Near-optimal Distributional Reinforcement Learning towards Risk-sensitive Control

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider finite episodic Markov decision processes aiming at the entropic risk measure (EntRM) of return for risk-sensitive control. We identify two properties of the EntRM that enable risk-sensitive distributional dynamic programming. We propose two novel distributional reinforcement learning (DRL) algorithms, including a model-free one and a model-based one, that implement optimism through two different schemes. We prove that both of them attain $\tilde{\mathcal{O}}(\frac{\exp(|\beta|H)-1}{|\beta|H}H\sqrt{HS^2AT})$ regret upper bound, where S is the number of states, A the number of states, H the time horizon and T the number of total time steps. It matches RSVI2 proposed in [22] with a much simpler regret analysis. To the best of our knowledge, this is the first regret analysis of DRL, which theoretically verifies the efficacy of DRL for risk-sensitive control. Finally, we improve the existing lower bound by proving a tighter bound of $\Omega(\frac{\exp(\beta H/6)-1}{\beta H}H\sqrt{SAT})$ for $\beta > 0$ case, which recovers the tight lower bound $\Omega(H\sqrt{SAT})$ in the risk-neutral setting.

1 Introduction

Standard reinforcement learning (RL) [45] seeks to find an optimal policy that maximizes the expectation of return. It is also called risk-neutral RL since the objective is the mean functional of the return distribution. However, in some high-stakes applications including finance [15, 6], medical treatment [21] and operations [16] etc, the decision-maker tends to be risk-sensitive with the goal of maximizing some risk measure of return distribution.

In this paper, we consider the problem of optimizing the exponential risk measure (EntRM) in the episodic and finite MDP setting for risk-sensitive control. The entropic risk measure can trade-off between the expectation and the variance, and adjusts the risk-sensitiveness by control a risk parameter (see Equation 1). Ever since the seminal work of [29], risk-sensitive RL based on the EntRM has been applied across a wide range of domains [43, 37, 27]. Most of the existing approaches, however, involve complicated algorithmic design to deal with the non-linearity of the EntRM.

Distributional reinforcement learning (DRL) [4] has demonstrated its superior performance over traditional methods in some difficult tasks [14, 13] under risk-neutral setting. Different from the value-based approaches, it learns the whole return distribution instead of a real-valued value function. Given the entire return distribution, it is natural to leverage the distributional information to optimize a risk measure other than expectation [13, 44, 33]. Despite of the intrinsic connection between DRL and risk-sensitive RL, it is surprising that existing works on risk-sensitive control via DRL approaches ([13, 34, 1]) lack regret analysis. Consequently, it is challenging to evaluate and improve these DRL algorithms in terms of sample-efficiency, which brings about a reasonable question

Can distributional reinforcement learning attain near-optimal regret for risk-sensitive control?

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In this work, we answer this question positively by providing two DRL algorithms with provably regret guarantees. We devise two novel DRL algorithms with principled exploration schemes for risk-sensitive control in the tabular MDP setting. In particular, the proposed algorithms implement the principle of optimism in the face of uncertainty (OFU) at the distributional level to balance the exploration-exploitation trade-off. By providing the first regret analysis of DRL, we theoretically verifies the efficacy of DRL for risk-sensitive control. Therefore, our work bridge the gap between DRL and risk-sensitive RL with regard to sample complexity.

Main contributions. We summarize our main contributions in the following.
1. We build a risk-sensitive distributional dynamic programming (RS-DDP) framework. To be more specific, we choose the entropic risk measure (EntRM) of the return distribution as our objective. By identifying two key properties of EntRM, We establish distributional Bellman optimality equation for risk-sensitive control.
2. We propose two DRL algorithms that enforce the OFU principle in a distributional fashion through two different schemes. We provide $\tilde{\mathcal{O}}(\frac{\exp(|\beta|H)-1}{|\beta|}H\sqrt{S^2AK})$ regret upper bound, which matches the best existing result of RSVI2 in [22]. It is the first regret analysis of DRL algorithm in the finite episodic MDP in the risk-sensitive setting. Compared to [22], our algorithm does not involve complicated bonus design, and our analysis are conceptually cleaner and easier to interpret.
3. We fill the gaps in the proof of lower bound in [23]. To the best of our knowledge, [23] only implies a lower bound $\Omega(\frac{\exp(|\beta|H/2)-1}{|\beta|}\sqrt{K})$ rather the claimed bound $\Omega(\frac{\exp(|\beta|H/2)-1}{|\beta|}\sqrt{T})$. The resulting lower bound is independent of S and A and is loose with a factor of $\sqrt{H}$. We overcome these issues by proving a tight lower bound of $\Omega(\frac{\exp(\beta H/6)-1}{\beta H}H\sqrt{SAT})$ for $\beta > 0$. Note that the lower bound is tight in the risk-neutral setting ($\beta \to 0$).

Related work. Following the paper [4], DRL has witnessed a rapid growth of study in literature [40, 14, 13, 2, 32]. Most of these works focus on improving the performance in the risk-neutral setting, with a few exceptions [13, 34, 1]. However, none of these works study the sample complexity.

A rich body of work studies risk-sensitive RL with the EntRM [7, 8, 10, 9, 3, 11, 12, 18, 17, 19, 24, 28, 30, 33, 35, 36, 38, 39, 42, 43]. In particular, [29] is the first to introduce the ERM as risk-sensitive objective in MDP. However, they either assume known transition and reward or consider infinite-horizon setting without sample-complexity considerations.

Two works are closely related to ours [23, 22] under precisely the same setting. [23] is the first to study the risk-sensitive episodic MDP, which provides the first algorithms and regret guarantees. Nevertheless, the regret upper bounds contain a dispensable factor of $\exp(|\beta|H^2)$. Additionally, their lower bound proof contains mistakes, and the corrected proof suggests a weaker bound. [22] improves the algorithm by removing the additional $\mathcal{O}(\exp(|\beta|H^2))$ factor. However, the regret analysis is complicated, and the lower bound is not fixed. A very recent work ([1]) independently proposes a risk-sensitive DDP framework, but their work is fundamentally different from ours. The risk measure considered in [1] is the conditional value at risk (CVaR), and they focus on the infinite horizon setting. Due to the space limit, we provide detailed comparisons with [23, 22, 1] in Appendix A.

2 Preliminaries

Notations. We write $[M:N] \triangleq \{M, M+1, ..., N\}$ and $[N] \triangleq [1:N]$ for any positive integers $M \leq N$. We adopt the convention that $\sum_{i=n}^{m} a_i \triangleq 0$ if $n > m$ and $\prod_{i=n}^{m} a_i \triangleq 1$ if $n > m$. We use $\mathbb{I}\{\cdot\}$ to denote the indicator function. For any $x \in \mathbb{R}$, we define $[x]^+ \triangleq \max\{x, 0\}$. We define the step function with parameter c as $\psi_c(x) \triangleq \mathbb{I}\{x \geq c\}$. Note that ψ_c represents the CDF of a deterministic variable taking value c. We denote by $\mathscr{D}([a,b])$, $\mathscr{D}_M$ and $\mathscr{D}$ the set of distributions supported on $[a,b]$, $[0,M]$ and the set of all distributions respectively. For a random variable (r.v.) X, we use $\mathbb{E}[X]$ and $\mathbb{V}[X]$ to denote its expectation and variance. For two r.v.s, we denote by $X \perp Y$ if X is independent of Y. We use $\tilde{\mathcal{O}}(\cdot)$ to denote $\mathcal{O}(\cdot)$ omitting logarithmic factors.

Episodic MDP. An episodic MDP is identified by $\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, (P_h)_{h\in[H]}, (\mathcal{R}_h)_{h\in[H]}, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P_h : \mathcal{S} \times \mathcal{A} \times \to \Delta(\mathcal{S})$ the probability transition kernel at step h, $\mathcal{R}_h : \mathcal{S} \times \mathcal{A} \to \mathscr{D}([0,1])$ the collection of reward distributions 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, Nature selects an initial state s_1^k arbitrarily. In step h, the agent takes action a_h^k and observes random reward $R_h^k(s_h^k, a_h^k) \sim \mathcal{R}_h(s_h^k, a_h^k)$ and reaches the next state $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$. The episode terminates at $H+1$ with $R_{H+1}^k = 0$, then the agent proceeds to next episode.

For each $(k,h) \in [K] \times [H]$, we denote by $\mathcal{H}_h^k \triangleq \left(s_1^1, a_1^1, s_2^1, a_2^1, \ldots, s_H^1, a_H^1, \ldots, s_h^k, a_h^k\right)$ the (random) history up to step h episode k. We define $\mathcal{F}_k \triangleq \mathcal{H}_H^{k-1}$ as the history up to episode $k-1$. We describe the interaction between the algorithm and MDP in two levels. In the level of episode, we define an algorithm as a sequence of function $\mathscr{A} \triangleq (\mathscr{A}_k)_{k\in[K]}$, each mapping $\mathcal{F}_k$ to a policy $\mathscr{A}_k(\mathcal{F}_k) \in \Pi$. We denote by $\pi^k \triangleq \mathscr{A}_k(\mathcal{F}_k)$ the policy at episode k. In the level of step, a (deterministic) policy π is defined as a sequence of functions $\pi = (\pi_h)_{h\in[H]}$ with $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$.

Entropic risk measure. EntRM is a well-known risk measure in risk-sensitive decision-making, including mathematical finance [25], Markovian decision processes [3]. The EntRM value of a r.v. $X \sim F$ with coefficient $\beta \neq 0$ is defined as

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

With slight abuse of notations, we write $U_\beta(F) = U_\beta(X)$ for $X \sim F$. For β with small absolute value, using Taylor's expansion we have

$$U_\beta(X) = \mathbb{E}[X] + \frac{\beta}{2}\mathbb{V}[X] + \mathcal{O}(\beta^2). \tag{1}$$

Hence for a decision-maker who aims at maximizing the EntRM value, she tends to be risk-seeking (favoring high uncertainty in X) if $\beta > 0$ and risk-averse (favoring low uncertainty in X) if $\beta < 0$. $|\beta|$ controls the risk-sensitivity. It exactly recovers mean as the risk-neutral objective when $\beta \to 0$.

3 Risk-sensitive Distributional Dynamic Programming

[4, 40] has discussed the *infinite-horizon* distributional dynamic programming in the *risk-neutral* setting, which will be referred to as the classical DDP. There is a big gap between the risk-sensitive MDP and the risk-neutral one. In this section, we establish the novel DDP framework for risk-sensitive control.

We start with defining the return for a policy π starting from state-action pair (s,a) at step h

$$Z_h^\pi(s,a) \triangleq \sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}),\ s_h = s, a_{h'} = \pi_{h'}(s_{h'}), s_{h'+1} \sim P_{h'}(\cdot|s_{h'}, a_{h'}).$$

Define $Y_h^\pi(s) \triangleq Z_h^\pi(s, \pi_h(s))$. There are three sources of randomness in $Z_h^\pi(s,a)$: the reward $R_h(s,a)$, the transition P^π and the next-state return $Y_{h+1}^\pi(s_{h+1})$. Denote by $\nu_h^\pi(s)$ and $\eta_h^\pi(s,a)$ the cumulative distribution function (CDF) corresponding to $Y_h^\pi(s)$ and $Z_h^\pi(s,a)$ respectively. To the end of risk-sensitive control, we define the action-value function of a policy π at step h as $Q_h^\pi(s,a) \triangleq U_\beta(Z_h^\pi(s,a))$, i.e. the EntRM value of the return distribution, for each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. The value function is defined as $V_h^\pi(s) \triangleq Q_h^\pi(s, \pi_h(s)) = U_\beta(Y_h^\pi(s))$.

We focus on the control setting, in which the goal is to find an optimal policy to maximize the value function, i.e.

$$\pi^* \triangleq \arg\max_{(\pi_1,\ldots,\pi_H)\in\Pi} V_1^{\pi_1\cdots\pi_H}(s).$$

We write $\pi = (\pi_1, ..., \pi_H)$ to emphasize that it is a multi-stage maximization problem. Direct search suffers exponential computational complexity. In the risk-neutral case, the *principle of optimality* holds, i.e.,the optimal policy of tail sub-problem is the tail optimal policy [5]. Therein the multi-stage maximization problem can be reduced to a multiple single-stage maximization problem. However, the principle does not always hold for general risk measures. For example, the optimal policy for CVaR may be non-Markovian/history-dependent ([41]).

We identify two key properties of EntRM, upon which we retain the principle of optimality.

Lemma 1. *The EntRM satisfies the following properties:*

- *Additive:* $X \perp Y \Rightarrow U_\beta(X+Y) = U_\beta(X) + U_\beta(Y),\ \forall X, Y.$
- *Monotonicity-preserving:* $\forall F_1, F_2, G \in \mathscr{D},\ \forall \theta \in [0,1],$
$$U_\beta(F_2) \le U_\beta(F_1) \Rightarrow U_\beta((1-\theta)F_2 + \theta G) \le U_\beta((1-\theta)F_1 + \theta G).$$

The proof is given in Appendix B. In particular, the additivity entails that the EntRM value of the current return $Z_h^\pi(s,a)$ equals the sum of the immediate value of $R_h(s,a)$ and the value of the future return $Y_h^\pi(s')$, i.e.,
$$U_\beta(Z_h^\pi(s,a)) = U_\beta(R_h(s,a)) + U_\beta(Y_h^\pi(s')).$$
The monotonicity-preserving property together with the additivity suggests that the optimal future return $Y_h^*(s')$ consists in the optimal current return $Z_h^*(s,a)$
$$Z_h^*(s,a) = R_h(s,a) + Y_h^*(s').$$
These observations implies the principle of optimality.

Proposition 1 (Principle of optimality)**.** *Let $\pi^* = \{\pi_1^*, \pi_2^*, ..., \pi_H^*\}$ be an optimal policy and assume when we visit some state s using policy π at time-step h with positive probability. Consider the sub-problem defined by the the following maximization problem*
$$\max_{\pi \in \Pi} V_h^\pi(s) = U_\beta(\mathcal{R}_h(s,a)) + U_\beta([P_h \nu_{h+1}^\pi](s,a)).$$
Then the truncated optimal policy $\{\pi_h^*, \pi_{h+1}^*, ..., \pi_H^*\}$ is optimal for this sub-problem.

The proof is given in Appendix E. It further induces the distributional Bellman optimality equation.

Proposition 2 (Distributional Bellman optimality equation)**.** *For arbitrary initial state s_1, the optimal policy $(\pi_h^*)_{h\in[H]}$ is given by the following backward recursions:*
$$\begin{aligned}&\nu_{H+1}^*(s) = \psi_0,\ \eta_h^*(s,a) = [P_h \nu_{h+1}^*](s,a) * f_h(\cdot|s,a),\\ &\pi_h^*(s) = \arg\max_{a\in\mathcal{A}} Q_h^*(s,a) = U_\beta(\eta_h^*(s,a)), \nu_h^*(s) = \eta_h^*(s, \pi_h^*(s)),\end{aligned} \tag{2}$$
where $f_h(s,a)$ is the probability density function of $R_h(s,a)$. Furthermore, the sequence $(\eta_h^*)_{h\in[H]}$ and $(\nu_h^*)_{h\in[H]}$ are the sequence of distributions corresponding to the optimal returns at each step.

The proof is given in Appendix E. For simplicity, we define the distributional Bellman operator $\mathcal{B}(P,\mathcal{R}) : \mathscr{D}^{\mathcal{S}} \to \mathscr{D}^{\mathcal{S}\times\mathcal{A}}$ with associated model $(P,\mathcal{R}) = (P(s,a), \mathcal{R}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ as
$$[\mathcal{B}(P,\mathcal{R})\nu](s,a) \triangleq [P\nu](s,a) * f_h(\cdot|s,a),\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$
Hence we can rewrite Equation 2 in a compact form:
$$\begin{aligned}&\nu_{H+1}^*(s) = \psi_0,\ \eta_h^*(s,a) = [\mathcal{B}(P_h,\mathcal{R}_h)\nu_{h+1}^*](s,a),\\ &\pi_h^*(s) = \arg\max_{a\in\mathcal{A}} U_\beta(\eta_h^*(s,a)),\ \nu_h^*(s) = \eta_h^*(s,\pi_h^*(s)), \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H].\end{aligned} \tag{3}$$
Finally, we define the regret of an algorithm $\mathscr{A}$ interacting with an MDP $\mathcal{M}$ for K episodes as
$$\text{Regret}(\mathscr{A}, \mathcal{M}, K) \triangleq \sum_{k=1}^K V_1^*(s_1^k) - V_h^{\pi^k}(s_1^k).$$
Note that the regret is a random variable since π^k is a random quantity. We denote by $\mathbb{E}[\text{Regret}(\mathscr{A}, \mathcal{M}, K)]$ the expected regret. We will omit $\boldsymbol{\pi}$ and $\mathcal{M}$ if it is clear from the context.

