Exponential Bellman Equation and Improved Regret Bounds for Risk-Sensitive Reinforcement Learning

Yingjie Fei[1] **Zhuoran Yang**[2] **Yudong Chen**[3] **Zhaoran Wang**[1]
[1] Northwestern University; yf275@cornell.edu, zhaoranwang@gmail.com
[2] Princeton University; zy6@princeton.edu
[3] University of Wisconsin-Madison; yudong.chen@wisc.edu

Abstract

We study risk-sensitive reinforcement learning (RL) based on the entropic risk measure. Although existing works have established non-asymptotic regret guarantees for this problem, they leave open an exponential gap between the upper and lower bounds. We identify the deficiencies in existing algorithms and their analysis that result in such a gap. To remedy these deficiencies, we investigate a simple transformation of the risk-sensitive Bellman equations, which we call the exponential Bellman equation. The exponential Bellman equation inspires us to develop a novel analysis of Bellman backup procedures in risk-sensitive RL algorithms, and further motivates the design of a novel exploration mechanism. We show that these analytic and algorithmic innovations together lead to improved regret upper bounds over existing ones.

1 Introduction

Risk-sensitive reinforcement learning (RL) is important for practical and high-stake applications, such as self-driving and robotic surgery. In contrast with standard and risk-neutral RL, it optimizes some risk measure of cumulative rewards instead of their expectation. One foundational framework for risk-sensitive RL maximizes the entropic risk measure of the reward, which takes the form of

$$V^{\pi} = \frac{1}{\beta} \log\{\mathbb{E}_{\pi}[e^{\beta R}]\},$$

with respect to the policy π, where $\beta \neq 0$ is a given risk parameter and R denotes the cumulative rewards.

Recently, the works of [20, 21] investigate the online setting of the above risk-sensitive RL problem. Under K-episode MDPs with horizon length of H, they propose two model-free algorithms, namely RSVI and RSQ, and prove that their algorithms achieve the regret upper bound (with its informal form given by)

$$\text{Regret}(K) \lesssim e^{|\beta| H^2} \cdot \frac{e^{|\beta| H} - 1}{|\beta| H} \sqrt{\text{poly}(H) \cdot K}$$

without assuming knowledge of the transition distribution or access to a simulator. They also provide a lower bound (informally presented as)

$$\text{Regret}(K) \gtrsim \frac{e^{|\beta| H'} - 1}{|\beta| H} \sqrt{\text{poly}(H) \cdot K}$$

that any algorithm has to incur, where H' is a linear function of H. Despite the non-asymptotic nature of their results, it is not hard to see that a wide gap exists between the two bounds. Specifically,

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

the upper bound has an additional $e^{|\beta|H^2}$ factor compared to the lower bound, and even worse, this factor is dominating in the upper bound since the quadratic exponent in $e^{|\beta|H^2}$ makes it exponentially larger than $\frac{e^{|\beta|H}-1}{|\beta|H}$ even for moderate values of $|\beta|$ and H. It is unclear whether the factor of $e^{|\beta|H^2}$ is intrinsic in the upper bound.

In this paper, we show that the additional factor in the upper bound is not intrinsic for the upper bound and can be eliminated by a refined algorithmic design and analysis. We identify two deficiencies in the existing algorithms and their analysis: (1) the main element of the analysis follows existing analysis of risk-neutral RL algorithms, which fails to exploit the special structure of the Bellman equations of risk-sensitive RL; (2) the existing algorithms use an excessively large bonus that results in the exponential blow-up in the regret upper bound.

To address the above shortcomings, we consider a simple transformation of the Bellman equations analyzed so far in the literature, which we call the *exponential Bellman equation*. A distinctive feature of the exponential Bellman equation is that they associate the instantaneous reward and value function of the next step in a multiplicative way, rather than in an additive way as in the standard Bellman equations. From the exponential Bellman equation, we develop a novel analysis of the Bellman backup procedure for risk-sensitive RL algorithms that are based on the principle of optimism. The analysis further motivates a novel exploration mechanism called *doubly decaying* bonus, which helps the algorithms adapt to their estimation error over each horizon step while at the same time exploring efficiently. These discoveries enable us to propose two model-free algorithms for RL with the entropic risk measure based on the novel bonus. By combining the new analysis and bonus design, we prove that the preceding algorithms attain nearly optimal regret bounds under episodic and finite-horizon MDPs. Compared to prior results, our regret bounds feature an exponential improvement with respect to the horizon length and risk parameter, removing the factor of $e^{|\beta|H^2}$ from existing upper bounds. This significantly narrows the gap between upper bounds and the existing lower bound of regret.

In summary, we make the following theoretical contributions in this paper.

1. We investigate the gap between existing upper and lower regret bounds in the context of risk-sensitive RL, and identify deficiencies of the existing algorithms and analysis;
2. We consider the exponential Bellman equation, which inspires us to propose a novel analysis of the Bellman backup procedure for RL algorithms based on the entropic risk measure. It further motivates a novel bonus design called doubly decaying bonus. We then design two model-free risk-sensitive RL algorithms equipped with the novel bonus.
3. The novel analytic framework and bonus design together enable us to prove that the preceding algorithms achieve nearly optimal regret bounds, which improve upon existing ones by an exponential factor in terms of the horizon length and risk sensitivity.

2 Related works

The problem of RL with respect to the entropic risk measure is first proposed by the classical work of [24], and has since inspired a large body of studies [2, 4–8, 13, 16–18, 22, 23, 25, 26, 31, 33, 37, 38, 40, 41, 43]. However, the algorithms from this line of works require knowledge of the transition kernel or assume access to a simulator of the underlying environment. Theoretical properties of these algorithms are investigated based on these assumptions, but the results are mostly of asymptotic nature, which do not shed light on their dependency on key parameters of the environment and agent.

The work of [20] represents the first effort to investigate the setting where transitions are unknown and simulators of the environment are unavailable. It establishes the first non-asymptotic regret or sample complexity guarantees under the tabular setting. Building upon [20], the authors of [21] extend the results to the function approximation setting, by considering linear and general function approximations of the underlying MDPs. Nevertheless, as discussed in Section 1, both works leave open an exponential gap between the regret upper and lower bounds, which the present work aims to address via novel algorithms and analysis motivated by the exponential Bellman equation.

We remark that although the exponential Bellman equation has been previously investigated in the literature of risk-sensitive RL [2, 5], this is the first time that it is explored for deriving regret and sample complexity guarantees of risk-sensitive RL algorithms. In Appendix A, we also make connections between risk-sensitive RL and distributional RL through the exponential Bellman equation.

Notations. For a positive integer n, we let $[n] \coloneqq \{1, 2, \dots, n\}$. 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, and 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. For functions $f, g : \mathcal{U} \to \mathbb{R}$, where $\mathcal{U}$ denotes their domain, we write $f \geq g$ if $f(u) \geq g(u)$ for any $u \in \mathcal{U}$. We denote by $\mathbb{I}\{\cdot\}$ the indicator function.

3 Problem background

3.1 Episodic and finite-horizon MDP

The setting of episodic Markov decision processes can be denoted by MDP$(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $H \in \mathbb{Z}_{>0}$ 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 transition kernels and reward functions, respectively. We let $S \coloneqq |\mathcal{S}|$ and $A \coloneqq |\mathcal{A}|$, and we assume $S, A < \infty$. We let $P_h(\cdot \,|\, s, a)$ denote the probability distribution over successor states of step $h+1$ if action a is executed in state s at step h. We assume that the reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is deterministic. We also assume that both $\mathcal{P}$ and $\mathcal{R}$ are unknown to learning agents.

Under the setting of an episodic MDP, the agent aims to learn the optimal policy by interacting with the environment throughout $K > 0$ episodes, described as follows. At the beginning of episode k, an initial state s_1^k is selected by the environment and we assume s_1^k stays the same for all $k \in [K]$. In each step $h \in [H]$ of episode k, the agent observes state $s_h^k \in \mathcal{S}$, executes an action $a_h^k \in \mathcal{A}$, and receives a reward equal to $r_h(s_h^k, a_h^k)$ from the environment. The MDP then transitions into state s_{h+1}^k randomly drawn from the transition kernel $P_h(\cdot \,|\, s_h^k, a_h^k)$. The episode terminates at step $H+1$, in which the agent does not take actions or receive rewards. We define a policy $\pi = \{\pi_h\}_{h \in [H]}$ as a collection 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 the episode.

3.2 Risk-sensitive RL

For each $h \in [H]$, we define the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π as the cumulative utility of the agent at state s of step h under the entropic risk measure, assuming that the agent commits to policy π in later steps. Specifically, we define

$$\forall (h, s) \in [H] \times \mathcal{S}, \quad V_h^\pi(s) \coloneqq \frac{1}{\beta} \log \left\{ \mathbb{E} \left[e^{\beta \sum_{i=h}^H r_i(s_i, \pi_i(s_i))} \,\middle|\, s_h = s \right] \right\}, \tag{1}$$

where $\beta \neq 0$ is a given risk parameter. The agent aims to maximize his cumulative utility in step 1, that is, to find a policy π such that $V_1^\pi(s)$ is maximized for all state $s \in \mathcal{S}$. Under this setting, if $\beta > 0$, the agent is risk-seeking and if $\beta < 0$, the agent is risk-averse. Furthermore, as $\beta \to 0$ the agent tends to be risk-neutral and $V_h^\pi(s)$ tends to the classical value function.

We may also define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is the cumulative utility of the agent who follows policy π, conditional on a particular state-action pair; formally, this is given by

$$\forall (h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}, \quad Q_h^\pi(s, a) \coloneqq \frac{1}{\beta} \log \left\{ \mathbb{E} \left[e^{\beta \sum_{i=h}^H r_i(s_i, a_i)} \,\middle|\, s_h = s, a_h = a \right] \right\}, \tag{2}$$

Under some mild regularity conditions [2], there always exists an optimal policy, which we denote as π^*, that yields the optimal value $V_h^*(s) \coloneqq \sup_\pi V_h^\pi(s)$ for all $(h, s) \in [H] \times \mathcal{S}$.

