A Tighter Problem-Dependent Regret Bound for Risk-Sensitive Reinforcement Learning

Xiaoyan Hu
The Chinese University of Hong Kong

Ho-fung Leung
The Chinese University of Hong Kong

Abstract

We study the regret for risk-sensitive reinforcement learning (RL) with the exponential utility in the episodic MDP. Recent works establish both a lower bound $\Omega((e^{|\beta|(H-1)/2}-1)\sqrt{SAT}/|\beta|)$ and the best known (upper) bound $\tilde{O}((e^{|\beta|H}-1)\sqrt{H^2SAT}/|\beta|)$, where H is the length of the episode, S the size of state space, A the size of action space, T the total number of timesteps, and β the risk parameter. The gap between the upper and the lower bound is exponential and hence is unsatisfactory. In this paper, we show that a variant of UCB-ADVANTAGE algorithm reduces a factor of $\sqrt{H}$ from the best previously known bound in any arbitrary MDP. To further sharpen the regret bound, we introduce a brand new mechanism of regret analysis and derive a problem-dependent regret bound without prior knowledge of the MDP from the algorithm. This bound is much tighter in MDPs with special structures. Particularly, we show that a regret that matches the information-theoretic lower bound up to logarithmic factors can be attained within a rich class of MDPs, which improves an exponential factor over the best previously known bound. Further, we derive a novel information-theoretic lower bound of $\Omega(\max_{h\in[H]} c^*_{v,h+1}\sqrt{SAT}/|\beta|)$, where $\max_{h\in[H]} c^*_{v,h+1}$ is a problem-dependent statistic. This lower bound shows that the problem-dependent regret bound achieved by the algorithm is optimal in its dependence on $\max_{h\in[H]} c^*_{v,h+1}$.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

1 INTRODUCTION

Risk-sensitive reinforcement learning (RL) studies the problem of an agent interacting with an unknown environment and making decisions based on both expected reward and *risk* (Howard and Matheson, 1972). In contrast to conventional RL, the need for treatment of risk in various real-world applications such as portfolio optimization (Kuroda and Nagai, 2002) and automated driving (Bernhard et al., 2019) motivates a line of studies in risk-sensitive RL. Recent studies from psychology and neuroscience also employ risk-sensitive RL to understand how the attitude to risk influences humans' decision process (Niv et al., 2012).

Several criteria are proposed to measure risk in RL, including mean-variance criterion (Sani et al., 2012) and conditional value-at-risk (Cassel et al., 2018; Huang and Haskell, 2021). In this paper, we follow a line of studies (Howard and Matheson, 1972; Masi and Stettner, 1999; Borkar, 2002; Cavazos-Cadena and Hernández-Hernández, 2011; Osogami, 2012; Maillard, 2013) that considers risk-sensitive RL with the exponential utility (EU), where the agent aims to maximize the following risk-sensitive objective function under a policy π

$$V^\pi := \frac{1}{\beta}\ln\{\mathbb{E}_\pi[e^{\beta R}]\} = \mathbb{E}_\pi[R] + \frac{\beta}{2}\mathbb{V}_\pi[R] + O(\beta^2) \quad (1)$$

where R is the random reward and β is the risk parameter. To show how this utility function incorporates risk, we apply Taylor's expansion in the second equation, which show clearly the expected reward $\mathbb{E}_\pi[R]$ and the variance of the reward $\mathbb{V}_\pi[R]$ when adopting policy π. Intuitively, the agent is said to be risk-seeking, risk-neutral, or risk-averse when $\beta > 0$, $\beta = 0$, or $\beta < 0$, respectively. The EU criterion has been applied to many real-world applications such as inventory control (Bouakiz and Sobel, 1992) and financial markets (Rásonyi and Sayit, 2022). In some cases, it can be more appropriate and advantageous than other risk criteria (Smith and Chapman, 2021). To model the uncertainty in the environment, we adopt the Markov Decision Processes (MDP) framework, where the agent sequentially observes the state, takes an action, receives a reward, and transits to the next state. Since the reward function and the transition kernel are unknown to the agent, a great challenge is the trade-off between exploration and exploitation, i.e., the agent faces a dilemma between exploring the unknown environment at the risk of gaining poor utility (in order to improve long-term performance), and maximizing the expected utility. Building upon this fundamental feature, several performance metrics are proposed to evaluate the efficiency of learning algorithms, including sample complexity of exploration (Kakade, 2003), average loss (Strehl and Littman, 2005), and Bayesian regret (Osband et al., 2013). In this paper, we follow a line of studies (Bartlett and Tewari, 2009; Jaksch et al., 2010; Osband and Roy, 2016; Azar et al., 2017; Jin et al., 2018a) that aims to minimize the *regret*, that is, the difference between the expected utility brought by following the optimal policy and that obtained using the learning algorithm. Our goal is to design a learning algorithm that achieves the information-theoretic lower bound of the regret, which is optimal in the minimax sense (Lattimore and Szepesvári, 2020).

While provably efficient learning algorithms are largely studied in risk-neutral RL (Kearns and Singh, 2002; Strehl and Littman, 2008; Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018a), it is not until recently that this problem is addressed in risk-sensitive RL under the MDP framework. Fei, Yang, Chen, Wang and Xie (2020) provide the first regret analysis and show that the Risk-Sensitive Q-learning (RSQ) attains a regret bound of $\tilde{O}(e^{|\beta|(H^2+H)}(e^{|\beta|H}-1)\sqrt{H^2SAT}/|\beta|)$, where polylogarithmic factors (of H, S, A, T, β, etc.) are hidden in $\tilde{O}(\cdot)$ notation. They also establish the information-theoretic lower bound in Theorem 3:[1]

$$\text{Regret}(T) \geq \Omega\left(\frac{e^{\frac{|\beta|(H-1)}{2}}-1}{|\beta|}\sqrt{SAT}\right) \qquad (2)$$

Fei, Yang, Chen and Wang (2021a) eliminate the factor $e^{|\beta|(H^2+H)}$ from the previous regret bound by utilizing the exponential Bellman equation and designing novel bonus terms. They then show that RSQ2, a modified version of RSQ, attains the best previously known upper bound $\tilde{O}((e^{|\beta|H}-1)\sqrt{H^2SAT}/|\beta|)$.[2] Although this greatly improves their previous result, there remains an exponential gap of at least $(e^{|\beta|H/2}+1)\sqrt{H^2e^{|\beta|}}$ compared to the information-theoretic lower bound (2), which makes it unsatisfactory, particularly for large $|\beta|$ and H.

The reason why previous studies fail in attaining the information-theoretic lower bound (2) is that their mechanism of regret analysis destroys the structure of the risk-sensitive objective function (1). Specifically, regret analysis in the previous study relies on the convexity of the exponential function, i.e., $x-y \leq (e^{\beta x}-e^{\beta y})/\beta$ when $x \geq y \geq 0$ and $\beta > 0$. This leads to a regret bound that takes a recursive form that makes it impossible to avoid the factor $e^{|\beta|H}$ (see Inequality (13) and further discussion in Section 4). To address this problem, we exploit the structure of the risk-sensitive objective function (1) and introduce a brand new mechanism of regret analysis based on both the concavity of the logarithm, i.e., $\ln x - \ln y \leq (x-y)/y$ when $x \geq y$, and the reference-advantage decomposition technique (Sidford et al., 2018; Zhang et al., 2020). When we apply it to a modified version of UCB-ADVANTAGE (Zhang et al., 2020), we establish the recursive form (18), which avoids the factor $e^{|\beta|H}$, and derive a problem-dependent regret bound, i.e., a regret bound that depends on the structure of MDP without prior knowledge of the MDP from the algorithm. Then, we show that the information-theoretic lower bound (2) can be attained under a mild condition.

In summary, we make the following contributions:

1. We carefully analyze the structure of the risk-sensitive objective function (1) and design a brand new mechanism to analyze the regret, which builds upon the concavity of the logarithm and the reference-advantage decomposition technique. When we apply this mechanism to a modified version of UCB-ADVANTAGE, we show that it avoids the factor $e^{|\beta|H}$ in the regret bound, unlike those utilizing the exponential Bellman equation in the previous studies.

2. In Theorem 1, we derive a problem-dependent regret bound without prior knowledge of the MDP from the algorithm. This bound improves a factor of $\sqrt{H}$ in any arbitrary MDP over the best previously known bound and can be much tighter in MDPs with special structures. Further, we show in Corollary 1.1 that within a rich class of MDPs, this problem-dependent regret bound translates to $\tilde{O}((e^{|\beta|(H-1)/2}-1)\sqrt{SAT}/|\beta|)$, which improves a factor of at least $(e^{|\beta|H/2}+1)\sqrt{H^2e^{|\beta|}}$ over the best previously known bound. This shows that a regret that matches the information-theoretic lower bound up to logarithmic factors can already be achieved within a wide range of problem instances.

3. We establish a novel information-theoretic lower bound of $\Omega(\max_h c^*_{v,h+1}\sqrt{SAT}/|\beta|)$ in Theorem 2, where $\max_h c^*_{v,h+1}$ is a problem-dependent statistic. This lower bound shows that the problem-dependent regret bound attained by the algorithm is optimal in its dependence on $\max_h c^*_{v,h+1}$. When compared to a problem-dependent regret bound established for risk-neutral RL (Zanette and Brunskill, 2019), our results show that the regret bound in the risk-sensitive RL

[1] Note that Fei et al. (2020) only give the proof for the specific case of $S=3$ and $A=2$, where S is the number of states and A is the number of actions. We generalize their result to any S and A. The detailed proof is left to Appendix H.

[2] The bound we present here is in a form slightly different from that the authors establish in their paper. This will be discussed in Section 5.

is not necessarily a monotonic function of the per-step conditional variance of the optimal (exponential) value function.

1.1 Related Works

The problem of risk-sensitive Markov decision processes is first proposed by Howard and Matheson (1972), where value iteration and policy iteration are applied to learning the optimal policy. Following this seminal work, a line of studies has been conducted (Masi and Stettner, 1999; Borkar, 2002; Cavazos-Cadena and Hernández-Hernández, 2011; Osogami, 2012; Bäuerle and Rieder, 2014; Chow et al., 2015; Huang and Haskell, 2021; A. and Fu, 2021). However, these works assume either a known transition kernel or access to a generative model that samples from the transition kernel in $O(1)$ time given any state-action pair. In contrast, we study the setting where the transition kernel is unknown, which poses great challenges to learning and adapts to many real-world scenes. We remark that there is another interesting line of study on risk-sensitive Multi-armed Bandit (MAB) (Maillard, 2013; Zimin et al., 2014; Cassel et al., 2018). However, learning in MDP is fundamentally different from that in MAB due to its longer planning horizon and unknown transition kernel.

In the MDP setting, Fei et al. (2020) provide the first regret analysis of risk-sensitive RL, where two provably efficient model-free algorithms, Risk-Sensitive Value Iteration (RSVI) and Risk-Sensitive Q-learning (RSQ), are proposed, and the information-theoretic lower bound is studied. As an effort to incorporate function approximation techniques, Fei, Yang, and Wang (2021b) study the MDP where each transition kernel admits a linear feature representation. They propose two algorithms and a sublinear regret is attained. Fei et al. (2021a) also exploit the structure of the exponential Bellman equation and design a doubly-decaying bonus. Their modified algorithms, RSQ2 and RSVI2, succeed in further minimizing the regret bound. Later, the gap-dependent regret bound of RSQ2 and RSVI2 is studied by Fei and Xu (2022). However, these works leave an exponential gap between the regret bounds and information-theoretic lower bound (2), which is unsatisfactory when $|\beta|$ and H are large. Recently, risk-sensitive RL has also been studied by Zhang, Yang, and Wang (2021) in the linear-quadratic (LQ) game, who prove that an actor-critic algorithm converges to the optimal policy.

2 PRELIMINARIES

2.1 Episodic MDP

An episodic and finite-horizon MDP (Bertsekas, 2009) is a quintuple $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the fixed length of each episode, $\mathcal{P} = \{P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}_{h \in [H]}$ the transition kernel where $\Delta(\mathcal{S})$ the space of probability simplex on $\mathcal{S}$, and $r = \{r_h : \mathcal{S} \times \mathcal{A} \to [0,1]\}_{h \in [H]}$ the deterministic reward function.[3] We assume that both $\mathcal{P}$ and r are unknown to the agent. The agent interacts with the MDP for K episodes. Without loss of generality, we assume that the initial state s_1 is *fixed*.[4] Let the (deterministic) policy of the kth episode be $\pi^k = \{\pi_h^k : \mathcal{S} \to \mathcal{A}\}_{h \in [H]}$. At timestep h of episode k, the agent observes state s_h^k, executes the action $a_h^k = \pi_h^k(s_h^k)$, obtains a reward $r_h(s_h^k, a_h^k)$, and transits to state s_{h+1}^k with probability $P_h(s_{h+1}^k | s_h^k, a_h^k)$. The episode ends at timestep $H+1$. We denote by $S := |\mathcal{S}|$ the size of the state space, and $A := |\mathcal{A}|$ the size of the action space. We also define $T := KH$ as the total timesteps.

2.2 Risk-sensitive Reinforcement Learning

In risk-sensitive RL with the exponential utility (Fei et al., 2020, 2021a), the value function is defined for all $(h, s) \in [H] \times \mathcal{S}$ and policy π as

$$V_h^\pi(s) := \frac{1}{\beta} \ln \left\{ \mathbb{E}_\pi \left[e^{\beta \cdot \sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})} \middle| s_h = s \right] \right\}$$

where $\beta \neq 0$ is the risk parameter of the exponential utility. Further, we define the Q-function for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ as

$$\begin{aligned} & Q_h^\pi(s,a) \\ :=& \frac{1}{\beta} \ln \left\{ \mathbb{E}_\pi \left[e^{\beta \cdot \sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})} \middle| s_h = s, a_h = a \right] \right\} \end{aligned}$$

For any function f defined on $\mathcal{S}$, define the operator $[P_h f](s,a) := \mathbb{E}_{s' \sim P_h(s'|s,a)} f(s')$. The Bellman equation of the policy π is hence given by

$$\begin{aligned} & Q_h^\pi(s,a) := r_h(s,a) + \frac{1}{\beta} \ln \left\{ [P_h e^{\beta \cdot V_{h+1}^\pi}](s,a) \right\} \qquad (3) \\ & V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)), \; V_{H+1}^\pi(s) = 0. \end{aligned}$$

It can be shown that $V_h^\pi(s), Q_h^\pi(s,a) \in [0, H-h+1]$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Let $\mathbb{V}([P_h e^{\beta \cdot V_{h+1}^\pi}](s,a)) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[(e^{\beta \cdot V_{h+1}^\pi(s')} - [P_h e^{\beta \cdot V_{h+1}^\pi}](s,a))^2]$ denote the per-step conditional variance of the exponential value function (conditioned on (h,s,a)). The maximum per-step conditional coefficient of variation (CV) of the exponential value function (conditioned on (h,s,a)) is defined as

$$c_{v,h+1}^\pi := \max_{s,a} \frac{\sqrt{\mathbb{V}([P_h e^{\beta \cdot V_{h+1}^\pi}](s,a))}}{[P_h e^{\beta \cdot V_{h+1}^\pi}](s,a)} \qquad (4)$$

Note that we have $c_{v,H+1}^\pi = 1$ for any policy π. In addition, we point out that for any policy π and $h \in [H]$, it holds that

[3] We use the notation that $[N] := \{1, 2, ..., N\}$, for any positive integer N.

[4] Note that any H-length episodic MDP with a stochastic initial state is equivalent to an $(H+1)$-length MDP with a dummy initial state s_0.