4 Algorithm

For a better understanding of the readers, we present our algorithms under the assumption that the reward is *deterministic and known*[1]. The algorithms for the case of random reward are given

[1]The algorithms for random reward enjoy the regret bounds of the same order.

in Appendix C. We denote by $\{r_h(s,a)\}_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$ the reward functions. For the case of deterministic reward, the Bellman update in Equation 2 takes the form

$$\eta_h^*(s,a) = [P_h\nu_{h+1}^*](s,a)(\cdot - r_h(s,a)),$$

since adding a deterministic reward $r_h(s,a)$ corresponds to shifting the distribution $[P_h\nu_{h+1}^*](s,a)$ by an amount of $r_h(s,a)$. We thus define the distributional Bellman operator $\mathcal{B}(P,\mathcal{R}) : \mathscr{D}^{\mathcal{S}} \to \mathscr{D}^{\mathcal{S}\times\mathcal{A}}$ with associated model $(P,r) = (P(s,a), r(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ as

$$[\mathcal{B}(P,r)\nu](s,a) \triangleq [P\nu](s,a)(\cdot - r_h(s,a)),\ \forall(s,a)\in\mathcal{S}\times\mathcal{A}.$$

We propose two DRL algorithms in this section, including a model-free algorithm and a model-based algorithm. We first introduce the **M**odel- **F**ree **R**isk-sensitive **O**ptimistic **D**istribution **I**teration (`RODI-MF`) in Algorithm 1. For completeness, we introduce some additional notations here. For two CDFs F and G over reals, we define the supremum distance between them $\|F-G\|_\infty \triangleq \sup_x |F(x)-G(x)|$. We define the ℓ_1 distance between two probability mass functions (PMFs) P and Q as $\|P-Q\|_1 \triangleq \sum_i |P_i - Q_i|$. We denote by $B_\infty(F,c) := \{G\in\mathscr{D} | \|G-F\|_\infty \le c\}$ the supremum norm ball centered at F with radius c. With slight abuse of notations, we denote by $B_1(P,c)$ the l_1 norm ball centered at P with radius c.

4.1 Algorithm overview

4.1.1 RODI-MF

In each episode, the algorithm includes the planning phase (Line 4-12) and the interaction phase (Line 13-17).

Planning phase. In a high level, the algorithm implements an optimistic version of approximate DDP from step $H+1$ to step 1 in each episode. In Line (5-7), it performs sample-based Bellman update. To make it clear, we introduce the superscript k to the variables of Algorithm 1 in episode k. For example, η_h^k denotes η_h in episode k. Specifically, for those visited state-action pairs, we claim that Line 6 is equivalent to a model-based Bellman update. Denote by $\mathbb{I}_h^k(s,a) \triangleq \mathbb{I}\{(s_h^k,a_h^k)=(s,a)\}$. Fix a tuple (s,a,k,h) such that $N_h^k(s,a)\ge 1$. We denote by $\hat{P}_h^k(\cdot|s,a)$ the empirical transition model

$$\hat{P}_h^k(s'|s,a) = \frac{1}{N_h^k(s,a)} \sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\cdot\mathbb{I}\{s_{h+1}^\tau = s'\}.$$

Observe that for any $\nu\in\mathscr{D}^{\mathcal{S}}$, we have

$$\begin{aligned}
\left[\hat{P}_h^k\nu\right](s,a) &= \sum_{s'\in\mathcal{S}} \hat{P}_h^k(s'|s,a)\nu(s') = \frac{1}{N_h^k(s,a)} \sum_{s'\in\mathcal{S}}\sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\cdot\mathbb{I}\{s_{h+1}^\tau = s'\}\nu(s') \\
&= \frac{1}{N_h^k(s,a)} \sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\cdot\sum_{s'\in\mathcal{S}}\mathbb{I}\{s_{h+1}^\tau = s'\}\nu(s_{h+1}^\tau) \\
&= \frac{1}{N_h^k(s,a)} \sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\nu(s_{h+1}^\tau).
\end{aligned}$$

Hence the update formula in Line 6 of Algorithm 1 can be rewritten as

$$\eta_h^k(s,a) = \left[\hat{P}_h^k\nu_{h+1}^k\right](s,a)(\cdot - r_h(s,a)) = \left[\mathcal{B}(\hat{P}_h^k, r_h)\nu_{h+1}^k\right](s,a),$$

implying the equivalence to a model-based Bellman update with empirical model $\hat{P}_h^k$. Alternatively, the unvisited (s,a) remains to be the return distribution corresponding to the highest possible reward $H+1-h$. The algorithm then computes the optimism constants (Line 8) and enforces OFU through the distributional optimism operator c_h^k (Line 9) to obtain the optimistically plausible return distribution η_h^k. The choice of c_h^k will be discussed later. The optimistic return distributions yields the optimistic value function, from which the algorithm generates the greedy policy π_h^k. The policy π_h^k will be used in the interaction phase.

Interaction phase. In Line (15-16), the agent interacts with the environment using policy π and updates the counts N_h based on new observations.

Algorithm 1 `RODI-MF`

1: Input: T and δ
2: Initialize $N_h(\cdot,\cdot) \leftarrow 0$; $\eta_h(\cdot,\cdot), \nu_h(\cdot) \leftarrow \psi_{H+1-h}$ for all $h \in [H]$
3: **for** $k = 1 : K$ **do**
4: **for** $h = H : 1$ **do**
5: **if** $N_h(\cdot,\cdot) > 0$ **then**
6: $\eta_h(\cdot,\cdot) \leftarrow \frac{1}{N_h(\cdot,\cdot)} \sum_{\tau \in [k-1]} \mathbb{I}_h^\tau(\cdot,\cdot)\nu_{h+1}(s_{h+1}^\tau)(\cdot - r_h(\cdot,\cdot))$
7: **end if**
8: $c_h(\cdot,\cdot) \leftarrow \sqrt{\frac{2S}{N_h(\cdot,\cdot)\vee 1}\iota}$
9: $\eta_h(\cdot,\cdot) \leftarrow \mathrm{O}^\infty_{c_h(\cdot,\cdot)}\eta_h(\cdot,\cdot)$
10: $\pi_h(\cdot) \leftarrow \arg\max_a U_\beta(\eta_h(\cdot,a))$
11: $\nu_h(\cdot) \leftarrow \eta_h(\cdot,\pi_h(\cdot))$
12: **end for**
13: Receive s_1^k
14: **for** $h = 1 : H$ **do**
15: $a_h^k \leftarrow \pi_h(s_h^k)$ and transit to s_{h+1}^k
16: $N_h(s_h^k,a_h^k) \leftarrow N_h(s_h^k,a_h^k) + 1$
17: **end for**
18: **end for**

4.1.2 RODI-MB

We introduce the second algorithm **M**odel- **B**ased **R**isk-sensitive **O**ptimistic **D**istribution **I**teration (`RODI-MB`). Algorithm 2 is a model-based algorithm because it requires to explicitly maintaining the empirical transition model in each episode. However, it can be reduced to a *non-distributional* reinforcement learning algorithm that deals with the one-dimensional values instead of the distributions, which saves the computational complexity and space complexity. Likewise, the algorithm includes the planning phase (Line 4-10) and the interaction phase (Line 11-15).

Planning phase. Analogous to Algorithm 1, the algorithm also performs approximate DDP together with the OFU principle. First, it applies the distributional optimistic operator to the empirical transition model $\hat{P}_h^k$ to get the optimistic transition model $\tilde{P}_h^k$. Then the algorithm uses $\tilde{P}_h^k$ to execute Bellman update to generate the optimistic return distributions η_h^k. The remaining steps are the same as Algorithm 1.

Interaction phase. In Line (13-14), the agent interacts with the environment using policy π^k and updates the counts N_h^{k+1} and empirical transition model $\hat{P}_h^{k+1}$ based on the new observations.

Algorithm 2 `RODI-MB`

1: Input: T and δ
2: $N_h^1(\cdot,\cdot) \leftarrow 0$; $\hat{P}_h^1(\cdot,\cdot) \leftarrow \frac{1}{S}\mathbf{1}$ for all $h \in [H]$
3: **for** $k = 1 : K$ **do**
4: $\nu_{H+1}^k(\cdot) \leftarrow \psi_0$
5: **for** $h = H : 1$ **do**
6: $\tilde{P}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^1_{c_h^k(\cdot,\cdot)}\hat{P}_h^k(\cdot,\cdot)$
7: $\eta_h^k(\cdot,\cdot) \leftarrow \left[\mathcal{B}\left(\tilde{P}_h^k, r_h\right)\nu_{h+1}^k\right](\cdot,\cdot)$
8: $\pi_h^k(\cdot) \leftarrow \arg\max_a U_\beta(\eta_h^k(\cdot,a))$
9: $\nu_h^k(\cdot) \leftarrow \eta_h^k(\cdot,\pi_h^k(\cdot))$
10: **end for**
11: Receive s_1^k
12: **for** $h = 1 : H$ **do**
13: $a_h^k \leftarrow \pi_h^k(s_h^k)$ and transit to s_{h+1}^k
14: Compute $N_h^{k+1}(\cdot,\cdot)$ and $\hat{P}_h^{k+1}(\cdot,\cdot)$
15: **end for**
16: **end for**

Algorithm 3 `ROVI`

1: Input: T and δ
2: $N_h^1(\cdot,\cdot) \leftarrow 0$; $\hat{P}_h^1(\cdot,\cdot) \leftarrow \frac{1}{S}\mathbf{1}$ for all $h \in [H]$
3: **for** $k = 1 : K$ **do**
4: $W_{H+1}^k(\cdot) \leftarrow 1$
5: **for** $h = H : 1$ **do**
6: $\tilde{P}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^1_{c_h^k(\cdot,\cdot)}\hat{P}_h^k(\cdot,\cdot)$
7: $J_h^k(\cdot,\cdot) \leftarrow e^{\beta r_h(\cdot,\cdot)}\left[\tilde{P}_h^k W_{h+1}^k\right](\cdot,\cdot)$
8: $W_h^k(\cdot) \leftarrow \max_a J_h^k(\cdot,a)$
9: **end for**
10: Receive s_1^k
11: **for** $h = 1 : H$ **do**
12: $a_h^k \leftarrow \arg\max_a J_h^k(s_h^k,a)$ and transit to s_{h+1}^k
13: Compute $N_h^{k+1}(\cdot,\cdot)$ and $\hat{P}_h^{k+1}(\cdot,\cdot)$
14: **end for**
15: **end for**

Equivalence to ROVI. **R**isk-sensitive **O**ptimistic **V**alue **I**teration (ROVI) is a non-distributional algorithm that deals with the real-valued value function rather than the distribution. It is motivated by the *exponential Bellman equation* proposed by [22]. We define the functional exponential EntRM (EERM) E_β as the EntRM after the exponential transformation

$$E_\beta(F) \triangleq \exp(\beta(U_\beta(F))) = \int_{\mathbb{R}} \exp(\beta x) dF(x).$$

Define the exponential value functions $W_h(s) \triangleq E_\beta(\nu_h(s))$ and $J_h(s,a) \triangleq E_\beta(\eta_h(s,a))$ for all (s,a,h)s. Applying EERM to Equation 3 yields the exponential Bellman equation

$$\begin{aligned} &J_h^*(s,a) = \exp(\beta r_h(s,a))[P_h W_{h+1}^*](s,a), \\ &W_h^*(s) = \text{sign}(\beta) \max_a \text{sign}(\beta) J_h^*(s,a),\ W_{H+1}^*(s) = 1. \end{aligned} \tag{4}$$

To verify the equivalence, it is sufficient to show that J_h^k in Algorithm 3 corresponds to the exponential function of η_h^k in Algorithm 2. Observe that E_β is linear in F, hence it follows that

$$\begin{aligned} E_\beta(\eta_h^k(s,a)) &= E_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right](s,a)(\cdot - r_h(s,a))\right) = \exp(\beta r_h(s,a)) \cdot \left[\tilde{P}_h^k E_\beta(\nu_{h+1}^k)\right](s,a) \\ &= \exp(\beta r_h(s,a)) \left[\tilde{P}_h^k W_{h+1}^k\right](s,a) = J_h^k(s,a). \end{aligned}$$

The two algorithms generate the policy sequence in the same way, implying that their trajectories $\mathcal{H}_H^K$ follow the same distribution. The formal statement is given in Appendix E.

4.2 Distributional Optimism

It is common to add a bonus to the reward to ensure optimism in the risk-neutral setting. Specifically, the bonus is closely related to the level of uncertainty, which is quantified by the concentration inequality. Yet, this type of optimism cannot be adapted to the distributional setup. As one of our technical novelty, the *distributional optimism* is introduced for algorithmic design and regret analysis. In particular, we specify two types of distributional optimism operators, which map a statistically plausible distribution (either the empirical model or the return distribution) to a optimistically plausible distribution. Either of them is applied by Algorithm 2 or Algorithm 1.

Distributional optimism on the return distribution (in Algorithm 1). For two CDFs F and G, we say that F is more optimistic than G (w.r.t. EntRM) if $U_\beta(F) \geq U_\beta(G)$. This reflects the intuition that the more optimistic distribution should own larger EntRM value. Following [31], we define the distributional optimism operator $\mathrm{O}_c^\infty : \mathscr{D}([a,b]) \mapsto \mathscr{D}([a,b])$ with level $c \in (0,1)$ as

$$(\mathrm{O}_c^\infty F)(x) \triangleq [F(x) - c\mathbb{I}_{[a,b)}(x)]^+.$$

The optimistic operator shifts the input F down by at most c over $[a,b)$, and retain the value 1 at b. It ensures that $\mathrm{O}_c^\infty F$ remains in $\mathscr{D}([a,b])$ and dominates all the other CDFs in $\mathscr{D}([a,b])$ in the sense that $(\mathrm{O}_c^\infty F)(x) \leq G(x)$ for any $G \in B_\infty(F,c)$. Since EntRM is monotonic, it holds that

$$U_\beta(\mathrm{O}_c^\infty F) \geq U_\beta(G),\ \forall G \in B_\infty(F,c).$$

Hence $\mathrm{O}_c^\infty F$ is the most optimistic distribution in the infinity ball $B_\infty(F,c)$. In other words, for any CDF F and G satisfying $\|F - G\|_\infty \leq c$, we have $\mathrm{O}_c^\infty G \succeq F$. When specialized to the return distributions, we can apply the distributional optimism operator to the estimated return distribution η_h^k (Line 9 of Algorithm 1) with the constant c_h^k to ensure $U_\beta(\eta_h^k(s,a)) \geq U_\beta(\eta_h^*(s,a))$. The constant c_h^k quantifies uncertainty in the model estimation, i.e., $\left\|\hat{P}_h^k(s,a) - P_h(s,a)\right\|_1$.

Distributional optimism on the model (in Algorithm 2). Given the model, we consider the optimism among the space of PMFs rather than CDFs. Using the ℓ_1 concentration inequality [46], we get a concentration bound of the empirical PMF of model: with probability at least $1-\delta$,

$$\left\|\widehat{P}_h^k(s,a) - P_h(s,a)\right\|_1 \leq c_h^k(s,a) = \sqrt{\frac{2S}{N_h^k(s,a)} \log\frac{1}{\delta}} = \tilde{\mathcal{O}}\left(\sqrt{\frac{2S}{N_h^k(s,a)}}\right).$$

We wish to obtain a optimistic transition model $\tilde{P}_h^k(s,a)$ from the empirical one $\widehat{P}_h^k(s,a)$. To be more specific, the return distribution η_h^k computed from $\tilde{P}_h^k(s,a)$ and ν_{h+1}^k should be more optimistic than the optimal one $\eta_h^*(s,a)$ with high probability. We thus define the distributional optimism operator $\mathrm{O}_c^1 : \mathscr{D}(\mathcal{S}) \mapsto \mathscr{D}(\mathcal{S})$ with level c and future return $\nu \in \mathscr{D}^{\mathcal{S}}$ as

$$\mathrm{O}_c^1\left(\widehat{P}(s,a),\nu\right) \triangleq \arg\max_{P\in B_1(\widehat{P}(s,a),c)} U_\beta([P\nu]).$$

The ERM satisfy an interesting property that enables an efficient approach to perform O_c^1 (see Appendix B). The following holds by using the induction

$$\begin{aligned} U_\beta\left(\eta_h^k(s,a)\right) = r_h(s,a) + U_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right][s,a]\right) &\geq r_h(s,a) + U_\beta\left(\left[P_h \nu_{h+1}^k\right][s,a]\right) \\ &\geq r_h(s,a) + U_\beta\left(\left[P_h \nu_{h+1}^*\right][s,a]\right) \\ &= U_\beta(\eta_h^*(s,a)), \end{aligned}$$

which verify the optimism of $\eta_h^k(s,a)$ over $\eta_h^*(s,a)$.