Bellman equations. For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, the Bellman equation associated with a policy π is given by

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

for $h \in [H]$. In Equation (3), it can be seen that the action value Q_h^π of step h is a non-linear function of the value function V_{h+1}^π of the later step. This is in contrast with the linear Bellman equations in the risk-neutral setting ($\beta \to 0$), where $Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_{s'}[V_{h+1}^\pi(s')]$. Based on Equation (3), for $h \in [H]$, the Bellman optimality equation is given by

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

Exponential Bellman equation. We introduce the *exponential Bellman equation*, which is an exponential transformation of Equations (3) and (4) (by taking exponential on both sides): for any policy π and tuple (h,s,a), we have

$$
e^{\beta \cdot Q_h^\pi(s,a)} = \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)} \left[e^{\beta (r_h(s,a) + V_{h+1}^\pi(s'))} \right]. \tag{5}
$$

When $\pi = \pi^*$, we obtain the corresponding optimality equation

$$
e^{\beta \cdot Q_h^*(s,a)} = \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)} \left[e^{\beta (r_h(s,a) + V_{h+1}^*(s'))} \right]. \tag{6}
$$

Note that Equation (5) associates the current and future cumulative utilities (Q_h^π and V_{h+1}^π) in a multiplicative way. An implication of Equation (5) is that one may estimate $e^{\beta \cdot Q_h^\pi(s,a)}$ by a quantity of the form

$$
w_h(s,a) = \text{SampAvg}(\{ e^{\beta(r_h(s_h,a_h) + V_{h+1}(s_{h+1}))} : (s_h,a_h) = (s,a) \}) \tag{7}
$$

given some estimate of the value function V_{h+1}. Here, we denote by SampAvg($\mathcal{X}$) the sample average computed over elements in the set $\mathcal{X}$ throughout past episodes, and it can be seen as an empirical MGF of cumulative rewards from step $h+1$. Equation (5) also suggests the following policy improvement procedure for a risk-sensitive policy π:

$$
\pi_h(s) \leftarrow \operatorname*{argmax}_{a' \in \mathcal{A}} Q_h(s,a') = \begin{cases} \operatorname{argmax}_{a' \in \mathcal{A}} e^{\beta \cdot Q_h(s,a')}, & \text{if } \beta > 0 \\ \operatorname{argmin}_{a' \in \mathcal{A}} e^{\beta \cdot Q_h(s,a')}, & \text{if } \beta < 0, \end{cases} \tag{8}
$$

where Q_h denotes some estimated action-value function, possibly obtained from the quantity w_h.

In the next section, we will discuss how the exponential Bellman equation (5) inspires the development of a novel analytic framework for risk-sensitive RL. Before proceeding, we introduce a performance metric for the agent. For each episode k, recall that s_1^k is the initial state chosen by the environment and let π^k be the policy of the agent at the beginning of episode k. Then the difference $V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)$ is called the *regret* of the agent in episode k. Therefore, after K episodes, the total regret for the agent is given by

$$
\text{Regret}(K) := \sum_{k \in [K]} [V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)], \tag{9}
$$

which serves as the key performance metric studied in this paper.

4 Analysis of risk-sensitive RL

4.1 Mechanism of existing analysis

In this section, we provide an informal overview of the mechanism underlying the existing analysis of risk-sensitive RL. Let us focus on the case $\beta > 0$ for simplicity of exposition; similar reasoning holds for $\beta < 0$. A key step in the existing regret analysis of RL algorithms is to establish a recursion on the difference $V_h^k - V_h^{\pi^k}$ over $h \in [H]$, where V_h^k is the iterate of an algorithm in step h of episode k and $V_h^{\pi^k}$ is the value function of the policy used in episode k. Such approach can be commonly found in the literature of algorithms that use the upper confidence bound [27, 28], in which the recursion takes the form of

$$
V_h^k - V_h^{\pi^k} \le V_{h+1}^k - V_{h+1}^{\pi^k} + \psi_h^k, \tag{10}
$$

for $\beta \to 0$ and some quantity ψ_h^k. The work of [20], which studies the risk-sensitive setting under the entropic risk measure, also follows this approach and derives regret bounds by establishing the recursion of the form

$$V_h^k - V_h^{\pi^k} \le e^{\beta H}\left(V_{h+1}^k - V_{h+1}^{\pi^k}\right) + \frac{1}{\beta}\tilde{b}_h^k + e^{\beta H}\tilde{m}_h^k, \tag{11}$$

where $\tilde{b}_h^k$ denotes the bonus which enforces the upper confidence bound and leads to the inequality $V_h^k \ge V_h^\pi$ for any policy π, and $\tilde{m}_h^k$ is part of a martingale difference sequence. The derivation of Equation (11) is based on the Bellman equation (3), which shows that the action value $Q_h^{\pi^k}$ is the sum of the reward r_h and the entropic risk measure of $V_{h+1}^{\pi^k}$. Following [20], we may then unroll the recursion (11) from $h = H$ to $h = 1$ to get

$$V_1^k - V_1^{\pi^k} \le \frac{1}{\beta} e^{\beta H^2} \sum_h \tilde{b}_h^k + e^{\beta H^2} \sum_h \tilde{m}_h^k, \tag{12}$$

given that $V_{H+1}^k = V_{H+1}^{\pi^k} = 0$. Using the inequality $\text{Regret}(K) \le \sum_k (V_1^k - V_1^{\pi^k})$, $\sum_{k,h} \tilde{b}_h^k \lesssim (e^{\beta H} - 1)\sqrt{K}$ and $\sum_{k,h} \tilde{m}_h^k \lesssim \sqrt{K}$, we obtain the regret bound in [20] as $\text{Regret}(K) \lesssim e^{\beta H^2} \frac{e^{\beta H}-1}{\beta H}\sqrt{K}$. Therefore, it can be seen that the dominating factor $e^{\beta H^2}$ in their regret bound originates in Equation (12), which can be further traced back to the exponential factor $e^{\beta H}$ in the error dynamics (11).

4.2 Refined approach via exponential Bellman equation

While the existing analysis in (11) is motivated by the Bellman equation of the form given in (3), we propose to work on the exponential Bellman equation (5). Equation (5) operates on the quantities $e^{\beta \cdot Q_h^\pi}$ and $e^{\beta \cdot V_{h+1}^\pi}$, which can be thought of as the MGFs of the current and future values, while the reward function r_h is involved as a multiplicative term. This motivates us to derive a new recursion:

$$e^{\beta \cdot V_h^k} - e^{\beta \cdot V_h^{\pi^k}} \le e^{\beta \cdot r_h^k}\left(e^{\beta \cdot V_{h+1}^k} - e^{\beta \cdot V_{h+1}^{\pi^k}}\right) + b_h^k + m_h^k, \tag{13}$$

where b_h^k, m_h^k denote some bonus and martingale terms, respectively, and r_h^k stands for the reward in step h of episode k. Unrolling Equation (13) yields

$$e^{\beta \cdot V_1^k} - e^{\beta \cdot V_1^{\pi^k}} \le \sum_h e^{\beta \cdot D_h^k}(b_h^k + m_h^k), \tag{14}$$

where $D_h^k = \sum_{i \in [h-1]} r_i^k$. In words, the error of $e^{\beta \cdot V_1^k} - e^{\beta \cdot V_1^{\pi^k}}$ is bounded by the weighted sum of bonus and martingale difference terms, where the weights are given by $e^{\beta \cdot D_h^k}$, the exponential rewards up to step $h - 1$. We may then apply a localized linearization of the logarithmic function, which gives $\text{Regret}(K) \le \frac{1}{\beta}\sum_k (e^{\beta \cdot V_1^k} - e^{\beta \cdot V_1^{\pi^k}})$, and arrives at a regret upper bound (the formal regret bounds will be established in Theorems 1 and 2 below). Different from Equation (11) where rewards are only implicitly encoded in V_h^k, in Equation (13) rewards are explicitly involved in the error dynamics via an exponential term.

To see why Equation (13) is intuitively correct, we may divide both sides of the equation by β and take $\beta \to 0$. By doing so, we should expect to obtain quantities from the error dynamics (10) of risk-neutral RL. Since the function $f_\beta(x) = (e^{\beta x} - 1)/\beta$ satisfies that $f_\beta(x) \to x$ as $\beta \to 0$ for any fixed x, we have

$$\lim_{\beta \to 0} \frac{1}{\beta}(e^{\beta \cdot V_h^k} - e^{\beta \cdot V_h^{\pi^k}}) = V_h^k - V_h^{\pi^k},$$

$$\lim_{\beta \to 0} \frac{1}{\beta}(e^{\beta \cdot r_h^k}(e^{\beta \cdot V_{h+1}^k} - e^{\beta \cdot V_{h+1}^{\pi^k}})) = r_h^k + V_{h+1}^k - (r_h^k + V_{h+1}^{\pi^k}) = V_{h+1}^k - V_{h+1}^{\pi^k},$$

recovering terms in (10). Therefore, the recursion (13) can be seen as generalizing those in the analysis of risk-neutral RL.

By comparing Equations (13) and (11), we see that while both error dynamics are derived from the same underlying Bellman equation, they inspire drastically different forms of recursion. Note that

Algorithm 1 RSVI2

1: $Q_h(\cdot,\cdot), V_h(\cdot) \leftarrow H-h+1$, $N_h(\cdot,\cdot) \leftarrow 0$ and $w_h(\cdot,\cdot) \leftarrow 0$ for all $h \in [H+1]$
2: **for** episode $k = 1, \ldots, K$ **do**
3:     **for** step $h = H, \ldots, 1$ **do**
4:         **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $N_h(s,a) \geq 1$ **do**
5:             $w_h(s,a) \leftarrow \frac{1}{N_h(s,a)} \sum_{\tau \in [k-1]} \mathbb{I}\{(s_h^\tau, a_h^\tau) = (s,a)\} \cdot e^{\beta[r_h(s,a) + V_{h+1}(s_{h+1}^\tau)]}$
6:             $b_h(s,a) \leftarrow c|e^{\beta(H-h+1)} - 1| \sqrt{\frac{S\log(HSAK/\delta)}{N_h(s,a)}}$ where $c > 0$ is a universal constant
7:             $G_h(s,a) \leftarrow \begin{cases} \min\{w_h(s,a) + b_h(s,a),\ e^{\beta(H-h+1)}\}, & \text{if } \beta > 0 \\ \max\{w_h(s,a) - b_h(s,a),\ e^{\beta(H-h+1)}\}, & \text{if } \beta < 0 \end{cases}$
8:             $V_h(s) \leftarrow \max_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h(s,a')\}$
9:         **end for**
10:     **end for**
11:     $\forall h \in [H]$, take $a_h \leftarrow \text{argmax}_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h(s_h, a')\}$; observe $r_h(s_h, a_h)$, s_{h+1}
12:     Add 1 to $N_h(s_h, a_h)$
13: **end for**

the multiplicative factor $e^{\beta \cdot r_h^k}$ in Equation (13) is milder than the factor $e^{\beta H}$ in Equation (11), since $r_h^k \in [0,1]$. This is the source of an improvement of our refined analysis over existing works. On the other hand, the success of applying the error dynamics (13) in our analysis crucially depends on the choice of bonus terms $\{b_h^k\}$, as an improper choice would blow up the error $e^{\beta \cdot V_1^k} - e^{\beta \cdot V_1^{\pi^k}}$. This observation motivates our novel bonus design, as we explain next in Section 5.

5 Algorithms

5.1 Overview of algorithms

In this section, we propose two model-free algorithms for RL with the entropic risk measure. We first present RSVI2, which is based on value iteration, in Algorithm 1. The algorithm has two main stages: it first estimates the value function using data accumulated up to episode $k-1$ (Line 3–10) and then executes the estimated policy to collect new trajectory (Line 11). In value function estimation, it computes the weights w_h, or the empirical MGF of some estimated cumulative rewards evaluated at β, which can be seen as a simple moving average over $\tau \in [k-1]$. Therefore, Line 5 functions as a concrete implementation of Equation (7) where the sample average is instantiated as a simple moving average. Then in Line 7, it computes an augmented estimate G_h by combining w_h with a bonus term b_h (defined in Line 6). This is followed by thresholding to put G_h in the proper range. Note that G_h is an optimistic estimator of the quantity $e^{\beta \cdot Q_h^\pi}$ in Equation (5): the construction of G_h is augmented by b_h so that it encourages exploration of rarely visited state-action pairs in future episodes, and thereby follows the principle of Risk-Sensitive Optimism in the Face of Uncertainty [20]. When $\beta < 0$, the bonus is subtracted from w_h, since a higher level of optimism corresponds to a *smaller* value of the estimate. In addition, Line 11 follows the reasoning of policy improvement suggested in Equation (8).

