Journal of Machine Learning Research 25 (2024) 1-56 Submitted 11/22; Revised 6/24; Published 7/24

Bridging Distributional and Risk-sensitive Reinforcement Learning with Provable Regret Bounds

Hao Liang haoliang1@link.cuhk.edu.cn
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen

Zhi-Quan Luo luozq@cuhk.edu.cn
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen

Editor: Csaba Szepesvari

Abstract

We study the regret guarantee for risk-sensitive reinforcement learning (RSRL) via distributional reinforcement learning (DRL) methods. In particular, we consider finite episodic Markov decision processes whose objective is the entropic risk measure (EntRM) of return. By leveraging a key property of the EntRM, the independence property, we establish the risk-sensitive distributional dynamic programming framework. We then propose two novel DRL algorithms that implement optimism through two different schemes, including a model-free one and a model-based one.

We prove that they both attain $\tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H)-1}{|\beta|}H\sqrt{S^2AK}\right)$ regret upper bound, where S, A, K, H, $T=KH$, and β represent the number of states, actions, episodes, time horizon, number of total time-steps and risk parameter respectively. It matches RSVI2 (Fei et al., 2021), with novel distributional analysis that focuses on the distributions of returns rather than the risk values associated with these returns. To the best of our knowledge, this is the first regret analysis that bridges DRL and RSRL in terms of sample complexity.

To address the computational inefficiencies inherent in the model-free DRL algorithm, we propose an alternative DRL algorithm with distribution representation. This approach effectively represents any bounded distribution using a refined distribution class. It significantly amplifies computational efficiency while maintaining the established regret bounds.

We also prove a tighter minimax lower bound of $\Omega\left(\frac{\exp(\beta H/6)-1}{\beta}\sqrt{SAT}\right)$ for the $\beta>0$ case, which recovers the tight lower bound $\Omega(H\sqrt{SAT})$ in the risk-neutral setting.

Keywords: distributional reinforcement learning, risk-sensitive reinforcement learning, regret bounds, episodic MDP, entropic risk measure

1. Introduction

Standard reinforcement learning (RL) seeks to find an optimal policy that maximizes the expected return (Sutton and Barto, 2018). This approach is often referred to as risk-neutral RL, as it focuses on the mean functional of the return distribution. However, in high-stakes applications, such as finance (Davis and Lleo, 2008; Bielecki et al., 2000), medical treatment

(Ernst et al., 2006), and operations research (Delage and Mannor, 2010), decision-makers are often risk-sensitive and aim to maximize a risk measure of the return distribution.

Since the pioneering work of Howard and Matheson (1972), risk-sensitive reinforcement learning (RSRL) based on the exponential risk measure (EntRM) has been applied to a wide range of domains (Shen et al., 2014; Nass et al., 2019; Hansen and Sargent, 2011). EntRM offers a trade-off between the expected return and its variance and allows for the adjustment of risk-sensitivity through a risk parameter. However, existing approaches typically require complicated algorithmic designs to handle the non-linearity of EntRM.

Distributional reinforcement learning (DRL) has demonstrated superior performance over traditional methods in some challenging tasks under a risk-neutral setting (Bellemare et al., 2017; Dabney et al., 2018b,a). Unlike value-based approaches, DRL learns the entire return distribution instead of a real-valued value function. Given the distributional information, it is natural to leverage it to optimize a risk measure other than expectation (Dabney et al., 2018a; Singh et al., 2020; Ma et al., 2020). Despite the intrinsic connection between DRL and RSRL, existing works on RSRL via DRL approaches lack regret analysis (Dabney et al., 2018a; Ma et al., 2021; Achab and Neu, 2021). Consequently, it is challenging to evaluate and improve these DRL algorithms in terms of sample efficiency. Additionally, DRL can be computationally demanding as return distributions are typically infinite-dimensional objects. This complexity raises a pertinent question:

Is it feasible for DRL to attain near-optimal regret in RSRL while preserving computational efficiency?

In this work, we answer this question positively by providing computationally efficient DRL algorithms with regret guarantee. We have developed two types of DRL algorithms, both designed to be computationally efficient, and equipped with principled exploration strategies tailored for tabular EntRM-MDP. Notably, these proposed algorithms apply the principle of optimism in the face of uncertainty (OFU) at a distributional level, effectively balancing the exploration-exploitation dilemma. By conducting the first regret analysis in the field of DRL, we bridge the gap between computationally efficient DRL and RSRL, especially in terms of sample complexity. Our work paves the way for deeper understanding and improving the efficiency of RSRL through the lens of distributional approaches.

1.1 Related Work

Related work in DRL has rapidly grown since Bellemare et al. (2017), with numerous studies aiming to improve performance in the risk-neutral setting (see Rowland et al., 2018; Dabney et al., 2018b,a; Barth-Maron et al., 2018; Yang et al., 2019; Lyle et al., 2019; Zhang et al., 2021). However, only a few works have considered risk-sensitive behavior, including Dabney et al. (2018a); Ma et al. (2021); Achab and Neu (2021). None of these works have addressed sample complexity.

A large body of work has investigated RSRL using the EntRM in various settings (Borkar, 2001, 2002; Borkar and Meyn, 2002; Borkar, 2010; Bäuerle and Rieder, 2014; Di Masi et al., 2000; Di Masi and Stettner, 2007; Cavazos-Cadena and Hernández-Hernández, 2011; Jaśkiewicz, 2007; Ma et al., 2020; Mihatsch and Neuneier, 2002; Osogami, 2012; Patek, 2001; Shen et al., 2013, 2014). However, these works either assume known transition and reward or consider infinite-horizon settings without considering sample complexity.

Our work is related to two recent studies by Fei et al. (2020) and Fei et al. (2021) in the same setting. Fei et al. (2020) introduced the first regret-guaranteed algorithms for risk-sensitive episodic Markov decision processes (MDPs), but their regret upper bounds contain an unnecessary factor of $\exp(|\beta|H^2)$ and their lower bound proof contains errors, leading to a weaker bound. While Fei et al. (2021) refined their algorithm by introducing a doubly decaying bonus that effectively removes the $\exp(|\beta|H^2)$ factor[1], the issue with the lower bound was not resolved. Achab and Neu (2021) proposed a risk-sensitive deep deterministic policy gradient framework, but their work is fundamentally different from ours as they consider the conditional value at risk and focus on discounted MDP with infinite horizon settings. Moreover, Achab and Neu (2021) assumes that the model is known.

