Cascaded Gaps: Towards Logarithmic Regret for Risk-Sensitive Reinforcement Learning

Yingjie Fei [1] **Ruitu Xu** [2]

Abstract

In this paper, we study gap-dependent regret guarantees for risk-sensitive reinforcement learning based on the entropic risk measure. We propose a novel definition of sub-optimality gaps, which we call cascaded gaps, and we discuss their key components that adapt to the underlying structures of the problem. Based on the cascaded gaps, we derive non-asymptotic and logarithmic regret bounds for two model-free algorithms under episodic Markov decision processes. We show that, in appropriate settings, these bounds feature exponential improvement over existing ones that are independent of gaps. We also prove gap-dependent lower bounds, which certify the near optimality of the upper bounds.

1. Introduction

We study the problem of risk-sensitive reinforcement learning (RL) based on the entropic risk measure, in which we aim to identify a decision making rule (or policy) $\widehat{\pi}$ that solves the following optimization problem:

$$\max_{\pi} \left\{ V^{\pi} = \frac{1}{\beta} \log(\mathbb{E}_{\pi}\, e^{\beta R}) \right\}, \tag{1.1}$$

where R denotes the cumulative reward and $\beta \neq 0$ is the risk parameter that induces risk-seeking learning when $\beta > 0$ and risk-averse learning when $\beta < 0$. The (standard) risk-neutral objective function used in RL, which is simply $\mathbb{E}_{\pi}[R]$, can be recovered from Eq. (1.1) by setting $\beta \to 0$. Moreover, the objective of (1.1) in the form of entropic risk measure admits a Taylor expansion $V^{\pi} = \mathbb{E}_{\pi}[R] + \frac{\beta}{2}\mathrm{Var}_{\pi}(R) + O(\beta^2)$, which represents a trade-off between the expectation and the variance (and possibly higher-order statistics) of the reward. Several lines of research on related problems have witnessed fruitful applications in a wide range of domains, including neuroscience (Niv et al., 2012; Shen et al., 2014), robotics (Nass et al., 2019; Williams et al., 2016; 2017), economics (Hansen and Sargent, 2011), etc. The formulation (1.1) has been related to notions of robustness (Osogami, 2012; Hansen and Sargent, 2011; Föllmer and Knispel, 2011) and bounded rationality (Simon, 1955; Ortega and Stocker, 2016) in decision making and behavioral studies. A thermodynamic view on such formulation has also been proposed for understanding sequential decision making systems (Ortega and Braun, 2013).

For problem (1.1), much recent work has been devoted to designing algorithms that attain finite-sample regret bounds under Markov decision processes (MDPs). Although the existing bounds are nearly optimal in the minimax sense, they are overly pessimistic as they generally fail to exploit particular structures of the underlying MDPs, such as sub-optimality gaps, which quantify the easiness of learning optimal policies under the MDPs. On the other hand, while previous work has explored and provided gap-dependent results for risk-neutral RL, it is unclear how the sub-optimality gaps should be constructed in the risk-sensitive setting. In particular, the definition of existing sub-optimality gaps, as we will elaborate in Section 4, crucially hinges on the linear structures of the risk-neutral setting, which no longer hold in the risk-sensitive setting characterized by the non-linear objective (1.1). It therefore begs the following natural questions: 1) how sub-optimality gaps should be characterized in risk-sensitive RL, and 2) whether we can obtain refined bounds on regret and sample complexity by taking advantage of the gap structures.

To answer the above questions, we study gap-dependent regret bounds for risk-sensitive RL based on the entropic risk measure. We begin by identifying two key conditions for a proper definition of sub-optimality gaps for risk-sensitive RL: Bellman difference condition and risk consistency condition. The Bellman difference condition states that the gaps induce a Bellman equation in which they play the role of reward functions; the risk consistency condition stipulates that the gaps stay on the same order of magnitude for

*Alphabetical order [1]Bloomberg, New York, USA [2]Department of Statistics and Data Science, Yale University, USA. Correspondence to: Yingjie Fei <yf275@cornell.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

both risk-averse and risk-seeking settings given fixed risk sensitivity $|\beta|$, and they reduce to risk-neutral gaps as $|\beta|$ vanishes. Motivated by the two conditions, we propose a novel characterization of sub-optimality gaps for risk-sensitive RL, which we call *cascaded gaps*. Cascaded gaps consist of three key components: 1) the difference of rewards along trajectories controlled by an optimal policy, 2) the reward functions evaluated along a free trajectory (not controlled by any policy), and 3) a normalization factor that depends on the risk parameter. The first two components together exhibit a cascading property and address the Bellman difference condition, whereas the third component facilitates risk consistency.

Based on the cascaded gaps, we derive non-asymptotic regret bounds for two existing risk-sensitive RL algorithms, RSVI2 and RSQ2, that scale logarithmically in the number of episodes and decay in the cascaded gaps. The proof is based on a unified framework for both algorithms. We demonstrate that under proper settings, our regret bounds attain an exponential improvement over existing results with respect to the number of episodes, as well as an exponential improvement in terms of risk sensitivity and horizon over existing sample complexity bounds. We further show that the provided upper bounds are nearly optimal by deriving compatible lower bounds. To the best of our knowledge, this is the first work that studies sub-optimality gaps in risk-sensitive RL with the entropic risk measure and derives gap-dependent regret bounds.

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

1. We propose a novel notion of sub-optimality gaps for risk-sensitive RL based on the entropic risk measure, which we call cascaded gaps. We discuss essential components of cascaded gaps tailored to the unique structure of risk-sensitive RL, and compare them with sub-optimality gaps in the risk-neutral setting.

2. We prove logarithmic regret bounds that adapt to the sub-optimality gaps for two existing risk-sensitive RL algorithms. The bounds are achieved via a unified framework for both algorithms, and they imply exponential improvements in both regret and sample complexity under appropriate settings.

3. We further derive lower bounds that nearly match the upper bounds, thereby showing that the upper bounds are nearly optimal.

Notation. We write shorthand $[n] := 1, \dots, n$ for any $n \in \mathbb{Z}_+$. For any series of variables $\{v_i\}_{i\in[n]}$, we define the notation $\mathrm{poly}(v_1, \dots, v_n) := c_0 \prod_{i\in[n]} v_i^{c_i}$ and $\mathrm{polylog}(v_1, \dots, v_n) := c_0 \prod_{i\in[n]} \log(v_i)^{c_i}$ for some positive universal constants $\{c_i\}_{i\geq 0}$. For $x > 0$, we write $\widetilde{O}(x)$ to denote $O(x\,\mathrm{polylog}(x))$; we define $\widetilde{\Omega}(x)$ in a similar way. Unless otherwise specified, $\log$ denotes the natural logarithm and $\log_2$ denotes the logarithm with base 2. For any functions f and g with the same domain, we write $f \leq g$ to mean $f(y) \leq g(y)$ for all y in the domain. We use notation $\phi(n) \lesssim \varphi(n)$ (or $\phi(n) \gtrsim \varphi(n)$) for functions ϕ and φ that satisfy $\phi(n) \leq C\varphi(n)$ (or $\phi(n) \geq c\varphi(n)$) for every $n \in \mathbb{Z}_+$ with some universal constant $C > 0$ (or $c > 0$); further, we write $\phi(n) \asymp \varphi(n)$ to mean $\phi(n) \lesssim \varphi(n)$ and $\phi(n) \gtrsim \varphi(n)$.

2. Related Works

Initiated by Howard and Matheson (1972); Jacobson (1973), risk-sensitive RL based on the entropic risk measure has been the focus of long-standing research efforts for the past decades (Eriksson and Dimitrakakis, 2019; Borkar, 2002; Borkar and Meyn, 2002; Coraluppi and Marcus, 1999; Osogami, 2012; Shen et al., 2013; Whittle, 1990; Mihatsch and Neuneier, 2002; Bäuerle and Rieder, 2014; Cavazos-Cadena and Fernández-Gaucherand, 2000; Fleming and McEneaney, 1995; Hernández-Hernández and Marcus, 1996; Di Masi and Stettner, 1999; Borkar, 2001). Most related to our work are perhaps those by Fei et al. (2020a; 2021b;a): under the episodic and finite-horizon MDPs, they propose computationally efficient algorithms for risk-sensitive RL and provide finite-sample and nearly optimal regret guarantees in both tabular and linear settings. These results are general, holding without access to transitions or simulators. However, they fail to exploit particular structures of the underlying MDPs, such as sub-optimality gaps, and are therefore overly conservative under certain settings.

For risk-neutral RL, a series of works has established non-asymptotic and gap-dependent regret bounds for optimistic algorithms, starting from Simchowitz and Jamieson (2019). Specifically, logarithmic regret are derived for optimistic Q-learning (Yang et al., 2021) and value iteration (He et al., 2021). Despite these recent developments, it remains unclear whether the defintion of sub-optimality gaps in risk-neutral RL is appropriate for the risk-sensitive setting, which the present work aims to address.

3. Preliminaries

3.1. Episodic and Finite-Horizon MDPs

We focus on the setting of tabular MDPs, represented by a tuple $(\mathcal{S}, \mathcal{A}, H, K, \mathcal{P}, r)$. Here, $\mathcal{S}$ denotes the set of available states with cardinality $|\mathcal{S}| = S$, $\mathcal{A}$ the set of actions available to the agent with cardinality $|\mathcal{A}| = A$, K the number of episodes, H the horizon, $\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]}$ the set of transition kernels, and $r = \{r_h\}_{h\in[H]}$ the set of reward functions. We assume that reward $r_h : \mathcal{S}\times\mathcal{A} \to [0,1]$ is deterministic for every step h. Without loss of generality, the agent starts at a fixed state $s_1^k = s_1$ in each episode $k \in [K]$. For episode $k \in [K]$ and step $h \in [H]$, it takes action a_h^k at state s_h^k and receives reward $r_h^k(s_h^k, a_h^k)$. Then the environment transitions into s_{h+1}^k with probability equal to $\mathcal{P}_h(s_{h+1}^k|s_h^k, a_h^k)$.

3.2. Risk-Sensitive RL

We define policy $\pi := \{\pi_h : \mathcal{S} \to \mathcal{A}\}$ as a collection of functions that map states to actions. In risk-sensitive RL based on entropic risk measure, we define the state-value function with respect to any π:

$$V_h^\pi(s) := \frac{1}{\beta}\log\left\{\mathbb{E}\left[e^{\beta\sum_{i=h}^H r_i(s_i,\pi_i(s_i))}\right] \mid s_h = s\right\},$$

for each $h \in [H]$ and $s \in \mathcal{S}$, where the expectation is taken over the transition kernel $\mathcal{P}$. The quantity $\beta \neq 0$ is the risk parameter of the entropic risk measure. In particular, $\beta > 0$ yields a risk-seeking value function, while $\beta < 0$ induces a risk-averse value function. The risk-neutral definition of the value function $\widetilde{V}_h^\pi(s) := \mathbb{E}[\sum_{i=h}^H r_i(s_i,\pi_i(s_i)) \mid s_h = s]$ can be recovered through taking $\beta \to 0$. Similarly, we define the corresponding action-value function as

$$\begin{aligned} &Q_h^\pi(s,a) \\ &\quad := \frac{1}{\beta}\log\left\{\mathbb{E}\left[e^{\beta\sum_{i=h}^H r_i(s_i,\pi_i(s_i))}\right] \mid s_h = s, a_h = a\right\}. \end{aligned}$$

Note that we omit the dependency of V_h^π and Q_h^π on β for simplicity. Consequently, the Bellman equation for risk-sensitive RL is given by

$$Q_h^\pi(s,a) = r_h(s,a) + \frac{1}{\beta}\log\mathbb{E}_{s'\sim\mathcal{P}_h(\cdot\mid s,a)}\left[e^{\beta\cdot V_{h+1}^\pi(s')}\right], \tag{3.1}$$

which relates the action-value function Q_h^π to the state-value function V_{h+1}^π of the next step. Note that the Bellman equation is non-linear in the value function due to the non-linearity of the entropic risk measure. It can be shown that there always exists an optimal policy π^* with the optimal value $V_h^*(s) := V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(s)$ for every $h \in [H]$ and $s \in \mathcal{S}$; we also write $Q_h^* := Q_h^{\pi^*}$ for $h \in [H]$.

Under episodic MDPs, the agent aims to learn an optimal policy π^* by interacting with the environment for K episodes. We measure the performance of the agent that follows policies $\{\pi^k\}_{k\in[K]}$ via the notion of regret, which is defined as

$$\mathcal{R}(K) := \sum_{k\in[K]}(V_1^* - V_1^{\pi^k})(s_1^k).$$

4. Cascaded Gaps

4.1. Bellman Difference Condition

Since both regret and sub-optimality gaps represent some notion of sub-optimality with respect to an optimal policy π^*, it would be instrumental to associate the two through a unified lens. We do so by introducing the following condition, which later plays a key role in our analysis.

Condition 4.1 (Bellman Difference Condition)**.** *Let $\{\widehat{V}_h^\pi\}_h$ be some value functions and $\widehat{\pi}^*$ be a corresponding optimal policy. We say that gap functions $\{\mathrm{gap}_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}\}_{h\in[H]}$ satisfy the Bellman difference condition if, for any policy π and tuple $(h,s) \in [H]\times\mathcal{S}$, there exists some map f and $Z_h^\pi := f(\widehat{V}_h^\pi)$ such that*

$$D_h^\pi(s) = \mathrm{gap}_h(s,a) + \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[D_{h+1}^\pi(s')],$$

where $D_h^\pi := Z_h^{\widehat{\pi}^*} - Z_h^\pi$ and $a := \pi_h(s)$.

Condition 4.1 stipulates that for any fixed policy π, the gaps induce a form of Bellman equation where the action follows policy π. The function D_h^π, itself being the difference of two functionals with respect to $\widehat{\pi}^*$ (optimal with respect to $\{\widehat{V}_h^\pi\}_h$) and π, takes the role of the value function, and the gap takes the role of the reward function. Indeed, Condition 4.1 associates the sub-optimality induced by D_h^π with that embedded in gap_h. The condition also suggests that when $\pi = \widehat{\pi}^*$, we have $D_h^{\widehat{\pi}^*} = 0$ and therefore $\mathrm{gap}_h(s, \widehat{\pi}_h^*(s)) = 0$.

As an example, we show that the sub-optimality gaps defined in risk-neutral RL meets Condition 4.1. Recall the risk-neutral value functions

$$\begin{aligned} \widetilde{Q}_h^\pi(s,a) &:= \mathbb{E}\left[\sum_{i=h}^H r_i(s_i,\pi_i(s_i)) \mid s_h = s, a_h = a\right], \\ \widetilde{V}_h^\pi(s) &:= \widetilde{Q}_h^\pi(s,\pi_h(s)), \end{aligned}$$

for $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ and policy π, with $\widetilde{Q}_h^*$ and $\widetilde{V}_h^*$ being the corresponding optimal value functions. In existing literature, the sub-optimality gaps for risk-neutral RL are given by

$$\widetilde{\Delta}_h(s,a) := \widetilde{V}_h^*(s) - \widetilde{Q}_h^*(s,a), \tag{4.1}$$

for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ (Simchowitz and Jamieson, 2019; Yang et al., 2021; He et al., 2021). Note that the gap in Eq. (4.1) computes the difference in values between the optimal action $\pi^*(s)$ and action a. As stated and proved below, it satisfies Condition 4.1 in the risk-neutral setting.