Next, we introduce RSQ2 in Algorithm 2, which is based on Q-learning. Similar to Algorithm 1, it consists of value estimation (Line 8–11) and policy execution (Line 6) steps. By combining Lines 9 and 10, we see that Algorithm 2 computes the optimistic estimate G_h as a projection of an exponential moving average of empirical MGFs:

$$G_h(s_h, a_h) \leftarrow \Pi_h\{\text{EMA}(\{e^{\beta[r_h(s_h,a_h) + V_{h+1}(s_{h+1})]}\})\}, \tag{15}$$

where Π_h denotes a projection that depends on step h. In particular, Line 9 can be interpreted as a computation of empirical MGFs evaluated at β and thus a concrete implementation of Equation (7) using an exponential moving average. This is in contrast with the simple moving average update in Algorithm 1.

Although Algorithms 1 and 2 are inspired by RSVI and RSQ of [20], respectively, we note that the main novelty of our algorithms lies in the bonus terms (b_h in Algorithm 1 and $b_{h,t}$ in Algorithm 2), which we call the *doubly decaying* bonus. We discuss this new bonus design in the following.

Algorithm 2 RSQ2

$Q_h(\cdot,\cdot), V_h(\cdot) \leftarrow H-h+1$ if $\beta>0$; $Q_h(\cdot,\cdot), V_h(\cdot) \leftarrow 0$ otherwise, for all $h \in [H+1]$
$N_h(\cdot,\cdot) \leftarrow 0$ for all $h \in [H]$; $\alpha_u \leftarrow \frac{H+1}{H+u}$ for $u \in \mathbb{Z}$
for episode $k = 1, \ldots, K$ **do**
  Receive the initial state s_1
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \mathrm{argmax}_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h(s_h, a')\}$ and observe $r_h(s_h, a_h)$ and s_{h+1}
    Add 1 to $N_h(s_h, a_h)$; $t \leftarrow N_h(s_h, a_h)$
    $b_{h,t} \leftarrow c|e^{\beta(H-h+1)} - 1|\sqrt{\frac{H \log(HSAK/\delta)}{t}}$ for some universal constant $c > 0$
    $w_h(s_h, a_h) \leftarrow (1-\alpha_t) \cdot G_h(s_h, a_h) + \alpha_t \cdot e^{\beta[r_h(s_h,a_h) + V_{h+1}(s_{h+1})]}$
    $G_h(s_h, a_h) \leftarrow \begin{cases} \min\{w_h(s_h,a_h) + \alpha_t b_{h,t},\ e^{\beta(H-h+1)}\}, & \text{if } \beta > 0 \\ \max\{w_h(s_h,a_h) - \alpha_t b_{h,t},\ e^{\beta(H-h+1)}\}, & \text{if } \beta < 0 \end{cases}$
    $V_h(s_h) \leftarrow \max_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h(s_h, a')\}$
  end for
end for

5.2 Doubly decaying bonus

Let us focus on $\beta > 0$ for this discussion. In optimism-based algorithms, the bonus term is used to enforce the upper confidence bound in order to encourage sufficient exploration in uncertain environments. It takes the form of a multiplier times a factor that is inversely proportional to visit counts $\{N_h\}$. Our bonus follows this structure and is given by

$$b_h(s,a) \propto (e^{\beta(H-h+1)} - 1)\sqrt{\frac{1}{N_h(s,a)}}, \tag{16}$$

ignoring factors that do not vary in (h, s, a). In Equation (16), the quantity $e^{\beta(H-h+1)}$ plays the role of the multiplier and $\sqrt{1/N_h(s,a)}$ is the factor that decreases in the visit count. While the component $\sqrt{1/N_h(s,a)}$ is common in bonus terms, our new bonus is designed to shrink its multiplier deterministically and exponentially across the horizon steps, as $e^{\beta(H-h+1)} - 1$ decreases from $e^{\beta H} - 1$ in step $h = 1$ to $e^{\beta} - 1$ in step $h = H$. This is in sharp contrast with the bonus terms typically found in risk-neutral RL algorithms, where the multipliers are kept constant in h (usually as a constant multiple of H). Furthermore, our bonus design is also in contrast with that in RSVI and RSQ proposed by [20], whose multiplier is $e^{\beta H} - 1$ and kept fixed along the horizon. Because b_h decays both in the visit count $N_h(s,a)$ (across episodes) and the multiplier $e^{\beta(H-h+1)} - 1$ (across the horizon), we name it as *doubly decaying* bonus. We remark that this is a novel feature of Algorithms 1 and 2, compared to RSVI and RSQ. Let us discuss how this new exploration mechanism is motivated from the error dynamics (14).

Motivation of exponential decay. From Equation (14), we see that the error of the iterate is bounded by the sum of weighted bonus terms, where the weights are of the form $e^{\beta \cdot D_h}$ and $D_h \in [0, h-1]$. Choosing $b_h \propto e^{\beta(H-h+1)} - 1$ ensures that the weighted bonus is on the order of $e^{\beta H} - 1$ at maximum. On the other hand, if we use the bonus as in [20], which is proportional to $e^{\beta H} - 1$, then we would end up with a multiplicative factor $e^{2\beta H} - 1$ in regret, which is exponentially larger than $e^{\beta H} - 1$. An alternative way to understanding the exponential decay of our bonus is as follows. At step h, the estimated value function is $V_h \in [0, H-h+1]$, which implies $e^{\beta \cdot V_h} \in [1, e^{\beta(H-h+1)}]$. The iterate G_h (of Algorithm 1 or 2) is used to estimate $e^{\beta \cdot Q_h^\pi}$, with its estimation error given by

$$|e^{\beta \cdot Q_h^\pi} - G_h| \approx |e^{\beta \cdot Q_h^\pi} - \widehat{\mathbb{P}}_h e^{\beta(r_h + V_{h+1})}| \le e^{\beta(H-h+1)} - 1,$$

where $\widehat{\mathbb{P}}_h$ denotes an empirical average operator over historical data in step h. Therefore, the estimation error of G_h shrinks exponentially across the horizon. Since bonus is used to compensate for and dominate the estimation error, the minimal order of b_h required is thus $e^{\beta(H-h+1)} - 1$, which is exactly the multiplier in Equation (16).

As a passing note, we remark that the decaying multiplier is not necessary in risk-neutral RL algorithms, since the estimation error therein satisfies $|Q_h - \widehat{\mathbb{P}}_h(r_h + V_{h+1})| \le H - h + 1$, which is upper bounded by H for all $h \in [H]$. This implies that it suffices to simply set the bonus multiplier as a constant multiple of H. In contrast, as we have explained, the estimation error of our algorithms decays exponentially in step h, and an adaptive and exponentially decaying bonus is needed.

Comparison with Bernstein-type bonus. We also compare our bonus in Equation (16) with the Bernstein-type bonus commonly used to improve sample efficiency of risk-neutral RL algorithms [1, 27]. The Bernstein-type bonus takes the form of

$$\bar{b}_h(s,a) \propto \sqrt{\frac{H + \widehat{\mathrm{Var}}(V_{h+1})}{N_h(s,a)}} + o\left(\sqrt{\frac{1}{N_h(s,a)}}\right), \tag{17}$$

where $\widehat{\mathrm{Var}}(\cdot)$ denotes an empirical variance operator over historical data and $o(\cdot)$ denotes a vanishing term as $N_h(s,a) \to \infty$. Our bonus in Equation (16) is different from the Bernstein-type bonus in Equation (17) in mechanism: our bonus features the multiplier $e^{\beta(H-h+1)} - 1$ which decays exponentially and deterministically over $h \in [H]$, whereas the Bernstein-type bonus uses $\sqrt{H + \widehat{\mathrm{Var}}(V_{h+1})}$ as the multiplier (ignoring the vanishing term). The term $\widehat{\mathrm{Var}}(V_{h+1})$ depends on the trajectory of the learning process. Therefore the multiplier is stochastic and stays on the polynomial order of H across the horizon. Moreover, it is unclear how the multiplier behaves in terms of step h.

6 Main results

In this section, we present and discuss our main theoretical results for Algorithms 1 and 2.

Theorem 1. *For any $\delta \in (0,1]$, with probability at least $1-\delta$ there exists a universal constant $c > 0$ (used in Algorithm 1), such that the regret of Algorithm 1 is bounded by*

$$\mathrm{Regret}(K) \lesssim \frac{e^{|\beta|H} - 1}{|\beta|H} \sqrt{H^4 S^2 AK \log^2(HSAK/\delta)}.$$

Theorem 2. *For any $\delta \in (0,1]$, with probability at least $1-\delta$ and when K is sufficiently large, there exists a universal constant $c > 0$ (used in Algorithm 2) such that the regret of Algorithm 2 obeys*

$$\mathrm{Regret}(K) \lesssim \frac{e^{|\beta|H} - 1}{|\beta|H} \sqrt{H^3 SAK \log(HSAK/\delta)}.$$

The proof of the two theorems are provided in Appendices B and C, respectively. Note that the above results generalize those in the literature of risk-neutral RL: when $\beta \to 0$, we recover the same regret bounds of LSVI in [28] and Q-learning in [27].

Let us discuss the connections between our results and those in [20]. The work of [20] proposes two algorithms, RSVI and RSQ, that attain the regret bound

$$\mathrm{Regret}(K) \lesssim e^{|\beta|H^2} \cdot \frac{e^{|\beta|H} - 1}{|\beta|H} \sqrt{\mathrm{poly}(H) \cdot K}, \tag{18}$$

and a lower bound incurred by any algorithm

$$\mathrm{Regret}(K) \gtrsim \frac{e^{|\beta|H'} - 1}{|\beta|H} \sqrt{\mathrm{poly}(H) \cdot K}, \tag{19}$$

where H' is a linear function in H; for simplicity of presentation, we exclude polynomial dependencies on other parameters and logarithmic factors from the two bounds. In particular, the proof of the lower bound is based on reducing an hard instance of MDP to a multi-armed bandit. It is a priori unclear whether the extra exponential factor $e^{|\beta|H^2}$ in the upper bound (18) is fundamental in the MDP setting, or is due to suboptimal analysis or algorithmic design. We would like to mention that although one trivial way of avoiding the $e^{|\beta|H^2}$ factor in the upper bound (18) is to use a sufficiently small $|\beta|$ in the algorithms of [20] (e.g., $|\beta| \le \frac{1}{H^2}$ so that $e^{|\beta|H^2} \lesssim 1$), such a small $|\beta|$ defeats the very purpose of have an appropriate degree of risk-sensitivity in the algorithms. Hence, an answer for *all* $\beta \neq 0$ would be desirable.