5 Regret Analysis

5.1 Regret upper bounds

Theorem 1 (Regret upper bound of `RODI-MF`)**.** *For any $\delta \in (0,1)$, with probability $1-\delta$, the regret of Algorithm 1 under deterministic reward or Algorithm 4 under random reward is bounded as*

$$\textit{Regret}(\texttt{RODI-MF}, K) \leq \mathcal{O}\left(\frac{1}{|\beta|} L_H H \sqrt{S^2 AK \log(4SAT/\delta)}\right) = \tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H)-1}{|\beta|} H\sqrt{S^2AK}\right).$$

The proof is given in Appendix D.

Theorem 2 (Regret upper bound of `RODI-MB/ROVI`)**.** *For any $\delta \in (0,1)$, with probability $1-\delta$, the regret of Algorithm 1/Algorithm 3 under deterministic reward or Algorithm 4/Algorithm 6 under random reward is bounded as*

$$\begin{aligned} \textit{Regret}(\texttt{RODI-MF}, K) = \textit{Regret}(\texttt{ROVI}, K) &\leq \mathcal{O}(\frac{1}{|\beta|} L_H H \sqrt{S^2 AK \log(4SAT/\delta)}) \\ &= \tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H)-1}{|\beta|} H\sqrt{S^2AK}\right). \end{aligned}$$

The proof is given in Appendix D. The above results match the best-known results in [22]. In particular, our algorithms attain exponentially improved regret bounds than those of RSVI and RSQ in [23] with a factor of $\exp(|\beta|H^2)$. By choosing $|\beta| = \mathcal{O}(1/H)$, we can eliminate the exponential term and achieve polynomial regret bound akin to the risk-neutral setting.

Compared to the traditional/non-distributional analysis dealing with one-dimensional values, our analysis is distribution-centered, called the *distributional analysis*. The distributional analysis deals with the distributions of the return rather than the risk measure values of the return. For example, it involves the operations of the distributions, the optimism between different distributions, the error caused by estimation of distribution, etc. These distributional aspects fundamentally differ from the traditional analysis that deals with the one-dimensional scalars (value functions). Now we recap the technical novelty of our analysis in the following.

Lipschitz continuity and linearity. We identify two important properties of EERM that establishes the regret upper bounds, including the Lipschitz continuity and linearity. Denote by L_M the Lipschitz constant of the EERM $E_\beta : \mathscr{D}([0,M]) \to \mathbb{R}$ with respect to the infinity norm $\|\cdot\|_\infty$. Lemma 2 provides a *tight* Lipschitz constant of EERM. The Lipschitz constant relates the difference between distributions to the difference measured by their EERM values.

Lemma 2 (Lipschitz property of EERM)**.** *E_β is Lipschitz continuous with respect to the supremum norm over $\mathscr{D}_M$ with $L_M = \exp(|\beta|M) - 1$. Moreover, L_M is tight in terms of both $|\beta|$ and M.*

Notice that $\lim_{\beta\to 0} L_M = 0$, which coincides with the fact that $\lim_{\beta\to 0} E_\beta = 1$. The linearity of EERM is a key property that sharpens the regret bounds. In contrast, EntRM is non-linear in the distribution, which could induce a factor of $\exp(|\beta|H)$ when controlling the error propagation across time-steps. It would further lead to a compounding factor of $\exp(|\beta|H^2)$ in the regret bound. In summary, the Lipschitz continuity property enables the regret upper bounds of DRL algorithms, and the linearity tightens the bound.

Distributional optimism. Another technical novelty in our analysis is the optimism in the face of uncertainty at the distributional level. The traditional analysis uses the OFU to construct a sequence of optimistic value functions. However, our analysis implements the *distributional optimism* that yields a sequence of optimistic return distributions. In particular, we first define a high probability event, under which the true return distribution concentrates around the estimated one with a certain confidence radius. Then we apply the distributional optimism operator to obtain the optimistically plausible return distribution and the optimistic EntRM value. Hence the regret can be bounded by the surrogate regret, with the optimal EntRM value replaced by

$$\text{Regret}(K) = \sum_{k=1}^{K} \frac{1}{\beta} \log\left(W_1^*(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right) \leq \frac{1}{\beta}\sum_{k=1}^{K} W_1^k(s_1^k) - W_1^{\pi^k}(s_1^k).$$

Distributional analysis vs. non-distributional analysis. When analyzing Algorithm 2/Algorithm 3, proving the regret bound of either algorithm suffices due to their equivalence relation. Since Algorithm 3 is a non-distributional algorithm, one may consider using the standard analysis that does not involve distributions. However, we show that this induces a factor of $\frac{1}{|\beta|}\exp(|\beta|H)$, which explodes as $|\beta| \to 0$. We overcome this issue by invoking a novel distributional analysis of Algorithm 2, leading to the desired factor of $\frac{1}{|\beta|}(\exp(|\beta|H) - 1)$.

Although we focus on the algorithms for the deterministic reward in the main text, the regret upper bounds also hold for case of random reward. Algorithm 4, Algorithm 5 and Algorithm 6 corresponds to Algorithm 1, Algorithm 2 and Algorithm 3 respectively (cf. Appendix C).

5.2 Regret lower bound

We provide more details of the mistakes in the lower bound of [23] in Appendix D. The proof of [23] reduces the regret lower bound to the two-armed bandit regret lower bound. Since the two-armed bandit is a special case of MDP with $S = 1$, $A = 2$ and $H = 1$, the reduction-based proof only leads to a lower bound independent of S, A, and H. Instead, our tight lower bound follows a totally different roadmap motivated by [20]. [20] proves the tight minimax lower bound $H\sqrt{SAT}$ for risk-neutral MDP. However, the generalization to risk-sensitive MDP is non-trivial. The main technical challenge is due to the non-linearity of EntRM. The proof in [23] heavily relies on the linearity of expectation, allowing the exchange between taking the risk measure (expectation) and the summation. In the risk-sensitive setting, the non-linearity of EntRM requires new proof techniques.

Assumption 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 3 (Tighter lower bound)**.** *Assume Assumption 1 holds and $\beta > 0$. Let $\bar{L} \triangleq (1 - \frac{1}{A})(S - 3) + \frac{1}{A}$. Then for any algorithm $\mathscr{A}$, there exists an MDP $\mathcal{M}_{\mathscr{A}}$ such that for $K \geq 2\exp(\beta(H - \bar{H} - d))\bar{H}\bar{L}A$ we have*

$$\mathbb{E}[\textit{Regret}(\mathscr{A}, \mathcal{M}_{\mathscr{A}}, K)] \geq \frac{1}{72\sqrt{6}} \frac{\exp(\beta H/6) - 1}{\beta H} H\sqrt{SAT}.$$

The proof is given in Appendix D. Theorem 3 recovers the tight lower bound for standard episodic MDP, implying that the exponential dependence on $|\beta|$ and H in the upper bounds is indispensable. Yet, it is not clear whether a similar lower bound holds for $\beta < 0$, which is left as a future direction.

6 Conclusion

We propose a risk-sensitive distributional dynamic programming framework. We devise two novel DRL algorithms, including a model-free one and a model-based one, which implement the OFU principle at the distributional level to balance the exploration and exploitation trade-off under the risk-sensitive setting. We prove that both attain near-optimal regret upper bounds compared with our improved lower bound.

There are several promising future directions. The current regret upper bound has an additional factor $\sqrt{HS}$ compared with the lower bound. It might be possible to remove the factor by designing new algorithms or improving the analysis. Besides, it is interesting to extend the DRL algorithm from tabular MDP to linear function approximation setting. Finally, it will be meaningful to investigate whether the DDP framework holds for other risk measures.

References

[1] Mastane Achab and Gergely Neu. Robustness and risk management via distributional dynamic programming. *arXiv preprint arXiv:2112.15430*, 2021.

[2] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

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

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

[5] Dimitri P Bertsekas et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

[6] Tomasz R Bielecki, Stanley R Pliska, and Michael Sherris. Risk sensitive asset allocation. *Journal of Economic Dynamics and Control*, 24(8):1145–1177, 2000.

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

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

[9] Vivek S Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5, 2010.

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

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

[12] Stefano P Coraluppi and Steven I Marcus. Risk-sensitive, minimax, and mixed risk-neutral/minimax control of markov decision processes. In *Stochastic analysis, control, optimization and applications*, pages 21–40. Springer, 1999.

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

[14] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] Mark Davis and Sébastien Lleo. Risk-sensitive benchmarked asset management. *Quantitative Finance*, 8(4):415–426, 2008.

[16] Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

[17] Giovanni B Di Masi et al. Infinite horizon risk sensitive control of discrete time markov processes with small risk. *Systems & control letters*, 40(1):15–20, 2000.

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

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

[20] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

[21] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

[22] Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *arXiv preprint arXiv:2006.13827*, 2020.

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

[25] Hans Föllmer and Alexander Schied. Stochastic finance. In *Stochastic Finance*. de Gruyter, 2016.

[26] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

[27] Lars Peter Hansen and Thomas J Sargent. Robustness. In *Robustness*. Princeton university press, 2011.

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

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

[30] Anna Jaśkiewicz. Average optimality for risk-sensitive control with general state space. *The annals of applied probability*, 17(2):654–675, 2007.

[31] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[32] Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019.

[33] Xiaoteng Ma, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2004.14547*, 2020.

[34] Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[35] Steven I Marcus, Emmanual Fernández-Gaucherand, Daniel Hernández-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pages 263–279. Springer, 1997.

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

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

[38] Takayuki Osogami. Robustness and risk-sensitivity in markov decision processes. *Advances in Neural Information Processing Systems*, 25:233–241, 2012.

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

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

[41] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

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

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

[44] Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Learning for Dynamics and Control*, pages 958–968. PMLR, 2020.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[46] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Negative Social Impact

This script may provide better guidance for risk-sensitive reinforcement learning community. It would have certain negative social impact if the proposed algorithms are deployed for illegal usage.

A Comparisons with Related Works

Comparison with [1] We summarize the differences between our work and [1] as follows.

- Setting. [1] considers the discounted MDP with infinite horizon, but we consider the episodic MDP setting. Moreover, [1] assumes that the model is known, while we propose DRL algorithms when the model is unknown (i.e., the learning). Neither RL algorithms suitable for unknown model nor sample complexity guarantee is provided in their work.
- Risk measure. [1] establish the risk-sensitive DDP framework using the risk measure Conditional Value at Risk, while our work considers the entropic risk measure.

Comparison with [23, 22] [21,22] solved the risk-sensitive MDP problem using *valued-based* RL, which estimates and constructs the optimistic version of the (EntRM) value function. [21] proposed the RSVI2 algorithm that improved upon [22] and achieved the best result with the regret upper bound of $\tilde{\mathcal{O}}(\frac{\exp(|\beta|H)-1}{|\beta|}H\sqrt{S^2AK})$. The significance of the proposed algorithms is three-fold.

- Our algorithms are the first distributional reinforcement learning algorithms with provably regret guarantees, suggesting that DRL can work well and even matches the performance of the SOTA value-based RL algorithm for risk-sensitive control in terms of sample complexity. The idea of leveraging the distributional information for risk-sensitivity purposes is natural since the risk measure value is obtained by applying the risk measure/functional to the return distribution. However, existing works on risk-sensitive control via DRL approaches [12, 31, 1] lack regret analysis. Thus, it is difficult to evaluate and improve their algorithms for sample efficiency. Therefore, our algorithms with near-optimal regret upper bounds bridge the gap between the DRL and risk-sensitive MDP in the theoretic RL community.
- Compared with [21], our algorithms are simpler and easier to interpret, leading to clean regret analysis. [21] implements optimism by adding a bonus to the risk measure value function. It designed an exploration mechanism called doubly decaying bonus to remove the $\exp(|\beta|H^2)$ factor from [22]. The doubly decaying bonus decays across the episode and the horizon, which is complicated and not straightforward. Instead, our algorithms implement the distributional optimism by iteratively constructing the optimistic return distribution. The distributional optimism does not involve a complicated bonus design. It only requires a simple application of distributional optimism operator with a constant decaying across the episode. Moreover, the doubly decaying bonus obscures the regret analysis, while our distributional-based analysis is clean and easy to follow.
- Our algorithm may be generalized to risk-sensitive MDP with other risk measures. The analysis of [22,23] is particularly suitable for the EntRM. It is unclear whether it is possible to extend to other risk measures. Under the distributional perspective, our algorithm maintains a sequence of optimistically plausible estimates of the return distribution. Since the distributional information suffices to deal with any risk measure, our algorithm may motivate the design of similar algorithms for other risk measures.

B Further Statements about the Properties

B.1 Proof of properties of EntRM

Proof of Lemma 1. We only prove the case that $\beta > 0$. The case that $\beta < 0$ follows analogously. For any two independent random variables X and Y, we have

$$\begin{aligned}U_\beta(X+Y) = \frac{1}{\beta}\log\mathbb{E}[\exp(\beta(X+Y))] &= \frac{1}{\beta}\log\mathbb{E}[\exp(\beta X)\cdot\exp(\beta Y)]\\ &= \frac{1}{\beta}\log\mathbb{E}[\exp(\beta X)] + \frac{1}{\beta}\log\mathbb{E}[\exp(\beta Y)]\\ &= U_\beta(X) + U_\beta(Y),\end{aligned}$$

therefore ERM is additive.

For any two distributions F_1 and F_2 such that $U_\beta(F_1) > U_\beta(F_2)$, we have

$$U_\beta(F_1) = \frac{1}{\beta}\log\int_{\mathbb{R}}\exp(\beta x)dF_1(x) > \frac{1}{\beta}\log\int_{\mathbb{R}}\exp(\beta x)dF_2(x) = U_\beta(F_1),$$

which implies $\int_{\mathbb{R}}\exp(\beta x)dF_1(x) > \int_{\mathbb{R}}\exp(\beta x)dF_2(x)$. Thus for any distribution G, it follows that

$$\begin{aligned}
U_\beta(\theta F_1 + (1-\theta)G) &= \frac{1}{\beta}\log\int_{\mathbb{R}}\exp(\beta x)d(\theta F_1(x) + (1-\theta)G(x)) \\
&= \frac{1}{\beta}\log\left(\theta\int_{\mathbb{R}}\exp(\beta x)dF_1(x) + (1-\theta)\int_{\mathbb{R}}\exp(\beta x)dG(x)\right) \\
&> \frac{1}{\beta}\log\left(\theta\int_{\mathbb{R}}\exp(\beta x)dF_2(x) + (1-\theta)\int_{\mathbb{R}}\exp(\beta x)dG(x)\right) \\
&= U_\beta(\theta F_2 + (1-\theta)G).
\end{aligned}$$

For any distributions F and G such that $U_\beta(F) > U_\beta(G)$ and $\theta > \theta'$, it holds that

$$\begin{aligned}
&\int_{\mathbb{R}}\exp(\beta x)d(\theta F(x) + (1-\theta)G(x)) - \int_{\mathbb{R}}\exp(\beta x)d(\theta' F(x) + (1-\theta')G(x)) \\
&= (\theta - \theta')\left(\int_{\mathbb{R}}\exp(\beta x)dF(x) - \int_{\mathbb{R}}\exp(\beta x)dG(x)\right) > 0.
\end{aligned}$$

Since $t \mapsto \frac{1}{\beta}\log(t)$ is a strictly monotonic mapping, we have $U_\beta(\theta F + (1-\theta)G) > U_\beta(\theta' F + (1-\theta')G)$. Hence ERM satisfies the monotonicity-preserving property. □

B.2 Monotonicity preserving

We state some lemmas about the monotonicity-preserving property and their proofs here. Note that the results hold for general risk measures satisfying the monotonicity-preserving property. They will be used in the proof of Proposition 1 and Proposition 2.