$c^{\pi}_{v,h+1} \le O(e^{|\beta|(H-h)/2})$ (See Appendix D). Compared to risk-neutral RL, the non-linearity between Q^{π}_h and V^{π}_{h+1} in the Bellman equation (3) poses great challenge to both algorithmic design and regret analysis (Fei et al., 2020). To address this problem, Fei et al. (2021a) establish the exponential Bellman equation

$$e^{\beta\cdot Q^{\pi}_h(s,a)} = e^{\beta\cdot r_h(s,a)}[P_h e^{\beta\cdot V^{\pi}_{h+1}}](s,a) \qquad (5)$$

Under some mild conditions (Bäuerle and Rieder, 2014), there exists an policy π^* that attains the optimal value function, i.e., $V^*_h(s) = \sup_\pi V^{\pi}_h(s)$ for all $(h,s) \in [H]\times\mathcal{S}$. The agent aims to minimize the *regret* within K episodes that is given by

$$\text{Regret}(T) := \sum_{k=1}^{K}(V^*_1(s_1) - V^{\pi^k}_1(s_1)). \qquad (6)$$

3 THE UCB-ADVANTAGE ALGORITHM FOR RISK-SENSITIVE RL

UCB-ADVANTAGE (Zhang et al., 2020) is a model-free RL algorithm that features upper confidence bound (UCB), reference-advantage decomposition, and advantage-based update rule. The idea is to first learn an optimistic estimation of the optimal value function denoted by V^{ref} and use it for later updates. With carefully designed bonus terms and update rules, it is shown that $V^{\text{ref}}(s) - V^*(s)$ can be upper bounded (with high probability) once the state s is visited more than N_0 times (Zhang et al., 2020, Corollary 6). This algorithm matches the information-theoretic lower bound up to logarithm factors for risk-neutral RL. However, to our best knowledge, its potential in risk-sensitive RL has not been studied. Adapting UCB-ADVANTAGE, we present Algorithm 1 for risk-sensitive RL with $\beta > 0$.[5]

Algorithm 1 utilizes a *stage*-based update rule. Quantities $Q_h(s,a)$ and $V_h(s)$ are updated only at the end of stage i, when the state-action pair (s,a) has been visited for l_i times (lines 9-17), where $l_1 = 1$, and $l_i = l_{i-1} + \lfloor(1+1/H)^i\rfloor$ for $i \ge 2$. We also define $l_0 = 0$ for convenience (note that $l_0 \notin \mathcal{L}$). This lazy update scheme ensures low local switching cost and adapts to various real-world settings (Bai et al., 2019). One difference between Algorithm 1 and UCB-ADVANTAGE is that the reference value V^{ref}_h is set twice (lines 18-20). The algorithm keeps track of two types of accumulators, the *global* ones and the *intra-stage* ones. The *global* accumulators are maintained all along the process, which include the total number of visits $N_h(s,a)$ of each state-action pair (s,a) and the fol-

Algorithm 1 UCB-ADVANTAGE FOR RISK-SEEKING RL ($\beta > 0$)

1: **Initialize:** the failure probability p in Lemmas 1 and 2; $\alpha \leftarrow \min\{e^{-\beta H}, e^{-H}\}$; $\alpha' \leftarrow \sqrt{HSA}$; $\iota \leftarrow \ln(2/p)$; risk parameter $\beta > 0$; set all accumulators to 0; $V_h(s) \leftarrow H$, $Q_h(s,a) \leftarrow H$, $V^{\text{ref}}_h(s) \leftarrow H$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$; $V_{H+1} \leftarrow 0, V^{\text{ref}}_{H+1} \leftarrow 0$; $\mathcal{L} := \{l_i \mid l_1 = 1, l_i = l_{i-1} + \lfloor(1+1/H)^i\rfloor, i = 2,3,...\}$
2: **for** episodes $k \leftarrow 1,2,...,K$ **do**
3: Receive s_1
4: **for** $h \leftarrow 1,2,...,H$ **do**
5: Take action $a_h \leftarrow \arg\max_a Q_h(s_h,a)$.
6: Observe the next state s_{h+1}.
7: $n := N_h(s_h,a_h) \xleftarrow{+} 1, \check{n} := \check{N}_h(s_h,a_h) \xleftarrow{+} 1$
8: Update accumulators by rules (7∼11).
9: **if** $n \in \mathcal{L}$ **then**
10: $b^k_h \leftarrow c_1\sqrt{\frac{\sigma^{\text{ref}}/n-(u^{\text{ref}}/n)^2}{n}\iota} + c_2\sqrt{\frac{\check{\sigma}/\check{n}-(\check{\Delta}/\check{n})^2}{\check{n}}\iota} + c_3\cdot e^{\beta H}\left(\frac{\iota}{n}+\frac{\iota}{\check{n}}+\frac{\iota^{\frac{3}{4}}}{n^{\frac{3}{4}}}+\frac{\iota^{\frac{3}{4}}}{\check{n}^{\frac{3}{4}}}\right)$
11: $\bar{b}^k_h \leftarrow 2\sqrt{\frac{e^{2\beta H}}{\check{n}}\iota}$
12: $Q'_h(s_h,a_h) \leftarrow r_h(s_h,a_h) + \frac{1}{\beta}\ln\left(\min\left\{\frac{\check{u}}{\check{n}}+\bar{b}_h, \frac{u^{\text{ref}}}{n}+\frac{\check{\Delta}}{\check{n}}+b_h\right\}\right)$
13: $Q_h(s_h,a_h) \leftarrow \min\{Q'_h(s_h,a_h), Q_h(s_h,a_h)\}$
14: $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$
15: $\check{N}_h(s_h,a_h) \leftarrow 0$
16: $\check{\Delta}_h(s_h,a_h), \check{u}_h(s_h,a_h), \check{\sigma}_h(s_h,a_h) \leftarrow 0$
17: **end if**
18: **if** $\sum_a N_h(s_h,a) \in \{N_0(\alpha), N_0(\alpha')\}$ **then**
19: $V^{\text{ref}}_h(s_h) \leftarrow V_h(s_h)$
20: **end if**
21: **end for**
22: **end for**

[5]We illustrate our core idea with the case of a positive β. When $\beta < 0$, the algorithm and the proofs of theorems need to be slightly modified, as discussed in the Appendix G.

lowing two accumulators:

$$u^{\text{ref}} := u_h^{\text{ref}}(s_h, a_h) \xleftarrow{+} e^{\beta \cdot V_{h+1}^{\text{ref}}(s_{h+1})} \tag{7}$$

$$\sigma^{\text{ref}} := \sigma_h^{\text{ref}}(s_h, a_h) \xleftarrow{+} e^{2\beta \cdot V_{h+1}^{\text{ref}}(s_{h+1})} \tag{8}$$

The *intra-stage* accumulators are maintained only within a stage. They will be reset once the stage completes (lines 15-16). These include the number of visits $\check{N}_h(s,a)$ of each state-action pair (s,a) within the current stage and the following three accumulators:

$$\check{u} := \check{u}_h(s_h, a_h) \xleftarrow{+} e^{\beta \cdot V_{h+1}(s_{h+1})} \tag{9}$$

$$\check{\sigma} := \check{\sigma}_h(s_h, a_h) \xleftarrow{+} (e^{\beta \cdot V_{h+1}(s_{h+1})} - e^{\beta \cdot V_{h+1}^{\text{ref}}(s_{h+1})})^2 \tag{10}$$

$$\check{\Delta} := \check{\Delta}_h(s_h, a_h) \xleftarrow{+} e^{\beta \cdot V_{h+1}(s_{h+1})} - e^{\beta \cdot V_{h+1}^{\text{ref}}(s_{h+1})} \tag{11}$$

We add a superscript k to these accumulators and other quantities in the algorithm to indicate their values at timestep h of the kth episode, i.e., Q_h^k, V_h^k, $V_h^{\text{ref},k}$, N_h^k, $u_h^{\text{ref},k}$, $\check{u}_h^k$, $\sigma_h^{\text{ref},k}$, $\check{\sigma}_h^k$, $\check{\Delta}_h^k$, b_h^k, and $\bar{b}_h^k$. Let n_h^k be the number of visits of (s_h^k, a_h^k) prior to the current stage, i.e., $n_h^k = l_j$, where the subscript j is a non-negative integer that satisfies $l_j < N_h^k \le l_{j+1}$. Note that (s_h^k, a_h^k) can be visited at most once in each episode. Among these n_h^k visits, we denote by $l_{h,i}^k$ the (index of) episode of the ith visit. Similarly, let $\check{n}_h^k$ be the number of visits of (s_h^k, a_h^k) during the last stage that $Q_h(s_h^k, a_h^k)$ is updated, i.e., $\check{n}_h^k = l_j - l_{j-1}$ if $j \ge 1$ (there is no "last stage" when $j = 0$). Among these $\check{n}_h^k$ visits, we denote by $\check{l}_{h,i}^k$ the episode of the ith visit. Hence, we can rewrite the notations in Equations (7∼11) as follows.

$$u_h^{\text{ref},k} = \sum_{i=1}^{n_h^k} e^{\beta \cdot V_{h+1}^{\text{ref}, l_{h,i}^k}(s_{h+1}^{l_{h,i}^k})}, \; \sigma_h^{\text{ref},k} = \sum_{i=1}^{n_h^k} e^{2\beta \cdot V_{h+1}^{\text{ref}, l_{h,i}^k}(s_{h+1}^{l_{h,i}^k})}$$

$$\check{u}_h^k = \sum_{i=1}^{\check{n}_h^k} e^{\beta \cdot V_{h+1}^{\check{l}_{h,i}^k}(s_{h+1}^{\check{l}_{h,i}^k})}$$

$$\check{\sigma}_h^k = \sum_{i=1}^{\check{n}_h^k} (e^{\beta \cdot V_{h+1}^{\check{l}_{h,i}^k}(s_{h+1}^{\check{l}_{h,i}^k})} - e^{\beta \cdot V_{h+1}^{\text{ref}, \check{l}_{h,i}^k}(s_{h+1}^{\check{l}_{h,i}^k})})^2$$

$$\check{\Delta}_h^k = \sum_{i=1}^{\check{n}_h^k} (e^{\beta \cdot V_{h+1}^{\check{l}_{h,i}^k}(s_{h+1}^{\check{l}_{h,i}^k})} - e^{\beta \cdot V_{h+1}^{\text{ref}, \check{l}_{h,i}^k}(s_{h+1}^{\check{l}_{h,i}^k})})$$

Let $\mathbb{I}[\cdot]$ denote the indicator function. We derive some useful properties of Algorithm 1, which facilitate the analysis of its regret bound. To start with, the following lemma states that with high probability, $Q_h^k(s,a)$ output from the algorithm is an optimistic estimation of the optimal Q-function $Q_h^*(s,a)$. (The detailed proof can be found in Appendix A.)

Lemma 1 (Optimism)**.** *Let $p \in (0,1)$ denote the failure probability. For any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, with probability at least $1 - 2(e^{2\beta H} T^3 + 3)p$, it holds that $Q_h^*(s,a) \le Q_h^{k+1}(s,a) \le Q_h^k(s,a)$. Therefore, we have that $V_h^*(s) = \max_a Q_h^*(s,a) \le Q_h^k(s, \arg\max_a Q_h^*(s,a)) \le V_h^k(s)$.*

Next, the following two lemmas show that when a state s is "sufficiently" visited, the differences between two quantities output from the algorithm ($V_h^k(s)$ and $V_h^{\text{ref},k}(s)$) and the optimal value function $V_h^*(s)$ can be bounded. (The detailed proofs can be found in Appendix B.)

Lemma 2 (Bounded estimation error)**.** *Conditioned on the successful events of Lemma 1, for any $\gamma \in (0, e^{\beta H}]$, with probability $(1 - Tp)$ it holds that $\sum_{k=1}^K \mathbb{I}[e^{\beta \cdot V_h^k(s)} - e^{\beta \cdot V_h^*(s)} \ge \gamma] \le O\left(e^{4\beta H} H^3 SA/\gamma^2\right)$.*

Lemma 3 (Good reference values)**.** *Conditioned on the successful events of Lemmas 1 and 2, it holds that $e^{\beta \cdot V_h^*(s)} \le e^{\beta \cdot V_h^{\text{ref},k}(s)} \le e^{\beta \cdot V_h^*(s)} + \gamma$ if $n_h^k(s) \ge N_0(\gamma) := c_4 e^{4\beta H} H^3 SA\iota/\gamma^2$, where c_4 is a sufficiently large constant for analysis. Therefore, we have that $\beta(V_h^k(s) - V_h^*(s)) \le \beta(V_h^{\text{ref},k}(s) - V_h^*(s)) \le \gamma$ when state s is visited by more than $N_0(\gamma)$ times.*

Note that Lemmas 1, 2, and 3 are generalizations of Proposition 4 and Lemmas 5 and 6 in the work of Zhang et al. (2020) for risk-seeking RL, respectively.

4 REGRET ANALYSIS

In this section, we shall adopt a widely used method for computing the regret bound of algorithms based on UCB (Azar et al., 2017; Jin et al., 2018b). By Lemma 1, we first note that Regret$(T) \le \sum_{k=1}^K (V_1^k - V_1^{\pi^k})$, where V_h^k is the value at timestep h of the kth episode, and $V_h^{\pi^k}$ is the value function of the policy used at episode k (see Equation (3)). Building upon the convexity of the exponential function, i.e., $x - y \le (e^{\beta x} - e^{\beta y})/\beta$ for $x \ge y \ge 0$, and note that $V_1^k \ge V_1^* \ge V_1^{\pi^k}$ for any $k \in [K]$ from Lemma 1, Fei et al. (2021a) first derive $V_1^k - V_1^{\pi^k} \le (e^{\beta \cdot V_1^k} - e^{\beta \cdot V_1^{\pi^k}})/\beta$. Then, they utilize the exponential Bellman equation (5) to establish the following recursive form (with constant terms omitted),

$$\begin{aligned} &\frac{1}{\beta} \sum_{k=1}^K (e^{\beta \cdot V_h^k} - e^{\beta \cdot V_h^{\pi^k}}) \\ \le &\frac{1}{\beta} \sum_{k=1}^K \left(e^{\beta}(1 + \frac{1}{H})(e^{\beta \cdot V_{h+1}^k} - e^{\beta \cdot V_{h+1}^{\pi^k}}) + B_h^k + M_h^k \right) \end{aligned} \tag{12}$$

Iterating this recursive form over h, the regret is bounded by only the bonus terms B_h^k and the martingale terms M_h^k, that is,

$$\text{Regret}(T) \le \frac{1}{\beta} \sum_{k=1}^K (e^{\beta \cdot V_1^k} - e^{\beta \cdot V_1^{\pi^k}})$$

$$
\begin{aligned}
&\leq \frac{1}{\beta}\sum_{k=1}^{K}\left(e^{\beta}(1+\frac{1}{H})(e^{\beta\cdot V_2^k}-e^{\beta\cdot V_2^{\pi^k}})+B_1^k+M_1^k\right)\\
&\leq \frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}e^{\beta(h-1)}(1+\frac{1}{H})^{h-1}\left(B_h^k+M_h^k\right)\\
&\leq \tilde{O}\left(\frac{e^{\beta H}-1}{\beta}\sqrt{H^2SAT}\right) \qquad (13)
\end{aligned}
$$

Unfortunately, the recursive form (12) is undesirable because an extra factor e^{β} is introduced after each rollout of h, which makes it impossible to attain the information-theoretic lower bound (2). To address this problem, in this paper we propose a new mechanism of regret analysis. When we apply it to Algorithm 1, we establish the following recursive form

$$
\begin{aligned}
&\sum_{k=1}^{K}(V_h^k-V_h^{\pi^k})\\
\leq&\frac{1}{\beta}\sum_{k=1}^{K}\lambda_{h+1}(1+\frac{1}{H})^2(V_{h+1}^k-V_{h+1}^{\pi^k})\\
&+\frac{1}{\beta}\sum_{k=1}^{K}\frac{b_h^k+m_h^k}{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}
\end{aligned}
$$

where b_h^k are the bonus terms, m_h^k are the martingale terms, and λ_{h+1} is the problem-dependent statistic (to be defined in Equation (16)), which avoids the factor of $e^{\beta H}$ in the regret bound. Therefore, the rest of this section will be devoted to establishing this recursive form. Define $\delta_h^k := V_h^k(s_h^k)-V_h^*(s_h^k)$ and $\varrho_h^k := e^{\beta\cdot V_h^k(s_h^k)}-e^{\beta\cdot V_h^{\text{ref},k}(s_h^k)}$. Let $V_h^{\text{REF}}(s)$ be the reference value of any $(s,h)\in\mathcal{S}\times[H]$ when the Kth episode is finished. Recall that $[P_hf](s,a) := \mathbb{E}_{s'\sim P_h(s'|s,a)}f(s')$, for any function f defined on $\mathcal{S}$. We denote by $[\widehat{P}_hf](s_h^k,a_h^k) := f(s_{h+1}^k)$ the empirical counterpart of $[P_hf]$. For convenience, we rewrite $l_{h,i}^k$ and $\check{l}_{h,i}^k$ as l_i and $\check{l}_i$, respectively, when the context is clear. Following the update rules (7∼11), we obtain,

$$
\begin{aligned}
&\zeta_h^k := V_h^k(s_h^k)-V_h^{\pi^k}(s_h^k)\\
\leq&Q_h^k(s_h^k,a_h^k)-Q_h^{\pi^k}(s_h^k,a_h^k)\\
\leq&\mathbb{I}[n_h^k=0]H+\frac{1}{\beta}\ln\left(\frac{u_h^{\text{ref},k}}{n_h^k}+\frac{\check{\Delta}_h^k}{\check{n}_h^k}+b_h^k\right)\\
&-\frac{1}{\beta}\ln\left([P_he^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)\right)\\
=&\mathbb{I}[n_h^k=0]H-\frac{1}{\beta}\ln\left([P_he^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)\right)\\
&+\frac{1}{\beta}\ln\left(\frac{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\right)\\
&+\frac{1}{\beta}\ln\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}e^{\beta\cdot V_{h+1}^{\text{ref},l_i}(s_{h+1}^{l_i})}+\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\varrho_{h+1}^{\check{l}_i}+b_h^k\right)
\end{aligned}
$$

$$
\begin{aligned}
\leq&\mathbb{I}[n_h^k=0]H-\frac{1}{\beta}\ln\left([P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)\right)\\
&+\frac{1}{\beta}\ln\left(\left[P_h(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}e^{\beta\cdot V_{h+1}^{\text{ref},l_i}}+\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\varrho_{h+1}^{\check{l}_i})\right](s_h^k,a_h^k)\right.\\
&\left.+2b_h^k\right)+\lambda_{h+1}(\zeta_{h+1}^k-\delta_{h+1}^k)+\frac{1}{\beta}\phi_{h+1}^k
\end{aligned} \qquad (14)
$$

$$
\begin{aligned}
\leq&\mathbb{I}[n_h^k=0]H+\lambda_{h+1}(\zeta_{h+1}^k-\delta_{h+1}^k)+\frac{1}{\beta}\vartheta_{h+1}^k\\
&+\frac{\lambda_{h+1}}{\beta}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(e^{\beta\delta_{h+1}^{\check{l}_i}}-1)
\end{aligned} \qquad (15)
$$

where

$$
\lambda_{h+1}=\max_{\pi,s,a,s':P_h(s'|s,a)>0}\frac{e^{\beta\cdot V_{h+1}^{\pi}(s')}}{[P_he^{\beta\cdot V_{h+1}^{\pi}}](s,a)} \qquad (16)
$$

$$
\psi_{h+1}^k := \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}[P_h(e^{\beta\cdot V_{h+1}^{\text{ref},l_i}}-e^{\beta\cdot V_{h+1}^{\text{REF}}})](s_h^k,a_h^k)
$$

$$
\xi_{h+1}^k := \frac{\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}[(P_h-\widehat{P}_h)(e^{\beta\cdot V_{h+1}^{\check{l}_i}}-e^{\beta\cdot V_{h+1}^*})](s_h^k,a_h^k)}{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}
$$

$$
\phi_{h+1}^k := \ln\left(\frac{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}{e^{\beta\lambda_{h+1}\cdot V_{h+1}^*(s_{h+1}^k)}}\cdot\frac{e^{\beta\lambda_{h+1}\cdot V_{h+1}^{\pi^k}(s_{h+1}^k)}}{[P_he^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)}\right)
$$

$$
\vartheta_{h+1}^k := \frac{2b_h^k}{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}+\psi_{h+1}^k+\xi_{h+1}^k+\phi_{h+1}^k
$$

Here, the first line of Inequality (14) is implied by the successful events in the proof of Lemma 1 (See Appendix A). Notice that $\psi_{h+1}^k\geq 0$ and $[P_he^{\beta\cdot V_{h+1}^*}](s,a)\geq 1$ for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$. A key step in our mechanism is to derive Inequality (15), where we utilize the concavity of the logarithm, i.e., $\ln x-\ln y\leq(x-y)/y$ when $x\geq y$, and the definition of λ_{h+1}. To further bound $\sum_{k=1}^{K}\zeta_h^k$, another key step is to use the reference-advantage technique to handle the last term of Inequality (15). Summing this term over k and note that λ_{h+1} is invariant to k, we derive

$$
\begin{aligned}
&\frac{\lambda_{h+1}}{\beta}\sum_{k=1}^{K}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(e^{\beta\delta_{h+1}^{\check{l}_i}}-1)\\
\leq&(1+\frac{1}{H})\lambda_{h+1}\sum_{j=1}^{K}\frac{e^{\beta\delta_{h+1}^j}-1}{\beta\delta_{h+1}^j}\delta_{h+1}^j\\
\leq&(1+\frac{1}{H})\lambda_{h+1}\sum_{k=1}^{K}\frac{e^{\alpha}-1}{\alpha}\delta_{h+1}^k\\
&+(1+\frac{1}{H})\lambda_{h+1}\sum_{k=1}^{K}\mathbb{I}[n_h^k<N_0(\alpha)]\frac{e^{\beta H}-1}{\beta}
\end{aligned} \qquad (17)
$$

where $\alpha=\min\{e^{-\beta H},e^{-H}\}$ is an input of Algorithm 1.

