Regret Bounds for Risk-sensitive Reinforcement Learning with Lipschitz Dynamic Risk Measures

Hao Liang [1 2] **Zhi-Quan Luo** [1 2]

Abstract

We study finite episodic Markov decision processes incorporating dynamic risk measures to capture risk sensitivity. To this end, we present two model-based algorithms applied to *Lipschitz* dynamic risk measures, a wide range of risk measures that subsumes spectral risk measure, optimized certainty equivalent, and distortion risk measures, among others. We establish both regret upper bounds and lower bounds. Notably, our upper bounds demonstrate optimal dependencies on the number of actions and episodes while reflecting the inherent trade-off between risk sensitivity and sample complexity. Additionally, we substantiate our theoretical results through numerical experiments.

1. Introduction

Standard reinforcement learning (RL) aims to identify an optimal policy that maximizes the expected return (Sutton & Barto, 2018). This approach is commonly known as risk-neutral RL since it prioritizes the mean value of the uncertain return. However, in domains characterized by high-stakes scenarios, such as finance (Davis & Lleo, 2008; Bielecki et al., 2000), medical treatment (Ernst et al., 2006), and operations research (Delage & Mannor, 2010), decision-makers exhibit risk-sensitive behavior and strive to optimize a risk measure associated with the return. For example, in the field of finance, investors have different risk appetites, and their investment decisions should consider risk factors such as market volatility, potential losses, and downside risks. Using risk-sensitive reinforcement learning, portfolio managers can optimize their investment strategies by balancing risk and return to meet their clients' risk preferences.

One classical framework that addresses risk sensitivity in Markov decision processes (MDPs) is the *static risk measure*. In this framework, the value of a policy is defined as a risk measure applied to the cumulative reward across all stages. Among the commonly used static risk measures are the entropic risk measure (ERM)(Howard & Matheson, 1972; Föllmer & Knispel, 2011) and the conditional value at risk (CVaR) (Rockafellar et al., 2000), along with several others. However, except for the ERM, the static risk measure generally does not satisfy the Bellman equation. Consequently, obtaining the optimal policy becomes computationally challenging, even when the MDP model is known.

As an extension of the static risk measure, the *dynamic risk measure* (DRM) (Ruszczyński, 2010) is constructed by recursively applying the risk measure to the reward at each stage. This recursive formulation naturally allows for the derivation of a dynamic programming equation and thus circumvents the computational burden. Furthermore, DRMs have the advantage of yielding time-consistent optimal policies, a property that is particularly justified in financial applications (Osogami, 2012). By ensuring time consistency, DRMs provide a more robust framework for decision-making in safety-critical applications, such as clinical treatment, where risk sensitivity at all stages is of paramount importance (Du et al., 2023).

Our work focuses on studying risk-sensitive reinforcement learning (RSRL) with a general DRM in the tabular and episodic MDP setting, in which the agent interacts with an unknown MDP with finite states and actions in an episodic manner. We make the mildest assumption that the risk measure used is Lipschitz continuous with respect to certain metric, which we refer to as the Lipschitz DRM. The Lipschitz risk measure encompasses a wide range of classes of risk measures in practical applications, including spectral risk measure (SRM), distortion risk measure, and optimized certainty equivalent (OCE), among others. Additionally, the Lipschitz DRM is also a broader class of risk measures compared to convex and coherent measures since any finite convex risk measure satisfies the Lipschitz property (Föllmer & Knispel, 2013). As a result, our framework encompasses various RL settings, such as risk-neutral RL, RSRL with ERM, RSRL with dynamic CVaR, and RL with

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen [2]Shenzhen Research Institute of Big Data. Correspondence to: Hao Liang <haoliang1@link.cuhk.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

dynamic OCE [1].

The use of Lipschitz risk measures introduces additional technical challenges that need to be addressed. One such challenge arises in the algorithmic design phase when designing exploration bonuses for generic nonlinear risk measures. The standard techniques commonly used in risk-neutral settings, such as Hoeffding inequality or Bernstein-type concentration bounds, are not directly applicable as they only deal with the concentration of mean values. To overcome this issue, previous works, such as (Du et al., 2023) and (Xu et al., 2023), design the exploration bonus based on specific properties of the risk measures they consider. For instance, (Du et al., 2023) chooses the exploration bonus for dynamic CVaR based on a classical concentration bound specific to CVaR, while (Xu et al., 2023) exploits the optimization representation of OCE and uses the concavity of the utility function to construct the bonus for OCE. We exploit the Lipschitz property of the risk measure to relate the value difference to the supremum distance and then apply the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Massart, 1990) to bound the deviation between one distribution and its empirical version.

Another challenge arises in deriving a recursion of the suboptimality gap across stages in the proof of regret upper bounds. Standard analysis in the risk-neutral setting relies on the linearity of the mean to obtain the recursion, which cannot be directly adapted to the risk-sensitive setting. To tackle this challenge, we leverage the Lipschitz property to establish a relationship between the value difference and the Wasserstein distance between two probability distributions. Specifically, we use a *transport inequality* to bound the Wasserstein distance between two probability mass functions (PMFs) with identical probability mass but different support by the expected difference between their supports. By incorporating the Lipschitz property, we obtain a recursion of the suboptimality gap, where the Lipschitz constant appears as a multiplicative factor. By addressing these challenges, we are able to design efficient algorithms and provide regret upper bounds.

We summarize our main contributions as follows:

1. We propose two model-based algorithms for RSRL with Lipschitz DRM. These algorithms incorporate the principle of optimism in the face of uncertainty (OFU) in different ways to facilitate efficient learning. To the best of our knowledge, this is the first work that investigates RSRL using general DRM without making the simulator assumption.

2. We provide worst-case and gap-dependent regret upper bounds for the proposed algorithms. Notably, the regret bounds are optimal in terms of the number of actions (A), and the number of episodes (K). They are dependent on the product of the Lipschitz constants of the risk measures at all stages, capturing the inherent trade-off between risk sensitivity and sample complexity.

3. We establish the minimax and gap-dependent lower bounds for episodic MDPs with general DRM. These lower bounds are tight in terms of A, K, and the number of states (S). Moreover, they reveal a constant factor that depends on the specific risk measure employed.

1.1. Related Work

RSRL without regret bounds. General DRM applied to MDP is presented in (Ruszczyński, 2010; Shen et al., 2013; Chu & Zhang, 2014; Asienkiewicz & Jaśkiewicz, 2017; Bäuerle & Glauner, 2022). However, these works typically assume that the model of the MDP is known, whereas our paper focuses on studying regret guarantees for RSRL in the presence of an unknown MDP. While there are studies such as (Coache & Jaimungal, 2021; Coache et al., 2022) that explore RSRL with dynamic convex risk measures and dynamic spectral risk measures, respectively, their work does not consider regret guarantees.

Regret bounds for RSRL with static risk measures. (Fei et al., 2020) provide the first regret bound for risk-sensitive tabular MDPs using the ERM. This result is further improved upon by (Fei et al., 2021), where they remove the exponential factor dependence on the episode length. (Fei & Xu, 2022) present the first gap-dependent regret bounds under this framework. (Liang & Luo, 2022) propose distributional reinforcement learning algorithms for RSRL with ERM, matching the results obtained in (Fei et al., 2021). (Bastani et al., 2022) consider RSRL with the objective of the spectral risk measure, where conditional value at risk (CVaR) is a special case. Furthermore, (Wang et al., 2023) improve upon the regret bound obtained in (Bastani et al., 2022) in terms of the number of states and episode length. Note that our work cannot be directly compared to these works since static risk measures and DRMs are different settings.

Regret bounds for RSRL with DRMs. (Du et al., 2023) provides the first regret bound for RSRL using DRMs, specifically focusing on dynamic conditional value at risk (CVaR). A very recent work (Xu et al., 2023) that investigates RSRL using dynamic OCE. OCE is a class of risk measures that encompasses several well-known measures, including ERM, CVaR, and mean-variance. They propose a value iteration algorithm based on the idea of upper confidence bound and derive regret upper bounds for their algorithm, as well as a minimax lower bound. (Lam et al., 2023) focuses on dynamic coherent risk measures in the context of non-linear function approximation. They propose an algorithm that leverages UCB-based value functions with

[1]Further details can be found in Section 1.1

non-linear function approximation and prove a sub-linear regret upper bound. However, their work relies on the assumption of a weak simulator, which allows for generating an arbitrary number of next states from any given state. It remains unclear whether such assumptions can be removed in the tabular setting.

Our work contributes to this branch of literature. In contrast to (Lam et al., 2023), our work does not rely on specific assumptions about the risk measure estimator or concentration bounds. Additionally, our approach considers a broader class of DRMs by focusing on Lipschitz DRMs, which encompasses a wider range of risk measures compared to coherent ones.

Our paper is organized as follows. We first introduce some background and problem formulations in Section 2. We then propose our algorithms in Section 3, which is followed by our main results in Section 4. The proof sketch of our main theorem in given in Section 5. The numerical experiments are shown in Section 6. Finally, we conclude our paper in Section 7.

2. Preliminaries

Notations. We write $[N] := \{1, 2, ..., N\}$ for any positive integers N. We use $\mathbb{I}\{\cdot\}$ to denote the indicator function. We denote by $a \vee b := \max\{a, b\}$. We use the notation $\tilde{\mathcal{O}}(\cdot)$ to represent $\mathcal{O}(\cdot)$ with logarithmic factors omitted. For two real numbers $a < b$, the notation $\mathscr{D}([a, b])$ refers to the space of all probability distributions that are bounded over the interval $[a, b]$. For a discrete set $x = \{x_1, \cdots, x_n\}$ and a probability vector $P = (P_1, \cdots, P_n)$, the notation (x, P) represents the discrete distribution where $\mathbb{P}(X = x_i) = P_i$.

Static risk measure. A (static) risk measure is a mapping $\rho : \mathscr{X} \to \mathbb{R}$ that assigns a real number to each random variable in the set $\mathscr{X}$, which satisfies certain properties of the following. It quantifies the risk associated with a random outcome.

- monotonicity: $X \preceq Y \Rightarrow \rho(X) \le \rho(Y)$,
- translation-invariance: $\rho(X + c) = \rho(X) + c,\ c \in \mathbb{R}$,
- super-additivity: $\rho(X + Y) \ge \rho(X) + \rho(Y)$,
- positive homogeneity: $\rho(\alpha X) = \alpha\rho(X)$ for $\alpha \ge 0$
- concavity: $\rho(\alpha X+(1-\alpha)Y) \ge \alpha\rho(X)+(1-\alpha)\rho(Y)$
- law-invariance: $F_X = F_Y \Rightarrow \rho(X) = \rho(Y)$.

Two intrinsic properties of risk measures are monotonicity and translation-invariance. *Coherent* risk measures, introduced by (Artzner et al., 1999), are a widely used class of risk measures that satisfy super-additivity and positive homogeneity in addition. Coherent risk measures capture important concepts such as diversification and risk pooling. *Concave* risk measures generalize coherent risk measures by relaxing the requirements of super-additivity and positive homogeneity to concavity. Concave risk measures are more flexible and can capture a wider range of risk preferences.

Lipschitz risk measures, on the other hand, form an even broader class of risk measures, which encompass both coherent and concave risk measures. They allow for more general functional forms and provide a flexible framework for capturing risk in various settings. Lipschitz risk measures satisfy the law-invariance property, therefore we overload notations and write $\rho(F_X) := \rho(X)$ for $X \sim F_X$.

Lipschitz continuity. For two cumulative distribution functions (CDFs) F and G, their supremum distance is defined as

$$\|F - G\|_\infty \triangleq \sup_{x\in\mathbb{R}} |F(x) - G(x)|\,.$$

For two distributions F, G over the reals, the Wasserstein distance between them coincides with their ℓ_1 distance (Bhat & LA, 2019)

$$W_1(F, G) = \|F - G\|_1 \triangleq \int_{-\infty}^{\infty} |F(x) - G(x)|\, dx.$$

A risk measure ρ is said to be Lipschitz continuous with respect to a distance $\|\cdot\|_p$ ($p = 1$ or $p = \infty$) over the set of distributions $\mathscr{D}([a, b])$ if

$$\rho(F)-\rho(G) \le L_p(\rho, [a, b])\, \|F - G\|_p\,, \forall F, G \in \mathscr{D}([a, b]).$$

Here, $L_p(\rho, [a, b])$ is the Lipschitz constant associated with the risk measure ρ over the interval $[a, b]$, and it represents the maximum rate of change of the risk measure with respect to the distance metric. The Lipschitz continuity property provides a way to quantify the sensitivity of the risk measure to changes in the underlying distributions. A larger Lipschitz constant indicates a greater sensitivity or variability of the risk measure values with respect to changes in the distributions.

To gain some intuition, we present the Lipschitz constant values for several popular risk measures over the interval $[0, M]$ in Table 1. For interested readers, please refer to Appendix A for the formal definitions of the risk measures and more detailed discussions about the Lipschitz constants.

Episodic MDP. An episodic MDP is defined by a tuple $\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, (P_h)_{h\in[H]}, (r_h)_{h\in[H]}, H)$, where $\mathcal{S}$ is the finite state space with cardinality $S \triangleq |\mathcal{S}|$, $\mathcal{A}$ the finite action space with cardinality $A \triangleq |\mathcal{A}|$, $P_h : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0, 1]$ the probability transition kernel at step h, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$

Table 1. Lipschitz constants of typical risk measures

Lipschitz constant	CVaR	distortion risk measure	ERM	OCE[2]
$L_1([0,M])$	$\frac{1}{\alpha}$	$\max g'(x)$	$\exp(\lvert\beta\rvert M)$	$u'(-M)$
$L_\infty([0,M])$	$\frac{M}{\alpha}$	$M \max g'(x)$	$\frac{\exp(\lvert\beta\rvert M)-1}{\lvert\beta\rvert}$	$u(-M)$

the reward functions at step h, and H the length of one episode. The agent interacts with the environment for K episodes. At the beginning of episode k, an initial state s_1^k is arbitrarily selected. In step h, the agent takes action a_h^k based on the state s_h^k, according to its policy. The policy is represented by a (deterministic) sequence of functions $\pi = (\pi_h)_{h\in[H]}$, where each π_h maps from $\mathcal{S}$ to $\mathcal{A}$. The agent observes the reward $r_h(s_h^k, a_h^k)$ and transitions to the next state $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$. The episode terminates at $H+1$, after which the agent proceeds to the next episode.

Dynamic programming with DRM. The dynamic risk measure is defined via a recursive application of static risk measures $(\rho_h)_{h\in[H-1]}$ (Ruszczyński, 2010). The (risk-sensitive) value function of a policy π at step h is defined recursively

$$
\begin{aligned}
Q_h^\pi(s_h, a_h) &= r_h(s_h, a_h) + \rho_h(V_{h+1}^\pi(s_{h+1})) \\
V_h^\pi(s_h) &= Q_h^\pi(s_h, \pi_h(s_h)), V_{H+1}^\pi(s_{H+1}) = 0.
\end{aligned} \tag{1}
$$

where ρ_h is taken over the *next-state value* $V_{h+1}^\pi(s_{h+1})$, i.e.,

$$
\begin{aligned}
V_{h+1}^\pi(s_{h+1}) \sim \left(V_{h+1}^\pi, P_h(s,a)\right) \implies \\
\rho_h(V_{h+1}^\pi(s_{h+1})) = \rho_h\left(\left(V_{h+1}^\pi, P_h(s,a)\right)\right).
\end{aligned}
$$

We refer to the distribution of $V_{h+1}^\pi(s_{h+1})$ as the *(next-state) value distribution* $\left(V_{h+1}^\pi, P_h(s,a)\right)$. For convenience, we write $\rho(x,P) = \rho((x,P))$, thus we write $\rho_h\left(V_{h+1}^\pi, P_h(s,a)\right)$. By incorporating the risk measure ρ_h into the recursive formulation, the dynamic risk measure framework provides a way to account for risk preferences and evaluate the risk-sensitive value function of a policy at each time step. When the risk measure ρ_h specializes in the mean (i.e., taking the expectation), Equation 1 reduces to the standard Bellman equation.