1.2 Contributions

This paper makes the following primary contributions:
1. Formulation of a Risk-Sensitive Distributional Dynamic Programming (RS-DDP) framework. This framework introduces a distributional Bellman optimality equation tailored for EntRM-MDP, leveraging a key property—the independence property[2]—of the EntRM.
2. Proposal of computationally efficient DRL algorithms that enforce the OFU principle in a distributional fashion, along with regret upper bounds of $\tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H)-1}{|\beta|}H\sqrt{S^2AK}\right)$. The DRL algorithms not only outperform existing methods empirically but are also supported by theoretical justifications. Furthermore, this marks the first instance of analyzing a DRL algorithm within a finite episodic EntRM-MDP setting.
3. Filling of gaps in the lower bound in Fei et al. (2020), resulting in a tight lower bound of $\Omega\left(\frac{\exp(\beta H/6)-1}{\beta}\sqrt{SAT}\right)$ for $\beta > 0$. This lower bound is dependent of S and A and recovers the tight lower bound in the risk-neutral setting (as $\beta \to 0$).

Algorithm	Regret bound	Time	Space
RSVI	$\tilde{\mathcal{O}}\left(\exp(\lvert\beta\rvert H^2)\frac{\exp(\lvert\beta\rvert H)-1}{\lvert\beta\rvert}\sqrt{HS^2AT}\right)$	$\mathcal{O}\left(TS^2A\right)$	$\mathcal{O}\left(HSA+T\right)$
RSVI2	$\tilde{\mathcal{O}}\left(\frac{\exp(\lvert\beta\rvert H)-1}{\lvert\beta\rvert}\sqrt{HS^2AT}\right)$		
RODI-Rep			
RODI-MF		$\mathcal{O}(KS^H)$	$\mathcal{O}(S^H)$
RODI-MB		$\mathcal{O}\left(TS^2A\right)$	$\mathcal{O}(HS^2A)$
lower bound	$\Omega\left(\frac{\exp(\beta H/6)-1}{\beta}\sqrt{SAT}\right)$	-	-

Table 1: Regret bounds and computational complexity comparisons.

2. Preliminaries

We provide the technical background in this section.

[1] A detailed comparison with Fei et al. (2021) is given in Section 8.
[2] The independence property will be formally introduced in Section 3.

2.1 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 denote by δ_c the Dirac measure at c. Similarly, ψ_c represents the unit step function at c, corresponding to the CDF of δ_c. We denote by $\Delta(\mathcal{S})$, $\mathscr{D}(a, b)$, $\mathscr{D}_M$ and $\mathscr{D}$ the set of distributions supported on a set $\mathcal{S}$, $[a, b]$, $[0, M]$ and the set of all distributions respectively. We use $(x_1, x_2; p)$ to denote a binary r.v. taking values x_1 and x_2 with probability $1-p$ and p. For a discrete set $x = \{x_1, \cdots, x_n\}$ and a probability vector $p = (p_1, \cdots, p_n)$, the notation (x, p) represents the discrete distribution with $\mathbb{P}(X = x_i) = p_i$. For a discrete distribution $\eta = (x, p)$, we use $|\eta| = |x|$ to denote the number of atoms of the distribution η. We use $\tilde{\mathcal{O}}(\cdot)$ to denote $\mathcal{O}(\cdot)$ omitting logarithmic factors. A table of notation is provided in Appendix A.

2.2 Episodic MDP

An episodic MDP is identified by $\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, (P_h)_{h\in[H]}, (r_h)_{h\in[H]}, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the probability transition kernel at step h, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ the collection of reward functions at step h and H the length of one episode. The agent interacts with the environment for K episodes. At the beginning of episode k, Nature selects an initial state s_1^k arbitrarily. In step h, the agent takes action a_h^k and observes deterministic reward $r_h(s_h^k, a_h^k)$ 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 $r_{H+1} = 0$, 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 of 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 a sequence of functions $\pi = (\pi_h)_{h\in[H]}$ with $\pi_h : \mathcal{S} \to \mathcal{A}$.

2.3 Risk Measure

Consider two random variables $X \sim F$ and $Y \sim G$. We assert that Y dominates X, and correspondingly, G dominates F, denoted as $Y \succeq X$ and $G \succeq F$, if and only if for every real number x, the inequality $F(x) \geq G(x)$ holds true. A risk measure, ρ, is a function mapping a set of random variables, denoted as $\mathscr{X}$, to the real numbers. This mapping adheres to several crucial properties:

- **M**onotonicity (**M**): $X \preceq Y \Rightarrow \rho(X) \leq \rho(Y),\ \forall X, Y \in \mathscr{X}$,
- **T**ranslation-**i**nvariance (**TI**): $\rho(X + c) = \rho(X) + c,\ \forall X \in \mathscr{X},\ \forall c \in \mathbb{R}$,
- **D**istribution-**i**nvariance (**DI**): $F_{X_1} = F_{X_2} \Rightarrow \rho(X_1) = \rho(X_2)$.

A mapping $\rho : \mathscr{X} \to \mathbb{R}$ qualifies as a risk measure if it satisfies both (**M**) and (**TI**). Additionally, a risk measure that also adheres to **(DI)** is termed a distribution-invariant risk measure. Distribution-invariant risk measure allows us to simplify notation by denoting $\rho(F_X)$ as $\rho(X)$.

We direct our attention to a specific distribution-invariant risk measure, EntRM, a prominent risk measure in domains requiring risk-sensitive decision-making, such as mathematical finance (Föllmer and Schied, 2016), Markovian decision processes (Bäuerle and Rieder, 2014). For a random variable $X \sim F$ and a non-zero coefficient β, the EntRM is defined as:

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

Given that EntRM satisfies **(DI)**, we can denote $U_\beta(F)$ as $U_\beta(X)$ for $X \sim F$. When β possesses a small absolute value, employing Taylor's expansion yields

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

Therefore, a decision-maker aiming to maximize the EntRM value demonstrates risk-seeking behavior (preferring higher uncertainty in X) when $\beta > 0$, and risk-averse behavior (preferring lower uncertainty in X) when $\beta < 0$. The absolute value of β dictates the sensitivity to risk, with the measure converging to the mean functional as β approaches zero.