Here, the first inequality is implied by the stage-based update of the algorithm (See Equation (15) and Inequality (16) in (Zhang et al., 2020)). The second inequality utilizes the reference-advantage technique, i.e., the successful events of Lemma 3, and the fact that $(e^x - 1)/x$ is non-decreasing when $x \geq 0$. Next, summing the both sides of Inequality (15) over k and together with Inequality (17), we derive the following recursive form,

$$\sum_{k=1}^{K} \zeta_h^k \leq HSA + \frac{1}{\beta}\sum_{k=1}^{K} \vartheta_{h+1} + (1+\frac{1}{H})^2 \lambda_{h+1} \sum_{k=1}^{K} \zeta_{h+1}^k + (1+\frac{1}{H})\frac{e^{\beta H}-1}{\beta}\lambda_{h+1} S \cdot N_0(\alpha) \tag{18}$$

where we utilize the fact that $(e^\alpha - 1)/\alpha < 1 + 1/H$,[6] $\sum_{k=1}^K \mathbb{I}[n_h^k = 0] \leq SA$, $\sum_{k=1}^K \mathbb{I}[n_h^k < N_0(\alpha)] \leq S \cdot N_0(\alpha)$, and $\zeta_h^k \geq \delta_h^k$ for any $(h,k) \in [H]\times[K]$. Define $\Lambda_h = \lambda_{h+1}\Lambda_{h-1}$ for $h \in [H-1]$ and $\Lambda_0 = 1$. Using the similar trick in Inequality (13), we iterate the recursive form (18) over h and derive

$$\mathrm{Regret}(T) \leq \frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{2(h-1)}\Lambda_{h-1}\vartheta_{h+1}^k + C \tag{19}$$

where $C := \sum_{h=1}^H (1+1/H)^{2(h-1)}\Lambda_{h-1}(\lambda_{h+1}(1+1/H)S\cdot N_0(\alpha)(e^{\beta H}-1)/\beta + HSA)$.

5 MAIN RESULTS

In Section 4, we analyze the regret for risk-seeking RL ($\beta > 0$). The counterpart case of risk-averse RL ($\beta < 0$) is similar, and is presented in the Appendix G. We are now ready to state the main results. We first present a problem-dependent regret bound that depends on the structure of the MDP without prior knowledge of the MDP from the algorithm, as follows. (The detailed proof can be found in Appendix C.)

Theorem 1 (Problem-dependent regret bound)**.** *For any $p \in (0,1)$, with probability at least $1-p$ and when T is sufficiently large, the regret of Algorithm 1 is bounded by the minimum between*

$$\tilde{O}\left(\frac{e^{|\beta|H}-1}{|\beta|}\sqrt{HSAT}\right) \tag{20}$$

and

$$\tilde{O}\left(\frac{1}{|\beta|}\max_{h\in[H]}\left\{\Lambda_{h-1}\cdot c_{v,h+1}^*\right\}\sqrt{\max\{SA,H\}HT}\right) \tag{21}$$

where $c_{v,h+1}^* := c_{v,h+1}^{\pi^*}$ is the maximum per-step conditional coefficient of variation (CV) defined in Equation (4) of the exponential optimal value function, λ_{h+1} is given by Equation (16), and $\Lambda_h = \lambda_{h+1}\Lambda_{h-1}$ for $h \in [H-1]$ where $\Lambda_0 = 1$.

The first term (20) shows that our algorithm improves over the best previously known bound $\tilde{O}((e^{|\beta|H} - 1)\sqrt{H^2SAT}/|\beta|)$ in any arbitrary MDP by a factor of $\sqrt{H}$.[7] As $\beta \to 0$, it translates to $\tilde{O}(\sqrt{H^3SAT})$ and recovers the regret bound of Q-learning with UCB-Bernstein for risk-neutral RL (Jin et al., 2018a). While the first term holds in the worst case and is invariant to the structure of the MDP, the second term (21) is problem-dependent. It implies that the regret bound can be significantly improved (tightened) in MDPs with special structure, i.e., small $c_{v,h+1}^*$ and λ_{h+1}. Next, we show in Corollary 1.1 that within a class of MDPs, our algorithm improves a factor of at least $(e^{|\beta|H/2}+1)\sqrt{H^2e^{|\beta|}}$ over the best previously known bound (Fei et al., 2021a). (The detailed proof can be found in Appendix D.)

Corollary 1.1. *When $SA \geq H$ and $\max_h \lambda_{h+1} \leq H^{-\frac{1}{2}}e^{\frac{|\beta|}{2}}$, the regret bound in Theorem 1 translates to*

$$\tilde{O}\left(\frac{e^{\frac{|\beta|(H-1)}{2}}-1}{|\beta|}\sqrt{SAT}\right)$$

Corollary 1.1 states that, given a particular β, our algorithm attains a regret bound of $\tilde{O}((e^{|\beta|(H-1)/2} - 1)\sqrt{SAT}/|\beta|)$ within a class of problems. Particularly, this class of problems includes a subset of MDPs where it holds that $\min_{s,a,s':P_h(s'|s,a)>0} P_h(s'|s,a) \geq \sqrt{H}e^{-|\beta|/2}$. (In this case, we have that $\max \lambda_{h+1} \leq (\min_{s,a,s':P_h(s'|s,a)>0} P_h(s'|s,a))^{-1} \leq H^{-1/2}e^{|\beta|/2}$.) Note that the RHS of the inequality $e^{-|\beta|/2}$ is close to zero for relatively large $|\beta|$. This class of problems contains a wide range of MDPs. Therefore, Corollary 1.1 shows that a regret that matches the information-theoretic lower bound (2) up to logarithmic factors can already be achieved within a wide range of MDPs.

To further interpret the dependence in the regret bound (21) on the maximum per-step conditional covariance $c_{v,h+1}^*$ of the exponential value function, we establish a novel problem-dependent lower bound as follows, which shows that such a dependence is unavoidable in the worst case. (The detailed proof can be found in Appendix E.)

[6] Recall that $\alpha = \min\{e^{-\beta H}, e^{-H}\} \leq e^{-H}$. Since we have $\ln(\frac{1}{\alpha}) < \frac{1}{\alpha}$, we obtain that $\frac{e^\alpha-\alpha-1}{\alpha}\ln(\frac{1}{\alpha}) < \frac{e^\alpha-\alpha-1}{\alpha^2} < e-2 < 1$. Dividing both sides by $\ln(\frac{1}{\alpha})$ and rearranging the terms, we derive that $\frac{e^\alpha-1}{\alpha} < 1 + 1/\ln(\frac{1}{\alpha}) \leq 1+\frac{1}{H}$.

[7] Note that the original regret bound established in (Fei et al., 2021a, Theorem 2) is $\tilde{O}((e^{|\beta|H}-1)\sqrt{H^3SAK}/|\beta|H)$. However, we find that there is a typo in its proof (Appendix. C), where an extra $\sum_{h\in[H]}$ is included in the LHS of the first line in Inequality (38). Hence, adopting their definition of δ_h^k, we have that $\sum_{k=1}^K \delta_1^k \leq \tilde{O}((e^{|\beta|H}-1)\sqrt{H^2SAT})$ after iterating Inequality (37) over h. Therefore, the established regret bound should be $\tilde{O}((e^{|\beta|H}-1)\sqrt{H^2SAT}/|\beta|)$ instead.

Theorem 2 (Problem-dependent information-theoretic lower bound)**.** *For any $t \in [0, (H-1)/2]$ and $c_v^t := e^{|\beta|t}$, define the class of problems*

$$\mathcal{M}(c_v^t) := \{M : \textit{Exists deterministic } \pi^* \textit{ of } M \textit{ such that } \max_h c_{v,h+1}^{\pi^*} = O(c_v^t)\} \quad (22)$$

Then, for sufficiently large K, there exists an absolute constant c_0 and a problem instance $M \in \mathcal{M}(c_v^*)$ such that for any online algorithm, it holds that

$$\text{Regret}(T, \mathcal{M}(t)) \geq \Omega\left(\frac{c_v^t}{|\beta|}\sqrt{SAT}\right) \quad (23)$$

When $t = (H-1)/2$ and $|\beta|(H-1) \geq \ln 4$, the problem-dependent information-theoretic lower bound (23) translates to

$$\Omega\left(\frac{e^{\frac{|\beta|(H-1)}{2}} - 1}{|\beta|}\sqrt{SAT}\right)$$

Theorem 2 shows that for the class of MDPs where $\max_h c_{v,h+1}^*$ has the order $O(e^{|\beta|t})$, the worst-case regret is at least $\Omega(e^{|\beta|t}\sqrt{SAT}/|\beta|)$. Compared to the problem-invariant lower bound (2) proposed by Fei et al. (2020), our problem-dependent lower bound (23) is tighter since it always holds that $\max_h c_{v,h+1}^* \leq O(e^{|\beta|(H-1)/2})$ in any MDP. In the worst case of $t = (H-1)/2$, our problem-dependent regret bound recovers the information-theoretic lower bound (2). Further,it shows that the regret bound (21) is unimprovable in its dependence on $\max_h c_{v,h+1}^*$ in the worst case. In risk-neutral RL, the problem-dependent regret bound established by Zanette and Brunskill indicates that the regret bound is a monotonic function of the (maximum) per-step conditional variance of the optimal value function when the reward function is deterministic (Zanette and Brunskill, 2019). One may wonder if this is still the case when the exponential value function is used in risk-sensitive RL. By the definition of $c_{v,h+1}^* := \max_{s,a}([P_h e^{\beta \cdot V_{h+1}^*}](s,a))^{-1}(\mathbb{V}([P_h e^{\beta \cdot V_{h+1}^*}](s,a)))^{1/2}$, the regret bound in an MDP is not necessarily a monotonic function of the per-step conditional variance $\mathbb{V}([P_h e^{\beta \cdot V_{h+1}^*}](s,a))$ of the exponential optimal value function, as it also depends on the expected exponential optimal value function $[P_h e^{\beta \cdot V_{h+1}^*}](s,a)$. In contrast, our algorithm may attain a higher regret bound in an MDP with a smaller $\mathbb{V}([P_h e^{\beta \cdot V_{h+1}^*}](s,a))$, provided that there is an even smaller $[P_h e^{\beta \cdot V_{h+1}^*}](s,a)$. However, we are unaware of whether the dependence on λ_{h+1} in the regret bound is necessary and we leave it as future work.

6 CONCLUSIONS

In this paper, we study the regret bound of risk-sensitive RL with the exponential utility function in an episodic MDP setting. We introduce a brand new mechanism and use it to analyze the regret of a modified version of UCB-ADVANTAGE (Zhang et al., 2020), which avoids the factor $e^{|\beta|H}$ in the regret bound, unlike those utilizing the exponential Bellman equation in the previous studies. We derive a problem-dependent regret bound without prior knowledge of the MDP from the algorithm. This bound improves a factor of $\sqrt{H}$ over the best previously known bound in any arbitrary MDP. In MDPs with special structure, this bound can be even tighter. Further, we show that this problem-dependent regret bound translates to $\tilde{O}((e^{|\beta|(H-1)/2} - 1)\sqrt{SAT}/|\beta|)$ within a rich class of MDPs, which improves the best previously known bound by at least a factor of $(e^{|\beta|H/2} + 1)\sqrt{H^2 e^{|\beta|}}$. This shows that a regret bound that matches the information-theoretic lower bound up to logarithmic factors can be attained within a wide range of problem instances. Further, we establish a novel information-theoretic lower bound of $\Omega(\max_h c_{v,h+1}^*\sqrt{SAT}/|\beta|)$, where where $\max_h c_{v,h+1}^*$ is a problem-dependent statistic. It shows that the regret bound attained by the algorithm is optimal in its dependence on $\max_h c_{v,h+1}^*$. When compared to the problem-dependent regret bound established by Zanette and Brunskill (2019) for risk-neutral RL, our results show that the regret bound in the risk-sensitive RL is not necessarily a monotonic function of the per-step conditional variance of the optimal (exponential) value function.

Acknowledgements

The work presented in this paper is partially supported by a research grant from the Research Grants Council, Hong Kong, China (RGC Ref. No. CUHK 14206820). The authors are grateful to the anonymous reviewers for their helpful comments.

References

A., P. L. and Fu, M. (2021). Risk-sensitive reinforcement learning.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient Q-learning with low switching cost. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bartlett, P. L. and Tewari, A. (2009). Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the*

Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, page 35–42, Arlington, Virginia, USA. AUAI Press.

Bernhard, J., Pollok, S., and Knoll, A. (2019). Addressing inherent uncertainty: Risk-sensitive behavior generation for automated driving using distributional reinforcement learning. *2019 IEEE Intelligent Vehicles Symposium (IV)*.

Bertsekas, D. P. (2009). *Neuro-Dynamic Programming*, pages 2555–2560. Springer US, Boston, MA.

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

Bouakiz, M. and Sobel, M. J. (1992). Inventory control with an exponential utility criterion. *Operations Research*, 40(3):603–608.

Bäuerle, N. and Rieder, U. (2014). More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120.

Cassel, A., Mannor, S., and Zeevi, A. (2018). A general approach to multi-armed bandits under risk criteria. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1295–1306. PMLR.

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

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2015). Risk-sensitive and robust decision-making: a cvar optimization approach. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Fei, Y. and Xu, R. (2022). Cascaded gaps: Towards logarithmic regret for risk-sensitive reinforcement learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6392–6417. PMLR.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. (2021a). Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. (2020). Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22384–22395. Curran Associates, Inc.

Fei, Y., Yang, Z., and Wang, Z. (2021b). Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3198–3207. PMLR.

Howard, R. A. and Matheson, J. E. (1972). Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369.

Huang, W. and Haskell, W. B. (2021). Stochastic approximation for risk-aware Markov decision processes. *IEEE Transactions on Automatic Control*, 66(3):1314–1320.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018a). Is Q-learning provably efficient? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018b). Is Q-learning provably efficient?

Kakade, S. M. (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232.

Kuroda, K. and Nagai, H. (2002). Risk-sensitive portfolio optimization on infinite time horizon. *Stochastics and Stochastic Reports*, 73(3-4):309–331.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Maillard, O.-A. (2013). Robust risk-averse stochastic multi-armed bandits. In Jain, S., Munos, R., Stephan, F., and Zeugmann, T., editors, *Algorithmic Learning Theory*, pages 218–233, Berlin, Heidelberg. Springer Berlin Heidelberg.

Masi, G. B. D. and Stettner, L. (1999). Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM J. Control Optim.*, 38(1):61–78.

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

Osband, I. and Roy, B. V. (2016). On lower bounds for regret in reinforcement learning.

Osband, I., Russo, D., and Van Roy, B. (2013). (More) Efficient reinforcement learning via posterior sampling. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Osogami, T. (2012). Robustness and risk-sensitivity in Markov decision processes. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Rásonyi, M. and Sayit, H. (2022). Exponential utility maximization in small/large financial markets.

Sani, A., Lazaric, A., and Munos, R. (2012). Risk-aversion in multi-armed bandits. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018). Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, page 770–787, USA. Society for Industrial and Applied Mathematics.

Smith, K. M. and Chapman, M. P. (2021). On exponential utility and conditional value-at-risk as risk-averse performance criteria.