The (risk-sensitive) optimal policy is defined as the policy that maximizes the value function, i.e., $\pi^* = \arg\max_\pi V_1^\pi$. Consequently, the optimal value function is defined as $V_h^*(s) = V_h^{\pi^*}(s)$ and $Q_h^*(s,a) = Q_h^{\pi^*}(s,a)$. (Ruszczyński, 2010) shows that an optimal Markovian policy exists, and the optimal value functions can be computed recursively. The Bellman optimality equation is given by

$$
\begin{aligned}
Q_h^*(s_h, a_h) &= r_h(s_h, a_h) + \rho_h(V_{h+1}^*, P_h(s_h, a_h)) \\
V_h^*(s_h) &= \max_{a\in\mathcal{A}} Q_h^*(s_h, a_h), V_{H+1}^*(s_{H+1}) = 0.
\end{aligned} \tag{2}
$$

The optimal policy is the greedy policy with respect to the optimal action-value function, i.e., $\pi_h^*(s) = \arg\max_{a\in\mathcal{A}} Q_h^*(s,a)$.

Regret. We define the regret of an algorithm `alg` interacting with an MDP $\mathcal{M}$ for K episodes as

$$
\text{Regret}(\texttt{alg}, \mathcal{M}, K) \triangleq \sum_{k=1}^K V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k).
$$

The regret quantifies the accumulated suboptimality gap of an algorithm compared to the optimal policy. It is a random variable due to the randomness introduced by π^k. We denote the expected regret by $\mathbb{E}[\text{Regret}(\texttt{alg}, \mathcal{M}, K)]$. We may omit the notation $\boldsymbol{\pi}$ and $\mathcal{M}$ when clear from the context.

3. Algorithm

In this section, we present two model-based algorithms that incorporate the OFU principle. They aim to strike a balance between exploration and exploitation during the learning process. Both algorithms belong to the model-based algorithm class as they maintain an empirical model of the environment during the learning process. We make the following assumption on the DRM, which our algorithms apply to:

Assumption 3.1. For each $h \in [H]$, ρ_h is Lipschitz continuous with respect to the $\|\cdot\|_1$ and $\|\cdot\|_\infty$, and satisfies $\rho_h : \mathscr{D}([a,b]) \to [a,b]$.

Remark 3.2. The second condition in Assumption 3.1 is mild since it is satisfied by common risk measures. For example, it is easy to check that CVaR and ERM satisfies this condition.

For simplicity, we drop ρ from the notations and write $L_{p,h} \triangleq L_p(\rho_h, [0, H-h])$ for $h \in [H-1]$. For two probability mass functions (PMFs) P and Q with the same support, we overload notations and denote by $\|P - Q\|_1 := \sum_i |P_1 - Q_i|$ their ℓ_1 distance.

3.1. UCBVI-DRM

`UCBVI-DRM` uses the bonus term to ensure optimism in the estimation of the value function, considering the non-linearity of the risk measure. In each step h of episode k, the optimistic value function is obtained by adding a bonus term b_h^k to the empirical value. The empirical value is constructed by approximating the Bellman optimality equation

(Equation 2) with empirical model. The empirical model is maintained and updated based on the visiting counts

$$N_h^k(s,a) \triangleq \sum_{\tau=1}^{k-1} \mathbb{I}\left\{(s_h^\tau, a_h^\tau) = (s,a)\right\},$$

$$N_h^k(s,a,s') \triangleq \sum_{\tau=1}^{k-1} \mathbb{I}\left\{(s_h^\tau, a_h^\tau, s_{h+1}^\tau) = (s,a,s')\right\}.$$

The empirical model $\hat{P}_h^k$ for step h in episode k is set to be the visiting frequency

$$\hat{P}_h^k(s'|s,a) = \frac{N_h^k(s,a,s')}{N_h^k(s,a) \vee 1}.$$

The bonus term b_h^k is composed of two factors: the estimation error of the next-state value distribution and the Lipschitz constant of the risk measure. The estimation error can be bounded as

$$\left\|(\hat{P}_h^k, V_{h+1}^*) - (P_h, V_{h+1}^*)\right\|_\infty \leq \sqrt{\frac{\iota}{2(N_h^k(\cdot,\cdot) \vee 1)}},$$

where ι is a confidence level to be specified later. This error term captures the uncertainty in the empirical model. The Lipschitz constant $L_{\infty,h}$ of the risk measure reflects its sensitivity to changes in the underlying distributions. Multiplying these two factors together yields the exploration bonus term, which is added to the empirical value function estimate. This bonus term encourages exploration in situations where the model estimation error is large or the risk measure is sensitive. By carefully designing the bonus term, UCBVI-DRM achieves optimism in its value function estimates, promoting exploration while considering the nonlinearity of the risk measure. This allows the algorithm to balance exploration and exploitation, taking into account the uncertainty in the model and the smoothness of the risk measure.

Remark 3.3. Algorithm 1 provides a general framework that subsumes other algorithms such as ICVaR-VI in (Du et al., 2023) for dynamic CVaR and OCE-VI in (Xu et al., 2023) for dynamic OCE. In particular, our bonus term matches theirs by setting $L_{\infty,h} = (H-h)/\alpha$ for CVaR and $L_{\infty,h} = u(-(H-h))$ for OCE. Therefore, Algorithm 1 generalizes these algorithms and provides a unified framework for addressing different risk measures.

3.2. OVI-DRM

The OVI-DRM algorithm (see Algorithm 2) is a model-based algorithm which injects the optimism in the estimated model. It operates at a high level as follows. For each step h in episode k, the algorithm constructs an optimistically estimated transition model $\tilde{P}_h^k$ based on a high probability concentration bound on the empirical transition model $\hat{P}_h^k$.

Algorithm 1 UCBVI-DRM

```
Input: (ρ_h)_{h∈[H−1]}, T and δ
Initialize N_h^k(·,·) ← 0, P̂_h^k(·,·) ← (1/S)1, ι ← log(SAT)
for k = 1 : K do
    V_H^k(·) ← max_a r_H(·,a)
    for h = H − 1 : 1 do
        b_h^k(·,·) ← L_{∞,h} sqrt( ι / (2(N_h^k(·,·)∨1)) )
        Q_h^k(·,·) ← r_h(·,·) + ρ_h(V_{h+1}^k, P̂_h^k(·,·)) + b_h^k(·,·)
        V_h^k(·) ← max_{a∈A} Q_h^k(·,a)
    end for
    Receive s_1^k
    for h = 1 : H do
        Take action a_h^k ← argmax_{a∈A} Q_h^k(s_h^k, a) and transition to s_{h+1}^k
        Update N_h^k(·,·) and P̂_h^k(·,·)
    end for
end for
```

This optimistic model allows for exploration and promotes optimism in the face of uncertainty. Using the optimistic model $\tilde{P}_h^k$, the algorithm approximates the Bellman optimality equation (Equation 2) to obtain optimistic value functions Q_h^k.

The optimistic model $\tilde{P}_h^k$ is obtained by applying a subroutine called OM (Optimistic Model) that takes the empirical model $\hat{P}_h^k(s,a)$, the value at the next step V_{h+1}^k, and a confidence radius $c_h^k(s,a)$ as input. The subroutine constructs an optimistic model within an ℓ_1 norm ball around $\hat{P}_h^k(s,a)$

$$\left\|P_h(s,a) - \hat{P}_h^k(s,a)\right\|_1 \leq c_h^k(s,a).$$

c_h^k represents the confidence radius around the empirical model within which the true model lies with high probability. Due to the space limit, we defer the details of the subroutine OM (Algorithm 3) in Appendix B. Note that OM is computationally efficient with complexity $\mathcal{O}(S\log S)$.

The optimism of the model induces optimism in the value estimates

$$(\tilde{P}_h^k(s,a), V_{h+1}^k) \succeq (P_h(s,a), V_{h+1}^k) \succeq (P_h(s,a), V_{h+1}^*) \implies Q_h^k(s,a) \geq Q_h^*(s,a).$$

By using an optimistically estimated model, the OVI-DRM algorithm promotes exploration and encourages the agent to take actions that have the potential for higher values.

4. Main Results

For convenience, we define $\tilde{L}_{1,t} \triangleq \prod_{i=1}^t L_{1,i}$ for $t \in [H]$.

Algorithm 2 `OVI-DRM`

Input: $(\rho_h)_{h\in[H-1]}$, T and δ
Initialize $N_h^k(\cdot,\cdot) \leftarrow 0$, $\hat{P}_h^k(\cdot,\cdot) \leftarrow \frac{1}{S}\mathbf{1}$
for $k = 1 : K$ **do**
  $V_H^k(\cdot) \leftarrow \max_a r_H(\cdot, a)$
  for $h = H-1 : 1$ **do**
    $c_h^k(\cdot,\cdot) \leftarrow \sqrt{\frac{2S}{N_h^k(\cdot,\cdot)\vee 1}\iota}$
    $\tilde{P}_h^k(\cdot,\cdot) \leftarrow \mathtt{OM}\left(\hat{P}_h^k(\cdot,\cdot), V_h^k, c_h^k(\cdot,\cdot)\right)$
    $Q_h^k(\cdot,\cdot) \leftarrow r_h(\cdot,\cdot) + \rho_h(V_{h+1}^k, \tilde{P}_h^k(\cdot,\cdot))$
    $V_h^k(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_h^k(\cdot, a)$
  end for
  Receive s_1^k
  for $h = 1 : H$ **do**
    Take action $a_h^k \leftarrow \arg\max_{a\in\mathcal{A}} Q_h^k(s_h^k, a)$ and transition to s_{h+1}^k
    Update $N_h^k(\cdot,\cdot)$ and $\hat{P}_h^k(\cdot,\cdot)$
  end for
end for

4.1. Worst-case Regret Bounds

Theorem 4.1 (Worst-case regret upper bound)**.** *Fix $\delta \in (0,1)$. Suppose Assumption 3.1 holds. Algorithm 1 and Algorithm 2 satisfies for any MDP $\mathcal{M}$*

$$\textit{Regret}(K) \leq \tilde{\mathcal{O}}\left(\sum_{h=1}^{H-1} L_{\infty,h}\tilde{L}_{1,h-1}\cdot\sqrt{S^2AK}\right)$$

with probability at least $1-\delta$, where $\iota \triangleq \log(4SAT/\delta)$.

The proof sketch of Theorem 4.1 is shown in Section 5.

Remark 4.2. In the risk-neutral setting, The Lipschitz constants take $L_{\infty,h} = H-h$ and $L_{1,h} = 1$, which leads to the bound of $\tilde{\mathcal{O}}\left(H^2\sqrt{S^2AK}\right)$.

Theorem 4.3 (Minimax Lower Bound)**.** *For any algorithm `alg`, there exists an MDP $\mathcal{M}$ such that for sufficiently large K*

$$\mathbb{E}[\textit{Regret}(\mathtt{alg}, \mathcal{M}, K)] \geq \Omega\left(c_\rho H\sqrt{SAT}\right),$$

where c_ρ is a constant dependent on the risk measure[3].

The construction of proof is based on (Domingues et al., 2021).

Comparisons and Discussions. We compare our regret bounds with that of (Xu et al., 2023) in the dynamic OCE setting. By instantiating the Lipschitz constants of OCE with $L_{\infty,h} = u(-H+h)$ and $L_{1,h} = u'_{-}(-H+h)$, their bound can be translated into $\tilde{\mathcal{O}}\left(\sum_{h=1}^{H-1} L_{\infty,h}\sqrt{\tilde{L}_{1,h-1}S^2AK}\right)$. Our bound matches their result with additional factors $\sqrt{\tilde{L}_{1,h-1}}$. This is because they employ a change-of-measure technique based on the concave optimization representation of OCE to bound derive a tighter recursion of value gaps, which cannot be easily extended to general risk measures. The Algorithm 1, however, still enjoy the same regret bound as that in (Xu et al., 2023) since our algorithm reduces to `OCE-VI` for the dynamic OCE. Furthermore, numerical experiments in Appendix D shows that Algorithm 2 empirically outperforms Algorithm 1 for dynamic CVaR problem. Our lower bound also matches the lower bound in (Xu et al., 2023), both of which are tight in S, A, K, H and depend on some constant related to the risk measure.

In the dynamic CVaR setting, which is a special case of dyamic OCE, our upper bound matches the bound $\tilde{\mathcal{O}}\left(H^2\sqrt{S^2AK}/\sqrt{\alpha^H}\right)$ in (Du et al., 2023) up to a factor of $1/\sqrt{\alpha^H}$. Algorithm 1 subsumes `ICVaR-VI` for the dynamic CVaR problem. In contrast to (Du et al., 2023) that provides a algorithm-dependent lower bound, we provide minimax and gap-dependent lower bound. Furthermore, (Lam et al., 2023) considers dynamic coherent risk measures and non-linear function approximation, and their regret bounds are derived under the assumption of a weak simulator. As a result, the regret bounds provided in our work are not directly comparable to theirs, even if their results are specialized to the tabular MDP setting.

Theorem 4.1 and Theorem 4.3 imply that RSRL with Lipschitz DRM can achieve regret bound that is minimax-optimal in terms of K and A. Specifically, the gap between the upper and lower bounds is determined by two factors: $\sqrt{S}$ and a multiplicative Lipschitz constant $\tilde{L}_{1,H}$. we provide more technical details behind this in the proof sketch. Achieving further improvements in these factors can be challenging, especially for general risk measures, under the mild Lipschitz assumption.

4.2. Gap-dependent Regret Bounds

Fix $h \in [H]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, the *sub-optimality gap* of (s,a) at step h is defined as $\Delta_h(s,a) \triangleq V_h^*(s,a) - Q_h^*(s,a)$. The *minimum sub-optimality gap* is defined as the minimum non-zero gap

$$\Delta_{\min} \triangleq \min_{h,s,a}\left\{\Delta_h(s,a) : \Delta_h(s,a) > 0\right\}.$$

Theorem 4.4 (Gap-dependent regret upper bound)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, Algorithm 1 and*

[3]For more details, please refer to Appendix D.

Algorithm 2 satisfy

$$\text{Regret}(K) \leq \mathcal{O}\left(\frac{S^2AH\left(\sum_{h=1}^{H-1}\tilde{L}_{1,h-1}L_{\infty,h}\right)^2}{\Delta_{min}}\log(SAT)\right).$$

We follow the standard convention in the literature for the gap-dependent lower bound. The lower bound is stated for algorithms that have sublinear worst-case regret. Specifically, we say that an algorithm `alg` is α-uniformly good if for any MDP $\mathcal{M}$, there exists a constant $C_{\mathcal{M}} > 0$ such that $\text{Regret}(\texttt{alg}, \mathcal{M}, K) \leq C_{\mathcal{M}}K^{\alpha}$. The construction of proof is based on (Simchowitz & Jamieson, 2019).

Theorem 4.5 (Gap-dependent regret lower bound)**.** *There exists an MDP $\mathcal{M}$ such that any α-uniformly good algorithm `alg` satisfies*

$$\lim_{K\to\infty}\frac{\text{Regret}(\texttt{alg},\mathcal{M},K)}{\log K} = \Omega\left((1-\alpha)\sum_{(s,a):\Delta_1(s,a)>0}\frac{(c_\rho H)^2}{\Delta_1(s,a)}\right)$$

To our knowledge, Theorem 4.4 provides the first result showing that RSRL with DRMs can achieve $\log T$-type regret. The lower bound shows that for sufficiently large K, the logarithmic dependence on K is unavoidable. Notably, it implies that our algorithms have a tight dependency on A and K. Furthermore, the presence of a constant factor in both the upper and lower bounds suggests that the specific choice of risk measure can significantly impact their performance.

5. Proof Sketch of Theorem 4.1

For simplicity, we only provide the proof sketch for `UCBVI-DRM`. The proof structure builds upon the framework established in (Azar et al., 2017), but we introduce new techniques to address the specific challenges posed by nonlinear risk measures: (i) the Lipschitz continuity w.r.t. $\|\cdot\|_\infty$ together with the DKW inequality to ensure the optimism (step 1), and (ii) the Lipschitz continuity w.r.t. $\|\cdot\|_1$ together with a transport inequality (see Lemma 5.1) to obtain recursions of suboptimality gap.

Step 1: establish optimism. We first show that $V_1^k(s_1^k) \geq V_1^*(s_1^k), \forall k \in [K]$ with high probability. Using the Lipschitz property of ρ_h w.r.t. $\|\cdot\|_\infty$ and the DKW inequality (Fact 5)

$$\begin{aligned}&\rho_h\left(V_{h+1}^*, P_h(s,a)\right) - \rho_h\left(V_{h+1}^*, \hat{P}_h^k(s,a)\right)\\&\leq L_{\infty,H-h}\left\|\left(V_{h+1}^*, P_h(s,a)\right) - \left(V_{h+1}^*, \hat{P}_h^k(s,a)\right)\right\|_\infty\\&\leq L_{\infty,H-h}\sqrt{\frac{\iota}{2(N_h^k(s,a)\vee 1)}} = b_h^k(s,a).\end{aligned}$$

The results follows from the monotonicity of ρ_h and induction.