2.4 Risk-neutral Distributional Dynamic Programming Revisited

Bellemare et al. (2017); Rowland et al. (2018) have discussed the *infinite-horizon* distributional dynamic programming in the *risk-neutral* setting, which will be referred to as the classical DDP. Now we adapt their results to the finite horizon setting. We define the return for a policy π starting from the state-action pair (s, a) at step h as follows:

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

We then define $Y_h^\pi(s) \triangleq Z_h^\pi(s, \pi_h(s))$, noting that both $Z_h^\pi(s,a)$ and $Y_h^\pi(s)$ are random variables. It follows immediately that:

$$Z_h^\pi(s,a) = r_h(s,a) + Y_{h+1}^\pi(S'), \quad S' \sim P_h(\cdot \mid s, a).$$

There are two sources of randomness in $Z_h^\pi(s,a)$: the transition P_h^π and the next-state return Y_{h+1}^π. 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. Rewriting the random variable in the form of CDF, we have the distributional Bellman equation

$$\eta_h^\pi(s,a) = \sum_{s'} P_h(s'|s,a)\nu_{h+1}^\pi(s')(\cdot - r_h(s,a)), \nu_h^\pi(s) = \eta_h^\pi(s, \pi_h(s)),$$

where $\nu_{h+1}^\pi(s')(\cdot - r_h(s,a))$ denotes the CDF, $\nu_{h+1}^\pi(s')$, shifted by the reward $r_h(s,a)$. The distributional Bellman equation outlines the backward recursion of the return distribution under a fixed policy. Our focus is primarily on risk-neutral control, aiming to maximize the mean value of the return, as represented by:

$$\pi^*(s) \triangleq \arg\max_{(\pi_1,\ldots,\pi_H)\in\Pi} \mathbb{E}[Z_1^\pi(s)]$$

Here, $\pi = (\pi_1, ..., \pi_H)$ signifies that this is a multi-stage maximization problem. An exhaustive search approach is impractical due to its exponential computational complexity. However, the principle of optimality applies, suggesting that the optimal policy for any tail sub-problem coincides with the tail of the optimal policy, as discussed in (Bertsekas et al., 2000). This principle enables the reduction of the multi-stage maximization problem into several single-stage maximization problems. Let $Z^* = Z^{\pi^*}$ and $Y^* = Y^{\pi^*}$ denote the optimal return. The risk-neutral Bellman optimality equation can be expressed as follows:

$$\begin{aligned} Z_h^*(s,a) &= r_h(s,a) + Y_{h+1}^*(S'), S' \sim P_h(\cdot|s,a) \\ \pi_h^*(s) &= \arg\max_a \mathbb{E}[Z_h^*(s,a)], Y_h^*(s) = Z_h^*(s,\pi_h^*(s)). \end{aligned}$$

For simplicity, we define $[P\nu](s,a) \triangleq \sum_{s'} P(s'|s,a)\nu(s')$, where P represents the transition kernel and ν denotes the return distribution. Rewriting the above equation in the form of distributions, we get:

$$\begin{aligned} \eta_h^*(s,a) &= [P_h \nu_{h+1}^*](s,a)(\cdot - r_h(s,a)) \\ \pi_h^*(s) &= \arg\max_a \mathbb{E}[\eta_h^*(s,a)], \nu_h^*(s) = \eta_h^*(s,\pi_h^*(s)). \end{aligned}$$

3. Risk-sensitive Distributional Dynamic Programming

In this section, we establish a distributional dynamic programming framework for risk-sensitive control. For the risk-sensitive purpose, we define the action-value function of a policy π at step h as $Q_h^\pi(s,a) \triangleq U_\beta(Z_h^\pi(s,a))$, which is 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, that is,

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

In the risk-sensitive setting, however, the principle of optimality does not always hold for general risk measures. For example, the optimal policy for CVaR-MDP may be non-Markovian or history-dependent (Shapiro et al., 2021). The principle of optimality for EntRM-MDP, in terms of risk values, has been identified in prior works (Kupper and Schachermayer, 2009; Bäuerle and Rieder, 2014). We revisit this principle through a distributional perspective, leveraging the well established *independence property*.

The independence property (also known as the independence axiom) is a well known concept in economics and decision theory (Von Neumann and Morgenstern, 1947; Dentcheva and Ruszczynski, 2013). For better illustration, we introduce some additional notations. We write $X \geq (>) Y$ or $F \geq (>) G$ if $U_\beta(X) \geq (>) U_\beta(Y)$ or $U_\beta(F) \geq (>) U_\beta(G)$. This is different from the notion of stochastic dominance $X \succeq Y$. In fact, **(M)** of EntRM implies

$$X \succeq Y \Longrightarrow U_\beta(X) \geq U_\beta(Y) \Longleftrightarrow X \geq Y.$$

Fact 1 (Independence property) *Let $F, G, H \in \mathscr{D}$ and $\theta \in (0,1)$. The following holds:*

$$F < G \Longrightarrow \theta F + (1-\theta)H < \theta G + (1-\theta)H.$$

*We say that the EntRM satisfies property **(I)** (also known as the independence property).*

Proof We only prove the case that $\beta > 0$. The case that $\beta < 0$ follows analogously. For any two distributions F and G such that $U_\beta(F) > U_\beta(G)$, we have

$$U_\beta(F) = \frac{1}{\beta}\log\int_{\mathbb{R}}\exp(\beta x)dF(x) > \frac{1}{\beta}\log\int_{\mathbb{R}}\exp(\beta x)dG(x) = U_\beta(G),$$

which implies $\int_{\mathbb{R}}\exp(\beta x)dF(x) > \int_{\mathbb{R}}\exp(\beta x)dG(x)$. Thus for any distribution H,

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

This finishes the proof. ■

Moreover, the property (**TI**) entails that the EntRM value of the current return $Z_h^\pi(s,a)$ equals the sum of the immediate reward $r_H(s,a)$ and the value of the future return $Y_h^\pi(s')$

$$U_\beta(Z_h^\pi(s,a)) = U_\beta(r_h(s,a) + Y_h^\pi(s')) = r_h(s,a) + U_\beta(Y_h^\pi(s')) = r_h(s,a) + U_\beta([P_h\nu_{h+1}^\pi](s,a)).$$

We will show that **(I)** and **(TI)** suggest that the optimal current return, $Z_h^*(s,a)$, is determined by the optimal future return, $Y_h^*(s')$,

$$Z_h^*(s,a) = r_h(s,a) + Y_h^*(s').$$

These observations implies the principle of optimality. 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 $h_1 < h_2 \le H$.