Strehl, A. L. and Littman, M. L. (2005). A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 856–863, New York, NY, USA. Association for Computing Machinery.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331. Learning Theory 2005.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR.

Zhang, Y., Yang, Z., and Wang, Z. (2021). Provably efficient actor-critic for risk-sensitive and robust adversarial rl: A linear-quadratic case. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2764–2772. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15198–15207. Curran Associates, Inc.

Zimin, A., Ibsen-Jensen, R., and Chatterjee, K. (2014). Generalized risk-aversion in stochastic multi-armed bandits.

A PROOF OF LEMMA 1

Proof. We prove by backward induction. Suppose that $Q_h^k(s,a) \geq Q_h^*(s,a)$ for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ at the k episode. If no update is conducted at episode $k+1$, then we have $e^{\beta \cdot Q_h^{k+1}(s,a)} = e^{\beta \cdot Q_h^k(s,a)} \geq e^{\beta \cdot Q_h^*(s,a)}$. Otherwise, we have

$$e^{\beta \cdot Q_h^{k+1}(s,a)} = \min \left\{ e^{\beta r_h(s,a)} \left(\frac{\check{u}}{\check{n}} + \bar{b}_h^k \right), e^{\beta r_h(s,a)} \left(\frac{u^{\text{ref}}}{n} + \frac{\check{\Delta}}{\check{n}} + b_h^k \right), e^{\beta \cdot Q_h^k(s,a)} \right\} \tag{24}$$

In the rest of the proof, we show that with high probability, the first two terms in the RHS of Equation (24) is no less than $e^{\beta \cdot Q_h^*(s,a)}$ (since this holds for the last term by assumption). Recall that p is the failure probability defined in Lemma 1 and $\iota = \log(\frac{2}{p})$. For the first case, by Azuma-Hoeffding's inequality, with probability at least $1-p$, it holds that

$$\begin{aligned}
e^{\beta \cdot \left(Q_h^{k+1}(s,a) - r_h(s,a)\right)} =& \frac{\check{u}}{\check{n}} + \bar{b}_h^k = \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} e^{\beta \cdot V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i})} + \bar{b}_h^k \\
\geq& \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} e^{\beta \cdot V_{h+1}^*(s_{h+1}^{\check{l}_i})} + 2\sqrt{\frac{e^{2\beta H}}{\check{n}} \iota} \\
\geq& \left[P_h e^{\beta \cdot V_{h+1}^*} \right](s,a) = e^{\beta \cdot (Q_h^*(s,a) - r_h(s,a))}
\end{aligned} \tag{25}$$

For the second case, we have

$$\begin{aligned}
& e^{\beta \cdot \left(Q_h^{k+1}(s,a) - r_h(s,a)\right)} = \frac{u^{\text{ref}}}{n} + \frac{\check{\Delta}}{\check{n}} + b_h^k \\
=& \left[P_h \left(\frac{1}{n} \sum_{i=1}^{n} e^{\beta \cdot V_{h+1}^{\text{ref}, l_i}} + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(e^{\beta \cdot V_{h+1}^{\check{l}_i}} - e^{\beta \cdot V_{h+1}^{\text{ref}, \check{l}_i}} \right) \right) \right](s,a) + \chi_1 + \chi_2 + b_h^k \\
\geq& \left[P_h \left(\frac{1}{\check{n}} \sum_{i=1}^{\check{n}} e^{\beta \cdot V_{h+1}^{\check{l}_i}} \right) \right](s,a) + \chi_1 + \chi_2 + b_h^k \\
\geq& \left[P_h e^{\beta \cdot V_{h+1}^*} \right](s,a) + \chi_1 + \chi_2 + b_h^k \\
=& e^{\beta \cdot (Q_h^*(s,a) - r_h(s,a))} + \chi_1 + \chi_2 + b_h^k
\end{aligned} \tag{26}$$

where

$$\chi_1 := \frac{1}{n} \sum_{i=1}^{n} \left[\left(\widehat{P}_h - P_h \right) e^{\beta \cdot V_{h+1}^{\text{ref}, l_i}} \right](s,a)$$

$$\chi_2 := \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left[\left(\widehat{P}_h - P_h \right) \left(e^{\beta \cdot V_{h+1}^{\check{l}_i}} - e^{\beta \cdot V_{h+1}^{\text{ref}, \check{l}_i}} \right) \right](s,a)$$

Next, it suffices to show that $b_h > |\chi_1| + |\chi_2|$. By (Zhang et al., 2020, Lemma 10) with $c = e^{\beta H}$ and $\epsilon = \frac{1}{T^2}$, with probability at least $1 - 2(e^{2\beta H} T^3 + 1)p$, it holds that

$$|\chi_1| \leq 2\sqrt{\frac{\iota \sum_{i=1}^{n} \mathbb{V}\left(\left[P_h e^{\beta \cdot V_{h+1}^{\text{ref}, l_i}} \right](s,a) \right)}{n^2}} + 2\frac{\sqrt{\iota}}{Tn} + \frac{2e^{\beta H} \iota}{n} \tag{27}$$

$$|\chi_2| \leq 2\sqrt{\frac{\iota \sum_{i=1}^{\check{n}} \mathbb{V}\left(\left[P_h \left(e^{\beta \cdot V_{h+1}^{\check{l}}} - e^{\beta \cdot V_{h+1}^{\text{ref}, \check{l}_i}} \right) \right](s,a) \right)}{\check{n}^2}} + 2\frac{\sqrt{\iota}}{T\check{n}} + \frac{2e^{\beta H} \iota}{\check{n}} \tag{28}$$

Further, following the exact analysis in the proof of (Zhang et al., 2020, Lemma 12), with probability at least $1-2p$, it holds that

$$\sum_{i=1}^{n} \mathbb{V}\left(\left[P_h e^{\beta \cdot V_{h+1}^{\text{ref}, l_i}} \right] \right) \leq n\nu^{\text{ref}} + 3e^{2\beta H} \sqrt{n\iota} \tag{29}$$

where

$$\nu^{\text{ref}} := \frac{\sigma^{\text{ref}}}{n} - \left(\frac{u^{\text{ref}}}{n}\right)^2 \tag{30}$$

Combining Inequality (27) and Inequality (29), we have

$$|\chi_1| \leq 2\sqrt{\frac{\nu^{\text{ref}}\iota}{n}} + \frac{5e^{\beta H}\iota^{\frac{3}{4}}}{n^{\frac{3}{4}}} + \frac{2\sqrt{\iota}}{Tn} + \frac{2e^{\beta H}\iota}{n} \tag{31}$$

Similarly, by (Zhang et al., 2020, Lemma 13), with probability at least $1-2p$, it holds that

$$|\chi_2| \leq 2\sqrt{\frac{\check{\nu}\iota}{\check{n}}} + \frac{5e^{\beta H}\iota^{\frac{3}{4}}}{\check{n}^{\frac{3}{4}}} + \frac{2\sqrt{\iota}}{T\check{n}} + \frac{2e^{\beta H}\iota}{\check{n}} \tag{32}$$

where

$$\check{\nu} := \frac{\check{\sigma}}{\check{n}} - \left(\frac{\check{\Delta}}{\check{n}}\right)^2 \tag{33}$$

Let $c_1 = 2, c_2 = 2,$ and $c_3 = 5$ in the construction of b_h (line 10 of Algorithm 1). With probability at least $1-2(e^{2\beta H}T^3+3)p$, we have that $b_h^k \geq |\chi_1| + |\chi_2|$, which concludes the proof. $\square$

B PROOF OF LEMMA 2

Proof. For convenience, we define $\tau_h^k := e^{\beta\cdot V_h^k(s_h^k)} - e^{\beta\cdot V_h^*(s_h^k)}$. Similar to the proof of (Zhang et al., 2020, Lemma 5), we will establish that for any weight sequence $\{w^k\}_{k=1}^K$ such that $w^k \geq 0$, it holds that

$$\sum_{k=1}^K w^k\tau_h^k \leq 240H^{\frac{3}{2}}e^{2\beta H}\sqrt{||w||_\infty \cdot SA||w||_1\iota} + 3e^{2\beta H}HSA||w||_\infty \tag{34}$$

where $||w||_\infty = \max_k w^k$ and $||w||_1 = \sum_k w^k$. Note that if Inequality (34) holds, then replacing w^k by $\mathbb{I}[\tau_h^k \geq \gamma]$ yields

$$\gamma\sum_{k=1}^K \mathbb{I}[\tau_h^k \geq \gamma] \leq \sum_{k=1}^K \mathbb{I}[\tau_h^k \geq \gamma]\tau_h^k \leq 240H^{\frac{3}{2}}e^{2\beta H}\sqrt{SA\iota\sum_{k=1}^K \mathbb{I}[\tau_h^k \geq \gamma]} + 3e^{2\beta H}HSA$$

which leads to

$$\sum_{k=1}^K \mathbb{I}[\tau_h^k \geq \gamma] \leq O\left(\frac{e^{4\beta H}H^3SA\iota}{\gamma^2}\right) \tag{35}$$

and concludes the proof. Now, we will prove Inequality (34). Conditioned on the successful events of Lemma 1, we have

$$\begin{aligned}
\tau_h^k =& e^{\beta\cdot V_h^k(s_h^k)} - e^{\beta\cdot V_h^*(s_h^k)} \leq e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^*(s_h^k,a_h^k)} \\
\leq& \mathbb{I}[n_h^k = 0]e^{\beta H} + e^\beta\left(\bar{b}_h^k + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} e^{\beta\cdot V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i})} - [P_h e^{\beta\cdot V_{h+1}^*}](s_h^k, a_h^k)\right) \\
\leq& \mathbb{I}[n_h^k = 0]e^{\beta H} + e^\beta\left(2\bar{b}_h^k + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} e^{\beta\cdot V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i})} - e^{\beta\cdot V_{h+1}^*(s_{h+1}^{\check{l}_i})}\right) \\
=& \mathbb{I}[n_h^k = 0]e^{\beta H} + e^\beta(2\bar{b}_h^k + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\tau_{h+1}^{\check{l}_i})
\end{aligned}$$

Let $\tilde{w}^k = \sum_{j=1}^K \frac{w^j}{\check{n}_h^j}\sum_{i=1}^{\check{n}_h^j}\mathbb{I}[k = \check{l}_{h,i}^j]$. By the exact analysis of Inequality (17), we have

$$\sum_{k=1}^K w^k\tau_h^k \leq e^\beta(2\sum_{k=1}^K w^k\bar{b}_h^k + \sum_{k=1}^K \tilde{w}^k\tau_{h+1}^k) + e^{\beta H}SA||w||_\infty \tag{36}$$

To bound the first term in the RHS of Inequality (36), following the exact analysis in the proof of (Zhang et al., 2020, Lemma 5), we obtain that

$$\sum_{k=1}^{K} w^k \bar{b}_h^k \le 20\sqrt{e^{2\beta H}}\iota(1+\frac{1}{H})\sqrt{||w||_\infty \cdot HSA||w||_1} \tag{37}$$

Plugging Inequality (37) into Inequality (36) yields

$$\sum_{k=1}^{K} w^k \tau_h^k \le e^{\beta}(80e^{\beta H}\sqrt{||w||_\infty \cdot HSA||w||_1\iota} + \sum_{k=1}^{K}\tilde{w}^k\tau_{h+1}^k) + e^{\beta H}SA||w||_\infty \tag{38}$$

Iterating Inequality (38) from H to h, and using the fact that $||\tilde{w}||_\infty \le (1+\frac{1}{H})||w||_\infty$ and $||\tilde{w}||_1 = ||w||_1$, we obtain

$$\sum_{k=1}^{K} w^k \tau_h^k \le 240He^{2\beta H}\sqrt{||w||_\infty \cdot HSA||w||_1\iota} + 3e^{2\beta H}HSA||w||_\infty$$

which concludes the proof. □

C PROOF OF THEOREM 1

Proof. We consider the case of $\beta > 0$.[8] To begin with, Term (20) is derived by the following lemma. (The detailed proof can be found in Appendix E.)

Lemma 4. *With probability at least $1 - O(e^{2\beta H}T^3 p \cdot T(HSA)^2 p')$ and when T is sufficiently large, it holds that*

$$\begin{aligned}\text{Regret}(T) \le& \frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{2(h-1)}\Lambda_{h-1}\left(\frac{2b_h^k}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)} + \psi_{h+1}^k + \xi_{h+1}^k + \phi_{h+1}^k\right)\\ \le& \tilde{O}\left(\frac{1}{\beta}\max_{h\in[H]}\left\{\Lambda_{h-1}\cdot c_{v,h+1}^*\right\}\sqrt{\max\{SA,H\}HT}\right)\end{aligned}$$

where $c_{v,h+1}^* = \max_{s,a}([P_h e^{\beta\cdot V_{h+1}^*}](s,a))^{-1}\sqrt{\mathbb{V}([P_h e^{\beta\cdot V_{h+1}^*}](s,a))}$ is the maximum per-step conditional coefficient of variation (CV) defined in Equation (4) of the exponential optimal value function. Here, $p' \in (0,1)$ defined in the proof is the failure probability of events that are independent of the successful events of Lemmas 1 and 2.

To derive Term (21), note that

$$\text{Regret}(T) \le \sum_{k=1}^{K}\left(V_1^k(s_1) - V_1^{\pi^k}(s_1)\right) \le \frac{1}{\beta}\left(e^{\beta\cdot V_1^k(s_1)} - e^{\beta\cdot V_1^{\pi^k}(s_1)}\right)$$

Next, we establish the following recursive form for the exponential Bellman equation

$$\begin{aligned}&\varsigma_h^k := e^{\beta\cdot V_h^k(s_h^k)} - e^{\beta\cdot V_h^{\pi^k}(s_h^k)} \le e^{\beta\cdot Q_h^k(s_h^k,a_h^k)} - e^{\beta\cdot Q_h^{\pi^k}(s_h^k,a_h^k)}\\ \le& \mathbb{I}[n_h^k=0]e^{\beta H} + e^{\beta}\left(\left(\frac{u_h^{\text{ref},k}}{n_h^k} + \frac{\check{\Delta}_h^k}{\check{n}_h^k} + b_h^k\right) - [P_h e^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)\right)\\ =& \mathbb{I}[n_h^k=0]e^{\beta H} + e^{\beta}\left(\left(\frac{u_h^{\text{ref},k}}{n_h^k} + \frac{\check{\Delta}_h^k}{\check{n}_h^k} + b_h^k\right) - [P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)\right)\\ &+ e^{\beta}\left(e^{\beta\cdot V_{h+1}^*(s_{h+1}^k)} - e^{\beta\cdot V_{h+1}^{\pi^k}(s_{h+1}^k)} + [(P_h-\widehat{P}_h)(e^{\beta\cdot V_{h+1}^*} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k,a_h^k)\right)\\ \le& \mathbb{I}[n_h^k=0]e^{\beta H} + e^{\beta}\left(\left[(P_h-\widehat{P}_h)\left(\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(e^{\beta\cdot V_{h+1}^{\check{l}_i}} - e^{\beta\cdot V_{h+1}^*}\right)\right)\right](s_h^k,a_h^k) + \psi_{h+1}^k + 2b_h^k\right)\end{aligned}$$

[8] We provide the modified algorithm and a sketch of proof for a negative β in Appendix G.

$$+ e^{\beta}\left(\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(e^{\beta\cdot V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i})} - e^{\beta\cdot V_{h+1}^{*}(s_{h+1}^{\check{l}_i})}\right) + e^{\beta\cdot V_{h+1}^{*}(s_{h+1}^{k})} - e^{\beta\cdot V_{h+1}^{\pi^k}(s_{h+1}^{k})}\right)$$
$$+ e^{\beta}[(P_h - \widehat{P}_h)(e^{\beta\cdot V_{h+1}^{*}} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k, a_h^k) \tag{39}$$

Summing over k and using the similar trick in Inequality (17), we derive

$$\sum_{k=1}^{K}\varsigma_h^k \le e^{\beta H}SA + e^{\beta}(1+\frac{1}{H})\sum_{k=1}^{K}\varsigma_{h+1}^k + \sum_{k=1}^{K} e^{\beta}[(P_h - \widehat{P}_h)(e^{\beta\cdot V_{h+1}^{*}} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k, a_h^k)$$
$$+ \sum_{k=1}^{K} e^{\beta}\left(\left[(P_h - \widehat{P}_h)\left(\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(e^{\beta\cdot V_{h+1}^{\check{l}_i}} - e^{\beta\cdot V_{h+1}^{*}}\right)\right)\right](s_h^k, a_h^k) + \psi_{h+1}^k + 2b_h^k\right) \tag{40}$$

Iterating over h, we obtain

$$\sum_{k=1}^{K}\varsigma_1^k \le e^{2\beta H}HSA + \sum_{k=1}^{K}\sum_{h=1}^{H}\left[e^{\beta}(1+\frac{1}{H})\right]^{h-1}[(P_h - \widehat{P}_h)(e^{\beta\cdot V_{h+1}^{*}} - e^{\beta\cdot V_{h+1}^{\pi^k}})](s_h^k, a_h^k)$$
$$+ \sum_{k=1}^{K}\sum_{h=1}^{H}\left[e^{\beta}(1+\frac{1}{H})\right]^{h-1}\left[(P_h - \widehat{P}_h)\left(\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(e^{\beta\cdot V_{h+1}^{\check{l}_i}} - e^{\beta\cdot V_{h+1}^{*}}\right)\right)\right](s_h^k, a_h^k)$$
$$+ \sum_{k=1}^{K}\sum_{h=1}^{H}\left[e^{\beta}(1+\frac{1}{H})\right]^{h-1}(\psi_{h+1}^k + 2b_h^k) \tag{41}$$