Step 2: regret decomposition. We define $\Delta_h^k \triangleq V_h^k - V_h^{\pi^k} \in \mathbb{R}^{\mathcal{S}}$ and $\delta_h^k \triangleq \Delta_h^k(s_h^k)$. The optimism implies that the regret can be bounded by the surrogate regret

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^K V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\\&\leq \sum_{k=1}^K V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) = \sum_{k=1}^K \delta_1^k.\end{aligned}$$

We write $r_h^k \triangleq r_h(s_h^k, a_h^k)$, $b_h^k \triangleq b_h^k(s_h^k, a_h^k)$, $N_h^k \triangleq N_h^k(s_h^k, a_h^k)$, $\hat{P}_h^k(s_h^k) \triangleq \hat{P}_h^k(s_h^k, a_h^k)$, and $P_h^{\pi^k} \triangleq P_h(s_h^k, a_h^k)$. We decompose δ_h^k as follows

$$\begin{aligned}\delta_h^k = b_h^k &+ \underbrace{\rho_h\left(V_{h+1}^k, \hat{P}_h^k\right) - \rho_h\left(V_{h+1}^k, P_h^{\pi^k}\right)}_{(a)}\\&+ \underbrace{\rho_h\left(V_{h+1}^k, P_h^{\pi^k}\right) - \rho_h\left(V_{h+1}^{\pi^k}, P_h^{\pi^k}\right)}_{(b)}.\end{aligned}$$

Bounding term (a) and term (b) requires new techniques compared with risk-neutral setting. To deal with the nonlinearity, we relate the two value difference terms to the distances between value distribution via Lipschitzness of the risk measure.To bound term (a), we use the Lipschitz property w.r.t. $\|\cdot\|_\infty$ to get

$$\begin{aligned}(a) &\leq L_{\infty,h}\left\|\left(V_{h+1}^k, \hat{P}_h^k\right) - \left(V_{h+1}^k, P_h^{\pi^k}\right)\right\|_\infty\\&\leq L_{\infty,h}\left\|\hat{P}_h^k - P_h^{\pi^k}\right\|_1 \leq L_{\infty,h}\cdot c_h^k.\end{aligned}$$

Due to the linearity of expectation, standard analysis bound term (b) by the sum of the vale gap at next step δ_{h+1}^k and a martingale noise, and then the recursion of the value gap is obtained. The derivation of a recursion in the presence of nonlinear ρ_h, however, leads to the main technical challenge. Our key observation to overcome the difficulty is a simple transport inequality in the following.

Lemma 5.1. $\|(x,P) - (y,P)\|_1 \leq \sum_{i=1}^n P_i\,|x_i - y_i|\,.$

Together with the Lipschitz property w.r.t. $\|\cdot\|_1$, we have

$$\begin{aligned}(b) &\leq L_{1,h} \left\| \left(V_{h+1}^k, P_h^{\pi^k}\right) - \left(V_{h+1}^{\pi^k}, P_h^{\pi^k}\right) \right\|_1 \\ &\leq L_{1,h} \sum_{s' \in \mathcal{S}} P_h^{\pi^k}(s') \left| V_{h+1}^k(s') - V_{h+1}^{\pi^k}(s') \right| \\ &= L_{1,h} \sum_{s' \in \mathcal{S}} P_h^{\pi^k}(s') \left(V_{h+1}^k(s') - V_{h+1}^{\pi^k}(s') \right) \\ &= L_{1,h} \cdot P_h^{\pi^k} \Delta_{h+1}^k \triangleq L_{1,h}(\epsilon_h^k + \delta_{h+1}^k),\end{aligned}$$

where $\epsilon_h^k \triangleq P_h^{\pi^k} \Delta_{h+1}^k - \Delta_{h+1}^k(s_{h+1}^k)$ is a martingale difference sequence, and the first equality is due to $V_{h+1}^k(s') \geq V_{h+1}^*(s') \geq V_{h+1}^{\pi^k}(s')$ for all s'. We bound δ_h^k recursively

$$\begin{aligned}\delta_h^k &\leq L_{\infty,h} \cdot c_h^k + L_{1,h}(\epsilon_h^k + \delta_{h+1}^k) + b_h^k \\ &\leq 2L_{\infty,h} \cdot c_h^k + L_{1,h}(\epsilon_h^k + \delta_{h+1}^k) + c_h^k.\end{aligned}$$

Step 3: putting together. Unrolling the recursion and summing up over K episodes yields

$$\begin{aligned}\text{Regret}(K) &\leq \sum_{k \in [K]} \delta_1^k \\ &\leq 2 \sum_{h=1}^{H-1} L_{\infty,h} \tilde{L}_{1,h-1} \sum_{k=1}^{K} c_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H-1} \tilde{L}_{1,h} \epsilon_h^k.\end{aligned}$$

We bound the first term via a pigeonhole argument and bound the second term by the Azuma-Hoeffding inequality. The finial result follows from a union bound.

6. Experiments

In this section, we provide some numerical results to validate the empirical performance of our algorithms. We compare our algorithms to the algorithms `UCBVI` (Azar et al., 2017) for risk-neutral RL and `RSVI2` (Fei et al., 2021) for RSRL with ERM.

In our experiments, we focus on an MDP with $S = 3$ states, A actions, and horizon H, which is similar to the construction in (Du et al., 2023). The major difference is that we consider a non-stationary MDP. The MDP consists of a fixed initial state denoted as state 0, and three additional states denoted as states 1, 2, and 3. In step $2 \leq h \leq H$, the three states generate reward 1,0 and 0.4, respectively. The agent starts from state 0 in the first step, takes action from $[A]$, and then transitions to one of three states $\{1, 2, 3\}$ in the next step. Any action in $[A - 1]$ leads to a uniform transition to state 1 and state 2. The optimal action A leads to a transition to state 2 and state 3 with probability 0.001 and 0.999.

We consider the dynamic CVaR with the homogeneous and in-homogeneous setting. For the homogeneous CVaR, the

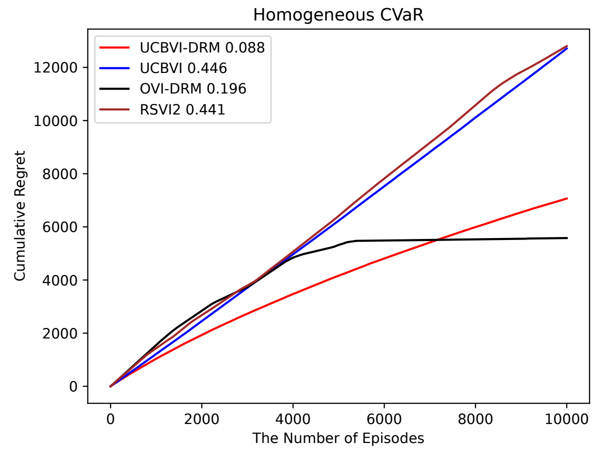

Figure 1. Comparison of different algorithms for homogeneous CVaR.

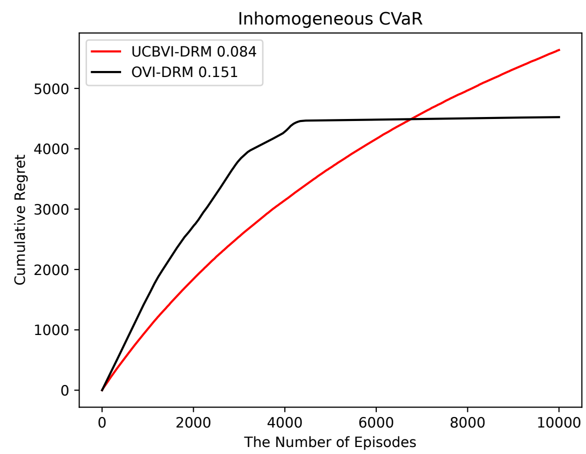

Figure 2. Comparison of different algorithms for in-homogeneous CVaR.

coefficients α at all steps are identical, while for the in-homogeneous CVaR, the coefficients α at different steps could be different. We set $\delta = 0.005$, $A = 5$, $H = 5$ and $K = 10000$. We set $\alpha = (0.05, 0.05, 0.05, 0.05)$ and $\alpha = (0.09, 0.08, 0.07, 0.05)$ for the homogeneous and in-homogeneous CVaR, respectively. We perform 5 independent runs for each algorithm.

As shown in Figure 1, `OVI-DRM` and `UCBVI-DRM` enjoy sublinear regret while the risk-neutral algorithm `UCBVI` and `RSVI2` suffer linear regret. In particular, `OVI-DRM` outperforms `UCBVI-DRM` because it achieves a better balance between exploration and exploitation. Figure 2 only plots the results for our algorithms. It shows that `UCBVI-DRM` can also learn the optimal policy in the in-homogeneous CVaR setting.

7. Conclusions

We propose two model-based algorithms for the broad class of Lipschitz DRMs. To establish the efficacy of our algorithms, we provide theoretical guarantees in the form of worst-case and gap-dependent regret upper bounds. To complement our upper bounds, we also establish regret lower bounds. These lower bounds demonstrate the inherent difficulty of the problem.

There are several promising future directions. It might be possible to improve the regret upper bounds by designing new algorithms or improving the analysis. Currently, our algorithms and analysis are primarily focused on tabular MDPs. However, extending the results to the setting of function approximation, such as linear function approximation, is an important and challenging task. The nonlinearity of risk measures poses a significant obstacle in this context. One potential approach to address this issue is to leverage techniques like value-targeted regression, as proposed in (Ayoub et al., 2020; Jia et al., 2020), and integrate them into our framework.

Acknowledgements

We thank all the anonymous reviewers for their helpful comments and suggestions. The work of Zhi-quan Luo was supported in part by the National Key Research and Development Project under grant 2022YFA1003900 and in part by the Guangdong Provincial Key Laboratory of Big Data Computing.

References

Acerbi, C. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

Acerbi, C. and Tasche, D. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Asienkiewicz, H. and Jaśkiewicz, A. A note on a new class of recursive utilities in markov decision processes. *Applicationes Mathematicae*, 44:149–161, 2017.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Balbás, A., Garrido, J., and Mayoral, S. Properties of distortion risk measures. *Methodology and Computing in Applied Probability*, 11(3):385–399, 2009.

Bastani, O., Ma, J. Y., Shen, E., and Xu, W. Regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 36259–36269, 2022.

Bäuerle, N. and Glauner, A. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966, 2022.

Ben-Tal, A. and Teboulle, M. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Bhat, S. P. and LA, P. Concentration of risk measures: A wasserstein distance approach. *Advances in neural information processing systems*, 32, 2019.

Bielecki, T. R., Pliska, S. R., and Sherris, M. Risk sensitive asset allocation. *Journal of Economic Dynamics and Control*, 24(8):1145–1177, 2000.

Chu, S. and Zhang, Y. Markov decision processes with iterated coherent risk measures. *International Journal of Control*, 87(11):2286–2293, 2014.

Coache, A. and Jaimungal, S. Reinforcement learning with dynamic convex risk measures. *arXiv preprint arXiv:2112.13414*, 2021.

Coache, A., Jaimungal, S., and Cartea, Á. Conditionally elicitable dynamic risk measures for deep reinforcement learning. *arXiv preprint arXiv:2206.14666*, 2022.

Davis, M. and Lleo, S. Risk-sensitive benchmarked asset management. *Quantitative Finance*, 8(4):415–426, 2008.

Delage, E. and Mannor, S. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Du, Y., Wang, S., and Huang, L. Provably efficient risk-sensitive reinforcement learning: Iterated cvar and worst path. In *The Eleventh International Conference on Learning Representations*, 2023.

Ernst, D., Stan, G.-B., Goncalves, J., and Wehenkel, L. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 667–672. IEEE, 2006.

Fei, Y. and Xu, R. Cascaded gaps: Towards logarithmic regret for risk-sensitive reinforcement learning. In *International Conference on Machine Learning*, pp. 6392–6417. PMLR, 2022.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446, 2021.

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

Föllmer, H. and Knispel, T. Convex risk measures: Basic facts, law-invariance and beyond, asymptotics for large portfolios. In *Handbook of the fundamentals of financial decision making: Part II*, pp. 507–554. World Scientific, 2013.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

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

Huang, A., Leqi, L., Lipton, Z., and Azizzadenesheli, K. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34: 23714–23726, 2021.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Lam, T., Verma, A., Low, B. K. H., and Jaillet, P. Risk-aware reinforcement learning with coherent risk measures and non-linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023.

Liang, H. and Luo, Z.-Q. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *arXiv preprint arXiv:2210.14051*, 2022.

Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Osogami, T. Iterated risk measures for risk-sensitive markov decision processes with discounted cost. *arXiv preprint arXiv:1202.3755*, 2012.

Prashanth, L. and Bhat, S. P. A wasserstein distance approach for concentration of empirical risk estimates. *The Journal of Machine Learning Research*, 23(1):10830–10890, 2022.

Rockafellar, R. T., Uryasev, S., et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Rudloff, B., Sass, J., and Wunderlich, R. Entropic risk constraints for utility maximization. *Festschrift in celebration of prof. Dr. Wilfried Grecksch's 60th birthday*, pp. 149–180, 2008.

Ruszczyński, A. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

Shen, Y., Stannat, W., and Obermayer, K. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Wang, K., Kallus, N., and Sun, W. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *arXiv preprint arXiv:2302.03201*, 2023.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Wirch, J. L. and Hardy, M. R. Distortion risk measures: Coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, pp. 15–17, 2001.

Xu, W., Gao, X., and He, X. Regret bounds for markov decision processes with recursive optimized certainty equivalents. *arXiv preprint arXiv:2301.12601*, 2023.

A. Risk Measures

A.1. Definitions

Conditional Value at Risk (CVaR) The CVaR value (Rockafellar et al., 2000) at level $\alpha \in (0, 1)$ for a distribution F is defined as

$$C_\alpha(F) \triangleq \sup_{\nu \in \mathbb{R}} \left\{ \nu - \frac{1}{\alpha} \mathbb{E}_{X \sim F}[(\nu - X)^+] \right\}.$$

(Acerbi & Tasche, 2002) showed that when F is a continuous distribution, $C_\alpha(F) = \mathbb{E}_{X \sim F}[X | X \leq F^{-1}(\alpha)]$.

Spectral risk measure (SRM) SRM is class of risk measures that generalizes CVaR via adopting a non-constant weighting function over the quantiles (Acerbi, 2002). The SRM value of F is defined as

$$S_\phi(F) \triangleq \int_0^1 \phi(y) F^{-1}(y) dy,$$

where $\phi : [0, 1] \to [0, \infty)$ is the weighting function. (Acerbi, 2002) showed that an SRM is coherent if ϕ is is decreasing and satisfies that $\int_0^1 \phi(y) dy = 1$. SRM can be viewed as a weighted average of the quantiles, with weight specified by $\phi(y)$. In fact, $S_\phi(F)$ specializes in $C_\alpha(F)$ for $\phi(y) = \frac{1}{\alpha} \mathbb{I}\{0 \leq y \leq \alpha\}$.

Distortion risk measure For a distribution $F \in \mathscr{D}([0, \infty))$, the distortion risk measure (Balbás et al., 2009; Wirch & Hardy, 2001) $\rho_g(F)$ is defined as

$$\rho_g(F) \triangleq \int_0^\infty g(1 - F(x)) dx,$$

where $g : [0, 1] \to [0, 1]$ is a continuous increasing function with $g(0) = 0$ and $g(1) = 1$. We refer to g as the distortion function. Distortion risk measure is coherent if and only if g is convex. Similar to SRM, distortion risk measure can also recover CVaR by choosing proper g.

Entropic risk measure (ERM) ERM adjusts the risk attitude of the user through the exponential utility function. In particular, the ERM value of F with coefficient $\beta \neq 0$ is defined as

$$U_\beta(F) \triangleq \frac{1}{\beta} \log(\mathbb{E}_{X \sim F}[\exp(\beta X)]) = \frac{1}{\beta} \log \left(\int_{\mathbb{R}} \exp(\beta x) dF(x) \right).$$

Notably, ERM is the prime example of a convex risk measure which is not coherent (Rudloff et al., 2008).

