efficiently implemented by computing sample means of $e^{\beta[r_h(s,a) + V_{h+1}(s')]}$ over those state-action pairs that the algorithm has visited. Therefore, it can also be interpreted as estimating the sample means of exponentiated Q-values under visitation measures induced by the transition kernels $\{P_h\}$. This is a typical feature of the family of batch algorithms, to which Algorithm 1 belongs. Then, in Line 10, the algorithm uses the intermediate value w_h to compute the estimate Q_h, by adding/subtracting bonus b_h and thresholding the sum/difference at $e^{\beta(H-h+1)}$, depending on the sign of β. It is not hard to see that the logarithmic-exponential transformation in Line 10 conforms and adapts to the non-linearity in Bellman equations (3) and (4). In addition, the thresholding operator ensures that the estimated action-value function Q_h of step h stays in the range $[0, H - h + 1]$ and so does the estimated value function V_h in Line 11. This is to enforce the estimates Q_h and V_h to be on the same scale as the optimal Q_h^* and V_h^*.

Besides the logarithmic-exponential transformation, another distinctive feature of Algorithm 1 is the way the bonus term $b_h > 0$ is incorporated in Line 10. At first sight, it might appear counter-intuitive to *subtract* b_h from w_h when $\beta < 0$. We demonstrate next that subtracting bonus when $\beta < 0$ in fact implements the idea of OFU in a risk-sensitive fashion.

Algorithm 1 RSVI

Input: number of episodes $K \in \mathbb{Z}_{>0}$, confidence level $\delta \in (0, 1]$, and risk parameter $\beta \neq 0$

1: $Q_h(s,a) \leftarrow H - h + 1$ and $N_h(s,a) \leftarrow 0$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$
2: $Q_{H+1}(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
3: Initialize datasets $\{\mathcal{D}_h\}$ as empty
4: **for** episode $k = 1, \ldots, K$ **do**
5: $\quad V_{H+1}(s) \leftarrow 0$ for each $s \in \mathcal{S}$
6: $\quad$ **for** step $h = H, \ldots, 1$ **do** ▷ *value estimation*
7: $\quad\quad$ Update w_h via Equation (8)
8: $\quad\quad$ **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $N_h(s,a) \geq 1$ **do**
9: $\quad\quad\quad b_h(s,a) \leftarrow c_\gamma \left|e^{\beta H} - 1\right| \sqrt{\frac{S \log(2SAT/\delta)}{N_h(s,a)}}$ for some universal constant $c_\gamma > 0$
10: $\quad\quad\quad Q_h(s,a) \leftarrow \begin{cases} \frac{1}{\beta} \log\left[\min\{e^{\beta(H-h+1)}, w_h(s,a) + b_h(s,a)\}\right], & \text{if } \beta > 0; \\ \frac{1}{\beta} \log\left[\max\{e^{\beta(H-h+1)}, w_h(s,a) - b_h(s,a)\}\right], & \text{if } \beta < 0 \end{cases}$
11: $\quad\quad\quad V_h(s) \leftarrow \max_{a' \in \mathcal{A}} Q_h(s,a')$
12: $\quad\quad$ **end for**
13: $\quad$ **end for**
14: $\quad$ **for** step $h = 1, \ldots, H$ **do** ▷ *policy execution*
15: $\quad\quad$ Take action $a_h \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_h(s_h, a)$ and observe $r_h(s_h, a_h)$ and s_{h+1}
16: $\quad\quad$ $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$
17: $\quad\quad$ Insert (s_h, a_h, s_{h+1}) into $\mathcal{D}_h$
18: $\quad$ **end for**
19: **end for**

Risk-Sensitive Upper Confidence Bound. For the purpose of illustration, let us consider a "promising" state $s^+ \in \mathcal{S}$ at step h that allows us to transition to states $\{s'\}$ in the next step with high values $\{V_{h+1}(s')\}$ regardless of actions taken. This means that the intermediate value $w_h(s^+, \cdot) \propto \sum_{s'} e^{\beta \cdot V_{h+1}(s')}$ tends to be *small,* given that $\beta < 0$ and $\{V_{h+1}(s')\}$ are large. By subtracting a positive b_h from w_h, we obtain an even smaller quantity $w_h(s^+, \cdot) - b_h(s^+, \cdot)$. We can then deduce that $Q_h(s^+, \cdot) \approx \frac{1}{\beta} \log[w_h(s^+, \cdot) - b_h(s^+, \cdot)]$ is *larger* compared to $\frac{1}{\beta} \log[w_h(s^+, \cdot)]$ which does not incorporate bonus, since the logarithmic function is monotonic and again $\beta < 0$ (we ignore thresholding for the moment). Therefore, subtracting bonus serves as a UCB for $\beta < 0$. Since the exact form of the UCB depends on both the magnitude and sign of β (as shown in Lines 9 and 10), we name it Risk-Sensitive Upper Confidence Bound (RS-UCB) and this results in what we call Risk-Sensitive Optimism in the Face of Uncertainty (RS-OFU).

3.2 Risk-Sensitive Q-learning

Although Algorithm 1 is model-free, it requires storage of historical data $\{\mathcal{D}_h\}$ and computation over them (Line 7). A more efficient class of algorithms is Q-learning algorithms, which update Q values in an online fashion as each state-action pair is encountered. We therefore propose Risk-Sensitive Q-learning (RSQ) and formally describe it in Algorithm 2.

Mechanism of Algorithm 2. Algorithm 2 is based on Q-learning with UCB studied in the work of [38] and we use the same learning rates therein

$$\alpha_t := \frac{H+1}{H+t} \tag{9}$$

for every integer $t \geq 1$. Similar to Algorithm 1, Algorithm 2 consists of the policy execution step (Line 6) and value estimation step (Lines 9–11). Line 9 updates the intermediate value w_h in an online fashion, in constrast with the batch update in Line 7 of Algorithm 1, and Algorithm 2 can thus be seen as an online algorithm. Line 10 then applies the same logarithmic-exponential transform to the intermediate value and bonus as in Algorithm 1. Note the similar way we use the bonus term b_t in estimating Q-values in Line 10 of Algorithm 2 as in Line 10 of Algorithm 1. Algorithm 2 therefore also implements RS-UCB and follows the principle of RS-OFU.

Algorithm 2 RSQ

Input: number of episodes $K \in \mathbb{Z}_{>0}$, confidence level $\delta \in (0, 1]$, learning rates $\{\alpha_t\}$ and risk parameter $\beta \neq 0$

$Q_h(s,a), V_h(s,a) \leftarrow H - h + 1$ and $N_h(s,a) \leftarrow 0$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$
$Q_{H+1}(s,a), V_{H+1}(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
for episode $k = 1, \ldots, K$ **do**
  Receive the initial state s_1
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \operatorname{argmax}_{a' \in \mathcal{A}} Q_h(s_h, a')$, and observe $r_h(s_h, a_h)$ and s_{h+1}
    $t = N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$
    $b_t \leftarrow c\left|e^{\beta H} - 1\right| \sqrt{\frac{H \log(SAT/\delta)}{t}}$ for some sufficiently large universal constant $c > 0$
    $w_h(s_h, a_h) \leftarrow (1 - \alpha_t) e^{\beta \cdot Q_h(s_h, a_h)} + \alpha_t e^{\beta [r_h(s_h, a_h) + V_{h+1}(s_{h+1})]}$
    $Q_h(s_h, a_h) \leftarrow \begin{cases} \frac{1}{\beta} \log\left[\min\{e^{\beta(H-h+1)}, w_h(s_h, a_h) + \alpha_t b_t\}\right], & \text{if } \beta > 0; \\ \frac{1}{\beta} \log\left[\max\{e^{\beta(H-h+1)}, w_h(s_h, a_h) - \alpha_t b_t\}\right], & \text{if } \beta < 0 \end{cases}$
    $V_h(s_h) \leftarrow \max_{a' \in \mathcal{A}} Q_h(s_h, a')$
  end for
end for

Comparisons of Algorithms 1 and 2. It is interesting to compare the bonuses used in Algorithms 1 and 2. The bonuses in both algorithms depend on the risk parameter β through a common factor $\left|e^{\beta H} - 1\right|$. A careful analysis (see our proofs in appendices) on the bonuses and the value estimation steps reveals that the effective bonuses added to the estimated value function is proportional to $\frac{e^{|\beta| H} - 1}{|\beta|}$. This means that the more risk-seeking/averse an agent is (or the larger $|\beta|$ is), the larger bonus it needs to compensate for its uncertainty over the environment. Such risk sensitivity of the bonus is also reflected in the regret bounds; see Theorems 1 and 2 below. Also, it is not hard to see that both algorithms have polynomial time and space complexities in S, A, K and H. Moreover, thanks to its online update procedure, Algorithm 2 is more efficient than Algorithms 1 in both time and space complexities, since it does not require storing historical data (in particular, $\{\mathcal{D}_h\}$ of Algorithm 1) nor computing statistics based on them for value estimation.

4 Main results

In this section, we first present regret bounds for Algorithms 1 and 2, and then we complement the results with a lower bound on regret that any algorithm has to incur.

4.1 Regret upper bounds

The following theorem gives an upper bound for regret incurred by Algorithm 1. Let $T \coloneqq KH$ be the total number of timesteps for which an algorithm is run, and recall the function $\lambda(u) \coloneqq (e^{3u} - 1)/u$.

Theorem 1. *For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the regret of Algorithm 1 is bounded by*

$$\text{Regret}(K) \lesssim \lambda(|\beta| H^2) \cdot \sqrt{H^3 S^2 A T \log^2(2SAT/\delta)}.$$

The proof is given in Appendix C. We see that the result of Theorem 1 adapts to both risk-seeking ($\beta > 0$) and risk-averse ($\beta < 0$) settings through a common factor of $\lambda(|\beta| H^2)$.

As $\beta \to 0$, the setting of risk-sensitive RL tends to that of standard and risk-neutral RL, and we have an immediate corollary to Theorem 1 as a precise characterization.

Corollary 1. *Under the setting of Theorem 1 and when $\beta \to 0$, with probability at least $1 - \delta$, the regret of Algorithm 1 is bounded by*

$$\text{Regret}(K) \lesssim \sqrt{H^3 S^2 A T \log^2(2SAT/\delta)}.$$

Proof. The result follows from Theorem 1 and the fact that $\lim_{\beta \to 0} \lambda(|\beta| H^2) = 3$. □

The result in Corollary 1 recovers the regret bound of [4, Theorem 2] under the standard RL setting and is nearly optimal compared to the minimax rates presented in [3, Theorems 1 and 2]. Corollary 1 also reveals that Theorem 1 interpolates between the risk-sensitive and risk-neutral settings.

Next, we give a regret upper bound for Algorithm 2 in the following theorem.

Theorem 2. *For any $\delta \in (0, 1]$, with probability at least $1 - \delta$ and when T is sufficiently large, the regret of Algorithm 2 is bounded by*

$$\text{Regret}(K) \lesssim \lambda(|\beta|H^2) \cdot \sqrt{H^4SAT \log(SAT/\delta)}.$$

The proof is given in Appendix E. Similarly to Theorem 1, Theorem 2 also covers both risk-seeking and risk-averse settings via the same factor $\lambda(|\beta|H^2)$, which gives the risk-neutral bound when $\beta \to 0$ as shown in the following.

Corollary 2. *Under the setting of Theorem 2 and when $\beta \to 0$, with probability at least $1 - \delta$, the regret of Algorithm 2 is bounded by*

$$\text{Regret}(K) \lesssim \sqrt{H^4SAT \log(SAT/\delta)}.$$

The proof follows the same reasoning as in that of Corollary 1. According to Corollary 2, the regret upper bound for Algorithm 2 matches the nearly optimal result in [38, Theorem 2] under the risk-neutral setting. As such, Theorems 1 and 2 strictly generalizes the existing nearly optimal regret bounds (up to polynomial factors).

The crux of the proofs of both Theorems 1 and 2 lies in a local linearization argument for the non-linear Bellman equations and non-linear updates of the algorithms, in which action-value and value functions are related by a logarithmic-exponential transformation. Although logarithmic and exponential functions are not Lipschitz globally, we show that they are locally Lipschitz in the domain of our interest, and their combined local Lipschitz factors turn out to be the exponential factors in the theorems. Once the Bellman equations and algorithm estimates are linearized, we can apply standard techniques in RL to obtain the final regret. It is noteworthy that, as suggested by [38], the regret bounds in Theorems 1 and 2 can automatically be translated into sample complexity bounds in the probably approximately correct (PAC) setting, which did not previously exist even given access to a simulator.