By slightly modifying the proof of Lemma 4, we have that with probability at least $(1 - O(e^{2\beta H}T^3 p\cdot T(HSA)^2 p'))$ and when T is sufficiently large,

$$\sum_{k=1}^{K}\varsigma_1^k \le \tilde{O}\left((e^{\beta H}-1)\sqrt{HSAT}\right) \tag{42}$$

We set $p \leftarrow p/\text{poly}(e^{2\beta H}, T)$ and $p' \leftarrow 1/\text{poly}(H, S, A, T)$ and conclude the proof. □

D PROOF OF COROLLARY 1.1

Proof. To prove this corollary, we first show that $c_{v,h+1}^* \le O(e^{|\beta|(H-h)/2})$. When $\beta > 0$ and a fixed h, assume that $[P_h e^{\beta\cdot V_{h+1}^*}](s,a)$ has an order of $O\left(e^{\beta t}\right)$ where $t \in [0, H-h]$. For any state s' that $P_h(s'|s,a) > 0$, we denote by $O(e^{\beta q_{s'}})$ the order of $e^{\beta\cdot V_{h+1}^*(s')}$ where $q_{s'} \in [0, H-h]$. If $q_{s'} > t$, we have that $P_h(s'|s,a) = O(e^{\beta(t-q_{s'})})$. By simple calculation, we derive that $\mathbb{V}([P_h e^{\beta\cdot V_{h+1}^*}](s_h,a_h)) \le O(\sum_{s':q_{s'}>t}(e^{\beta q_{s'}} - e^{\beta t})^2 e^{\beta(t-q_{s'})} + e^{2\beta t}) = O(e^{\beta(\max_{s'} q_{s'}+t)})$. Therefore, we have that $c_{v,h+1}^* \le O(e^{\beta(\max_{s'} q_{s'}-t)/2}) \le O(e^{\beta(H-h)/2})$. When $\beta < 0$, we only have to replace $\max_{s'} q_{s'}$ by $\min_{s'} q_{s'}$ and derive that $c_{v,h+1}^* \le O(e^{-\beta(H-h)/2})$. Next, we show that when $\max_h \lambda_{h+1} \le H^{-1/2}e^{|\beta|/2}$, we have that $\max_h\{\Lambda_{h-1}\cdot c_{v,h+1}^*\} \le H^{-1/2}e^{|\beta|(H-1)/2}$. Since we have that $\Lambda_{h-1} = \prod_{i=1}^{h-1}\lambda_{i+1} \le H^{-1/2}e^{|\beta|(h-1)/2}$ for $h \ge 2$ and $\Lambda_0 = 1$. Hence, we have that $\max_h\{\Lambda_{h-1}\cdot c_{v,h+1}^*\} \le H^{-1/2}e^{|\beta|(H-1)/2}$, which concludes the proof. □

E PROOF OF THEOREM 2

Proof. Given any $0 \le t \le (H-1)/2$ and $c_v^t = e^{|\beta|t}$, we show that there is an problem instance $M \in \mathcal{M}(c_v^t)$ such that *any* online algorithm suffers a regret $\Omega(c_v^t\sqrt{SAT}/|\beta|)$, where the class of problems $\mathcal{M}(c_v^t)$ is defined in Equation (22). Since the proof for $\beta > 0$ and $\beta < 0$ is similar, we focus on the case $\beta > 0$. Inspired by the proof of (Fei et al., 2020, Theorem 3), we construct the following hard instance M:

- The state space is $\mathcal{S} := \{s_i\}_{i\in[S]} \cup \{s_g, s_b\}$. There are S "bandit states" $\{s_i\}_{i\in[S]}$, one "good state" s_g, and one "bad state" s_b.

- The action space is $\mathcal{A} := [A]$.
- Transition kernel $\mathcal{P}$: Each bandit state $s_i, i \in [S]$ can only transit to either the "good state" s_g or the "bad state" s_b. Particularly, for some fixed $a^* \in [A]$, it holds that $P_h(s_g|s_i, a^*) = \delta + \epsilon$ and $P_h(s_b|s_i, a^*) = 1 - \delta - \epsilon$ for some $\delta, \epsilon > 0$ that will be specified later and any $h \in [H]$. For any $a \neq a^*$, we have that $P_h(s_g|s_i, a^*) = \delta$ and $P_h(s_b|s_i, a^*) = 1 - \delta$ for any $h \in [H]$. Both the "good state" s_g and the "bad state" s_b are absorbing, i.e., $P_h(s_g|s_g, a) = P_h(s_b|s_b, a) = 1$ for any $(h, a) \in [H] \times [A]$.
- Reward function r: At each bandit state $s_i, i \in [S]$ and the "bad state" s_b, all actions yield no reward, i.e., $r_h(s_b, a) = r_h(s_i, a) = 0$ for any $(h, a) \in [H] \times [A]$. In addition, at the "good state" s_g, any action yields a reward 1, i.e., $r_h(s_g, a) = 1$ for any $(h, a) \in [H] \times [A]$.
- Initial state distribution is uniform on the bandit states, i.e., $S_1 \sim \text{Unif}\{s_i\}_{i \in [S]}$.

We first show that this problem instance M belongs to the class of problems $\mathcal{M}(t)$ when δ and ϵ are carefully selected. Let $p_1 := \delta + \epsilon$ and $p_2 := \delta$. Note that the optimal policy π^* is to select arm a^* with probability 1 at any bandit state at the first timestep (and to select any arbitrary action in the rest of the episode since the agent transits to either s_g or s_b after the first timestep and is absorbed in that state in the rest of the episode), we have that

$$\max_h c_{v,h+1}^{\pi^*} = c_{v,2}^{\pi^*} = \frac{\sqrt{p_1 - p_1^2}(e^{\beta(H-1)} - 1)}{p_1 e^{\beta(H-1)} + (1 - p_1)}$$

where the first equation holds by the fact that $c_{v,h+1}^{\pi^*} = 0$ for any $h = 2, ..., H$. Hence, if p_1 is selected such that

$$\max_h c_{v,h+1}^{\pi^*} = \frac{\sqrt{p_1 - p_1^2}(e^{\beta(H-1)} - 1)}{p_1(e^{\beta(H-1)} - 1) + 1} = O(c_v^t)$$

then we have that $M \in \mathcal{M}(t)$. This can be achieved by setting $p_1 = O(e^{2\beta(t-H+1)})$. In addition, by the construction of M, at each bandit state s_i at the first timestep, the MDP can be reduced to an A-armed bandit where all arms are i.i.d. $(H-1) \cdot \text{Ber}(\delta)$, but one arm a^* is i.i.d. $(H-1) \cdot \text{Ber}(\delta + \epsilon)$ for some $\delta, \epsilon > 0$. Therefore, we can simply work on this MAB problem instead of the original problem instance M. Consider that the agent interacts with the MAB for K episodes. For the kth episode, let S_1^k be the initial state and we denote by $\pi^k : \mathcal{S} \to \Delta([A])$ the policy used by the agent, where $\pi^k(a|S_1^k) =: \mathbb{P}(\pi^k = a)$ is the probability that arm a is selected according π^k. Therefore, we have that for any $k \in [K]$

$$V_1^*(S_1^k) = \frac{1}{\beta} \ln \mathbb{E} e^{\beta \cdot r(a^*)}$$

$$V_1^{\pi^k}(S_1^k) = \frac{1}{\beta} \ln \mathbb{E} e^{\beta \cdot r_k} = \frac{1}{\beta} \ln \left(\sum_{a \in [A]} \mathbb{P}(\pi^k = a) \cdot \mathbb{E} e^{\beta \cdot r(a)} \right)$$

where we denote by r_k the (random) reward received at the kth episode following policy π^k and we denote by $r(a)$ the (random) reward when pulling the ath arm. For any $a \in [A], a \neq a^*$ and $k \in [K]$, let

$$\begin{aligned}
\Delta :=& \frac{\mathbb{E} e^{\beta \cdot r(a^*)} - \mathbb{E} e^{\beta \cdot r(a)}}{\mathbb{E} e^{\beta \cdot r(a^*)}} \\
=& \frac{p_1 e^{\beta(H-1)} + (1 - p_1) - [p_2 e^{\beta(H-1)} + (1 - p_2)]}{\mathbb{E} e^{\beta \cdot r(a^*)}} = \frac{\epsilon(e^{\beta(H-1)} - 1)}{\mathbb{E} e^{\beta \cdot r(a^*)}}
\end{aligned}$$

Let a_k denote the arm pulled by the agent at the kth episode. Hence, the regret at the kth episode is

$$\begin{aligned}
V_1^*(S_1^k) - V_1^{\pi^k}(S_1^k) =& \frac{1}{\beta} \ln \left(\frac{\sum_{a \in [A]} \mathbb{P}(\pi^k = a) \cdot \mathbb{E} e^{\beta \cdot r(a^*)}}{\sum_{a \in [A]} \mathbb{P}(\pi^k = a) \cdot \mathbb{E} e^{\beta \cdot r(a)}} \right) \\
=& \frac{1}{\beta} \ln \left(1 + \frac{\sum_{a \in [A], a \neq a^*} \mathbb{P}(\pi^k = a) \cdot \left(\mathbb{E} e^{\beta \cdot r(a^*)} - \mathbb{E} e^{\beta \cdot r(a)} \right)}{\sum_{a \in [A]} \mathbb{P}(\pi^k = a) \cdot \mathbb{E} e^{\beta \cdot r(a)}} \right) \\
\geq& \frac{1}{\beta} \ln \left(1 + \frac{\sum_{a \in [A], a \neq a^*} \mathbb{P}(\pi^k = a) \cdot \left(\mathbb{E} e^{\beta \cdot r(a^*)} - \mathbb{E} e^{\beta \cdot r(a)} \right)}{\mathbb{E} e^{\beta \cdot r(a^*)}} \right)
\end{aligned}$$

$$=\frac{1}{\beta}\ln\left(1+\mathbb{E}\{\mathbb{I}[a_k\neq a^*]|a_k\sim\pi^k(\cdot|S_1^k)\}\Delta\right)$$
$$\geq\frac{1}{2\beta}\cdot\mathbb{E}\{\mathbb{I}[a_k\neq a^*]|a_k\sim\pi^k(\cdot|S_1^k)\}\cdot\Delta$$

where the last inequality holds by $\ln(1+x)\geq x/2$ for any $x\in[0,1]$. Therefore, we further have that

$$\mathbb{E}_M\left[\sum_{k=1}^K\left(V_1^*(S_1^k)-V_1^{\pi^k}(S_1^k)\right)\right]\geq\frac{1}{2\beta}\cdot\sum_{k=1}^K\mathbb{E}_{M,\pi^k}\left\{\mathbb{I}[a_k\neq a^*]\right\}\cdot\Delta \tag{43}$$

where the expectation $\mathbb{E}_{\pi^k}$ is w.r.t the randomness during the algorithm execution within MDP M. To further derive a lower bound of the RHS of Inequality (43), we first consider the regret when the agent is uninformative about the optimal action a^*.

Regret of an uninformative agent. We consider a problem instance M_0 that has the same construction as the above problem instance M except that there is no "special" action a^*, i.e., it holds that $P_h(s_g|s_i,a^*)=\delta$ and $P_h(s_b|s_i,a^*)=1-\delta$ for any $(h,i)\in[H]\times[S]$. When the agent interacts with M_0, she is uninformative in the sense no information is provided on the action a^*. Therefore, we have that

$$\sum_{k=1}^K\mathbb{E}_{M_0,\pi^k}\left\{\mathbb{I}[a_k\neq a^*]\right\}=\sum_{k=1}^K\frac{A-1}{A}=K\left(1-\frac{1}{A}\right) \tag{44}$$

We now establish that, if $\epsilon=P_h(s_g|s_i,a^*)-P_h(s_g|s_i,a),a\neq a^*$ is sufficiently small, then over a limited time horizon, the observation from interacting with the problem instance M cannot be significantly different from the observation from interacting with the problem instance M_0. If that is the case, then $\sum_{k=1}^K\mathbb{E}_{M,\pi^k}\{\mathbb{I}[a_k\neq a^*]\}$ should be close to $\sum_{k=1}^K\mathbb{E}_{M_0,\pi^k}\{\mathbb{I}[a_k\neq a^*]\}$. To formalize this idea, we first introduce some useful notations. Note that each episode starts at a random bandit state and the rewards of the arms at these bandit states are independent and identically distributed. Therefore, we can consider each bandit state $s_i,i\in[S]$ independently. We denote by $\mathcal{H}_k^i=(s_i,a_1,r_1,\cdots,s_i,a_{k-1},r_{k-1})$ any possible sequence of histories starting from state s_i from interacting with the problem instance M. Similarly, we define $\tilde{\mathcal{H}}_k^i=(s_i,\tilde{a}_1,\tilde{r}_1,\cdots,s_i,\tilde{a}_{k-1},\tilde{r}_{k-1})$ any possible sequence of histories from interacting with the problem instance M_0. Since the initial distribution is uniform over all bandit states, each state s_i is expected to be visited at the first timestep by K/S times. Let $B_{k,i}^{K/S}:=(r_k,\cdots,r_{K/S})$ and $\tilde{B}_{k,i}^{K/S}:=(\tilde{r}_k,\cdots,\tilde{r}_{K/S})$. We define $P(b_{k,i}^{K/S}|\mathcal{H}_k):=\mathbb{P}(B_{k,i}^{K/S}=b_{k,i}^{K/S}|\mathcal{H}_k)$ and $\tilde{P}(b_{k,i}^{K}|\tilde{\mathcal{H}}_k):=\mathbb{P}(\tilde{B}_{k,i}^{K/S}=b_{k,i}^{K/S}|\tilde{\mathcal{H}}_k)$. To quantify the difference between these two distributions, we employ the following notion of KL divergence

$$d_{KL}\left(\tilde{P}(b_{k,i}^{K/S}|\tilde{\mathcal{H}}_k^i),P(b_{k,i}^{K/S}|\mathcal{H}_k^i)\right)=\mathbb{E}\left[\sum_{b_{k,i}^{K/S}}\tilde{P}(b_{k,i}^{K/S}|\tilde{\mathcal{H}}_k^i)\ln\left(\frac{\tilde{P}(b_{k,i}^{K/S}|\tilde{\mathcal{H}}_k^i)}{P(b_{k,i}^{K/S}|\mathcal{H}_k^i)}\right)\right]$$

Applying the chain rule of KL divergence, we obtain that

$$d_{KL}\left(\tilde{P}(b_{1,i}^{K/S}|\tilde{\mathcal{H}}_k^i),P(b_{1,i}^{K/S}|\mathcal{H}_k^i)\right)=\sum_{k=1}^{K/S}d_{KL}\left(\tilde{P}(b_{k,i}^{k}|\tilde{\mathcal{H}}_k^i),P(b_{k,i}^{k}|\mathcal{H}_k^i)\right)$$
$$=\sum_{k=1}^{K/S}\mathbb{P}[\tilde{a}_k=a^*]\left(\delta\ln\left(\frac{\delta}{p_1}\right)+(1-\delta)\ln\left(\frac{1-\delta}{1-p_1}\right)\right)$$
$$=\frac{K}{SA}\left(\delta\ln\left(\frac{\delta}{p_1}\right)+(1-\delta)\ln\left(\frac{1-\delta}{1-p_1}\right)\right)\leq\frac{K}{SA}\frac{\epsilon^2}{\delta\ln 2}$$

where the last inequality holds by (Osband and Roy, 2016, Proposition 1) Let $n_{K/S}^i(a^*):=\sum_{k=1}^{K/S}\mathbb{I}[a_k=a^*]$ denote the (random) number of times that arm a^* is chosen in these K/S episodes that starts from the bandit state s_i in the problem instance M. Similarly, we denote by $\tilde{n}_{K/S}^i(a^*):=\sum_{k=1}^{K/S}\mathbb{I}[\tilde{a}_k=a^*]$ the (random) number of times that arm a^* is chosen in these K/S episodes that starts from the bandit state s_i in the problem instance M_0. Using Pinsker's inequality, we have that

$$\mathbb{E}\left[\frac{n_{K/S}^i(a^*)}{K/S}-\frac{\tilde{n}_{K/S}^i(a^*)}{K/S}\right]\leq\sqrt{\frac{1}{2}d_{KL}\left(\tilde{P}(b_{k,i}^{K/S}),P(b_{k,i}^{K/S})\right)}$$

Since $\mathbb{E}[\tilde{n}^i_{K/S}(a^*)] = K/(SA)$ due to the fact that the agent is uninformative, it holds that

$$\mathbb{E}\left[\frac{n^i_{K/S}(a^*)}{K/S}\right] \le \sqrt{\frac{1}{2}d_{KL}\left(\tilde{P}(b^{K/S}_{k,i}), P(b^{K/S}_{k,i})\right)} + \frac{1}{A} \tag{45}$$

Since Inequality (45) holds for any arbitrary bandit state $s_i, i \in [S]$. Therefore, if $\delta \in [0, 1/2]$ and $\epsilon \le 1 - 2\delta$, then through a simple substitution in deriving Inequality (44), we have that

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}_{M_0,\pi^k}\left\{\mathbb{I}[a_k \ne a^*]\right\} = &\sum_{i\in[S]} \frac{K}{S}\left(1 - \frac{1}{A} - \sqrt{\frac{1}{2}d_{KL}\left(\tilde{P}(b^{K/S}_{k,i}), P(b^{K/S}_{k,i})\right)}\right)\\ \ge &K\left(1 - \frac{1}{A} - \sqrt{\frac{K}{SA}\frac{\epsilon^2}{2\delta}}\right)\end{aligned}$$