Optimized certainty equivalent (OCE) The OCE (Ben-Tal & Teboulle, 2007) value of F associated with a utility function u is given by

$$C_u(F) \triangleq \sup_\lambda \{\lambda + \mathbb{E}_{X \sim F}[u(X - \lambda)]\} = \sup_\lambda \left\{ \lambda + \int_{\mathbb{R}} u(x - \lambda) dF(x) \right\},$$

where u is a non-decreasing, closed utility function that satisfies $u(0) = 0$ and $1 \in \partial u(0)$. THE OCE is risk-averse (risk-seeking) if and only u is concave (convex). OCE subsumes important examples of popular risk measures, including the ERM and CVaR.

A.2. Lipschitz Property

We summarizes the Lipschitz constants of common risk measures over a finite interval $[a, b]$ in Table 2. (Prashanth & Bhat, 2022) provides the Lipschitz constants of SRM, OCE, and distortion risk measure w.r.t. the Wasserstein distance or $\|\cdot\|_1$. (Huang et al., 2021) provides the Lipschitz constants of distortion risk measure w.r.t. $\|\cdot\|_\infty$.

For completeness, we will derive the Lipschitz constants of SRM and OCE w.r.t. $\|\cdot\|_\infty$ in the following. Fact 1 offers a simple way to derive the Lipschitz constant of a risk measure w.r.t. $\|\cdot\|_1$ based on that w.r.t. the $\|\cdot\|_\infty$. Therefore, the Lipschitz constants of SRM w.r.t. $\|\cdot\|_\infty$ can take $L_\infty(M_\phi, [a, b]) = (b - a) \cdot L_1(M_\phi, [a, b]) = (b - a) \max |\phi(x)|$. As a special case, we have $L_\infty(C_\alpha, [0, M]) = \frac{b-a}{\alpha}$.

Fact 1. If a functional ρ has Lipschitz constant $L_1([a,b])$ over $\mathscr{D}([a,b])$, then it has Lipschitz constant $L_\infty([a,b]) = L_1([a,b])(b-a)$.

Proof. Suppose ρ has Lipschitz constant $L_1([a,b])$ over $\mathscr{D}([a,b])$, then

$$|\rho(F)-\rho(G)| \leq L_1([a,b]) \left\|F-G\right\|_1 \leq L_1([a,b]) \left\|F-G\right\|_\infty (b-a), \forall F,G \in \mathscr{D}([a,b]).$$

This implies that $L_\infty([a,b]) = L_1([a,b])(b-a)$ is a valid choice. □

Fact 2. The Lipschitz constants of OCE w.r.t. $\left\|\cdot\right\|_\infty$ is $L_\infty(C_u,[a,b]) = -u(a-b)$ for concave utility function and $L_\infty(C_u,[a,b]) = u(b-a)$ for convex utility function.

Proof. Let $\lambda_1, \lambda_2 \in [a,b]$ satisfy

$$C_u(F) = \lambda_1 + \int_a^b u(x-\lambda_1)dF(x) = \max_{\lambda\in[a,b]} \lambda + \int_a^b u(x-\lambda)dF(x)$$
$$C_u(G) = \lambda_2 + \int_a^b u(x-\lambda_2)dG(x) = \max_{\lambda\in[a,b]} \lambda + \int_a^b u(x-\lambda)dG(x).$$

Without loss generality, we assume $C_u(F) > C_u(G)$. It holds that

$$\begin{aligned}
C_u(F) - C_u(G) &= \lambda_1 + \int_a^b u(x-\lambda_1)dF(x) - \lambda_2 - \int_a^b u(x-\lambda_2)dG(x) \\
&\leq \lambda_1 + \int_a^b u(x-\lambda_1)dF(x) - \lambda_1 - \int_a^b u(x-\lambda_1)dG(x) \\
&= \int_a^b u(x-\lambda_1)dF(x) - \int_a^b u(x-\lambda_1)dG(x) \\
&= u(x-\lambda_1)F(x)|_a^b - \int_a^b F(x)du(x-\lambda_1) - u(x-\lambda_1)G(x)|_a^b + \int_a^b G(x)du(x-\lambda_1) \\
&= \int_a^b (G(x)-F(x))du(x-\lambda_1) \\
&\leq \int_a^b du(x-\lambda_1) \cdot \left\|F-G\right\|_\infty \\
&= (u(b-\lambda_1) - u(a-\lambda_1)) \left\|F-G\right\|_\infty \\
&\leq \max_{\lambda\in[a,b]} (u(b-\lambda) - u(a-\lambda)) \left\|F-G\right\|_\infty = L_\infty(C_u,[a,b]) \left\|F-G\right\|_\infty,
\end{aligned}$$

where the second inequality is due to that u is non-decreasing. For concave utility function, we can bound the last term as

$$\max_{\lambda\in[a,b]} (u(b-\lambda) - u(a-\lambda)) = u(b-b) - u(a-b) = -u(a-b).$$

For convex utility function, we can bound the last term as

$$\max_{\lambda\in[a,b]} (u(b-\lambda) - u(a-\lambda)) = u(b-a) - u(a-a) = u(b-a).$$

□

Fact 3. The Lipschitz constants of OCE w.r.t. $\left\|\cdot\right\|_1$ is $L_1(C_u,[a,b]) = u'(a-b)$ for concave utility function and $L_1(C_u,[a,b]) = u'(b-a)$ for convex utility function.

Proof. Let $\lambda_1, \lambda_2 \in [a,b]$ satisfy

$$C_u(F) = \lambda_1 + \int_a^b u(x-\lambda_1)dF(x) = \max_{\lambda\in[a,b]} \lambda + \int_a^b u(x-\lambda)dF(x)$$
$$C_u(G) = \lambda_2 + \int_a^b u(x-\lambda_2)dG(x) = \max_{\lambda\in[a,b]} \lambda + \int_a^b u(x-\lambda)dG(x).$$

Table 2. Lipschitz constants of common risk measures

Lipschitz constant	CVaR	SRM	distortion risk measure	OCE[4]
$L_1([a,b])$	$\frac{1}{\alpha}$	$\max \phi(x)$	$\max g'(x)$	$u'(a-b)$
$L_\infty([a,b])$	$\frac{b-a}{\alpha}$	$(b-a)\max \phi(x)$	$(b-a)\max g'(x)$	$-u(a-b)$

Algorithm 3 `OM`

1: Input: $P = (P(s_1), \dots, P(s_S)), V = (V(s_1), \dots, V(s_S))$ and $c > 0$
2: Sorting: let $V' = (V(s_{(1)}), \cdots, V(s_{(S)}))$ such that $V(s_{(1)}) \leq V(s_{(2)}) \leq \cdots \leq V(s_{(S)})$
3: Permutation: let $P' = (P(s_{(1)}), \cdots, P(s_{(S)}))$
4: Transport: sequentially move probability mass $\frac{c}{2} \wedge 1$ of the leftmost states to $s_{(S)}$

Without loss generality, we assume $C_u(F) > C_u(G)$. It holds that

$$\begin{aligned}
C_u(F) - C_u(G) &= \lambda_1 + \int_a^b u(x-\lambda_1)dF(x) - \lambda_2 - \int_a^b u(x-\lambda_2)dG(x) \\
&\leq \int_a^b u(x-\lambda_1)dF(x) - \int_a^b u(x-\lambda_1)dG(x) \\
&= \int_a^b (G(x)-F(x))du(x-\lambda_1) \\
&= \int_a^b (G(x)-F(x))u'(x-\lambda_1)dx \\
&\leq \max_{\lambda\in[a,b],x\in[a,b]} u'(x-\lambda) \int_a^b |G(x)-F(x)|dx \\
&= L_1(C_u,[a,b]) \left\|F-G\right\|_1,
\end{aligned}$$

where the second inequality is due to the non-negativity of u'. For concave utility function, we can bound the last term as

$$\max_{\lambda\in[a,b],x\in[a,b]} u'(x-\lambda) = u'(a-b).$$

For convex utility function, we can bound the last term as

$$\max_{\lambda\in[a,b],x\in[a,b]} u'(x-\lambda) = u'(b-a).$$

□

OCE subsumes ERM when $u(x) = \frac{\exp(\beta x)-1}{\beta}$. In particular, $L_\infty(U_\beta,[a,b]) = -u(a-b) = -\frac{\exp(\beta(a-b))-1}{\beta} = \frac{\exp(|\beta|(b-a))-1}{|\beta|}$ for concave utility ($\beta < 0$) and $L_\infty(U_\beta,[a,b]) = u(b-a) = \frac{\exp(\beta(b-a))-1}{\beta}$ for convex utility ($\beta > 0$).

B. Subroutine

We present the subroutine `OM` used in Algorithm 2 in this section. Fix an (s,a,k,k), `OM` takes the empirical model $\hat{P}_h^k(s,a)$, the value at the next step V_{h+1}^k, and a confidence radius $c_h^k(s,a)$ as input and outputs the optimistic model $\tilde{P}_h^k(s,a)$. For a PMF P and a real number $c > 0$, denote by $B_1(P,c) \triangleq \{P' | \left\|P'-P\right\|_1 \leq c\}$ the ℓ_1 norm ball centered at P with radius c.

Recall that we $F \succeq G$ denotes that $F(x) \leq G(x), \forall x \in \mathbb{R}$. Lemma B.1 shows that `OM` can output an optimistic model $\tilde{P}$ whose value distribution dominates those generated by the model within the concentration ball.

Lemma B.1. *Let P, V, c be the input of `OM` and $\tilde{P}$ be the output. It holds that*

$$(\tilde{P}, V) \succeq (Q, V), \forall Q \in B_1(P,c).$$

Proof. For simplicity, let $P = (P_1, \cdots, P_n)$, $V \in \mathbb{R}^n$ satisfying $V_1 \leq V_2 \cdots \leq V_n$. Observe that the CDF (P, V) is a piecewise constant function. Hence it suffices to show that

$$\sum_{j=1}^{i} \tilde{P}_j \leq \sum_{j=1}^{i} Q_j, \forall i \in [n], \forall Q \in B_1(P, c).$$

Let $l \triangleq \min\left\{i | \sum_{j=1}^{i} P_j \geq \frac{c}{2}\right\}$. There are two cases.

Case 1: $P_n + \frac{c}{2} \leq 1$. Since $\sum_{j=1}^{l-1} P_j < \frac{c}{2}$ and $\sum_{j=1}^{l} P_j \geq \frac{c}{2}$, we have

$$\tilde{P}_i = \begin{cases} 0, i \in [l-1] \\ \sum_{j=1}^{l} P_j - \frac{c}{2}, i = l \\ P_i, l+1 \leq i \leq n-1 \\ P_n + \frac{c}{2}, i = n \end{cases}$$

and thus

$$\sum_{j=1}^{i} \tilde{P}_j = \begin{cases} 0, i \in [l-1] \\ \sum_{j=1}^{l} P_j - \frac{c}{2}, l \leq i \leq n-1 \\ 1, i = n \end{cases}$$

For any $Q \in B_1(P, c)$, it holds that

$$\sum_{j=1}^{i} \tilde{P}_j \leq \sum_{j=1}^{i} Q_j, \forall i \in [n].$$

Otherwise $\sum_{j=1}^{k} Q_j < \sum_{j=1}^{k} P_j - \frac{c}{2}$ for some $l \leq k \leq n-1$, which implies $\sum_{j=k+1}^{n} Q_j = 1 - \sum_{j=1}^{k} Q_j > 1 - \sum_{j=1}^{k} P_j + \frac{c}{2} = \sum_{j=1}^{k} P_j + \frac{c}{2}$. This leads to a contradiction $\|P - Q\|_1 = \sum_{j \in [n]} |P_j - Q_j| \geq |\sum_{j=1}^{k} P_j - \sum_{j=1}^{k} Q_j| + |\sum_{j=k+1}^{n} P_j - \sum_{j=k+1}^{n} Q_j| > c$.

Case 2: $P_n + \frac{c}{2} > 1$. In this case, $\tilde{P}_j = 0$ for $j \in [n-1]$ and $\tilde{P}_n = 1$. It is obvious that

$$\sum_{j=1}^{i} \tilde{P}_j \leq \sum_{j=1}^{i} Q_j, \forall i \in [n], \forall Q \in B_1(P, c).$$

Therefore, we have

$$(\tilde{P}, V) \succeq (Q, V), \forall Q \in B_1(P, c).$$

□

Lemma B.1 together with the monotonicity of ρ_h implies that the output $\tilde{P}_h^k(s, a)$ satisfies

$$\rho_h(\tilde{P}_h^k(s, a), V_{h+1}^k) \geq \rho_h(P', V_{h+1}^k),\ \forall P' \in B_1(\hat{P}_h^k(s, a), c_h^k(s, a)).$$

In Appendix C, we will show that $P_h(s, a) \in B_1(\hat{P}_h^k(s, a), c_h^k(s, a))$ with high probability. Suppose $V_{h+1}^k(s) \geq V_{h+1}^*(s), \forall s$. It follows that

$$\begin{aligned} Q_h^k(s, a) &= r_h(s, a) + \rho_h(\tilde{P}_h^k(s, a), V_{h+1}^k) \geq r_h(s, a) + \rho_h(P_h(s, a), V_{h+1}^k) \\ &\geq r_h(s, a) + \rho_h(P_h(s, a), V_{h+1}^*) = Q_h^*(s, a), \forall (s, a). \end{aligned}$$

The second inequality is due to the monotonicity of ρ_h together with the fact that

$$V_{h+1}^k \geq V_{h+1}^* \Longrightarrow (P_h(s, a), V_{h+1}^k) \succeq (P_h(s, a), V_{h+1}^*).$$

Then we have $V_h^k(s) = \max Q_h^k(s, a) \geq \max Q_h^*(s, a) = V_h^*(s), \forall s$. By induction, we obtain

$$V_h^k(s) \geq V_h^*(s), \forall (k, h, s).$$

This implies that the value functions induced by the optimistic models are indeed optimistic compared to the optimal value functions. The computational complexity of `OM` is $O(S \log(S))$, since the computational complexity of each step is $O(S \log(S))$, $O(S)$, and $O(S)$.

C. Proofs of Regret Upper Bounds

C.1. Worst-case Regret Upper Bound

We first prove the worst-case regret upper bound for Algorithm 1.

C.1.1. Proof for Algorithm 1

Step 1: verify optimism. Fix an arbitrary $\delta \in (0,1)$. Define the good event $\mathcal{G}_\delta$ as

$$\mathcal{G}_\delta \triangleq \left\{ \left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{2S}{N_h^k(s,a) \vee 1}\iota}, \forall (s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \right\},$$

under which the empirical model concentrates around the true model under $\|\cdot\|_1$.

Lemma C.1 (High probability good event)**.** *The event $\mathcal{G}_\delta$ holds with probability at least* $1-\delta$.

Fact 4 (ℓ_1 concentration bound, (Weissman et al., 2003)). Let P be a probability distribution over a finite discrete measurable space $(\mathcal{X}, \Sigma)$. Let $\widehat{P}_n$ be the empirical distribution of P estimated from n samples. Then with probability at least $1-\delta$,

$$\left\| \widehat{P}_n - P \right\|_1 \leq \sqrt{\frac{2|\mathcal{X}|}{n} \log \frac{1}{\delta}}.$$

Lemma C.1 does not directly follow from a union bound together with Fact 4 since the case $N_h^k(s,a) = 0$ need to be checked.