Proposition 1 (Principle of optimality) *Let $\pi^* = \pi^*_{1:H}$ be an optimal policy. Fixing $h \in [H]$, then the truncated optimal policy $\pi^*_{h:H}$ is optimal for the sub-problem:*

$$\pi^*_{h:H} = \arg\max_{\pi_{h:H}\in\Pi_{h:H}} V_h^\pi.$$

Remark 2 *While the principle of optimality for EntRM-MDP has been identified in prior works, we revisit this principle through a distributional perspective. In particular, we derive Proposition 1 through **(I)**. This distributional perspective of dynamic programming will facilitate the algorithm design and regret analysis in distributional RL.*

Proof 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 $\tilde{s}_{h-1}$ with $P_{h-1}(\tilde{s}_h|\tilde{s}_{h-1}, \pi^*_{h-1}(\tilde{s}_{h-1})) > 0$ such that

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

where the inequality is due to **(I)** of U_β. It follows that $(\pi^*_{h-1}, \tilde{\pi}_{h:H})$ is a strictly better policy than $\pi^*_{h-1:H}$ for the subproblem from $h-1$ to H. Using induction, we deduce that $(\pi^*_{1:h-1}, \tilde{\pi}_{h:H})$ is a strictly better policy than $\pi^* = \pi^*_{1:H}$. This is contradicted to the assumption that π^* is an optimal policy. ■

Furthermore, the principle of optimality induces the *distributional Bellman optimality equation* in the risk-sensitive setting.

Proposition 3 (Distributional Bellman optimality equation) *The optimal policy π^* is given by the following backward recursions:*

$$\begin{aligned} &\nu^*_{H+1}(s) = \psi_0,\ \ \eta^*_h(s,a) = [P_h\nu^*_{h+1}](s,a)(\cdot - r_h(s,a)), \\ &\pi^*_h(s) = \arg\max_{a\in\mathcal{A}} Q^*_h(s,a) = U_\beta(\eta^*_h(s,a)), \nu^*_h(s) = \eta^*_h(s, \pi^*_h(s)), \end{aligned} \tag{2}$$

where $F(\cdot - c)$ denotes the CDF obtained by shifting F to the right by c. Furthermore, the sequence $(\eta^*_h)_{h\in[H]}$ and $(\nu^*_h)_{h\in[H]}$ represent the sequence of distributions corresponding to the optimal returns at each step.

Proof Throughout the proof we omit $*$ for the ease of notation. The proof follows from induction. Notice that $\eta_h(s_h)$ and $V_h(s_h)$ are the return distribution and value function for state s_h at step h following policy $\pi_{h:H}$ respectively. At step H, it is obvious that π_H is the optimal policy that maximizes the EntRM value at the final step. Fixing $h \in [H-1]$, we 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, $\forall \pi'_{h+1:H}, \forall s_{h+1}$:

$$U_\beta(\nu_{h+1}(s_{h+1})) = U_\beta(\nu_{h+1}^{\pi_{h+1:H}}(s_{h+1})) \geq U_\beta(\nu_{h+1}^{\pi'_{h+1:H}}(s_{h+1})).$$

It follows that $\forall s_h$,

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

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. ■

For simplicity, we define the *distributional Bellman operator* $\mathcal{T}(P, r) : \mathscr{D}^{\mathcal{S}} \to \mathscr{D}^{\mathcal{S}\times\mathcal{A}}$ with associated model $(P, r) = (P(s, a), r(s, a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ as

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

Denote by $\mathcal{T}_h \triangleq \mathcal{T}(P_h, r_h)$, then we can rewrite Equation 2 in a compact form:

$$\begin{aligned}
&\nu^*_{H+1}(s) = \psi_0,\ \eta^*_h(s, a) = [\mathcal{T}_h \nu^*_{h+1}](s, a), \\
&\pi^*_h(s) = \arg\max_{a\in\mathcal{A}} U_\beta(\eta^*_h(s, a)),\ \nu^*_h(s) = \eta^*_h(s, \pi^*_h(s)).
\end{aligned} \tag{3}$$

Discussion about the independence property Another property closely related to **(I)** is the tower (**T**) property (Kupper and Schachermayer, 2009).

Definition 4 (Tower property) *A risk measure ρ satisfies the tower property if for two r.v.s X and Y, we have*

$$\rho(X) = \rho(\rho(X|Y)),$$

where $\rho(\cdot|Y)$ is taken w.r.t. the conditional distribution.

We can show that the following implications hold

$$\textbf{(T)} \Longrightarrow \textbf{(I)} \Longrightarrow \text{Dynamic Programming (DP).}$$

(T) $\Longrightarrow$ **(I)**: Suppose **(T)** holds. Let $X_1 \sim F, X_2 \sim H, Y_1 \sim G, Y_2 \sim H$. Let $I \sim (1, 2; 1-\theta)$ be a binary r.v. independent of X and Y. Given $F \leq G$, we have

$$\begin{aligned}
\textbf{(DI)} &\Longrightarrow U_\beta(X_1) = U_\beta(F) \leq U_\beta(G) = U_\beta(Y_1) \\
&\Longrightarrow U_\beta(X_I|I) \sim (U_\beta(X_1), U_\beta(X_2); 1-\theta) \preceq (U_\beta(Y_1), U_\beta(Y_2); 1-\theta) \sim U_\beta(Y_I|I).
\end{aligned}$$

Next, Fact 2 implies

$$X_I \sim \theta F + (1-\theta)H, Y_I \sim \theta G + (1-\theta)H.$$

Fact 2 (Mixture distribution) *Let $X_i \sim F_i$ for $i \in [n]$. Let I be a discrete r.v.* independent of $(X_i)_i$ *with $\mathbb{P}(I = i) = \theta_i$. Then $X_I \sim \sum_{i\in[n]} \theta_i F_i$.*