Plugging in Inequality (43), we obtain that

$$\begin{aligned}&\mathbb{E}_M\left[\sum_{k=1}^{K}\left(V_1^*(S_1^k) - V_1^{\pi^k}(S_1^k)\right)\right] \ge \frac{1}{2\beta}K\left(1 - \frac{1}{A} - \sqrt{\frac{K}{SA}\frac{\epsilon^2}{2\delta}}\right)\cdot\Delta\\ \ge&\frac{1}{2\beta}K\left(1 - \frac{1}{A} - \sqrt{\frac{K}{SA}\frac{\epsilon^2}{2\delta}}\right)\frac{\epsilon(e^{\beta(H-1)} - 1)}{\mathbb{E}e^{\beta\cdot r(a^*)}}\\ =&\frac{1}{8\beta}\frac{\sqrt{\delta}(e^{\beta(H-1)} - 1)}{\mathbb{E}e^{\beta\cdot r(a^*)}}\sqrt{SAK} \text{ by setting } \epsilon^2 = \frac{\delta SA}{8K}\\ \ge&\frac{1}{8\beta}\frac{\sqrt{p_1 - p_1^2}(e^{\beta(H-1)} - 1)}{\mathbb{E}e^{\beta\cdot r(a^*)}}\sqrt{SAK} \text{ for sufficiently large } K \text{ such that } \epsilon \le (\delta+\epsilon)^2\\ =&\frac{1}{8\beta}\cdot\max_h c^{\pi^*}_{v,h+1}\cdot\sqrt{SAK}\\ =&\Omega\left(\frac{c_v^t}{\beta}\sqrt{SAK}\right)\end{aligned}$$

Further, since the transition kernel is timestep-dependent by definition, i.e., $P_1, P_2, ..., P_H$ may not be identical. We augment the state from S to be HS as in the proof of (Jin et al., 2018b, Theorem 3). Recall that $T := KH$. Since the case of $\beta < 0$ can be proved similarly, therefore, we conclude that

$$\text{Regret}(T, \mathcal{M}(t)) \ge \Omega\left(\frac{e^{|\beta|t}}{|\beta|}\sqrt{SAT}\right)$$

Note that $\max_h c^{\pi^*}_{v,h+1} \le O(e^{|\beta|(H-1)/2})$ for any MDP (See Appendix D). Hence, when $|\beta|(H-1)$ is sufficiently large, this bound translates to

$$\Omega\left(\frac{e^{\frac{|\beta|(H-1)}{2}} - 1}{|\beta|}\sqrt{SAT}\right)$$

in the worst case, which concludes the proof. ∎

F PROOF OF LEMMA 4

Let $p' \in (0, 1)$ denote the failure probability of events that are *independent* of the successful events of Lemmas 1 and 2. In the rest of the proof, we define $\eta_h := (1 + \frac{1}{H})^{2(h-1)}\Lambda_{h-1}$ and $\iota' := \ln(2/p')$.

F.1 Upper Bound ψ^k_{h+1} Term.

Lemma 5. *With probability at least $1 - (HSA + 1)p'$, it holds that*

$$\frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\psi^k_{h+1} \le \frac{1}{\beta}\Lambda_{H-1}(\ln(T) + 1)\left(N^{\alpha'}_{p'}\cdot e^{\beta H}HSA + 2N_0(\alpha')\cdot H^{\frac{3}{2}}S^{\frac{3}{2}}A^{\frac{1}{2}} + 2\sqrt{HSAT\iota'}\right)$$

where $\alpha' = \sqrt{HSA}$ is the input of Algorithm 1 and $N_{p'}^{\alpha'}$ is defined in Equation (6).

Proof. Let $\gamma \in (0, e^{\beta H}]$. Define $\varphi_{h+1}^k(s,a,\gamma) := \mathbb{I}[\vee_{s':P_h(s'|s,a)>S^{-1}e^{-\beta H}\gamma}(N_h^k(s') < N_0(\gamma))]$. That is, $\varphi_{h+1}^k(s,a,\gamma) = 1$ means that there exists some state s' such that $P_h(s'|s,a) > S^{-1}e^{-\beta H}$ is visited by less than $N_0(\gamma)$ times. For convenience, we denote $\bar{\varphi}_{h+1}^k(s,a,\gamma) := \mathbb{I}[\wedge_{s':P_h(s'|s,a)>S^{-1}e^{-\beta H}\gamma}(N_h^k(s') \geq N_0(\gamma))] = 1 - \varphi_{h+1}^k(s,a,\gamma)$. Again, $\bar{\varphi}_{h+1}^k(s,a,\gamma) = 1$ means that *any* such states of taking action a at state s at timestep h are visited by more than $N_0(\gamma)$ times. We have that

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\psi_{h+1}^k \\
\leq& e^2\Lambda_{H-1}\sum_{h=1}^{H}\sum_{j=1}^{H}\left[P_h(e^{\beta\cdot V_{h+1}^{\text{ref},j}} - e^{\beta\cdot V_{h+1}^{\text{REF}}})\right](s_h^j,a_h^j)\sum_{k=1}^{K}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{I}[l_{h,i}^k = j] \\
\leq& 2e^2\Lambda_{H-1}(\ln(T)+1)\sum_{h=1}^{H}\sum_{k=1}^{K}\left[P_h(e^{\beta\cdot V_{h+1}^{\text{ref},k}} - e^{\beta\cdot V_{h+1}^{\text{REF}}})\right](s_h^k,a_h^k) \qquad (46) \\
\leq& 2e^2\Lambda_{H-1}(\ln(T)+1)\left(\sum_{h=1}^{H}\sum_{k=1}^{K}e^{\beta H}\varphi_{h+1}^k(s_h^k,a_h^k,\alpha') + \sum_{h=1}^{H}\sum_{k=1}^{K}\bar{\varphi}_{h+1}^k(s_h^k,a_h^k,\alpha')\left[P_h(e^{\beta\cdot V_{h+1}^{\text{ref},k}} - e^{\beta\cdot V_{h+1}^{\text{REF}}})\right](s_h^k,a_h^k)\right) \\
& \qquad (47)
\end{aligned}
$$

where Inequality (46) follows from the same trick in the derivation of (Zhang et al., 2020, Inequality (58)). To further obtain an upper bound, we first state an important result.

Lemma 6. *Let $p' \in (0,1)$ denote the failure probability and $\gamma \in (0, e^{\beta H}]$. We define*

$$
\begin{aligned}
N_{p'}^{\gamma} :=& \min\{n \in N^+ \mid n \cdot S^{-1}e^{-\beta H}\gamma - \sqrt{2Sn\ln(2/p')} > N_0(\gamma)\} \\
=& \left(\sqrt{\frac{N_0(\gamma)}{S^{-1}e^{-\beta H}\gamma} + \frac{2S\ln(2/p')}{4S^-2e^{-2\beta H}\gamma^2}} + \frac{\sqrt{2S\ln(2/p')}}{2S^{-1}e^{-\beta H}\gamma}\right)^2
\end{aligned} \qquad (48)
$$

where $N_0(\gamma) = c_4e^{4\beta H}H^3SA\iota/\gamma^2$ is defined in Lemma 3. For any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, if $N_h^j(s,a) \geq N_{p'}^{\gamma}$ holds at episode j, then we have that $N_{h+1}^j(s') := \sum_{a\in\mathcal{A}} N_{h+1}^j(s',a)\ N_h^j(s,a) \geq N_0(\gamma)$ for any $s' \in \mathcal{S}$ such that $P_h(s'|s,a) \geq S^{-1}e^{-\beta H}\gamma$.

Proof. The proof relies on the L_1 deviation bound for a multinomial distribution (Weissman et al., 2003), which is stated as follows without proof.

Lemma 7. *Let $p' \in (0,1)$ denote the failure probability. For any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, it holds that*

$$
P\left(|N_{h+1}(s') - N_h(s,a)\cdot P_h(s'|s,a)| \leq \sqrt{2SN_h(s,a)\cdot\ln(\frac{2}{p'})}\right) \geq 1-p', \forall s' \in \mathcal{S}
$$

where $N_{h+1}(s')$ is the number of visits to state s' at timestep $h+1$ after taking action a at state s at timestep h, and $N_h(s,a)$ is the number of visits to (h,s,a).

Note that $N_{h+1}(s') \geq \frac{1}{P_h(s'|s,a)}\left(N_h(s,a) - \sqrt{2SN_h(s,a)\cdot\ln(\frac{2}{p'})}\right)$. Letting the RHS no less than $N_0(\gamma)$ yields the result. □

Intuitively, Lemma 6 states that any state s' that can be reached by taking action a at state s at timestep h with probability no less than $S^{-1}e^{-\beta H}\gamma$ is visited more than $N_0(\gamma)$ times when (h,s,a) is experienced $N_{p'}^{\gamma}$ times. Therefore, we have that

$$
\sum_{h=1}^{H}\sum_{k=1}^{K}\varphi_{h+1}^k(s_h^k,a_h^k,\alpha') \leq HSA\cdot N_{p'}^{\alpha} \qquad (49)
$$

Further, note that when $\bar{\varphi}_{h+1}^k(s,a,\gamma)=1$, we have that

$$
\begin{aligned}
&\left[P_h(e^{\beta\cdot V_{h+1}^{\mathrm{ref},k}}-e^{\beta\cdot V_{h+1}^{\mathrm{REF}}})\right](s,a)\\
=&\sum_{s':P_h(s'|s,a)\geq S^{-1}e^{-\beta H}\gamma}P_h(s'|s,a)(e^{\beta\cdot V_{h+1}^{\mathrm{ref},k}}-e^{\beta\cdot V_{h+1}^{\mathrm{REF}}})+\sum_{s':P_h(s'|s,a)<S^{-1}e^{-\beta H}\gamma}P_h(s'|s,a)(e^{\beta\cdot V_{h+1}^{\mathrm{ref},k}}-e^{\beta\cdot V_{h+1}^{\mathrm{REF}}})\\
\leq&\sum_{s':P_h(s'|s,a)\geq S^{-1}e^{-\beta H}\gamma}P_h(s'|s,a)\gamma+\sum_{s':P_h(s'|s,a)<S^{-1}e^{-\beta H}\gamma}S^{-1}e^{-\beta H}\gamma\cdot e^{\beta H}\\
\leq&\gamma+S\cdot S^{-1}e^{-\beta H}\gamma\cdot e^{\beta H}=2\gamma
\end{aligned}
$$

Therefore, we derive that

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\bar{\varphi}_{h+1}^k(s_h^k,a_h^k,\alpha')\left[P_h(e^{\beta\cdot V_{h+1}^{\mathrm{ref},k}}-e^{\beta\cdot V_{h+1}^{\mathrm{REF}}})\right](s_h^k,a_h^k)\\
\leq&2\alpha'\sum_{h=1}^{H}\sum_{k=1}^{K}\bar{\varphi}_{h+1}^k\left(s_h^k,a_h^k,\alpha'\right)\left(\mathbb{I}[N_h^k(s)<N_0(\alpha')]+\left[(P_h-\widehat{P}_h)\mathbb{I}[N_h^k(s)<N_0(\alpha')]\right](s_h^k,a_h^k)\right)\\
\leq&2N_0(\alpha')\cdot\alpha' HS+2\sqrt{\alpha'^2T\iota'}
\end{aligned}
\tag{50}
$$

which concludes the proof. □

F.2 Upper Bound ξ_{h+1}^k Term.

Lemma 8. *With probability at least $1-((T+1)HSAp')$, it holds that*

$$
\frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\xi_{h+1}^k\leq\frac{1}{\beta}\Lambda_{H-1}\cdot O\left(e^{\beta H}N_{p'}^{\sqrt{H}}HSA+\sqrt{HT\iota'}+\sqrt{HSAT\iota'}\right)
$$

where $N_{p'}^{\sqrt{H}}$ is defined in Equation (6).

Proof. Define $\theta_{h+1}^k=\eta_h\sum_{k=1}^K(1/\check{n}_h^k)\sum_{i=1}^{\check{n}_h^k}\mathbb{I}[\check{l}_{h,i}^k=j]$, $\tilde{\theta}_{h+1}^k=\eta_h\cdot\lfloor(1+1/H)x_h^j\rfloor/x_h^j$, and x_h^j is the number of elements in the current stage with respect to (s_h^j,a_h^j,h). Similar to the analysis in the proof of Lemma 5, we derive

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\xi_{h+1}^k\\
=&\sum_{h=1}^{H}\sum_{j=1}^{K}\eta_h\frac{\left[\left(P_h-\widehat{P}_h\right)(e^{\beta\cdot V_{h+1}^j}-e^{\beta\cdot V_{h+1}^*})\right](s_h^j,a_h^j)}{\left[P_he^{\beta\cdot V_{h+1}^*}\right](s_h^j,a_h^j)}\sum_{k=1}^{K}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbb{I}[\check{l}_{h,i}^k=j]\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}\theta_{h+1}^k\frac{\left[\left(P_h-\widehat{P}_h\right)(e^{\beta\cdot V_{h+1}^k}-e^{\beta\cdot V_{h+1}^*})\right](s_h^k,a_h^k)}{\left[P_he^{\beta\cdot V_{h+1}^*}\right](s_h^k,a_h^k)}\\
\leq&\sum_{h=1}^{H}\sum_{k=1}^{K}\theta_{h+1}^k\left(\varphi_{h+1}^k(s_h^k,a_h^k,\gamma)e^{\beta H}+\bar{\varphi}_{h+1}^k(s_h^k,a_h^k,\gamma)\frac{\left[\left(P_h-\widehat{P}_h\right)(e^{\beta\cdot V_{h+1}^k}-e^{\beta\cdot V_{h+1}^*})\right](s_h^k,a_h^k)}{\left[P_he^{\beta\cdot V_{h+1}^*}\right](s_h^k,a_h^k)}\right)\\
\leq&\Lambda_{H-1}\cdot O\left(e^{\beta H}N_{p'}^{\gamma}HSA\right)+\sum_{h=1}^{H}\sum_{k=1}^{K}\theta_{h+1}^k\bar{\varphi}_{h+1}^k(s_h^k,a_h^k,\gamma)\frac{\left[\left(P_h-\widehat{P}_h\right)(e^{\beta\cdot V_{h+1}^k}-e^{\beta\cdot V_{h+1}^*})\right](s_h^k,a_h^k)}{\left[P_he^{\beta\cdot V_{h+1}^*}\right](s_h^k,a_h^k)}\\
=&\Lambda_{H-1}\cdot O\left(e^{\beta H}N_{p'}^{\gamma}HSA\right)+\sum_{h=1}^{H}\sum_{k=1}^{K}\tilde{\theta}_{h+1}^k\bar{\varphi}_{h+1}^k(s_h^k,a_h^k,\gamma)\frac{\left[\left(P_h-\widehat{P}_h\right)(e^{\beta\cdot V_{h+1}^k}-e^{\beta\cdot V_{h+1}^*})\right](s_h^k,a_h^k)}{\left[P_he^{\beta\cdot V_{h+1}^*}\right](s_h^k,a_h^k)}
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{h=1}^{H}\sum_{k=1}^{K}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\bar{\varphi}_{h+1}^k(s_h^k,a_h^k,\gamma)\frac{\left[\left(P_h-\widehat{P}_h\right)\left(e^{\beta\cdot V_{h+1}^k}-e^{\beta\cdot V_{h+1}^*}\right)\right](s_h^k,a_h^k)}{\left[P_h e^{\beta\cdot V_{h+1}^*}\right](s_h^k,a_h^k)}\\
&\le\Lambda_{H-1}\cdot O\left(e^{\beta H}N_{p'}^{\gamma}HSA+\sqrt{\gamma^2T\iota'}+\sqrt{\gamma^2 SAT\iota'}\right)
\end{aligned}
\tag{51}
$$

Here, the second inequality holds with probability $(1-HSAp')$ by lemma 7 and a union bound over all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$. The last inequality happens with probability $(1-(T+1)p')$ by Azuma-Hoeffding's inequality following a similar analysis in the proof of Lemma 15 in (Zhang et al., 2020). Set $\gamma\leftarrow\sqrt{H}$ and we conclude the proof. □

F.3 Upper Bound ϕ_{h+1}^k Term.

Lemma 9. *With probability $(1-p')$, it holds that*

$$
\frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\phi_{h+1}^k\le\Lambda_{H-1}\cdot O\left(\sqrt{H^2T\iota'}\right)
$$

Proof. Define

$$
z_h^\pi(s,a):=\frac{[P_h e^{\beta\cdot V_{h+1}^\pi}](s,a)}{e^{\beta\lambda_{h+1}\cdot[P_hV_{h+1}^\pi](s,a)}}
\tag{52}
$$

By the definition of λ_{h+1} in Equation (16), we have

$$
\frac{\partial z_h(s,a)}{\partial V_{h+1}^\pi(s')}\propto e^{\beta\cdot V_{h+1}^\pi(s')}-\lambda_{h+1}[P_he^{\beta\cdot V_{h+1}^\pi}](s,a)\le 0
\tag{53}
$$

Therefore, we have for any policy π and any $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$

$$
\frac{[P_he^{\beta\cdot V_{h+1}^*}](s,a)}{e^{\beta\lambda_{h+1}\cdot[P_hV_{h+1}^*](s,a)}}\le\frac{[P_he^{\beta\cdot V_{h+1}^\pi}](s,a)}{e^{\beta\lambda_{h+1}\cdot[P_hV_{h+1}^\pi](s,a)}}
\tag{54}
$$

Notice that $\eta_h\lambda_{h+1}=(1+1/H)^{2(h-1)}\Lambda_{h-1}\lambda_{h+1}=(1+1/H)^{2(h-1)}\Lambda_h\le e^2\Lambda_{H-1}$. By Azuma-Hoeffding's inequality, we can easily derive

$$
\begin{aligned}
&\frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\phi_{h+1}^k\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\left(\frac{1}{\beta}\ln\left(\frac{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}{e^{\beta\lambda_{h+1}\cdot V_{h+1}^*(s_{h+1}^k)}}\right)-\frac{1}{\beta}\ln\left(\frac{[P_he^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)}{e^{\beta\lambda_{h+1}\cdot V_{h+1}^{\pi^k}(s_{h+1}^k)}}\right)\right)\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\frac{1}{\beta}\ln\left(\frac{z_h^{\pi^*}(s_h^k,a_h^k)}{z_h^{\pi^k}(s_h^k,a_h^k)}\right)+\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\lambda_{h+1}\left[(P_h-\widehat{P}_h)(V_{h+1}^*-V_{h+1}^{\pi^k})\right](s_h^k,a_h^k)\\
\le&O\left(\Lambda_{H-1}\sqrt{H^2T\iota'}\right)
\end{aligned}
\tag{55}
$$

□

F.4 Upper Bound b_h^k Term.