In view of Theorems 1 and 2, we see that our Algorithms 1 and 2 achieve regret bounds that are exponentially sharper than those of RSVI and RSQ. In particular, our results eliminate the $e^{|\beta|H^2}$ factor from Equation (18) thanks to the novel analysis and doubly decaying bonus in our algorithms, which are inspired by the exponential Bellman equation (5). As a result, our bounds significantly narrow the gap between upper bounds and the lower bound (19).

Acknowledgments and Disclosure of Funding

We thank the reviewers for their constructive feedback. Z. Yang acknowledges Simons Institute (Theory of Reinforcement Learning). Y. Chen is partially supported by NSF grant CCF-1704828 and CAREER Award CCF-2047910. Z. Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports.

References

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

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

[3] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

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

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

[6] Vivek S. Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, pages 55–60, 2010.

[7] 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.

[8] Rolando Cavazos-Cadena and Daniel Hernández-Hernández. Discounted approximations for risk-sensitive average criteria in Markov decision chains with finite state space. *Mathematics of Operations Research*, 36(1):133–146, 2011.

[9] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. In *Advances in Neural Information Processing Systems*, 2021.

[10] Lin Chen, Bruno Scherrer, and Peter L Bartlett. Infinite-horizon offline reinforcement learning with linear function approximation: Curse of dimensionality and algorithm. *arXiv preprint arXiv:2103.09847*, 2021.

[11] Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. In *International Conference on Learning Representations*, 2021.

[12] Lin Chen, Qian Yu, Hannah Lawrence, and Amin Karbasi. Minimax regret of switching-constrained online convex optimization: No phase transition. In *Advances in Neural Information Processing Systems*, 2020.

[13] Stefano P. Coraluppi and Steven I. Marcus. Risk-sensitive, minimax, and mixed risk-neutral/minimax control of Markov decision processes. In *Stochastic Analysis, Control, Optimization and Applications*, pages 21–40. Springer, 1999.

[14] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[15] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] 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.

[17] Giovanni B. Di Masi and Lukasz Stettner. Infinite horizon risk sensitive control of discrete time Markov processes with small risk. *Systems & Control Letters*, 40(1):15–20, 2000.

[18] Giovanni B. Di Masi and Łukasz Stettner. Infinite horizon risk sensitive control of discrete time Markov processes under minorization property. *SIAM Journal on Control and Optimization*, 46(1):231–252, 2007.

[19] Amir-massoud Farahmand. Value function in frequency domain and the characteristic value iteration algorithm. In *Advances in Neural Information Processing Systems*, 2019.

[20] Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *Advances in Neural Information Processing Systems*, 2020.

[21] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pages 3198–3207. PMLR, 2021.

[22] 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.

[23] 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.

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

[25] Wenjie Huang and William B Haskell. Stochastic approximation for risk-aware markov decision processes. *IEEE Transactions on Automatic Control*, 66(3):1314–1320, 2020.

[26] Anna Jaśkiewicz. Average optimality for risk-sensitive control with general state space. *The Annals of Applied Probability*, 17(2):654–675, 2007.

[27] 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.

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

[29] Joel Z Leibo, Cyprien de Masson d'Autume, Daniel Zoran, David Amos, Charles Beattie, Keith Anderson, Antonio García Castañeda, Manuel Sanchez, Simon Green, Audrunas Gruslys, et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018.

[30] Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *SIAM Journal on Optimization*, 29(3):1879–1907, 2019.

[31] 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.

[32] Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *International Conference on Machine Learning*, pages 4424–4434. PMLR, 2019.

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

[34] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *International Conference on Machine Learning*, 2010.

[35] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

[36] 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.

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

[38] Stephen D. Patek. On terminating Markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386, 2001.

[39] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

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

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

[42] Ganlin Song, Ruitu Xu, and John Lafferty. Convergence and alignment of gradient descentwith random back propagation weights. *arXiv preprint arXiv:2106.06044*, 2021.

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

[44] Ruitu Xu, Lin Chen, and Amin Karbasi. Meta learning in the continuous time limit. In *International Conference on Artificial Intelligence and Statistics*, pages 3052–3060. PMLR, 2021.

[45] Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32:6193–6202, 2019.

Appendices

A Connections to distribution RL

In this appendix, we establish connections between risk-sensitive RL and distributional RL via the lens of the exponential Bellman equation.

Distributional RL has been studied in the line of works [3, 15, 19, 32, 34, 35, 39, 45]. The framework of distributional RL is built upon the following key equation, namely the distributional Bellman equation:

$$\forall h \in [H], \quad Z_h^\pi(s,a) \stackrel{d}{=} R_h(s,a) + Z_{h+1}^\pi(X', U'), \tag{20}$$

for a fixed policy π, where $Z_{H+1}^\pi(\cdot,\cdot) := 0$, $X' \sim P_h(\cdot \mid s,a)$, $U' \sim \pi(\cdot \mid X')$ and R_h is the reward distribution in step h. Here, we use $\stackrel{d}{=}$ to denote equality in distribution. It can be seen that $Z_h^\pi(s,a)$ is the distribution of cumulative rewards under policy π at step h, when the state and action (s,a) are visited in step h. Based on Equation (20), a distributional Bellman optimality operator $\mathbb{T}_h$ is given by

$$[\mathbb{T}_h Z](s,a) \stackrel{d}{:=} R_h(s,a) + Z_{h+1}(X', \operatorname*{argmax}_{a' \in \mathcal{A}} \mathbb{E}[Z_{h+1}(X', a')]), \tag{21}$$

where again $X' \sim P_h(\cdot \mid s,a)$. Note that in Equation (21), the optimal action is greedy with respect to the *expectation* of the distribution Z_{h+1}. Most existing distributional RL algorithms work with distribution estimates such as quantiles [14, 15] or empirical distribution functions [3, 39].

Now recall the exponential Bellman equation (5), which takes the form

$$\forall h \in [H], \quad e^{\mu \cdot Q_h^\pi(s,a)} = \mathbb{E}_{X'}[e^{\mu(r_h(s,a) + V_{h+1}^\pi(X'))}], \tag{22}$$

for any fixed $\mu \in \mathbb{R}$, where $V_{H+1}^\pi(\cdot) := 0$, $X' \sim P_h(\cdot \mid s,a)$ and r_h is the deterministic reward function by our assumption. Given the definitions (1) and (2) with β replaced by μ, we note that both Q_h^π and V_{h+1}^π in the above equation depend on the value of μ (which we omit for simplicity of notations). Then by the definition of Q_h^π in Equation (2), one sees that $\{e^{\mu \cdot Q_h^\pi} : \mu \in \mathbb{R}\}$ represents the MGF of the cumulative rewards at step h when policy π is executed. Hence, the exponential Bellman equation for risk-sensitive RL provides an instantiation of Equation (20) through the MGF of rewards.

B Proof of Theorem 1

First, we set some notations and definitions. Define $\iota := \log(2HSAK/\delta)$ for a given $\delta \in (0,1]$. We adopt the shorthands $\mathbb{I}_h^\tau(s,a) := \mathbb{I}\{(s_h^\tau, a_h^\tau) = (s,a)\}$ and $r_h^\tau := r_h(s_h^\tau, a_h^\tau)$ for $(\tau, h) \in [K] \times [H]$. We let $N_h^k(s,a)$ be the visit count of (h,s,a) at the beginning of episode k. We denote by V_h^k, G_h^k, b_h^k the values of V_h, G_h, b_h after the updates in step h of episode k, respectively. We also set $Q_h^k = \frac{1}{\beta}\log\{G_h^k\}$.

For the time being we consider $\beta > 0$. For $h \in [H]$, we define

$$\delta_h^k := e^{\beta V_h^k(s_h^k)} - e^{\beta V_h^{\pi^k}(s_h^k)},$$
$$\zeta_{h+1}^k := [P_h(e^{\beta[r_h(s_h^k,a_h^k) + V_{h+1}^k(s')]} - e^{\beta[r_h(s_h^k,a_h^k) + V_{h+1}^{\pi^k}(s')]})](s_h^k, a_h^k) - e^{\beta r_h(s_h^k,a_h^k)}\delta_{h+1}^k,$$

where $[P_h f](s,a) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[f(s')]$ for any $f : \mathcal{S} \to \mathbb{R}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. It can be seen that b_h^k in Algorithm 1 can be equivalently defined as

$$b_h^k := c(e^{\beta(H-h+1)} - 1)\sqrt{\frac{S\iota}{\max\{1, N_h^k(s_h^k, a_h^k)\}}}, \tag{23}$$

where c is the universal constant from Lemma 2. For any $(k,h) \in [K] \times [H]$, we have

$$\delta_h^k \stackrel{(i)}{=} (e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^{\pi^k}})(s_h^k, a_h^k)$$

$$\overset{(ii)}{=} \left[\min\{e^{\beta(H-h+1)}, (w_h^k + b_h^k)(s_h^k, a_h^k)\} - \mathbb{E}_{s' \sim P_h(\cdot \mid s_h^k, a_h^k)} e^{\beta[r_h(s_h^k, a_h^k) + V_{h+1}^k(s')]}\right]$$

$$+ \left[\mathbb{E}_{s' \sim P_h(\cdot \mid s_h^k, a_h^k)} e^{\beta[r_h(s_h^k, a_h^k) + V_{h+1}^k(s')]} - \mathbb{E}_{s' \sim P_h(\cdot \mid s_h^k, a_h^k)} e^{\beta[r_h(s_h^k, a_h^k) + V_{h+1}^{\pi^k}(s')]}\right]$$

$$\overset{(iii)}{\leq} 2b_h^k + [P_h(e^{\beta[r_h(s_h^k, a_h^k) + V_{h+1}^k(s')]} - e^{\beta[r_h(s_h^k, a_h^k) + V_{h+1}^{\pi^k}(s')]})](s_h^k, a_h^k)$$

$$= 2b_h^k + e^{\beta \cdot r_h(s_h^k, a_h^k)} \delta_{h+1}^k + \zeta_{h+1}^k \tag{24}$$

In the above equation, step (i) holds by the construction of Algorithm 1 and the definition of $V_h^{\pi^k}$ in Equation (3); step (ii) holds by Equations (29) and (30). step (iii) holds on the event of Lemma 2; the last step follows from Lemma 4.