It follows that

$$\begin{aligned} U_\beta(X_I|I) \preceq U_\beta(Y_I|I) &\Longrightarrow U_\beta(U_\beta(X_I|I)) \leq U_\beta(U_\beta(Y_I|I)) \\ &\Longrightarrow U_\beta(X_I) \leq U_\beta(Y_I) \\ &\Longrightarrow U_\beta(\theta F + (1-\theta)H)) \leq U_\beta(\theta G + (1-\theta)H), \end{aligned}$$

where the first and second implication is due to **(M)** and **(T)** of U_β.
(T) $\Longrightarrow$ DP: Fix $h \in [H-1]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Using **(T)** leads to a decomposition of the *risk-sensitive value function* as follows:

$$\begin{aligned} Q_h(s,a) = U_\beta\left(Z_h(s,a)\right) = U_\beta\left(r_h(s,a) + Y_{h+1}(S)\right) &= r_h(s,a) + U_\beta\left(Y_{h+1}(S)\right) \\ &= r_h(s,a) + U_\beta\left(U_\beta\left(Y_{h+1}(S)|S\right)\right) \\ &= r_h(s,a) + U_\beta\left(V_{h+1}(S)\right). \end{aligned}$$

We call it the risk-sensitive *Value* Bellman equation, which relates the action-value function Q_h at step h to the next-step value functions V_{h+1}. We can further derive the *Value Bellman optimality equation* with **(M)** and **(T)**. Observe that

$$\begin{aligned} V_{h+1}(s) \geq V'_{h+1}(s), \forall s &\Longrightarrow V_{h+1}(S)|S \succeq V'_{h+1}(S)|S \\ &\Longrightarrow U_\beta\left(V_{h+1}(S)\right) \geq U_\beta\left(V'_{h+1}(S)\right) \\ &\Longrightarrow Q_h(s,a) \geq Q'_h(s,a), \end{aligned}$$

which implies the Value Bellman optimality equation

$$\begin{aligned} Q^*_h(s,a) &= r_h(s,a) + U_\beta\left(V^*_{h+1}(S)\right) \\ \pi^*_h(s) &= \arg\max Q^*_h(s,a), V^*_h(s) = Q^*_h(s, \pi^*_h(s)). \end{aligned}$$

We make the following summary.
(i) Both **(T)** and **(I)** imply the principle of dynamic programming, but **(I)** is considered a weaker assumption than **(T)**. This indicates that while both properties support the formulation of DP, **(I)** does so under less stringent conditions.
(ii) **(I)** inspires a distributional perspective in EntRM-MDP, leading to the concept of DDP. This perspective involves running DP in the language of random variables or distributions, as opposed to traditional scalar values. In contrast, **(T)** primarily supports the classical Value Bellman equation. it is important to note that both distributional and classical DP contribute to the derivation of optimal policies. However, in our work, **(I)** plays a crucial and irreplaceable role. DDP, enabled by **(I)**, facilitates the algorithm design and regret analysis in distributional RL, which is not achievable solely with **(T)**.

Performance metric Finally, the regret of an algorithm $\mathscr{A}$ interacting with an MDP $\mathcal{M}$ for K episodes is defined as

$$\mathrm{Regret}(\mathscr{A}, \mathcal{M}, K) \triangleq \sum_{k=1}^{K} V^*_1(s^k_1) - V^{\pi^k}_h(s^k_1).$$

Note that the regret is a random variable since π^k is a random quantity. We denote by $\mathbb{E}[\mathrm{Regret}(\mathscr{A}, \mathcal{M}, K)]$ the expected regret. We will omit $\mathscr{A}$ and $\mathcal{M}$ for simplicity.

4. RODI-MF

In this section, we introduce **M**odel-**F**ree **R**isk-sensitive **O**ptimistic **D**istribution **I**teration (`RODI-MF`), as detailed in Algorithm 1.

Algorithm 1 `RODI-MF`

```
 1: Input: $T$ and $\delta$
 2: Initialize $N_h(\cdot,\cdot) \leftarrow 0$; $\eta_h(\cdot,\cdot), \nu_h(\cdot) \leftarrow \delta_{H+1-h}$ $\forall h \in [H]$
 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
```

We begin by establishing additional notations. For two Cumulative Distribution Functions (CDFs) F and G, the supremum distance between them is defined as $\|F - G\|_\infty \triangleq \sup_x |F(x) - G(x)|$. We define the ℓ_1 distance between two Probability Mass Functions (PMFs) with the same support P and Q as $\|P - Q\|_1 \triangleq \sum_i |P_i - Q_i|$. Furthermore, the set $B_\infty(F, c) := \{G \in \mathscr{D} | \|G - F\|_\infty \leq c\}$ denotes the supremum norm ball of CDFs centered at F with radius c. Analogously, $B_1(P, c)$ represents the ℓ_1 norm ball of PMFs centered at P with radius c.

In each episode, Algorithm 1 comprises two distinct phases: the planning phase and the interaction phase. During the planning phase, the algorithm executes an optimistic variant of the approximate Risk-Sensitive Distributional Dynamic Programming (RS-DDP), progressing backward from step $H+1$ to step 1 within each episode. This process results in a policy to be employed during the subsequent interaction phase. We offer further details about the two phases as follows:

Planning phase (Line 4-12). The algorithm undertakes a sample-based distributional Bellman update in Lines 5-7. To clarify, we append the episode index k to the variables in Algorithm 1 corresponding to episode k. For instance, η_h^k represents η_h in episode k. Specifically, for visited state-action pairs, Line 6 essentially performs an approximate DDP. Let $\mathbb{I}_h^k(s,a) \triangleq \mathbb{I}\{(s_h^k, a_h^k) = (s,a)\}$ and $N_h^k(s,a) \triangleq \sum_{\tau\in[k-1]} \mathbb{I}_h^{\tau}(s,a)$. For a given tuple (s, a, k, h) with $N_h^k(s,a) \geq 1$, , the empirical transition model $\hat{P}_h^k(\cdot|s,a)$ is defined as:

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

For any $\nu \in \mathscr{D}^{\mathcal{S}}$, the following holds:

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

Thus, the update formula in Line 6 of Algorithm 1 can be reformulated as:

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

Conversely, for unvisited state-action pairs, the return distribution remains aligned with the highest plausible reward $H+1-h$. Subsequently, the algorithm calculates the optimism constants c_h^k (Line 8) and applies the distributional optimism operator $\mathrm{O}_{c_h^k}^{\infty}$ (Line 9) to obtain the optimistically plausible return distribution η_h^k. The distributional optimism operator shifts a specified amount, c_h^k, of the leftmost probability mass of the input distribution to the rightmost end, thereby generating a more optimistic distribution. A formal definition of this operator will be presented in Section 4.2. The optimistic return distribution yields the optimistic value function, from which the algorithm derives the greedy policy π_h^k to be applied during the interaction phase.