In the risk-sensitive setting where β is bounded away from 0, our regret bounds of Theorems 1 and 2 depend exponentially in the horizon length H and the risk sensitivity $|\beta|$. In what follows, we argue that such exponential dependence is unavoidable.

4.2 Regret lower bound

We now present a fundamental lower bound on the regret, which complements the upper bounds in Theorems 1 and 2.

Theorem 3. *If $|\beta|\,(H-1)$ and K are sufficiently large, the regret of any policy obeys*

$$\text{Regret}(K) \gtrsim \lambda(|\beta|(H-1)/6) \cdot \sqrt{HT}.$$

The proof is given in Appendix F. In the proof, we construct an MDP that can be reduced to a bandit problem. We then show that any bandit algorithm has to incur an expected regret, in terms of the logarithmic-exponential objective, that grows as predicted in Theorem 3.

Theorem 3 shows that the exponential dependence on the $|\beta|$ and H in Theorems 1 and 2 is essentially indispensable. In addition, it features a sub-linear dependence on T through the $\tilde{O}(\sqrt{T})$ factor. In view of Theorem 3, therefore, both Theorems 1 and 2 are nearly optimal in their dependence on β, H and T. One should contrast Theorem 3 with Lemma 1, which shows that the worst-case regret is linear in H and T. Such a linear regret can be attained by any trivial algorithm that does not learn at all. In sharp contrast, in order to achieve the optimal $\sqrt{T}$ scaling (which by standard arguments implies a finite sample-complexity bound), an algorithm must incur a regret that is exponential in H. Therefore, our results show a (perhaps surprising) tradeoff between risk sensitivity and sample efficiency.

Broader Impact

This work contributes to the risk-awareness of machine learning and improves the way RL algorithms handle risks arising from uncertain environments. We have proposed two efficient and model-free algorithms for risk-sensitive RL with the exponential utility. We show that both algorithms follow the principle of Risk-Sensitive Optimism in the Face of Uncertainty (RS-OFU), and they achieve nearly optimal regret bounds with respect to the risk parameter, horizon length and total number of timesteps.

Acknowledgments and Disclosure of Funding

Y. Chen is partially supported by NSF grants 1657420 and 1704828. Q. Xie is partially supported by NSF grant 1955997. The other co-authors have no funding to disclose.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. JMLR.org, 2017.

[2] Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

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

[4] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. *arXiv preprint arXiv:2002.04017*, 2020.

[5] Arnab Basu and Mrinal K. Ghosh. Zero-sum risk-sensitive stochastic differential games. *Mathematics of Operations Research*, 37(3):437–449, 2012.

[6] Arnab Basu and Mrinal Kanti Ghosh. Zero-sum risk-sensitive stochastic games on a countable state space. *Stochastic Processes and their Applications*, 124(1):961–983, 2014.

[7] Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

[8] Nicole Bäuerle and Ulrich Rieder. Zero-sum risk-sensitive stochastic games. *Stochastic Processes and their Applications*, 127(2):622–642, 2017.

[9] Vivek S. Borkar. A sensitivity formula for risk-sensitive cost and the actor-critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001.

[10] Vivek S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2):294–311, 2002.

[11] Vivek S. Borkar and Sean P. Meyn. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

[12] Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1-3):33–57, 1996.

[13] Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Conference on Learning Theory*, pages 1295–1306, 2018.

[14] Rolando Cavazos-Cadena and Emmanuel Fernández-Gaucherand. The vanishing discount approach in Markov chains with risk-sensitive criteria. *IEEE Transactions on Automatic Control*, 45(10):1800–1816, 2000.

[15] Rolando Cavazos-Cadena and Daniel Hernández-Hernández. The vanishing discount approach in a class of zero-sum finite games with risk-sensitive average criterion. *SIAM Journal on Control and Optimization*, 57(1):219–240, 2019.

[16] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. In *Advances in Neural Information Processing Systems*, pages 3509–3517, 2014.

[17] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[18] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

[19] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems*, pages 1522–1530, 2015.

[20] William R Clements, Benoît-Marie Robaglia, Bastien Van Delft, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

[21] Stefano P. Coraluppi and Steven I. Marcus. Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica*, 35(2):301–309, 1999.

[22] Eric V. Denardo, Eugene A Feinberg, and Uriel G Rothblum. The multi-armed bandit, with constraints. *Annals of Operations Research*, 208(1):37–62, 2013.

[23] Eric V. Denardo, Haechurl Park, and Uriel G. Rothblum. Risk-sensitive and risk-neutral multiarmed bandits. *Mathematics of Operations Research*, 32(2):374–394, 2007.

[24] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *arXiv preprint arXiv:1710.07283*, 2017.

[25] Giovanni B. Di Masi and Lukasz Stettner. Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, 38(1):61–78, 1999.

[26] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R Jovanović. Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*, 2020.

[27] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[28] Hannes Eriksson and Christos Dimitrakakis. Epistemic risk-sensitive reinforcement learning. *arXiv preprint arXiv:1906.06273*, 2019.

[29] Emmanuel Fernández-Gaucherand and Steven I. Marcus. Risk-sensitive optimal control of hidden Markov models: Structural results. *IEEE Transactions on Automatic Control*, 42(10):1418–1422, 1997.

[30] Wendell H Fleming and William M McEneaney. Risk-sensitive control on an infinite time horizon. *SIAM Journal on Control and Optimization*, 33(6):1881–1915, 1995.

[31] Michael Fu et al. Risk-sensitive reinforcement learning: A constrained optimization viewpoint. *arXiv preprint arXiv:1810.09126*, 2018.

[32] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[33] Daniel Hernández-Hernández and Steven I. Marcus. Risk sensitive control of Markov processes in countable state space. *Systems & Control Letters*, 29(3):147–155, 1996.

[34] Ronald A. Howard and James E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

[35] Wenjie Huang, Pham Viet Hai, and William B. Haskell. Model and algorithm for time-consistent risk-aware Markov games. *arXiv preprint arXiv:1901.04882*, 2019.

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

[37] Anna Jaśkiewicz and Andrzej S. Nowak. Stationary Markov perfect equilibria in risk sensitive stochastic overlapping generations models. *Journal of Economic Theory*, 151:411–447, 2014.

[38] 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.

[39] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

[40] Margriet B. Klompstra. Nash equilibria in risk-sensitive dynamic games. *IEEE Transactions on Automatic Control*, 45(7):1397–1401, 2000.

[41] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[42] Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

[43] Steven I. Marcus, Emmanual Fernández-Gaucherand, Daniel Hernández-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive Markov decision processes. In *Systems and Control in the Twenty-first Century*, pages 263–279. Springer, 1997.

[44] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[45] Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2-3):267–290, 2002.

[46] Yael Niv, Jeffrey A. Edlund, Peter Dayan, and John P. O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562, 2012.

[47] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

[48] Takayuki Osogami. Robustness and risk-sensitivity in Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 233–241, 2012.

[49] Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual optimization: Stochastically constrained Markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.

[50] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[51] Yun Shen, Wilhelm Stannat, and Klaus Obermayer. Risk-sensitive Markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

[52] Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 2014.

[53] Wen Sun, Debadeepta Dey, and Ashish Kapoor. Safety-aware algorithms for adversarial contextual bandit. In *International Conference on Machine Learning*, pages 3280–3288. JMLR. org, 2017.

[54] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, pages 1468–1476, 2015.

[55] Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

[56] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[57] Qingda Wei. Nonzero-sum risk-sensitive finite-horizon continuous-time stochastic games. *Statistics & Probability Letters*, 147:96–104, 2019.

[58] Peter Whittle. *Risk-sensitive Optimal Control*, volume 20. Wiley New York, 1990.

[59] Tengyang Xie, Bo Liu, Yangyang Xu, Mohammad Ghavamzadeh, Yinlam Chow, Daoming Lyu, and Daesub Yoon. A block coordinate ascent algorithm for mean-variance optimization. In *Advances in Neural Information Processing Systems*, pages 1065–1075, 2018.

[60] Jia Yuan Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[61] Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *arXiv preprint arXiv:1910.09496*, 2019.

[62] Liyuan Zheng and Lillian J Ratliff. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*, 2020.

[63] Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.

Appendices

A Preliminaries

We set some notations and shorthands before the proofs. For both Algorithms 1 and 2, we let s_h^k, a_h^k, w_h^k, Q_h^k and V_h^k denote the values of s_h, a_h, w_h, Q_h and V_h in episode k, and we denote by N_h^k the value of N_h at the end of episode $k-1$. For Algorithm 1, we let $\mathcal{D}_h^k$ be the value of $\mathcal{D}_h$ at the end of episode $k-1$. Next, we introduce a simple yet powerful result.

Fact 1. *Consider $x, y, b \in \mathbb{R}$ such that $x \geq y$.*

(a) if $y \geq g$ for some $g > 0$, then $\log(x) - \log(y) \leq \frac{1}{g}(x-y)$;

(b) Assume further that $y \geq 0$. If $b \geq 0$ and $x \leq u$ for some $u > 0$, then $e^{bx} - e^{by} \leq be^{bu}(x-y)$; if $b < 0$, then $e^{by} - e^{bx} \leq (-b)(x-y)$.

Proof. The results follow from Lipschitz continuity of the functions $x \mapsto \log(x)$ and $x \mapsto e^{bx}$. □

We record a simple fact about exponential factors.

Fact 2. *Define $\lambda_0 \coloneqq \frac{e^{|\beta|H}-1}{|\beta|}$ and $\lambda_2 \coloneqq e^{|\beta|(H^2+H)}$. Then we have $\lambda_0\lambda_2 H \leq \frac{e^{3|\beta|H^2}-1}{|\beta|}$.*

B Proof warmup for Theorem 1

First, we set some notations and definitions. Define $d \coloneqq SA$, $\iota \coloneqq \log(2dT/\delta)$ for a given $\delta \in (0,1]$, and I to be the $d \times d$ identity matrix. To streamline some parts of the proof, we define $\phi(s,a)$ to be a vector in $\mathbb{R}^d$ whose (s,a)-th entry is equal to one and other entries equal to zero (so $\phi(s,a)$ is a canonical basis of $\mathbb{R}^{SA}$). Also let Λ_h^k be a diagonal matrix in $\mathbb{R}^{d\times d}$ with each (s,a)-th diagonal entry equal to $\max\{N_h^{k-1}(s,a), 1\}$. It can be seen that Λ_h^k is positive definite. We adopt the shorthands $\phi_h^\tau \coloneqq \phi(s_h^\tau, a_h^\tau)$ and $r_h^\tau \coloneqq r_h(s_h^\tau, a_h^\tau)$ for $(\tau, h) \in [K] \times [H]$.

From now on, we fix a tuple $(k,h) \in [K]\times[H]$ and then fix $(s,a) \in \mathcal{S}\times\mathcal{A}$ such that $N_h^{k-1}(s,a) \geq 1$. We also fix a policy π. We set

$$w_h^\pi = e^{\beta \cdot Q_h^\pi(\cdot,\cdot)}. \tag{10}$$

It can be verified that by the definition of $\phi(s,a)$, we have

$$\begin{aligned} Q_h^\pi(s,a) &= \frac{1}{\beta}\log\left(e^{\beta\cdot Q_h^\pi(s,a)}\right) \\ &= \frac{1}{\beta}\log\left(\left\langle \phi(s,a), e^{\beta\cdot Q_h^\pi(\cdot,\cdot)}\right\rangle\right) \\ &= \frac{1}{\beta}\log\left(\langle \phi(s,a), w_h^\pi\rangle\right), \end{aligned} \tag{11}$$

as well as

$$w_h^\pi(s,a) = e^{\beta\cdot Q_h^\pi(s,a)} = \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta\cdot Q_h^\pi(s_h^\tau, a_h^\tau)}\right]\right\rangle, \tag{12}$$

where the last step follows from the definition of Λ_h^k.