Proposition 4.2. *The sub-optimality gaps $\{\widetilde{\Delta}_h\}_{h\in[H]}$ for risk-neutral RL satisfy Condition 4.1 with $Z_h^\pi := \widetilde{V}_h^\pi$.*

Proof. Recall that in the risk-neutral setting, the Bellman equation for any policy π is given by

$$\widetilde{Q}_h^\pi(s,a') = r_h(s,a') + \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a')}[\widetilde{V}_{h+1}^\pi(s')] \quad (4.2)$$

for any $(h,s,a') \in [H]\times\mathcal{S}\times\mathcal{A}$. We fix a tuple (h,s,a) where $a=\pi_h(s)$, and let $Z_h^\pi := \widetilde{V}_h^\pi$ so that $D_h^\pi = \widetilde{V}_h^* - \widetilde{V}_h^\pi$ in Condition 4.1. From the definition (4.1) of $\widetilde{\Delta}_h$, we have

$$\begin{aligned}
\widetilde{\Delta}_h(s,a) &= \widetilde{V}_h^*(s) - \widetilde{Q}_h^*(s,a)\\
&= \widetilde{V}_h^*(s) - \widetilde{V}_h^\pi(s) + \widetilde{V}_h^\pi(s) - \widetilde{Q}_h^*(s,a)\\
&= \widetilde{V}_h^*(s) - \widetilde{V}_h^\pi(s) + \widetilde{Q}_h^\pi(s,a) - \widetilde{Q}_h^*(s,a)\\
&\overset{(i)}{=} \widetilde{V}_h^*(s) - \widetilde{V}_h^\pi(s)\\
&\quad + \Big[r_h(s,a) + \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[\widetilde{V}_{h+1}^\pi(s')]\Big]\\
&\quad - \Big[r_h(s,a) + \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[\widetilde{V}_{h+1}^*(s')]\Big]\\
&= \widetilde{V}_h^*(s) - \widetilde{V}_h^\pi(s)\\
&\quad - \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[\widetilde{V}_{h+1}^*(s') - \widetilde{V}_{h+1}^\pi(s')]\\
&= D_h^\pi(s) - \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[D_{h+1}^\pi(s')],
\end{aligned}$$

where step (i) is due to the Bellman equation (4.2). $\square$

Given Proposition 4.2, a connection between regret and sub-optimality gaps can be established: since regret in the risk-neutral setting is defined as $\widetilde{\mathcal{R}}(K) := \sum_{k\in[K]}(\widetilde{V}_1^* - \widetilde{V}_1^{\pi^k})(s_1^k)$, we have $\widetilde{\mathcal{R}}(K) = \sum_{k\in[K]} D_1^{\pi^k}(s_1^k)$ (with $D_h^{\pi^k}$ as implied in Proposition 4.2). In words, the regret can be written as the sum of $\{D_1^{\pi^k}\}$ defined for the Bellman difference condition, in which sub-optimality gaps take the role of rewards.

The proof of Proposition 4.2 exploits the Bellman equations under the risk-neutral setting and, in particular, the linearity of $\widetilde{Q}_h^\pi$ in terms of r_h and $\widetilde{V}_{h+1}^\pi$. However, such linear properties are not available in risk-sensitive RL, as seen in Eq. (3.1), where the non-linearity is induced by the entropic risk measure $X \mapsto \frac{1}{\beta}\log(\mathbb{E}[e^{\beta X}])$. This suggests that a simple definition of sub-optimality gaps such as Eq. (4.1) may not be appropriate, and an alternative definition is necessary.

4.2. Cascading Structure

To introduce sub-optimality gaps for risk-sensitive RL, we need a few additional notations. We denote by τ a trajectory of length H, which is a series of state-action pairs $\{(s_j,a_j)\}_{j\in[H]}$, and we let $\mathcal{T}$ be the set of all possible trajectories. For any trajectory $\tau\in\mathcal{T}$ and $h\in[H]$, we let τ_h denote the trajectory that consists of the first h elements in τ, and we define the set $\mathcal{T}_h := \{\tau_h : \tau\in\mathcal{T}\}$. Note that $\tau_H = \tau$ and $\mathcal{T}_H = \mathcal{T}$. We also let τ_0 be an empty trajectory and $\mathcal{T}_0 := \{\tau_0\}$. We further define the cumulative reward function R on trajectories such that $R(\tau_0) := 0$ and $R(\tau_h) := \sum_{j\in[h]} r_j(s_j,a_j)$ for $h\in[H]$ and $\tau\in\mathcal{T}$.

Motivated by the discussion in Section 4.1, we propose the following definition of sub-optimality gaps for risk-sensitive RL. For any step h and trajectory τ, we let

$$\begin{aligned}
&\Delta_{h,\beta}(s,a;\tau_{h-1})\\
&\quad := \psi_\beta \cdot e^{\beta\cdot R(\tau_{h-1})}\cdot\big[e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot Q_h^*(s,a)}\big],
\end{aligned} \quad (4.3)$$

where $\psi_\beta := 1/\beta$ for $\beta>0$ and $\psi_\beta := e^{-\beta H}/\beta$ for $\beta<0$. Here, we slightly abuse the gap definition by augmenting it with additional dependency on β and τ_{h-1} (based on Condition 4.1). Note that $\Delta_{h,\beta}\ge 0$ for any $\beta\ne 0$.

Let us remark on several noteworthy properties of this gap definition. First, in contrast with $\widetilde{\Delta}_h$ defined in Eq. (4.1) for the risk-neutral setting, which only depends on π^* and a single state-action pair (s,a) at step h, the gap $\Delta_{h,\beta}$ defined in Eq. (4.3) additionally depends on the trajectory prior to step h. Specifically, the gap consists of two components: the factor $e^{\beta\cdot R(\tau_{h-1})}$, which is with respect to an uncontrolled trajectory τ_{h-1} up to step $h-1$, as well as a quantity $e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot Q_h^*(s,a)}$, which is with respect to the trajectory controlled by an optimal π^*. This means that $\Delta_{h,\beta}$ contains both uncontrolled and optimally controlled trajectories. Second, given a trajectory τ and for $\beta>0$, as h increases, the multiplicative factor $e^{\beta\cdot R(\tau_{h-1})}\in[1,e^{\beta(h-1)}]$ is non-decreasing in h and the exponential value functions $e^{\beta\cdot V_h^*(s)}$, $e^{\beta\cdot Q_h^*(s,a)}\in[1,e^{\beta(H-h+1)}]$ are non-increasing in h; vice versa for $\beta<0$. See Fig. 1 for an illustration of this property. In view of their special structure, we name these gaps as *cascaded gaps*.

We will soon discuss the factor ψ_β, another distinctive and important feature of cascaded gaps, but for now let us show that the gaps satisfy Condition 4.1.

Proposition 4.3. *For any $\beta\ne 0$, we have that $\{\Delta_{h,\beta}\}_{h\in[H]}$ satisfy Condition 4.1 with $Z_h^\pi := e^{\beta(R(\tau_{h-1})+V_h^\pi(s))}$.*

Proof. Let us consider an arbitrary policy π and fix a tuple (h,s,a) such that $a=\pi_h(s)$. We also fix a trajectory τ whose h-th element is (s,a). We have

$$\begin{aligned}
&\Delta_{h,\beta}(s,a;\tau_{h-1})\\
&= e^{\beta\cdot R(\tau_{h-1})}\big[e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot Q_h^*(s,a)}\big]\\
&= e^{\beta\cdot R(\tau_{h-1})}\big[e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot V_h^\pi(s)}\big]\\
&\quad + e^{\beta\cdot R(\tau_{h-1})}\big[e^{\beta\cdot V_h^\pi(s)} - e^{\beta\cdot Q_h^*(s,a)}\big]\\
&= e^{\beta\cdot R(\tau_{h-1})}\big[e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot V_h^\pi(s)}\big]\\
&\quad + e^{\beta\cdot R(\tau_{h-1})}\big[e^{\beta\cdot Q_h^\pi(s,a)} - e^{\beta\cdot Q_h^*(s,a)}\big]\\
&\overset{(i)}{=} e^{\beta\cdot R(\tau_{h-1})}\big[e^{\beta\cdot V_h^*(s)} - e^{\beta\cdot V_h^\pi(s)}\big]
\end{aligned}$$

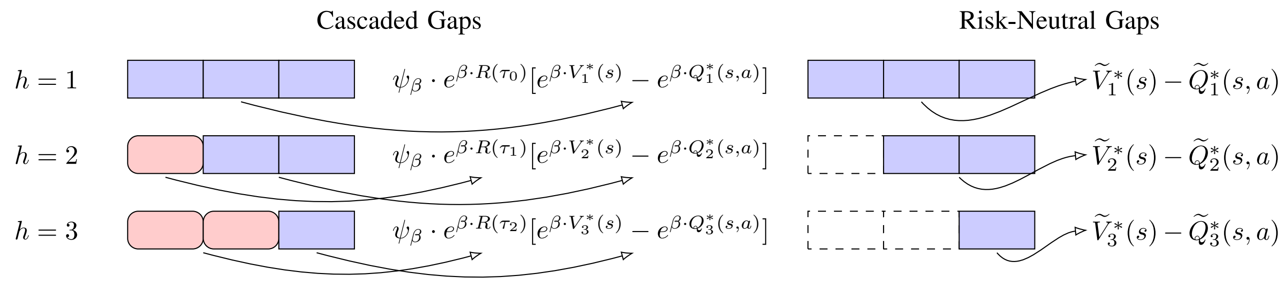

Figure 1. A comparison of the cascaded gaps (4.3) in the risk-sensitive setting ($\beta > 0$) and risk-neutral gaps (4.1) for $H = 3$. The blue blocks illustrate π^*-controlled trajectories, whereas the red blocks illustrate uncontrolled trajectories. Note that for the top cascaded gap, the uncontrolled trajectory part $e^{\beta \cdot R(\tau_0)} = 1$ since $R(\tau_0) = 0$ by definition.

$$
\begin{aligned}
&\quad + e^{\beta \cdot R(\tau_{h-1})} \left[e^{\beta \cdot r_h(s,a)} \, \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot|s,a)} [e^{\beta \cdot V_{h+1}^\pi(s')}] \right] \\
&\quad - e^{\beta \cdot R(\tau_{h-1})} \left[e^{\beta \cdot r_h(s,a)} \, \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot|s,a)} [e^{\beta \cdot V_{h+1}^*(s')}] \right] \\
&= e^{\beta \cdot R(\tau_{h-1})} [e^{\beta \cdot V_h^*(s)} - e^{\beta \cdot V_h^\pi(s)}] \\
&\quad - \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot|s,a)} [e^{\beta \cdot R(\tau_h)} (e^{\beta \cdot V_{h+1}^*(s')} - e^{\beta \cdot V_{h+1}^\pi(s')})] \\
&= D_h^\pi(s) - \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot|s,a)} [D_{h+1}^\pi(s')],
\end{aligned}
$$

where step (i) holds by taking exponential on both sides of the Bellman equation (3.1), and the last step holds by the definition of D_h^π in Condition 4.1 and that of Z_h^π. □

The proof crucially exploits the multiplicative property of the Bellman equation (3.1) raised to exponential[1] as well as the cascading structure of $\Delta_{h,\beta}$. Based on $\Delta_{h,\beta}$, we define the *minimal cascaded gap* $\Delta_{\min,\beta}$ as the minimum non-zero cascaded gap over tuples $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ and trajectories $\tau \in \mathcal{T}$, *i.e.*,

$$
\begin{aligned}
&\Delta_{\min,\beta} \\
&\quad := \min_{h,s,a,\tau} \{ \Delta_{h,\beta}(s, a; \tau_{h-1}) : \Delta_{h,\beta}(s, a; \tau_{h-1}) \neq 0 \},
\end{aligned} \tag{4.4}
$$

For any fixed β, the minimal gap serves as a measure for the difficulty of the corresponding MDP problem. We assume $\Delta_{\min,\beta} > 0$ throughout the paper to avoid triviality.

4.3. Normalization for Risk Consistency

One might notice that our notion of cascaded gaps is not the only gap definition that satisfies Condition 4.1. Indeed, another candidate for the gap definition would be $\Delta'_{h,\beta}(s,a) := \mathrm{sign}(\beta) \cdot e^{\beta \cdot R(\tau_{h-1})} [e^{\beta \cdot V_h^*(s)} - e^{\beta \cdot Q_h^*(s,a)}]$, with the only difference, compared to $\Delta_{h,\beta}$, being that it replaces the normalization factor ψ_β with the sign of β. It is not hard to show that this alternative definition also meets Condition 4.1.

[1]The result of the transformation is known as the exponential Bellman equation (Fei et al., 2021a).

Yet, we demonstrate that the normalizer ψ_β is crucial for the gap $\Delta_{h,\beta}$ to showcase *risk consistency*: the gap has the same order of magnitude when $|\beta|$ is fixed and recovers the risk neutral gap $\widetilde{\Delta}_h$ as $|\beta| \to 0$. To illustrate this point (as well as the deficiency of the alternative $\Delta'_{h,\beta}$), let us consider an MDP with arbitrary transition kernels and its reward function satisfying $r_h(s,a) = 1$ for $(h, s, a) \in [H-1] \times \mathcal{S} \times \mathcal{A}$, $r_H(s, a^*) = 1$ for some action $a^* \in \mathcal{A}$, and $r_H(s,a) = 0$ for $\mathcal{A} \setminus \{a^*\}$. That is, this MDP has all its rewards equal to 1 except for the last step when sub-optimal actions are taken (which yields zero rewards).

For $\beta > 0$, the alternative gap $\Delta'_{h,\beta}$ of the above MDP is on the order of $e^{\beta H} - 1$ (which grows exponentially in β), but for $\beta < 0$, its order is of $1 - e^{\beta H}$ (which is upper bounded by 1 for any $\beta < 0$). Therefore, the magnitude of $\Delta'_{h,\beta}$ is inconsistent under different signs of β. On the other hand, it can be verified that our definition $\Delta_{h,\beta}$ is on the same order of $(e^{|\beta|H} - 1)/|\beta|$ for *all* $\beta \neq 0$, thanks to the risk-dependent normalization factor ψ_β. In addition, as $\beta \to 0$, we have $\Delta_{h,\beta}(s, a; \tau_{h-1}) \to V_h^*(s) - Q_h^*(s,a) = \widetilde{\Delta}_h(s,a)$ for any (h, s, a, τ) by L'Hospital's rule, thereby recovering the definition of sub-optimality gaps in the risk-neutral setting; nevertheless, $\Delta'_{h,\beta}$ tends to 0 and becomes degenerate as $\beta \to 0$.

5. Algorithms

We consider two model-free algorithms for risk-sensitive RL, RSVI2 (Algorithm 1) and RSQ2 (Algorithm 2), both of which are proposed in Fei et al. (2021a).