Interaction phase (Line 14-17). In Lines 15-16, the agent interacts with the environment under policy π^k and refreshes the counts N_h^k based on newly gathered observations.

4.1 Connection to Exponential Utility

Our analysis explores the relationship between EntRM and Exponential Utility (EU). The EU is defined as follows:

$$E_\beta(F) \triangleq e^{\beta U_\beta(F)} = \int_{\mathbb{R}} e^{\beta x} dF(x),$$

where it serves as an exponential transformation of the EntRM. Notably, this transformation preserves the order in the sense that for any non-zero β, $\forall \beta \neq 0$,

$$U_\beta(F) \geq U_\beta(G) \Longleftrightarrow \mathrm{sign}(\beta)E_\beta(F) \geq \mathrm{sign}(\beta)E_\beta(G).$$

Leveraging this property, we derive the distributional Bellman optimality equation in terms of EU as follows:

$$\begin{aligned}\nu_{H+1}^*(s) = \psi_0,\ \eta_h^*(s,a) = [P_h\nu_{h+1}^*](s,a)(\cdot - r_h(s,a)),\\ \pi_h^*(s) = \arg\max_{a\in\mathcal{A}} \mathrm{sign}(\beta)E_\beta(\eta_h^*(s,a)),\ \nu_h^*(s) = \eta_h^*(s,\pi_h^*(s)).\end{aligned} \tag{4}$$

Proposition 5 (Equivalence between EntRM and EU) *The policy π^*, generated by Equation 4, is optimal for both the EntRM and EU. Moreover, the return distribution generated aligns with the optimal return distribution for EntRM.*

Proof The proof employs induction. The only difference between Equation 4 and Equation 2 lies in the policy generation step. For clarity, denote the quantities generated by the respective equations as $(\cdot)^*$ and $\tilde{(\cdot)}^*$. The base case with $\eta_H^* = \tilde{\eta}_H^*$ is evident. It follows that $\pi_H^*(s) = \tilde{\pi}_H^*(s)$ for each s, due to the preserved order under the exponential transformation. Subsequently, it holds that $\nu_H^* = \tilde{\nu}_H^*$. Assuming $\nu_{h+1}^* = \tilde{\nu}_{h+1}^*$ for $h+1 \in [2:H]$, we establish that $\eta_h^* = \tilde{\eta}_h^*$, $\pi_h^* = \tilde{\pi}_h^*$ and $\nu_h^* = \tilde{\nu}_h^*$. This completes the induction process. ■

We further present two important properties of EU, instrumental in formulating the regret upper bounds: *Lipschitz continuity* and *linearity.* Denote by L_M the Lipschitz constant of $E_\beta : \mathscr{D}_M \to \mathbb{R}$ with respect to the infinity norm $\|\cdot\|_\infty$, which satisfies:

$$E_\beta(F) - E_\beta(G) \leq L_M \|F - G\|_\infty, \forall F, G \in \mathscr{D}_M.$$

Lemma 6 establishes a *tight* Lipschitz constant for EU, linking the distance between distributions to the difference in their EU values.

Lemma 6 (Lipschitz property of EU) *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 with respect to both β and M.*

The proof is deferred to Appendix B. It is worth noting that as β approaches zero, L_M tends to zero, aligning with the observation that $\lim_{\beta\to 0} E_\beta = 1$. Another key property of EU is the linearity:

$$E_\beta(\theta F + (1-\theta)G) = \theta E_\beta(F) + (1-\theta)E_\beta(G).$$

This property significantly refines the regret bounds. In contrast, the non-linearity of EntRM may result in an exponential factor of $\exp(|\beta|H)$ in error propagation across time steps, potentially leading to a compounded factor of $\exp(|\beta|H^2)$ in the regret bound.

4.2 Distributional Optimism over the Return Distribution

For the purpose of clarity, we will focus on the scenario where $\beta > 0$ in the subsequent discussion. The case for $\beta < 0$ can be approached using analogous reasoning. We commence by formally defining optimism at the distributional level.

Definition 7 *Given two CDFs F and G, we say that F is more optimistic than G if $F \geq G$.*

This definition aligns with the intuitive notion that a more optimistic distribution should possess a larger EntRM value. Given that the exponential transformation preserves order, F is more optimistic than G if and only if $E_\beta(F) \geq E_\beta(G)$. Following the methodology of Keramati et al. (2020), we introduce the distributional optimism operator $\mathrm{O}_c^\infty : \mathscr{D}(a,b) \mapsto \mathscr{D}(a,b)$ for a level $c \in (0,1)$ as

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

This operator shifts the distribution F down by a maximum of c over $[a,b)$, ensuring that the resulting function $\mathrm{O}_c^\infty F$ remains a valid CDF in $\mathscr{D}(a,b)$ and optimistically dominates all other CDFs within the confidence ball $B_\infty(F,c)$[3]. In particular, $\mathrm{O}_c^\infty F$ represents maximum permissible downward adjustment under the constraints of the infinity ball, ensuring that for any distribution $G \in B_\infty(F,c)$

$$(\mathrm{O}_c^\infty F)(x) \leq G(x), \forall x \in \mathbb{R}.$$

Lemma 8 *Let ρ be a functional (not necessarily a risk measure) satisfying* **(M)***. 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*

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

Proof Consider any $G \in \mathscr{D}([a,b]) \cap B_\infty(F,c)$. By the definition of $B_\infty(F,c)$, we have $\sup_{x\in[a,b]} |F(x) - G(x)| \leq c$. Therefore, for any $x \in [a,b]$, we have $G(x) \geq \max(F(x)-c, 0) = (\mathrm{O}_c^\infty F)(x)$. Since $\mathrm{O}_c^\infty F$ dominates any G in $\mathscr{D}([a,b]) \cap B_\infty(F,c)$ and considering **(M)** of ρ, we arrive at the conclusion. ■

We define the EU 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). 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). We define $\iota = \log(SAT/\delta)$ for any $\delta \in (0,1)$ and introduce the *good event* as follows:

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

This event encapsulates the scenario where the empirical distributions concentrates around the true distributions with respect to the ℓ_1 norm. Leveraging Lemma 12, **(M)** of EU, and inductive reasoning, we arrive at Proposition 9, which asserts that the sequence $W_1^k(s_1^k)_{k\in[K]}$ is consistently optimistic compared to the optimal value sequence $W_1^*(s_1^k)_{k\in[K]}$.