Proof. Fix some $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$. If $N_h^k(s,a) = 0$, then we have $\hat{P}_h^k(\cdot|s,a) = \frac{1}{S}\mathbf{1}$. A simple calculation yields that for any $P_h(\cdot|s,a)$

$$\left\| \frac{1}{S}\mathbf{1} - P_h(\cdot|s,a) \right\|_1 \leq 2 \leq \sqrt{2S \log(1/\delta)}.$$

It follows that

$$\mathbb{P}\left(\left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{2S}{N_h^k(s,a) \vee 1} \log(1/\delta)} \,\middle|\, N_h^k(s,a) = 0 \right) = 1 > 1 - \delta.$$

The event is true for the unseen state-action pairs. Now we consider the case that $N_h^k(s,a) > 0$. By Fact 4 , we have that for any integer $n \geq 1$

$$\mathbb{P}\left(\left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{2S}{N_h^k(s,a)} \log(1/\delta)} \,\middle|\, N_h^k(s,a) = n \right) \geq 1 - \delta.$$

Thus we have

$$\begin{aligned}
&\mathbb{P}\left(\left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{2S \log(1/\delta)}{N_h^k(s,a)}} \right) \\
&= \sum_{n=0,1,\cdots} \mathbb{P}\left(\left\| \hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{2S \log(1/\delta)}{N_h^k(s,a) \vee 1}} \,\middle|\, N_h^k(s,a) = n \right) \mathbb{P}(N_h^k(s,a) = n) \\
&\geq (1-\delta) \sum_{n=0,1,\cdots} \mathbb{P}(N_h^k(s,a) = n) = 1 - \delta.
\end{aligned}$$

Applying a union bound over all (s,a,k,h) and rescaling δ leads to the result. □

Lemma C.2 (Range of V^*)**.** *For any MDP, it holds that $V_h^*(s) \in [0, H+1-h]$ for all $(s,h) \in \mathcal{S} \times [H+1]$.*

Proof. The proof follows from induction and Assumption 3.1. Observe that $V_{H+1}^* = 0$. Suppose $V_{h+1}^*(s) \in [0, H-h]$ for any s, then we have

$$0 \le Q_h^*(s,a) = r_h(s,a) + \rho_h(V_{h+1}^*, P_h(s,a)) \le 1 + H - h,$$

where the inequalities are due to the Assumption 3.1. Then we have $V_h^*(s) = \max_a Q_h^*(s,a) \in [0, H+1-h]$. The induction is completed. □

Fact 5 (DKW inequality for discrete distribution). Let $\hat{P}_n$ be the empirical PMF for (x, P) with n samples, then w.p. at least $1-\delta$

$$\left\|(x,P) - (x, \hat{P}_n)\right\|_\infty \le \sqrt{\frac{\log(2/\delta)}{2n}}.$$

We remark that we can also derive a bound by Fact 6: w.p. $1-\delta$

$$\left\|(x,P) - (x, \hat{P}_n)\right\|_\infty \le \left\|P - \hat{P}_n\right\|_1 \le \sqrt{\frac{2m\log(2/\delta)}{n}}.$$

However, this bound is looser than that from Fact 5 with a factor of $\sqrt{m}$.

Lemma C.3 (Optimistic value function)**.** *Conditioned on event $\mathcal{G}_\delta$, the sequence $\{V_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 1 satisfies $V_1^k(s_1^k) \ge V_1^*(s_1^k), \forall k \in [K]$.*

Proof. The proof follows from induction. Fix $k \in [K]$. It is evident that the inequality holds when $h = H+1$. Suppose the inequality holds for $h+1$. It follows that for any (s,a)

$$\begin{aligned} Q_h^k(s,a) &= r_h(s,a) + \rho_h\left(V_{h+1}^k, \hat{P}_h^k(s,a)\right) + b_h^k(s,a) \\ &\ge r_h(s,a) + \rho_h\left(V_{h+1}^*, \hat{P}_h^k(s,a)\right) + b_h^k(s,a) \\ &\ge r_h(s,a) + \rho_h(V_{h+1}^*, P_h(s,a)) = Q_h^*(s,a). \end{aligned}$$

The first inequality is due to the monotonicity of ρ_h and the induction hypothesis, and the second one follows from that

$$\begin{aligned} \rho_h(V_{h+1}^*, P_h(s,a)) - \rho_h(V_{h+1}^*, \hat{P}_h^k(s,a)) &\le L_\infty(\rho_h, H-h)\left\|(V_{h+1}^*, P_h(s,a)) - \left(V_{h+1}^*, \hat{P}_h^k(s,a)\right)\right\|_\infty \\ &\le L_\infty(\rho_h, H-h)\sqrt{\frac{\iota}{2(N_h^k(s,a)\vee 1)}} = b_h^k(s,a), \end{aligned}$$

where the first inequality follows from the Lipschitz property of ρ_h and Lemma C.2, and the second one is due to the DKW inequality (Fact 5). □

Fact 6. Let (x,P) and (x,Q) be two discrete distributions with the same support $x = (x_1, \cdots, x_m)$ and F, G be their CDFs respectively. It holds that

$$\|F - G\|_\infty \le \|P - Q\|_1.$$

Proof. Without loss of generality, we assume that $x_1 \le x_2 \cdots \le x_n$. By definition,

$$\begin{aligned} \|F-G\|_\infty &= \sup_{x\in\mathbb{R}} |F(x) - G(x)| = \max_{i\in[n]} |F(x_i) - G(x_i)| = \max_{i\in[n]} \left|\sum_{j\in[i]} P_j - \sum_{j\in[i]} Q_j\right| \\ &\le \max_{i\in[n]} \sum_{j\in[i]} |P_j - Q_j| = \sum_{j\in[n]} |P_j - Q_j| = \|P-Q\|_1. \end{aligned}$$

The second equality is due to that F and G are piecewise constant functions that only differ at $x_1, \cdots, x_n$. This would lead a worse bonus term with a factor of $\sqrt{S}$. □

Remark C.4. Alternatively, we have

$$\begin{aligned}\rho_h(V^*_{h+1}, P_h(s,a)) - \rho_h(V^*_{h+1}, \hat{P}^k_h(s,a)) &\leq L_\infty(\rho_h, H-h) \left\| (V^*_{h+1}, P_h(s,a)) - \left(V^*_{h+1}, \hat{P}^k_h(s,a)\right) \right\|_\infty \\ &\leq L_\infty(\rho_h, H-h) \left\| P_h(s,a) - \hat{P}^k_h(s,a) \right\|_1 \\ &\leq L_\infty(\rho_h, H-h) \sqrt{\frac{2S}{(N^k_h(s,a) \vee 1)}\iota},\end{aligned}$$

where the second inequality is due to Fact 6, the third inequality is due to Lemma C.1.

Step 2: regret decomposition. We introduce the key technical lemma here.

Lemma C.5. *Let (x, P) and (y, P) be two discrete distributions, where $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$. It holds that*

$$\|(x,P) - (y,P)\|_1 \leq \sum_{i \in [n]} P_i \left| x_i - y_i \right|.$$

Proof. By the definition of Wasserstein distance between two discrete distributions, we have

$$\begin{aligned}\|F - G\|_1 &= \inf_{\sum_j \lambda_{i,j} = P_i, \sum_i \lambda_{i,j} = P_j} \sum_i \sum_j \lambda_{i,j} \left| x_i - y_j \right| \\ &\leq \sum_i \sum_j \delta_{i,j} P_i \left| x_i - y_j \right| \\ &= \sum_i P_i \sum_j \delta_{i,j} \left| x_i - y_i \right| \\ &= \sum_i P_i \left| x_i - y_i \right|.\end{aligned}$$

The inequality holds since $\{\delta_{i,j} P_i\}_{i,j}$ is a valid coupling

$$\sum_j \delta_{i,j} P_i = P_i, \ \sum_i \delta_{i,j} P_i = P_j.$$

□

We define $\Delta^k_h \triangleq V^k_h - V^{\pi^k}_h \in [-(H+1-h), H+1-h]^{\mathcal{S}}$ and $\delta^k_h \triangleq \Delta^k_h(s^k_h)$. For any (s,h) and any π, we let $P^\pi_h(\cdot|s) \triangleq P_h(\cdot|s, \pi_h(s))$. The regret can be bounded as

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^K V^*_1(s^k_1) - V^{\pi^k}_1(s^k_1) = \sum_{k=1}^K V^*_1(s^k_1) - V^k_1(s^k_1) + V^k_1(s^k_1) - V^{\pi^k}_1(s^k_1) \\ &\leq \sum_{k=1}^K V^k_1(s^k_1) - V^{\pi^k}_1(s^k_1) = \sum_{k=1}^K \delta^k_1.\end{aligned}$$

For simplicity, we write $r^k_h \triangleq r_h(s^k_h, \pi^k_h(s^k_h))$, $b^k_h \triangleq b^k_h(s^k_h, \pi^k_h(s^k_h))$, $N^k_h \triangleq N^k_h(s^k_h, \pi^k_h(s^k_h))$ and $\hat{P}^k_h(s^k_h) \triangleq \hat{P}^k_h(s^k_h, \pi^k_h(s^k_h))$. For any $h \in [H-1]$, we decompose δ^k_h as follows

$$\begin{aligned}\delta^k_h &= \rho_h\left(V^k_{h+1}, \hat{P}^k_h(s^k_h)\right) + b^k_h - \rho_h\left(V^{\pi^k}_{h+1}, P^{\pi^k}_h(s^k_h)\right) \\ &= \underbrace{\rho_h\left(V^k_{h+1}, \hat{P}^k_h(s^k_h)\right) - \rho_h\left(V^k_{h+1}, P^{\pi^k}_h(s^k_h)\right)}_{(a)} + \underbrace{\rho_h\left(V^k_{h+1}, P^{\pi^k}_h(s^k_h)\right) - \rho_h\left(V^{\pi^k}_{h+1}, P^{\pi^k}_h(s^k_h)\right)}_{(b)} + b^k_h.\end{aligned}$$

Using the Lipschitz property of ρ_h,

$$
\begin{aligned}
(a) &\le L_\infty(\rho_h, H-h) \left\| \left(V_{h+1}^k, \hat{P}_h^k(s_h^k)\right) - \left(V_{h+1}^k, P_h^{\pi^k}(s_h^k)\right) \right\|_\infty \\
&\le L_\infty(\rho_h, H-h) \left\| \hat{P}_h^k(s_h^k) - P_h^{\pi^k}(s_h^k) \right\|_1 \\
&\le L_\infty(\rho_h, H-h) \sqrt{\frac{2S}{N_h^k \vee 1} \iota}.
\end{aligned}
$$

Applying Lemma C.5 yields that

$$
\begin{aligned}
(b) &\le L_1(\rho_h, H-h) \left\| \left(V_{h+1}^k, P_h^{\pi^k}(s_h^k)\right) - \left(V_{h+1}^{\pi^k}, P_h^{\pi^k}(s_h^k)\right) \right\|_1 \\
&\le L_1(\rho_h, H-h) \sum_{s' \in \mathcal{S}} P_h^{\pi^k}(s'|s_h^k) \left| V_{h+1}^k(s') - V_{h+1}^{\pi^k}(s') \right| \\
&= L_1(\rho_h, H-h) \sum_{s' \in \mathcal{S}} P_h^{\pi^k}(s'|s_h^k) \left(V_{h+1}^k(s') - V_{h+1}^{\pi^k}(s') \right) \\
&= L_1(\rho_h, H-h) \left[P_h^{\pi^k} \Delta_{h+1}^k \right] (s_h^k) \\
&\triangleq L_1(\rho_h, H-h)(\epsilon_h^k + \delta_{h+1}^k),
\end{aligned}
$$

where $\epsilon_h^k \triangleq [P_h^{\pi^k} \Delta_{h+1}^k](s_h^k) - \Delta_{h+1}^k(s_{h+1}^k)$ is a martingale difference sequence with $\epsilon_h^k \in [-2(H-h), 2(H-h)]$ a.s. for all $(k,h) \in [K] \times [H]$. The first equality is due to that $V_{h+1}^k(s') \ge V_{h+1}^*(s') \ge V_{h+1}^{\pi^k}(s')$ for all s'.

Now we can bound δ_h^k recursively

$$
\begin{aligned}
\delta_h^k &\le L_\infty(\rho_h, H-h) \sqrt{\frac{2S}{N_h^k \vee 1} \iota} + L_1(\rho_h, H-h)(\epsilon_h^k + \delta_{h+1}^k) + L_\infty(\rho_h, H-h) \sqrt{\frac{\iota}{2(N_h^k \vee 1)}} \\
&\le 2 L_\infty(\rho_h, H-h) e_h^k + L_1(\rho_h, H-h)(\epsilon_h^k + \delta_{h+1}^k),
\end{aligned}
$$

where we define $e_h^k \triangleq \sqrt{\frac{2S}{N_h^k \vee 1} \iota}$ in the last line. Repeating the procedure, we obtain

$$
\begin{aligned}
\delta_1^k &\le 2 \sum_{h=1}^{H-1} L_\infty(\rho_h, H-h) \prod_{i=1}^{h-1} L_1(\rho_i, H-i) e_h^k + \sum_{h=1}^{H-1} \prod_{i=1}^{h} L_1(\rho_i, H-i) \epsilon_h^k + \prod_{h=1}^{H-1} L_1(\rho_h, H-h) \delta_H^k \\
&= 2 \sum_{h=1}^{H-1} L_\infty(\rho_h, H-h) \prod_{i=1}^{h-1} L_1(\rho_i, H-i) e_h^k + \sum_{h=1}^{H-1} \prod_{i=1}^{h} L_1(\rho_i, H-i) \epsilon_h^k,
\end{aligned}
$$

where the last step is because $\delta_H^k = Q_H^k - Q_H^* = r_H - r_H = 0$.

Step 3: putting together. The total regret is bounded as

$$
\text{Regret}(K) \le \sum_{k \in [K]} \delta_1^k \le 2 \sum_{h=1}^{H-1} L_\infty(\rho_h, H-h) \prod_{i=1}^{h-1} L_1(\rho_i, H-i) \sum_{k=1}^{K} e_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H-1} \prod_{i=1}^{h} L_1(\rho_i, H-i) \epsilon_h^k.
$$

The first term can be bounded as

$$\begin{aligned}
2\sum_{h=1}^{H-1} L_\infty(\rho_h, H-h)\prod_{i=1}^{h-1} L_1(\rho_i, H-i)\sum_{k=1}^{K} e_h^k &= 2\sum_{h=1}^{H-1} L_\infty(\rho_h, H-h)\prod_{i=1}^{h-1} L_1(\rho_i, H-i)\sum_{k=1}^{K}\sqrt{\frac{2S}{N_h^k \vee 1}\iota} \\
&\leq 4\sum_{h=1}^{H-1} L_\infty(\rho_h, H-h)\prod_{i=1}^{h-1} L_1(\rho_i, H-i)\sqrt{S^2AK\iota} \\
&\triangleq 4\sum_{h=1}^{H-1} L_{\infty,h}\prod_{i=1}^{h-1} L_{1,i}\sqrt{S^2AK\iota} \\
&\triangleq 4\sum_{h=1}^{H-1} L_{\infty,h}\tilde{L}_{1,h-1}\sqrt{S^2AK\iota},
\end{aligned}$$

where we denote by $L_{\infty,h} = L_\infty(\rho_h, H-h)$ and $\tilde{L}_{1,h-1} = \prod_{i=1}^{h-1} L_{1,i}$ for simplicity. We can bound the second term by Azuma-Hoeffding inequality: with probability at least $1-\delta'$, the following holds

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H-1}\prod_{i=1}^{h} L_1(\rho_i, H-i)\epsilon_h^k = \sum_{k=1}^{K}\sum_{h=1}^{H-1}\tilde{L}_{1,h}\epsilon_h^k &\leq \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H-1}\frac{(2(H-h)\tilde{L}_{1,h})^2}{2}\log(1/\delta')} \\
&= \sqrt{\sum_{h=1}^{H-1}(H-h)^2\tilde{L}_{1,h}^2}\sqrt{2K\log(1/\delta')}
\end{aligned}$$

Using a union bound and let $\delta = \delta' = \frac{\tilde{\delta}}{2}$, we have that with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{Regret}(K) &\leq 4\sum_{h=1}^{H-1} L_{\infty,h}\tilde{L}_{1,h-1}\sqrt{S^2AK\iota} + \sqrt{\sum_{h=1}^{H-1}(H-h)^2\tilde{L}_{1,h}^2}\sqrt{2K\log(1/\delta')} \\
&= \tilde{\mathcal{O}}\left(\sum_{h=1}^{H-1} L_{\infty,h}\tilde{L}_{1,h-1}\sqrt{S^2AK}\right).
\end{aligned}$$

The equality is due to that

$$\begin{aligned}
\sum_{h=1}^{H-1} L_{\infty,h}\tilde{L}_{1,h-1} \geq \sqrt{\sum_{h=1}^{H-1} L_{\infty,h}^2\tilde{L}_{1,h-1}^2} &= \sqrt{\sum_{h=1}^{H-1}((H-h)L_{1,h})^2\tilde{L}_{1,h-1}^2} \\
&= \sqrt{\sum_{h=1}^{H-1}(H-h)^2\tilde{L}_{1,h}^2},
\end{aligned}$$

where the first inequality comes from the non-negativity of $L_{\infty,h}\tilde{L}_{1,h-1}$, and the first equality is due to the choice $L_{\infty,h} = L_{1,h}(H-h)$.