Algorithm 1 is based on value iteration that features an optimistic estimate Q_h of the action value with a bonus term. In episode k, we compute at each step h the sample average

$$
\begin{aligned}
w_h(s,a) \leftarrow \frac{1}{N_h(s,a)} &\sum_{i \in [k-1]} \mathbb{I}\{(s_h^i, a_h^i) = (s,a)\} \\
&\cdot e^{\beta [r_h(s,a) + V_{h+1}(s_{h+1}^i)]}
\end{aligned} \tag{5.1}
$$

over prior episodes for all visited state-action pairs (s, a). The bonus is given by

$$b_h(s,a) \leftarrow c\left|e^{\beta(H-h+1)} - 1\right|\sqrt{\frac{S\log(2SAHK/\delta)}{N_h(s,a)}}, \tag{5.2}$$

where $c > 0$ is a universal constant. It decays in both step h and the number of visits N_h, thus also known as the doubly decaying bonus (Fei et al., 2021a), and enforces the principle of *Risk-Sensitive Optimism in the Face of Uncertainty* that encourages more exploration of less frequently visited state-action pairs. We then compute the optimistic estimate of the action-value function through

$$Q_h(s,a) \leftarrow \frac{1}{\beta}\log(G_h(s,a)), \tag{5.3}$$

where

$$G_h(s,a) \leftarrow \begin{cases} \min\{e^{\beta(H-h+1)}, w_h(s,a) + b_h(s,a)\}, & \text{if } \beta > 0; \\ \max\{e^{\beta(H-h+1)}, w_h(s,a) - b_h(s,a)\}, & \text{if } \beta < 0. \end{cases}$$

Note that for $\beta > 0$, the addition of the bonus term b_h represents optimism in risk-seeking decision making, whereas for $\beta < 0$ the subtraction of the bonus term corresponds to optimism in risk-averse decision making. Finally, in the policy execution stage, action a_h is taken following the policy that maximizes $Q_h(s_h, \cdot)$ over $\mathcal{A}$.

Algorithm 1 RSVI2

Require: number of episodes K, confidence level $\delta \in (0, 1]$, and risk parameter $\beta \neq 0$

$Q_h(s,a), V_h(s) \leftarrow H - h + 1$, $w_h(s,a) \leftarrow 0$, and $N_h(s,a) \leftarrow 0$ for all $(h,s,a) \in [H+1] \times \mathcal{S} \times \mathcal{A}$
for episode $k = 1, \ldots, K$ **do**
  for step $h = H, \ldots, 1$ **do**
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $N_h(s,a) \geq 1$ **do**
      Update $w_h(s,a)$ following (5.1)
      Update $b_h(s,a)$ following (5.2)
      Update $Q_h(s,a)$ following (5.3)
      $V_h(s) \leftarrow \max_{a' \in \mathcal{A}} Q_h(s,a')$
    end for
  end for
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \arg\max_{a \in \mathcal{A}} Q_h(s_h, a)$ and observe $r_h(s_h, a_h)$ and s_{h+1}
    $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$
  end for
end for

Algorithm 2 RSQ2

Require: number of episodes K, confidence level $\delta \in (0, 1]$, and risk parameter $\beta \neq 0$

$Q_h(s,a), V_h(s) \leftarrow H - h + 1$ if $\beta > 0$; $Q_h(s), V_h(s,a) \leftarrow 0$ if $\beta < 0$, for all $(h,s,a) \in [H+1] \times \mathcal{S} \times \mathcal{A}$
$N_h(s,a) \leftarrow 0$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, and $\alpha_t \leftarrow \frac{H+1}{H+t}$ for all $t \in \mathbb{Z}_+$
for episode $k = 1, \ldots, K$ **do**
  Receive the initial state s_1
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \arg\max_{a' \in \mathcal{A}} Q_h(s_h, a')$, and observe $r_h(s_h, a_h)$ and s_{h+1}
    $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$
    $t \leftarrow N_h(s_h, a_h)$
    Update $w_h(s_h, a_h)$ following (5.4)
    Update $b_{h,t}$ following (5.5)
    Update $Q_h(s_h, a_h)$ following (5.6)
    $V_h(s_h) \leftarrow \max_{a' \in \mathcal{A}} Q_h(s_h, a')$
  end for
end for

On the other hand, Algorithm 2 follows the paradigm of Q-learning. In step h it computes the (exponential) moving average estimate

$$w_h(s_h,a_h) \leftarrow (1-\alpha_t)G_h(s_h,a_h) + \alpha_t \cdot e^{\beta[r_h(s_h,a_h) + V_{h+1}(s_{h+1})]} \tag{5.4}$$

through online updates instead of batch updates as used in Algorithm 1. However, it uses a similar doubly decaying bonus term

$$b_{h,t} \leftarrow c\left|e^{\beta(H-h+1)} - 1\right|\sqrt{\frac{H\log(2SAHK/\delta)}{t}} \tag{5.5}$$

for some universal constant $c > 0$, in enforcing optimism for efficient exploration. Similarly, the optimistic estimation of the value function is set as

$$Q_h(s_h,a_h) \leftarrow \frac{1}{\beta}\log(G_h(s_h,a_h)), \tag{5.6}$$

where the update on the exponential value function and truncation are given by

$$G_h(s_h,a_h) \leftarrow \begin{cases} \min\{e^{\beta(H-h+1)}, w_h(s_h,a_h) + \alpha_t b_t\}, & \text{if } \beta > 0; \\ \max\{e^{\beta(H-h+1)}, w_h(s_h,a_h) - \alpha_t b_t\}, & \text{if } \beta < 0. \end{cases}$$

6. Main Results

In this section, we present gap-dependent regret bounds for risk-sensitive RL. We first provide regret upper bounds for Algorithms 1 and 2, and then we present a regret lower bound that any algorithm has to incur. For notational simplicity, we write $\Delta_{\min} := \Delta_{\min,\beta}$ and $\Delta_h := \Delta_{h,\beta}$ in short by dropping their dependency on β.

6.1. Regret Upper Bounds

The following theorem provides the gap-dependent performance of Algorithm 1.

Theorem 6.1. *For any fixed $\delta \in (0, 1]$, with probability at least $1 - \delta$, the regret of Algorithm 1 is upper bounded by*

$$\mathcal{R}(K) \lesssim \frac{(e^{|\beta|H} - 1)^2 H^3 S^2 A}{|\beta|^2 \Delta_{\min}} \log(HSAK/\delta)^2.$$

Moreover, the expected regret is upper bounded by

$$\mathbb{E}[\mathcal{R}(K)] \lesssim \frac{(e^{|\beta|H} - 1)^2 H^3 S^2 A}{|\beta|^2 \Delta_{\min}} \log(HSAK)^2.$$

The proof is provided in Appendix B.2. The above bounds are general as they hold for any $\beta \neq 0$. They also imply results obtained under the risk-neutral setting when $|\beta| \to 0$. This is verified given that $(e^{|\beta|H} - 1)/|\beta| \to H$ and $\Delta_{\min} \to \widetilde{\Delta}_{\min}$ (where we let $\widetilde{\Delta}_{\min}$ denote the minimal sub-optimality gap for the risk-neutral setting). It can thus be seen that when $|\beta| \to 0$, Theorem 6.1 provides a result that matches the risk-neutral bound $O((H^5 d^3/\widetilde{\Delta}_{\min}) \log(HSAK/\delta)^2)$ (where $d = SA$ under our setting) in He et al. (2021, Theorem 4.4) with respect to K and H.

Next we provide regret guarantees for Algorithm 2.

Theorem 6.2. *For any fixed $\delta \in (0, 1]$, with probability at least $1 - \delta$, the regret of Algorithm 2 is upper bounded by*

$$\mathcal{R}(K) \lesssim \frac{(e^{|\beta|H} - 1)^2 H^4 SA}{|\beta|^2 \Delta_{\min}} \log(HSAK/\delta).$$

Moreover, the expected regret is upper bounded by

$$\mathbb{E}[\mathcal{R}(K)] \lesssim \frac{(e^{|\beta|H} - 1)^2 H^4 SA}{|\beta|^2 \Delta_{\min}} \log(HSAK).$$

The proof is provided in Appendix B.3. Note that the above regret bounds have the same factor $\frac{(e^{|\beta|H}-1)^2}{|\beta|^2 \Delta_{\min}}$ as in Theorem 6.1; we will show in Section 6.2 that such dependency is nearly optimal. Applying the same argument as for Theorem 6.1, when $|\beta| \to 0$, Theorem 6.2 recovers the risk-neutral bound $O((H^6 SA/\widetilde{\Delta}_{\min}) \log(SAHK))$ proved in Yang et al. (2021, Theorem 3.1) for a Q-learning algorithm.

While the above discussion focuses on the case $|\beta| \to 0$, we also have the following result for $|\beta| \leq 1/H$, which is more general.

Corollary 6.3. *For any fixed $\delta \in (0, 1]$, if $|\beta| \leq \frac{1}{H}$, then with probability at least $1 - \delta$ the regret of Algorithms 1 and 2 is upper bounded by*

$$\mathcal{R}(K) \lesssim \begin{cases} \frac{H^5 S^2 A}{\Delta_{\min}} \log(HSAK/\delta)^2, & \textit{for Algorithm 1;} \\ \frac{H^6 SA}{\Delta_{\min}} \log(HSAK/\delta), & \textit{for Algorithm 2.} \end{cases}$$

The expected regret of the two algorithms can be bounded similarly.

Proof. The result follows from Theorems 6.1 and 6.2 by using the fact that the function $f(b) = \frac{e^{bx}-1}{b}$ is increasing on $(0, \infty)$ for any $x > 0$ and $f(\frac{1}{x}) = (e - 1)x \lesssim x$. $\square$

Corollary 6.3 states that as long as $|\beta|$ is sufficiently small, the regret of both algorithms can be bounded by quantities that are polynomial in H (ignoring the possible H-dependence of $\Delta_{\min}$).

Comparison with existing works on risk-sensitive RL. Let us place Theorems 6.1 and 6.2 into the context of known results for risk-sensitive RL. For ease of notation, we define the shorthand $\mathrm{poly}(H, S, A; K, 1/\delta) := \mathrm{poly}(H, S, A) \cdot \mathrm{polylog}(K, 1/\delta)$. Combining our results with existing regret bounds in Fei et al. (2021a, Theorems 1 and 2), we have

$$\mathcal{R}(K) \lesssim \frac{e^{|\beta|H} - 1}{|\beta|} \cdot \mathrm{poly}(H, S, A; K, 1/\delta) \cdot \min\left\{\frac{e^{|\beta|H} - 1}{|\beta|\Delta_{\min}}, K^{1/2}\right\}. \tag{6.1}$$

We see that the gap-dependent regret bounds in Theorems 6.1 and 6.2 trade off the polynomial dependency on K in the $\widetilde{O}(K^{1/2})$-regret (proved by Fei et al. (2021a)) with a factor of $\frac{e^{|\beta|H}-1}{|\beta|\Delta_{\min}}$. Since $\Delta_{\min} \in (0, \frac{1}{|\beta|}(e^{|\beta|H} - 1)]$, we may write $\Delta_{\min} = \frac{\mu}{|\beta|}(e^{|\beta|H} - 1)$ for some $\mu \in (0, 1]$. Then for $\mu \asymp 1$, the above regret bound becomes

$$\mathcal{R}(K) \lesssim \frac{e^{|\beta|H} - 1}{|\beta|} \mathrm{poly}(H, S, A; K, 1/\delta).$$

Under this setting, we attain an *exponential* improvement in K over the existing regret bounds in (Fei et al., 2021a), reducing the polynomial dependency on K (specifically the $\widetilde{O}(K^{1/2})$ dependency) to a logarithmic one. In sharp contrast, the regret bounds of Fei et al. (2021a) that are independent of sub-optimality gaps, *i.e.*,

$$\mathcal{R}(K) \leq \min\left\{HK, \frac{e^{|\beta|H} - 1}{|\beta|} K^{1/2} \mathrm{poly}(H, S, A; K, 1/\delta)\right\}.$$

must incur the exponential factor $\frac{e^{|\beta|H}-1}{|\beta|}$ for gaining only a polynomial improvement in K. When $\mu \lesssim \frac{\log(K)}{\sqrt{K}}$, the regret bound (6.1) is dominated by the existing $\widetilde{O}(K^{1/2})$ bound.

Our gap-dependent regret bounds also imply an exponential improvement in terms of sample complexity. Based on an argument in Jin et al. (2018), our Theorems 6.1 and 6.2 imply that Algorithms 1 and 2 find ε-optimal policies in the PAC setting with $\widetilde{\Omega}\big(\frac{(e^{|\beta| H}-1)^2}{|\beta|^2\Delta_{\min}\varepsilon}\operatorname{poly}(H,S,A)\big)$ samples for any $\varepsilon > 0$. On the other hand, the regret bounds in Fei et al. (2021a) suggest sample complexity bounds on the order of $\widetilde{\Omega}\big(\frac{(e^{|\beta| H}-1)^2}{|\beta|^2\varepsilon^2}\operatorname{poly}(H,S,A)\big)$. Hence, when $\varepsilon = \widetilde{O}(\frac{|\beta|\Delta_{\min}}{e^{|\beta| H}-1})$, our results translate to an exponential improvement in $|\beta|$ and H in sample complexity bounds compared to those of Fei et al. (2021a).

6.2. Regret Lower Bounds

Below we present regret lower bounds that complement the upper bounds in Theorems 6.1 and 6.2.

Theorem 6.4. *If $|\beta|(H-1) \geq \log 4$, $H \geq 2$, $\Delta_{\min} \leq \frac{1}{8|\beta|}$, and $K \asymp \frac{1}{|\beta|^2\Delta_{\min}^2}(e^{|\beta|(H-1)}-1)$, then for any algorithm it holds that*

$$\mathbb{E}[\mathcal{R}(K)] \gtrsim \frac{e^{|\beta|(H-1)}-1}{|\beta|^2\Delta_{\min}};$$

if $|\beta|(H-1) \leq \log H$, $H \geq 8$, $\Delta_{\min} \leq \frac{1}{4|\beta| H}(e^{|\beta|(H-1)}-1)$, and $K \asymp \frac{1}{H|\beta|^2\Delta_{\min}^2}(e^{|\beta|(H-1)}-1)^2$, then for any algorithm it holds that

$$\mathbb{E}[\mathcal{R}(K)] \gtrsim \frac{H}{\Delta_{\min}}.$$

We provide the proof in Appendix C. When β is sufficiently large, Theorem 6.4 provides a lower bound with exponential dependence on $|\beta|$ and H, thus nearly matching the upper bound in Theorem 6.1 in terms of the exponential dependency and up to a logarithmic factor in K. Compared with the upper bound, the lower bound falls short of a term of $e^{|\beta|(H-1)}-1$ as well as polynomial factors in other parameters; it is not yet clear whether there exists a fundamental gap between the two bounds, and we leave the investigation for future work.

On the other hand, when $|\beta|$ is sufficiently small, we achieve a lower bound that depends only polynomially on H and is independent of β (beyond potential dependence in $\Delta_{\min}$). Consequently, this result nearly matches that of Corollary 6.3. Compared with existing risk-neutral lower bound of He et al. (2021, Theorem 5.4), our result specializes in the tabular setting and holds in the regime of non-vanishing β, while theirs adapts to linear function approximation but only in the risk-neutral regime ($|\beta| \to 0$).

To the best of our knowledge, this work presents the first non-asymptotic and gap-dependent regret bounds for risk-sensitive RL based on the entropic risk measure.

6.3. A Unified Framework

In existing literature, algorithms based on value iteration and Q-learning are often analyzed in independent ways due to their distinctive characteristics and update mechanism. We instead employ a unified framework for analyzing the regret of Algorithms 1 and 2. To that end, we focus on the high-probability regret bounds, and the expectation bound can be obtained as a by-product of the analysis. For each $k \in [K]$, let us define $\tau^k := \{(s_h^k, a_h^k)\}_{h\in[H]}$ to be the sample trajectory in episode k. Thanks to Proposition 4.3 that the cascaded gaps $\{\Delta_h\}$ satisfy Condition 4.1, we may derive the following lemma on regret using a standard concentration result.