Let us define

$$q_1^+ \coloneqq \begin{cases} \left\langle \phi(s,a), w_h^k\right\rangle + b_h^k(s,a), & \text{if } \beta > 0, \\ \left\langle \phi(s,a), w_h^k\right\rangle - b_h^k(s,a), & \text{if } \beta < 0, \end{cases}$$

$$q_1 \coloneqq \begin{cases} \min\{e^{\beta(H-h+1)}, q_1^+\}, & \text{if } \beta > 0, \\ \max\{e^{\beta(H-h+1)}, q_1^+\}, & \text{if } \beta < 0. \end{cases}$$

By the definition of Λ_h^k and ϕ_h^k, observe that

$$w_h^k(s,a) = \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]} \right] \right\rangle. \tag{13}$$

Define

$$G_0 \coloneqq (Q_h^k - Q_h^\pi)(s,a) = \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\left\{\langle \phi(s,a), w_h^\pi\rangle\right\}, \tag{14}$$

and our goal is to derive lower and upper bounds for G_0. From Equation (14), we have

$$\begin{aligned}
G_0 &= \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\left\{ \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta\cdot Q_h^\pi(s_h^\tau, a_h^\tau)} \right] \right\rangle \right\} \\
&= \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\left\{ \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s_h^\tau, a_h^\tau)} e^{\beta[r_h^\tau + V_{h+1}^\pi(s')]} \right] \right\rangle \right\} \\
&=: \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_3\}.
\end{aligned}$$

The first step above holds by Equation (12), and the second step follows from Equation (3). In order to control G_0, we define an intermediate quantity

$$q_2 \coloneqq \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s_h^\tau, a_h^\tau)} e^{\beta[r_h^\tau + V_{h+1}^k(s')]} \right] \right\rangle;$$

in words, q_2 replaces the quantity V_{h+1}^π in q_3 by V_{h+1}^k. It can be seen that

$$G_0 = G_1 + G_2, \tag{15}$$

where

$$\begin{aligned}
G_1 &\coloneqq \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_2\}, \\
G_2 &\coloneqq \frac{1}{\beta}\log\{q_2\} - \frac{1}{\beta}\log\{q_3\}.
\end{aligned} \tag{16}$$

Note that G_0, G_1 and G_2 are all well-defined, according to the following result.

Lemma 2. *We have $q_i \in [\min\{1, e^{\beta(H-h+1)}\}, \max\{1, e^{\beta(H-h+1)}\}]$ for $i \in [3]$.*

Proof. We prove the result by focusing on q_1. By the definitions of Λ_h^k and ϕ, the (s,a)-th entry of the vector $(\Lambda_h^k)^{-1}\sum_{\tau\in[k-1]}\phi_h^\tau \cdot u_h^\tau$ equals $\frac{1}{N_h^{k-1}(s,a)}\sum_{\tau\in[k-1]} u_h^\tau \cdot \mathbb{I}\{(s_h^\tau, a_h^\tau) = (s,a)\}$ for any sequence $\{u_h^\tau\}_{\tau\in[k-1]}$. Then, the result follows from the fact that $e^{\beta[r_h^\tau + V_{h+1}^k(s')]} \in [\min\{1, e^{\beta(H-h)}\}, \max\{1, e^{\beta(H-h)}\}]$ for $(\tau, s') \in [K]\times\mathcal{S}$ and the definition of q_1. □

Therefore, we have the following equivalent form of Equation (14):

$$(Q_h^k - Q_h^\pi)(s,a) = G_1 + G_2. \tag{17}$$

Thanks to the identity (17), our goal is now to control G_1 and G_2, which is done in the following lemma.

Lemma 3. *For all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$ that satisifies $N_h^{k-1}(s,a) \geq 1$, there exist universal constants $c_1, c_\gamma > 0$ (where c_γ is used in Line 9 of Algorithm 1) such that*

$$0 \leq G_1 \leq c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot d\sqrt{\iota}\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}$$

with probability at least $1-\delta/2$. Furthermore, if $V_{h+1}^k(s') \geq V_{h+1}^\pi(s')$ for all $s' \in \mathcal{S}$, then we have

$$0 \leq G_2 \leq e^{|\beta|H} \cdot \mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)}[V_{h+1}^k(s') - V_{h+1}^\pi(s')].$$

Proof. **Case** $\beta > 0$**.** To control G_1, we note that $N_h^{k-1}(s,a) = \phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)$ and by Equation (13) we can compute

$$
\begin{aligned}
&\left|q_1^+ - q_2 - b_h^k(s,a)\right| \\
&= \left|\left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]} - \mathbb{E}_{s'\sim P_h(\cdot\,|\,s_h^\tau,a_h^\tau)} e^{\beta[r_h^\tau + V_{h+1}^k(s')]}\right]\right\rangle\right| \\
&= \left|\frac{1}{N_h^{k-1}(s,a)} \sum_{(s,a,s^+)\in\mathcal{D}_h^{k-1}} e^{\beta[r_h(s,a)+V_{h+1}^k(s^+)]} - \mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta[r_h(s,a)+V_{h+1}^k(s')]}\right| \\
&\le \frac{1}{N_h^{k-1}(s,a)} \sum_{(s,a,s^+)\in\mathcal{D}_h^{k-1}} \left|e^{\beta[r_h(s,a)+V_{h+1}^k(s^+)]} - \mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta[r_h(s,a)+V_{h+1}^k(s')]}\right| \\
&\le \frac{1}{N_h^{k-1}(s,a)} \sum_{t\in\left[N_h^{k-1}(s,a)\right]} c'\left|e^{\beta H} - 1\right| \sqrt{\frac{S\iota}{t}} \\
&\le \frac{1}{N_h^{k-1}(s,a)} \int_{t\in\left[0,N_h^{k-1}(s,a)\right]} c'\left|e^{\beta H} - 1\right| \sqrt{\frac{S\iota}{t}}\mathrm{d}t \\
&= \frac{1}{N_h^{k-1}(s,a)} \cdot c\left|e^{\beta H} - 1\right| \sqrt{S\iota \cdot N_h^{k-1}(s,a)} \\
&= c\left|e^{\beta H} - 1\right| \sqrt{S\iota} \cdot \sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)},
\end{aligned}
$$

where the fourth step holds by Lemma 6, and the last step holds by the definition of Λ_h^k; in the above, $c' > 0$ is a universal constant and $c = 2c'$. If we choose $c_\gamma = c$ in the definition of $b_h^k(s,a)$ in Line 9 of Algorithm 1, we have

$$
0 \le q_1^+ - q_2 \le 2c \cdot \left|e^{\beta H} - 1\right| \sqrt{S\iota} \cdot \sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}.
$$

Therefore, we have $q_1 \ge q_2$, and thus $G_1 \ge 0$, by the first inequality above, the definition of q_1 and Lemma 2 (in particular, $q_2 \le e^{\beta(H-h+1)}$). By Lemma 2 and Fact 1(a) (with $g = 1$, $x = q_1$ and $y = q_2$), we have

$$
G_1 \le \frac{1}{\beta}(q_1 - q_2) \le \frac{1}{\beta}(q_1^+ - q_2),
$$

which together with the second inequality displayed above implies the desired upper bound on G_1.

Now we control the term G_2. For $\beta > 0$, it is not hard to see that the assumption $V_{h+1}^k(s') \ge V_{h+1}^\pi(s')$ for all $s' \in \mathcal{S}$ implies that $q_2 \ge q_3$ and therefore $G_2 \ge 0$. We also have

$$
\begin{aligned}
G_2 &\le \frac{1}{\beta}(q_2 - q_3) \\
&\le e^{\beta H}\left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s_h^\tau,a_h^\tau)}[V_{h+1}^k(s') - V_{h+1}^\pi(s')]\right]\right\rangle \\
&= e^{|\beta|H}\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)}[V_{h+1}^k(s') - V_{h+1}^\pi(s')],
\end{aligned}
$$

where the first step holds by Fact 1(a) (with $g = 1$, $x = q_2$, and $y = q_3$) and the fact that $q_2 \ge q_3 \ge 1$ (with the last inequality suggested by Lemma 2), and the second step holds by Fact 1(b) (with $b = \beta$, $x = r_h^\tau + V_{h+1}^k(s)$, and $y = r_h^\tau + V_{h+1}^\pi(s)$) and $H \ge r_h^\tau + V_{h+1}^k(s) \ge r_h^\tau + V_{h+1}^\pi(s) \ge 0$.

Case $\beta < 0$**.** Similar to the case of $\beta > 0$, we have

$$
\begin{aligned}
&\left|q_1^+ - q_2 + b_h^k(s,a)\right| \\
&\le c \cdot \left|e^{\beta H} - 1\right| \sqrt{S\iota} \cdot \sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}.
\end{aligned}
$$

If we choose $c_\gamma = c$ in the definition of $b_h^k(s,a)$ in Line 9 of Algorithm 1, the above equation implies

$$
0 \le q_2 - q_1^+ \le 2c \cdot \left|e^{\beta H} - 1\right| \sqrt{S\iota} \cdot \sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}.
$$

Therefore, we have $q_1 \leq q_2$, and thus $G_1 \geq 0$, by the first inequality displayed above, the definition of q_1 and Lemma 2 (in particular, $q_2 \geq e^{\beta(H-h+1)}$). By Lemma 2 and Fact 1(a) (with $g = e^{\beta H}$, $x = q_2$ and $y = q_1$), we further have

$$\begin{aligned} G_1 &= \frac{1}{(-\beta)} \left(\log\{q_2\} - \log\{q_1\}\right) \\ &\leq \frac{e^{-\beta H}}{|\beta|}(q_2 - q_1) \\ &\leq \frac{e^{-\beta H}}{|\beta|}(q_2 - q_1^+), \end{aligned}$$

which together with the second inequality displayed above and the fact that $\left|e^{\beta H} - 1\right| = 1 - e^{\beta H}$ implies the desired upper bound on G_1.

Next we control G_2. The assumption $V_{h+1}^k(s') \geq V_{h+1}^\pi(s')$ for all $s' \in \mathcal{S}$ implies that $q_2 \leq q_3$ and therefore $G_2 \geq 0$. We also have

$$\begin{aligned} G_2 &= \frac{1}{(-\beta)} \left(\log\{q_3\} - \log\{q_2\}\right) \\ &\leq \frac{e^{-\beta H}}{(-\beta)}(q_3 - q_2) \\ &\leq e^{|\beta| H} \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau \in [k-1]} \phi_h^\tau \left[\mathbb{E}_{s' \sim P_h(\cdot \mid s_h^\tau, a_h^\tau)}[V_{h+1}^k(s') - V_{h+1}^\pi(s')]\right] \right\rangle \\ &= e^{|\beta| H} \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)}[V_{h+1}^k(s') - V_{h+1}^\pi(s')], \end{aligned}$$

where the second step holds by Fact 1(a) (with $g = e^{\beta H}$, $x = q_3$, and $y = q_2$) and the fact that $q_3 \geq q_2 \geq e^{\beta H}$ (with the last inequality suggested by Lemma 2), and the third step holds by Fact 1(b) (with $b = \beta$, $x = r_h^\tau + V_{h+1}^k(s)$, and $y = r_h^\tau + V_{h+1}^\pi(s)$) and $r_h^\tau + V_{h+1}^k(s) \geq r_h^\tau + V_{h+1}^\pi(s) \geq 0$.

The proof is hence completed. □

The next lemma establishes the dominance of Q_h^k over Q_h^*.

Lemma 4. *On the event of Lemma 3, we have $Q_h^k(s,a) \geq Q_h^\pi(s,a)$ for all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.*

Proof. For the purpose of the proof, we set $Q_{H+1}^\pi(s,a) = Q_{H+1}^*(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We fix a tuple $(k,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$ and use strong induction on h. The base case for $h = H+1$ is satisfied since $(Q_{H+1}^k - Q_{H+1}^\pi)(s,a) = 0$ for $k \in [K]$ by definition. Now we fix an $h \in [H]$ and assume that $0 \leq (Q_{h+1}^k - Q_{h+1}^*)(s,a)$. Moreover, by the induction assumption we have

$$V_{h+1}^k(s) = \max_{a' \in \mathcal{A}} Q_{h+1}^k(s,a') \geq \max_{a' \in \mathcal{A}} Q_{h+1}^\pi(s,a') \geq V_{h+1}^\pi(s). \tag{18}$$

We also assume that (s,a) satisfies $N_h^{k-1}(s,a) \geq 1$, since otherwise $Q_h^k(s,a) = H - h + 1 \geq Q_h^\pi(s,a)$ and we are done. This assumption and Equation (18) together imply $G_2 \geq 0$ by Lemma 3. We also have $G_1 \geq 0$ on the event of Lemma 3. Therefore, it follows that $(Q_h^k - Q_h^\pi)(s,a) \geq 0$ by Equation (17). The induction is completed and so is the proof. □

Lemma 4 leads to an immediate and important corollary.

Lemma 5. *For any $\delta \in (0,1]$, with probability at least $1 - \delta/2$, we have $V_h^k(s) \geq V_h^\pi(s)$ for all $(k,h,s) \in [K] \times [H] \times \mathcal{S}$.*

Proof. The result follows from Lemma 4 and Equation (18). □

B.1 Supporting lemmas

We first present a concentration result.