Using the fact that $V_{H+1}^k(s) = V_{H+1}^{\pi^k}(s) = 0$ and that $r_h(\cdot, \cdot) \in [0, 1]$, we can expand the recursion in Equation (24) and get

$$\delta_1^k \leq \sum_{h \in [H]} e^{\beta(h-1)} \zeta_{h+1}^k + 2 \sum_{h \in [H]} e^{\beta(h-1)} b_h^k.$$

Summing the above display over $k \in [K]$ gives

$$\sum_{k \in [K]} \delta_1^k \leq \sum_{k \in [K]} \sum_{h \in [H]} e^{\beta(h-1)} \zeta_{h+1}^k + 2 \sum_{k \in [K]} \sum_{h \in [H]} e^{\beta(h-1)} b_h^k \tag{25}$$

Let us now control the two terms in Equation (25). Note that $\{\zeta_{h+1}^k\}$ is a martingale difference sequence satisfying $|\zeta_h^k| \leq 2H$ for all $(k, h) \in [K] \times [H]$. By the Azuma-Hoeffding inequality, we have for any $t > 0$,

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

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

$$\sum_{k \in [K]} \sum_{h \in [H]} e^{\beta(h-1)} \zeta_{h+1}^k \leq (e^{\beta H} - 1)\sqrt{2HK \log(2/\delta)} \leq (e^{\beta H} - 1)\sqrt{2HK\iota}, \tag{26}$$

where $\iota = \log(2HSAK/\delta)$. For the second term in Equation (25), recall the definition of b_h^k in Equation (23), and we can derive

$$\begin{aligned}
\sum_{k \in [K]} \sum_{h \in [H]} e^{\beta(h-1)} b_h^k &\leq \sum_{k \in [K]} \sum_{h \in [H]} c(e^{\beta H} - 1)\sqrt{S\iota} \sqrt{\frac{1}{\max\{1, N_h^k(s_h^k, a_h^k)\}}} \\
&= c(e^{\beta H} - 1)\sqrt{S\iota} \sum_{k \in [K]} \sum_{h \in [H]} \sqrt{\frac{1}{\max\{1, N_h^k(s_h^k, a_h^k)\}}} \\
&\overset{(i)}{\leq} c(e^{\beta H} - 1)\sqrt{S\iota} \sum_{h \in [H]} \sqrt{K} \sqrt{\sum_{k \in [K]} \frac{1}{\max\{1, N_h^k(s_h^k, a_h^k)\}}} \\
&\leq c(e^{\beta H} - 1)\sqrt{S\iota}\sqrt{2H^2SAK\iota},
\end{aligned} \tag{27}$$

where step (i) follows the Cauchy-Schwarz inequality and the last step holds by the pigeonhole principle. Plugging Equations (26) and (27) back to Equation (25) yields

$$\begin{aligned}
\sum_{k \in [K]} \delta_1^k &\leq (e^{\beta H} - 1)\sqrt{2HK\iota} + 2c(e^{\beta H} - 1)\sqrt{2H^2S^2AK\iota^2} \\
&\lesssim (e^{\beta H} - 1)\sqrt{2H^2S^2AK\iota^2},
\end{aligned}$$

The proof for $\beta > 0$ is completed by invoking Lemma 9 on the event of Lemma 4. We note that the proof of $\beta < 0$ follows a similar procedure and is therefore omitted.

B.1 Auxiliary lemmas

Let us fix a pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Recall from Algorithm 1 that

$$w_h^k(s,a) = \frac{1}{N_h^k(s,a)} \sum_{\tau \in [k-1]} \mathbb{I}_h^\tau(s,a) \left[e^{\beta[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]} \right]. \tag{28}$$

If $N_h^k(s,a) \geq 1$, we define

$$q_{h,1}^{k,+}(s,a) := \begin{cases} w_h^k(s,a) + b_h^k(s,a), & \text{if } \beta > 0, \\ w_h^k(s,a) - b_h^k(s,a), & \text{if } \beta < 0. \end{cases}$$

$$q_{h,1}^{k}(s,a) := \begin{cases} \min\{q_{h,1}^{k,+}(s,a), e^{\beta(H-h+1)}\}, & \text{if } \beta > 0, \\ \max\{q_{h,1}^{k,+}(s,a), e^{\beta(H-h+1)}\}, & \text{if } \beta < 0, \end{cases}$$

and if $N_h^k(s,a) = 0$, we let

$$q_{h,1}^{k,+}(s,a) = q_{h,1}^{k}(s,a) := e^{\beta(H-h+1)}.$$

Also define

$$q_{h,2}^k(s,a) := \begin{cases} \frac{1}{N_h^k(s,a)} \sum_{\tau \in [k-1]} \mathbb{I}_h^\tau(s,a) \left[\mathbb{E}_{s' \sim P_h(\cdot \,|\, s_h^\tau, a_h^\tau)} e^{\beta[r_h^\tau + V_{h+1}^k(s')]} \right], & \text{if } N_h^k(s,a) \geq 1, \\ e^{\beta(H-h+1)} \text{ if } \beta > 0; \qquad 1 \text{ if } \beta < 0, & \text{if } N_h^k(s,a) = 0, \end{cases}$$

and for any policy π,

$$q_{h,3}^{k,\pi}(s,a) := \begin{cases} \frac{1}{N_h^k(s,a)} \sum_{\tau \in [k-1]} \mathbb{I}_h^\tau(s,a) \left[\mathbb{E}_{s' \sim P_h(\cdot \,|\, s_h^\tau, a_h^\tau)} e^{\beta[r_h^\tau + V_{h+1}^\pi(s')]} \right], & \text{if } N_h^k(s,a) \geq 1, \\ e^{\beta \cdot Q_h^\pi(s,a)} & \text{if } N_h^k(s,a) = 0. \end{cases}$$

It can be seen that

$$q_{h,2}^k(s,a) = \mathbb{E}_{s' \sim P_h(\cdot \,|\, s,a)} e^{\beta[r_h(s,a) + V_{h+1}^k(s')]} \tag{29}$$

when $N_h^k(s,a) \geq 1$, and

$$q_{h,3}^{k,\pi}(s,a) = e^{\beta Q_h^\pi(s,a)} = \mathbb{E}_{s' \sim P_h(\cdot \,|\, s,a)} e^{\beta[r_h(s,a) + V_{h+1}^\pi(s')]} \tag{30}$$

for all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$ by the exponential Bellman equation (5). We have that if $\beta > 0$,

$$e^{\beta Q_h^k} - e^{\beta Q_h^\pi} = q_{h,1}^k - q_{h,3}^{k,\pi} = (q_{h,1}^k - q_{h,2}^k) + (q_{h,2}^k - q_{h,3}^{k,\pi}), \tag{31}$$

and if $\beta < 0$,

$$e^{\beta Q_h^\pi} - e^{\beta Q_h^k} = q_{h,3}^{k,\pi} - q_{h,1}^k = (q_{h,3}^{k,\pi} - q_{h,2}^k) + (q_{h,2}^k - q_{h,1}^k). \tag{32}$$

Let us state a uniform concentration result.

Lemma 1. *Define* $\iota := \log(2HSAK/\delta)$ *and*

$$\bar{\mathcal{V}}_{h+1} := \left\{ \bar{V}_{h+1} : \mathcal{S} \to \mathbb{R} \mid \forall s \in \mathcal{S},\ \bar{V}_{h+1}(s) \in [0, H-h] \right\}.$$

For any $\delta \in (0,1]$*, there extsts a universal constant* $c_0 > 0$ *such that with probability* $1-\delta$*, we have*

$$\left| \frac{1}{N_h^k(s,a)} \sum_{\tau \in [k-1]} \mathbb{I}_h^\tau(s,a) \left[e^{\beta[r_h^\tau + \bar{V}(s_{h+1}^\tau)]} - \mathbb{E}_{s' \sim P_h(\cdot \,|\, s_h^\tau, a_h^\tau)} e^{\beta[r_h^\tau + \bar{V}(s')]} \right] \right|$$
$$\leq c_0 (e^{\beta(H-h+1)} - 1) \sqrt{\frac{S\iota}{\max\{1, N_h^k(s,a)\}}}$$

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

Proof. The result is a simple adaptation of [20, Lemma 6]. □

We now control the difference $q_{h,1}^k - q_{h,2}^k$.

Lemma 2. *Recall the definition of b_h^k from Algorithm 1. For all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, there exists some universal constant $c > 0$ (where c is used in Line 6 of Algorithm 1) such that the following holds with probability at least $1 - \delta/2$: if $\beta > 0$, we have*

$$0 \le (q_{h,1}^k - q_{h,2}^k)(s,a) \le 2b_h^k,$$

and if $\beta < 0$, we have

$$0 \le (q_{h,2}^k - q_{h,1}^k)(s,a) \le 2b_h^k.$$

Proof. Let us fix a tuple $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

Case $\beta > 0$**.** For $N_h^k(s,a) = 0$, we have $q_{h,1}^k \le e^{\beta(H-h+1)}$ and $q_{h,2}^k \ge 1$ by construction and the result follows immediately. Now we assume $N_h^k(s,a) \ge 1$. By Equation (28) we can compute

$$\begin{aligned}
&\left|(q_{h,1}^{k,+} - b_h^k - q_{h,2}^k)(s,a)\right| \\
&= \left| \frac{1}{N_h^k(s,a)} \sum_{\tau \in [k-1]} \mathbb{I}_h^\tau(s,a) \left[e^{\beta[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]} - \mathbb{E}_{s' \sim P_h(\cdot \,|\, s_h^\tau, a_h^\tau)}[e^{\beta[r_h^\tau + V_{h+1}^k(s')]}] \right] \right| \\
&\le c_0 (e^{\beta(H-h+1)} - 1) \sqrt{\frac{S\iota}{\max\{1, N_h^k(s,a)\}}},
\end{aligned}$$

where the last step holds by Lemma 1 with $c_0 > 0$ being a universal constant. Setting c in b_h^k to be equal to c_0, we have

$$0 \le (q_{h,1}^{k,+} - q_{h,2}^k)(s,a) \le 2b_h^k.$$

Therefore, we have $q_{h,1}^k \ge q_{h,2}^k$ by the first inequality above, the definition of $q_{h,1}^k$ and the property $q_{h,2}^k \le e^{\beta(H-h+1)}$. Also, since $q_{h,1}^{k,+} \ge q_{h,1}^k$, it holds that $q_{h,1}^k - q_{h,2}^k \le q_{h,1}^{k,+} - q_{h,2}^k$. The conclusion follows.