Lemma 6.5. *For Algorithms 1 and 2 and any fixed $\delta \in (0,1]$, it holds with probability at least $1-\delta/2$ that*

$$\begin{aligned}\mathcal{R}(K) \lesssim \sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k, a_h^k; \tau_{h-1}^k) \\ + \frac{e^{|\beta| H}-1}{|\beta|}H\log(\log K/\delta).\end{aligned}$$

The proof is given in Appendix B.1. In the above lemma, the regret $\mathcal{R}(K)$ plays a role similar to the expectation of the random variable $\sum_{k,h}\Delta_h(s_h^k, a_h^k; \tau_{h-1}^k)$, while the second term on the RHS can be interpreted as the deviation of the random variable from its expectation. With Lemma 6.5 in place, it remains to bound the first term of RHS for both algorithms. We do so in the next two lemmas.

Lemma 6.6. *For Algorithm 1 and any $\delta \in (0,1]$, it holds with probability at least $1-\delta/2$ that*

$$\begin{aligned}&\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k, a_h^k; \tau_{h-1}^k) \\ &\quad\lesssim \frac{(e^{|\beta| H}-1)^2H^3S^2A}{|\beta|^2\Delta_{\min}}\log(2HSAK/\delta)^2.\end{aligned}$$

Lemma 6.7. *For Algorithm 2 and any $\delta \in (0,1]$, it holds with probability at least $1-\delta/2$ that*

$$\begin{aligned}&\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k, a_h^k; \tau_{h-1}^k) \\ &\quad\lesssim \frac{(e^{|\beta| H}-1)^2H^4SA}{|\beta|^2\Delta_{\min}}\log(2HSAK/\delta).\end{aligned}$$

We provide the proofs in Appendices B.2.2 and B.3.2. By combining Lemma 6.5 with Lemmas 6.6 and 6.7, we arrive at Theorems 6.1 and 6.2, respectively. In addition, we remark that the bounds on expected regret can also be obtained from Lemmas 6.6 and 6.7 for corresponding algorithms by simple calculations.

7. Conclusion

We study gap-dependent regret for risk-sensitive RL with the entropic risk measure under episodic and finite-horizon MDPs. We propose a novel definition of sub-optimality gaps, named as cascaded gaps, tailored to the unique characteristics of risk-sensitive RL. We prove gap-dependent lower bounds on the regret to be incurred by any algorithm, and provide nearly matching upper bounds for two existing model-free algorithms. Under proper settings, we demonstrate that our upper bounds imply exponential improvement in bounds of both regret and sample complexity over existing results.

References

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

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

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

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

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

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. In *ICLR*, 2021.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pages 1670–1680. PMLR, 2020.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. In *NeurIPS*, 2021a.

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

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

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

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

Yingjie Fei and Yudong Chen. Exponential error rates of sdp for block models: Beyond grothendieck's inequality. *IEEE Transactions on Information Theory*, 65(1):551–571, 2018a.

Yingjie Fei and Yudong Chen. Hidden integrality of sdp relaxations for sub-gaussian mixture models. In *Conference On Learning Theory*, pages 1931–1965. PMLR, 2018b.

Yingjie Fei and Yudong Chen. Achieving the bayes error rate in synchronization and block models by sdp, robustly. *IEEE Transactions on Information Theory*, 66(6): 3929–3953, 2020.

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

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33:6743–6754, 2020b.

Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2111.03947*, 2021a.

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

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

Hans Föllmer and Thomas Knispel. Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 11 (02n03):333–351, 2011.

Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton university press, 2011.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

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

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

David Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic control*, 18(2):124–131, 1973.

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

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

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

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

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence*, pages 129–139. PMLR, 2021a.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. *arXiv preprint arXiv:2110.12727*, 2021b.

Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. In *Advances in Neural Information Processing Systems*, 2021c.

Yifei Min, Tianhao Wang, Ruitu Xu, Zhaoran Wang, Michael I Jordan, and Zhuoran Yang. Learn to match with no regret: Reinforcement learning in markov matching markets. *arXiv preprint arXiv:2203.03684*, 2022.

David Nass, Boris Belousov, and Jan Peters. Entropic risk measure in policy search. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1101–1106. IEEE, 2019.

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

Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469 (2153):20120683, 2013.

Pedro A Ortega and Alan A Stocker. Human decision-making under limited time. *arXiv preprint arXiv:1610.01698*, 2016.

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

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

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

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32: 1153–1162, 2019.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

Ganlin Song, Ruitu Xu, and John Lafferty. Convergence and alignment of gradient descent with random back propagation weights. *arXiv preprint arXiv:2106.06044*, 2021.

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

Grady Williams, Paul Drews, Brian Goldfain, James M Rehg, and Evangelos A Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1433–1440. IEEE, 2016.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721. IEEE, 2017.

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

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

A. Additional Definitions

Before diving into the proofs, we would like to provide additional definitions on several notion of gaps. We start with the definition of policy-controlled trajectories and sample trajectories as series of state-action pairs; we then define several notions of semi-normalized gaps that we use only in the proofs.

For a policy π, we define the π-controlled trajectory $\tau^\pi \coloneqq \{(s_j, \pi_j(s_j))\}_{j\in[H]}$ as a series of state-action pairs where the action follows π at every state. We define τ^k be the sample trajectory of episode k, *i.e.*, $\tau^k \coloneqq \{(s_j^k, a_j^k)\}_{j\in[H]}$. Let us introduce some additional notion of gaps, based upon cascaded gaps, to assist our proofs. Without loss of generality, we fix $\beta \neq 0$, a trajectory τ, and $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. We define the semi-normalized sub-optimality gap as

$$\overline{\Delta}_h(s, a; \tau_{h-1}) \coloneqq \frac{1}{\beta} e^{\beta \cdot \sum_{j=1}^{h-1} r_j(s_j, a_j)} \left[e^{\beta \cdot V_h^*(s)} - e^{\beta \cdot Q_h^*(s,a)} \right],$$

and we also pair the semi-normalized gap with a semi-normalizer

$$\bar{\psi}_\beta \coloneqq \begin{cases} 1, & \beta > 0; \\ e^{-\beta H}, & \beta < 0. \end{cases}$$

Note that the cascaded gap Δ_h satisfies that $\Delta_h(s, a; \tau_{h-1}) = \bar{\psi}_\beta \cdot \overline{\Delta}_h(s, a; \tau_{h-1})$, which can be regarded as a further level of normalization. For any policy π, we also define the π-controlled sub-optimality gap as

$$\overline{\Delta}_h^\pi(s, a; \tau_{h-1}) \coloneqq \frac{1}{\beta} e^{\beta \cdot \sum_{j=1}^{h-1} r_j(s_j, a_j)} \left[e^{\beta \cdot V_h^*(s)} - e^{\beta \cdot Q_h^\pi(s,a)} \right],$$

which characterizes the sub-optimality of policy π with respect to the optimal policy π^*. Similar to the semi-normalized sub-optimality gap, we define the *normalized* π-controlled sub-optimality gap to be $\Delta_h^\pi(s, a; \tau_{h-1}) \coloneqq \bar{\psi}_\beta \cdot \overline{\Delta}_h^\pi(s, a; \tau_{h-1})$, where the semi-normalizer is applied. Notice that $V_h^*(s) \geq Q_h^*(s, a) \geq Q_h^\pi(s, a)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ by definition, and the gaps are always non-negative quantities due to the monotonicity of exponential function and the normalization factor $\frac{1}{\beta}$. The semi-normalizer $\bar{\psi}_\beta$ is designed to keep the gaps on the same magnitude for both $\beta > 0$ and $\beta < 0$.

We introduce a notion of optimism gap that represents the difference between the optimistic estimation Q_h^k by the algorithm and the optimal value function V_h^*. Similar to the cascaded gap, we define the semi-normalized optimism gap as

$$\overline{\Delta}_h^k(s, a; \tau_{h-1}) \coloneqq \frac{1}{\beta} e^{\beta \cdot \sum_{j=1}^{h-1} r_j(s_j, a_j)} \left[e^{\beta \cdot Q_h^k(s,a)} - e^{\beta \cdot V_h^*(s)} \right],$$

and the *normalized* optimism gap as $\Delta_h^k(s, a; \tau_{h-1}) \coloneqq \bar{\psi}_\beta \cdot \overline{\Delta}_h^k(s, a; \tau_{h-1})$, with the same semi-normalizer applied.

Moreover, we define the (semi-normalized) minimal sub-optimality gap to be the minimal non-zero semi-normalized sub-optimality gap over the tuple (h, s, a, τ):

$$\overline{\Delta}_{\min} \coloneqq \min_{h,s,a,\tau} \{ \overline{\Delta}_h(s, a; \tau_{h-1}) : \overline{\Delta}_h(s, a; \tau_{h-1}) \neq 0 \}.$$

Note that the dependency on β is implicit here. With the above definition, we recall the minimal sub-optimality gap from Eq. (4.4), and have that $\Delta_{\min} = \bar{\psi}_\beta \overline{\Delta}_{\min}$.

In the subsequent proofs we will leverage a peeling argument, for which we define a series of end points $\{\rho_n\}_{n=1}^N$, where $\rho_n \coloneqq 2^n \Delta_{\min}$, and they generate a series of intervals $\{I_n\}_{n=1}^N$ with $I_n \coloneqq [\rho_{n-1}, \rho_n)$ for all $n \in [N]$.

Recall that s_1^k is defined as the state in the first step of episode k; since we assume fixed initial state s_1 for all episodes, we have $s_1^k = s_1$. We introduce the notion of exponential regret that sums over all the episodes the difference between exponential value functions of the optimal policy π^* and that of any policy π^k. Specifically, for any episodic MDP with K episodes, the exponential regret of policy $\{\pi^k\}_{k=1}^K$ is defined as $\mathcal{E}(K) \coloneqq \frac{1}{\beta} \sum_{k\in[K]} [e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k)$.

B. Proofs of Upper Bounds

B.1. Proof of Lemma 6.5

In this proof, we assume $\beta > 0$ without loss of generality, the proof where $\beta < 0$ can be similarly carried out. Let us denote $Z_k \coloneqq \sum_{h\in[H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) - \frac{1}{\beta}[e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k)$. Following from Lemma B.1, we have $\{Z_k\}_{k\in[K]}$ being a martingale difference sequence with respect to the filtration $\mathcal{F}_k$ that represents all the randomness up to episode k. Further recall that the semi-normalized sub-optimality gap for any trajectory

$$\overline{\Delta}_h(s_h, a_h; \tau_{h-1}) = \frac{1}{\beta} e^{\beta \cdot \sum_{j=1}^{h-1} r_j(s_j, a_j)} \left[e^{\beta \cdot V_h^*(s_h)} - e^{\beta \cdot Q_h^*(s_h, a_h)} \right] \geq 0,$$

and we can thus control the magnitude of Z_k by

$$|Z_k| \leq \sum_{h \in [H]} |\overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k)| \leq \frac{H}{|\beta|} |e^{\beta H} - 1| =: B_\beta.$$

For any trajectory $\{\tau_{h-1}^k\}_{h,k}$, if the exponential regret $\mathcal{E}(K) = \frac{1}{\beta} \sum_{k \in [K]} [e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k) \leq B_\beta$, then the sum of Z_k can be lower bounded through applying the definition of Z_k:

$$\sum_{k \in [K]} Z_k \geq -\frac{1}{\beta} \sum_{k \in [K]} (e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}})(s_1^k) \geq -B_\beta.$$

Otherwise, if $\mathcal{E}(K) > B_\beta$, we lower bound the sum $\sum_{k \in [K]} Z_k$ following Freedman inequality from Lemma D.5. More specifically, notice that given the filtration $\mathcal{F}_k$, the variance $\chi = \sum_{k \in [K]} \mathbb{E}[Z_k^2 \mid \mathcal{F}_k]$ over all Z_k's is upper bounded by

$$\begin{aligned}
\chi &\overset{(i)}{\leq} \sum_{k \in [K]} \mathbb{E}[(Z_k + \frac{1}{\beta}(e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}})(s_1^k))^2 \mid \mathcal{F}_k] \\
&= \sum_{k \in [K]} \mathbb{E}\Big[\Big(\sum_{h \in [H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) \Big)^2 \mid \mathcal{F}_k \Big] \\
&\overset{(ii)}{\leq} \sum_{k \in [K]} B_\beta \cdot \mathbb{E}\Big[\sum_{h \in [H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) \mid \mathcal{F}_k \Big] \\
&= B_\beta \sum_{k \in [K]} \frac{1}{\beta} (e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}})(s_1^k) \\
&= B_\beta \cdot \mathcal{E}(K),
\end{aligned}$$

where step (i) follows from $\mathbb{E}[(X - \mathbb{E}\, X)^2] \leq \mathbb{E}\, X^2$ for any random variable X, and step (ii) follows from the fact that $\sum_{h \in [H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) \leq B_\beta$. For any $\varsigma > 0$ and $\varrho \in \mathbb{Z}_+$, we let

$$v_i := \frac{2^i}{K} B_\beta \cdot \mathcal{E}(K) = 2^i \frac{H}{|\beta|^2} (e^{\beta H} - 1)^2$$

and

$$u_i := \sqrt{2^{i+1} \frac{H}{|\beta|^2} (e^{\beta H} - 1)^2 \varsigma} + \frac{2H |e^{\beta H} - 1| \varsigma}{3|\beta|}$$

for each $i \in [\varrho]$, and the corresponding concentration inequality $\mathbb{P}[\sum_{k \in [K]} Z_k \leq -u_i,\ \chi \leq v_i] \leq e^{-\varsigma}$ follows from Lemma D.5. Let us denote the shorthand $U := 2\sqrt{\mathcal{E}(K) \frac{H}{|\beta|} |e^{\beta H} - 1| \varsigma} + \frac{2H|e^{\beta H} - 1|\varsigma}{3|\beta|}$ and $\underline{B}_\beta := \frac{1}{|\beta|} |e^{\beta H} - 1| \leq B_\beta$, then it holds that

$$\mathbb{P}\Big[\sum_{k \in [K]} Z_k \leq -U,\ \ \mathcal{E}(K) > B_\beta \Big] \leq \mathbb{P}\Big[\sum_{k \in [K]} Z_k \leq -U,\ \ \mathcal{E}(K) > \underline{B}_\beta \Big],$$

and we can bound the RHS following a peeling argument. Notice that event $\mathcal{G} \subseteq \bigcup_{i=1}^{\varrho} \mathcal{G}_i$, where we denote the events $\mathcal{G} := \{ \frac{1}{|\beta|} |e^{\beta H} - 1| < \mathcal{E}(K) \leq \frac{K}{|\beta|} |e^{\beta H} - 1| \}$ and $\mathcal{G}_i := \{ \frac{2^{i-1}}{|\beta|} |e^{\beta H} - 1| < \mathcal{E}(K) \leq \frac{2^i}{|\beta|} |e^{\beta H} - 1| \}$ for all $i \in [\varrho]$. It follows the definition that

$$\mathbb{P}\Big[\sum_{k \in [K]} Z_k \leq -U,\ \ \mathcal{E}(K) > \underline{B}_\beta \Big] \overset{(i)}{=} \mathbb{P}\Big[\sum_{k \in [K]} Z_k \leq -U,\ \ \chi \leq B_\beta \cdot \mathcal{E}(K),\ \ \mathcal{G} \Big]$$

$$\overset{(ii)}{\leq} \sum_{i=1}^{\varrho} \mathbb{P}\Big[\sum_{k\in[K]} Z_k \leq -U,\ \ \chi \leq B_\beta \cdot \mathcal{E}(K),\ \ \mathcal{G}_i \Big]$$
$$\overset{(iii)}{\leq} \sum_{i=1}^{\varrho} \mathbb{P}\Big[\sum_{k\in[K]} Z_k \leq -u_i,\ \ \chi \leq v_i \Big]$$
$$\leq \varrho e^{-\varsigma},$$