Lemma 6. *Define*

$$\bar{\mathcal{V}}_{h+1} \coloneqq \left\{\bar{V}_{h+1} : \mathcal{S} \to \mathbb{R} \mid \forall s \in \mathcal{S},\ \bar{V}_{h+1}(s) \in [\min\{e^{\beta(H-h)}, 1\}, \max\{e^{\beta(H-h)}, 1\}]\right\}.$$

There exists a universal constant $c > 0$ such that with probability $1-\delta$, we have

$$\left|e^{\beta[r_h(s_h^k,a_h^k)+\bar{V}(s_{h+1}^k)]} - \mathbb{E}_{s'\sim P_h(\cdot|s_h^k,a_h^k)} e^{\beta[r_h(s_h^k,a_h^k)+\bar{V}(s')]}\right| \le c\left|e^{\beta H}-1\right|\sqrt{\frac{S\iota}{N_h^k(s,a)}}$$

for all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$ and all $\bar{V} \in \bar{\mathcal{V}}_{h+1}$.

Proof. The proof follows the same reasoning as [4, Lemma 12]. □

The next few lemmas help control $\sum_{k\in[K]} (\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k$.

Lemma 7 ([39, Lemma D.2])**.** *Let $\{\phi_t\}_{t\ge 0}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\sup_{t\ge 0}\|\phi_t\| \le 1$. Let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix with $\lambda_{\min}(\Lambda_0) \ge 1$. For any $t \ge 0$, we define $\Lambda_t \coloneqq \Lambda_0 + \sum_{i\in[t]} \phi_i\phi_i^\top$. Then, we have*

$$\log\left[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right] \le \sum_{i\in[t]} \phi_i^\top \Lambda_{i-1}^{-1}\phi_i \le 2\log\left[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right].$$

Lemma 8. *Recall the definitions of ϕ_h^k and Λ_h^k. For any $h \in [H]$, we have*

$$\sum_{k\in[K]} (\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k \le 2d\iota,$$

where $\iota = \log(2dT/\delta)$

Proof. Define $\Gamma_h^k \coloneqq \lambda I + \sum_{\tau\in[k-1]} \phi_h^\tau (\phi_h^\tau)^\top$ with $\lambda = 1$. It is not hard to see that by the definition of Λ_h^k we have $\Lambda_h^k \preceq \Gamma_h^k$ for $h \in [H]$. Since $\lambda_{\min}(\Gamma_h^k) \ge 1$ and $\|\phi_h^k\| \le 1$ for all $(k,h) \in [K]\times[H]$, by Lemma 7 we have for any $h \in [H]$ that

$$\sum_{k\in[K]} (\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k \le \sum_{k\in[K]} (\phi_h^k)^\top (\Gamma_h^k)^{-1}\phi_h^k \le 2\log\left[\frac{\det(\Gamma_h^{k+1})}{\det(\Gamma_h^1)}\right].$$

Furthermore, note that $\|\Gamma_h^{k+1}\| = \|\lambda I + \sum_{\tau\in[k]} \phi_h^k(\phi_h^k)^\top\| \le \lambda + k$. This implies

$$\sum_{k\in[K]} (\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k \le 2d\log\left[\frac{\lambda+k}{\lambda}\right] \le 2d\iota,$$

as desired. □

C Proof of Theorem 1

Define $\delta_h^k \coloneqq V_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)$, and $\zeta_{h+1}^k \coloneqq \mathbb{E}_{s'\sim P_h(\cdot\,|\,s_h^k,a_h^k)}[V_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')] - \delta_{h+1}^k$. For any $(k,h) \in [K]\times[H]$, we have

$$\begin{aligned}
\delta_h^k &= (Q_h^k - Q_h^{\pi_k})(s_h^k, a_h^k) \\
&\le c_1 \cdot \frac{e^{|\beta|H}-1}{|\beta|} \cdot \sqrt{S\iota}\sqrt{\phi(s_h^k,a_h^k)^\top(\Lambda_h^k)^{-1}\phi(s_h^k,a_h^k)} \\
&\quad + e^{|\beta|H} \cdot \mathbb{E}_{s'\sim P_h(\cdot\,|\,s_h^k,a_h^k)}[V_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')]
\end{aligned}$$

$$= c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot \sqrt{S\iota} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} + e^{|\beta|H} (\delta_{h+1}^k + \zeta_{h+1}^k). \tag{19}$$

In the above equation, the first step holds by the construction of Algorithm 1 and the definition of $V_h^{\pi_k}$ in Equation (3); the second step is a consequence of combining Equation (17) as well as Lemmas 3 and 5; the last step follows from the definitions of δ_h^k and ζ_{h+1}^k.

Noting that $V_{H+1}^k(s) = V_{H+1}^{\pi_k}(s) = 0$ and the fact that $\delta_{h+1}^k + \zeta_{h+1}^k \geq 0$ implied by Lemma 5, we can continue by expanding the recursion in Equation (19) and get

$$\delta_1^k \leq \sum_{h \in [H]} e^{(|\beta|H)h} \zeta_{h+1}^k + c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot \sum_{h \in [H]} e^{(|\beta|H)(h-1)} \sqrt{S\iota} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)}. \tag{20}$$

Therefore, we have

$$\begin{aligned} \text{Regret}(K) &= \sum_{k \in [K]} \left[(V_1^* - V_1^{\pi_k})(s_1^k) \right] \\ &\leq \sum_{k \in [K]} \delta_1^k \\ &\leq e^{|\beta|H^2} \sum_{k \in [K]} \sum_{h \in [H]} \zeta_{h+1}^k \\ &\quad + c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot e^{|\beta|H^2} \cdot \sqrt{S\iota} \sum_{k \in [K]} \sum_{h \in [H]} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)}, \end{aligned} \tag{21}$$

where the second step holds by Lemma 5 with π therein set to the optimal policy, and in the last step we applied Equation (20) along with the Cauchy-Schwarz inequality.

We proceed to control the two terms in Equation (21). Since the construction of V_h^k is independent of the new observation s_h^k in episode k, we have that $\{\zeta_{h+1}^k\}$ is a martingale difference sequence satisfying $\left|\zeta_h^k\right| \leq 2H$ for all $(k, h) \in [K] \times [H]$. By the Azuma-Hoeffding inequality, we have for any $t > 0$,

$$\mathbb{P}\left(\sum_{k \in [K]} \sum_{h \in [H]} \zeta_{h+1}^k \geq t \right) \leq \exp\left(-\frac{t^2}{2T \cdot H^2} \right).$$

Hence, with probability $1 - \delta/2$, there holds

$$\sum_{k \in [K]} \sum_{h \in [H]} \zeta_{h+1}^k \leq \sqrt{2TH^2 \cdot \log(2/\delta)} \leq 2H\sqrt{T\iota}, \tag{22}$$

where $\iota = \log(2dT/\delta)$. For the second term in Equation (21), we apply Lemma 8 and the Cauchy-Schwarz inequality to obtain

$$\begin{aligned} &\sum_{k \in [K]} \sum_{h \in [H]} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \\ &\leq \sum_{h \in [H]} \sqrt{K} \sqrt{\sum_{k \in [K]} \phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \\ &\leq H\sqrt{2dK\iota}. \end{aligned} \tag{23}$$

Plugging Equations (22) and (23) back to Equation (21) yields

$$\text{Regret}(K) \leq e^{|\beta|H^2} \cdot 2H\sqrt{T\iota} + c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot e^{|\beta|H^2} \cdot H\sqrt{2dSK\iota^2}$$

$$\leq (c_1+2)\cdot\frac{e^{|\beta|H}-1}{|\beta|}\cdot e^{|\beta|H^2}\cdot\sqrt{2dHST\iota^2},$$

where the last step holds since $\frac{e^{|\beta|H}-1}{|\beta|}\geq H$. The proof is completed in view of Fact 2 and the identity $d=SA$.

D Proof warmup for Theorem 2

Recall the learning rates $\{\alpha_t\}$ defined in Equation (9). Define the quantities

$$\alpha_t^0 \coloneqq \prod_{j=1}^{t}(1-\alpha_j), \qquad \alpha_t^i \coloneqq \alpha_i\prod_{j=i+1}^{t}(1-\alpha_j) \tag{24}$$

for integers $i,t\geq 1$. By convention, we set $\alpha_t^0=1$ and $\sum_{i\in[t]}\alpha_t^i=0$ if $t=0$, and $\alpha_t^i=\alpha_i$ if $t<i+1$. Define the shorthand $\iota\coloneqq\log(SAT/\delta)$ for $\delta\in(0,1]$.

The following fact describes some key properties of the learning rates $\{\alpha_t\}$.

Fact 3. *The following properties hold for α_t^i.*

(a) $\frac{1}{\sqrt{t}}\leq\sum_{i\in[t]}\frac{\alpha_t^i}{\sqrt{i}}\leq\frac{2}{\sqrt{t}}$ *for every integer* $t\geq 1$.

(b) $\max_{i\in[t]}\alpha_t^i\leq\frac{2H}{t}$ *and* $\sum_{i\in[t]}(\alpha_t^i)^2\leq\frac{2H}{t}$ *for every integer* $t\geq 1$.

(c) $\sum_{t=i}^{\infty}\alpha_t^i=1+\frac{1}{H}$ *for every integer* $i\geq 1$.

(d) $\sum_{i\in[t]}\alpha_t^i=1$ *and* $\alpha_t^0=0$ *for every integer* $t\geq 1$, *and* $\sum_{i\in[t]}\alpha_t^i=0$ *and* $\alpha_t^0=1$ *for* $t=0$.

Proof. The first three facts can be found in [38, Lemma 4.1], and the last one follows from direct calculation in view of Equation (24). □

We also present a lemma that controls the deviation of the exponentiated value function from its expectation.

Lemma 9. *There exists a universal constant $c>0$ such that for any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$ and $k_1,\ldots,k_t<k$ with $t=N_h^k(s,a)$, we have*

$$\left|\frac{1}{\beta}\sum_{i\in[t]}\alpha_t^i\left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]}-\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)}e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}\right]\right|$$
$$\leq\frac{c\left|e^{\beta H}-1\right|}{|\beta|}\sqrt{\frac{H\iota}{t}}.$$

with probability at least $1-\delta$, and

$$\frac{1}{|\beta|}\sum_{i\in[t]}\alpha_t^i b_i\in\left[\frac{c\left|e^{\beta H}-1\right|}{|\beta|}\sqrt{\frac{H\iota}{t}},\frac{2c\left|e^{\beta H}-1\right|}{|\beta|}\sqrt{\frac{H\iota}{t}}\right].$$

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

$$\psi(i,k,h,s,a)\coloneqq e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]}-\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)}e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}$$

Let us fix a tuple $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$. It can be seen that $\{\mathbb{I}(k_i\leq K)\cdot\psi(i,k,h,s,a)\}_{i\in[\tau]}$ for $\tau\in[K]$ is a martingale difference sequence. By the Azuma-Hoeffding inequality and a union bound over $\tau\in[K]$, we have with probability at least $1-\delta/(HSA)$, for all $\tau\in[K]$,

$$\left|\sum_{i\in[\tau]}\alpha_\tau^i\cdot\mathbb{I}(k_i\leq K)\cdot\psi(i,k,h,s,a)\right|$$

$$\leq \frac{c\left|e^{\beta H}-1\right|}{2}\sqrt{\iota\sum_{i\in[\tau]}(\alpha_\tau^i)^2} \leq c\left|e^{\beta H}-1\right|\sqrt{\frac{H\iota}{\tau}}$$

where $c>0$ is some universal constant, the first step holds since $r_h(s,a)+V_{h+1}^*(s')\in[0,H]$ for $s'\in\mathcal{S}$, and the last step follows from Fact 3(b). Since the above equation holds for all $\tau\in[K]$, it also holds for $\tau=t=N_h^k(s,a)\leq K$. Note that $\mathbb{I}(k_i\leq K)=1$ for all $i\in[N_h^k(s,a)]$. Therefore, applying another union bound over $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, we have that the following holds for all $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$ and with probability at least $1-\delta$:

$$\left|\sum_{i\in[t]}\alpha_\tau^i\cdot\psi(i,k,h,s,a)\right|\leq c\left|e^{\beta H}-1\right|\sqrt{\frac{H\iota}{t}}, \tag{25}$$

where $t=N_h^k(s,a)$. Using the fact that $r_h+V_{h+1}^*\in[0,H]$, we have

$$\begin{aligned}&\left|\frac{1}{\beta}\sum_{i\in[t]}\alpha_t^i\left[\mathbb{E}_{s'\sim\hat{P}_h^{k_i}(\cdot\,|\,s,a)}e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}-\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)}e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}\right]\right|\\ =&\left|\frac{1}{\beta}\sum_{i\in[t]}\alpha_t^i\cdot\psi(i,k,h,s,a)\right|\leq\frac{c\left|e^{\beta H}-1\right|}{|\beta|}\sqrt{\frac{H\iota}{t}}.\end{aligned}$$