Lemma 10. *Recall that $p\in(0,1)$ is the failure probability defined in Lemmas 1 and 2 and $\iota=\log(2/p)$. With probability at least $1-4p$, it holds that*

$$
\begin{aligned}
&\frac{2}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h\frac{b_h^k}{[P_he^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\\
\le&\frac{1}{\beta}\cdot O\Bigg(\max_h\{\Lambda_{h-1}\cdot c_{v,h+1}^*\}\sqrt{HSAT\iota}\\
&\qquad+\Lambda_{H-1}\left(\sqrt{e^{\beta H}\alpha HSAT\iota}+\sqrt{\alpha^2HSAT\iota}+e^{\beta H}(HSA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}+e^{\beta H}HSA\sqrt{SN_0(\alpha)\iota}\ln(T)\right)\Bigg)
\end{aligned}
$$

Proof. Define $\nu_h^{\mathrm{ref},k} = \frac{\sigma^{\mathrm{ref},k}}{n_h^k} - (\frac{u^{\mathrm{ref},k}}{n_h^k})^2$ and $\check{\nu}_h^k = \frac{\check{\sigma}_h^k}{\check{n}_h^k} - (\frac{\check{\Delta}_h^k}{\check{n}_h^k})^2$. By the definition of b_h^k, we have that

$$\begin{aligned}
&2\sum_{h=1}^{H}\sum_{k=1}^{K}\eta_h \frac{b_h^k}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)} \\
\le& 2e^2 \sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\Lambda_{h-1}}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\left(c_1\sqrt{\frac{\nu_h^{\mathrm{ref},k}}{n_h^k}\iota} + c_2\sqrt{\frac{\check{\nu}_h^k}{\check{n}_h^k}\iota}\right) \\
&+ 2e^2\Lambda_{H-1}\sum_{h=1}^{H}\sum_{k=1}^{K}\left(c_3\left(\frac{e^{\beta H}\iota}{n_h^k} + \frac{e^{\beta H}\iota}{\check{n}_h^k} + \frac{e^{\beta H}\iota^{\frac{3}{4}}}{(n_h^k)^{\frac{3}{4}}} + \frac{e^{\beta H}\iota^{\frac{3}{4}}}{(\check{n}_h^k)^{\frac{3}{4}}}\right)\right) \\
\le& O\Bigg(\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\Lambda_{h-1}}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\left(\sqrt{\frac{\nu_h^{\mathrm{ref},k}}{n_h^k}\iota} + \sqrt{\frac{\check{\nu}_h^k}{\check{n}_h^k}\iota}\right) \\
&+ e^{\beta H}\Lambda_{H-1}H^2 SA\ln(T)\iota + e^{\beta H}\Lambda_{H-1}H^{\frac{3}{2}}(SA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}\Bigg)
\end{aligned} \tag{56}$$

The following Lemma is a counterpart of Lemma 18 in (Zhang et al., 2020).

Lemma 11. *With probability at least $1-4p$, it holds that*

$$\nu_h^{\mathrm{ref},k} - \mathbb{V}([P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)) \le 4e^{\beta H}\alpha + \frac{6e^{2\beta H}SN_0(\alpha)}{n_h^k} + 14e^{2\beta H}\sqrt{\frac{\iota}{n_h^k}}$$

Next, we bound the first two terms respectively. For the first term, we have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\Lambda_{h-1}}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\sqrt{\frac{\nu_h^{\mathrm{ref},k}}{n_h^k}\iota} \\
\le& \sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\Lambda_{h-1}}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\sqrt{\frac{\mathbb{V}\left([P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)\right)}{n_h^k}\iota} \\
&+ \Lambda_{H-1}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\left(\frac{4e^{\beta H}\alpha}{n_h^k} + \frac{6e^{2\beta H}N_0(\alpha)\cdot S}{(n_h^k)^2} + 14e^{2\beta H}\frac{\sqrt{\iota}}{(n_h^k)^{\frac{3}{2}}}\right)\iota} \\
\le& O\Bigg(\max_{s,a,h}\frac{\Lambda_{h-1}\sqrt{\mathbb{V}\left([P_h e^{\beta\cdot V_{h+1}^*}](s,a)\right)}}{[P_h e^{\beta\cdot V_{h+1}^*}](s,a)}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{1}{n_h^k}\iota} \\
&+ \Lambda_{H-1}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\left(\frac{4e^{\beta H}\alpha}{n_h^k} + \frac{6e^{2\beta H}N_0(\alpha)\cdot S}{(n_h^k)^2} + 14e^{2\beta H}\frac{\sqrt{\iota}}{(n_h^k)^{\frac{3}{2}}}\right)\iota}\Bigg) \\
\le& O\Bigg(\max_{s,a,h}\frac{\Lambda_{h-1}\sqrt{\mathbb{V}\left([P_h e^{\beta\cdot V_{h+1}^*}](s,a)\right)}}{[P_h e^{\beta\cdot V_{h+1}^*}](s,a)}\sqrt{HSAT\iota} \\
&+ \Lambda_{H-1}\left(\sum_{s,a,h}\sqrt{N_h^{K+1}(s,a)e^{\beta H}\alpha\iota} + e^{\beta H}HSA\sqrt{N_0(\alpha)\cdot S\iota}\ln(T) + e^{\beta H}(HSA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}\right)\Bigg) \\
\le& O\Bigg(\max_h\{\Lambda_{h-1}\cdot c_{v,h+1}^*\}\sqrt{HSAT\iota} \\
&+ \Lambda_{H-1}\left(\sqrt{e^{\beta H}\alpha HSAT\iota} + e^{\beta H}HSA\sqrt{N_0(\alpha)\cdot S\iota}\ln(T) + e^{\beta H}(HSA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}\right)\Bigg)
\end{aligned} \tag{57}$$

where we utilize the definition that $c_{v,h+1}^* = \max_{s,a}([P_h e^{\beta \cdot V_{h+1}^*}](s,a))^{-1}\sqrt{\mathbb{V}([P_h e^{\beta \cdot V_{h+1}^*}](s,a))}$ in Equation (4). For the second term, since we have that

$$
\begin{aligned}
\check{\nu}_h^k \leq & \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(e^{\beta \cdot V_{h+1}^{\mathrm{ref},\check{l}_i}(s_{h+1}^{\check{l}_i})} - e^{\beta \cdot V_{h+1}^*(s_{h+1}^{\check{l}_i})}\right) \\
\leq & \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(e^{2\beta H} w_{h+1}^{\check{n}_h^k}(s_{h+1}^{\check{l}_i}) + \alpha^2) \leq \frac{1}{\check{n}_h^k} e^{2\beta H} N_0(\alpha)\cdot S + \alpha^2
\end{aligned} \tag{58}
$$

We derive that

$$
\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K}\Lambda_{h-1}\sqrt{\frac{\check{\nu}_h^k}{\check{n}_h^k}\iota} \leq & \Lambda_{H-1}\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\sqrt{\frac{\alpha^2}{\check{n}_h^k}\iota} + \frac{\sqrt{e^{2\beta H} S N_0(\alpha)\iota}}{\check{n}_h^k}\right) \\
\leq & O\left(\Lambda_{H-1}\left(\sqrt{\alpha^2 HSAT\iota} + e^{\beta H} HSA\sqrt{N_0(\alpha)\cdot S\iota}\ln(T)\right)\right)
\end{aligned} \tag{59}
$$

□

F.5 Putting Everything Together

Recall that $\alpha = e^{-\beta H}$ and $\alpha' = \sqrt{HSA}$. Note that $c_{v,H+1}^* = 1$. When T is sufficiently large, we conclude that,

$$
\begin{aligned}
& \frac{1}{\beta}\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{2(h-1)}\Lambda_{h-1}\left(\frac{2b_h^k}{[P_h e^{\beta \cdot V_{h+1}^*}](s_h^k, a_h^k)} + \psi_{h+1}^k + \xi_{h+1}^k + \phi_{h+1}^k\right) \\
\leq & \frac{1}{\beta}\cdot O\Bigg(\max_h\{\Lambda_{h-1}\cdot c_{v,h+1}^*\}\sqrt{HSAT\iota} \\
& \quad + \Lambda_{H-1}\left(\sqrt{e^{\beta H}\alpha HSAT\iota} + \sqrt{\alpha^2 HSAT\iota} + e^{\beta H}(HSA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}\right) \\
& \quad + \Lambda_{H-1}\left(e^{\beta H} HSA\left(N_{p'}^{\alpha'} + \sqrt{N_0(\alpha)\cdot S\iota}\right) + N_0(\alpha)\cdot H^{\frac{3}{2}}S^{\frac{3}{2}}A^{\frac{1}{2}} + \sqrt{HSAT\iota'}\right)\ln(T) \\
& \quad + \Lambda_{H-1}\left(e^{\beta H} N_{p'}^{\sqrt{H}} HSA + \sqrt{HT\iota'} + \sqrt{HSAT\iota'} + \beta\sqrt{H^2 T\iota'}\right)\Bigg) \\
\leq & \frac{1}{\beta}\cdot O\left(\max_{h\in[H]}\left\{\Lambda_{h-1}\cdot c_{v,h+1}^*\right\}\sqrt{\max\{SA,H\}HT\iota'}\ln(T)\right) \\
= & \frac{1}{\beta}\cdot \tilde{O}\left(\max_{h\in[H]}\left\{\Lambda_{h-1}\cdot c_{v,h+1}^*\right\}\sqrt{\max\{SA,H\}HT}\right)
\end{aligned}
$$

which concludes the proof.

G MODIFIED ALGORITHM AND A SKETCH OF REGRET ANALYSIS FOR RISK-AVERSE RL

G.1 UCB-ADVANTAGE FOR RISK-AVERSE RL

The modified algorithm for risk-averse RL is presented in Algorithm 2. We provide the counterparts of Lemmas 1, 2, and 3 as follows, which can be proved similarly.

Lemma 12 (Optimism)**.** *Let $p \in (0,1)$, for any $s,a,h,k \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, with probability at least $1-2(T^3+3)p$, it holds that $Q_h^*(s,a) \le Q_h^k(s,a) \le Q_h^k(s,a)$. Therefore, we have that $V_h^*(s) = \max_a Q_h^*(s,a) \le Q_h^k(s, \arg\max_a Q_h^*(s,a)) \le V_h^k(s)$.*

Lemma 13 (Bounded estimation error)**.** *Conditioned on the successful events of Lemma 12, for any $\gamma \in (0,1]$, with probability $(1-Tp)$ it holds that $\sum_{k=1}^{K} \mathbb{I}[e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot V_h^k(s)} \ge \gamma] \le O\left(e^{4\beta H} H^3 SA/\gamma^2\right)$.*

Lemma 14 (Good reference values)**.** *Conditioned on the successful events of Lemma 12 and Lemma 13, it holds that $e^{\beta\cdot V_h^*(s)} \ge e^{\beta\cdot V_h^{\mathrm{ref},k}(s)} \ge e^{\beta\cdot V_h^*(s)} - \gamma$ if $n_h^k(s) \ge N_0(\gamma) := c_4 e^{4\beta H} H^3 SA\iota/\gamma^2$, where c_4 is a sufficiently large constant for analysis.*

Note that when $\beta < 0$, if $Q_h^k(s,a)$ is an optimistic estimation of $Q_h^*(s,a)$, i.e., $Q_h^k(s,a) \ge Q_h^*(s,a)$, then we have that $e^{\beta\cdot Q_h^k(s,a)} \le e^{\beta\cdot Q_h^*(s,a)}$. Therefore, Lemmas 13 and 14 states that $e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot V_h^k(s)} \ge \gamma$, which is slightly different from Lemmas 2 and 3.

Algorithm 2 UCB-ADVANTAGE FOR RISK-AVERSE RL ($\beta < 0$)

Initialize: $\alpha \leftarrow \min\{e^{-H}, \frac{e^{\beta H}}{4(H+1)}\}$; $\alpha' \leftarrow e^{\beta H}\sqrt{HSA}$; $\iota \leftarrow \ln(2/p)$ where p is the failure probability in Lemmas 1 and 2; risk parameter $\beta < 0$; set all accumulators to 0; $V_h(s) \leftarrow H-h+1$, $Q_h(s,a) \leftarrow H-h+1$, $V_h^{\mathrm{ref}}(s) \leftarrow H-h+1$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$; $V_{H+1} \leftarrow 0$, $V_{H+1}^{\mathrm{ref}} \leftarrow 0$; $\mathcal{L} := \{l_i \mid l_1 = 1, l_i = l_{i-1} + \lfloor (1+1/H)^i \rfloor, i = 2,3,...\}$
for episodes $k \leftarrow 1,2,...,K$ **do**
  Receive s_1
  for $h \leftarrow 1,2,...,H$ **do**
    Take action $a_h \leftarrow \arg\max_a Q_h(s_h,a)$ and observe the next state s_{h+1}
    $n := N_h(s_h,a_h) \xleftarrow{+} 1$, $\check{n} := \check{N}_h(s_h,a_h)) \xleftarrow{+} 1$, and update by rules (9), (10), and (11)
    if $n \in \mathcal{L}$ **then**
      $b_h \leftarrow c_1\sqrt{\frac{\sigma^{\mathrm{ref}}/n - (u^{\mathrm{ref}}/n)^2}{n}\iota} + c_2\sqrt{\frac{\check{\sigma}/\check{n} - (\check{\Delta}/\check{n})^2}{\check{n}}\iota} + c_3\left(\frac{\iota}{n} + \frac{\iota}{\check{n}} + \frac{\iota^{\frac{3}{4}}}{n^{\frac{3}{4}}} + \frac{\iota^{\frac{3}{4}}}{\check{n}^{\frac{3}{4}}}\right)$
      $\bar{b}_h \leftarrow 2\sqrt{\frac{1}{\check{n}}\iota}$
      $\tilde{b}_h \leftarrow 2\sqrt{\frac{1}{n}\iota}$
      $z_h \leftarrow \max\left\{e^{\beta(H-h+1)}, \frac{\check{u}}{\check{n}} - \bar{b}_h, \frac{u^{\mathrm{ref}}}{n} - \tilde{b}_h, \frac{u^{\mathrm{ref}}}{n} + \frac{\check{\Delta}}{\check{n}} - b_h\right\}$
      $Q_h(s_h,a_h) \leftarrow \min\left\{r_h(s_h,a_h) + \frac{1}{\beta}\ln(z_h), Q_h(s_h,a_h)\right\}$
      $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$
      $\check{N}_h(s_h,a_h), \check{\Delta}_h(s_h,a_h), \check{u}_h(s_h,a_h), \check{\sigma}_h(s_h,a_h) \leftarrow 0$
    end if
    if $\sum_a N_h(s_h,a) = N_0(\alpha)$ or $\sum_a N_h(s_h,a) = N_0(\alpha')$ **then**
      $V_h^{\mathrm{ref}}(s_h) \leftarrow V_h(s_h)$
    end if
  end for
end for

G.2 A Sketch of Regret Analysis of Algorithm 2

Next, we provide a sketch of regret analysis. Observe that when $\beta < 0$, we have that

$$\mathrm{Regret}(T) \le \sum_{k=1}^{K}\left(V_1^k(s_1) - V_1^{\pi^k}(s_1)\right) \le \frac{e^{|\beta|H}}{|\beta|}\sum_{k=1}^{K}\left(e^{\beta\cdot V_1^{\pi^k}(s_1)} - e^{\beta\cdot V_1^k(s_1)}\right)$$

Hence, we can derive Term (20) of Theorem 1 by a similar analysis in Appendix C. To derive Term (21) of Theorem 1, we establish the recursive form that is similar to Inequality (18). Note that by Lemma 12, we have that

$$\begin{aligned}\zeta_h^k :=& V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)\\ \le& \frac{1}{\beta}\ln(z_h^k) - \frac{1}{\beta}\ln\left([P_h e^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)\right)\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{|\beta|}\ln\left([P_h e^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)\right)-\frac{1}{|\beta|}\ln\left([P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)\right)+\frac{1}{|\beta|}\ln\left([P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)\right)-\frac{1}{|\beta|}\ln(z_h^k)\\
=&\lambda_{h+1}\left(V_{h+1}^{*}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)\right)-\frac{1}{|\beta|}\left(\ln\left(\frac{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}{e^{\beta\cdot V_{h+1}^{*}(s_{h+1}^k)}}\right)-\ln\left(\frac{[P_h e^{\beta\cdot V_{h+1}^{\pi^k}}](s_h^k,a_h^k)}{e^{\beta\cdot V_{h+1}^{\pi^k}(s_{h+1}^k)}}\right)\right)\\
&+\frac{1}{|\beta|}\ln\left([P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)\right)-\frac{1}{|\beta|}\ln(z_h^k)\\
\le&\lambda_{h+1}\zeta_{h+1}^k-\lambda_{h+1}\delta_{h+1}^k-\frac{1}{|\beta|}\phi_{h+1}^k+\frac{1}{|\beta|}\ln\left([P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)\right)-\frac{1}{|\beta|}\ln(z_h^k)
\end{aligned}
\tag{60}
$$