Case $\beta < 0$**.** We have, similar to the previous case, that

$$\left|(q_{h,1}^{k,+} - b_h^k - q_{h,2}^k)(s,a)\right| \le c_0 (1 - e^{\beta(H-h+1)}) \sqrt{\frac{S\iota}{\max\{1, N_h^k(s,a)\}}}.$$

Choosing $c = c_0$ in the definition of $b_h^k(s,a)$ leads to

$$0 \le (q_{h,2}^k - q_{h,1}^{k,+})(s,a) \le 2b_h^k.$$

This implies $q_{h,2}^k \ge q_{h,1}^{k,+}$, and since $q_{h,1}^k, q_{h,2}^k \ge e^{\beta(H-h+1)}$, we also have $q_{h,2}^k \ge q_{h,1}^k$. In addition, since $q_{h,1}^{k,+} \le q_{h,1}^{k,}$, it also holds that $q_{h,2}^k - q_{h,1}^k \le q_{h,2}^k - q_{h,1}^{k,+}$. Then the conclusion of this case follows. ☐

Lemma 3. *On the event of Lemma 2, for all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$ and any policy π, we have*

$$\begin{cases} e^{\beta \cdot Q_h^k(s,a)} \ge e^{\beta \cdot Q_h^\pi(s,a)}, & \text{if } \beta > 0, \\ e^{\beta \cdot Q_h^k(s,a)} \le e^{\beta \cdot Q_h^\pi(s,a)}, & \text{if } \beta < 0. \end{cases}$$

Proof. We focus on the case of $\beta > 0$ since the proof for $\beta < 0$ is very similar. 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 $e^{\beta \cdot Q_{H+1}^k(s,a)} = e^{\beta \cdot Q_{H+1}^\pi(s,a)} = 1$ for $k \in [K]$ by definition. Now we fix an $h \in [H]$ and assume that $e^{\beta \cdot Q_{h+1}^k(s,a)} \ge e^{\beta \cdot Q_{h+1}^*(s,a)}$. Moreover, by the induction assumption we have

$$e^{\beta \cdot V_{h+1}^k(s)} = \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^k(s,a')} \ge \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^\pi(s,a')} \ge e^{\beta \cdot V_{h+1}^\pi(s)}. \tag{33}$$

We also assume that (s,a) satisfies $N_h^k(s,a) \ge 1$, since otherwise $e^{\beta \cdot Q_h^k(s,a)} = e^{\beta(H-h+1)} \ge e^{\beta \cdot Q_h^\pi(s,a)}$ and we are done. This assumption and Equation (33) together imply $q_{h,2}^k \ge q_{h,3}^{k,\pi}$ by Lemma 2. We also have $q_{h,1}^k \ge q_{h,2}^k$ on the event of Lemma 2. Therefore, it follows that $e^{\beta \cdot Q_h^k(s,a)} \ge e^{\beta \cdot Q_h^\pi(s,a)}$ by Equation (31) and we have completed the induction. ☐

Lemma 4. *For all* $(k,h,s) \in [K]\times[H]\times\mathcal{S}$ *and any* $\delta \in (0,1]$*, with probability at least* $1-\delta/2$*, we have*

$$\begin{cases} e^{\beta\cdot V_h^k(s)} \geq e^{\beta\cdot V_h^{\pi}(s)}, & \text{if } \beta > 0, \\ e^{\beta\cdot V_h^k(s)} \leq e^{\beta\cdot V_h^{\pi}(s)}, & \text{if } \beta < 0. \end{cases}$$

Proof. The result follows from Lemma 3 and Equation (33). □

C Proof of Theorem 2

We first lay out some additional notations to facilitate our proof. Let N_h^k, G_h^k, V_h^k be the N_h, G_h, V_h functions at the beginning of the episode k, before t is updated. We also set $Q_h^k := \frac{1}{\beta} G_h^k$. We let $\widehat{P}_h^k(\cdot \mid s,a)$ denote the delta function centered at s_{h+1}^k for all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$. This means $\mathbb{E}_{s'\sim\widehat{P}_h^k(\cdot \mid s,a)}[f(s')] = f(s_{h+1}^k)$ for any $f:\mathcal{S}\to\mathbb{R}$. Denote by $n_h^k := N_h^k(s_h^k, a_h^k)$. Recall from Algorithm 2, the learning rate is defined as

$$\alpha_t := \frac{H+1}{H+t}, \tag{34}$$

for $t \in \mathbb{Z}$.

For now we consider the case for $\beta > 0$. We define the following quantities to ease the notations for the proof:

$$\begin{aligned} \delta_h^k &:= e^{\beta\cdot V_h^k(s_h^k)} - e^{\beta\cdot V_h^{\pi^k}(s_h^k)}, \\ \phi_h^k &:= e^{\beta\cdot V_h^k(s_h^k)} - e^{\beta\cdot V_h^{*}(s_h^k)}, \\ \xi_{h+1}^k &:= [(P_h - \widehat{P}_h^k)(e^{\beta\cdot V_{h+1}^*} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k, a_h^k). \end{aligned}$$

For each fixed $(k,h) \in [K]\times[H]$, we let $t = N_h^k(s_h^k, a_h^k)$. Then it holds that

$$\begin{aligned} \delta_h^k &\overset{(i)}{=} e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^{\pi^k}(s_h^k,a_h^k)} \\ &= [e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^*(s_h^k,a_h^k)}] + [e^{\beta\cdot Q_h^*(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^{\pi^k}(s_h^k,a_h^k)}] \\ &\overset{(ii)}{=} [e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^*(s_h^k,a_h^k)}] + e^{\beta\cdot r_h(s_h^k,a_h^k)}[P_h(e^{\beta\cdot V_{h+1}^*} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k,a_h^k) \\ &\overset{(iii)}{\leq} [e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^*(s_h^k,a_h^k)}] + e^{\beta}[P_h(e^{\beta\cdot V_{h+1}^*} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k,a_h^k) \\ &= [e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^*(s_h^k,a_h^k)}] + e^{\beta}(\delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k) \\ &\overset{(iv)}{\leq} \alpha_t^0(e^{\beta(H-h+1)} - 1) + 2\gamma_{h,t} + \sum_{i\in[t]} \alpha_t^i \cdot e^{\beta}\left[e^{\beta\cdot V_{h+1}^{k_i}(s_{h+1}^{k_i})} - e^{\beta\cdot V_{h+1}^*(s_{h+1}^{k_i})}\right] \\ &\quad + e^{\beta}(\delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k) \\ &= \alpha_t^0(e^{\beta(H-h+1)} - 1) + 2\gamma_{h,t} + \sum_{i\in[t]} \alpha_t^i \cdot e^{\beta}\phi_{h+1}^{k_i} \\ &\quad + e^{\beta}(\delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k) \end{aligned} \tag{35}$$

where step (i) holds since $V_h^k(s_h^k) = \max_{a'\in\mathcal{A}} Q_h^k(s_h^k,a') = Q_h^k(s_h^k,a_h^k)$ and $V_h^{\pi^k}(s_h^k) = Q_h^{\pi^k}(s_h^k, \pi_h^k(s_h^k)) = Q_h^{\pi^k}(s_h^k,a_h^k)$; step (ii) holds by the exponential Bellman equation (5); step (iii) holds since $V_{h+1}^* \geq V_{h+1}^{\pi^k}$ implies $e^{\beta\cdot V_{h+1}^*} \geq e^{\beta\cdot V_{h+1}^{\pi^k}}$ given that $\beta > 0$; step (iv) holds on the event of Lemma 8 (with $\gamma_{h,t}$ defined therein).

We bound each term in (35) one by one. First, we have

$$\sum_{k\in[K]} \alpha_{n_h^k}^0 (e^{\beta(H-h+1)} - 1) = (e^{\beta(H-h+1)} - 1)\sum_{k\in[K]} \mathbb{I}\{n_h^k = 0\}$$

$$\leq (e^{\beta(H-h+1)} - 1)SA.$$

The second term in (35) can be bounded by

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

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 $(s_h^k, a_h^k) = (s_h^{k'}, a_h^{k'})$. For the first time we visit $(s_h^{k'}, a_h^{k'})$ we have $n_h^k = n_h^{k'} + 1$, for the second time we have $n_h^k = n_h^{k'} + 2$, and etc. Therefore, we may continue the above display as

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

where the last step follows Fact 1(c). Collecting the above results and plugging them into Equation (35), we have

$$\begin{aligned}
\sum_{k\in[K]} \delta_h^k &\leq (e^{\beta(H-h+1)} - 1)SA + \left(1 + \frac{1}{H}\right) e^{\beta} \sum_{k\in[K]} \phi_{h+1}^k \\
&\quad + \sum_{k\in[K]} e^{\beta}(\delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k) + 2 \sum_{k\in[K]} \gamma_{h,n_h^k} \\
&\leq (e^{\beta(H-h+1)} - 1)SA + \left(1 + \frac{1}{H}\right) e^{\beta} \sum_{k\in[K]} \delta_{h+1}^k \\
&\quad + \sum_{k\in[K]} (2\gamma_{h,n_h^k} + e^{\beta} \xi_{h+1}^k),
\end{aligned} \tag{36}$$

where the last step holds since $\delta_{h+1}^k \geq \phi_{h+1}^k$ (due to the fact that $\beta > 0$ and $V_{h+1}^* \geq V_{h+1}^{\pi^k}$).

Now, we unroll the quantity $\sum_{k\in[K]} \delta_h^k$ recursively in the form of Equation (36), and get

$$\begin{aligned}
\sum_{k\in[K]} \delta_1^k &\leq \sum_{h\in[H]} \left[\left(1 + \frac{1}{H}\right) e^{\beta} \right]^{h-1} \left[(e^{\beta(H-h+1)} - 1)SA + \sum_{k\in[K]} (2\gamma_{h,n_h^k} + e^{\beta} \xi_{h+1}^k) \right] \\
&= \sum_{h\in[H]} \left(1 + \frac{1}{H}\right)^{h-1} \left[(e^{\beta H} - e^{\beta(h-1)})SA + \sum_{k\in[K]} (2e^{\beta(h-1)} \gamma_{h,n_h^k} + e^{\beta h} \xi_{h+1}^k) \right] \\
&= \sum_{h\in[H]} \left(1 + \frac{1}{H}\right)^{h-1} \left[(e^{\beta H} - e^{\beta(h-1)})SA + \sum_{k\in[K]} 2e^{\beta(h-1)} \gamma_{h,n_h^k} \right] \\
&\quad + \sum_{h\in[H]} \sum_{k\in[K]} \left(1 + \frac{1}{H}\right)^{h-1} e^{\beta h} \xi_{h+1}^k \\
&\leq e \left[(e^{\beta H} - 1)HSA + \sum_{k\in[K]} \sum_{h\in[H]} 2e^{\beta(h-1)} \gamma_{h,n_h^k} \right] + \sum_{h\in[H]} \sum_{k\in[K]} \left(1 + \frac{1}{H}\right)^{h-1} e^{\beta h} \xi_{h+1}^k,
\end{aligned} \tag{37}$$

where the first step uses the fact that $\delta^k_{H+1} = 0$ for $k \in [K]$; the last step holds since $(1 + 1/H)^h \leq (1 + 1/H)^H \leq e$ for all $h \in [H]$. By the pigeonhole principle, for any $h \in [H]$ we have

$$\begin{aligned}\sum_{k\in[K]} \sum_{h\in[H]} e^{\beta(h-1)} \gamma_{n_h^k} &\lesssim (e^{\beta H} - 1) \sum_{k\in[K]} \sqrt{\frac{H\iota}{n_h^k}} \\ &\lesssim (e^{\beta H} - 1) \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{n\in[N_h^K(s,a)]} \sqrt{\frac{H\iota}{n}} \\ &\lesssim (e^{\beta H} - 1)\sqrt{HSAK\iota} \end{aligned} \tag{38}$$

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 and the fact that $\left|\left(1 + \frac{1}{H}\right)^{h-1} e^{\beta h}\xi^k_{h+1}\right| \leq e(e^{\beta H} - 1)$ for $h \in [H]$ together imply that with probability at least $1 - \delta$, we have

$$\left|\sum_{h\in[H]} \sum_{k\in[K]} \left(1 + \frac{1}{H}\right)^{h-1} e^{\beta h}\xi^k_{h+1}\right| \lesssim (e^{\beta H} - 1)\sqrt{HK\iota}. \tag{39}$$

Plugging Equations (38) and (39) into (37), we have

$$\sum_{k\in[K]} \delta^k_1 \lesssim (e^{\beta H} - 1)\sqrt{HSAK\iota},$$

when K is large enough. Invoking Lemma 9 completes the proof for the case $\beta > 0$.