To prove the result for $\frac{1}{|\beta|}\sum_{i\in[t]}\alpha_t^i b_i$, we recall the definition of $\{b_t\}$ in Line 8 of Algorithm 2 and compute

$$\begin{aligned}\frac{1}{|\beta|}\sum_{i\in[t]}\alpha_t^i b_i&=\frac{c\left|e^{\beta H}-1\right|}{|\beta|}\sum_{i\in[t]}\alpha_t^i\sqrt{\frac{H\iota}{i}}\\ &\in\left[\frac{c\left|e^{\beta H}-1\right|}{|\beta|}\sqrt{\frac{H\iota}{t}},\frac{2c\left|e^{\beta H}-1\right|}{|\beta|}\sqrt{\frac{H\iota}{t}}\right]\end{aligned}$$

where the last step holds by Fact 3(a). □

We fix a tuple $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$ with $k_i\leq k$ being the episode in which (s,a) is taken the i-th time at step h. Let us define

$$\begin{aligned}q_1^+&\coloneqq\begin{cases}\alpha_t^0e^{\beta(H-h+1)}+\sum_{i\in[t]}\alpha_t^i\left[e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]}+b_i\right], & \text{if }\beta>0,\\ \alpha_t^0e^{\beta(H-h+1)}+\sum_{i\in[t]}\alpha_t^i\left[e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]}-b_i\right], & \text{if }\beta<0,\end{cases}\\ q_1&\coloneqq\begin{cases}\min\{e^{\beta(H-h+1)},q_1^+\}, & \text{if }\beta>0,\\ \max\{e^{\beta(H-h+1)},q_1^+\}, & \text{if }\beta<0,\end{cases}\end{aligned}$$

and

$$\begin{aligned}q_2^+&\coloneqq\begin{cases}\alpha_t^0e^{\beta(H-h+1)}+\sum_{i\in[t]}\alpha_t^i\left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]}+b_i\right], & \text{if }\beta>0,\\ \alpha_t^0e^{\beta(H-h+1)}+\sum_{i\in[t]}\alpha_t^i\left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]}-b_i\right], & \text{if }\beta<0,\end{cases}\\ q_2&\coloneqq\begin{cases}\min\{e^{\beta(H-h+1)},q_2^+\}, & \text{if }\beta>0,\\ \max\{e^{\beta(H-h+1)},q_2^+\}, & \text{if }\beta<0,\end{cases}\\ q_2'&\coloneqq\alpha_t^0e^{\beta(H-h+1)}+\sum_{i\in[t]}\alpha_t^i\left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]}\right],\end{aligned}$$

and

$$q_3\coloneqq\alpha_t^0e^{\beta\cdot Q_h^*(s,a)}+\sum_{i\in[t]}\alpha_t^i\left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)}e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}\right].$$

We have a simple fact on q_2 and q_2'.

Fact 4. *If $\beta > 0$, we have $q_2' \le q_2$; if $\beta < 0$, we have $q_2' \ge q_2$.*

Proof. We focus on the case of $\beta > 0$. Note that $r_h(s,a) + V_{h+1}^*(s_{h+1}^{k_i}) \in [0, H-h+1]$, which implies $e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} \le e^{\beta(H-h+1)}$. We also have $\alpha_t^0, \sum_{i\in[t]} \alpha_t^i \in \{0,1\}$ with $\alpha_t^0 + \sum_{i\in[t]} \alpha_t^i = 1$ by Fact 3(d). These together imply that $q_2' \le e^{\beta H}$ and $q_2' - q_2^+ = -\sum_{i\in[t]} \alpha_t^i b_i \le 0$ by definition of b_i in Line 8 of Algorithm 2. Therefore, $q_2' \le \min\{e^{\beta(H-h+1)}, q_2^+\} = q_2$. The case of $\beta < 0$ can be proved in a similar way and thus omitted. □

Next, we establish a representation of the performance difference $(Q_h^k - Q_h^*)(s,a)$ using the quantities q_1 and q_3.

Lemma 10. For any $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$, let $t = N_h^k(s,a)$ and suppose (s,a) was previously taken at step h of episodes $k_1, \dots, k_t < k$. We have

$$(Q_h^k - Q_h^*)(s,a) = \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_3\}.$$

Proof. The Bellman optimality equation (4) implies

$$e^{\beta\cdot Q_h^*(s,a)} = e^{\beta\cdot r_h(s,a)}\left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta\cdot V_{h+1}^*(s')}\right].$$

By Fact 3(d), we have

$$e^{\beta\cdot Q_h^*(s,a)} = \alpha_t^0 e^{\beta\cdot Q_h^*(s,a)} + \sum_{i\in[t]} \alpha_t^i e^{\beta\cdot r_h(s,a)}\left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta\cdot V_{h+1}^*(s')}\right] = q_3$$

for each integer $t \ge 0$, and therefore

$$Q_h^*(s,a) = \frac{1}{\beta}\log\{q_3\}. \tag{26}$$

We finish the proof by combining Equation (26) and the fact that $Q_h^k(s,a) = \frac{1}{\beta}\log\{q_1\}$, which follows from Line 10 of Algorithm 2. □

We define the quantities

$$\begin{aligned} G_1 &\coloneqq \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_2\}, \\ G_2 &\coloneqq \frac{1}{\beta}\log\{q_2\} - \frac{1}{\beta}\log\{q_3\}, \end{aligned} \tag{27}$$

It is not hard to see that $(Q_h^k - Q_h^*)(s,a) = G_1 + G_2$ by Lemma 10. The next lemma establishes upper and lower bounds for $(Q_h^k - Q_h^*)(s,a)$.

Lemma 11. *For all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$ such that $t = N_h^k(s,a) \ge 1$, let*

$$\gamma_t \coloneqq 2\sum_{i\in[t]} \alpha_t^i b_i \cdot \begin{cases} \frac{1}{|\beta|}, & \text{if } \beta > 0, \\ \frac{e^{-\beta H}}{|\beta|}, & \text{if } \beta < 0, \end{cases}$$

and with probability at least $1-\delta$ we have

$$0 \le (Q_h^k - Q_h^*)(s,a) \le \alpha_t^0 H e^{|\beta|H} + \sum_{i\in[t]} \alpha_t^i e^{|\beta|H}\left[V_{h+1}^{k_i}(s_{h+1}^{k_i}) - V_{h+1}^*(s_{h+1}^{k_i})\right] + 2\gamma_t,$$

where $k_1, \dots, k_t < k$ are the episodes in which (s,a) was taken at step h, and $\gamma_t \le \frac{4c(e^{|\beta|H}-1)}{|\beta|}\sqrt{\frac{H\iota}{t}}$.

Proof. We prove the lower bound for $(Q_h^k - Q_h^*)(s,a)$ and then use it to prove the upper bound.

Lower bound for $Q^k - Q^*$.

For the purpose of the proof, we set $Q^k_{H+1}(s,a) = Q^*_{H+1}(s,a) = 0$ for all $(k,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$. We fix a $(s,a) \in \mathcal{S} \times \mathcal{A}$ and use strong induction on k and h. Without loss of generality, we assume that there exists a (k,h) such that $(s,a) = (s^k_h, a^k_h)$ (that is, (s,a) has been taken at some point in Algorithm 2), since otherwise $Q^k_h(s,a) = H - h + 1 \geq Q^*_h(s,a)$ for all $(k,h) \in [K] \times [H]$ and we are done. The base case for $k = 1$ and $h = H+1$ is satisfied since $(Q^{k'}_{H+1} - Q^*_{H+1})(s,a) = 0$ for $k' \in [K]$ by definition. We fix a $(k,h) \in [K] \times [H]$ and assume that $0 \leq (Q^{k_i}_{h+1} - Q^*_{h+1})(s,a)$ for each $k_1, \ldots, k_t < k$ (here $t = N^k_h(s,a)$). Then we have for $i \in [t]$ that

$$V^{k_i}_{h+1}(s) = \max_{a' \in \mathcal{A}} Q^{k_i}_{h+1}(s,a') \geq \max_{a' \in \mathcal{A}} Q^*_{h+1}(s,a') = V^*_{h+1}(s).$$

Recall the quantities G_1 and G_2 defined in Equation (27). The above equation implies $G_1 \geq 0$. We also have $G_2 \geq 0$ by the fact $Q^*_h(s,a) \leq H$ and on the event of Lemma 9. Therefore, it follows that $(Q^k_h - Q^*_h)(s,a) = G_1 + G_2 \geq 0$. The induction is completed and we have proved that $0 \leq (Q^k_h - Q^*_h)(s,a)$ for all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

Upper bound for $Q^k - Q^*$.

Let us fix a $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. Since $0 \leq (Q^k_h - Q^*_h)(s,a)$, we have for $i \in [t]$ that

$$V^{k_i}_{h+1}(s) = \max_{a' \in \mathcal{A}} Q^{k_i}_{h+1}(s,a') \geq \max_{a' \in \mathcal{A}} Q^*_{h+1}(s,a') = V^*_{h+1}(s).$$

Case $\beta > 0$. We have

$$\begin{aligned}
G_1 &= \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_2\} \\
&\leq \frac{1}{\beta}(q_1 - q_2) \\
&\leq \frac{1}{\beta}(q_1^+ - q_2') \\
&\leq \frac{1}{\beta}\sum_{i\in[t]} \alpha^i_t \left[e^{\beta[r_h(s,a) + V^{k_i}_{h+1}(s^{k_i}_{h+1})]} - e^{\beta[r_h(s,a) + V^*_{h+1}(s^{k_i}_{h+1})]}\right] + \frac{1}{\beta}\sum_{i\in[t]} \alpha^i_t b_i \\
&\leq e^{|\beta|H} \sum_{i\in[t]} \alpha^i_t \left[(V^{k_i}_{h+1} - V^*_{h+1})(s^{k_i}_{h+1})\right] + \gamma_t,
\end{aligned}$$

where the second step holds by Fact 1(a) with $g = 1$ and the fact that $V^{k_i}_{h+1}(s) \geq V^*_{h+1}(s)$ and by noticing that $\alpha^0_t, \sum_{i\in[t]} \alpha^i_t \in \{0,1\}$ with $\alpha^0_t + \sum_{i\in[t]} \alpha^i_t = 1$ by Fact 3(d) (so that $q_1 \geq q_2$), the third step holds since by definition $q_1^+ \geq q_1$ and by Fact 4 $q_2' \leq q_2$, and the last step holds by Fact 1(b) and the fact that $H \geq r_h(s,a) + V^{k_i}_{h+1}(s) \geq r_h(s,a) + V^*_{h+1}(s) \geq 0$. For G_2, we have

$$\begin{aligned}
G_2 &= \frac{1}{\beta}\log\{q_2\} - \frac{1}{\beta}\log\{q_3\} \\
&\leq \frac{1}{\beta}(q_2 - q_3) \\
&\leq \frac{1}{\beta}(q_2^+ - q_3) \\
&= \frac{\alpha^0_t}{\beta}\left[e^{\beta H} - e^{\beta \cdot Q^*_h(s,a)}\right] + \frac{1}{\beta}\sum_{i\in[t]} \alpha^i_t b_i \\
&\quad + \frac{1}{\beta}\sum_{i\in[t]} \alpha^i_t \left[e^{\beta[r_h(s,a) + V^*_{h+1}(s^{k_i}_{h+1})]} - \mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta[r_h(s,a) + V^*_{h+1}(s')]}\right] \\
&\leq \alpha^0_t H e^{|\beta|H} + \gamma_t,
\end{aligned}$$

In the above, the second step holds by Fact 1(a) with $g = 1$ and

$$\sum_{i\in[t]} \alpha^i_t b_i \geq \left|\sum_{i\in[t]} \alpha^i_t \left[e^{\beta[r_h(s,a) + V^*_{h+1}(s^{k_i}_{h+1})]} - \mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta[r_h(s,a) + V^*_{h+1}(s')]}\right]\right|$$

on the event of Lemma 9 (so that $q_2 \geq q_3$); the third step holds by Fact 4; the last step holds by Fact 1(b) and $Q_h^*(s,a) \in [0,H]$ and on the event of Lemma 9.