Lemma 3. *Let* T *be a risk measure satisfying the monotonicity-preserving property and* $n \geq 2$ *be an arbitrary integer. If* $\mathrm{T}(F_i) \geq \mathrm{T}(G_i), \forall i \in [n]$ *(and* $\mathrm{T}(F_j) \neq \mathrm{T}(G_j)$ *for some* $j \in [n]$*) then* $\mathrm{T}\left(\sum_{i=1}^n \theta_i F_i\right) \geq (>) \mathrm{T}(\sum_{i=1}^n \theta_i G_i)$ *for any* $\theta \in \Delta_n$ *(and* $\theta_j \neq 0$*).*

Proof. The proof follows from induction. Note that $\sum_{i=1}^n \theta_i F_i = \theta_1 F_1 + (1-\theta_1)\sum_{i=2}^n \frac{\theta_i}{1-\theta_1}F_i$ and $\sum_{i=2}^n \frac{\theta_i}{1-\theta_1}F_i \in \mathscr{D}$, therefore by the definition of MP we have $\mathrm{T}(\sum_{i=1}^n \theta_i F_i) \geq \mathrm{T}(\theta_1 G_1 + \sum_{i=2}^n \theta_i F_i)$. Suppose that for some $k \in [n-1]$ it holds that $\mathrm{T}(\sum_{i=1}^n \theta_i F_i) \geq \mathrm{T}(\sum_{i=1}^k \theta_i G_i + \sum_{i=k+1}^n \theta_i F_i)$. Since

$$\begin{aligned}
\sum_{i=1}^k \theta_i G_i + \sum_{i=k+1}^n \theta_i F_i &= \theta_{k+1}F_{k+1} + \sum_{i=1}^k \theta_i G_i + \sum_{i=k+2}^n \theta_i F_i \\
&= \theta_{k+1}F_{k+1} + (1-\theta_{k+1})\left[\sum_{i=1}^k \frac{\theta_i}{1-\theta_{k+1}}G_i + \sum_{i=k+2}^n \frac{\theta_i}{1-\theta_{k+1}}F_i\right]
\end{aligned}$$

and $\frac{1}{1-\theta_{k+1}}\left[\sum_{i=1}^k \theta_i G_i + \sum_{i=k+2}^n \theta_i F_i\right] \in \mathscr{D}$, it follows that

$$\mathrm{T}\left(\sum_{i=1}^n \theta_i F_i\right) \geq \mathrm{T}\left(\sum_{i=1}^k \theta_i G_i + \sum_{i=k+1}^n \theta_i F_i\right) \geq \mathrm{T}\left(\sum_{i=1}^{k+1} \theta_i G_i + \sum_{i=k+2}^n \theta_i F_i\right).$$

The induction is completed. If in addition for some $j \in [n]$ it holds that $\mathrm{T}(F_j) > \mathrm{T}(G_j)$, the proof follows analogously by replacing the inequality to the strict one and the fact that $\theta_j > 0$. □

Lemma 4 (Monotonicity-preserving under pairwise transport)**.** *Let* T *be a risk measure satisfying the monotonicity-preserving property. Suppose* $n \geq 2$ *and* $(F_i)_{i\in[n]}$ *satisfies* $\mathrm{T}(F_1) \leq \mathrm{T}(F_2)... \leq \mathrm{T}(F_n)$*. For any* $\theta, \theta' \in \Delta_n$ *and any* $1 \leq i < j \leq n$ *such that*

$$\begin{cases} \theta'_i \leq \theta_i, \\ \theta'_j \geq \theta_j, \\ \theta'_k = \theta_k, \quad k \neq i, j \end{cases}$$

It holds that $\mathrm{T}(\sum_{i=1}^n \theta_i F_i) \leq \mathrm{T}(\sum_{i=1}^n \theta'_i F_i)$*.*

Proof. Observe that

$$\begin{aligned} \sum_{k=1}^n \theta'_k F_k = \theta'_i F_i + \theta'_j F_j + \sum_{k\neq i,j} \theta'_k F_k &= \theta'_i F_i + \theta'_j F_j + \sum_{k\neq i,j} \theta_k F_k \\ &= (\theta'_i F_i + \theta'_j F_j) + (1 - \theta_i - \theta_j) \sum_{k\neq i,j} \theta_k F_k. \end{aligned}$$

By the definition of the monotonicity-preserving property, it suffices to prove $\mathrm{T}(\frac{1}{\theta_i+\theta_j}(\theta'_i F_i + \theta'_j F_j)) \geq \mathrm{T}(\frac{1}{\theta_i+\theta_j}(\theta_i F_i + \theta_j F_j))$. The result follows from the definition and the fact that $\mathrm{T}(F_i) \leq \mathrm{T}(F_j)$ and $\theta'_i \leq \theta_i$. □

Lemma 5 (Monotonicity-preserving under block-wise transport)**.** *Suppose* $n \geq 2$ *and* $(F_i)_{i\in[n]}$ *satisfies* $\mathrm{T}(F_1) \leq \mathrm{T}(F_2)... \leq \mathrm{T}(F_n)$*. It holds that* $\mathrm{T}(\sum_{i=1}^n \theta_i F_i) \leq \mathrm{T}(\sum_{i=1}^n \theta'_i F_i)$ *for any* $\theta, \theta' \in \Delta_n$ *satisfying* $\exists k \in [n], \theta'_i \leq \theta_i$ *if* $i \leq k$ *and* $\theta'_i \geq \theta_i$ *otherwise.*

Proof. Fix $k \in [n]$. We rewrite the assumption imposed to θ' as $\theta'_i = \theta_i - \delta_i$ for $i \leq k$ and $\theta'_i = \theta_i + \delta_i$ for $i > k$, where each $\delta_i \geq 0$. It will be shown that there exists a sequence $\{\theta^l\}_{l\in[k]}$ satisfying $\theta^0 = \theta$ and $\theta^k = \theta'$ such that $\mathrm{T}(\theta^l) \leq \mathrm{T}(\theta^{l+1})$, then the proof shall be completed.

The sequence is constructed as follows: at the l-th iteration, we transport probability mass δ_l of θ_l to the probability mass of $k+1, ..., n$. Specifically, we start from moving to the least number $i_l \geq i_{l-1}$ that satisfy $\theta^{l-1}_{i_l} < \theta'_{i_l}$ and sequentially move to the next one if there is remaining mass. The iteration stops until all the mass δ_l are transported. Repeating the procedure for k times we obtain $\theta^k = \theta'$. The inequality $\mathrm{T}(\theta^l) \leq \mathrm{T}(\theta^{l+1})$ for each iteration follows from Lemma 4. □

B.3 Proof of properties of EERM

Proof of Lemma 2. We only provide the proof for the case $\beta > 0$. The case $\beta < 0$ follows from analogous arguments. For any $F, G \in \mathscr{D}_M$, without loss of generality we assume $\int_0^M G(x) d\exp(\beta x) - \int_0^M F(x) d\exp(\beta x) \geq 0$, otherwise we switch the order.

$$\begin{aligned} |E_\beta(F) - E_\beta(G)| &= \left| \int_0^M \exp(\beta x) dF(x) - \int_0^M \exp(\beta x) dG(x) \right| \\ &= \left| \exp(\beta x) F(x)|_0^M - \int_0^M F(x) d\exp(\beta x) - \exp(\beta x) G(x)|_0^M + \int_0^M G(x) d\exp(\beta x) \right| \\ &= \int_0^M (G(x) - F(x)) d\exp(\beta x) \\ &\leq \int_0^M |G(x) - F(x)| \, d\exp(\beta x) \\ &\leq \|F - G\|_\infty \int_0^M 1 d\exp(\beta x) \\ &= (\exp(\beta M) - 1) \|F - G\|_\infty . \end{aligned}$$

To show the tightness of the constant, consider two scaled Bernoulli distributions $F = (1-\mu_1)\psi_0 + \mu_1\psi_M$ and $G = (1-\mu_2)\psi_0 + \mu_2\psi_M$ with $\Delta := \mu_1 - \mu_2 > 0$, where $\mu_1, \mu_2 \in (0,1)$ are some constants to be determined. It holds that

$$\begin{aligned} E_\beta(F) - E_\beta(G) &= \mu_1 \exp(\beta M) + 1 - \mu_1 - (\mu_2 \exp(\beta M) + 1 - \mu_2) \\ &= (\mu_1 - \mu_2)(\exp(\beta M) - 1) \\ &= \|F - G\|_\infty (\exp(\beta M) - 1). \end{aligned}$$

where the last equality holds since $\|F-G\|_\infty = F(0) - G(0) = \mu_1 - \mu_2 = \Delta$ (independent of M). More formally, we have

$$\inf_{M>0,\beta>0} \sup_{F,G\in\mathscr{D}_M} \frac{|E_\beta(F) - E_\beta(G)|}{\|F-G\|_\infty} = \exp(\beta M) - 1.$$

□

C Algorithms for the Random Reward

We present the algorithms for the random reward in this section, which share the same intuitions as the deterministic reward case. Therefore we focus on clarifying their differences here. We denote by $\delta(\cdot)$ the Dirac delta function.

C.1 RODI-MF

In each episode, the algorithm includes the planning phase (Line 4-12) and the interaction phase (Line 13-17). We highlight two key differences in the planning phase. We introduce the superscript k to the variables of Algorithm 4 in episode k. The first difference is that the algorithm *implicitly* maintains the empirical reward distribution in addition to the empirical transition model

$$\hat{\mathcal{R}}_h^k(s,a) = \frac{\sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\delta(\cdot - R_h^\tau)}{N_h^k(s,a)}.$$

Analogous to the previous setting, we claim that Line 6 is equivalent to a model-based Bellman update for those visited (s,a)s. Fix an (s,a,k,h) such that $N_h^k(s,a) \geq 1$. We have shown that for any $\nu \in \mathscr{D}^{\mathcal{S}}$,

$$\left[\hat{P}_h^k \nu\right](s,a) = \frac{1}{N_h^k(s,a)} \sum_{\tau\in[k-1]} \mathbb{I}_h^\tau(s,a)\nu(s_{h+1}^\tau).$$

Hence the update formula in Line 6 of Algorithm 4 can be rewritten as

$$\eta_h^k(s,a) = \left[\hat{P}_h^k \nu_h^k\right](s,a) * \hat{\mathcal{R}}_h^k(s,a) = [\mathcal{B}(\hat{P}_h^k, \hat{\mathcal{R}}_h^k)\nu](s,a).$$

Alternatively, the unvisited (s,a) remains to be the return distribution corresponding to the highest possible reward $H+1-h$. The second difference is that the optimism constant $c_h^k(s,a)$ is increased by an amount of $\sqrt{\frac{1}{2N_h^k(s,a)\vee 1}\iota}$, which corresponds to the estimation error arisen from the unknown reward distribution. The additional term is a lower order term, implying that the regret upper bound of Algorithm 4 is in the same order as that of Algorithm 1.

Algorithm 4 `RODI-MF` (for the random reward)

Input: T and δ
Initialize $N_h(\cdot,\cdot) \leftarrow 0$; $\eta_h(\cdot,\cdot), \nu_h(\cdot) \leftarrow \psi_{H+1-h}$ for all $h \in [H]$
for $k = 1 : K$ **do**
  for $h = H : 1$ **do**
    if $N_h(\cdot,\cdot) > 0$ **then**
      $\eta_h(\cdot,\cdot) \leftarrow \frac{1}{(N_h(\cdot,\cdot))^2} \sum_{\tau,\tau' \in [k-1]^2} \mathbb{I}_h^{\tau}(\cdot,\cdot)\mathbb{I}_h^{\tau'}(\cdot,\cdot)\nu_{h+1}(s_{h+1}^{\tau})(\cdot - R_h^{\tau'}(\cdot,\cdot))$
    end if
    $c_h(\cdot,\cdot) \leftarrow \sqrt{\frac{2S}{N_h(\cdot,\cdot)\vee 1}\iota} + \sqrt{\frac{1}{2N_h(\cdot,\cdot)\vee 1}\iota}$
    $\eta_h(\cdot,\cdot) \leftarrow \mathrm{O}^{\infty}_{c_h(\cdot,\cdot)}\eta_h(\cdot,\cdot)$
    $\pi_h(\cdot) \leftarrow \arg\max_a U_\beta(\eta_h(\cdot,a))$
    $\nu_h(\cdot) \leftarrow \eta_h(\cdot,\pi_h(\cdot))$
  end for
  Receive s_1^k
  for $h = 1 : H$ **do**
    $a_h^k \leftarrow \pi_h(s_h^k)$ and transit to s_{h+1}^k
    $N_h(s_h^k,a_h^k) \leftarrow N_h(s_h^k,a_h^k) + 1$
  end for
end for

C.2 RODI-MB

We provide a model-based algorithm (Algorithm 5), which is equivalent to a *nearly classical* algorithm (Algorithm 5). We emphasize the difference between Algorithm 5 and Algorithm 2. For each (s,a), it applies the distributional optimism operators $\mathrm{O}^1_{c_{h,1}^k(s,a)}$ and $\mathrm{O}^\infty_{c_{h,2}^k(s,a)}$ to the empirical transition model $\hat{P}_h^k(s,a)$ and the empirical reward distribution $\hat{\mathcal{R}}_h^k(s,a)$ respectively, in which $c_{h,1}^k(s,a)$ and $c_{h,2}^k(s,a)$ are set to be $\sqrt{\frac{2S}{N_h^k(s,a)\vee 1}\iota}$ and $\sqrt{\frac{1}{2N_h^k(s,a)\vee 1}\iota}$. Note that the $c_{h,2}^k(s,a)$ is a lower order term in comparison to $c_{h,1}^k(s,a)$, implying that the regret upper bound of Algorithm 5 is in the same order as that of Algorithm 2.

Remark 1. *Algorithm 5 is not a fully classical algorithm because it explicitly maintains the reward distributions for all state-action pairs. However, it does not involve the distributional Bellman update that takes the return distributions for all states as input and outputs the return distributions for all state-action pairs. Hence it still reduces considerable computation complexity and space complexity, which makes more close to the classical algorithm rather than the distributional algorithm.*

Equivalence to `ROVI` Define the exponential value functions $W_h(s) \triangleq E_\beta(\nu_h(s))$ and $J_h(s,a) \triangleq E_\beta(\eta_h(s,a))$ for all (s,a,h)s. Observe that for two independent r.v.s $X \sim F$ and $Y \sim G$, we have

$$E_\beta(F * g) = E_\beta(X+Y) = E_\beta(X)E_\beta(Y),$$

where g is the PDF of G. Applying EERM to Equation 2 yields the exponential Bellman equation

$$\begin{aligned} &J_h^*(s,a) = E_\beta(R_h(s,a))[P_h W_{h+1}^*](s,a), \\ &W_h^*(s) = \mathrm{sign}(\beta)\max_a \mathrm{sign}(\beta)J_h^*(s,a),\ W_{H+1}^*(s) = 1. \end{aligned} \tag{5}$$

We will show that J_h^k in Algorithm 6 corresponds to the exponential value function of η_h^k in Algorithm 5. Observe that

$$\begin{aligned} E_\beta(\eta_h^k(s,a)) &= E_\beta\left(\left[\tilde{P}_h^k\nu_{h+1}^k\right](s,a) * \tilde{\mathcal{R}}_h^k(s,a)\right) = E_\beta(\tilde{\mathcal{R}}_h^k(s,a)) \cdot \left[\tilde{P}_h^k E_\beta(\nu_{h+1}^k)\right](s,a) \\ &= E_\beta(\tilde{\mathcal{R}}_h^k(s,a))\left[\tilde{P}_h^k W_{h+1}^k\right](s,a) = J_h^k(s,a). \end{aligned}$$

The two algorithms generate the policy sequence in the same way. The formal statement is given in Appendix E.