where step (i) follows from the fact that $\mathcal{E}(K) \leq K|e^{\beta H} - 1|/|\beta|$, step ($ii$) follows from stratifying the feasible range into $\varrho = \lceil \log K \rceil$ layers and applying union bound over all $i \in [\varrho]$, and step (iii) follows from relaxing the quantity of $\mathcal{E}(K)$ within the stratified range $2^{i-1}|e^{\beta H} - 1|/|\beta| < \mathcal{E}(K) \leq 2^i|e^{\beta H} - 1|/|\beta|$ for each $i \in [\varrho]$. Combining both cases, with probability at least $1 - \varrho \cdot e^{-\varsigma}$, we have a lower bound

$$\sum_{k\in[K]} Z_k \geq \min\{-U, -B_\beta\} \geq -U - B_\beta.$$

Recall that $\sum_{k\in[K]} Z_k = \sum_{k\in[K]} \sum_{h\in[H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) - \mathcal{E}(K)$. The lower bound on $\sum_{k\in[K]} Z_k$ implies an upper bound on exponential regret:

$$\mathcal{E}(K) \leq 2\sqrt{\mathcal{E}(K) \cdot \frac{H}{|\beta|}|e^{\beta H} - 1|\varsigma} + \sum_{k\in[K]} \sum_{h\in[H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) + \frac{2H|e^{\beta H} - 1|\varsigma}{3|\beta|} + \frac{H}{|\beta|}|e^{\beta H} - 1|,$$

and a sufficient condition gives that with probability at least $1 - \delta/2$

$$\mathcal{E}(K) \leq 2 \sum_{k\in[K]} \sum_{h\in[H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) + \frac{16H|e^{\beta H} - 1|\varsigma}{3|\beta|} + \frac{2H}{|\beta|}|e^{\beta H} - 1|,$$

where $\varsigma = \log(2\lceil \log K \rceil/\delta)$. Following Lemma D.1, with probability at least $1 - \delta/2$, the total regret $\mathcal{R}(K)$ is bounded by

$$\begin{aligned}
\mathcal{R}(K) &\leq \bar{\psi}_\beta \cdot \mathcal{E}(K) \\
&\leq 2\bar{\psi}_\beta \sum_{k\in[K]} \sum_{h\in[H]} \overline{\Delta}_h(s_h^k, a_h^k; \tau_{h-1}^k) + \frac{(e^{|\beta|H} - 1)(16H \log(2\lceil \log K \rceil/\delta) + 6H)}{3|\beta|} \\
&= 2 \sum_{k\in[K]} \sum_{h\in[H]} \Delta_h(s_h^k, a_h^k; \tau_{h-1}^k) + \frac{(e^{|\beta|H} - 1)(16H \log(2\lceil \log K \rceil/\delta) + 6H)}{3|\beta|}.
\end{aligned}$$

Lemma B.1. *The quantity* $\frac{1}{\beta}[e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k)$ *for episode* $k \in [K]$ *admits the following decomposition:*

$$\frac{1}{\beta}[e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k) = \mathbb{E}\Big[\sum_{h\in[H]} \overline{\Delta}_h(s_h, \pi_h^k(s_h); \tau_{h-1}^{\pi^k}) \mid \mathcal{F}_k \Big].$$

Proof. For any episode $k \in [K]$, we have

$$\frac{1}{\beta}[e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k) = \overline{\Delta}_1(\tau_1^{\pi^k}) + \frac{1}{\beta} e^{\beta \cdot r_1(s_1, \pi_1^k(s_1))} \mathbb{E}_{s_2}[(e^{\beta \cdot V_2^*} - e^{\beta \cdot V_2^{\pi^k}})(s_2) \mid \mathcal{F}_k],$$

where the equality is due to Proposition 4.3 and the expectation is taken over the transition probability $\mathcal{P}_1(\cdot \mid s_1, \pi_1^k(s_1))$ given the policy π^k. We expand the RHS of the equation recursively to get

$$\frac{1}{\beta}[e^{\beta \cdot V_1^*} - e^{\beta \cdot V_1^{\pi^k}}](s_1^k) = \sum_{h\in[H]} \mathbb{E}\Big[\overline{\Delta}_h(s_h, \pi_h^k(s_h); \tau_{h-1}^{\pi^k}) \mid \mathcal{F}_k \Big],$$

where the sub-optimality gap is defined over the entire trajectory $\tau_{h-1}^{\pi^k}$ and the expectation is over all trajectories reachable under the policy π^k and transition probability $\mathcal{P}_h(\cdot \mid s_h, \pi_h^k(s_h))$ for all $h \in [H]$. $\square$

In the sections below, we provide for each algorithm a near-optimal upper bound of the sum of cascaded gaps $\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)$ following a peeling argument, a widely used technique for empirical processes (Yang et al., 2021; He et al., 2021). The final results, *i.e.*, Theorems 6.1 and 6.2, follow from plugging in the upper bounds of the sum of cascaded gaps into Lemma 6.5.

B.2. Upper Bounds for Algorithm 1

B.2.1. Proof of Theorem 6.1

High-probability regret bound. Following Lemmas 6.5 and 6.6, with probability at least $1-\delta$ we have

$$\begin{aligned}\mathcal{R}(K) &\lesssim \sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)+\frac{(e^{|\beta|H}-1)H\log(\log K/\delta)}{|\beta|}\\ &\lesssim \frac{(e^{|\beta|H}-1)^2H^3S^2A\log(2HSAK/\delta)^2}{|\beta|^2\Delta_{\min}},\end{aligned}$$

where the last inequality is due to $\Delta_{\min}\le\frac{1}{|\beta|}(e^{|\beta|H}-1)$.

Expected regret bound. Recall from Lemma D.1 that $\mathcal{R}(K)\le\bar{\psi}_\beta\mathcal{E}(K)$. Since Lemma 6.6 holds with probability at least $1-\delta/2$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(K)] &\le \bar{\psi}_\beta\,\mathbb{E}[\mathcal{E}(K)]\\ &= \bar{\psi}_\beta\,\mathbb{E}\Big[\sum_{k\in[K]}\sum_{h\in[H]}\overline{\Delta}_h(s_h^k,a_h^k;\tau_{h-1}^k)\Big]\\ &= \sum_{\tau_{h-1}^k}\mathbb{P}[\tau_{h-1}^k]\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)\\ &\overset{(i)}{\le} \sum_{n\in[N]}\rho_n\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)\in I_n\}+\frac{\delta}{2|\beta|}\cdot HK(e^{|\beta|H}-1)\\ &\le \sum_{n\in[N]}\rho_n\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)\ge\rho_{n-1}\}+\frac{\delta}{2|\beta|}\cdot HK(e^{|\beta|H}-1)\\ &\overset{(ii)}{\lesssim} \sum_{n\in[N]}\frac{(e^{|\beta|H}-1)^2H^3S^2A\log(4HSAK)^2}{2^n|\beta|^2\Delta_{\min}}+\frac{\delta}{|\beta|}\cdot HK(e^{|\beta|H}-1)\\ &\lesssim \frac{(e^{|\beta|H}-1)^2H^3S^2A}{|\beta|^2\Delta_{\min}}\log(2HSAK)^2,\end{aligned}$$

where step (i) follows from stratifying the range of Δ_h^k into $N:=\lceil\log_2(\frac{1}{|\beta|}(e^{|\beta|H}-1)/\Delta_{\min})\rceil$ slices with end points $\{\rho_n\}_{n=1}^N$, where we define $\rho_n:=2^n\Delta_{\min}$ and interval $I_n:=[\rho_{n-1},\rho_n)$ for all $n\in[N]$; step (ii) follows from Lemma B.3; the last inequality is due to $\Delta_{\min}\le\frac{1}{|\beta|}(e^{|\beta|H}-1)$ and taking $\delta=\frac{1}{HK}$.

B.2.2. Proof of Lemma 6.6

For any $h\in[H]$ and $k\in[K]$, we have $V_h^*(s_h^k)=Q_h^*(s_h^k,\pi_h^*(s_h^k))$ and $\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)\le\frac{1}{|\beta|}(e^{|\beta|H}-1)$. Let us define $N:=\lceil\log_2(\frac{1}{|\beta|}(e^{|\beta|H}-1)/\Delta_{\min})\rceil$. Following Lemma B.2, with probability at least $1-\delta/2$, we have

$$\begin{aligned}\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k) &\overset{(i)}{\le} \sum_{k\in[K]}\sum_{h\in[H]}\sum_{n\in[N]}\rho_n\cdot\mathbb{I}\{\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)\in I_n\}\\ &\overset{(ii)}{\le} \sum_{n\in[N]}\rho_n\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h^{\pi^k}(s_h^k,a_h^k;\tau_{h-1}^k)\ge\rho_{n-1}\}\end{aligned}$$

$$\overset{(iii)}{\lesssim} \sum_{n\in[N]} \rho_n \frac{(e^{|\beta|H}-1)^2 H^3 S^2 A \log(2HSAK/\delta)^2}{4^{n-1}|\beta|^2\Delta_{\min}^2}$$
$$\lesssim \frac{(e^{|\beta|H}-1)^2 H^3 S^2 A \log(2HSAK/\delta)^2}{|\beta|^2\Delta_{\min}},$$

where step (i) is due to the peeling argument that stratifies the range of $\Delta_{\min}$ into N slices with end points $\{\rho_n\}_{n=1}^N$, step (ii) follows from $Q_h^*(s_h^k, a_h^k) \geq Q_h^{\pi}(s_h^k, a_h^k)$, and step ($iii$) follows from Lemma B.2.

Lemma B.2. *Under Algorithm 1, with probability at least* $1-\delta/2$*, we have for any* $n \in \mathbb{Z}_+$

$$\sum_{k\in[K]}\sum_{h\in[H]} \mathbb{I}\{\Delta_h^{\pi^k}(s_h^k, a_h^k; \tau_{h-1}^k) \geq \rho_n\} \lesssim \frac{(e^{|\beta|H}-1)^2 H^3 S^2 A}{4^n|\beta|^2\Delta_{\min}^2}\log(2HSAK/\delta)^2.$$

Proof. Let us denote $M_{h,n}$ to be the number of episodes such that the sub-optimality of the episode at step h is no less than ρ_n, *i.e.*, $M_{h,n} := \sum_{k\in[K]} \mathbb{I}\{\Delta_h^{\pi^k}(s_h^k, a_h^k; \tau_{h-1}^k) \geq \rho_n\}$. Especially, $k_1 < \ldots < k_{M_{h,n}} < k$ denote the selected indices of previous episodes such that $\Delta_h^{\pi^{k_i}}(s_h^{k_i}, a_h^{k_i}; \tau_{h-1}^{k_i}) \geq \rho_n$ at step h, and we further define $R_h^{k_i} := \sum_{j=1}^h r_j(s_j^{k_i}, a_j^{k_i})$ to be the sum of rewards for the first h steps within the selected episodes. For the convenience of notation, we use ϑ to denote the logarithmic factor $\log(2HSAK/\delta)$. Let us also define a shorthand $[\mathcal{P}_h V](s,a) := \mathbb{E}_{s'}[V(s')]$ with respect to $\mathcal{P}_h$ for any value function $V : \mathcal{S} \to \mathbb{R}$ and state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Notice that we can make a recursive upper bound on the gap between the optimistic value function $Q_h^{k_i}$ and policy controlled value function $Q_h^{\pi^{k_i}}$. Recall that the Bellman equation for each $k \in [K]$ is given by

$$e^{\beta Q_h^{\pi^k}(s_h^k, a_h^k)} = \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot\,|\,s_h^k, a_h^k)}\, e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^{\pi^k}(s')]},$$

and the update rule is

$$e^{\beta Q_h^k}(s_h^k, a_h^k) = \min\{e^{\beta(H-h+1)}, (w_h^k + b_h^k)(s_h^k, a_h^k)\},$$

as specified in the Algorithm 1. In particular, we have

$$w_h^k(s_h^k, a_h^k) \leq \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot\,|\,s_h^k, a_h^k)}\, e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^k(s')]} + b_h^k(s_h^k, a_h^k)$$

with high probability due to Lemma D.3. Combining these together, we get

$$(e^{\beta Q_h^k} - e^{\beta Q_h^{\pi^k}})(s_h^k, a_h^k) \leq \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot\,|\,s_h^k, a_h^k)}\, e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^k(s')]} - \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot\,|\,s_h^k, a_h^k)}\, e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^{\pi^k}(s')]} + 2b_h^k(s_h^k, a_h^k).$$

Hence, the recursive upper bound is constructed as

$$\begin{aligned}\sum_{i\in[M_{h,n}]} \bar{\psi}_\beta e^{\beta\cdot R_{h-1}^{k_i}}(e^{\beta\cdot Q_h^{k_i}} - e^{\beta\cdot Q_h^{\pi^{k_i}}})(s_h^{k_i}, a_h^{k_i}) &\overset{(i)}{\leq} \bar{\psi}_\beta \sum_{i\in[M_{h,n}]} e^{\beta\cdot R_{h-1}^{k_i}} e^{\beta\cdot r_h(s_h^{k_i}, a_h^{k_i})}(e^{\beta\cdot Q_{h+1}^{k_i}} - e^{\beta\cdot Q_{h+1}^{\pi^{k_i}}})(s_{h+1}^{k_i}, a_{h+1}^{k_i}) \\ &\quad + \bar{\psi}_\beta \sum_{i\in[M_{h,n}]} e^{\beta\cdot R_{h-1}^{k_i}} 2b_h^{k_i} + \bar{\psi}_\beta \sum_{i\in[M_{h,n}]} e^{\beta\cdot R_{h-1}^{k_i}}\zeta_{h+1}^{k_i} \\ &\leq \sum_{i\in[M_{h,n}]} \bar{\psi}_\beta e^{\beta\cdot R_h^{k_i}}(e^{\beta\cdot Q_{h+1}^{k_i}} - e^{\beta\cdot Q_{h+1}^{\pi^{k_i}}})(s_{h+1}^{k_i}, a_{h+1}^{k_i}) \\ &\quad + e^{\beta\cdot(h-1)} \sum_{i\in[M_{h,n}]} 2\bar{\psi}_\beta b_h^{k_i} + e^{\beta\cdot(h-1)} \sum_{i\in[M_{h,n}]} \bar{\psi}_\beta \zeta_{h+1}^{k_i},\end{aligned}$$

where step (i) follows from adding and subtracting $\mathbb{E}_{s'}\, e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^k(s')]}$ at the same time and the fact that by construction, $a_{h+1}^k = \pi_{h+1}^k(s_{h+1}^k)$ and

$$V_{h+1}^{\pi^k}(s_{h+1}^k) = Q_{h+1}^{\pi^k}(s_{h+1}^k, a_{h+1}^k),$$

$$V_{h+1}^k(s_{h+1}^k) = Q_{h+1}^k(s_{h+1}^k, a_{h+1}^k).$$

Recall that for each episode k and step h we defined the bonus term $b_h^k \coloneqq c|e^{\beta(H-h+1)} - 1|\sqrt{\frac{S\vartheta}{\max\{1, N_h^k(s_h^k, a_h^k)\}}}$, where c is a universal constant, and here we also define

$$\begin{aligned}\zeta_{h+1}^k \coloneqq{}& [\mathcal{P}_h(e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^k(s')]} - e^{\beta[r_h(s_h^k,a_h^k)+V_{h+1}^{\pi^k}(s')]})](s_h^k, a_h^k)\\ &- e^{\beta\cdot r_h(s_h^k,a_h^k)}(e^{\beta\cdot V_{h+1}^k} - e^{\beta\cdot V_{h+1}^{\pi^k}})(s_{h+1}^k).\end{aligned}$$