Case $\beta < 0$**.** We have

$$\begin{aligned}
G_1 &= \frac{1}{(-\beta)} \log\{q_2\} - \frac{1}{(-\beta)} \log\{q_1\} \\
&\leq \frac{e^{-\beta H}}{(-\beta)}(q_2 - q_1) \\
&\leq \frac{e^{-\beta H}}{(-\beta)}(q_2' - q_1^+) \\
&= \frac{e^{-\beta H}}{(-\beta)} \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} - e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]} \right] + \frac{e^{-\beta H}}{(-\beta)} \sum_{i\in[t]} \alpha_t^i b_i \\
&\leq e^{|\beta|H} \sum_{i\in[t]} \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) \right] + \gamma_t,
\end{aligned}$$

where the second step holds by Fact 1(a) with $g = e^{\beta H}$ and the fact that $V_{h+1}^{k_i}(s) \geq V_{h+1}^*(s)$ (so that $q_2 \geq q_1$), the third step holds since $q_2' \geq q_2$ by Fact 4 and $q_1^+ \leq q_1$ by definition, and the last step holds by Fact 1(b) and the fact that $H \geq r_h(s,a) + V_{h+1}^{k_i}(s) \geq r_h(s,a) + V_{h+1}^*(s) \geq 0$. For G_2, we have

$$\begin{aligned}
G_2 &= \frac{1}{(-\beta)} \log\{q_3\} - \frac{1}{(-\beta)} \log\{q_2\} \\
&\leq \frac{e^{-\beta H}}{(-\beta)}(q_3 - q_2) \\
&\leq \frac{e^{-\beta H}}{(-\beta)}(q_3 - q_2^+) \\
&= \frac{e^{-\beta H}}{(-\beta)} \alpha_t^0 \left[e^{\beta \cdot Q_h^*(s,a)} - e^{\beta H} \right] + \frac{e^{-\beta H}}{(-\beta)} \sum_{i\in[t]} \alpha_t^i b_i \\
&\quad + \frac{e^{-\beta H}}{(-\beta)} \sum_{i\in[t]} \alpha_t^i \left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta[r_h(s,a)+V_{h+1}^*(s')]} - e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} \right] \\
&\leq e^{-\beta H} \alpha_t^0 \left[H - Q_h^*(s,a) \right] + \frac{2e^{-\beta H}}{(-\beta)} \sum_{i\in[t]} \alpha_t^i b_i \\
&\leq \alpha_t^0 H e^{|\beta|H} + \gamma_t.
\end{aligned}$$

where the second step holds by Fact 1(a) given $q_3 \geq q_2$, the second to the last step holds by Fact 1(b), the fact that $Q_h^*(s,a) \leq H$ and on the event of Lemma 9, and the last step holds by the definition of γ_t.

Combining the bounds of G_1 and G_2 with the identity $(Q_h^k - Q_h^*)(s,a) = G_1 + G_2$ yields the upper bound for $(Q_h^k - Q_h^*)(s,a)$. The proof is completed in view of Lemma 9 and the definition of γ_t that imply

$$\gamma_t \leq \frac{4c(e^{|\beta|H} - 1)}{|\beta|} \sqrt{\frac{H\iota}{t}}.$$

□

E Proof of Theorem 2

We first introduce some notations. Let $\mathcal{G}$ be a discrete space. Define the shorthand

$$\mathsf{lse}_\beta(P, f) \coloneqq \frac{1}{\beta} \log\left\{ \mathbb{E}_{x\sim P} \left[\exp\left(\beta \cdot f(x) \right) \right] \right\}, \tag{28}$$

for a probability distribution P supported on $\mathcal{G}$ and function $f : \mathcal{G} \to \mathbb{R}$. We record a useful lemma that shows $\mathsf{lse}_\beta(\cdot, \cdot)$ is Lipschitz continuous in the second argument.

Lemma 12. *Let $\mathcal{G}$ be a discrete space and $\bar{f} \geq 0$ be a non-negative number. Let the functions $f, f' : \mathbb{R}^d \mapsto [0, \bar{f}]$ be such that $f(x) \geq f'(x)$ for all $x \in \mathbb{R}^d$. Also let P be a probability distribution supported on $\mathcal{G}$. We have*

$$\mathsf{lse}_\beta(P, f) - \mathsf{lse}_\beta(P, f') \leq e^{|\beta|\bar{f}} \cdot \mathbb{E}_{x\sim P}[f(x) - f'(x)].$$

The proof is given in Appendix E.1.

Define $\hat{P}_h^k(\cdot \mid s, a)$ to be the delta function centered at s_{h+1}^k for all $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, and this means $\mathbb{E}_{s'\sim\hat{P}_h^k(\cdot \mid s,a)}[f(s')] = f(s_{h+1}^k)$ for any function $f : \mathcal{S} \to \mathbb{R}$. We let

$$\delta_h^k \coloneqq (V_h^k - V_h^{\pi_k})(s_h^k) \quad \text{and} \quad \phi_h^k \coloneqq (V_h^k - V_h^*)(s_h^k).$$

Also define

$$\xi_{h+1}^k \coloneqq [(P_h - \hat{P}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})](s_h^k, a_h^k).$$

Note that For each $(k, h) \in [K] \times [H]$, we have

$$\begin{aligned}
\delta_h^k &= (Q_h^k - Q_h^{\pi_k})(s_h^k, a_h^k) \\
&= (Q_h^k - Q_h^*)(s_h^k, a_h^k) + (Q_h^* - Q_h^{\pi_k})(s_h^k, a_h^k) \\
&\leq \alpha_t^0 H e^{|\beta|H} + \sum_{i\in[t]} \alpha_t^i e^{|\beta|H} \phi_{h+1}^{k_i} + 2\gamma_t \\
&\quad + [\mathsf{lse}(P_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^*) - \mathsf{lse}(P_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^{\pi_k})] \\
&\leq \alpha_t^0 H e^{|\beta|H} + \sum_{i\in[t]} \alpha_t^i e^{|\beta|H} \phi_{h+1}^{k_i} + 2\gamma_t + e^{|\beta|H}[P_h(V_{h+1}^* - V_{h+1}^{\pi_k})](s_h^k, a_h^k) \\
&= \alpha_t^0 H e^{|\beta|H} + \sum_{i\in[t]} \alpha_t^i e^{|\beta|H} \phi_{h+1}^{k_i} + 2\gamma_t + e^{|\beta|H}(\delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k), \qquad (29)
\end{aligned}$$

where the third step holds by Lemma 11 and the Bellman equations (3) and (4), the fourth step holds by Lemma 12 and the fact that $0 \leq V_{h+1}^{\pi_k}(s) \leq V_{h+1}^*(s) \leq H$ for all $s \in \mathcal{S}$, and the last step follows by defintion that $\delta_{h+1}^k - \phi_{h+1}^k = (V_{h+1}^* - V_{h+1}^{\pi_k})(s_{h+1}^k) = [\hat{P}_h^k(V_{h+1}^* - V_{h+1}^{\pi_k})](s_h^k, a_h^k)$ and the definition of ξ_{h+1}^k.

We now compute $\sum_{k\in[K]} \delta_h^k$ for a fixed $h \in [H]$. Denote by $n_h^k \coloneqq N_h^k(s_h^k, a_h^k)$ and we have

$$\sum_{k\in[K]} \alpha_{n_h^k}^0 H e^{|\beta|H} = H e^{|\beta|H} \sum_{k\in[K]} \mathbb{I}\{n_h^k = 0\} \leq H e^{|\beta|H} SA.$$

Then we turn to control the second term in Equation (29) summed over $k \in [K]$, that is,

$$\sum_{k\in[K]} \sum_{i\in[t]} \alpha_t^i e^{|\beta|H} \phi_{h+1}^{k_i} = e^{|\beta|H} \sum_{k\in[K]} \sum_{i\in[n_h^k]} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(s_h^k, a_h^k)},$$

where $k_i(s_h^k, a_h^k)$ denotes the episode in which (s_h^k, a_h^k) was taken at step h for the i-th time. We re-group the above summation in a different way. For every $k' \in [K]$, the term $\phi_{h+1}^{k'}$ appears in the summand with $k > k'$ if and only if $(s_h^k, a_h^k) = (s_h^{k'}, a_h^{k'})$. The first time it appears we have $n_h^k = n_h^{k'} + 1$, the second time it appears we have $n_h^k = n_h^{k'} + 2$, and etc. Therefore,

$$e^{|\beta|H} \sum_{k\in[K]} \sum_{i\in[n_h^k]} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(s_h^k, a_h^k)} \leq e^{|\beta|H} \sum_{k'\in[K]} \phi_{h+1}^{k'} \sum_{t\geq n_h^{k'}+1} \alpha_t^{n_h^{k'}} \leq e^{|\beta|H} \left(1 + \frac{1}{H}\right) \sum_{k'\in[K]} \phi_{h+1}^{k'},$$

where the last step follows Fact 3(c). Collecting the above results and plugging them into Equation (29), we have

$$\sum_{k\in[K]} \delta_h^k \leq H e^{|\beta|H} SA + e^{|\beta|H} \left(1 + \frac{1}{H}\right) \sum_{k\in[K]} \phi_{h+1}^k$$

$$
\begin{aligned}
&+ e^{|\beta|H} \sum_{k\in[K]} (\delta_{h+1}^k - \phi_{h+1}^k) + \sum_{k\in[K]} (2\gamma_{n_h^k} + e^{|\beta|H}\xi_{h+1}^k) \\
&\leq He^{|\beta|H}SA + e^{|\beta|H}\left(1+\frac{1}{H}\right)\sum_{k\in[K]}\delta_{h+1}^k \\
&\quad + \sum_{k\in[K]} (2\gamma_{n_h^k} + e^{|\beta|H}\xi_{h+1}^k), \qquad (30)
\end{aligned}
$$

where the last step holds since $\delta_{h+1}^k \geq \phi_{h+1}^k$ (due to the fact that $V_{h+1}^*(s) \geq V_{h+1}^{\pi_k}(s)$ for all $x \in \mathcal{S}$). Since it holds that

$$
\left[e^{|\beta|H}\left(1+\frac{1}{H}\right)\right]^H \leq e^{|\beta|H^2+1},
$$

we can expand the quantity $\sum_{k\in[K]}\delta_1^k$ recursively in the form of Equation (30), apply Holder's inequality and use the fact that $\delta_{H+1}^k = 0$ to get

$$
\sum_{k\in[K]}\delta_1^k \leq e^{|\beta|H^2+1}\left[H^2e^{|\beta|H}SA + \sum_{h\in[H]}\sum_{k\in[K]}(2\gamma_{n_h^k} + e^{|\beta|H}\xi_{h+1}^k)\right]. \qquad (31)
$$

By the pigeonhole principle, for any $h \in [H]$ we have

$$
\begin{aligned}
\sum_{k\in[K]}\gamma_{n_h^k} &\lesssim \frac{e^{|\beta|H}-1}{|\beta|}\sum_{k\in[K]}\sqrt{\frac{H\iota}{n_h^k}} \\
&= \frac{e^{|\beta|H}-1}{|\beta|}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n\in[N_h^K(s,a)]}\sqrt{\frac{H\iota}{n}} \\
&\lesssim \frac{e^{|\beta|H}-1}{|\beta|}\sqrt{HSAK\iota} \\
&= \frac{e^{|\beta|H}-1}{|\beta|}\sqrt{SAT\iota}, \qquad (32)
\end{aligned}
$$

where the third step holds since $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_h^K(s,a) = K$ and the RHS of the second step is maximized when $N_h^K(s,a) = K/(SA)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Finally, the Azuma-Hoeffding inequality implies that with probability at least $1-\delta$, we have

$$
\left|\sum_{h\in[H]}\sum_{k\in[K]}\xi_{h+1}^k\right| \lesssim H\sqrt{T\iota}. \qquad (33)
$$

Putting together Equations (32) and (33) and plugging them into (31), we have

$$
\begin{aligned}
\sum_{k\in[K]}\delta_1^k &\lesssim e^{|\beta|(H^2+H)}\cdot H^2SA \\
&\quad + e^{|\beta|H^2}\cdot\frac{e^{|\beta|H}-1}{|\beta|}\sqrt{H^2SAT\iota} \\
&\quad + e^{|\beta|(H^2+H)}\cdot H\sqrt{T\iota}. \\
&\leq e^{|\beta|(H^2+H)}\cdot H^2SA \\
&\quad + e^{|\beta|(H^2+H)}\cdot\frac{e^{|\beta|H}-1}{|\beta|}\sqrt{H^2SAT\iota}
\end{aligned}
$$

The proof is completed in view of Fact 2 and when T is sufficiently large.