Remark C.6. The following statement is not true

$$\|(x,P)-(y,P)\|_1 \leq \left|\sum_{i\in[n]} P_i(x_i - y_i)\right|.$$

Consider the case that $(x,P) = ((0,1),(\frac{1}{3},\frac{2}{3})$ and $(y,P) = ((\frac{1}{3},\frac{5}{6}),(\frac{1}{3},\frac{2}{3}))$. A simple calculation yields that $\sum_{i\in[n]} P_i(x_i - y_i) = 0$.

C.1.2. PROOF FOR ALGORITHM 2

Step 1: verify optimism.

Lemma C.7 (Optimistic value function)**.** *Conditioned on event $\mathcal{G}_\delta$, the sequence $\{V_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 2 satisfies $V_1^k(s_1^k) \geq V_1^*(s_1^k), \forall k \in [K]$.*

Proof. The proof follows from Appendix B. □

Step 2: regret decomposition. The regret can be bounded as

$$\text{Regret}(K) = \sum_{k=1}^K V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \leq \sum_{k=1}^K V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) = \sum_{k=1}^K \delta_1^k.$$

For any $h \in [H-1]$, we decompose δ_h^k as follows

$$\begin{aligned}\delta_h^k &= \rho_h\left(V_{h+1}^k, \tilde{P}_h^k(s_h^k)\right) - \rho_h\left(V_{h+1}^{\pi^k}, P_h^{\pi^k}(s_h^k)\right)\\ &= \underbrace{\rho_h\left(V_{h+1}^k, \tilde{P}_h^k(s_h^k)\right) - \rho_h\left(V_{h+1}^k, P_h^{\pi^k}(s_h^k)\right)}_{(a)} + \underbrace{\rho_h\left(V_{h+1}^k, P_h^{\pi^k}(s_h^k)\right) - \rho_h\left(V_{h+1}^{\pi^k}, P_h^{\pi^k}(s_h^k)\right)}_{(b)}.\end{aligned}$$

Using the Lipschitz property of ρ_h,

$$\begin{aligned}(a) &\leq L_\infty(\rho_h, H-h)\left\|\left(V_{h+1}^k, \tilde{P}_h^k(s_h^k)\right) - \left(V_{h+1}^k, P_h^{\pi^k}(s_h^k)\right)\right\|_\infty\\ &\leq L_\infty(\rho_h, H-h)\left\|\tilde{P}_h^k(s_h^k) - P_h^{\pi^k}(s_h^k)\right\|_1\\ &\leq L_\infty(\rho_h, H-h)\left(\left\|\tilde{P}_h^k(s_h^k) - \hat{P}_h^{\pi^k}(s_h^k)\right\|_1 + \left\|\hat{P}_h^k(s_h^k) - P_h^{\pi^k}(s_h^k)\right\|_1\right)\\ &\leq 2L_\infty(\rho_h, H-h)\sqrt{\frac{2S}{N_h^k \vee 1}\iota}.\end{aligned}$$

Using arguments similar to the proof for Algorithm 1

$$(b) \leq L_1(\rho_h, H-h)(\epsilon_h^k + \delta_{h+1}^k),$$

Now we can bound δ_h^k recursively

$$\begin{aligned}\delta_h^k &\leq 2L_\infty(\rho_h, H-h)\sqrt{\frac{2S}{N_h^k \vee 1}\iota} + L_1(\rho_h, H-h)(\epsilon_h^k + \delta_{h+1}^k)\\ &= 2L_\infty(\rho_h, H-h)e_h^k + L_1(\rho_h, H-h)(\epsilon_h^k + \delta_{h+1}^k).\end{aligned}$$

Repeating the procedure, we obtain

$$\delta_1^k \leq 2\sum_{h=1}^{H-1} L_\infty(\rho_h, H-h)\prod_{i=1}^{h-1} L_1(\rho_i, H-i)e_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h} L_1(\rho_i, H-i)\epsilon_h^k.$$

Step 3: putting together. The results follows from analogous arguments of the proof for Algorithm 1.

C.2. Gap-dependent Regret Upper Bound

Step 1: regret decomposition. The regret of each episode can be rewritten as the expected sum of sub-optimality gaps for each action:

$$\begin{aligned}(V_1^* - V_1^{\pi^k})(s_1^k) &= V_1^*(s_1^k) - Q_1^*(s_1^k, a_1^k) + (Q_1^* - Q_1^{\pi^k})(s_1^k, a_1^k)\\ &= \Delta_1(s_1^k, a_1^k) + [P_2(V_2^* - V_2^{\pi^k})](s_2^k, a_2^k)\\ &= \cdots = \mathbb{E}\left[\sum_{h=1}^H \Delta_h(s_h^k, a_h^k)\right].\end{aligned}$$

Step 2: optimism.

Lemma C.8. *With probability at least $1-\delta$, the following event holds*

$$0 \le (Q_h^k - Q_h^*)(s,a) \le 2b_h^k(s,a) + L_{1,h}[P_h(V_{h+1}^k - V_{h+1}^*)](s,a).$$

Proof.

$$\begin{aligned}
&(Q_h^k - Q_h^*)(s,a) = r_h(s,a) + \rho_h\left(V_{h+1}^k, \hat{P}_h^k(s,a)\right) + b_h^k(s,a) - r_h(s,a) - \rho_h\left(V_{h+1}^*, P_h(s,a)\right) \\
&= \underbrace{\rho_h\left(V_{h+1}^k, \hat{P}_h^k(s,a)\right) - \rho_h\left(V_{h+1}^k, P_h(s,a)\right)}_{(a)} + \underbrace{\rho_h\left(V_{h+1}^k, P_h(s,a)\right) - \rho_h\left(V_{h+1}^*, P_h(s,a)\right)}_{(b)} + b_h^k(s,a) \\
&\le L_{\infty,h}\left\|(V_{h+1}^k, \hat{P}_h^k(s,a)) - (V_{h+1}^k, P_h(s,a))\right\|_\infty + L_{1,h}\left\|(V_{h+1}^k, P_h(s,a)) - (V_{h+1}^*, P_h(s,a))\right\|_1 + b_h^k(s,a) \\
&\le L_{\infty,h}\left\|\hat{P}_h^k(s,a) - V_{h+1}^k, P_h(s,a)\right\|_1 + L_{1,h}[P_h(V_{h+1}^k - V_{h+1}^*)](s,a) + b_h^k(s,a) \\
&\le 2b_h^k(s,a) + L_{1,h}[P_h(V_{h+1}^k - V_{h+1}^*)](s,a)
\end{aligned}$$

□

Step 3: bound number of steps in each interval

Lemma C.9. *For any $n \in [N]$,*

$$C^n \triangleq \left|\left\{(k,h) : (Q_h^k - Q_h^*)(s_h^k, a_h^k) \in [2^{n-1}\Delta_{min}, 2^n\Delta_{min})\right\}\right| \le \mathcal{O}\left(\frac{HS^2A\iota\left(\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'}\right)^2}{4^n\Delta_{min}^2}\right).$$

Proof. For every $n \in [N]$, $h \in [H]$, define

$$\begin{aligned}
w_k^{(n,h)} &\triangleq \mathbb{I}\left\{(Q_h^k - Q_h^*)(s_h^k, a_h^k) \in [2^{n-1}\Delta_{\min}, 2^n\Delta_{\min})\right\} \\
C^{(n,h)} &\triangleq \sum_{k=1}^K w_k^{(n,h)}.
\end{aligned}$$

Observe that $w_k^{(n,h)} \le 1$ and $(w_k^{(n,h)})^2 = w_k^{(n,h)}$. Now we bound $\sum_{k=1}^K w_k^{(n,h)}(Q_h^k - Q_h^*)(s_h^k, a_h^k)$ from both sides. On the one hand, by Lemma ,

$$\begin{aligned}
\sum_{k=1}^K w_k^{(n,h)}(Q_h^k - Q_h^*)(s_h^k, a_h^k) &\le 4\sqrt{S^2A\iota C^{(n,h)}} \cdot \sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'} + \sqrt{2C^{(n,h)}\log\frac{1}{\delta'}} \cdot \sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{1,h'}(H-h') \\
&= \mathcal{O}\left(\sqrt{S^2A\iota C^{(n,h)}} \cdot \sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'}\right).
\end{aligned}$$

On the other hand, by the definition of $w_k^{(n,h)}$,

$$\sum_{k=1}^K w_k^{(n,h)}(Q_h^k - Q_h^*)(s_h^k, a_h^k) \ge \sum_{k=1}^K w_k^{(n,h)} 2^{n-1}\Delta_{\min} = 2^{n-1}\Delta_{\min} \cdot C^{(n,h)}.$$

Combining the two inequalities, we obtain

$$C^{(n,h)} \le \mathcal{O}\left(\frac{S^2A\iota\left(\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'}\right)^2}{4^n\Delta_{\min}^2}\right)$$

Finally, we have

$$C^{(n)} = \sum_{h=1}^{H} C^{(n,h)} \leq \mathcal{O}\left(\frac{S^2 A\iota \sum_{h=1}^{H}\left(\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'}\right)^2}{4^n\Delta_{\min}^2}\right) \leq \mathcal{O}\left(\frac{S^2 A\iota H\left(\sum_{h'=1}^{H-1}\tilde{L}_{1,h'-1}L_{\infty,h'}\right)^2}{4^n\Delta_{\min}^2}\right)$$

□

Lemma C.10. *For any positive sequence $\{w_k\}_{k\in[K]}$, it holds that for any $h \in [H]$*

$$\sum_{k=1}^{K} w_k(Q_h^k - Q_h^*)(s_h^k, a_h^k) \leq 4\sqrt{wS^2A\iota\sum_{k=1}^{K} w_k}\cdot\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'} + \sqrt{2\sum_{k=1}^{K} w_k^2\log\frac{1}{\delta'}}\cdot\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{1,h'}(H-h').$$

Proof. By Lemma 5,

$$\begin{aligned}\sum_{k=1}^{K} w_k(Q_h^k - Q_h^*)(s_h^k, a_h^k) &\leq \sum_{k=1}^{K} w_k\left(2L_{\infty,h}\sqrt{\frac{2S\iota}{N_h^k}} + L_{1,h}[P_h(V_{h+1}^k - V_{h+1}^*)](s_h^k, a_h^k)\right)\\ &= \underbrace{2L_{\infty,h}\sum_{k=1}^{K} w_k\sqrt{\frac{2S\iota}{N_h^k\vee 1}}}_{(a)} + \underbrace{L_{1,h}\sum_{k=1}^{K} w_k\epsilon_h^k}_{(b)} + L_{1,h}\sum_{k=1}^{K} w_k(V_{h+1}^k - V_{h+1}^*)(s_{h+1}^k)\\ &\leq (a) + (b) + L_{1,h}\sum_{k=1}^{K} w_k(Q_{h+1}^k - Q_{h+1}^*)(s_{h+1}^k, a_{h+1}^k),\end{aligned}$$

where $\epsilon_h^k \triangleq [P_h(V_{h+1}^k - V_{h+1}^*)](s_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^*)](s_{h+1}^k) \in [-2(H-h), 2(H-h)]$ is a martingale difference sequence w.r.t. $\mathcal{F}_h^k$ for any $h \in [H]$, i.e., $\mathbb{E}\left[\epsilon_h^k|\mathcal{F}_h^k\right] = 0$. Define $k(s,a,t) \triangleq \min\{k : N_h^k(s,a) \geq t\}$ the episode when (s,a) is visited t times at step h. We can bound term (a) as

$$\begin{aligned}(a) = 2L_{\infty,h}\sum_{k=1}^{K} w_k\sqrt{\frac{2S\iota}{N_h^k}} &= 2L_{\infty,h}\sqrt{2S\iota}\sum_{s,a}\sum_{k=1}^{K}\mathbb{I}\{(s_h^k, a_h^k) = (s,a)\}\frac{w_k}{\sqrt{N_h^k(s,a)\vee 1}}\\ &= 2L_{\infty,h}\sqrt{2S\iota}\sum_{s,a}\sum_{t=1}^{N_h^K(s,a)}\frac{w_{k(s,a,t)}}{\sqrt{t}}\\ &\leq 2L_{\infty,h}\sqrt{2S\iota}\sum_{s,a}\sum_{t=1}^{C_{s,a}/w}\frac{w}{\sqrt{t}}\\ &\leq 4L_{\infty,h}\sqrt{S\iota}\sum_{s,a}\sqrt{C_{s,a}w}\\ &\leq 4L_{\infty,h}\sqrt{wS^2A\iota\sum_{k=1}^{K} w_k},\end{aligned}$$

where $C_{s,a} \triangleq \sum_{t=1}^{N_h^K(s,a)} w_{k(s,a,t)}$ and $w_k \leq w$ for any k, and the last inequality follows from that $\sum_{s,a} C_{s,a} = \sum_{s,a}\sum_{t=1}^{N_h^K(s,a)} w_{k(s,a,t)} = \sum_{k=1}^{K} w_k$.

Since $\{\epsilon_h^k\}_{k\in[K]}$ is a MDS with $|\epsilon_h^k| \leq 2(H-h)$, we can bound term (b) by Azuma-Hoeffding inequality: w.p. $1-\delta'$

$$(b) = L_{1,h}\sum_{k=1}^{K} w_k\epsilon_h^k \leq L_{1,h}(H-h)\sqrt{2\sum_{k=1}^{K} w_k^2\log\frac{1}{\delta'}}.$$

Thus we can get a recursive bound

$$\sum_{k=1}^{K} w_k(Q_h^k - Q_h^*)(s_h^k, a_h^k) \leq 4L_{\infty,h}\sqrt{wS^2A\iota\sum_{k=1}^{K} w_k} + L_{1,h}(H-h)\sqrt{2\sum_{k=1}^{K} w_k^2 \log\frac{1}{\delta'}} + L_{1,h}\sum_{k=1}^{K} w_k(Q_{h+1}^k - Q_{h+1}^*)(s_{h+1}^k, a_{h+1}^k).$$

Unrolling the inequality yields

$$\sum_{k=1}^{K} w_k(Q_h^k - Q_h^*)(s_h^k, a_h^k) \leq \sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}\left(4L_{\infty,h'}\sqrt{wS^2A\iota\sum_{k=1}^{K} w_k} + L_{1,h'}(H-h')\sqrt{2\sum_{k=1}^{K} w_k^2\log\frac{1}{\delta'}}\right)$$
$$= 4\sqrt{wS^2A\iota\sum_{k=1}^{K} w_k}\cdot\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'} + \sqrt{2\sum_{k=1}^{K} w_k^2\log\frac{1}{\delta'}}\cdot\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{1,h'}(H-h')$$

□

Step 4: Bound the regret Denote by $\tau \triangleq (s_h^k, a_h^k)_{k,h}$ the trajectory. Define $\text{clip}[x|\delta] \triangleq x\mathbb{I}\{x \geq \delta\}$. Observe that

$$V_h^*(s_h^k) = \max_a Q_h^*(s_h^k, a) \leq \max_a Q_h^k(s_h^k, a) = Q_h^k(s_h^k, a_h^k),$$

thus we get

$$\Delta_h(s_h^k, a_h^k) = \text{clip}[V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)|\Delta_{\text{min}}] \leq \text{clip}[(Q_h^k - Q_h^*)(s_h^k, a_h^k)|\Delta_{\text{min}}].$$

$$\begin{aligned}
\text{Regret}(K) &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\Delta_h(s_h^k, a_h^k)\right] = \sum \mathbb{P}(\tau)\sum_{k=1}^{K}\sum_{h=1}^{H}\Delta_h(s_h^k, a_h^k|\tau) \\
&\leq \sum_{\tau\in E}\mathbb{P}(\tau)\sum_{k=1}^{K}\sum_{h=1}^{H}\text{clip}[(Q_h^k - Q_h^*)(s_h^k, a_h^k|\tau)|\Delta_{\text{min}}] + \sum_{\tau\in E^c}\mathbb{P}(\tau)KH^2 \\
&\leq \mathbb{P}(E)\sum_{n=1}^{N} 2^n\Delta_{\text{min}}C^{(n)} + \mathbb{P}(E^c)KH^2 \\
&\leq \sum_{n=1}^{N}\mathcal{O}\left(\frac{HS^2A\iota\left(\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'}\right)^2}{2^n\Delta_{\text{min}}}\right) + H \\
&= \mathcal{O}\left(\frac{HS^2A\iota\left(\sum_{h'=h}^{H-1}\prod_{i=h}^{h'-1} L_{1,i}L_{\infty,h'}\right)^2}{\Delta_{\text{min}}}\right).
\end{aligned}$$

D. Proofs of Regret Lower Bounds

D.1. Minimax Lower Bound

We make the following assumption (Domingues et al., 2021).