To simplify the notation, we denote $n_h^k \coloneqq \max\{1, N_h^k(s_h^k, a_h^k)\}$. Expanding the recursive inequality, with probability at least $1-\delta/2$ we have

$$\begin{aligned}\sum_{i\in[M_{h,n}]} e^{\beta\cdot R_{h-1}^{k_i}}(e^{\beta\cdot Q_h^{k_i}} - e^{\beta\cdot Q_h^{\pi^{k_i}}})(s_h^{k_i}, a_h^{k_i}) &\le 2\sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)}\bar{\psi}_\beta b_h^{k_i} + \sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)}\bar{\psi}_\beta \zeta_{h+1}^{k_i}\\ &\overset{(i)}{\le} 2c(e^{|\beta|H}-1)\sqrt{2H^2S^2AM_{h,n}\vartheta^2} + (e^{|\beta|H}-1)\sqrt{2HM_{h,n}\vartheta},\end{aligned} \tag{B.1}$$

where step (i) follows from two upper bounds on $\sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)} b_h^{k_i}$ and $\sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)}\zeta_{h+1}^{k_i}$. More specifically, for the summation on $b_h^{k_i}$ we have

$$\begin{aligned}\sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)} b_h^{k_i} &\le \sum_{h\in[H]}\sum_{i\in[M_{h,n}]} c|e^{\beta H}-1|\sqrt{\frac{S\vartheta}{n_h^{k_i}}}\\ &\overset{(i)}{\le} c|e^{\beta H}-1|\sqrt{S\vartheta}\sum_{h\in[H]}\sqrt{M_{h,n}}\sqrt{\sum_{i\in[M_{h,n}]}\frac{1}{n_h^{k_i}}}\\ &\le c|e^{\beta H}-1|\sqrt{S\vartheta}\sum_{h\in[H]}\sqrt{M_{h,n}}\sqrt{\sum_{s,a}\sum_{j=1}^{N_h^{M_{h,n}}(s,a)}\frac{1}{\max\{1,j\}}}\\ &\overset{(ii)}{\le} c|e^{\beta H}-1|\sqrt{S\vartheta}\sqrt{2H^2SAM_{h,n}},\end{aligned}$$

where step (i) follows from the Cauchy–Schwarz inequality, and step (ii) follows from the pigeonhole principle. Since each term of $e^{\beta(h-1)}\zeta_h^k$ can be controlled by $|e^{\beta(h-1)}\zeta_h^k| \le |e^{\beta H}-1|$ for all $k\in[K]$ and $h\in[H]$, the Azuma-Hoeffding inequality gives

$$\mathbb{P}\Big[\sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)}\zeta_{h+1}^{k_i} \ge \varepsilon\Big] \le \exp\Big(-\frac{\varepsilon^2}{2HM_{h,n}(e^{\beta H}-1)^2}\Big)$$

for any $\varepsilon > 0$, which means with probability at least $1-\delta/2$,

$$\sum_{h\in[H]}\sum_{i\in[M_{h,n}]} e^{\beta\cdot(h-1)}\zeta_{h+1}^{k_i} \le |e^{\beta H}-1|\sqrt{2HM_{h,n}\vartheta}.$$

At the same time, we provide for the optimism gap a lower bound as follows:

$$\begin{aligned}\sum_{i\in[M_{h,n}]}\bar{\psi}_\beta e^{\beta\cdot R_{h-1}^{k_i}}(e^{\beta\cdot Q_h^{k_i}} - e^{\beta\cdot Q_h^{\pi^{k_i}}})(s_h^{k_i}, a_h^{k_i}) &\overset{(i)}{\ge} \bar{\psi}_\beta\sum_{i\in[M_{h,n}]} e^{\beta\cdot R_{h-1}^{k_i}}(e^{\beta\cdot Q_h^{k_i}(s_h^{k_i},\pi_h^*(s_h^{k_i}))} - e^{\beta\cdot Q_h^{\pi^{k_i}}(s_h^{k_i},a_h^{k_i})})\\ &\overset{(ii)}{\ge} \bar{\psi}_\beta\sum_{i\in[M_{h,n}]} e^{\beta\cdot R_{h-1}^{k_i}}(e^{\beta\cdot Q_h^{*}(s_h^{k_i},\pi_h^*(s_h^{k_i}))} - e^{\beta\cdot Q_h^{\pi^{k_i}}(s_h^{k_i},a_h^{k_i})})\\ &= \sum_{i\in[M_{h,n}]}|\beta|\Delta_h^{\pi^{k_i}}(s_h^{k_i}, a_h^{k_i}; \tau_h^{k_i})\end{aligned}$$

$$\geq \rho_n|\beta|M_{h,n}, \tag{B.2}$$

where step (i) is due to the construction of Algorithm 1 and step (ii) follows from Lemma D.2.

Finally, we combine the upper bound (B.1) and lower bound (B.2) of $\sum_{i\in[M_{h,n}]}\bar{\psi}_\beta e^{\beta\cdot R_{h-1}}(e^{\beta\cdot Q_h^{k_i}}-e^{\beta\cdot Q_h^{\pi^{k_i}}})(s_h^{k_i},a_h^{k_i})$ to get

$$\rho_n|\beta|M_{h,n}\leq 2c(e^{|\beta|H}-1)\sqrt{2H^2S^2AM_{h,n}\vartheta^2}+(e^{|\beta|H}-1)\sqrt{2HM_{h,n}\vartheta}.$$

Solving for $M_{h,n}$ we get

$$M_{h,n}=\sum_{k\in[K]}\mathbb{I}\{\Delta_h^{\pi^k}(s_h^k,a_h^k;\tau_{h-1}^k)\geq\rho_n\}\lesssim\frac{(e^{|\beta|H}-1)^2H^2S^2A\log(2HSAK/\delta)^2}{4^n|\beta|^2\Delta_{\min}^2}.$$

□

B.3. Upper Bounds for Algorithm 2

B.3.1. Proof of Theorem 6.2

High-probability regret bound. By Lemmas 6.5 and 6.7, it holds with probability at least $1-\delta$ that

$$\begin{aligned}\mathcal{R}(K)&\lesssim\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k)+\frac{(e^{|\beta|H}-1)H\log(\log K/\delta)}{|\beta|}\\&\lesssim\frac{(e^{|\beta|H}-1)^2H^4SA\log(2HSAK/\delta)}{|\beta|^2\Delta_{\min}},\end{aligned}$$

where the last inequality is due to $\Delta_{\min}\leq\frac{1}{|\beta|}(e^{|\beta|H}-1)$.

Expected regret bound. From Lemma D.1 we have that $\mathcal{R}(K)\leq\bar{\psi}_\beta\cdot\mathcal{E}(K)$ and Lemma 6.7 holds with probability at least $1-\delta/2$. The expected regret can be bounded through

$$\begin{aligned}\mathbb{E}[\mathcal{R}(K)]&\leq\bar{\psi}_\beta\,\mathbb{E}[\mathcal{E}(K)]\\&=\bar{\psi}_\beta\,\mathbb{E}\left[\sum_{k\in[K]}\sum_{h\in[H]}\bar{\Delta}_h(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\right]\\&=\sum_{\tau_{h-1}^{\pi^k}}\mathbb{P}[\tau_{h-1}^{\pi^k}]\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\\&\overset{(i)}{\leq}\sum_{n\in[N]}\rho_n\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h^k(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\in I_n\}+\frac{\delta}{2|\beta|}HK(e^{|\beta|H}-1)\\&\overset{(ii)}{\lesssim}\sum_{n\in[N]}\frac{(e^{|\beta|H}-1)^2H^4SA\log(4HSAK)}{2^n|\beta|^2\Delta_{\min}}+\frac{\delta}{|\beta|}HK(e^{|\beta|H}-1)\\&\lesssim\frac{(e^{|\beta|H}-1)^2H^4SA}{|\beta|^2\Delta_{\min}}\log(HSAK),\end{aligned}$$

where step (i) follows from stratifying the range of Δ_h^k into $N:=\lceil\log_2(\frac{1}{|\beta|}(e^{|\beta|H}-1)/\Delta_{\min})\rceil$ slices with end points $\{\rho_n\}_{n=1}^N$; step (ii) follows from Lemma B.3; the last inequality is due to $\Delta_{\min}\leq\frac{1}{|\beta|}(e^{|\beta|H}-1)$ and taking $\delta=\frac{1}{HK}$.

B.3.2. Proof of Lemma 6.7

With the help of Lemma B.3, it holds with probability at least $1-\delta/2$ that

$$\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h^k(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\in I_n\}\lesssim\frac{(e^{|\beta|H}-1)^2H^4SA\log(2HSAK/\delta)}{4^{n-1}|\beta|^2\Delta_{\min}^2}.$$

Hence, with probability at least $1-\delta/2$, the sum of the cascaded gaps is bounded by

$$\begin{aligned}\sum_{k\in[K]}\sum_{h\in[H]}\Delta_h(s_h^k,a_h^k;\tau_{h-1}^k) &\le \sum_{n\in[N]}\rho_n\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h^k(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\in I_n\}\\ &\lesssim \sum_{n\in[N]}\frac{(e^{|\beta|H}-1)^2H^4SA\log(2HSAK/\delta)}{2^{n-1}|\beta|^2\Delta_{\min}}\\ &\lesssim \frac{(e^{|\beta|H}-1)^2H^4SA\log(2HSAK/\delta)}{|\beta|^2\Delta_{\min}},\end{aligned}$$

where $N\coloneqq\lceil\log_2(\frac{1}{|\beta|}(e^{|\beta|H}-1)/\Delta_{\min})\rceil$ and the last step follows from an infinite sum of geometric series. Recall the definition $\rho_n\coloneqq 2^n\Delta_{\min}$ and interval $I_n\coloneqq[\rho_{n-1},\rho_n)$ for all $n\in[N]$.

Lemma B.3. *Under Algorithm 2, with probability at least $1-\delta/2$, it holds that for any $n\in\mathbb{Z}_+$*

$$\sum_{k\in[K]}\sum_{h\in[H]}\mathbb{I}\{\Delta_h^k(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\in I_n\}\lesssim\frac{(e^{|\beta|H}-1)^2H^4SA}{4^n|\beta|^2\Delta_{\min}^2}\log(2HSAK/\delta).$$

Proof. We focus on the case of $\beta>0$; the case for $\beta<0$ follows a similar argument. We denote the shorthand $\vartheta\coloneqq\log(2HSAK/\delta)$. For every $h\in[H]$ and $n\in[N]$, we define

$$\overline{M}_{h,n}\coloneqq\sum_{k\in[K]}\mathbb{I}\{\Delta_h^k(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})\in I_n\}$$

to be the number of episodes where the corresponding gap falls into the interval I_n. For any $i\in[\overline{M}_{h,n}]$, we denote k_i to be the i-th episode with gap $\Delta_h^k(s_h^k,a_h^k;\tau_{h-1}^{\pi^k})$ lying in the interval I_n.

Recall that $N_h^k(s_h^k,a_h^k)$ is the number of visits on state-action pair (s_h^k,a_h^k) at step h prior to episode k, and $\gamma_{h,t}\coloneqq 2\sum_{i\in[t]}\alpha_t^i b_{h,i}$ is the corresponding bonus term for any given t. For the time being, we only consider step h, and we will ignore some of the subscripts on h for simplicity of notation. In particular, we define $t_i\coloneqq N_h^{k_i}(s_h^{k_i},a_h^{k_i})$ and $\varkappa(s,a,j)$ to be the episode where (s,a) is visited for the j-th time. We first apply Lemma D.4 to get an upper bound with three components:

$$\begin{aligned}\sum_{i\in[\overline{M}_{h,n}]}\beta\Delta_h^{k_i}(s_h^{k_i},a_h^{k_i};\tau_{h-1}^{\pi^{k_i}}) &\le e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\left(e^{\beta\cdot Q_h^{k_i}(s_h^{k_i})}-e^{\beta\cdot Q_h^*(s_h^{k_i},a_h^{k_i})}\right)\\ &\le e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\alpha_{t_i}^0(e^{\beta(H-h+1)}-1)+2e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\gamma_{h,t_i}\\ &\quad+e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\sum_{\ell\in[t_i]}\alpha_{t_i}^\ell\cdot\left(e^{\beta\cdot V_{h+1}^{k_\ell}(s_{h+1}^{k_\ell})}-e^{\beta\cdot V_{h+1}^*(s_{h+1}^{k_\ell})}\right).\end{aligned}$$

Especially, the first term on the RHS can be bounded by the number of state-action pairs, *i.e.*,

$$\sum_{i\in[\overline{M}_{h,n}]}\alpha_{t_i}^0(e^{\beta(H-h+1)}-1)\le\sum_{i\in[\overline{M}_{h,n}]}(e^{\beta(H-h+1)}-1)\cdot\mathbb{I}\{t_i=0\}\le(e^{\beta(H-h+1)}-1)SA, \tag{B.3}$$

where $\alpha_{t_i}^i=1$ only if $N_h^{k_i}(s_h^{k_i},a_h^{k_i})=0$, and $(s_h^{k_i},a_h^{k_i})\in\mathcal{S}\times\mathcal{A}$ only has SA choices. The second term can be similarly controlled by

$$\begin{aligned}2e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\gamma_{h,t_i} &\le 2e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}4c(e^{\beta(H-h+1)}-1)\sqrt{\frac{H\vartheta}{t_i}}\\ &\le 8e^{\beta(h-1)}(e^{\beta(H-h+1)}-1)c\sqrt{H\vartheta}\sum_{i\in[\overline{M}_{h,n}]}\frac{1}{\sqrt{N_h^{k_i}(s_h^{k_i},a_h^{k_i})}}\end{aligned}$$

$$\leq 8e^{\beta(h-1)}(e^{\beta(H-h+1)}-1)c\sqrt{H\vartheta}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{j=2}^{N_h^K(s,a)}\frac{\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa(s,a,j)=k_i\}}{\sqrt{j-1}},$$

where for each state-action pair (s,a), the weighted sum $\sum_{i=2}^{N_h^K(s,a)}\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa(s,a,j)=k_i\}/\sqrt{i-1}$ can be further bounded through

$$\sum_{i=2}^{N_h^K(s,a)}\frac{\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa(s,a,j)=k_i\}}{\sqrt{j-1}}\leq\sum_{i=1}^{L_{s,a}}\frac{1}{\sqrt{i}}\leq 2\sqrt{L_{s,a}},$$

where $L_{s,a}\coloneqq\sum_{j=1}^{N_h^K(s,a)}\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa(s,a,j)=k_i\}$. Then we can also get an upper bound on the second term as

$$\begin{aligned}2e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\gamma_{h,t_i}&\leq 16e^{\beta(h-1)}(e^{\beta(H-h+1)}-1)c\sqrt{H\vartheta}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{L_{s,a}}\\&\overset{(i)}{\leq}16e^{\beta(h-1)}(e^{\beta(H-h+1)}-1)c\sqrt{SA\overline{M}_{h,n}H\vartheta},\end{aligned}\tag{B.4}$$

where step (i) follows from $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}L_{s,a}=\overline{M}_{h,n}$. For the third term, by rearranging the order of summations and taking advantage of the fact that $V_h^k(s_h^k)=Q_h^k(s_h^k,a_h^k)$ and $V_h^*(s_h^k)\geq Q_{h+1}^*(s_{h+1}^k,a_{h+1}^k)$, we get

$$\begin{aligned}&\sum_{i\in[\overline{M}_{h,n}]}\sum_{\ell\in[t_i]}\alpha_{t_i}^\ell\left(e^{\beta\cdot V_{h+1}^{k_\ell}(s_{h+1}^{k_\ell})}-e^{\beta\cdot V_{h+1}^*(s_{h+1}^{k_\ell})}\right)\\&=\sum_{\ell\in[K]}\left(e^{\beta\cdot V_{h+1}^\ell}-e^{\beta\cdot V_{h+1}^*}\right)(s_{h+1}^\ell)\sum_{j=N_h^\ell(s_h^\ell,a_h^\ell)+1}^{N_h^K(s_h^\ell,a_h^\ell)}\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa_h(s_h^\ell,a_h^\ell,j)=k_i\}\cdot\alpha_j^{N_h^\ell(s_h^\ell,a_h^\ell)+1}\\&\leq\sum_{\ell\in[K]}\left(e^{\beta\cdot Q_{h+1}^\ell}-e^{\beta\cdot Q_{h+1}^*}\right)(s_{h+1}^\ell)\sum_{j=N_h^\ell(s_h^\ell,a_h^\ell)+1}^{N_h^K(s_h^\ell,a_h^\ell)}\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa_h(s_h^\ell,a_h^\ell,j)=k_i\}\cdot\alpha_j^{N_h^\ell(s_h^\ell,a_h^\ell)+1}.\end{aligned}$$