Proposition 9 (Optimism) *Conditioned on event $\mathcal{G}_\delta$, the sequence $\{W_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 1 are all greater than or equal to $W_1^*(s_1^k)$, i.e.,*

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

We first present a series of lemmas, specifically Lemma 10 through Lemma 12, which is used in the proof of Proposition 9.

Lemma 10 (High probability good event) *For any $\delta \in (0,1)$, the event $\mathcal{G}_\delta$ is true with probability at least $1-\delta$.*

We will verify the distributional optimism conditioned on $\mathcal{G}_\delta$.

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

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

[3] For a more comprehensive explanation, please refer to Liang and Luo (2023).

Proof

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

■

Lemma 12 (Bound on the optimistic constant) *For any bounded distributions $\{F_i\}_{i\in[n]}$ and any $\theta,\theta'\in\Delta_n$ it holds that if $c\geq\|\theta-\theta'\|_1$, then*

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

Proof Without loss of generality assume $F\in\mathscr{D}_M^n$. By Lemma 11, for any x

$$\left|\sum_{i=1}^{n}(\theta_i'-\theta_i)F_i(x)\right|\leq\left\|\theta'-\theta\right\|_1 .$$

For any $x\in[0,M+1)$,

$$\mathrm{O}_c^\infty\left(\sum_{i=1}^{n}\theta_i' F_i\right)(x)=\left[\sum_{i=1}^{n}\theta_i' F_i(x)-c\right]^+=\left[\sum_{i=1}^{n}\theta_i F_i(x)+\sum_{i=1}^{n}(\theta_i'-\theta_i)F_i(x)-c\right]^+$$
$$\leq\left[\sum_{i=1}^{n}\theta_i F_i(x)+\left\|\theta'-\theta\right\|_1-c\right]^+\leq\left[\sum_{i=1}^{n}\theta_i F_i(x)\right]^+=\sum_{i=1}^{n}\theta_i F_i(x).$$

■

Now we give the proof of Proposition 9.

Proof The proof proceeds by induction. We fix $k\in[K]$ and consider each stage h in reverse order. Consider the base case: for any (s,a) such that $N_H^k(s,a)>0$

$$J_H^k(s,a)=E_\beta(\eta_H^k(s,a))=E_\beta(\delta_{r_H(s,a)})=\exp(\beta r_H(s,a))=J_H^*(s,a).$$

This equality holds because the reward received at stage H is deterministic and hence the EU is simply the exponential of the scaled reward. For unvisited state-action pairs (s,a) with $N_H^k(s,a)=0$, the EU is given by:

$$J_H^k(s,a)=E_\beta(\eta_H^k(s,a))=E_\beta(\delta_1)=\exp(\beta)\geq J_H^*(s,a).$$

Here, the EU value defaults to $\exp(\beta)$, which is greater than or equal to the optimal EU value for any (s,a). Given these calculations, for any state s, the EU value at stage H satisfies $W_H^k(s)=\max_a J_H^k(s,a)\geq\max_a J_H^*(s,a)=W_H^*(s)$, establishing the base case. Assuming that for stage $h+1$, $W_{h+1}^k(s)\geq W_{h+1}^*(s)$ holds for all states s, we now consider stage h. For any visited state-action pair (s,a) with $N_h^k(s,a) > 0$, we apply Lemma 12 with $\theta = P_h(s,a), \theta' = \hat{P}_h^k(s,a)$, $F = \nu_{h+1}^k$ to obtain

$$[P_h \nu_{h+1}^k](s,a) \preceq \mathrm{O}^{\infty}_{c_h^k(s,a)}([\hat{P}_h^k \nu_{h+1}^k](s,a)),$$

given that $c_h^k(s,a) = \sqrt{\frac{2S}{N_h^k(s,a) \vee 1}\iota} \geq \left\| P_h(\cdot|s,a) - \hat{P}_h^k(\cdot|s,a) \right\|_1$ for $h \in [H-1]$. We have

$$\begin{aligned} J_h^k(s,a) &= E_\beta(\mathrm{O}^{\infty}_{c_h^k(s,a)}([\hat{P}_h^k \nu_{h+1}^k](s,a)(\cdot - r_h(s,a)))) \\ &= \exp(\beta r_h(s,a)) E_\beta(\mathrm{O}^{\infty}_{c_h^k(s,a)}([\hat{P}_h^k \nu_{h+1}^k](s,a))) \\ &\geq \exp(\beta r_h(s,a)) E_\beta([P_h \nu_{h+1}^k](s,a)) \\ &= \exp(\beta r_h(s,a)) \cdot [P_h W_{h+1}^k](s,a) \\ &\geq \exp(\beta r_h(s,a)) \cdot [P_h W_{h+1}^*](s,a) \\ &= J_h^*(s,a), \end{aligned}$$

where the first inequality is due to **(M)**, and the second inequality follows from the induction assumption. For unvisited state-action pairs (s,a) at stage h, the EU is calculated based on the maximum possible reward, ensuring that:

$$J_h^k(s,a) = E_\beta(\delta_{H+1-h}) = \exp(\beta(H+1-h)) \geq J_h^*(s,a).$$

Finally, aggregating these values for any state s at stage h, we obtain:

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

completing the induction step and thereby the proof. ■

4.3 Regret Upper Bound of `RODI-MF`

Theorem 13 (Regret upper bound of `RODI-MF`) *For any $\delta \in (0,1)$, with probability $1-\delta$, the regret of Algorithm 1 is bounded as*

$$\mathit{Regret}(\texttt{RODI-MF}, K) \leq \mathcal{O}\left(L_H(U_\beta) H \sqrt{S^2 A K \iota}\right) = \tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H) - 1}{|\beta|} H \sqrt{S^2 A K}\right),$$

where $L_H(U_\beta) = \frac{\exp(|\beta|H)-1}{|\beta|}$ is the Lipschitz constant of EntRM over $\mathscr{D}(0,H)$.