Algorithm 5 RODI-MB

Input: T and δ

$N_h^1(\cdot,\cdot) \leftarrow 0$; $(\hat{P}_h^1(\cdot,\cdot), \hat{\mathcal{R}}_h^1(\cdot,\cdot)) \leftarrow (\frac{1}{S}\mathbf{1}, \psi_{\frac{1}{2}})$ for all $h \in [H]$

for $k = 1 : K$ **do**

  $\nu_{H+1}^k(\cdot) \leftarrow \psi_0$

  for $h = H : 1$ **do**

    $\tilde{P}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^1_{c_{h,1}^k(\cdot,\cdot)} \hat{P}_h^k(\cdot,\cdot)$

    $\tilde{\mathcal{R}}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^\infty_{c_{h,2}^k(\cdot,\cdot)} \hat{\mathcal{R}}_h^k(\cdot,\cdot)$

    $\eta_h^k(\cdot,\cdot) \leftarrow [\mathcal{B}(\tilde{P}_h^k, \tilde{\mathcal{R}}_h^k)\nu_{h+1}^k](\cdot,\cdot)$

    $\pi_h^k(\cdot) \leftarrow \arg\max_a E_\beta(\eta_h^k(\cdot,a))$

    $\nu_h^k(\cdot) \leftarrow \eta_h^k(\cdot, \pi_h^k(\cdot))$

  end for

  Receive s_1^k

  for $h = 1 : H$ **do**

    $a_h^k \leftarrow \pi_h^k(s_h^k)$ and transit to s_{h+1}^k

    Compute $N_h^{k+1}(\cdot,\cdot)$, $\hat{P}_h^{k+1}(\cdot,\cdot)$ and $\hat{\mathcal{R}}_h^{k+1}(\cdot,\cdot)$

  end for

end for

Algorithm 6 ROVI

Input: T and δ

$N_h^1(\cdot,\cdot) \leftarrow 0$; $(\hat{P}_h^1(\cdot,\cdot), \hat{\mathcal{R}}_h^1(\cdot,\cdot)) \leftarrow (\frac{1}{S}\mathbf{1}, \psi_{\frac{1}{2}})$ for all $h \in [H]$

for $k = 1 : K$ **do**

  $W_{H+1}^k(\cdot) \leftarrow 1$

  for $h = H : 1$ **do**

    $\tilde{P}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^1_{c_{h,1}^k(\cdot,\cdot)} \hat{P}_h^k(\cdot,\cdot)$

    $\tilde{\mathcal{R}}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^\infty_{c_{h,2}^k(\cdot,\cdot)} \hat{\mathcal{R}}_h^k(\cdot,\cdot)$

    $J_h^k(\cdot,\cdot) \leftarrow E_\beta\left(\tilde{\mathcal{R}}_h^k(\cdot,\cdot)\right)\left[\tilde{P}_h^k W_{h+1}^k\right](\cdot,\cdot)$

    $W_h^k(\cdot) \leftarrow \max_a J_h^k(\cdot,a)$

  end for

  Receive s_1^k

  for $h = 1 : H$ **do**

    $a_h^k \leftarrow \arg\max_a J_h^k(s_h^k, a)$ and transit to s_{h+1}^k

    Compute $N_h^{k+1}(\cdot,\cdot)$ and $\hat{P}_h^{k+1}(\cdot,\cdot)$

  end for

end for

D Proof of Regret Bounds

D.1 Proof of Theorem 1

We only prove the case that the reward is random and $\beta > 0$. The proof can be readily adapted to other cases.

Step 1: Verify optimism. Denote by $\iota = \log(2SAT/\delta)$. For any $\delta \in (0,1)$, we define the good event as

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

under which the empirical distributions concentrates around the true distributions w.r.t. $\|\cdot\|_1$.

Lemma 6 (High probability good event)**.** *For any $\delta \in (0,1)$, the event $\mathcal{G}_\delta$ is true with probability at least $1-\delta$.*

Fact 1. *Let X be a random variable taking values over positive integers and E be an event. If $\mathbb{P}(E|X=i) \ge p$ for any $i = 1, 2, ...$, then $\mathbb{P}(E|X>0) \ge p$.*

Proof. $\mathbb{P}(E|X>0) = \frac{\mathbb{P}(E, X>0)}{\mathbb{P}(X>0)} = \frac{\sum_{i\ge 1}\mathbb{P}(E|X=i)\mathbb{P}(X=i)}{\sum_{i\ge1}\mathbb{P}(X=i)} \ge \frac{\sum_{i\ge1} p\mathbb{P}(X=i)}{\sum_{i\ge1}\mathbb{P}(X=i)} = p$. □

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 $(\hat{P}_h^k(\cdot|s,a), \hat{\mathcal{R}}_h^k(s,a)) = (\frac{1}{S}\mathbf{1}, \psi_{\frac{1}{2}})$. A simple calculation yields that for any $\mathcal{R}_h(s,a) \in \mathscr{D}([0,1])$ and any $P_h(\cdot|s,a)$

$$\left\|\psi_{\frac{1}{2}} - \mathcal{R}_h(s,a)\right\|_\infty \le \frac{1}{2} \le \sqrt{\frac{1}{2}\log(2SAT/\delta)}, \left\|\frac{1}{S}\mathbf{1} - P_h(\cdot|s,a)\right\|_1 \le 2 \le \sqrt{2S\log(2SAT/\delta)}.$$

It follows that

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

Thus the the event is true for the unseen state-action pairs. Now we consider the case that $N_h^k(s,a)>0$. By the DKW inequality, ℓ_1 concentration bound of empirical measure and a union bound, we have that for any $n\ge 1$

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

We use Fact 1 to get

$$\mathbb{P}\left(\left\|\hat{\mathcal{R}}_h^k(s,a)-\mathcal{R}_h(s,a)\right\|_\infty \le \sqrt{\frac{1}{2N_h^k(s,a)}\log(2/\delta)}, \left\|\hat{P}_h^k(\cdot|s,a)-P_h(\cdot|s,a)\right\|_1 \le \sqrt{\frac{2S}{N_h^k(s,a)}\log(2/\delta)}\,\middle|\, N_h^k(s,a)>0\right)\ge 1-\delta.$$

Taking the two cases into consideration

$$\begin{aligned}
&\mathbb{P}\left(\left\|\hat{\mathcal{R}}_h^k(s,a)-\mathcal{R}_h(s,a)\right\|_\infty \le \sqrt{\frac{\log(2/\delta)}{2N_h^k(s,a)}}, \left\|\hat{P}_h^k(\cdot|s,a)-P_h(\cdot|s,a)\right\|_1 \le \sqrt{\frac{2S\log(2/\delta)}{N_h^k(s,a)}}\right)\\
&=\mathbb{P}\left(\left\|\hat{\mathcal{R}}_h^k(s,a)-\mathcal{R}_h(s,a)\right\|_\infty \le \sqrt{\frac{\log(2/\delta)}{2(N_h^k(s,a)\vee 1)}}, \left\|\hat{P}_h^k(\cdot|s,a)-P_h(\cdot|s,a)\right\|_1\right.\\
&\left.\le \sqrt{\frac{2S\log(2/\delta)}{N_h^k(s,a)\vee 1}}\,\middle|\, N_h^k(s,a)=0\right)\mathbb{P}(N_h^k(s,a)=0)\\
&+\mathbb{P}\left(\left\|\hat{\mathcal{R}}_h^k(s,a)-\mathcal{R}_h(s,a)\right\|_\infty \le \sqrt{\frac{\log(2/\delta)}{2N_h^k(s,a)}}, \left\|\hat{P}_h^k(\cdot|s,a)-P_h(\cdot|s,a)\right\|_1 \le \sqrt{\frac{2S\log(2/\delta)}{N_h^k(s,a)}}\right.\\
&\left.|N_h^k(s,a)>0\right)\mathbb{P}(N_h^k(s,a)>0)\\
&\ge \mathbb{P}\left(N_h^k(s,a)=0)+(1-\delta)\mathbb{P}(N_h^k(s,a)>0\right)\ge 1-\delta.
\end{aligned}$$

Applying a union bound over all $(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$ and rescaling δ leads to the result. □

Lemma 6 suggests that $\mathcal{G}_\delta$ holds with probability $1-\delta$, therefore it suffices to prove the theorem conditioned on $\mathcal{G}_\delta$.

Lemma 7. *Let* T *be a functional (not necessarily a risk measure) satisfying the monotonicity, i.e.,* $\mathrm{T}(F)\le\mathrm{T}(G)$ *for any* $F\preceq G$*. For any* $G\in\mathscr{D}([a,b])$*, it holds that if* $G\in B_\infty(F,c)$*, then* $G\preceq \mathrm{O}_c^\infty F$*. Moreover, it holds that*

$$\mathrm{O}_c^\infty F\in\arg\max_{G\in B_\infty(F,c)\cap\mathscr{D}([a,b])}\mathrm{T}(G).$$

Proof. Let $G\in\mathscr{D}([a,b])\cap B_\infty(F,c)$. It follows from the definition of $B_\infty(F,c)$ that $\sup_{x\in[a,b]}|F(x)-G(x)|\le c$, therefore for any $x\in[a,b]$, $G(x)\ge\max(F(x)-c,0)=(\mathrm{O}_c^\infty F)(x)$. The monotonicity of T leads to the result. □

Notice that E_β is also monotonic, which will be used to establish the optimism of the EERM value sequence generated by the algorithm.

Lemma 8. *For any two distributions $F, G \in \mathscr{D}_M$ and any function $u : \mathbb{R} \to \mathbb{R}$, we have that*

$$|\mathbb{E}_F[u(X)] - \mathbb{E}_G[u(X)]| \le |u(M) - u(0)| \|F - G\|_\infty.$$

Proof. Observe that

$$\begin{aligned}
|\mathbb{E}_F[u(X)] - \mathbb{E}_G[u(X)]| &= \left| \int_0^M u(x) dF(x) - \int_0^M u(x) dG(x) \right| \\
&= \left| u(x)F(x)|_0^M - \int_0^M F(x) du(x) - u(x)G(x)|_0^M + \int_0^M G(x) du(x) \right| \\
&= \left| \int_0^M G(x) - F(x) du(x) \right| \\
&\le \left| \int_0^M du(x) \right| \|F - G\|_\infty = |u(M) - u(0)| \, \|F - G\|_\infty .
\end{aligned}$$

□

Lemma 9 (Bound on the optimistic constant)**.** *For any bounded distributions $\{F_i\}_{i\in[n]}$, any $G, G' \in \mathscr{D}([0,1])$ and any $\theta, \theta' \in \Delta_n$ it holds that if $c \ge \|\theta - \theta'\|_1 + \|G - G'\|_\infty$, then*

$$g * \sum_{i=1}^n \theta_i F_i \preceq \mathrm{O}_c^\infty \left(g' * \sum_{i=1}^n \theta_i' F_i \right),$$

where g and g' are the PDF of G and G' resp..

Proof. Without loss of generality assume $F \in \mathscr{D}_M^n$. For any $x \in [0, M+1)$,

$$\begin{aligned}
&\mathrm{O}_c^\infty \left(g' * \sum_{i=1}^n \theta_i' F_i \right)(x) = \left[\sum_{i=1}^n \theta_i' \int_0^1 F_i(x-r) g'(r) dr - c \right]^+ \\
&= \left[\sum_{i=1}^n \theta_i \int_0^1 F_i(x-r) g(r) dr + \sum_{i=1}^n \theta_i' \int_0^1 F_i(x-r) g'(r) dr - \sum \theta_i \int_0^1 F_i(x-r) g(r) dr - c \right]^+ \\
&= \left[\left(g * \sum_{i=1}^n \theta_i F_i \right)(x) + \sum_{i=1}^n \theta_i' \int_0^1 F_i(x-r) g'(r) dr - \sum_{i=1}^n \theta_i \int_0^1 F_i(x-r) g(r) dr - c \right]^+ .
\end{aligned}$$

It suffices to prove

$$c \ge \left| \sum_{i=1}^n \theta_i' \int_0^1 F_i(x-r) g'(r) dr - \sum_{i=1}^n \theta_i \int_0^1 F_i(x-r) g(r) dr \right|, \forall x \in [0, M+1].$$

We have $\forall x \in [0, M+1]$,

$$\begin{aligned}
\text{RHS} \le &\left| \sum_{i=1}^n \theta_i' \int_0^1 F_i(x-r) g'(r) dr - \sum_{i=1}^n \theta_i \int_0^1 F_i(x-r) g'(r) dr \right| \\
&+ \left| \sum_{i=1}^n \theta_i \int_0^1 F_i(x-r) g'(r) dr - \sum_{i=1}^n \theta_i \int_0^1 F_i(x-r) g(r) dr \right| \\
&\le \sum_{i=1}^n |\theta_i' - \theta_i| \int_0^1 F_i(x-r) g'(r) dr + \sum_{i=1}^n \theta_i \left| \int_0^1 F_i(x-r) g'(r) dr - \int_0^1 F_i(x-r) g(r) dr \right| \\
&\le \|\theta' - \theta\|_1 + \|G - G'\|_\infty ,
\end{aligned}$$

where the last inequality follows from that $\int_0^1 F_i(x-r)g'(r)dr \le \int_0^1 g'(r)dr = 1$ and the fact that

$$\left|\int_0^1 F_i(x-r)g'(r)dr - \int_0^1 F_i(x-r)g(r)dr\right| = |\mathbb{E}_G[F_i(x-R)] - \mathbb{E}_{G'}[F_i(x-R)| \le \|G-G'\|_\infty$$

due to Lemma 8. □

We define the EERM value produced by the algorithm as $W_h^k(s) \triangleq E_\beta(\nu_h^k(s))$ and $J_h^k(s,a) \triangleq E_\beta(\eta_h^k(s,a))$ for all (s,a,k,h)s. Similarly, we define $W_h^*(s) \triangleq E_\beta(\nu_h^*(s))$ and $J_h^*(s,a) \triangleq E_\beta(\eta_h^*(s,a))$ for all (s,a,h)s. Using Lemma 9, the monotonicity of EERM, and inductions, we arrives at Lemma 10, which guarantees the sequence $\{W_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 4 is indeed optimistic compared to the optimal value $\{W_1^*(s_1^k)\}_{k\in[K]}$.

Lemma 10 (Optimism)**.** *Conditioned on event $\mathcal{G}_\delta$, the sequence $\{W_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 4 are all greater than or equal to $W_1^*(s_1^k)$, i.e.,*

$$W_1^k(s_1^k) = E_\beta(\nu_1^k(s_1^k)) \ge E_\beta(\nu_1^*(s_1^k)) = W_1^*(s_1^k), \forall k \in [K].$$

Proof. The proof follows from induction. Fix $k \in [K]$. For $h = H$ we have that for any (s,a)

$$\begin{aligned} J_H^k(s,a) = E_\beta(\eta_H^k(s,a)) &= E_\beta(\mathrm{O}^\infty_{c_H^k(s,a)}(\hat{\mathcal{R}}_H^k(s,a))) \\ &\ge E_\beta(\mathcal{R}_H(s,a)) = J_H^*(s,a), \end{aligned}$$

where the inequality is due to Lemma 7 and the fact that $\mathcal{R}_H(s,a) \in B_\infty(\hat{\mathcal{R}}_H(s,a), c_H^k(s,a)) \cap \mathscr{D}_1$. Thus $W_H^k(s) = \max_a J_H^k(s,a) \ge \max_a J_H^*(s,a) = W_H^*(s), \forall s$. Now suppose for $h+1 \in [2:H]$, it holds that $W_{h+1}^k(s) \ge W_{h+1}^*(s), \forall s$. For each (s,a), we applying Lemma **??** with $\theta = P_h(s,a), \theta' = \hat{P}_h^k(s,a), F = \nu_{h+1}^k, G = \mathcal{R}_h(s,a)$ and $G' = \hat{\mathcal{R}}_h^k(s,a)$ to obtain

$$[P_h\nu_{h+1}^k](s,a) * f_{\mathcal{R}_h(s,a)} \preceq \mathrm{O}^\infty_{c_h^k(s,a)}([\hat{P}_h^k\nu_{h+1}^k](s,a) * f_{\hat{\mathcal{R}}_h^k(s,a)})$$

since $c_h^k(s,a) = \sqrt{\frac{2S}{N_h^k(s,a)\vee 1}\iota} + \sqrt{\frac{1}{2(N_h^k(s,a)\vee 1)}\iota} \ge \left\|P_h(\cdot|s,a) - \hat{P}_h^k(\cdot|s,a)\right\|_1 + \left\|\mathcal{R}_h(s,a) - \hat{\mathcal{R}}_h^k(s,a)\right\|_\infty$ for $h \in [H-1]$. It follows that

$$\begin{aligned} J_h^k(s,a) &= E_\beta(\mathrm{O}^\infty_{c_h^k(s,a)}([\hat{P}_h^k\nu_{h+1}^k](s,a) * f_{\hat{\mathcal{R}}_h^k(s,a)})) \\ &\ge E_\beta([P_h\nu_{h+1}^k](s,a) * f_{\mathcal{R}_h(s,a)}) \\ &= E_\beta(\mathcal{R}_h(s,a)) \cdot [P_h W_{h+1}^k](s,a) \\ &\ge E_\beta(\mathcal{R}_h(s,a)) \cdot [P_h W_{h+1}^*](s,a) \\ &= J_h^*(s,a), \forall (s,a), \end{aligned}$$

where the first inequality is due to the property (**M**), and the second inequality follows from the induction assumption. The second equality is due to Equation **??**. Finally it follows that for any s,

$$W_h^k(s) = \max_a J_h^k(s,a) \ge \max_a J_h^*(s,a) = W_h^*(s).$$

The induction is completed. □

Step 2: Regret decomposition.