Denote $\phi^\ell\coloneqq\sum_{j=N_h^\ell(s_h^\ell,a_h^\ell)+1}^{N_h^K(s_h^\ell,a_h^\ell)}\mathbb{I}\{\exists i\in[\overline{M}_{h,n}]:\varkappa_h(s_h^\ell,a_h^\ell,j)=k_i\}\cdot\alpha_j^{N_h^\ell(s_h^\ell,a_h^\ell)+1}$, and the above inequality turns into

$$\sum_{i\in[\overline{M}_{h,n}]}\sum_{\ell\in[t_i]}\alpha_{t_i}^\ell\left(e^{\beta\cdot V_{h+1}^{k_\ell}(s_{h+1}^{k_\ell})}-e^{\beta\cdot V_{h+1}^*(s_{h+1}^{k_\ell})}\right)\leq\sum_{\ell\in[K]}\phi^\ell\left(e^{\beta\cdot Q_{h+1}^\ell}-e^{\beta\cdot Q_{h+1}^*}\right)(s_{h+1}^\ell).$$

Recall that each element in $\{\phi^\ell\}_{\ell\in[K]}$ is bounded by $1+\frac{1}{H}$, and expand the recursive inequality

$$\begin{aligned}e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\left(e^{\beta\cdot Q_h^{k_i}(s_h^{k_i})}-e^{\beta\cdot Q_h^*(s_h^{k_i},a_h^{k_i})}\right)&\leq(e^{\beta H}-1)SA+16(e^{\beta H}-1)c\sqrt{SA\overline{M}_{h,n}H\vartheta}\\&\quad+e^{\beta(h-1)}\sum_{\ell\in[K]}\phi^\ell\left(e^{\beta\cdot Q_{h+1}^\ell}-e^{\beta\cdot Q_{h+1}^*}\right)(s_{h+1}^\ell)\end{aligned}$$

to get

$$\begin{aligned}e^{\beta(h-1)}\sum_{i\in[\overline{M}_{h,n}]}\left(e^{\beta\cdot Q_h^{k_i}(s_h^{k_i})}-e^{\beta\cdot Q_h^*(s_h^{k_i},a_h^{k_i})}\right)&\leq\sum_{w=0}^{H-h}SA(e^{\beta H}-1)\left(1+\frac{1}{H}\right)^w\\&\quad+\sum_{w=0}^{H-h}16(e^{\beta H}-1)c(1+1/H)^w\sqrt{SA\overline{M}_{h,n}H\vartheta}\\&\leq eHSA(e^{\beta H}-1)+16eH(e^{\beta H}-1)c\sqrt{SA\overline{M}_{h,n}H\vartheta}.\end{aligned}\tag{B.5}$$

With (B.3), (B.4), and (B.5), we obtain the complete upper bound

$$\sum_{i\in[\overline{M}_{h,n}]} \beta\Delta_h^{k_i}(s_h^{k_i}, a_h^{k_i}; \tau_{h-1}^{\pi^{k_i}}) \le e^{\beta(h-1)} \sum_{i\in[\overline{M}_{h,n}]} \left(e^{\beta\cdot Q_h^{k_i}(s_h^{k_i})} - e^{\beta\cdot Q_h^*(s_h^{k_i}, a_h^{k_i})}\right)$$
$$\le (e^{\beta H}-1)(eHSA + 16eHc\sqrt{SA\overline{M}_{h,n}H\vartheta}). \tag{B.6}$$

On the other side, we can also obtain a lower bound on the sum of gaps following the stratification of the empirical gap $\overline{\Delta}_h^k$:

$$\sum_{i\in[\overline{M}_{h,n}]} \beta\Delta_h^{k_i}(s_h^{k_i}, a_h^{k_i}; \tau_{h-1}^{\pi^{k_i}}) \ge \rho_{n-1}\beta\overline{M}_{h,n}. \tag{B.7}$$

We combine both the upper bound (B.6) and the lower bound (B.7) on $\sum_{i\in[\overline{M}_{h,n}]} \beta\Delta_h^{k_i}(s_h^{k_i}, a_h^{k_i}; \tau_{h-1}^{\pi^{k_i}})$ to get $\rho_{n-1}\beta\overline{M}_{h,n} \le (e^{\beta H}-1)(eHSA + 8eHc\sqrt{SA\overline{M}_{h,n}H\vartheta})$, which leads to a sufficient condition

$$\overline{M}_{h,n} = \sum_{k\in[K]} \mathbb{I}\{\Delta_h^k(s_h^k, a_h^k; \tau_{h-1}^{\pi^k}) \in I_n\} \lesssim \frac{(e^{\beta H}-1)^2 H^3 SA\vartheta}{4^n\beta^2\Delta_{\min}^2}.$$

Recall that $\vartheta = \log(2HSAK/\delta)$. Sum the equation above over $h \in [H]$ to get

$$\sum_{k\in[K]}\sum_{h\in[H]} \mathbb{I}\{\Delta_h^k(s_h^k, a_h^k; \tau_{h-1}^{\pi^k}) \in I_n\} \lesssim \frac{(e^{\beta H}-1)^2 H^4 SA\log(2HSAK/\delta)}{4^n\beta^2\Delta_{\min}^2}.$$

□

C. Lower Bounds

C.1. Proof of Theorem 6.4

We prove the two cases of the theorem in Lemmas C.1 and C.2, respectively. We construct two bandit problems such that for any policy π the maximum regret in these two problems is lower bounded. Let us assume the first bandit machine BANDIT I has two arms, where the first arm has reward $H-1$ with probability $p_1^{\mathbb{I}\{\beta>0\}}(1-p_1)^{\mathbb{I}\{\beta<0\}}$ and reward 0 with probability $(1-p_1)^{\mathbb{I}\{\beta>0\}}p_1^{\mathbb{I}\{\beta<0\}}$, whereas the second arm has reward $H-1$ with probability $p_2^{\mathbb{I}\{\beta>0\}}(1-p_2)^{\mathbb{I}\{\beta<0\}}$ and reward 0 with probability $(1-p_2)^{\mathbb{I}\{\beta>0\}}p_2^{\mathbb{I}\{\beta<0\}}$. Similarly, the second bandit machine BANDIT II is also assumed to have two arms with the same Bernoulli-type rewards, with corresponding probabilities q_1 and q_2, respectively.

It is not hard to see that a K-round bandit problem described above is equivalent to a K-episode and H-step MDP where the state space $\mathcal{S}$ has three elements: initial state s_0, absorbing state s_1, and absorbing state s_2. At the first step, two actions $a_1, a_2 \in \mathcal{A}$ are available to the state s_0. More specifically, if one takes action a_1, then with probability $p_1^{\mathbb{I}\{\beta>0\}}(1-p_1)^{\mathbb{I}\{\beta<0\}}$ for BANDIT I (or $q_1^{\mathbb{I}\{\beta>0\}}(1-q_1)^{\mathbb{I}\{\beta<0\}}$ for BANDIT II) the environment transitions into state s_1 and with probability $(1-p_1)^{\mathbb{I}\{\beta>0\}}p_1^{\mathbb{I}\{\beta<0\}}$ for BANDIT I (or $(1-q_1)^{\mathbb{I}\{\beta>0\}}q_1^{\mathbb{I}\{\beta<0\}}$ for BANDIT II) it transitions into state s_2. Similarly, if one takes action a_2, the environment transitions according to p_2 for BANDIT I (or q_2 for BANDIT II). Moreover, we define reward function $r_h(s_0, a) = 0$, $r_h(s_1, a) = 1$, and $r_h(s_2, a) = 0$. In short, taking action a_1 is equivalent to pulling the first arm on the corresponding bandit machine and taking action a_2 is equivalent to pulling the second arm.

Now we start focusing on the lower bound analysis of the bandit problem. In particular, we define the transition probability p_1, p_2, q_1, and q_2 such that the first arm is optimal on BANDIT I while the second arm is optimal on BANDIT II, *i.e.*,

$$p_2 = u_{\beta,H}, \quad p_1 = q_1 = p_2 + (-1)^{\mathbb{I}\{\beta<0\}}\xi, \quad q_2 = p_2 + (-1)^{\mathbb{I}\{\beta<0\}}\cdot 2\xi,$$

where we select a positive quantity $\xi \le \frac{1}{4}u_{\beta,H}$. The quantity $u_{\beta,H}$ is set to be $e^{-|\beta|(H-1)}$ for Lemma C.1 and $\frac{1}{H}$ for Lemma C.2.

Due to the design of the MDPs, $\overline{\Delta}_h(s, a; \tau_{h-1}) = 0$ for any $h \geq 2$ and τ_{h-1}. For state s_1, the minimal sub-optimality gap is therefore given by $\overline{\Delta}_1(s_1, a) \coloneqq \overline{\Delta}_1(s_1, a; \tau_0) = \frac{1}{|\beta|}(e^{\beta \cdot V_1^*(s_1)} - e^{\beta \cdot Q_1^*(s_1,a)})$ for some action a, which it is by design determined by a and the randomness of the environment. More specifically, $\overline{\Delta}_1(s, a) = 0$ if it takes the optimal action $a = a^*$ and the only non-zero sub-optimality gap is given by the action $a = a'$ that takes the sub-optimal arm a', *i.e.*,

$$\begin{aligned}
\overline{\Delta}_{\min} &= \frac{1}{|\beta|}\left|e^{\beta \cdot V_1^*(s_1)} - e^{\beta \cdot Q_1^*(s_1,a')}\right| \\
&= \frac{1}{|\beta|}\left|p_1 e^{\beta(H-1)} + (1-p_1) - p_2 e^{\beta(H-1)} - (1-p_2)\right| \\
&= \frac{1}{|\beta|}\left|(p_1 - p_2)e^{\beta(H-1)} - (p_1 - p_2)\right| \\
&= \frac{1}{|\beta|}|e^{\beta(H-1)} - 1|\xi.
\end{aligned}$$

Theorem 6.4 follows directly by combining Lemmas C.1 and C.2.

Lemma C.1. *If $|\beta|(H-1) \geq \log 4$, $\Delta_{\min} \leq \frac{1}{8|\beta|}$, and $K \asymp \frac{1}{|\beta|^2 \Delta_{\min}^2}(e^{|\beta|(H-1)} - 1)$, then the regret of any policy obeys*

$$\mathbb{E}[\mathcal{R}(K)] \gtrsim \frac{e^{|\beta|(H-1)} - 1}{|\beta|^2 \Delta_{\min}}.$$

Proof. Applying Lemma C.3 with $K = \lfloor p_2(1-p_2)/\xi^2 \rfloor$, we get

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(K)] &\gtrsim \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \frac{p_2(1-p_2)}{\xi} \\
&\overset{(i)}{\gtrsim} \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \frac{p_2}{\xi} \\
&\overset{(ii)}{\gtrsim} \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \frac{p_2|e^{\beta(H-1)} - 1|}{|\beta|\overline{\Delta}_{\min}} \\
&\overset{(iii)}{\gtrsim} \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \frac{p_2(e^{|\beta|(H-1)} - 1)}{|\beta|\Delta_{\min}} \\
&= \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \frac{1 - p_2}{|\beta|\Delta_{\min}} \\
&\overset{(iv)}{\gtrsim} \frac{e^{|\beta|(H-1)} - 1}{|\beta|^2 \Delta_{\min}},
\end{aligned}$$

where step (*i*) and step (*iv*) follow from $1 - p_2 \geq \frac{1}{2}$, step (*ii*) follows from $\overline{\Delta}_{\min} = |e^{\beta(H-1)} - 1|\xi/|\beta|$, and step (*iii*) is due to the definition of $\Delta_{\min}$, and the equality follows from $p_2 = e^{-|\beta|(H-1)}$. □

Lemma C.2. *If $H \geq 8$, $|\beta|(H-1) \leq \log H$, $\Delta_{\min} \leq \frac{1}{4|\beta|H}(e^{|\beta|(H-1)} - 1)$ and the number of episodes $K \asymp \frac{1}{H|\beta|^2\Delta_{\min}^2}(e^{|\beta|(H-1)} - 1)^2$, the regret of any policy obeys*

$$\mathbb{E}[\mathcal{R}(K)] \geq \frac{H}{\Delta_{\min}}.$$

Proof. Similar to the proof of Lemma C.1, we have $\overline{\Delta}_{\min} = \frac{1}{|\beta|}|e^{\beta(H-1)} - 1|\xi$. Note that we have $\xi \leq \frac{1}{4H}$ satisfied as $K \geq 16H$. Apply Lemma C.3 and take $K = \lfloor p_2(1-p_2)/\xi^2 \rfloor$, then it yields

$$\mathbb{E}[\mathcal{R}(K)] \gtrsim \frac{e^{|\beta|(H-1)} - 1}{|\beta|} \cdot \frac{p_2(1-p_2)}{\xi}$$

$$
\begin{aligned}
&\overset{(i)}{\gtrsim} \frac{e^{|\beta|(H-1)}-1}{|\beta|}\cdot\frac{1}{H\xi}\\
&= \frac{e^{|\beta|(H-1)}-1}{|\beta|}\cdot\frac{|e^{\beta(H-1)}-1|}{H|\beta|\overline{\Delta}_{\min}}\\
&= \frac{(e^{|\beta|(H-1)}-1)^2}{|\beta|^2 H\Delta_{\min}}\\
&\overset{(ii)}{\gtrsim} \frac{(H-1)^2}{H\Delta_{\min}}\\
&\gtrsim \frac{H}{\Delta_{\min}},
\end{aligned}
$$

where step (i) follows from $p_2 = \frac{1}{H}$ and step (ii) follows from $e^x - 1 \geq x$ for $x > 0$ and $e^{|\beta|(H-1)} \leq H$. $\square$

Lemma C.3. *For either case of*

1. $H \geq 2$, $|\beta|(H-1) \geq \log 4$, $p_2 = e^{-|\beta|(H-1)}$, *and* $0 < \xi \leq \frac{1}{4}e^{-|\beta|(H-1)}$*;*
2. $H \geq 8$, $|\beta|(H-1) \leq \log H$, $p_2 = \frac{1}{H}$, *and* $0 < \xi \leq \frac{1}{4H}$,

the regret of any policy obeys

$$\mathbb{E}[\mathcal{R}(K)] \geq \frac{K}{64|\beta|}\xi(e^{|\beta|(H-1)}-1)\exp\Big(-\frac{8K\xi^2}{p_2(1-p_2)}\Big).$$

Proof. Given the definition of the MDPs, for $u_{\beta,H} = e^{-|\beta|(H-1)}$, we have

$$p_2 = e^{-|\beta|(H-1)}, \quad p_1 = q_1 = p_2 + (-1)^{\mathbb{I}\{\beta<0\}}\xi, \quad q_2 = p_2 + (-1)^{\mathbb{I}\{\beta<0\}}\cdot 2\xi,$$

where we select a positive quantity $\xi \leq \frac{1}{4}e^{-|\beta|(H-1)}$ such that all the quantities listed above are bounded below by $\frac{1}{2}$ for $|\beta|(H-1) \geq \log 4$. Similarly for $u_{\beta,H} = \frac{1}{H}$, we have

$$p_2 = \frac{1}{H}, \quad p_1 = q_1 = p_2 + (-1)^{\mathbb{I}\{\beta<0\}}\xi, \quad q_2 = p_2 + (-1)^{\mathbb{I}\{\beta<0\}}\cdot 2\xi,$$

where we select a positive quantity $\xi \leq \frac{1}{4H}$ such that all the quantities are bounded below by $\frac{1}{2}$ for $H > 8$.