Let $\kappa_h^k := \mathbb{I}[[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)-z_h^k>\frac{e^{\beta H}}{H+1}]$. Note that when $\kappa_h^k=0$, we have that

$$
\frac{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)-z_h^k}{z_h^k}=\left(\frac{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)-z_h^k}-1\right)^{-1}\le(1+\frac{1}{H})\frac{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)-z_h^k}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}
$$

where the inequality holds by $(x-1)^{-1}\le(1+1/H)x^{-1}$ when $x\ge H+1$. Hence, we obtain that

$$
\begin{aligned}
&\frac{1}{|\beta|}\ln\left([P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)\right)-\frac{1}{|\beta|}\ln(z_h^k)\\
\le&H\cdot\kappa_h^k+(1-\kappa_h^k)\frac{1}{|\beta|}\frac{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)-z_h^k}{z_h^k}\\
\le&H\cdot\kappa_h^k+(1+\frac{1}{H})\frac{1}{|\beta|}\frac{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)-z_h^k}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}\\
\le&H\cdot\kappa_h^k+(1+\frac{1}{H})\frac{1}{|\beta|}\frac{\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}[P_h(e^{\beta\cdot V_{h+1}^{*}}-e^{\beta\cdot V_{h+1}^{\check{l}_i}})](s_h^k,a_h^k)-\psi_{h+1}^k+2b_h^k}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}
\end{aligned}
\tag{61}
$$

where the first inequality holds by $\ln(x)-\ln(y)\le\frac{x-y}{y}, x\ge y>0$. Further, by the exact analysis in deriving Inequality (17), we obtain that

$$
\begin{aligned}
&\frac{1}{|\beta|}\sum_{k=1}^{K}\frac{\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}[P_h(e^{\beta\cdot V_{h+1}^{*}}-e^{\beta\cdot V_{h+1}^{\check{l}_i}})](s_h^k,a_h^k)}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}\\
=&\frac{1}{|\beta|}\sum_{k=1}^{K}\frac{\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(e^{\beta\cdot V_{h+1}^{*}(s_{h+1}^k)}-e^{\beta\cdot V_{h+1}^{\check{l}_i}(s_{h+1}^k)})}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}+\frac{1}{\beta}\sum_{k=1}^{K}\xi_{h+1}^k\\
\le&\frac{\lambda_{h+1}}{\check{n}_h^k}\sum_{k=1}^{K}\sum_{i=1}^{\check{n}_h^k}\frac{e^{\beta\delta_{h+1}^{\check{l}_i}}-1}{\beta\delta_{h+1}^{\check{l}_i}}\delta_{h+1}^{\check{l}_i}+\frac{1}{\beta}\sum_{k=1}^{K}\xi_{h+1}^k\\
\le&(1+\frac{1}{H})\lambda_{h+1}\frac{e^{\alpha}-1}{\alpha}\sum_{k=1}^{K}\delta_{h+1}^k+(1+\frac{1}{H})\lambda_{h+1}\frac{e^{\beta H}-1}{\beta}\sum_{k=1}^{K}\mathbb{I}[n_h^k<N_0(\alpha)]+\frac{1}{\beta}\sum_{k=1}^{K}\xi_{h+1}^k
\end{aligned}
\tag{62}
$$

where $\alpha=\min\{e^{-H},\frac{e^{\beta H}}{4(H+1)}\}$ is an input of Algorithm 2. Hence, combining Inequalities (60), (61), and (62) yields

$$
\begin{aligned}
\sum_{k=1}^{K}\zeta_h^k\le&H\cdot\sum_{k=1}^{K}\kappa_h^k+(1+\frac{1}{H})\lambda_{h+1}\frac{e^{\beta H}-1}{\beta}\sum_{k=1}^{K}\mathbb{I}[n_h^k<N_0(\alpha)]\\
&+\lambda_{h+1}\sum_{k=1}^{K}\zeta_{h+1}^k+\left((1+\frac{1}{H})^2-1\right)\lambda_{h+1}\sum_{k=1}^{K}\delta_{h+1}^k\\
&-(1+\frac{1}{H})\frac{1}{\beta}\sum_{k=1}^{K}\frac{2b_h^k-\psi_{h+1}^k}{[P_h e^{\beta\cdot V_{h+1}^{*}}](s_h^k,a_h^k)}+\frac{1}{\beta}\sum_{k=1}^{K}\phi_{h+1}^k+\frac{1}{\beta}\sum_{k=1}^{K}\xi_{h+1}^k
\end{aligned}
$$

where we use the fact that $\frac{e^{\alpha}-1}{\alpha} \le 1+\frac{1}{H}$ (See Footnote 6 in the main paper). Note that $\zeta_h^k \ge \delta_{h+1}^k$ for any $(h,k) \in [H]\times[K]$. Therefore, we derive

$$\begin{aligned}\sum_{k=1}^{K}\zeta_h^k \le& H\cdot\sum_{k=1}^{K}\kappa_h^k + (1+\frac{1}{H})\lambda_{h+1}\frac{e^{\beta H}-1}{\beta}\sum_{k=1}^{K}\mathbb{I}[n_h^k < N_0(\alpha)] \\ &+ (1+\frac{1}{H})^2\lambda_{h+1}\sum_{k=1}^{K}\zeta_{h+1}^k + \frac{1}{\beta}\sum_{k=1}^{K}\left(\phi_{h+1}^k + \xi_{h+1}^k - (1+\frac{1}{H})\frac{2b_h^k - \psi_{h+1}^k}{[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k)}\right)\end{aligned} \tag{63}$$

which is the counterpart of the recursive form (18) for Algorithm 2. Following a similar analysis in the proof of Lemma 4, we can derive Term (21) for risk-averse RL. It remains to show that $\sum_{k=1}^{K}\kappa_h^k$ can be bounded by a constant. Note that by line 10 of Algorithm 2, we have that

$$z_h^k \ge \frac{u_h^{\mathrm{ref},k}}{n_h^k} - \tilde{b}_h$$

where $u_h^{\mathrm{ref},k} = \sum_{i=1}^{n_h^k} e^{\beta\cdot V_{h+1}^{\mathrm{ref},l_i}(s_{h+1}^{l_i})}$. Hence, we derive that

$$\begin{aligned}&[P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k) - z_h^k \\ \le& [P_h e^{\beta\cdot V_{h+1}^*}](s_h^k,a_h^k) - \frac{u_h^{\mathrm{ref},k}}{n_h^k} + \tilde{b}_h \\ \le& \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}[P_h(e^{\beta\cdot V_{h+1}^*} - e^{\beta\cdot V_{h+1}^{\mathrm{ref},l_i}})](s_h^k,a_h^k) + \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}[(P_h-\widehat{P}_h)e^{\beta\cdot V_{h+1}^{\mathrm{ref},l_i}}](s_h^k,a_h^k) + \tilde{b}_h \\ \le& \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}[P_h(e^{\beta\cdot V_{h+1}^*} - e^{\beta\cdot V_{h+1}^{\mathrm{ref},l_i}})](s_h^k,a_h^k) + \tilde{O}\left(\sqrt{\frac{1}{n_h^k}}\right) \\ \le& \frac{(1-e^{\beta H})N_p^{\alpha} + \alpha\cdot(n_h^k - N_p^{\alpha})}{n_h^k} + \tilde{O}\left(\sqrt{\frac{1}{n_h^k}}\right)\end{aligned} \tag{64}$$

where $\alpha = \min\{e^{-H}, \frac{e^{\beta H}}{4(H+1)}\}$ and N_p^{α} is a constant defined in Equation (48). Let N denote the minimum n_h^k such that both the first and the second term of Inequality (64) is smaller than $\frac{e^{\beta H}}{2(H+1)}$, which is also a constant. Therefore, $\sum_{k=1}^{K}\kappa_h^k \le \mathbb{I}[n_h^k \le N] \le NSA$.

H PROOF OF THE LOWER BOUND

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

$$\mathrm{Regret}(T) \ge \Omega\left(\frac{e^{\frac{|\beta|(H-1)}{2}}-1}{|\beta|}\sqrt{SAT}\right)$$

Proof. We first note that the key to proving the generalized information-theoretic lower bound (3) is the following lemma, which is the counterpart of (Osband and Roy, 2016, Theorem 1) for risk-sensitive multi-armed bandit (MAB). In fact, it corresponds to the hard instance in (Fei et al., 2020, Figure 2) in the proof of (Fei et al., 2020, Theorem 3) for arbitrary A.

Lemma 15. *Let* sup *be the supermum over all distributions of rewards such that for each $a = 1,\dots,A$ the rewards $r(1)_t,\dots,r(A)_t$ are i.i.d. and let* inf *be the infimum over all reinforcement learning algorithms. Then*

$$\inf\sup\left(\max_a v(a)K - \mathbb{E}\left[\sum_{t=1}^{K}v(\tilde{a}_t)\right]\right) \ge \frac{1}{72}\frac{e^{\frac{|\beta|\cdot(H-1)}{2}}-1}{|\beta|}\sqrt{AK}$$

where $v(a) = \frac{1}{\beta}\ln\{\mathbb{E}_{r(a)}[e^{\beta\cdot r(a)}]\}$.

Proof. Since the proof for $\beta > 0$ and $\beta < 0$ is similar, we focus on the case $\beta > 0$. We consider a A-armed bandit where all arms are i.i.d. $(H-1)\cdot \mathrm{Ber}(\delta)$, but one arm a^* is i.i.d. $(H-1)\cdot \mathrm{Ber}(\delta+\epsilon)$ for some $\delta, \epsilon > 0$. We define an auxiliary $\tilde{r}_t(a) = r_t(a)$ for all $a \neq a^*$, but with the rewards of the action a^* replaced by the draw $\tilde{r}_t \sim \mathrm{Ber}(\delta)$. We consider an auxiliary sequence of actions $\tilde{a}_t \sim \pi_t(\tilde{H}_t)$ for $\tilde{H}_t = (\tilde{a}_1, \tilde{r}_1, ..., \tilde{a}_{t-1}, \tilde{r}_{t-1})$ as the history generated by an agent with no feedback informing them about a^*. Let $n_K(a) := |\{a_t = a | t = 1, ..., K\}|$ and $\tilde{n}_K(a) := |\{\tilde{a}_t = a | t = 1, ..., K\}|$ denote the number of times arm a has been selected by time K under a_t and $\tilde{a}_t$, respectively. The following lemma is a counterpart of (Osband and Roy, 2016, Lemma 1) for risk-sensitive MAB.

Lemma 16. *(Regret of an Uninformed Agent). For all $\delta, \epsilon > 0$ and all learning algorithms π, it holds that*

$$\max_a v(a)K - \mathbb{E}\left[\sum_{t=1}^{K} v(\tilde{a}_t)\right] \leq \frac{A-1}{A} K\epsilon'$$

where

$$v(a) = \begin{cases} \frac{1}{\beta}\ln\left(\delta e^{\beta\cdot(H-1)} + (1-\delta)\right), & \text{if } a \neq a^* \\ \frac{1}{\beta}\ln\left((\delta+\epsilon) e^{\beta\cdot(H-1)} + (1-\delta-\epsilon)\right), & \text{if } a = a^* \end{cases}$$

$$\epsilon' = \frac{1}{\beta}\ln\left(\frac{(\delta+\epsilon)e^{\beta\cdot(H-1)} + (1-\delta-\epsilon)}{\delta e^{\beta\cdot(H-1)} + (1-\delta)}\right)$$

Proof. We have that

$$\begin{aligned} \max_a v(a)K - \mathbb{E}\left[\sum_{t=1}^{K} v(\tilde{a}_t)\right] &= \mathbb{E}\left[\sum_{a\neq a^*} \tilde{n}_K(a)\epsilon'\right] \\ &= \epsilon'(K - \tilde{n}_K(a^*)) \\ &= \epsilon' K\left(1 - \frac{1}{A}\right) \end{aligned}$$

where the last equation follows from a symmetry argument, since a^* is independent of $\tilde{n}_t(a)$ for all actions a, which concludes the proof. ☐

We now establish that, if ϵ is sufficiently small, then over a limited time horizon the distributions of $\tilde{r}_t(a_t)$ cannot be significantly different from the outcomes $r_t(a_t)$. We compare the conditional distributions over the choice of action P with the choice of action $\tilde{P}$ which would have arisen under the uninformative data $\tilde{H}_t$. To be more precise we define $P(z_t^K|H_t) := \mathbb{P}(r_t^K = z_t^K|H_t)$ and $\tilde{P}(z_t^K|\tilde{H}_t) := \mathbb{P}(\tilde{r}_t^K = z_t^K|\tilde{H}_t)$, where we denote by $r_t^K := (r_t(a_t)), ..., r_K(a_t))$ the sequence of rewards from time t to K and similarly for $\tilde{r}_t^K$. To quantify the difference between two distributions we utilize the following KL divergence

$$d_{KL}\left(\tilde{P}(z_t^K|\tilde{H}_t), P(z_t^K|H_t)\right) = \mathbb{E}\left[\sum_{z_t^K} \tilde{P}(z_t^K|\tilde{H}_t)\ln\left(\frac{\tilde{P}(z_t^K|\tilde{H}_t)}{P(z_t^K|H_t)}\right)\right]$$

By (Osband and Roy, 2016, Lemma 3) and through a simple substitution in Lemma 15, for all $\delta, \epsilon > 0$ and all learning algorithms π, it holds that

$$\max_a v(a)K - \mathbb{E}\left[\sum_{t=1}^{K} v(\tilde{a}_t)\right] \geq \epsilon' K\left(1 - \frac{1}{A} - \sqrt{\frac{1}{2}d_{KL}\left(\tilde{P}(z_t^K|\tilde{H}_t), P(z_t^K|H_t)\right)}\right) \tag{65}$$

Further, combining (Osband and Roy, 2016, Proposition 1) and Inequality (65), we have that

$$\max_a v(a)K - \mathbb{E}\left[\sum_{t=1}^{K} v(\tilde{a}_t)\right]$$

$$
\begin{aligned}
&\geq \epsilon' K \left(1 - \frac{1}{A} - \sqrt{\frac{\epsilon^2}{2\delta}\frac{K}{A}}\right) \text{ for all } \epsilon \\
&\geq \frac{1}{2\beta}\frac{\epsilon(e^{\beta\cdot(H-1)}-1)}{\delta(e^{\beta\cdot(H-1)}-1)+1} K \left(1 - \frac{1}{A} - \sqrt{\frac{\epsilon^2}{2\delta}\frac{K}{A}}\right) \text{ for } \epsilon \leq \delta \\
&\geq \frac{1}{2\beta}\frac{\epsilon(e^{\beta\cdot(H-1)}-1)}{3} K \left(1 - \frac{1}{A} - \sqrt{\frac{\epsilon^2}{2\delta}\frac{K}{A}}\right) \text{ by setting } \delta = e^{-\beta\cdot(H-1)} \\
&\geq \frac{e^{\beta\cdot(H-1)}-1}{6\beta}\sqrt{\frac{\delta A}{8K}} K \left(1 - \frac{1}{A} - \frac{1}{4}\right) \text{ by setting } \epsilon^2 = \frac{\delta A}{8K} \\
&\geq \frac{1}{72}\frac{e^{\frac{\beta\cdot(H-1)}{2}}-1}{\beta}\sqrt{AK}
\end{aligned} \tag{66}
$$

where the second inequality follows from the fact that $\ln(1+x) \geq x/2$ for $x \in [0,1]$. Therefore, we conclude the proof. $\square$

Next, we extend the Lemma 15 from MAB to reinforcement learning with $S \geq 2$. Consider a finite-horizon MDP that starts from state s_0. The agent ends up in states 1 to S with equal probability, independent of the action. At each such state $i = 1, ..., S$, the agent faces the hard instance constructed in the proof of Lemma 15. Since the expected number of times of visiting each state i is K/S, we derive the counterpart of Inequality (66) in the following,

$$
\begin{aligned}
&\sum_i \max_a v(a)\frac{K}{S} - \mathbb{E}\left[\sum_{t=1}^{K} v(\tilde{a}_t)\right] \\
&\geq \epsilon' K \left(1 - \frac{1}{A} - \sqrt{\frac{\epsilon^2}{2\delta}\frac{K}{SA}}\right) \\
&\geq \frac{e^{\beta\cdot(H-1)}-1}{6\beta}\sqrt{\frac{\delta SA}{8K}} K \left(1 - \frac{1}{A} - \frac{1}{4}\right) \text{ by setting } \delta = e^{-\beta\cdot(H-1)} \text{ and } \epsilon^2 = \frac{\delta SA}{8K} \\
&\geq \frac{1}{72}\frac{e^{\frac{\beta\cdot(H-1)}{2}}-1}{\beta}\sqrt{SAK}
\end{aligned}
$$

That is, the regret of interacting with this MDP for K episodes is lower bounded by

$$
\Omega\left(\frac{e^{\frac{|\beta|(H-1)}{2}}-1}{|\beta|}\sqrt{SAK}\right)
$$

Further, since the transition kernel is timestep-dependent by definition, i.e., $P_1, P_2, ..., P_H$ may not be the same. We augment the state from S to be HS as in the proof of (Jin et al., 2018b, Theorem 3). Recall that $T := KH$, we have that

$$
\text{Regret}(T) \geq \Omega\left(\frac{e^{\frac{|\beta|(H-1)}{2}}-1}{|\beta|}\sqrt{SAT}\right)
$$

which concludes the proof. $\square$