E.1 Proof of Lemma 12

We have the following two cases.

Case $\beta > 0$**.** We have

$$\begin{aligned}
\mathsf{lse}_\beta(P, f) - \mathsf{lse}_\beta(P, f') &\le \frac{1}{\beta}\mathbb{E}_{x\sim P}\left[e^{\beta\cdot f(x)} - e^{\beta\cdot f'(x)}\right] \\
&\le \frac{1}{\beta}\mathbb{E}_{x\sim P}\left[\beta e^{\beta\bar{f}}(f(x) - f'(x))\right] \\
&= e^{\beta\bar{f}}\cdot\mathbb{E}_{x\sim P}[f(x) - f'(x)],
\end{aligned}$$

where the first step holds by Fact 1(a) with $g = 1$ and the fact that $e^{\beta\cdot f(x)} \ge e^{\beta\cdot f'(x)} \ge 1$, and the second holds by Fact 1(b) with $u = \bar{f}$ and the fact that $f(x) \ge f'(x)$.

Case $\beta < 0$**.** We have

$$\begin{aligned}
\mathsf{lse}_\beta(P, f) - \mathsf{lse}_\beta(P, f') &= -\left[\mathsf{lse}_\beta(P, f') - \mathsf{lse}_\beta(P, f)\right] \\
&\le \frac{\exp(-\beta\bar{f})}{(-\beta)}\mathbb{E}_{x\sim P}\left[\exp(\beta\cdot f'(x)) - \exp(\beta\cdot f(x))\right] \\
&\le \frac{\exp(-\beta\bar{f})}{(-\beta)}\mathbb{E}_{x\sim P}\left[(-\beta)(f(x) - f'(x))\right] \\
&= \exp(-\beta\bar{f})\cdot\mathbb{E}_{x\sim P}[f(x) - f'(x)],
\end{aligned}$$

where the second step holds by Fact 1(a) with $g = e^{\beta\bar{f}}$ given that $x \in [e^{\beta\bar{f}}, 1]$, and the third step holds by Fact 1(b) and the fact $1 \ge e^{\beta\cdot f'(x)} \ge e^{\beta\cdot f(x)} > 0$.

F Proof of Theorem 3

We consider the following MDP as illustrated in Figure 2. For now, we focus on the case $\beta > 0$; we shall see soon that the construction for $\beta < 0$ can be done in a similar way. The MDP is equipped with $\mathcal{A} = \{a_1, a_2\}$ and $\mathcal{S} = \{s_1, s_2, s_3\}$, where state s_1 is the initial state, and states s_2 and s_3 are absorbing regardless of actions taken. The reward function satisfies that $r_h(s_2, a) = 1$ and $r_h(s_1, a) = r_h(s_3, a) = 0$ for all $h \in [H]$ and $a \in \mathcal{A}$. In Figure 2, step $H + 1$ is a virtual step that represents termination of an episode and generates no reward. At the initial state s_1, we may choose to take action a_1 or a_2. If a_1 is taken at state s_1, then we transition to s_2 with probability p_1 and to s_3 with probability $1 - p_1$. If a_2 is taken at state s_1, then we transition to s_2 with probability p_2 and to s_3 with probability $1 - p_2$. We interact with such an MDP for K episodes.

We note that the above K-episode MDP is equivalent to a K-round two-armed bandit with per-round reward ranging in $[0, H - 1]$, where the first transition in each episode of the MDP can be viewed as a pull of an arm in each round of the bandit. Therefore, the regret lower bound for the MDP can be proved using lower bound techniques for bandits. Our proof follows the same reasoning of [41, Theorem 15.2]. We start by discussing the setup for the proof under the cases $\beta > 0$ and $\beta < 0$.

For each $\rho \in [0, 1]$, let $\mathrm{Ber}(\rho)$ denote the Bernoulli distribution with parameter ρ. Let us fix a policy π. We first consider the case $\beta > 0$. We construct a pair of two-armed bandits, which we call ν_p and $\nu_{p'}$. The first bandit ν_p has $X_1 = (H - 1)\cdot\mathrm{Ber}(p_1)$ as the first arm and $X_2 = (H - 1)\cdot\mathrm{Ber}(p_2)$ as the second arm. The second bandit $\nu_{p'}$ has $X_1' = (H - 1)\cdot\mathrm{Ber}(p_1')$ as the first arm and $X_2' = (H - 1)\cdot\mathrm{Ber}(p_2')$ as the second arm. We let $p_2 < p_1 = p_1' < p_2'$ and $p_2 = e^{-\beta(H-1)}$. Let $\Delta := p_1 - p_2$ and we will choose $\Delta \le \frac{1}{4}e^{-\beta(H-1)}$ later in the proof. Note that when $|\beta(H - 1)|$ is large enough, we have $\Delta \le \frac{1}{100}$. Let $p_2' = p_1 + \Delta$, so that $p_2' = p_2 + 2\Delta$ and $p_2' \le \frac{1}{4}$.

We then consider $\beta < 0$. Let $p_2 = e^{\beta(H-1)}$, and set $p_1 = p_1' = p_2 - \Delta$ and $p_2' = p_2 - 2\Delta$ for some $\Delta \in (0, \frac{1}{4}e^{\beta(H-1)}]$ to be specified later. Similar to the case $\beta > 0$, we construct a pair of two-armed bandits ν_p and $\nu_{p'}$. The first bandit ν_p has $X_1 = (H-1)\cdot\mathrm{Ber}(1-p_1)$ as the first arm and $X_2 = (H-1)\cdot\mathrm{Ber}(1-p_2)$ as the second arm. The second bandit $\nu_{p'}$ has $X_1' = (H-1)\cdot\mathrm{Ber}(1-p_1')$ as the first arm and $X_2' = (H - 1)\cdot\mathrm{Ber}(1 - p_2')$ as the second arm. When $|\beta(H - 1)|$ is sufficiently

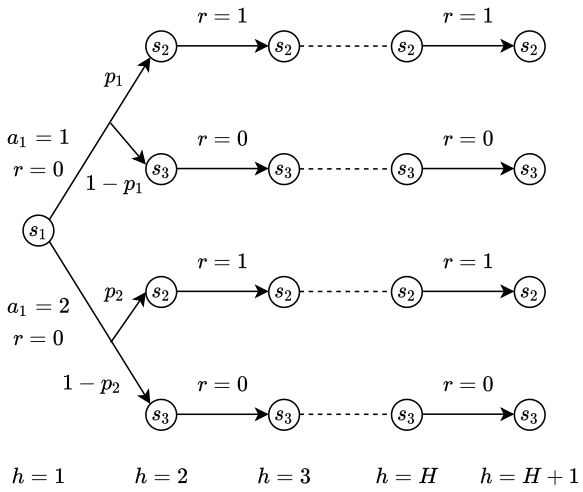

Figure 2: Illustration of the MDP for the lower bound proof for $\beta > 0$.

large, we have $1 - p_2' \geq 1 - p_1' = 1 - p_1 \geq 1 - p_2 = 1 - e^{\beta(H-1)} \geq \frac{1}{2}$ and $\Delta \leq \frac{1}{4} e^{\beta(H-1)}$ implies $\Delta \leq \frac{1}{100}$.

In the remaining of the section, we provide a unified proof for both cases of $\beta > 0$ and $\beta < 0$. We denote by $\mathbb{P}_{\pi,\nu_p}$ and $\mathbb{P}_{\pi,\nu_{p'}}$ the probability measures induced jointly by π and the two bandits, respectively. We will use the shorthands $\mathbb{P}_p := \mathbb{P}_{\pi,\nu_p}$ and $\mathbb{P}_{p'} := \mathbb{P}_{\pi,\nu_{p'}}$ for notational simplicity. Note that for both $\beta > 0$ and $\beta < 0$, the first arm is optimal for bandit ν_p, while the second is optimal for bandit $\nu_{p'}$. Let $T_a(K)$ be the number of times we have pulled the a-th arm of a bandit after we execute policy π for K rounds. It is clear that $\mathbb{E}_p[T_2(K)] \leq K$. Let $R_{\pi,\nu}(K)$ denote the regret of policy π after it is executed for K rounds in bandit ν.

By Lemmas 13 and 14, we have

$$\begin{aligned} R_{\pi,\nu_p}(K) &\gtrsim \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \Delta \cdot \mathbb{E}_p[T_2(K)] \\ &\geq \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \Delta \cdot \left[\frac{K}{2} \cdot \mathbb{P}_p(T_1(K) \leq K/2)\right], \end{aligned}$$

and

$$\begin{aligned} R_{\pi,\nu_{p'}}(K) &\gtrsim \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \Delta \cdot \mathbb{E}_{p'}[T_1(K)] \\ &\geq \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \Delta \cdot \left[\frac{K}{2} \cdot \mathbb{P}_{p'}(T_1(K) > K/2)\right]. \end{aligned}$$

We combine the above two displays and get

$$\begin{aligned} &\frac{1}{2}\left[R_{\pi,\nu_p}(K) + R_{\pi,\nu_{p'}}(K)\right] \\ &\gtrsim \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot K\Delta \left[\mathbb{P}_p(T_1(K) \leq K/2) + \mathbb{P}_{p'}(T_1(K) > K/2)\right] \\ &\geq \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot K\Delta \cdot \exp\left[-D_{\mathrm{KL}}(\mathbb{P}_p \| \mathbb{P}_{p'})\right] \\ &\geq \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot K\Delta \cdot \exp\left[-K \cdot \frac{8\Delta^2}{p_2(1-p_2)}\right] \end{aligned} \tag{34}$$

where the second step holds by the Bretagnolle–Huber inequality [41, Theorem 14.2], and the last step follows from the fact that $\mathbb{E}_p\left[T_2(K)\right] \leq K$ and Lemma 16. Now we set

$$\Delta := \sqrt{\frac{p_2(1-p_2)}{K}}.$$

Note that this choice of Δ ensures $\Delta \le \frac{1}{4}e^{-|\beta|(H-1)}$ as long as K is sufficiently large. Hence, continuing from (34) we have

$$\begin{aligned}
\frac{1}{2}\left[R_{\pi,\nu_p}(K) + R_{\pi,\nu_{p'}}(K)\right] &\gtrsim \frac{e^{|\beta|(H-1)}-1}{|\beta|}\cdot K\Delta \\
&\gtrsim \frac{e^{|\beta|(H-1)}-1}{|\beta|}\cdot \sqrt{p_2(1-p_2)K} \\
&\ge \frac{e^{|\beta|(H-1)}-1}{|\beta|}\cdot \sqrt{\frac{1}{2}e^{-\beta(H-1)}K} \\
&\gtrsim \frac{e^{|\beta|(H-1)/2}-1}{|\beta|}\sqrt{K},
\end{aligned}$$

where the third step holds since $p_2 = e^{-|\beta|(H-1)}$ and $1-p_2 \ge \frac{1}{2}$. The proof is completed by upper bounding the LHS of the above display by $\max\{R_{\pi,\nu_p}(K), R_{\pi,\nu_{p'}}(K)\}$, and recalling that $\lambda(u) = (e^{3u}-1)/u$ for $u>0$ and $T = KH$.