For any such MDP equivalent to the above bandit models and any policy π, let us define Γ_a to be the reward from taking action $a \in \mathcal{A}$. For notational convenience, we let a^* denote the optimal arm and a' denote the sub-optimal arm. The regret of such MDP in the k-th episode is given by

$$
\begin{aligned}
(V_1^* - V_1^{\pi^k})(s_1) &= \Big|\frac{1}{\beta}\log\mathbb{E}\, e^{\beta\Gamma_{a^*}} - \frac{1}{\beta}\log\Big(\sum_{a\in\mathcal{A}}\mathbb{P}[a^k = a]\,\mathbb{E}\, e^{\beta\Gamma_a}\Big)\Big|\\
&= \frac{1}{|\beta|}\Big|\log\frac{\sum_{a\in\mathcal{A}}\mathbb{P}[a^k=a]\,\mathbb{E}\, e^{\beta\Gamma_a}}{\mathbb{E}\, e^{\beta\Gamma_{a^*}}}\Big|\\
&\overset{(i)}{\geq} \frac{1}{|\beta|}\log\Big(1 + \frac{\mathbb{P}[a^k=a']|\,\mathbb{E}\, e^{\beta\Gamma_{a'}} - \mathbb{E}\, e^{\beta\Gamma_{a^*}}|}{\mathbb{E}\, e^{\beta\Gamma_{a^*}}}\Big)\\
&= \frac{1}{|\beta|}\log\Big(1 + \mathbb{E}[\mathbb{I}\{a^k=a'\}]\frac{|\,\mathbb{E}\, e^{\beta\Gamma_{a'}} - \mathbb{E}\, e^{\beta\Gamma_{a^*}}|}{\mathbb{E}\, e^{\beta\Gamma_{a^*}}}\Big)\\
&\overset{(ii)}{\geq} \frac{1}{2|\beta|}\frac{|\,\mathbb{E}\, e^{\beta\Gamma_{a'}} - \mathbb{E}\, e^{\beta\Gamma_{a^*}}|}{\mathbb{E}\, e^{\beta\Gamma_{a^*}}}\,\mathbb{E}[\mathbb{I}\{a^k=a'\}],
\end{aligned}
$$

where step (i) is due to $\mathbb{E}\, e^{\beta\Gamma_{a^*}} \ge \mathbb{E}\, e^{\beta\Gamma_{a'}}$ for $\beta > 0$, and step (ii) is due to $\log(1+x) \ge x/2$ for $x \in [0,1]$ and $\xi \le \frac{1}{4}p_2$. In particular, we have

$$\begin{aligned}\frac{|\,\mathbb{E}\, e^{\beta\Gamma_{a'}} - \mathbb{E}\, e^{\beta\Gamma_{a^*}}|}{\mathbb{E}\, e^{\beta\Gamma_{a^*}}} &= \frac{|(\mathbb{P}[a^*] - \mathbb{P}[a'])e^{\beta(H-1)} - (\mathbb{P}[a^*] - \mathbb{P}[a'])|}{\mathbb{P}[a^*]e^{\beta(H-1)} + (1 - \mathbb{P}[a^*])} \\ &= \frac{|\xi(e^{\beta(H-1)} - 1)|}{\mathbb{P}[a^*]e^{\beta(H-1)} + (1 - \mathbb{P}[a^*])} \\ &\overset{(i)}{\ge} \frac{1}{4}\xi(e^{|\beta|(H-1)} - 1),\end{aligned}$$

where step (i) follows from the definition of the bandits and the assumptions. Notice that the inequalities hold for both cases where $u_{\beta,H} = e^{-|\beta|(H-1)}$ and $u_{\beta,H} = \frac{1}{H}$. Notably, $1 - \mathbb{P}[a^*]$ dominates the denominator when $\beta > 0$ while being on the order of $e^{\beta(H-1)}$ when $\beta < 0$.

Let us denote the regret on BANDIT I with $\mathcal{R}_{\mathrm{I}}(K)$ and that on BANDIT II with $\mathcal{R}_{\mathrm{II}}(K)$. Combining the two inequalities above, we have

$$\begin{aligned}\max\{\mathbb{E}[\mathcal{R}_{\mathrm{I}}(K)] + \mathbb{E}[\mathcal{R}_{\mathrm{II}}(K)]\} &\overset{(i)}{\ge} \frac{1}{2}\,\mathbb{E}[\mathcal{R}_{\mathrm{I}}(K)] + \frac{1}{2}\,\mathbb{E}[\mathcal{R}_{\mathrm{II}}(K)] \\ &\overset{(ii)}{\ge} \frac{1}{16|\beta|}\xi(e^{|\beta|(H-1)} - 1)\sum_{k\in[K]}\big(\mathbb{E}_p[\mathbb{I}\{a^k = a'\}] + \mathbb{E}_q[\mathbb{I}\{a^k = a'\}]\big) \\ &\ge \frac{K}{64|\beta|}\xi(e^{|\beta|(H-1)} - 1)\exp\Big(-\frac{8K\xi^2}{p_2(1-p_2)}\Big),\end{aligned}$$

where step (i) follows from $\mathcal{R}(K) = \sum_{k\in[K]}(V_1^* - V_1^{\pi^k})(s_1)$ for each bandit, and step (ii) follows from Lemma C.4. $\square$

Lemma C.4. *Under the setup of Lemma C.3, we have*

$$\sum_{k\in[K]}\big(\mathbb{E}_p[\mathbb{I}\{a^k = a'\}] + \mathbb{E}_q[\mathbb{I}\{a^k = a'\}]\big) \ge \frac{K}{4}\exp\Big(-\frac{8K\xi^2}{p_2(1-p_2)}\Big).$$

Proof. Notice that

$$\begin{aligned}\sum_{k\in[K]}\big(\mathbb{E}_p[\mathbb{I}\{a^k = a'\}] + \mathbb{E}_q[\mathbb{I}\{a^k = a'\}]\big) &= \mathbb{E}_p\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a'\}\Big] + \mathbb{E}_q\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a'\}\Big] \\ &\overset{(i)}{\ge} \frac{K}{2}\,\mathbb{P}_p\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a_1\} \le \frac{K}{2}\Big] + \frac{K}{2}\,\mathbb{P}_q\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a_1\} > \frac{K}{2}\Big],\end{aligned}$$

where step (i) is due to the assumption that the optimal arm of BANDIT I is the first arm and the optimal arm of BANDIT II is the second arm. Following Bretagnolle-Huber inequality (Lattimore and Szepesvári (2020), Theorem 14.2), we have a lower bound in the form of an exponential divergence:

$$\mathbb{P}_p\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a_1\} \le \frac{K}{2}\Big] + \mathbb{P}_q\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a_1\} > \frac{K}{2}\Big] \ge \frac{1}{2}\exp(-D_{\mathrm{KL}}(\mathbb{P}_p \,\|\, \mathbb{P}_q)),$$

and the divergence between two probability measures can be upper bounded through the following argument. Let us denote $\hat{p} = p_2^{\mathbb{I}\{\beta>0\}}(1-p_2)^{\mathbb{I}\{\beta<0\}}$ and $\hat{q} = q_2^{\mathbb{I}\{\beta>0\}}(1-q_2)^{\mathbb{I}\{\beta<0\}}$, then we have

$$\begin{aligned}D_{\mathrm{KL}}(\mathbb{P}_p \,\|\, \mathbb{P}_q) &\overset{(i)}{=} \mathbb{E}_p\Big[\sum_{k\in[K]}\mathbb{I}\{a^k = a'\}\Big]\cdot D_{\mathrm{KL}}\big(\mathrm{Ber}(\hat{p})\|\mathrm{Ber}(\hat{q})\big) \\ &\overset{(ii)}{\le} K\hat{p}\log\Big(1 + \frac{\hat{p} - \hat{q}}{\hat{q}}\Big) + K(1-\hat{p})\log\Big(1 + \frac{\hat{q} - \hat{p}}{1 - \hat{q}}\Big)\end{aligned}$$

$$
\begin{aligned}
&\overset{(iii)}{\leq} K\hat{p}\frac{\hat{p}-\hat{q}}{\hat{q}} + K(1-\hat{p})\frac{\hat{q}-\hat{p}}{1-\hat{q}} \\
&= \frac{(\hat{q}-\hat{p})^2 K}{\hat{q}(1-\hat{q})} \\
&\overset{(iv)}{\leq} \frac{8\xi^2}{p_2(1-p_2)},
\end{aligned}
$$

where step (i) follows from Lattimore and Szepesvári (2020, Lemma 15.1), step (ii) follows from $\mathbb{E}_p[\sum_{k\in[K]}\mathbb{I}\{a^k = a'\}] \leq K$ and the definition of Kullback-Leibler (kl) divergence, step (iii) follows from $\log(1+x) \leq x$, and step (iv) follows from $|p_2 - q_2| = 2\xi$ and $p_2 \leq q_2 \leq \frac{1}{2}$ for $\beta > 0$ and $\frac{1}{2}p_2 \leq q_2 \leq p_2 \leq \frac{1}{2}$ for $\beta < 0$. □

D. Auxiliary Lemmas

Lemma D.1. *If $V_1^k(s_1) \geq V_1^{\pi^k}(s_1)$ for $k \in [K]$, then the regret is upper bounded by*

$$
\mathcal{R}(K) \leq \bar{\psi}_\beta \cdot \mathcal{E}(K).
$$

Proof. Recall that $\frac{d}{dx}\log x = \frac{1}{x}$ for all $x > 0$. Especially, $\frac{d}{dx}\log x \leq 1$ for all $x \geq 1$ and $\frac{d}{dx}\log x \leq e^{|\beta|H}$ for all $x \geq e^{-|\beta|H}$. The regret can be upper bounded by the corresponding exponential regret as follows:

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

where step (i) follows from the assumption that $V_1^k(s) \geq V_1^\pi(s)$ and step (ii) follows from mean value theorem. We provide the proof here for the sake of completeness, and similar proof can be found in Fei et al. (2021a). □

Lemma D.2. *For all $k \in [K]$, $h \in [H]$, state $s \in \mathcal{S}$, and $\delta > 0$, the following holds with probability at least $1 - \frac{\delta}{2}$:*

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

Proof. This is Fei et al. (2021a, Lemma 4). □

Lemma D.3. *Define $\overline{\mathcal{V}}_{h+1} \coloneqq \{\overline{V}_{h+1} : \mathcal{S} \to \mathbb{R} \mid \forall s \in \mathcal{S}, \overline{V}_{h+1}(s) \in [0, H-h]\}$. For any $\delta \in (0,1]$, there exists a universal constant $c_0 > 0$ such that with probability $1-\delta$, we have*

$$
\begin{aligned}
&\left| \frac{1}{N_h^k(s,a)} \sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\left[e^{\beta\left[r_h^\tau + \overline{V}\left(s_{h+1}^\tau\right)\right]} - \mathbb{E}_{s'\sim P_h\left(\cdot \,\mid\, s_h^\tau, a_h^\tau\right)} e^{\beta\left[r_h^\tau + \overline{V}\left(s'\right)\right]}\right] \right| \\
&\qquad \leq c_0\left(e^{\beta(H-h+1)} - 1\right)\sqrt{\frac{S\log(2HSAK/\delta)}{\max\left\{1, N_h^k(s,a)\right\}}}
\end{aligned}
$$

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

Proof. This is Fei et al. (2021a, Lemma 1). □

Lemma D.4. *For any episode $k \in [K]$, step $h \in [H]$, and state-action pair $(s_h^k, a_h^k) \in \mathcal{S} \times \mathcal{A}$ such that $t = N_h^k(s_h^k, a_h^k) \geq 1$, let $\gamma_{h,t} := 2\sum_{i\in[t]} \alpha_t^i b_{h,i}$ and $k_1, \ldots, k_t < k$ be the episodes in which (s_h^k, a_h^k) is visited at step h, then it holds with probability at least $1-\delta$ for any $\beta > 0$ that*

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

and for any $\beta < 0$ that

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

Proof. This is Fei et al. (2021a, Lemmas 3 and 8). □

Lemma D.5 (Freedman Inequality (Cesa-Bianchi and Lugosi (2006), Lemma A.7))**.** *Suppose $\{Z_i\}_{i=1}^n$ be a martingale difference sequence on filtration $\{\mathcal{F}_i\}_{i=1}^n$ such that Z_i is $\mathcal{F}_{i+1}$-measurable, $\mathbb{E}[Z_i \mid \mathcal{F}_i] = 0$, and $|Z_i| \leq B$ for some constant B. Define $\chi = \sum_{i=1}^n \mathbb{E}[Z_i^2 \mid \mathcal{F}_i]$, and it follows for any $u > 0$ and $v > 0$ that*

$$\mathbb{P}\Big[\sum_{i=1}^n Z_i \geq u,\, \chi \leq v\Big] \leq \exp\Big(\frac{-u^2}{2v + 2uB/3}\Big).$$

Future directions and broad impact. Given recent advancement in the research of deep neural networks, a promising direction of further research would be to investigate how neural approximation and its generalization properties (Chen and Xu, 2021; Chen et al., 2020; 2021a; Min et al., 2021a) would benefit risk-sensitive RL. Understanding and designing efficient algorithms for risk-sensitive RL in other settings, including shortest path problems (Min et al., 2021b), off-policy evaluation (Min et al., 2021c), offline learning (Chen et al., 2021b), matching (Min et al., 2022) and non-stationary learning (Fei et al., 2020b), may also be of great interest. Moreover, as risk-sensitive RL is closely related to human learning and behaviors, it would be intriguing to study how it synthesizes with relevant areas such as meta learning and bio-inspired learning (Xu et al., 2021; Song et al., 2021). Last but not least, exploring how risk sensitivity could be used to augment unsupervised learning algorithms (Fei and Chen, 2018a;b; 2020; Ling et al., 2019) would be an important future topic as well.