Assumption D.1. Assume $S \geq 6$, $A \geq 2$, and there exists an integer d such that $S = 3 + \frac{A^d - 1}{A-1}$. We further assume that $H \geq 3d$ and $\bar{H} \triangleq \frac{H}{3} \geq 1$.

Theorem D.2. *Assume Assumption D.1 holds. For any algorithm $\mathscr{A}$, there exists an MDP $\mathcal{M}_{\mathscr{A}}$ such that for sufficiently large K*

$$\mathbb{E}[\textit{Regret}(\mathscr{A}, \mathcal{M}_{\mathscr{A}}, K)] \geq \frac{\sqrt{p}}{27\sqrt{6}}c_{\rho,1}H\sqrt{SAT}.$$

Step 1. Fix an arbitrary algorithm $\mathscr{A}$. We introduce three types of special states for the hard MDP class: a waiting state s_w where the agent starts and may stay until stage $\bar{H}$, after that it has to leave; a good state s_g, which is absorbing and is the only rewarding state; a bad state s_b that is absorbing and provides no reward. The rest of $S-3$ states are part of a A-ary tree of depth $d-1$. The agent can only arrive s_w from the root node s_{root} and can only reach s_g and s_b from the leaves of the tree. Let $\bar{H} \in [H-d]$ be the first parameter of the MDP class. We define $\tilde{H} := \bar{H} + d + 1$ and $H' := H + 1 - \tilde{H}$. We denote by $\mathcal{L} := \{s_1, s_2, ..., s_{\bar{L}}\}$ the set of $\bar{L}$ leaves of the tree. For each $u^* := (h^*, \ell^*, a^*) \in [d+1 : \bar{H} + d] \times \mathcal{L} \times \mathcal{A}$, we define an MDP $\mathcal{M}_{u^*}$ as follows. The transitions in the tree are deterministic, hence taking action a in state s results in the a-th child of node s. The transitions from s_w are defined as

$$P_h\left(s_{\mathrm{w}} \mid s_{\mathrm{w}}, a\right) := \mathbb{I}\left\{a = a_{\mathrm{w}}, h \leq \bar{H}\right\} \quad \text{and} \quad P_h\left(s_{\text{root}} \mid s_{\mathrm{w}}, a\right) := 1 - P_h\left(s_{\mathrm{w}} \mid s_{\mathrm{w}}, a\right).$$

The transitions from any leaf $s_i \in \mathcal{L}$ are specified as

$$P_h\left(s_g \mid s_i, a\right) := p + \Delta_{u^*}\left(h, s_i, a\right) \quad \text{and} \quad P_h\left(s_b \mid s_i, a\right) := 1 - p - \Delta_{u^*}\left(h, s_i, a\right),$$

where $\Delta_{u^*}\left(h, s_i, a\right) := \epsilon\mathbb{I}\{(h, s_i, a) = (h^*, s_{\ell^*}, a^*)\}$ for some constants $p \in [0,1]$ and $\epsilon \in [0, \min(1-p, p)]$ to be determined later. p and ϵ are the second and third parameters of the MDP class. Observe that s_g and s_b are absorbing, therefore we have $\forall a, P_h\left(s_g \mid s_g, a\right) := P_h\left(s_b \mid s_b, a\right) := 1$. The reward is a deterministic function of the state

$$r_h(s,a) := \mathbb{I}\{s = s_g, h \geq \tilde{H}\}.$$

Finally, we define a reference MDP $\mathcal{M}_0$ which differs from the previous MDP instances only in that $\Delta_0(h, s_i, a) := 0$ for all (h, s_i, a). For each ϵ, p and $\bar{H}$, we define the MDP class

$$\mathcal{C}_{\bar{H},p,\epsilon} := \mathcal{M}_0 \cup \{\mathcal{M}_{u^*}\}_{u^* \in [d+1:\bar{H}+d] \times \mathcal{L} \times \mathcal{A}}.$$

For an MDP $\mathcal{M}_{u^*}$, the optimal policy $\pi^{*,\mathcal{M}_{u^*}}$ starts to traverse the tree at step h^*-d then chooses to reach the leaf s_{l^*} and performs action a^*. The optimal value in any of these MDPs is the same

$$\begin{aligned} V_1^{*,\mathcal{M}_{u^*}} &= V_{h^*}^{*,\mathcal{M}_{u^*}}(s_{l^*}) = Q_{h^*}^{*,\mathcal{M}_{u^*}}(s_{l^*}, a^*) = r_h(s_{l^*}, a^*) + \rho_{h^*}(V_{h^*+1}^{*,\mathcal{M}_{u^*}}, P_h(s_{l^*}, a^*)) \\ &= \rho_{h^*}((V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_g), V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_b)), (p+\epsilon, 1-p-\epsilon)). \end{aligned}$$

For simplicity, we may drop $\mathcal{M}_{u^*}$ from the notations. Notice that the agent must be in either of the absorbing states at step $h \geq \tilde{H} = \bar{H} + d + 1$. Observe that $V_H^{*,\mathcal{M}_{u^*}}(s_g) = r_H(s_g, a) = 1$ since $r_h(s_g, a) = 1$ for any a and any $h \geq \tilde{H}$, and $V_H^{*,\mathcal{M}_{u^*}}(s_b) = 0$. Thus we have:

$$Q_{H-1}^{*,\mathcal{M}_{u^*}}(s_g, a) = r_{H-1}(s_g, a) + \rho_{H-1}((V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_g), V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_b)), (1,0)) = 1 + V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_g) = 2, \forall a,$$

where the second equality follows from that $\rho_h(c) = c$ for a deterministic constant c. Therefore $V_{H-1}^{*,\mathcal{M}_{u^*}}(s_g) = 2$. Similarly we can get $V_{H-1}^{*,\mathcal{M}_{u^*}}(s_b) = 0$. It follows from inductions that $V_h^{*,\mathcal{M}_{u^*}}(s_g) = H + 1 - h$ and $V_h^{*,\mathcal{M}_{u^*}}(s_b) = 0$ for $h \geq \tilde{H}$. Moreover, observe that

$$V_{\tilde{H}-1}^{*,\mathcal{M}_{u^*}}(s_g) = 0 + \rho_{\tilde{H}-1}((V_{\tilde{H}}^{*,\mathcal{M}_{u^*}}(s_g), V_{\tilde{H}}^{*,\mathcal{M}_{u^*}}(s_b)), (1,0)) = V_{\tilde{H}}^{*,\mathcal{M}_{u^*}}(s_g) = H + 1 - \tilde{H} = H'$$

and $V_{\tilde{H}-1}^{*,\mathcal{M}_{u^*}}(s_b) = 0$. Then $V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_g) = \cdots = V_{\tilde{H}-1}^{*,\mathcal{M}_{u^*}}(s_g) = H'$ and $V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_b) = \cdots = V_{\tilde{H}-1}^{*,\mathcal{M}_{u^*}}(s_b) = 0$. Thus the optimal value

$$\begin{aligned} V_1^{*,\mathcal{M}_{u^*}} &= \rho_{h^*}((V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_g), V_{h^*+1}^{*,\mathcal{M}_{u^*}}(s_b)), (p+\epsilon, 1-p-\epsilon)) \\ &= \rho_{h^*}((V_{\tilde{H}}^{*,\mathcal{M}_{u^*}}(s_g), V_{\tilde{H}}^{*,\mathcal{M}_{u^*}}(s_b)), (p+\epsilon, 1-p-\epsilon)) \\ &= \rho_{h^*}((H', 0), (p+\epsilon, 1-p-\epsilon)) \end{aligned}$$

Consider the case that policy $\pi^k \neq \pi^*$. Then we have $(s_{h^*}^k, a_{h^*}^k) \neq (s_{l^*}, a^*)$. Analogously, we can get

$$V_h^{\pi^k}(s_g) = H + 1 - h, \; V_h^{*,\mathcal{M}_{u^*}}(s_b) = 0$$

for $h \geq \tilde{H}$. Suppose π^k arrives at the leaf node $s^k_{l^k}$ in step $l^k \in [1+d, \tilde{H}-1]$, then $V^{\pi^k}_{l^k+1}(s_g) = \cdots = V^{\pi^k}_{\tilde{H}}(s_g) = H+1-\tilde{H} = H'$ and $V^{\pi^k}_{l^k+1}(s_b) = \cdots = V^{\pi^k}_{\tilde{H}}(s_b) = 0$. Since $P_{l^k}(s_g|s^k_{l^k}, a^k_h) = p$,

$$V^{\pi^k}_1 = \rho_{l^k}((V^{\pi^k}_{l^k+1}(s_g), V^{\pi^k}_{l^k+1}(s_b)), (p, 1-p)) = \rho_{l^k}((H', 0), (p, 1-p))$$

Denote by $x^k_h := (s^k_h, a^k_h)$ for each (k,h), $x^* := (s_{\ell^*}, a^*)$ and $N_K(u^*) := \sum_{k=1}^K \mathbb{I}\{x^k_{h^*} = x^*\}$. It follows that

$$V^{\pi^k}_1 = \mathbb{I}\{x^k_{h^*} = x^*\} V^*_1 + \mathbb{I}\{x^k_{h^*} \neq x^*\} \rho_{l^k}((H', 0), (p, 1-p))$$

Define c_ρ as the constant that satisfies

$$\begin{aligned}\rho((H,0),(p,1-p)) - \rho((H,0),(q,1-q)) &\geq c_\rho \left\|((H,0),(p,1-p)) - ((H,0),(q,1-q))\right\|_1 \\ &= c_\rho H|p-q|.\end{aligned}$$

The expected regret of $\mathscr{A}$ in $\mathcal{M}_{u^*}$ can be bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[\text{Regret}(\mathscr{A},\mathcal{M}_{u^*},K)] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K V^*_1 - V^{\pi^k}_1\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \mathbb{I}\{x^k_{h^*} \neq x^*\}\left(\rho_{h^*}((H',0),(p+\epsilon,1-p-\epsilon)) - \rho_{l^k}((H',0),(p,1-p))\right)\right] \\
&\geq \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \mathbb{I}\{x^k_{h^*} \neq x^*\} c_\rho \left\|((H',0),(p+\epsilon,1-p-\epsilon)) - ((H',0),(p,1-p))\right\|_1\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \mathbb{I}\{x^k_{h^*} \neq x^*\} c_\rho H' \epsilon\right] \\
&= c_\rho \epsilon H'(K - \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[N_K(u^*)]),
\end{aligned}$$

Step 2. The maximum of the regret can be bounded below by the mean over all instances as

$$\begin{aligned}
\max_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \text{Regret}(\mathscr{A},\mathcal{M}_{u^*},K) &\geq \frac{1}{\bar{H}\bar{L}A} \sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \text{Regret}(\mathscr{A},\mathcal{M}_{u^*},K) \\
&\geq c_{\rho,1} H' K \epsilon \left(1 - \frac{1}{\bar{L}AK\bar{H}} \sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \mathbb{E}_{u^*}[N_K(u^*)]\right).
\end{aligned}$$

Observe that it can be further bounded if we can obtain an upper bound on $\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \mathbb{E}_{u^*}[N_K(u^*)]$, which can be done by relating each expectation to the expectation under the reference MDP $\mathcal{M}_0$.

Fact 7 (Lemma 1,(Garivier et al., 2019)). Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0,1]$, we have

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) \geq \mathrm{kl}(\mathbb{E}_1[Z], \mathbb{E}_2[Z]),$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively.

Fact 8 (Lemma 5, (Domingues et al., 2021)). Let $\mathcal{M}$ and $\mathcal{M}'$ be two MDPs that are identical except for their transition probabilities, denoted by P_h and P'_h, respectively. Assume that we have $\forall (s,a)$, $P_h(\cdot \mid s,a) \ll P'_h(\cdot \mid s,a)$. Then, for any stopping time τ with respect to $(I_k)_{k\geq 1}$ that satisfies $\mathbb{P}_{\mathcal{M}}[\tau < \infty] = 1$

$$\mathrm{KL}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\mathcal{M}'}) = \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H-1]} \mathbb{E}_{\mathcal{M}}[N^\tau_h(s,a)] \,\mathrm{KL}(P_h(\cdot \mid s,a), P'_h(\cdot \mid s,a)).$$

Fact 9 (Lemma 28, (Liang & Luo, 2022)). If $\epsilon \geq 0$, $p \geq 0$ and $p + \epsilon \in [0, \frac{1}{2}]$, then $\mathrm{kl}(p, p+\epsilon) \leq \frac{\epsilon^2}{2p(1-p)} \leq \frac{\epsilon^2}{p}$.

By applying Fact 7 with $Z = \frac{N_K(u^*)}{K} \in [0,1]$, we have

$$\mathrm{kl}\left(\frac{1}{K}\mathbb{E}_0\left[N_K(u^*)\right], \frac{1}{K}\mathbb{E}_{u^*}[N_K(u^*)]\right) \leq \mathrm{KL}\left(\mathbb{P}_0, \mathbb{P}_{u^*}\right).$$

By Pinsker's inequality, it implies that

$$\frac{1}{K}\mathbb{E}_{u^*}[N_K(u^*)] \leq \frac{1}{K}\mathbb{E}_0\left[N_K(u^*)\right] + \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathbb{P}_0, \mathbb{P}_{u^*}\right)}.$$

Since $\mathcal{M}_0$ and $\mathcal{M}_{u^*}$ only differs at stage h^* when $(s,a) = x^*$, it follows from Fact 8 that

$$\mathrm{KL}\left(\mathbb{P}_0, \mathbb{P}_{u^*}\right) = \mathbb{E}_0\left[N_K(u^*)\right]\mathrm{kl}(p, p+\varepsilon).$$

By Fact 9, we have $\mathrm{kl}(p, p+\epsilon) \leq \frac{\epsilon^2}{p}$ for $\epsilon \geq 0$ and $p + \epsilon \in [0, \frac{1}{2}]$. Consequently,

$$\begin{aligned}
&\frac{1}{K}\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \mathbb{E}_{u^*}[N_K(u^*)] \\
&\leq \frac{1}{K}\mathbb{E}_0\left[\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} N_K(u^*)\right] + \frac{\epsilon}{\sqrt{2p}}\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}}\sqrt{\mathbb{E}_0\left[N_K(u^*)\right]} \\
&\leq 1 + \frac{\epsilon}{\sqrt{2p}}\sqrt{\bar{L}AK\bar{H}},
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwartz inequality and that $\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} N_K(u^*) = K$. It follows that

$$\max_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \mathrm{Regret}(\mathscr{A}, \mathcal{M}_{u^*}, K) \geq c_{\rho,1}H'K\epsilon\left(1 - \frac{1}{\bar{L}A\bar{H}} - \frac{\frac{\epsilon}{\sqrt{2p}}\sqrt{\bar{L}AK\bar{H}}}{\bar{L}A\bar{H}}\right).$$

Step 3. Choosing $\epsilon = \sqrt{\frac{p}{2}}(1 - \frac{1}{LA\bar{H}})\sqrt{\frac{LA\bar{H}}{K}}$ maximizes the lower bound

$$\max_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \mathrm{Regret}(\mathscr{A}, \mathcal{M}_{u^*}, K) \geq \frac{\sqrt{p}}{2\sqrt{2}}c_{\rho,1}H'\left(1 - \frac{1}{\bar{L}A\bar{H}}\right)^2\sqrt{\bar{L}AK\bar{H}}.$$

Since $S \geq 6$ and $A \geq 2$, we have $\bar{L} = (1 - \frac{1}{A})(S-3) + \frac{1}{A} \geq \frac{S}{4}$ and $1 - \frac{1}{\bar{L}A\bar{H}} \geq 1 - \frac{1}{\frac{6}{4}\cdot 2} = \frac{2}{3}$. Choose $\bar{H} = \frac{H}{3}$ and use the assumption that $d \leq \frac{H}{3}$ to obtain that $H' = H - d - \bar{H} \geq \frac{H}{3}$. Now we choose arbitrary $p \leq \frac{1}{4}$ and $\epsilon = \sqrt{\frac{p}{2}}(1 - \frac{1}{\bar{L}A\bar{H}})\sqrt{\frac{LA\bar{H}}{K}} < \frac{1}{2\sqrt{2}}\sqrt{\frac{\bar{L}A\bar{H}}{K}} \leq \frac{1}{4}$ if $K \geq 2\bar{L}A\bar{H}$. Such choice of p and ϵ guarantees the assumption of Fact 9. Finally we use the fact that $\sqrt{\bar{L}AK\bar{H}} \geq \frac{1}{2\sqrt{3}}\sqrt{SAKH}$ to obtain

$$\max_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \mathrm{Regret}(\mathscr{A}, \mathcal{M}_{u^*}, K) \geq \frac{\sqrt{p}}{27\sqrt{6}}c_\rho H\sqrt{SAKH}.$$

D.2. Gap-dependent Lower Bound

Theorem D.3 (Gap-dependent regret lower bound)**.** *Let $S \geq 2$ and $A \geq 2$, and let $\{\delta_{s,a}\}_{s,a\in\mathcal{S}\times\mathcal{A}} \subset (0, \frac{H}{8})$ denote a set of gaps. For any $h \geq 1$, there exists an MDP $\mathcal{M}$ with $\mathcal{S} = [S+2]$ and $\mathcal{A} = [A]$ such that any α-uniformly good algorithm* `alg` *satisfies*

$$\lim_{K\to\infty}\frac{\mathit{Regret}(\texttt{alg}, \mathcal{M}, K)}{\log K} = \Omega\left((1-\alpha)\sum_{(s,a):\Delta_1(s,a)>0}\frac{(c_\rho H)^2}{\Delta_1(s,a)}\right)$$

We first fix an arbitrary α-uniformly good algorithm $\mathscr{A}$. For simplicity, we may drop $\mathscr{A}$ from the notations, e.g., $\mathbb{E}_{\mathcal{M}} = \mathbb{E}_{\mathcal{M},\mathscr{A}}$.