F.1 Auxiliary Lemmas

Lemma 13. *Let π be any policy and ν be any two-armed bandit with distinct arms. Let X_a denote the a-th arm of ν. Define $a^* \coloneqq \mathit{argmax}_{a\in\{a_1,a_2\}} \frac{1}{\beta}\log \mathbb{E}_\nu e^{\beta X_a}$ and $b \in \{a_1,a_2\}\backslash\{a^*\}$. Also define*

$$\delta_{b,\nu} \coloneqq \begin{cases} (\mathbb{E}_\nu e^{\beta X_{a^*}} - \mathbb{E}_\nu e^{\beta X_b})/\mathbb{E}_\nu e^{\beta X_{a^*}}, & \text{if } \beta > 0, \\ (\mathbb{E}_\nu e^{\beta X_b} - \mathbb{E}_\nu e^{\beta X_{a^*}})/\mathbb{E}_\nu e^{\beta X_{a^*}}, & \text{if } \beta < 0. \end{cases}$$

We have

$$R_{\pi,\nu}(K) \ge \frac{1}{2\,|\beta|}\delta_{b,\nu}\cdot \mathbb{E}_{\pi,\nu}\left[T_b(K)\right].$$

Proof. Let Y_k be the reward received at round k by executing π, and $\mathcal{A} = \{a^*, b\}$. We slightly abuse the notation by writing $\pi^k = a$ to mean that arm a is pulled in round k by executing π. Recall the definitions of the value functions V_1^* and V_1^π from (4) and (3), respectively. Since there is no state in bandit, we omit the arguments of the value functions. We observe that

$$V_1^* = \frac{1}{\beta}\log \mathbb{E}e^{\beta X_{a^*}}$$

and

$$V_1^{\pi^k} = \frac{1}{\beta}\log \mathbb{E}e^{\beta Y_k} = \frac{1}{\beta}\log\left\{\sum_{a\in\mathcal{A}} \mathbb{P}(\pi^k = a)\cdot \mathbb{E}e^{\beta X_a}\right\}.$$

In the RHS of the two displays above, the probability $\mathbb{P}(\cdot)$ is with respect to π and ν, and the expectation $\mathbb{E}[\cdot]$ is with respect to ν. Note that by the definitions of a^* and b, we have $\delta_{b,\nu} \in [0,1]$ for any $\beta \ne 0$.

For $\beta > 0$, we have

$$\begin{aligned}
V_1^* - V_1^{\pi^k} &= \frac{1}{\beta}\log\left\{\frac{\sum_{a\in\mathcal{A}} \mathbb{P}(\pi^k = a)\cdot \mathbb{E}e^{\beta X_{a^*}}}{\sum_{a\in\mathcal{A}} \mathbb{P}(\pi^k = a)\cdot \mathbb{E}e^{\beta X_a}}\right\} \\
&= \frac{1}{\beta}\log\left\{1 + \frac{\mathbb{P}(\pi^k = b)\cdot(\mathbb{E}e^{\beta X_{a^*}} - \mathbb{E}e^{\beta X_b})}{\sum_{a\in\mathcal{A}} \mathbb{P}(\pi^k = a)\cdot \mathbb{E}e^{\beta X_a}}\right\} \\
&\ge \frac{1}{\beta}\log\left\{1 + \frac{\mathbb{P}(\pi^k = b)\cdot(\mathbb{E}e^{\beta X_{a^*}} - \mathbb{E}e^{\beta X_b})}{\mathbb{E}e^{\beta X_{a^*}}}\right\} \\
&= \frac{1}{\beta}\log\left\{1 + \mathbb{E}\left[\mathbb{I}\{\pi^k = b\}\right]\cdot \delta_{b,\nu}\right\} \\
&\ge \frac{1}{2\beta}\cdot \mathbb{E}\left[\mathbb{I}\{\pi^k = b\}\right]\cdot \delta_{b,\nu},
\end{aligned}$$

where the last step holds since $\delta_{b,\nu} \in [0,1]$ and $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$. Summing both sides of the above display over $k \in [K]$ and noticing $T_b(K) = \sum_{k\in[K]} \mathbb{I}\{\pi^k = b\}$ yield the result.

For $\beta < 0$, we have

$$\begin{aligned}
V_1^* - V_1^{\pi^k} &= \frac{1}{|\beta|} \log \left\{ \sum_{a\in\mathcal{A}} \mathbb{P}(\pi^k = a) \cdot \mathbb{E}e^{\beta X_a} \right\} - \frac{1}{|\beta|} \log \mathbb{E}e^{\beta X_{a^*}} \\
&= \frac{1}{|\beta|} \log \left\{ \frac{\sum_{a\in\mathcal{A}} \mathbb{P}(\pi^k = a) \cdot \mathbb{E}e^{\beta X_a}}{\mathbb{E}e^{\beta X_{a^*}}} \right\} \\
&= \frac{1}{|\beta|} \log \left\{ 1 + \frac{\mathbb{P}(\pi^k = b) \cdot (\mathbb{E}e^{\beta X_b} - \mathbb{E}e^{\beta X_{a^*}})}{\mathbb{E}e^{\beta X_{a^*}}} \right\} \\
&= \frac{1}{|\beta|} \log \left\{ 1 + \mathbb{E}\left[\mathbb{I}\{\pi^k = b\}\right] \cdot \delta_{b,\nu} \right\} \\
&\geq \frac{1}{2\,|\beta|} \cdot \mathbb{E}\left[\mathbb{I}\{\pi^k = b\}\right] \cdot \delta_{b,\nu},
\end{aligned}$$

where the last step holds since $\delta_{b,\nu} \in [0,1]$ and $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$. Summing both sides of the above display over $k \in [K]$ and noticing $T_b(K) = \sum_{k\in[K]} \mathbb{I}\{\pi^k = b\}$ yield the result. □

Lemma 14. *Consider the setting of Lemma 13, and recall the bandits ν_p and $\nu_{p'}$ and the quantity Δ defined in Section F. For $\nu \in \{\nu_p, \nu_{p'}\}$, we have*

$$\delta_{b,\nu} \gtrsim \Delta(e^{|\beta|(H-1)} - 1).$$

Proof. We first consider the case $\beta > 0$. For $\nu = \nu_p$, we have

$$\begin{aligned}
\delta_{b,\nu} &= \frac{p_1 e^{\beta(H-1)} + (1-p_1) - [p_2 e^{\beta(H-1)} + (1-p_2)]}{p_1 e^{\beta(H-1)} + (1-p_1)} \\
&= \frac{\Delta(e^{\beta(H-1)} - 1)}{p_1 e^{\beta(H-1)} + (1-p_1)} \\
&\geq \frac{\Delta(e^{\beta(H-1)} - 1)}{3},
\end{aligned}$$

where the second step holds since $p_1 = p_2 + \Delta$, and the last step holds since $p_1 = p_2 + \Delta \leq 2e^{-\beta(H-1)}$ given $p_2 = e^{-\beta(H-1)}$ and $\Delta \leq \frac{1}{4}e^{-\beta(H-1)}$. For $\nu = \nu_{p'}$, we have

$$\begin{aligned}
\delta_{b,\nu} &= \frac{p_2' e^{\beta(H-1)} + (1-p_2') - [p_1' e^{\beta(H-1)} + (1-p_1')]}{p_2' e^{\beta(H-1)} + (1-p_2')} \\
&= \frac{\Delta(e^{\beta(H-1)} - 1)}{p_2' e^{\beta(H-1)} + (1-p_2')} \\
&\geq \frac{\Delta(e^{\beta(H-1)} - 1)}{4},
\end{aligned}$$

where the second step holds since $p_2' = p_1' + \Delta = p_1 + \Delta$, and the last step holds since $p_2' = p_1 + \Delta = p_2 + 2\Delta \leq 3e^{-\beta(H-1)}$ given $p_2 = e^{-\beta(H-1)}$ and $\Delta \leq e^{-\beta(H-1)}$.

Now we consider $\beta < 0$. For $\nu = \nu_p$, we have

$$\begin{aligned}
\delta_{b,\nu} &= \frac{(1-p_2) e^{\beta(H-1)} + p_2 - [(1-p_1) e^{\beta(H-1)} + p_1]}{(1-p_1) e^{\beta(H-1)} + p_1} \\
&= \frac{\Delta(1 - e^{\beta(H-1)})}{(1-p_1) e^{\beta(H-1)} + p_1} \\
&\geq \frac{\Delta(1 - e^{\beta(H-1)})}{2e^{\beta(H-1)}}
\end{aligned}$$

$$= \frac{\Delta(e^{-\beta(H-1)} - 1)}{2},$$

where the second step holds since $p_1 = p_2 - \Delta$, and the third step holds since $1 - p_1 \leq 1$ and $p_1 = p_2 - \Delta = e^{\beta(H-1)} - \Delta \leq e^{\beta(H-1)}$. For $\nu = \nu_{p'}$, we have

$$\begin{aligned}
\delta_{b,\nu} &= \frac{(1-p_1')e^{\beta(H-1)} + p_1' - [(1-p_2')e^{\beta(H-1)} + p_2']}{(1-p_2')e^{\beta(H-1)} + p_2'} \\
&= \frac{\Delta(1 - e^{\beta(H-1)})}{(1-p_2')e^{\beta(H-1)} + p_2'} \\
&\geq \frac{\Delta(1 - e^{\beta(H-1)})}{2e^{\beta(H-1)}} \\
&= \frac{\Delta(e^{-\beta(H-1)} - 1)}{2},
\end{aligned}$$

where the second step holds since $p_2' = p_1' - \Delta$, and the third step holds since $1 - p_2' \leq 1$ and $p_2' = p_2 - 2\Delta = e^{\beta(H-1)} - 2\Delta \leq e^{\beta(H-1)}$. We note $-\beta(H-1) = |\beta|\,(H-1)$ since $\beta < 0$ and the proof is completed. □

Lemma 15. *Under the setting of Section F, we have*

$$D_{KL}(\mathbb{P}_p \| \mathbb{P}_{p'}) \leq K \cdot \frac{8\Delta^2}{p_2(1-p_2)}.$$

Proof. For $\beta > 0$, we have

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_p \| \mathbb{P}_{p'}) &= \mathbb{E}_p\left[T_2(K)\right] \cdot D_{\mathrm{KL}}(\mathrm{Ber}(p_2) \| \mathrm{Ber}(p_2')) \\
&\leq K \cdot \frac{(p_2' - p_2)^2}{p_2'(1-p_2')} \\
&= K \cdot \frac{4\Delta^2}{p_2'(1-p_2')} \\
&\leq K \cdot \frac{4\Delta^2}{p_2(1-p_2)},
\end{aligned}$$

where the first step follows from [41, Lemma 15.1], the second step follows from the fact that $\mathbb{E}_p\left[T_2(K)\right] \leq K$ and Lemma 16, the third step follows from the identity $p_2' = p_2 + 2\Delta$, and the last step holds since $p_2 \leq p_2' \leq \frac{1}{2}$ and the function $x \mapsto x(1-x)$ is increasing on $[0, \frac{1}{2}]$.

For $\beta < 0$, we have

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_p \| \mathbb{P}_{p'}) &= \mathbb{E}_p\left[T_2(K)\right] \cdot D_{\mathrm{KL}}(\mathrm{Ber}(1-p_2) \| \mathrm{Ber}(1-p_2')) \\
&\leq K \cdot \frac{(p_2' - p_2)^2}{p_2'(1-p_2')} \\
&= K \cdot \frac{4\Delta^2}{p_2'(1-p_2')} \\
&\leq K \cdot \frac{8\Delta^2}{p_2(1-p_2)},
\end{aligned}$$

where the first step follows from [41, Lemma 15.1], the second step follows from the fact that $\mathbb{E}_p\left[T_2(K)\right] \leq K$ and Lemma 16, the third step follows from the identity $p_2' = p_2 - 2\Delta$, and the last step holds since $p_2 = e^{\beta(H-1)}$ and $\Delta \leq \frac{1}{4}e^{\beta(H-1)} = \frac{1}{4}p_2$ means $\frac{1}{2}p_2 \leq p_2' \leq p_2 \leq \frac{1}{2}$ which implies $p_2(1-p_2) \leq 2p_2'(1-p_2')$. □

Lemma 16. *Let q, q' be such that $0 \leq q' < q < 1$. We have*

$$D_{KL}(Ber(q') \| Ber(q)) \leq \frac{(q-q')^2}{q(1-q)}.$$

Proof. Let $\Delta_q \coloneqq q - q'$. The KL divergence can be upper bounded as follows:

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathrm{Ber}(q')\|\mathrm{Ber}(q)\right) &= q' \log\left(\frac{q'}{q}\right) + (1-q')\log\left(\frac{1-q'}{1-q}\right) \\
&= q' \log\left(1 + \frac{q'-q}{q}\right) + (1-q')\log\left(1 + \frac{q-q'}{1-q}\right) \\
&\overset{(i)}{\leq} q' \cdot \frac{q'-q}{q} + (1-q') \cdot \frac{q-q'}{1-q} \\
&= (\Delta_q - q) \cdot \frac{\Delta_q}{q} + (1 - q + \Delta_q) \cdot \frac{\Delta_q}{1-q} \\
&= \frac{\Delta_q^2}{q} + \frac{\Delta_q^2}{1-q} \\
&= \frac{\Delta_q^2}{q(1-q)},
\end{aligned}$$

where step (i) holds since $\log(1+x) \leq x$ for all $x > -1$. The proof is completed.

□