The proof is very similar for the case of $\beta < 0$, and one only needs to exchange the role of V_h^k and $V_h^{\pi^k}$ in the definitions of δ^k_h, ϕ^k_h, ξ^k_h, etc, to get the counterpart of Equation (35) and of the remaining analysis.

C.1 Auxiliary lemmas

Recall the learning rate α_t defined in Equation (34). We define

$$\alpha_t^0 := \prod_{j=1}^{t} (1 - \alpha_j), \qquad \alpha_t^i := \alpha_i \prod_{j=i+1}^{t} (1 - \alpha_j) \tag{40}$$

for integers $i, t \geq 1$. 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 $t < i + 1$.

In the following, we provide some useful facts about the learning rate.

Fact 1. *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 [27, Lemma 4.1], and the last one follows from direct calculation in view of Equation (40). □

Define the shorthand $\iota := \log(SAT/\delta)$ for $\delta \in (0, 1]$. 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 visited the i-th time at step h. Let us define

$$q_{h,1}^{k,+}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \begin{cases} \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]} + b_{h,i}\right], & \text{if } \beta > 0, \\ \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]} - b_{h,i}\right], & \text{if } \beta < 0, \end{cases}$$

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

and

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

$$q_{h,2}^{k,+}(s,a) \coloneqq \alpha_t^0 e^{\beta(H-h+1)} + \begin{cases} \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} + b_{h,i}\right], & \text{if } \beta > 0, \\ \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} - b_{h,i}\right], & \text{if } \beta < 0, \end{cases}$$

$$q_{h,2}^k(s,a) \coloneqq \begin{cases} \min\{q_{h,2}^{k,+}(s,a), e^{\beta(H-h+1)}\}, & \text{if } \beta > 0, \\ \max\{q_{h,2}^{k,+}(s,a), e^{\beta(H-h+1)}\}, & \text{if } \beta < 0, \end{cases}$$

and

$$q_{h,3}^k(s,a) \coloneqq \alpha_t^0 e^{\beta \cdot Q_h^*(s,a)} + \sum_{i\in[t]} \alpha_t^i \left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}\right].$$

We have a simple fact on $q_{h,2}^k$ and $q_{h,2}^{k,\circ}$.

Fact 2. *If $\beta > 0$, we have $q_{h,2}^{k,\circ}(\cdot,\cdot) \le q_{h,2}^k(\cdot,\cdot)$; if $\beta < 0$, we have $q_{h,2}^{k,\circ}(\cdot,\cdot) \ge q_{h,2}^k(\cdot,\cdot)$.*

Proof. We focus on the case where $\beta > 0$ and the case for $\beta < 0$ can be proved similarly. Note that $r_h(s,a) + V_{h+1}^*(s_{h+1}^{k_i}) \in [0, H-h+1]$ 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 1(d). Together they imply that $q_{h,2}^{k,\circ}(\cdot,\cdot) \le e^{\beta(H-h+1)}$ and $(q_{h,2}^{k,\circ} - q_{h,2}^{k,+})(\cdot,\cdot) = -\sum_{i\in[t]} \alpha_t^i b_{h,i} \le 0$ by definition of $b_{h,i}$ in Line 8 of Algorithm 2. Therefore, $q_{h,2}^{k,\circ}(\cdot,\cdot) \le \min\{e^{\beta(H-h+1)}, q_{h,2}^{k,+}(\cdot,\cdot)\} = q_{h,2}^k(\cdot,\cdot)$. □

Next, we write the difference $e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*}$ in terms of $q_{h,1}^k$ and $q_{h,3}^k$.

Lemma 5. For any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, suppose (s,a) was previously visited at step h of episodes $k_1, \ldots, k_t < k$. We have

$$(e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*})(s,a) = (q_{h,1}^k - q_{h,3}^k)(s,a).$$

Proof. For $e^{\beta \cdot Q_h^k}$, Line 10 of Algorithm 2 implies that

$$e^{\beta \cdot Q_h^k(s,a)} = q_{h,1}^k(s,a). \tag{41}$$

For $e^{\beta \cdot Q_h^*}$, we have from exponential Bellman equation (5) that

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

Let $t = N_h^k(s,a)$ and by Fact 1(d), we have

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

for each integer $t \ge 0$. By the definition of $q_{h,3}^k$ we have

$$e^{\beta \cdot Q_h^*(s,a)} = q_{h,3}^k(s,a). \tag{42}$$

The proof is completed by combining Equations (41) and (42). □

From Lemma 5, we can derive the decomposition

$$(e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*})(s,a) = (q_{h,1}^k - q_{h,2}^k)(s,a) + (q_{h,2}^k - q_{h,3}^k)(s,a) \tag{43}$$

if $\beta > 0$, and

$$(e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*})(s,a) = (q_{h,2}^k - q_{h,1}^k)(s,a) + (q_{h,3}^k - q_{h,2}^k)(s,a) \tag{44}$$

if $\beta < 0$. We have the following lemmas.

Lemma 6. *There exists a universal constant $c > 0$ in the definition of $b_{h,t}$ in Algorithm 2 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*

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

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

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

Proof. We focus on the case where $\beta > 0$ and the proof for $\beta < 0$ is similar. For any $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, define

$$\begin{aligned} \psi(i, k, h, s, a) &:= e^{\beta[r_h(s,a) + V_{h+1}^*(s_{h+1}^{k_i})]} - \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)} \big[e^{\beta[r_h(s,a) + V_{h+1}^*(s')]} \big] \\ &= \mathbb{E}_{s' \sim \hat{P}_h^{k_i}(\cdot \mid s,a)} \big[e^{\beta[r_h(s,a) + V_{h+1}^*(s')]} \big] - \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)} \big[e^{\beta[r_h(s,a) + V_{h+1}^*(s')]} \big] \end{aligned}$$

Let us fix a tuple $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. We have 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 $\tau \in [K]$, it holds that with probability at least $1 - \delta/(HSA)$, for all $\tau \in [K]$,

$$\left| \sum_{i \in [\tau]} \alpha_\tau^i \cdot \mathbb{I}(k_i \leq K) \cdot \psi(i, k, h, s, a) \right|$$
$$\leq \frac{c}{2} (e^{\beta(H-h+1)} - 1) \sqrt{\iota \sum_{i \in [\tau]} (\alpha_\tau^i)^2} \leq c (e^{\beta(H-h+1)} - 1) \sqrt{\frac{H\iota}{\tau}}$$

where $c > 0$ is some universal constant, the first step holds since $r_h(s,a) + V_{h+1}^*(s') \in [0, H-h+1]$ for $s' \in \mathcal{S}$, and the last step follows from Fact 1(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 $1 - \delta$:

$$\left| \sum_{i \in [t]} \alpha_\tau^i \cdot \psi(i, k, h, s, a) \right| \leq c (e^{\beta(H-h+1)} - 1) \sqrt{\frac{H\iota}{t}}, \tag{45}$$

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

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

For bounds on $\sum_{i \in [t]} \alpha_t^i b_{h,i}$, we recall the definition of $\{b_{h,t}\}$ in Line 8 of Algorithm 2 and compute

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

where the last step holds by Fact 1(a). $\square$

The next two lemmas compare the iterate $e^{\beta \cdot Q_h^k}$ (and $e^{\beta \cdot V_h^k}$) with the optimal exponential value function $e^{\beta \cdot Q_h^*}$ (and $e^{\beta \cdot V_h^*}$).

Lemma 7. *For all $(k, h, s, a)$and any $\delta \in (0, 1]$, it holds with probability at least $1 - \delta$ that*

$$\begin{cases} e^{\beta \cdot Q_h^k(s,a)} \geq e^{\beta \cdot Q_h^*(s,a)}, & \text{if } \beta > 0, \\ e^{\beta \cdot Q_h^k(s,a)} \leq e^{\beta \cdot Q_h^*(s,a)}, & \text{if } \beta < 0. \end{cases}$$

Proof. We focus on the case where $\beta > 0$ and the proof for $\beta < 0$ is similar. For the purpose of the proof, we set $Q_{H+1}^k(s, a) = Q_{H+1}^*(s, a) = 0$ 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 $(s, a) = (s_h^k, a_h^k)$ (that is, (s, a) has been visited at some point in Algorithm 2), since otherwise $e^{\beta \cdot Q_h^k(s,a)} = e^{\beta(H-h+1)} \geq e^{\beta \cdot Q_h^*(s,a)}$ 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 $e^{\beta \cdot Q_{H+1}^{k'}(s,a)} = e^{\beta \cdot Q_{H+1}^*(s,a)}$ for $k' \in [K]$ by definition. We fix a $(k, h) \in [K] \times [H]$ and assume that $e^{\beta \cdot Q_{h+1}^{k_i}(s,a)} \geq e^{\beta \cdot Q_{h+1}^*(s,a)}$ for each $k_1, \ldots, k_t < k$ (here $t = N_h^k(s, a)$). Then we have for $i \in [t]$ that

$$e^{\beta \cdot V_{h+1}^{k_i}(s)} = \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{k_i}(s,a')} \geq \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^*(s,a')} = e^{\beta \cdot V_{h+1}^*(s)},$$

where the first equality holds by the update procedure in Algorithm 2. Recall the decomposition in Equation (43). The above displayed equation implies $q_{h,1}^k \geq q_{h,2}^k$ by the definition of $q_{h,2}^k$. We also have $q_{h,2}^k \geq q_{h,3}^k$ by the fact $e^{\beta \cdot Q_h^*(s,a)} \leq e^{\beta(H-h+1)}$ and on the event of Lemma 6. Therefore, it follows that $(e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*})(s, a) \geq 0$ by Equation (43). The induction is completed. $\square$

Lemma 8. *For all $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$ such that $t = N_h^k(s, a) \geq 1$, let $\gamma_{h,t} := 2 \sum_{i \in [t]} \alpha_t^i b_{h,i}$ and let $k_1, \ldots, k_t < k$ be the episodes in which (s, a) is visited at step h. Then the following holds with probability at least $1 - \delta$: if $\beta > 0$, we have*

$$\begin{aligned} &(e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*})(s, a) \\ &\leq \alpha_t^0 \left[e^{\beta(H-h+1)} - 1 \right] + 2\gamma_{h,t} + \sum_{i \in [t]} \alpha_t^i e^\beta \left[e^{\beta \cdot V_{h+1}^{k_i}(s_{h+1}^{k_i})} - e^{\beta \cdot V_{h+1}^*(s_{h+1}^{k_i})} \right], \end{aligned}$$

and if $\beta < 0$, we have

$$\begin{aligned} &(e^{\beta \cdot Q_h^*} - e^{\beta \cdot Q_h^k})(s, a) \\ &\leq \alpha_t^0 \left[1 - e^{\beta(H-h+1)} \right] + 2\gamma_{h,t} + \sum_{i \in [t]} \alpha_t^i \left[e^{\beta \cdot V_{h+1}^*(s_{h+1}^{k_i})} - e^{\beta \cdot V_{h+1}^{k_i}(s_{h+1}^{k_i})} \right]. \end{aligned}$$

Furthermore, we have $\gamma_{h,t} \leq 4c \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{\frac{H\iota}{t}}$.