Lemma 11. *For any $F_i \in \mathscr{D}$ and any $\theta, \theta' \in \Delta_n$ with any $n \ge 2$, it holds that*

$$\left\|\sum_{i=1}^n \theta_i F_i - \sum_{i=1}^n \theta_i' F_i\right\|_\infty \le \|\theta - \theta'\|_1 .$$

Proof.

$$\begin{aligned}
\left\|\sum_{i=1}^n \theta_i F_i - \sum_{i=1}^n \theta_i' F_i\right\|_\infty &= \sup_{x\in\mathbb{R}} \left|\sum_{i=1}^n (\theta_i F - \theta_i') F_i(x)\right| \\
&\le \sup_{x\in\mathbb{R}} \sum_{i=1}^n |\theta_i - \theta_i'| F_i(x) \\
&\le \sum_{i=1}^n |\theta_i - \theta_i'| \\
&= \|\theta - \theta'\|_1 .
\end{aligned}$$

□

We define $\Delta_h^k \triangleq W_h^k - W_h^{\pi^k} = E_\beta(\nu_h^k) - E_\beta\left(\nu_h^{\pi^k}\right) \in D_h^S$ with

$$D_h \triangleq [1 - \exp(\beta(H+1-h)), \exp(\beta(H+1-h)) - 1]$$

and $\delta_h^k \triangleq \Delta_h^k(s_h^k)$. For any (s,h) and any π, we let $P_h^\pi(\cdot|s) := P_h(\cdot|s, \pi_h(s))$. Observe that the regret can be bounded as

$$\begin{aligned}
\mathrm{Regret}(K) &= \sum_{k=1}^K \frac{1}{\beta}\log\left(W_1^*(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right) \\
&= \sum_{k=1}^K \frac{1}{\beta}\log\left(W_1^*(s_1^k)\right) - \frac{1}{\beta}\log\left(V_1^k(s_1^k)\right) + \frac{1}{\beta}\log\left(W_1^k(s_1^k)\right) - \frac{1}{\beta}\log\left(V_1^{\pi^k}(s_1^k)\right) \\
&\le \sum_{k=1}^K \frac{1}{\beta}\log\left(W_1^k(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right) \\
&\le \frac{1}{\beta}\sum_{k=1}^K W_1^k(s_1^k) - W_1^{\pi^k}(s_1^k) = \frac{1}{\beta}\sum_{k=1}^K \delta_1^k.
\end{aligned}$$

We can decompose δ_h^k as follows

$$\begin{aligned}
\delta_h^k &= E_\beta\left(\nu_h^k(s_h^k)\right) - E_\beta\left(\nu_h^{\pi^k}(s_h^k)\right) \\
&= E_\beta\left(O_{c_h^k}\left(\left[\hat{P}_h^{\pi^k}\eta_{h+1}^k\right](s_h^k) * f_{\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)}\right)\right) - E_\beta\left(\left[P_h^{\pi^k}\nu_{h+1}^{\pi^k}\right](s_h^k) * f_{\mathcal{R}_h^{\pi^k}(s_h^k)}\right) \\
&= \underbrace{E_\beta\left(O_{c_h^k}\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)}\right)\right) - E_\beta\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)}\right)}_{(a)} \\
&\quad + \underbrace{E_\beta\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)}\right) - E_\beta\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\mathcal{R}_h^{\pi^k}(s_h^k)}\right)}_{(b)} \\
&\quad + \underbrace{E_\beta\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\mathcal{R}_h^{\pi^k}(s_h^k)}\right) - E_\beta\left(\left[P_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\mathcal{R}_h^{\pi^k}(s_h^k)}\right)}_{(c)} \\
&\quad + \underbrace{E_\beta\left(\left[P_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\mathcal{R}_h^{\pi^k}(s_h^k)}\right) - E_\beta\left(\left[P_h^{\pi^k}\nu_{h+1}^{\pi^k}\right](s_h^k) * f_{\mathcal{R}_h^{\pi^k}(s_h^k)}\right)}_{(d)}.
\end{aligned}$$

Using the Lipschitz property of EERM,

$$\begin{aligned}
(a) &\le L_{H+1-h}\left\|\mathrm{O}_{c_h^k}^\infty\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)}\right) - \left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) * f_{\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)}\right\|_\infty \\
&\le L_{H+1-h} c_h^k \\
&= (\exp(\beta(H+1-h)) - 1)\left(\sqrt{\frac{1}{2(N_h^k \vee 1)}\iota} + \sqrt{\frac{S}{(N_h^k \vee 1)}\iota}\right).
\end{aligned}$$

Define $e_h^k \triangleq \left\|\hat{P}_h^k(s_h^k) - P_h^{\pi^k}(s_h^k)\right\|_1$. We can bound (b) as

$$\begin{aligned}
(b) &= \left(E_\beta(\hat{\mathcal{R}}_h^{\pi^k}(s_h^k)) - E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))\right) \cdot E_\beta\left(\left[\hat{P}_h^k \nu_{h+1}^k\right](s_h^k)\right) \\
&\leq L_1 \left\|\hat{\mathcal{R}}_h^{\pi^k}(s_h^k) - \mathcal{R}_h^{\pi^k}(s_h^k))\right\|_\infty \left[\hat{P}_h^k W_{h+1}^k\right](s_h^k) \\
&\leq (\exp(\beta) - 1)\sqrt{\frac{1}{2(N_h^k \vee 1)}\iota} \exp(\beta(H-h)) \\
&\leq (\exp(\beta(H+1-h)) - 1)\sqrt{\frac{1}{2(N_h^k \vee 1)}\iota}.
\end{aligned}$$

We bound (c) as

$$\begin{aligned}
(c) &= E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))\left(E_\beta\left(\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k)\right) - E_\beta\left(\left[P_h^{\pi^k}\nu_{h+1}^k\right](s_h^k)\right)\right) \\
&\leq \exp(\beta) L_{H-h}\left\|\left[\hat{P}_h^{\pi^k}\nu_{h+1}^k\right](s_h^k) - \left[P_h^{\pi^k}\nu_{h+1}^k\right](s_h^k)\right\|_\infty \\
&\leq \exp(\beta)(\exp(\beta(H-h)) - 1)\left\|\hat{P}_h^{\pi^k}(s_h^k) - P_h^{\pi^k}(s_h^k)\right\|_1 \\
&\leq (\exp(\beta(H+1-h)) - 1)\sqrt{\frac{S}{(N_h^k \vee 1)}\iota},
\end{aligned}$$

where the second inequality is due to Lemma 11. By the linearity of EERM, We bound (d) as

$$\begin{aligned}
(d) &= E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))\left[P_h^{\pi^k}(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_h^k) \\
&= E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))\left[P_h^{\pi^k}\Delta_{h+1}^k\right](s_h^k) \\
&= E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))(\epsilon_h^k + \delta_{h+1}^k),
\end{aligned}$$

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 $\epsilon_h^k \in 2D_{h+1}$ a.s. for all $(k,h) \in [K] \times [H]$. Since

$$(b) + (c) \leq L_{H+1-h} c_h^k,$$

we can bound δ_h^k recursively as

$$\delta_h^k \leq 2L_{H+1-h}c_h^k + E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))(\epsilon_h^k + \delta_{h+1}^k).$$

Repeating the procedure, we can get

$$\begin{aligned}
\delta_1^k &\leq 2\sum_{h=1}^{H-1} L_{H+1-h}\prod_{i=1}^{h-1} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \prod_{i=1}^{H-1} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k \\
&\leq 2\sum_{h=1}^{H-1}(\exp(\beta(H+1-h)) - 1)\exp(\beta(h-1))c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \prod_{i=1}^{H-1} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k \\
&\leq 2\sum_{h=1}^{H-1}(\exp(\beta H) - 1)c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \prod_{i=1}^{H-1} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k.
\end{aligned}$$

It follows that

$$\sum_{k=1}^{K}\delta_1^k \leq 2(\exp(\beta(H+1))-1)\sum_{k=1}^{K}\sum_{h=1}^{H-1} c_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H-1}\prod_{i=1}^{h} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \sum_{k=1}^{K}\prod_{i=1}^{H-1} E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k.$$

Step 3: Bound each term. The first term can be bounded as

$$\begin{aligned}2(\exp(\beta(H+1))-1)\sum_{k=1}^{K}\sum_{h=1}^{H-1}c_h^k &= 2(\exp(\beta(H+1))-1)\sum_{h=1}^{H-1}\sum_{k=1}^{K}\left(\sqrt{\frac{1}{2(N_h^k\vee 1)}\iota}+\sqrt{\frac{S}{(N_h^k\vee 1)}\iota}\right)\\ &\le 3(\exp(\beta(H+1))-1)\sum_{h=1}^{H-1}\sqrt{2S^2AK\iota}\\ &= 3(\exp(\beta(H+1))-1)\sqrt{2S^2AK\iota}.\end{aligned}$$

Observe that

$$\begin{aligned}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k \in \exp(\beta h)D_h &= \exp(\beta h)[1-\exp(\beta(H+1-h)),\exp(\beta(H+1-h))-1]\\ &\subseteq [1-\exp(\beta(H+1)),\exp(\beta(H+1))-1],\end{aligned}$$

thus we can bound the second term by Azuma-Hoeffding inequality: with probability at least $1-\delta'$, the following holds

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k \le (\exp(\beta(H+1))-1)\sqrt{2KH\log(1/\delta')}.$$

We have

$$\begin{aligned}\sum_{k=1}^{K}\prod_{i=1}^{H-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k &\le \sum_{k=1}^{K}\exp(\beta(H-1))L_1c_H^k\\ &\le \sum_{k=1}^{K}\exp(\beta(H-1))(\exp(\beta)-1)\left(\sqrt{\frac{1}{2(N_h^k\vee 1)}\iota}+\sqrt{\frac{S}{(N_h^k\vee 1)}\iota}\right)\\ &\le 1.5(\exp(\beta H)-1)\sqrt{2S^2AK\iota}.\end{aligned}$$

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

$$\begin{aligned}\text{Regret}(K) &\le \frac{1}{\beta}\left(4.5(\exp(\beta(H+1))-1)\sqrt{2S^2AK\iota}+(\exp(\beta(H+1))-1)\sqrt{2KH\iota}\right)\\ &= \tilde{\mathcal{O}}\left(\frac{\exp(\beta H)-1}{\beta H}H\sqrt{HS^2AT}\right),\end{aligned}$$

where $\iota \triangleq \log(4SAT/\delta)$.

D.2 Proof of Theorem 2

We only prove the case that the reward is random. The proof can be readily adapted to the deterministic reward case.

Distributional analysis vs non-distributional analysis By Proposition 5, Algorithm 5 is equivalent to Algorithm 6. Since Algorithm 6 is a classical algorithm, it is thus natural to use the classical analysis to derive the regret bounds. That being said, we will show that the distributional analysis yields a tighter bound than the non-distributional analysis. In particular, **the latter one yields a regret bound that explodes as β approaches zero, but our analysis can recover the desired order when reduced to the risk-neutral setting.**

Step 1: Verify optimism. Lemma 6 suggests that $\mathcal{G}_\delta$ holds with probability $1-\delta$, therefore it suffices to prove the theorem conditioned on $\mathcal{G}_\delta$.

Lemma 12 (Optimistic transition model). *Fix (s,a,k,h). For any $P\in B_1(\hat{P}_h^k(s,a),c_{h,1}^k(s,a))$, we have*

$$E_\beta\left(\left[\tilde{P}_h^k\nu_{h+1}^k\right](s,a)\right)\ge E_\beta\left(\left[P\nu_{h+1}^k\right](s,a)\right).$$

Lemma 13 (Optimism)**.** *Conditioned on event $\mathcal{G}_\delta$, the sequence $\{W_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 5 are all greater than or equal to $W_1^*(s_1^k)$, i.e.,*

$$W_1^k(s_1^k) = E_\beta(\nu_1^k(s_1^k)) \geq E_\beta(\nu_1^*(s_1^k)) = W_1^*(s_1^k), \forall k \in [K].$$

Proof. The proof follows from induction. Fix $k \in [K]$. For $h = H$, we have that for any (s,a)

$$\begin{aligned} J_H^k(s,a) &= E_\beta(\eta_H^k(s,a)) = E_\beta(\mathrm{O}^\infty_{c^k_{H,2}(s,a)}(\hat{\mathcal{R}}_H^k(s,a))) \\ &\geq E_\beta(\mathcal{R}_H(s,a)) = J_H^*(s,a), \end{aligned}$$

where the inequality is due to Lemma 7 and the fact that $\mathcal{R}_H(s,a) \in B_\infty(\hat{\mathcal{R}}_H(s,a), c^k_{H,2}(s,a)) \cap \mathscr{D}_1$. Thus $W_H^k(s) = \max_a J_H^k(s,a) \geq \max_a J_H^*(s,a) = W_H^*(s), \forall s$. Now suppose for $h+1 \in [2:H]$, it holds that $W_{h+1}^k(s) \geq W_{h+1}^*(s), \forall s$. It follows that

$$\begin{aligned} J_h^k(s,a) &= E_\beta\left(\tilde{\mathcal{R}}_h^k(s,a)\right) E_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right](s,a)\right) \\ &\geq E_\beta\left(\mathcal{R}_h(s,a)\right) E_\beta\left(\left[P_h \nu_{h+1}^k\right](s,a)\right) \\ &\geq E_\beta\left(\mathcal{R}_h(s,a)\right) E_\beta\left(\left[P_h \nu_{h+1}^*\right](s,a)\right) \\ &= J_h^*(s,a), \forall (s,a), \end{aligned}$$

where the first inequality is due to Lemma 12, and the second inequality follows from the induction assumption. Since for any s,

$$W_h^k(s) = \max_a J_h^k(s,a) \geq \max_a J_h^*(s,a) = W_h^*(s),$$

The induction is completed. □

Step 2: Regret decomposition. We define $\Delta_h^k \triangleq W_h^k - W_h^{\pi^k} = E_\beta(\nu_h^k) - E_\beta\left(\nu_h^{\pi^k}\right) \in D_h^S$ with

$$D_h \triangleq [1 - \exp(\beta(H+1-h)), \exp(\beta(H+1-h)) - 1]$$

and $\delta_h^k \triangleq \Delta_h^k(s_h^k)$. For any (s,h) and any π, we let $P_h^\pi(\cdot|s) \triangleq P_h(\cdot|s, \pi_h(s))$. The regret can be bounded as

$$\begin{aligned} \mathrm{Regret}(K) &= \sum_{k=1}^K \frac{1}{\beta}\log\left(W_1^*(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right) \\ &= \sum_{k=1}^K \frac{1}{\beta}\log\left(W_1^*(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^k(s_1^k)\right) + \frac{1}{\beta}\log\left(W_1^k(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right) \\ &\leq \sum_{k=1}^K \frac{1}{\beta}\log\left(W_1^k(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right) \\ &\leq \frac{1}{\beta}\sum_{k=1}^K W_1^k(s_1^k) - W_1^{\pi^k}(s_1^k) = \frac{1}{\beta}\sum_{k=1}^K \delta_1^k. \end{aligned}$$

We can decompose δ_h^k as follows

$$
\begin{aligned}
\delta_h^k &= E_\beta(\nu_h^k(s_h^k)) - E_\beta(\nu_h^{\pi^k}(s_h^k)) \\
&= E_\beta\left(\tilde{\mathcal{R}}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right](s_h^k)\right) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[P_h^{\pi^k} \nu_{h+1}^{\pi^k}\right](s_h^k)\right) \\
&= \underbrace{E_\beta\left(\tilde{\mathcal{R}}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right](s_h^k)\right) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right](s_h^k)\right)}_{(a)} \\
&+ \underbrace{E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[\tilde{P}_h^k \nu_{h+1}^k\right](s_h^k)\right) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[P_h^{\pi^k} \nu_{h+1}^k\right](s_h^k)\right)}_{(b)} \\
&+ \underbrace{E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[P_h^{\pi^k} \nu_{h+1}^k\right](s_h^k)\right) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right) E_\beta\left(\left[P_h^{\pi^k} \nu_{h+1}^{\pi^k}\right](s_h^k)\right)}_{(c)} \\
&= \underbrace{E_\beta\left(\tilde{\mathcal{R}}_h^{\pi^k}(s_h^k)\right)\left[\tilde{P}_h^k W_{h+1}^k\right](s_h^k) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[\tilde{P}_h^k W_{h+1}^k\right](s_h^k)}_{(a)} \\
&+ \underbrace{E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[\tilde{P}_h^k W_{h+1}^k\right](s_h^k) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[P_h^{\pi^k} W_{h+1}^k\right](s_h^k)}_{(b)} \\
&+ \underbrace{E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[P_h^{\pi^k} W_{h+1}^k\right](s_h^k) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[P_h^{\pi^k} W_{h+1}^{\pi^k}\right](s_h^k)}_{(c)}.
\end{aligned}
$$

Both distributional analysis and non-distributional analysis seem to be viable to deal with (b), but the non-distributional analysis turns out to yield an unsatisfactory bound.