Step 1: construction of the hard instance. Our construction mirrors the lower bounds in . However, their instance is suited for homogeneous/stationary MDP. Define an MDP $\mathcal{M}$ with $\mathcal{S} = \{0\} \cup [S+2]$ and $\mathcal{A} = [A]$. Without loss of generality, we consider the case $H \geq 2$. Otherwise, it reduces to a bandit setting. We first specify the transition kernels. For the convenience of analysis, we introduce $s_0 = 0$ at stage $h = 0$ with $P_0(s|0) = \frac{1}{S}$ for any $s \in [S]$. In other words, the initial state s_1 is uniformly distributed over $[S]$. For $(s,a) \in [S] \times [A]$, let

$$P_1(S+1|s,a) = \frac{3}{4} - \frac{2\delta_{s,a}}{H-1} =: \frac{3}{4} - \tilde{\delta}_{s,a},\ P_1(S+1|s,a) = 1 - P_1(S+1|s,a).$$

Thus at stage 1, each state $s \in [S]$ can only transit to either state $S+1$ or $S+2$. Furthermore, we set state $S+1$ and $S+2$ to be absorbing state, i.e.

$$P_h(S+1|S+1,a) = P_h(S+2|S+2,a) = 1,\ \forall h \in [2:H-1], a \in [A].$$

Finally, we set the reward functions as

$$R(x,a) := \begin{cases} 1 & (x,a) = (S+1,1) \\ \frac{1}{2} & (x,a) = (S+2,1) \\ 0 & \text{otherwise.} \end{cases}$$

We assume that there exists a unique action $\pi^*(s)$ for each $s \in [S]$ such that $\delta_{s,\pi^*(s)} = 0$. We will see that such action is the optimal action. Note that $S{+}1$ and $S{+}2$ are absorbing states and the only two rewarding states, hence $V_h^*(S{+}1) = H{+}1{-}h$ and $V_h^*(S+2) = \frac{H+1-h}{2}$ for $h \in [2:H]$. It follows that for $x \in [S]$,

$$\begin{aligned} V_h^*(s) &= 0 + \rho_h(V_{h+1}^*, P_h(s, \pi_h^*(s))) = \rho_h\left(\left(H-h, \frac{H-h}{2}\right), \left(\frac{3}{4}, \frac{1}{4}\right)\right), \\ Q_h^*(s,a) &= 0 + \rho_h(V_{h+1}^*, P_h(s,a)) = \rho_h\left(\left(H-h, \frac{H-h}{2}\right), \left(\frac{3}{4} - \tilde{\delta}_{s,a}, \frac{1}{4} + \tilde{\delta}_{s,a}\right)\right), \end{aligned}$$

which implies that

$$\begin{aligned} \Delta_h(s,a) &= \rho_h\left(\left(H-h, \frac{H-h}{2}\right), \left(\frac{3}{4}, \frac{1}{4}\right)\right) - \rho_h\left(\left(H-h, \frac{H-h}{2}\right), \left(\frac{3}{4} - \tilde{\delta}_{s,a}, \frac{1}{4} + \tilde{\delta}_{s,a}\right)\right) \\ &\geq c_\rho \left\| \left(\left(H-h, \frac{H-h}{2}\right), \left(\frac{3}{4}, \frac{1}{4}\right)\right) - \left(\left(H-h, \frac{H-h}{2}\right), \left(\frac{3}{4} - \tilde{\delta}_{s,a}, \frac{1}{4} + \tilde{\delta}_{s,a}\right)\right) \right\|_1 \\ &= c_\rho \frac{H-h}{2} \tilde{\delta}_{s,a}. \end{aligned}$$

In particular, we have $\Delta_1(s,a) \geq c_\rho \frac{H-1}{2} \tilde{\delta}_{s,a} = c_{\rho,1}\delta_{s,a}$. Note that $\Delta_1(s,a)$ is only defined for $s \in [S]$.

Step 2: regret decomposition. The regret for algorithm $\mathscr{A}$ over MDP $\mathcal{M}$ can be decomposed as follows

$$\begin{aligned}
\text{Regret}(\mathscr{A}, \mathcal{M}, K) &= \mathbb{E}\left[\sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta_h(s_h^k, a_h^k)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} \mathbb{I}\left\{s_h^k = s, a_h^k = a\right\} \Delta_h(s_h^k, a_h^k)\right] \\
&= \sum_{h=1}^{H}\sum_{s,a} \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{I}\left\{s_h^k = s, a_h^k = a\right\}\right] \Delta_h(s,a) \\
&= \sum_{h=1}^{H}\sum_{s,a} \mathbb{E}\left[N_h^K(s,a)\right] \Delta_h(s,a) \\
&= \sum_{h=1}^{H}\sum_{s\in[S],a} \mathbb{E}\left[N_h^K(s,a)\right] \Delta_h(s,a).
\end{aligned}$$

where the last equality is due to that $\Delta_h(s,a)$ is only defined over $[S]$. For our hard instance, observe that $\mathbb{I}\left\{s_h^k = s\right\} = 0$ for $s \in [S]$ and $h \neq 1$. Therefore $N_h^K(s,a) = 0$ for $s \in [S]$ and $h \neq 1$, which implies

$$\text{Regret}(\mathscr{A}, \mathcal{M}, K) = \sum_{s\in[S],a} \mathbb{E}\left[N_1^K(s,a)\right] \Delta_1(s,a).$$

We claim that for any (s,a) such that $s \in [S]$ and $\Delta_1(s,a) > 0$, and any $K \geq K_0(\mathcal{M})$, it holds that

$$\mathbb{E}_{\mathscr{A},\mathcal{M}}\left(N_1^K(s,a)\right) \geq \Omega\left(\frac{1}{\tilde{\delta}_{s,a}^2}\log K\right) = \Omega\left(\frac{(c_{\rho,1}H)^2}{\Delta_1(s,a)^2}\log K\right).$$

It follows that

$$\text{Regret}(\mathscr{A}, \mathcal{M}, K) \geq \Omega\left(\sum_{s\in[S],a:\Delta_1(s,a)>0} \frac{(c_{\rho,1}H)^2}{\Delta_1(s,a)}\log K\right).$$

Step 3: bounding $\mathbb{E}\left[N_h^K(s,a)\right]$. Observe that $\tilde{\delta}_{s,a} = \frac{2}{H-1}\delta_{s,a} \in (0, \frac{1}{2})$ due to the assumption that $\delta_{s,a} \in (0, \frac{H}{8})$. By Fact 7 and Fact 8, let Z be a $\mathcal{F}^K$-measurable random variable, then it holds that

$$\text{kl}\left(\mathbb{E}_{\mathcal{M}}[Z], \mathbb{E}_{\mathcal{M}'}[Z]\right) \leq \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H-1]} \mathbb{E}_{\mathcal{M}}\left[N_h^K(s,a)\right] \text{KL}\left(P_h(s,a), P_h'(s,a)\right).$$

Now fix an arbitrary $(s,a) \in [S]\times[A]$. Define an MDP $\mathcal{M}_{s,a}'$ which differs from M only in that $P_1(S+1|s,a) = \frac{3}{4} + \eta$, where $\eta = \min\{\frac{1}{8}, \tilde{\delta}_{s,a}\}$. For simplicity, we write $\mathcal{M}' = \mathcal{M}_{s,a}'$. The following holds

$$\text{kl}\left(\mathbb{E}_{\mathcal{M}}[Z], \mathbb{E}_{\mathcal{M}'}[Z]\right) \leq \mathbb{E}_{\mathcal{M}}\left[N_1^K(s,a)\right] \text{KL}\left(P_1(s,a), P_1'(s,a)\right) = \mathbb{E}_{\mathcal{M}}\left[N_1^K(s,a)\right] \text{kl}\left(\frac{3}{4} - \tilde{\delta}_{s,a}, \frac{3}{4} + \eta\right).$$

Observe that $\frac{1}{4} < \frac{3}{4} - \tilde{\delta}_{s,a} < \frac{3}{4} + \eta < \frac{7}{8}$ and $\eta + \tilde{\delta}_{s,a} \leq 2\tilde{\delta}_{s,a}$, it follows from Fact 9 that

$$\text{kl}\left(\frac{3}{4} - \tilde{\delta}_{s,a}, \frac{3}{4} + \eta\right) \leq \frac{(\eta + \tilde{\delta}_{s,a})^2}{2(\frac{3}{4} - \tilde{\delta}_{s,a})(1 - \frac{3}{4} - \eta)} < 64\tilde{\delta}_{s,a}^2.$$

Now we have

$$\mathbb{E}_{\mathcal{M}}\left[N_1^K(s,a)\right] \geq \frac{1}{64\tilde{\delta}_{s,a}^2}\text{kl}\left(\mathbb{E}_{\mathcal{M}}[Z], \mathbb{E}_{\mathcal{M}'}[Z]\right) \geq \frac{(c_{\rho,1}(H-1))^2}{256\Delta_1^2(s,a)}\text{kl}\left(\mathbb{E}_{\mathcal{M}}[Z], \mathbb{E}_{\mathcal{M}'}[Z]\right).$$

We set $Z = \sum_{k=1}^{K} \frac{\mathbb{I}\{\pi_1^k(s)=a\}}{K} \in [0,1]$. Note that Z is indeed $\mathcal{F}^K$-measurable random variable since π^k is $\mathcal{F}^K$-measurable and (s,a) is fixed. Denote by Δ' the gap for MDP $\mathcal{M}'$. Observe that for $a' \neq a$,

$$\begin{aligned}\Delta_1'(s,a') &= \rho_h\left(\left(H-1, \frac{H-1}{2}\right), \left(\frac{3}{4}+\eta, \frac{1}{4}-\eta\right)\right) - \rho_h\left(\left(H-1, \frac{H-1}{2}\right), \left(\frac{3}{4}-\tilde{\delta}_{s,a'}, \frac{1}{4}+\tilde{\delta}_{s,a'}\right)\right) \\ &\geq c_\rho \frac{H-1}{2}(\eta + \tilde{\delta}_{s,a'}) \geq c_\rho \frac{H-1}{2}\eta.\end{aligned}$$

Under MDP $\mathcal{M}'$, action a is the unique optimal action for s, thus

$$\begin{aligned}\text{Regret}(\mathscr{A}, \mathcal{M}', K) &\geq \sum_{a' \neq a} \mathbb{E}_{\mathcal{M}'}\left[N_1^K(s,a')\right] \Delta_1'(s,a') \\ &\geq c_\rho \frac{H-1}{2}\eta \sum_{a'\neq a} \mathbb{E}_{\mathcal{M}'}\left[N_1^K(s,a')\right] \\ &= c_\rho \frac{H-1}{2}\eta \mathbb{E}_{\mathcal{M}'}\left[\sum_{k=1}^{K}\sum_{a'\neq a} \mathbb{I}\left(s_1^k = s, \pi_1^k(s_1^k) = a\right)\right] \\ &= c_\rho \frac{H-1}{2}\eta \mathbb{E}_{\mathcal{M}'}\left[\sum_{k=1}^{K}\left(\mathbb{I}\left(s_1^k = s\right) - \mathbb{I}\left(s_1^k = s, \pi_1^k(s_1^k) = a\right)\right)\right] \\ &= c_\rho \frac{H-1}{2}\eta \left(\frac{K}{S} - \mathbb{E}_{\mathcal{M}'}\left[\sum_{k=1}^{K} \mathbb{I}\left(s_1^k = s\right)\mathbb{I}\left(\pi_1^k(s) = a\right)\right]\right) \\ &= c_\rho \frac{H-1}{2}\eta \left(\frac{K}{S} - \sum_{k=1}^{K} \mathbb{E}_{\mathcal{M}'}\left[\mathbb{I}\left(s_1^k = s\right)\right]\mathbb{E}_{\mathcal{M}'}\left[\mathbb{I}\left(\pi_1^k(s) = a\right)\right]\right) \\ &= c_\rho \frac{H-1}{2}\frac{K}{S}\left(1 - \mathbb{E}_{\mathcal{M}'}[Z]\right),\end{aligned}$$

where the second to the last equality is due to the dependence between s_1^k and π_1^k. Since $\mathscr{A}$ is α-uniformly good algorithm, there exists $C_{\mathcal{M}'} > 0$ such that

$$c_\rho \frac{H-1}{2}\frac{K}{S}\left(1 - \mathbb{E}_{\mathcal{M}'}[Z]\right) \leq \text{Regret}(\mathscr{A}, \mathcal{M}', K) \leq C_{\mathcal{M}'}K^\alpha,$$

implying

$$1 - \mathbb{E}_{\mathcal{M}'}[Z] \leq \frac{2C_{\mathcal{M}'}S}{c_{\rho,1}(H-1)K^{1-\alpha}}.$$

We can also get

$$C_{\mathcal{M}}K^\alpha \geq \text{Regret}(\mathscr{A}, \mathcal{M}, K) \geq \mathbb{E}_{\mathcal{M}}\left[N_1^K(s,a)\right]\Delta_1(s,a) \geq \frac{K\Delta_1(s,a)}{S}\mathbb{E}_{\mathcal{M}}[Z],$$

which implies that $\mathbb{E}_{\mathcal{M}}[Z] \leq \frac{C_{\mathcal{M}}S}{\Delta_1(s,a)K^{1-\alpha}}$. Observe that

$$\text{kl}(x,y) \geq (1-x)\log\frac{1}{1-y} - \log 2.$$

It follows that

$$\text{kl}\left(\mathbb{E}_{\mathcal{M}}[Z], \mathbb{E}_{\mathcal{M}'}[Z]\right) \geq \left(1 - \frac{C_{\mathcal{M}}S}{\Delta_1(s,a)K^{1-\alpha}}\right)\left((1-\alpha)\log K - \log\frac{2C_{\mathcal{M}'}S}{c_{\rho,1}(H-1)}\right) - \log 2.$$

Step 4. We can also prove for the case $h \neq 1$ by modifying the transition kernels for state 0. For $h \neq 1$, we set the transition kernels as

$$P_l(0|0,a) = 1, \forall l \in [0:h-2], \forall a \in [A],\ P_{h-1}(s|0,a) = \frac{1}{S}, \forall s \in [S], \forall a \in [A].$$

In other words, the MDP is randomly initialized over $[S]$ at stage h rather than stage 1. For $(s, a) \in [S] \times [A]$, let

$$P_h(S+1|s,a) = \frac{3}{4} - \frac{2\delta_{s,a}}{H-1} =: \frac{3}{4} - \tilde{\delta}_{s,a},\ P_h(S+1|s,a) = 1 - P_h(S+1|s,a).$$

Finally, we still set $S+1, S+2$ to be absorbing states. Using similar arguments concludes the proof.