Proof. Note that by definition,

$$q_{h,1}^k(s, a) = e^{\beta \cdot Q_h^k(s,a)}, \quad q_{h,3}^k(s, a) = e^{\beta \cdot Q_h^*(s,a)}.$$

Let us fix a tuple $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. On the event of Lemma 7, we have

$$\begin{cases} e^{\beta \cdot Q_h^k(s,a)} \geq e^{\beta \cdot Q_h^*(s,a)}, & \text{if } \beta > 0, \\ e^{\beta \cdot Q_h^k(s,a)} \leq e^{\beta \cdot Q_h^*(s,a)}, & \text{if } \beta < 0. \end{cases}$$

This implies that for $i \in [t]$, if $\beta > 0$ then

$$e^{\beta \cdot V_{h+1}^{k_i}(s)} = \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{k_i}(s,a')} \geq \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^*(s,a')} = e^{\beta \cdot V_{h+1}^*(s)},$$

and if $\beta < 0$, then

$$e^{\beta \cdot V_{h+1}^{k_i}(s)} = \min_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{k_i}(s,a')} \leq \min_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^*(s,a')} = e^{\beta \cdot V_{h+1}^*(s)}.$$

Here, the first equalities for the above two displays follow from the update procedure in Algorithm 2.

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

$$
\begin{aligned}
(q_{h,1}^k - q_{h,2}^k)(s,a) &\overset{(i)}{\leq} (q_{h,1}^{k,+} - q_{h,2}^{k,\circ})(s,a) \\
&\overset{(ii)}{\leq} \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]} - e^{\beta[r_h(s,a)+V_{h+1}^{*}(s_{h+1}^{k_i})]} \right] + \sum_{i\in[t]} \alpha_t^i b_{h,i} \\
&\leq \sum_{i\in[t]} \alpha_t^i \cdot e^{\beta} \left[e^{\beta \cdot V_{h+1}^{k_i}(s_{h+1}^{k_i})} - e^{\beta \cdot V_{h+1}^{*}(s_{h+1}^{k_i})} \right] + \gamma_{h,t}
\end{aligned}
$$

where step (i) holds by the fact that $\alpha_t^0, \sum_{i\in[t]} \alpha_t^i \in \{0,1\}$ with $\alpha_t^0 + \sum_{i\in[t]} \alpha_t^i = 1$ by Fact 1(d) (so that $q_{h,1}^k \geq q_{h,2}^k$); step (ii) holds by definitions of $q_{h,1}^{k,+}$ and $q_{h,2}^{k,\circ}$; the last step holds since r_h is in $[0,1]$ entrywise and $V_{h+1}^{k_i}(s) \geq V_{h+1}^*(s)$. Moreover, we have

$$
\begin{aligned}
(q_{h,2}^k - q_{h,3}^k)(s,a) &\overset{(i)}{\leq} (q_{h,2}^{k,+} - q_{h,3}^k)(s,a) \\
&= \alpha_t^0 \left[e^{\beta(H-h+1)} - e^{\beta \cdot Q_h^*(s,a)} \right] + \sum_{i\in[t]} \alpha_t^i b_{h,i} \\
&\quad + \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} - \mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} [e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}] \right] \\
&\leq \alpha_t^0 \left[e^{\beta(H-h+1)} - 1 \right] + \gamma_{h,t},
\end{aligned}
$$

where step (i) holds by

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

on the event of Lemma 6 (so that $q_{h,2}^k \geq q_{h,3}^k$) and Fact 2; the last step holds by $Q_h^* \geq 0$ and on the event of Lemma 6.

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

$$
\begin{aligned}
(q_{h,2}^k - q_{h,1}^k)(s,a) &\overset{(i)}{\leq} (q_{h,2}^{k,\circ} - q_{h,1}^{k,+})(s,a) \\
&= \sum_{i\in[t]} \alpha_t^i \left[e^{\beta[r_h(s,a)+V_{h+1}^{*}(s_{h+1}^{k_i})]} - e^{\beta[r_h(s,a)+V_{h+1}^{k_i}(s_{h+1}^{k_i})]} \right] + \sum_{i\in[t]} \alpha_t^i b_i \\
&\leq \sum_{i\in[t]} \alpha_t^i \left[e^{\beta \cdot V_{h+1}^{*}(s_{h+1}^{k_i})} - e^{\beta \cdot V_{h+1}^{k_i}(s_{h+1}^{k_i})} \right] + \gamma_{h,t},
\end{aligned}
$$

where the step (i) holds since $q_{h,2}^{k,\circ} \geq q_{h,2}^k$ by Fact 2 and $q_{h,1}^{k,+} \leq q_{h,1}^k$ by definition, and the last step holds by the fact that $r_h(s,a) + V_{h+1}^{k_i}(s) \geq r_h(s,a) + V_{h+1}^*(s)$, that $e^{\beta \cdot r_h(s,a)} \leq 1$ given $\beta < 0$, and the definition of $\gamma_{h,t}$. In addition, we can derive

$$
\begin{aligned}
(q_{h,3}^k - q_{h,2}^k)(s,a) &\overset{(i)}{\leq} (q_{h,3}^k - q_{h,2}^{k,+})(s,a) \\
&= \alpha_t^0 \left[1 - e^{\beta \cdot Q_h^*(s,a)} \right] + \sum_{i\in[t]} \alpha_t^i b_i \\
&\quad + \sum_{i\in[t]} \alpha_t^i \left[\mathbb{E}_{s'\sim P_h(\cdot\,|\,s,a)} [e^{\beta[r_h(s,a)+V_{h+1}^*(s')]}] - e^{\beta[r_h(s,a)+V_{h+1}^*(s_{h+1}^{k_i})]} \right] \\
&\overset{(ii)}{\leq} \alpha_t^0 \left[1 - e^{\beta(H-h+1)} \right] + 2 \sum_{i\in[t]} \alpha_t^i b_i
\end{aligned}
$$

$$\leq \alpha_t^0 \left[1 - e^{\beta(H-h+1)}\right] + \gamma_{h,t}.$$

where step (i) holds since $q_{h,2}^k \geq q_{h,2}^{k,+}$, step (ii) holds on the event of Lemma 6, and the last step holds by the definition of $\gamma_{h,t}$.

Combining the above calculations with Equation (43) for the case where $\beta > 0$ (or Equation (44) for the case where $\beta < 0$) yields the upper bound for $(e^{\beta \cdot Q_h^k} - e^{\beta \cdot Q_h^*})(s,a)$ (or $(e^{\beta \cdot Q_h^*} - e^{\beta \cdot Q_h^k})(s,a)$). Furthermore, Lemma 6 and the definition of $\gamma_{h,t}$ together imply

$$\gamma_{h,t} \leq 4c \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{\frac{H\iota}{t}}.$$

The proof is completed. □

We present a simple inequality for the regret.

Lemma 9. *Suppose that for any $k \in [K]$ we have $V_1^k(s_1^k) \geq V_1^*(s_1^k)$. Then for $\beta > 0$, the regret is bounded by*

$$\text{Regret}(K) \leq \frac{1}{\beta} \sum_{k \in [K]} [e^{\beta \cdot V_1^k(s_1^k)} - e^{\beta \cdot V_1^{\pi^k}(s_1^k)}],$$

and for $\beta < 0$, the regret is bounded by

$$\text{Regret}(K) \leq \frac{e^{-\beta H}}{|\beta|} \sum_{k \in [K]} [e^{\beta \cdot V_1^{\pi^k}(s_1^k)} - e^{\beta \cdot V_1^k(s_1^k)}],$$

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

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k \in [K]} (V_1^* - V_1^{\pi^k})(s_1^k) \\
&\overset{(i)}{\leq} \sum_{k \in [K]} (V_1^k - V_1^{\pi^k})(s_1^k) \\
&= \sum_{k \in [K]} \left[\frac{1}{\beta} \log\{e^{\beta \cdot V_1^k(s_1^k)}\} - \frac{1}{\beta} \log\{e^{\beta \cdot V_1^{\pi^k}(s_1^k)}\}\right] \\
&\overset{(ii)}{\leq} \sum_{k \in [K]} \frac{1}{\beta} [e^{\beta \cdot V_1^k(s_1^k)} - e^{\beta \cdot V_1^{\pi^k}(s_1^k)}] \\
&= \frac{1}{\beta} \sum_{k \in [K]} [e^{\beta \cdot V_1^k(s_1^k)} - e^{\beta \cdot V_1^{\pi^k}(s_1^k)}],
\end{aligned}$$

where step (i) holds by our assumption, and step (ii) holds by the 1-Lipschitzness of the function $f(x) = \log x$ for $x \geq 1$ and note that our assumption implies that $V_1^k(s_1^k) \geq V_1^*(s_1^k) \geq V_1^{\pi^k}(s_1^k)$.

For $\beta < 0$, we similarly have

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k \in [K]} (V_1^* - V_1^{\pi^k})(s_1^k) \\
&\overset{(i)}{\leq} \sum_{k \in [K]} (V_1^k - V_1^{\pi^k})(s_1^k) \\
&= \sum_{k \in [K]} \left[\frac{1}{\beta} \log\{e^{\beta \cdot V_1^k(s_1^k)}\} - \frac{1}{\beta} \log\{e^{\beta \cdot V_1^{\pi^k}(s_1^k)}\}\right] \\
&= \sum_{k \in [K]} \left[\frac{1}{(-\beta)} \log\{e^{\beta \cdot V_1^{\pi^k}(s_1^k)}\} - \frac{1}{(-\beta)} \log\{e^{\beta \cdot V_1^k(s_1^k)}\}\right]
\end{aligned}$$

$$\overset{(ii)}{\leq} \sum_{k \in [K]} \frac{e^{-\beta H}}{(-\beta)} [e^{\beta \cdot V_1^{\pi^k}(s_1^k)} - e^{\beta \cdot V_1^k(s_1^k)}]$$
$$= \frac{e^{-\beta H}}{|\beta|} \sum_{k \in [K]} [e^{\beta \cdot V_1^{\pi^k}(s_1^k)} - e^{\beta \cdot V_1^k(s_1^k)}],$$

where step (i) holds by our assumption, and step (ii) holds by the $(e^{-\beta H})$-Lipschitzness of the function $f(x) = \log x$ for $x \geq e^{\beta H}$ and note that our assumption implies that $V_1^k(s_1^k) \geq V_1^*(s_1^k) \geq V_1^{\pi^k}(s_1^k)$. □

Broader impact and future directions. Risk-sensitive RL has close association with neuroscience, psychology and behavioral economics, as it has been applied to model human behaviors [36, 41]. Interestingly, this array of topics are also actively studied by researchers in the areas of meta learning [44], biologically inspired deep learning [42] and deep reinforcement learning [29]. It would be an exciting research direction to establish connections between these related areas through rigorous and theoretical analysis of deep learning [9, 11]. Motivated by the inertia of switching actions that is widely observed in human behaviors, the study of switching constrained algorithms [12] for risk-sensitive RL could be another promising direction for future investigation. Furthermore, to make our algorithms practical and efficient on large-scaled datasets collected in the aforementioned applications, it is imperative to enable offline learning procedures for risk-sensitive RL, possibly by techniques developed in the literature of offline RL [10]. It would also be of great interest to understand the landscape of the optimization problems [30] that arise in the offline learning setting.