Non-distributional analysis: Notice that $W_{h+1}^k(s) \le \exp(\beta(H-h)),\ \forall s$. Thus the following holds

$$
\begin{aligned}
(b) &= E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left(\left[\tilde{P}_h^k W_{h+1}^k\right](s_h^k) - \left[P_h^{\pi^k} W_{h+1}^k\right](s_h^k)\right) \\
&= E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left(\left[\left(\tilde{P}_h^k - P_h^{\pi^k}\right) W_{h+1}^k\right](s_h^k)\right) \\
&\le \exp(\beta)\left\|\tilde{P}_h^k - P_h^{\pi^k}\right\|_1 \max_s W_{h+1}^k(s) \\
&\le 2\exp(\beta(H+1-h)) c_{h,1}^k.
\end{aligned}
$$

Distributional analysis: Using the Lipschitz property of EERM, we have

$$
\begin{aligned}
(b) &\le L_{H+1-h}\left\|\left[\tilde{P}_h^k \nu_{h+1}^k\right](s_h^k)(\cdot - r_h^k) - \left[P_h^{\pi^k} \nu_{h+1}^k\right](s_h^k)(\cdot - r_h^k)\right\|_\infty \\
&\le L_{H+1-h}\left\|\tilde{P}_h^k - P_h^{\pi^k}\right\|_1 \\
&\le 2L_{H+1-h} c_{h,1}^k \\
&= 2(\exp(\beta(H+1-h)) - 1) c_{h,1}^k,
\end{aligned}
$$

where the second inequality is due to Lemma 11. The two types of analysis lead to different coefficients. Consider the risk-neutral setting $\beta \to 0$. For the distributional analysis, the coefficient appears in the regret bound as

$$
\lim_{\beta \to 0} \frac{\exp(\beta(H+1-h)) - 1}{\beta} = H + 1 - h,
$$

in contrast, the non-distributional analysis leads to that

$$
\lim_{\beta \to 0} \frac{\exp(\beta(H+1-h))}{\beta} = \infty.
$$

For small β, the distributional analysis recovers the order of the corresponding risk-neutral algorithm. However, the non-distributional analysis yields a exploding factor as $\beta \to 0$. Therefore, it is not proper to use the classical analysis to obtain the regret bound of Algorithm 6. We can bound (a) as

$$\begin{aligned}(a) &= \left(E_\beta\left(\tilde{\mathcal{R}}_h^{\pi^k}(s_h^k)\right) - E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\right)\left[\tilde{P}_h^k W_{h+1}^k\right](s_h^k)\\ &\le L_1\left\|\tilde{\mathcal{R}}_h^{\pi^k}(s_h^k) - \mathcal{R}_h^{\pi^k}(s_h^k)\right\|_\infty \cdot \exp(\beta(H-h))\\ &\le (\exp(\beta(H+1-h))-1)c_{h,2}^k,\end{aligned}$$

where the second inequality follows from the DKW inequality and the definition of $c_{h,2}^k$. Term (c) is bounded as

$$\begin{aligned}(c) &= E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[P_h^{\pi^k}(W_{h+1}^k - W_{h+1}^{\pi^k})\right](s_h^k)\\ &= E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)\left[P_h^{\pi^k}\Delta_{h+1}^k\right](s_h^k)\\ &= E_\beta\left(\mathcal{R}_h^{\pi^k}(s_h^k)\right)(\epsilon_h^k + \delta_{h+1}^k),\end{aligned}$$

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 $\epsilon_h^k \in 2D_{h+1}$ a.s. for all $(k,h) \in [K]\times[H]$. Denote by $c_h^k \triangleq c_{h,1}^k + c_{h,2}^k$. In summary, we can bound δ_h^k recursively as

$$\delta_h^k \le 2L_{H+1-h}c_h^k + E_\beta(\mathcal{R}_h^{\pi^k}(s_h^k))(\epsilon_h^k + \delta_{h+1}^k).$$

Repeating the procedure, we can get

$$\begin{aligned}\delta_1^k &\le 2\sum_{h=1}^{H-1} L_{H+1-h}\prod_{i=1}^{h-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \prod_{i=1}^{H-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k\\ &\le 2\sum_{h=1}^{H-1}(\exp(\beta(H+1-h))-1)\exp(\beta(h-1))c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \prod_{i=1}^{H-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k\\ &\le 2\sum_{h=1}^{H-1}(\exp(\beta H)-1)c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \prod_{i=1}^{H-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k.\end{aligned}$$

It follows that

$$\sum_{k=1}^{K}\delta_1^k \le 2(\exp(\beta(H+1))-1)\sum_{k=1}^{K}\sum_{h=1}^{H-1}c_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H-1}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k + \sum_{k=1}^{K}\prod_{i=1}^{H-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k.$$

Step 3: Bound each term. The first term can be bounded as

$$\begin{aligned}2(\exp(\beta(H+1))-1)\sum_{k=1}^{K}\sum_{h=1}^{H-1}c_h^k &= 2(\exp(\beta(H+1))-1)\sum_{h=1}^{H-1}\sum_{k=1}^{K}\left(\sqrt{\frac{1}{2(N_h^k\vee 1)}\iota} + \sqrt{\frac{S}{(N_h^k\vee 1)}\iota}\right)\\ &\le 3(\exp(\beta(H+1))-1)\sum_{h=1}^{H-1}\sqrt{2S^2AK\iota}\\ &= 3(\exp(\beta(H+1))-1)\sqrt{2S^2AK\iota}.\end{aligned}$$

Observe that

$$\begin{aligned}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k \in \exp(\beta h)D_h &= \exp(\beta h)[1-\exp(\beta(H+1-h)), \exp(\beta(H+1-h))-1]\\ &\subseteq [1-\exp(\beta(H+1)), \exp(\beta(H+1))-1],\end{aligned}$$

thus we can bound the second term by Azuma-Hoeffding inequality: with probability at least $1-\delta'$, the following holds

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1}\prod_{i=1}^{h}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\epsilon_h^k \le (\exp(\beta(H+1))-1)\sqrt{2KH\log(1/\delta')}.$$

We have

$$\sum_{k=1}^{K}\prod_{i=1}^{H-1}E_\beta(\mathcal{R}_i^{\pi^k}(s_i^k))\delta_H^k \le \sum_{k=1}^{K}\exp(\beta(H-1))L_1 c_H^k$$
$$\le \sum_{k=1}^{K}\exp(\beta(H-1))(\exp(\beta)-1)\left(\sqrt{\frac{1}{2(N_h^k\vee 1)}\iota}+\sqrt{\frac{S}{(N_h^k\vee 1)}\iota}\right)$$
$$\le 1.5(\exp(\beta H)-1)\sqrt{2S^2AK\iota}.$$

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

$$\mathrm{Regret}(K)\le \frac{1}{\beta}\left(4.5(\exp(\beta(H+1))-1)\sqrt{2S^2AK\iota}+(\exp(\beta(H+1))-1)\sqrt{2KH\iota}\right)$$
$$=\tilde{\mathcal{O}}\left(\frac{\exp(\beta H)-1}{\beta H}H\sqrt{HS^2AT}\right),$$

where $\iota \triangleq \log(4SAT/\delta)$.

In contrast, if we use non-distributional analysis, we will arrive at

$$\mathrm{Regret}(K)\le \tilde{\mathcal{O}}\left(\frac{\exp(\beta H)}{\beta}\sqrt{HS^2AT}\right),$$

which blows up as $\beta\to 0$.

D.3 Proof for regret lower bounds

Notations. We define $\mathrm{kl}(p,q) := p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}$ as the KL divergence between two Bernoulli distributions with parameters p and q.

D.3.1 Correction of Lower Bound

[23] presents the following lower bound.

Proposition 3 (Theorem 3,[23]). *For sufficiently large K and H, the regret of any algorithm obeys*

$$\mathbb{E}[\mathrm{Regret}(K)]\gtrsim \frac{e^{|\beta|H/2}-1}{|\beta|}\sqrt{T\log T}.$$

However, the lower bound itself and the proof are incorrect. The major mistake appears at the second inequality of the following statements in their proof

$$\mathbb{E}[\mathrm{Regret}(K)]\gtrsim \frac{\exp(\beta H/2)-1}{\beta}\sqrt{K\log(K)}$$
$$\gtrsim \frac{\exp(\beta H/2)-1}{\beta}\sqrt{KH\log(KH)}.$$

The authors establish the second inequality based on the following fact

Fact 2 (Fact 5,[23]). *For any $\alpha>0$, the function $f_\alpha := \frac{e^{\alpha x}-1}{x}, x>0$ is increasing and satisfies $\lim_{x\to 0} f_\alpha = \alpha$.*

In fact, we can only use Fact 2 to derive $\frac{\exp(\beta H/2)-1}{\beta}\gtrsim H$, which combined with the first inequality yields

$$\mathbb{E}[\mathrm{Regret}(K)]\gtrsim H\sqrt{KH\log(KH)}.$$

It is a weaker lower bound and does not feature the dependence on β. The best result we can get based on the original proof is that

Proposition 4 (Correction of Theorem 3,[23]). *For sufficiently large K and H, the regret of any algorithm obeys*

$$\mathbb{E}[\mathrm{Regret}(K)]\gtrsim \frac{e^{|\beta|H/2}-1}{|\beta|}\sqrt{K\log K}.$$

D.3.2 Proof of Theorem 3

We introduce some notations here. We define the probability measure induced by an algorithm $\mathscr{A}$ and an MDP instance $\mathcal{M}$ as

$$\mathbb{P}_{\mathscr{A}\mathcal{M}}(\mathcal{F}^{K+1}) := \prod_{k=1}^{K} \mathbb{P}_{\mathscr{A}_k(\mathcal{F}^k)\mathcal{M}}(\mathcal{I}_H^k | s_1^k),$$

where $\mathbb{P}_{\pi\mathcal{M}}$ is the probability measure induced by a policy π and $\mathcal{M}$, which is defined as

$$\mathbb{P}_{\pi\mathcal{M}}(\mathcal{I}_H | s_1) := \prod_{h=1}^{H} \pi_h(a_h|s_h) P_h^{\mathcal{M}}(s_{h+1}|s_h, a_h).$$

Note that the probability measure for the truncated history $\mathcal{H}_h^k$ can be obtained by marginalization

$$\mathbb{P}_{\mathscr{A}\mathcal{M}}(\mathcal{H}_h^k) = \mathbb{P}_{\mathscr{A}\mathcal{M}}(\mathcal{F}^k)\mathbb{P}_{\mathscr{A}_k(\mathcal{F}^k)\mathcal{M}}(\mathcal{I}_h^k).$$

We denote by $\mathbb{P}_{\mathscr{A}\mathcal{M}}$ and $\mathbb{E}_{\mathscr{A}\mathcal{M}}$ the probability measure and expectation induced by $\mathscr{A}$ and $\mathcal{M}$. For the sake of simplicity, the dependency on $\mathscr{A}$ and $\mathcal{M}$ may be dropped if it is clear in the context.

Fact 3 (Lemma 1, [26])**.** *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*

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

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

Fact 4 (Lemma 5, [20])**.** *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 $(I_k)_{k\geq1}$ that satisfies $\mathbb{P}_{\mathcal{M}}[\tau < \infty] = 1$*

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

Lemma 14. *If $\epsilon \geq 0$, $p \geq 0$ and $p + \epsilon \in [0, \frac{1}{2}]$, then* $\mathrm{kl}(p, p + \epsilon) \leq \frac{\epsilon^2}{2p(1-p)} \leq \frac{\epsilon^2}{p}$.

Proof. Fix $q \in [0, 1]$, let $h(p) := \mathrm{kl}(p, q)$. It is immediate that

$$h'(p) = \log\frac{p}{q} - \log\frac{1-p}{1-q},$$
$$h''(p) = \frac{1}{p(1-p)} > 0.$$

Therefore $h(p)$ is strictly convex, increasing in $(q, 1)$ and decreasing in $(0, q)$. By Taylor's expansion, we have that

$$h(p) = h(q) + h'(q)(p-q) + \frac{1}{2}h''(r)(p-q)^2 = \frac{(p-q)^2}{2r(1-r)}$$

for some $r \in [p, q]$ $(p < q)$ or $r \in [q, p]$ $(p > q)$. In particular, for any $\epsilon \geq 0$ such that $q = p + \epsilon \leq \frac{1}{2}$ it follows that

$$\mathrm{kl}(p, p+\epsilon) = \frac{(p-q)^2}{2r(1-r)}|_{q=p+\epsilon} = \frac{\epsilon^2}{2r(1-r)} \leq \frac{\epsilon^2}{2p(1-p)} \leq \frac{\epsilon^2}{p},$$

where the first inequality follows from the fact that $r \mapsto r(1-r)$ is increasing in $[p, p+\epsilon] \subset [0, \frac{1}{2}]$ and the second inequality is due to that $1 - p \geq \frac{1}{2}$. □

The proof of Theorem 3 adopts the same construction of hard MDP class $\mathcal{C}$ as [20].

Proof. We consider the case that $\beta > 0$. 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 $S - 3$ 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

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

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

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

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

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

The total expected ERM value of $\mathscr{A}$ is given by

$$\begin{aligned}
&\mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K U_\beta\left(\sum_{h=1}^H r_h(s_h^k, a_h^k)|\pi^k\right)\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \frac{1}{\beta}\log \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\exp\left(\beta \sum_{h=1}^H r_h(s_h^k, a_h^k)\right)\right]\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \frac{1}{\beta}\log \mathbb{E}_{\pi^k,\mathcal{M}_{u^*}}\left[\exp\left(\beta \sum_{h=\tilde{H}}^H \mathbb{I}\{s_h^k = s_g\}\right)\right]\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \frac{1}{\beta}\log \mathbb{E}_{\pi^k,\mathcal{M}_{u^*}}\left[\exp(\beta H' \mathbb{I}\{s_{\tilde{H}}^k = s_g\})\right]\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^K \frac{1}{\beta}\log(\exp(\beta H')\mathbb{P}_{\pi^k,\mathcal{M}_{u^*}}(s_{\tilde{H}}^k = s_g) + \mathbb{P}_{\pi^k,\mathcal{M}_{u^*}}(s_{\tilde{H}}^k = s_b))\right],
\end{aligned}$$

where the second equality follows from the fact that the reward is non-zero only after step $\tilde{H}$, the third equality is due to that the agent gets into absorbing state when $h \geq \tilde{H}$. Define $x_h^k := (s_h^k, a_h^k)$ for each (k,h) and $x^* := (s_{\ell^*}, a^*)$, then it is not hard to obtain that

$$\begin{aligned}
\mathbb{P}_{\pi^k,u^*}\left[s_{\tilde{H}}^k = s_g\right] &= \sum_{h=1+d}^{\bar{H}+d} p\mathbb{P}_{\pi^k,u^*}\left(s_h^k \in \mathcal{L}\right) + \mathbb{I}\left\{h = h^*\right\}\mathbb{P}_{\pi^k,u^*}(x_h^k = x^*)\varepsilon \\
&= p + \epsilon \mathbb{P}_{\pi^k,u^*}(x_{h^*}^k = x^*).
\end{aligned}$$

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 corresponding optimal value in any of the MDPs is $V^{*,\mathcal{M}_{u^*}} = \frac{1}{\beta}\log(\exp(\beta H')(p+\epsilon) + 1 - p - \epsilon)$. Define $p_{u^*}^k := \mathbb{P}_{\pi^k,u^*}(x_{h^*}^k = x^*)$, then the expected regret of $\mathscr{A}$ in $\mathcal{M}_{u^*}$ can be bounded below as

$$\begin{aligned}
&\mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[\text{Regret}(\mathscr{A},\mathcal{M}_{u^*},K)] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^{K} V^{*,\mathcal{M}_{u^*}} - U_\beta\left(\sum_{h=1}^{H} r_h(x_h^k)|\pi^k\right)\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^{K}\frac{1}{\beta}\log\frac{\exp(\beta H')(p+\epsilon)+1-p-\epsilon}{\exp(\beta H')(p+\epsilon p_{u^*}^k)+1-p-\epsilon p_{u^*}^k}\right] \\
&= \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^{K}\frac{1}{\beta}\log\left(1+\frac{\epsilon(1-p_{u^*}^k)(\exp(\beta H')-1)}{\exp(\beta H')(p+\epsilon p_{u^*}^k)+1-p-\epsilon p_{u^*}^k}\right)\right] \\
&\geq \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\sum_{k=1}^{K}\frac{1}{\beta}\log\left(1+\frac{\epsilon(1-p_{u^*}^k)(\exp(\beta H')-1)}{1+1}\right)\right] \\
&\geq \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}\left[\frac{\exp(\beta H')-1}{4\beta}\epsilon\sum_{k=1}^{K}(1-p_{u^*}^k)\right] \\
&= \frac{\exp(\beta H')-1}{4\beta}\epsilon\sum_{k=1}^{K}(1-\mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[p_{u^*}^k]) \\
&= \frac{\exp(\beta H')-1}{4\beta}K\epsilon\left(1-\frac{1}{K}\mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[N_K(u^*)]\right).
\end{aligned}$$

The first inequality holds by setting $p+\epsilon \leq \exp(-\beta H')$. The second inequality holds by letting $\epsilon \leq 2\exp(-\beta H')$ since $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$. The last equality follows from the fact that

$$\mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[p_{u^*}^k] = \mathbb{E}_{\mathscr{A},\mathcal{M}_{u^*}}[\mathbb{P}_{\pi^k,u^*}(x_{h^*}^k = x^*)] = \mathbb{P}_{\mathscr{A},u^*}(x_{h^*}^k = x^*) = \mathbb{E}_{\mathscr{A},u^*}[\mathbb{I}\{(x_{h^*}^k = x^*)\}]$$

and the definition of $N_K(u^*) := \sum_{k=1}^{K}\mathbb{I}\{x_{h^*}^k = x^*\}$.

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

$$\begin{aligned}
\max_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \text{Regret}(\mathscr{A},\mathcal{M}_{u^*},K) &\geq \frac{1}{\bar{H}\bar{L}A}\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} \text{Regret}(\mathscr{A},\mathcal{M}_{u^*},K) \\
&\geq \frac{\exp(\beta H')-1}{4\beta}K\epsilon\left(1-\frac{1}{\bar{L}AK\bar{H}}\sum_{u^* \in [d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}}\mathbb{E}_{u^*}[N_K(u^*)]\right).
\end{aligned}$$

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 $\mathcal{M}_0$.

By applying Fact 3 with $Z = \frac{N_K(u^*)}{K} \in [0,1]$, we have

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

By Pinsker's inequality, it implies that

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

Since $\mathcal{M}_0$ and $\mathcal{M}_{u^*}$ only differs at stage h^* when $(s,a) = x^*$, it follows from Fact 4 that

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

By Lemma 14, 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

$$\max_{u^* \in [d+1:\bar{H}+d] \times \mathcal{L} \times \mathcal{A}} \mathrm{Regret}(\mathscr{A}, \mathcal{M}_{u^*}, K) \geq \frac{\exp(\beta H') - 1}{4\beta} K\epsilon \left(1 - \frac{1}{\bar{L}A\bar{H}} - \frac{\frac{\epsilon}{\sqrt{2p}} \sqrt{\bar{L}AK\bar{H}}}{\bar{L}A\bar{H}} \right).$$

Choosing $\epsilon = \sqrt{\frac{p}{2}}(1 - \frac{1}{\bar{L}A\bar{H}})\sqrt{\frac{\bar{L}A\bar{H}}{K}}$ maximizes the lower bound

$$\max_{u^* \in [d+1:\bar{H}+d] \times \mathcal{L} \times \mathcal{A}} \mathrm{Regret}(\mathscr{A}, \mathcal{M}_{u^*}, K) \geq \frac{\sqrt{p}}{8\sqrt{2}} \frac{\exp(\beta H') - 1}{\beta} \left(1 - \frac{1}{\bar{L}A\bar{H}} \right)^2 \sqrt{\bar{L}AK\bar{H}}.$$

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 $p = \frac{1}{4}\exp(-\beta H')$ and $\epsilon = \sqrt{\frac{p}{2}}(1 - \frac{1}{\bar{L}A\bar{H}})\sqrt{\frac{\bar{L}A\bar{H}}{K}} \leq \frac{1}{2\sqrt{2}} \exp(-\beta H'/2)\sqrt{\frac{\bar{L}A\bar{H}}{K}} \leq \frac{1}{4}\exp(-\beta H')$ if $K \geq 2\exp(\beta H')\bar{L}A\bar{H}$. Such choice of p and ϵ guarantees the assumption of Lemma 14 and that $p + \epsilon \leq \exp(-\beta H')$, $\epsilon \leq 2\exp(-\beta H')$. Finally we use the fact that $\sqrt{\bar{L}AK\bar{H}} \geq \frac{1}{2\sqrt{3}}\sqrt{SAKH}$ to obtain

$$\max_{u^* \in [d+1:\bar{H}+d] \times \mathcal{L} \times \mathcal{A}} \mathrm{Regret}(\mathscr{A}, \mathcal{M}_{u^*}, K) \geq \frac{1}{72\sqrt{6}} \frac{\exp(\beta H/6) - 1}{\beta} \sqrt{SAKH}.$$

□

E Proof for Propositions

For notational simplicity, we write $\pi_{h_1:h_2} = \{\pi_{h_1}, \pi_{h_1+1}, \cdots, \pi_{h_2}\}$ for two positive integers $h_1 < h_2 \leq H$.

Proof of Proposition 1. Notice that there exists some optimal policy for sub-problems at each step, which will be shown in Proposition 2. Suppose that the truncated policy $\pi^*_{h:H}$ is not optimal for this subproblem, then there exists an optimal policy $\tilde{\pi}_{h:H}$ such that

$$\exists \tilde{s}_h \quad \text{occurring with positive probability,} \quad V_h^{\tilde{\pi}_{h:H}}(\tilde{s}_h) > V_h^{\pi^*_{h:H}}(\tilde{s}_h).$$

There exists a state $\tilde{s}_{h-1}$ which occurs with positive probability and $P_{h-1}(\tilde{s}_h | \tilde{s}_{h-1}, \pi^*_{h-1}(\tilde{s}_{h-1})) > 0$ such that

$$\begin{aligned} U_\beta(\nu_{h-1}^{\pi^*_{h-1}, \tilde{\pi}_{h:H}}(\tilde{s}_{h-1})) &= U_\beta(\mathcal{R}_{h-1}(\tilde{s}_{h-1}, \pi^*_{h-1}(\tilde{s}_{h-1}))) + U_\beta \left(\left[P_{h-1}\nu_h^{\tilde{\pi}_{h:H}} \right] (\tilde{s}_{h-1}, \pi^*_{h-1}(\tilde{s}_{h-1})) \right) \\ &> U_\beta(\mathcal{R}_{h-1}(\tilde{s}_{h-1}, \pi^*_{h-1}(\tilde{s}_{h-1}))) + U_\beta \left(\left[P_{h-1}\nu_h^{\pi^*_{h:H}} \right] (\tilde{s}_{h-1}, \pi^*_{h-1}(\tilde{s}_{h-1})) \right) \\ &= U_\beta \left(\nu_{h-1}^{\pi^*_{h-1:H}}(\tilde{s}_{h-1}) \right), \end{aligned}$$

where the inequality is due to the strict monotonicity preserving property of U_β. It follows that $\{\pi^*_{h-1}, \tilde{\pi}_h, , ..., \tilde{\pi}_H\}$ is a strictly better policy than $\{\pi^*_{h-1}, \pi^*_h, , ..., \pi^*_H\}$ for the subproblem from

$h-1$ to H. Suppose for $h'+1 \in [2, h-1]$, $\{\pi^*_{h'+1}, ..., \tilde{\pi}_h, , ..., \tilde{\pi}_H\}$ is a strictly better policy than $\{\pi^*_{h'+1}, ..., \pi^*_h, , ..., \pi^*_H\}$ for the sub-problem from $h'+1$ to H. Similarly we can obtain that $\{\pi^*_{h'}, ..., \tilde{\pi}_h, , ..., \tilde{\pi}_H\}$ is also a strictly better policy than $\{\pi^*_{h'}, ..., \pi^*_h, , ..., \pi^*_H\}$. Repeating the above arguments finally yields that $\{\pi^*_1, \pi^*_2, ..., \tilde{\pi}_h, , ..., \tilde{\pi}_H\}$ is a strictly better policy than $\pi^* = \{\pi^*_1, \pi^*_2, ..., \pi^*_H\}$. This is contradicted to the assumption that $\pi^* = \{\pi^*_1, \pi^*_2, ..., \pi^*_H\}$ is an optimal policy. □

Proof of Proposition 2. Throughout the proof we drop the dependence on $*$ for the ease of notation. The proof follows from induction. Notice that by distributional Bellman equation, $\eta_h(s_h)$ and $V_h(s_h)$ are the return distribution at state s_h at step h following policy $\pi_{h:H}$ and value function respectively. At step H, it is obvious that π_H is the optimal policy that maximizes the ERM value at the final step for each state $s_H \in \mathcal{S}$. Now fix $h \in [H-1]$, assume that $\pi_{h+1:H}$ is the optimal policy for the subproblem

$$V_{h+1}^{\pi_{h+1:H}}(s_{h+1}) = \max_{\pi'_{h+1:H}} V_{h+1}^{\pi'_{h+1:H}}(s_{h+1}), \forall s_{h+1}.$$

In other words,

$$\begin{aligned}U_\beta(\nu_{h+1}(s_{h+1})) = U_\beta(\nu_{h+1}^{\pi_{h+1:H}}(s_{h+1})) &= \max_{\pi'_{h+1:H}} U_\beta(\nu_{h+1}^{\pi'_{h+1:H}}(s_{h+1})) \\ &\geq U_\beta(\nu_{h+1}^{\pi'_{h+1:H}}(s_{h+1})), \forall \pi'_{h+1:H}, \forall s_{h+1}.\end{aligned}$$

It follows that $\forall s_h$,

$$\begin{aligned}V_h(s_h) &= Q_h(s_h, \pi_h(s_h)) = U_\beta(\nu_h^{\pi_{h:H}}(s_h)) = \max_{a_h} U_\beta(\eta_h(s_h, a_h)) \\ &= \max_{a_h}\{U_\beta(\mathcal{R}_h(s_h, a_h)) + U_\beta([P_h\nu_{h+1}](s_h, a_h))\} \\ &\geq \max_{a_h}\left\{U_\beta(\mathcal{R}_h(s_h, a_h)) + \max_{\pi'_{h+1:H}} U_\beta\left(\left[P_h\nu_{h+1}^{\pi'_{h+1:H}}\right](s_h, a_h)\right)\right\} \\ &= \max_{\pi'_h}\left\{U_\beta(\mathcal{R}_h(s_h, \pi'_h(s_h))) + \max_{\pi'_{h+1:H}} U_\beta\left(\left[P_h\nu_{h+1}^{\pi'_{h+1:H}}\right](s_h, a_h)\right)\right\} \\ &= \max_{\pi'_{h:H}}\left\{U_\beta(\mathcal{R}_h(s_h, \pi'_h(s_h))) + U_\beta\left(\left[P_h\nu_{h+1}^{\pi'_{h+1:H}}\right](s_h, \pi'_h(s_h))\right)\right\} \\ &= \max_{\pi'_{h:H}} U_\beta\left(\nu_h^{\pi'_{h+1:H}}(s_h)\right).\end{aligned}$$

Hence V_h is the optimal value function at step h and $\pi_{h:H}$ is the optimal policy for the sub-problem from h to H. The induction is completed. □

Definition 1. *For two algorithms $\mathscr{A}$ and $\tilde{\mathscr{A}}$, we say that $\mathscr{A}$ is equivalent to $\tilde{\mathscr{A}}$ (vice versa) if for any $k \in [K]$, any $\mathcal{F}_k$ it holds that $\mathscr{A}(\mathcal{F}_k) = \tilde{\mathscr{A}}(\mathcal{F}_k)$.*

It follows from the induction that the whole history/trajectory $\mathcal{F}_{K+1}$ generated by the interaction between each of two equivalent algorithms and any MDP instance follows the same distribution. Moreover, the two algorithms possess equal regret.

Proposition 5 (Equivalence between `ROVI` and `RODI-MB`)**.** *Algorithm 5 (Algorithm 2) is equivalent to Algorithm 6 (Algorithm 3).*

Proof. We only prove the case that $\beta > 0$. The case that $\beta < 0$ follows analogously. Fix an arbitrary $k \in [K]$ and $\mathcal{F}_k = \{s^1_1, a^1_1, R^1_1, \cdots, s^{k-1}_H, a^{k-1}_H, R^{k-1}_H\}$. Denote by $\mathscr{A}$ ($\tilde{\mathscr{A}}$) and $\{\pi^k_h\}$ ($\{\tilde{\pi}^k_h\}$) Algorithm 6 (Algorithm 5) and the associated policy sequence. It suffices to prove that π^k coincides with $\tilde{\pi}^k$ for the same history $\mathcal{F}_k$. By the definition of the two algorithms, we have

$$\tilde{\pi}^k_h(s) = \arg\max_a Q^k_h(s, a) = U_\beta(\eta^k_h(s, a)),\ \pi^k_h(s) = \arg\max_a J^k_h(s, a).$$

If $J^k_h(s, a) = E_\beta(\eta^k_h(s, a)) = \exp(\beta Q^k_h(s, a))$ for any (s, a), then $\pi^k_h = \tilde{\pi}^k_h$ due to the monotonicity of the exponential function. We will prove that $J^k_h(s, a) = E_\beta(\eta^k_h(s, a))$ by the induction. Notice that $J_H^k(s,a) = E_\beta(\eta_H^k(s,a))$. Assume that $J_h^k(s,a) = E_\beta(\eta_h^k(s,a))$ for some $h \in [H]$. It follows that $\pi_h^k = \tilde{\pi}_h^k$ and

$$\begin{aligned} W_h^k(s) &= \max_a J_h^k(s,a) = J_h^k(s,\pi_h^k(s)) = E_\beta(\eta_h^k(s,\pi_h^k(s))) = E_\beta(\eta_h^k(s,\tilde{\pi}_h^k(s))) \\ &= E_\beta(\nu_h^k(s)). \end{aligned}$$

Given the same history $\mathcal{F}_k$, the two algorithms share the empirical transition model $\hat{P}_{h-1}^k$, the empirical reward distribution $\hat{\mathcal{R}}_{h-1}^k$, the count N_{h-1}^k, and the optimism constants $c_{h-1,1}^k$, $c_{h-1,2}^k$. Therefore they also share the optimistic transition model $\tilde{P}_{h-1}^k$ as well as the optimistic reward distribution $\tilde{\mathcal{R}}_{h-1}^k$. According to the update formula of Algorithm 6, we have that for any (s,a)

$$\begin{aligned} J_{h-1}^k(s,a) &= E_\beta\left(\tilde{\mathcal{R}}_{h-1}^k(s,a)\right)\left[\tilde{P}_{h-1}^k W_h^k\right](s,a) = E_\beta\left(\tilde{\mathcal{R}}_{h-1}^k(s,a)\right) E_\beta\left(\left[\tilde{P}_{h-1}^k \nu_h^k\right](s,a)\right) \\ &= E_\beta\left(\left[\mathcal{B}(\tilde{P}_{h-1}^k, \tilde{\mathcal{R}}_{h-1}^k)\nu_h^k\right](s,a)\right) \\ &= E_\beta\left(\eta_{h-1}^k(s,a)\right). \end{aligned}$$

Thus the proof for the case of random reward is completed. The proof for the case of deterministic reward follows analogously.

□