Remark 14 *The regret bounds achieved by **`RODI-MF`** match the best-known results in Fei et al. (2021). In particular, **`RODI-MF`** attains exponentially improved regret bounds compared to RSVI and RSQ in Fei et al. (2020) with a factor of $\exp(|\beta|H^2)$.*

Remark 15 *For values of β that are close to zero, an expansion using Taylor's series reveals that the EntRM, $U_\beta(Z^\pi)$, can be approximated by the sum of the expected cumulative reward and a term proportional to the variance of the cumulative reward, with higher-order terms contributing insignificantly. Considering that the reward r_h lies in the interval $[0,1]$, both the expected cumulative reward and its variance are bounded by terms linear and quadratic in H, respectively. To balance the expected reward and the risk (as quantified by the variance), it is prudent to choose $\beta = \mathcal{O}(1/H)$.*

Remark 16 *By choosing $|\beta| = \mathcal{O}(1/H)$, we can eliminate the exponential dependency on H and achieve polynomial regret bound akin to the risk-neutral setting. Therefore, DRL can achieve $\mathrm{O}\left(H\sqrt{HS^2AT}\right)$ regret bound for RSRL with reasonable risk-sensitivity.*

Proof We first prove the case $\beta > 0$. Define $\Delta_h^k \triangleq W_h^k - W_h^{\pi^k} = E_\beta(\nu_h^k) - E_\beta\left(\nu_h^{\pi^k}\right) \in D_h^S$

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

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

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

where the last inequality follows from Lemma 32 and that both $W_1^k(s_1^k)$ and $W_1^{\pi^k}(s_1^k)$ are larger than 1. We can decompose δ_h^k as follows

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

Using the Lipschitz property of EU, we have

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

We can bound (b) as

$$\begin{aligned}
(b) &= \exp(\beta r_h^k)\left(E_\beta\left(\left[\hat{P}_h^{\pi^k} \nu_{h+1}^k\right](s_h^k)\right) - E_\beta\left(\left[P_h^{\pi^k} \nu_{h+1}^k\right](s_h^k)\right)\right) \\
&\leq \exp(\beta) L_{H-h} \left\| \left[\hat{P}_h^{\pi^k} \nu_{h+1}^k\right](s_h^k) - \left[P_h^{\pi^k} \nu_{h+1}^k\right](s_h^k) \right\|_\infty \\
&\leq \exp(\beta)(\exp(\beta(H-h)) - 1) \left\| \hat{P}_h^{\pi^k}(s_h^k) - P_h^{\pi^k}(s_h^k) \right\|_1 \\
&\leq (\exp(\beta(H+1-h)) - 1)\sqrt{\frac{2S}{(N_h^k \vee 1)}\iota},
\end{aligned}$$

where the second inequality is due to Lemma 11. Observe that

$$(c) = \exp(\beta r_h^k)\left[P_h^{\pi^k}(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_h^k) = \exp(\beta r_h^k)\left[P_h^{\pi^k} \Delta_{h+1}^k\right](s_h^k) = \exp(\beta r_h^k)(\epsilon_h^k + \delta_{h+1}^k),$$

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 $(k,h) \in [K] \times [H]$, and $e_h^k \triangleq \left\| \hat{P}_h^k(s_h^k) - P_h^{\pi^k}(s_h^k) \right\|_1$. Since $(a) + (b) \leq 2L_{H+1-h} c_h^k$, we can bound δ_h^k recursively as

$$\delta_h^k \leq 2L_{H+1-h} c_h^k + \exp(\beta r_h^k)(\epsilon_h^k + \delta_{h+1}^k).$$

Repeating the procedure, we get

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

Thus

$$\sum_{k=1}^{K} \delta_1^k \leq 2(\exp(\beta H) - 1)\sum_{k=1}^{K}\sum_{h=1}^{H-1} c_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H-1}\prod_{i=1}^{h} \exp(\beta r_i^k)\epsilon_h^k + \sum_{k=1}^{K} \exp(\beta(H-1))\delta_H^k.$$

Now we bound each term separably. The first term can be bounded as

$$\begin{aligned}2(\exp(\beta(H+1))-1)\sum_{k=1}^{K}\sum_{h=1}^{H-1}c_h^k &= 2(\exp(\beta(H+1))-1)\sum_{h=1}^{H-1}\sum_{k=1}^{K}\sqrt{\frac{2S}{(N_h^k\vee 1)}\iota}\\ &\le 4(\exp(\beta(H+1))-1)\sum_{h=1}^{H-1}\sqrt{2S^2AK\iota}\\ &= 4(\exp(\beta(H+1))-1)(H-1)\sqrt{2S^2AK\iota}.\end{aligned}$$

Observe that

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

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

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

The third term can be bounded as

$$\begin{aligned}&\sum_{k=1}^{K}\exp(\beta(H-1))\delta_H^k = \exp(\beta(H-1))\sum_{k=1}^{K}W_H^k(s_H^k)-W_H^{\pi^k}(s_H^k)\\ &=\exp(\beta(H-1))\sum_{k=1}^{K}\mathbb{I}\{N_H^k=0\}\exp(\beta)+\mathbb{I}\{N_H^k>0\}\exp(\beta r_H(s_H^k))-\exp(\beta r_H(s_H^k))\\ &\le \exp(\beta(H-1))(\exp(\beta)-1)\sum_{k=1}^{K}\mathbb{I}\{N_H^k=0\}\\ &\le (\exp(\beta H)-1)SA < (\exp(\beta(H+1))-1)SA\end{aligned}$$

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

$$\begin{aligned}\mathrm{Regret}(K) &\le \frac{\exp(\beta(H+1))-1}{\beta}\left(4(H-1)\sqrt{2S^2AK\iota}+\sqrt{2KH\iota}+SA\right)\\ &=\tilde{\mathcal{O}}\left(\frac{\exp(\beta H)-1}{\beta}\sqrt{HS^2AT}\right),\end{aligned}$$

where $\iota \triangleq \log(2SAT/\delta)$.

Now we consider the case $\beta < 0$. Using similar arguments, we arrive at

$$\begin{aligned}\operatorname{Regret}(K) &\le \sum_{k=1}^{K} \frac{1}{\beta}\log\left(W_1^k(s_1^k)\right) - \frac{1}{\beta}\log\left(W_1^{\pi^k}(s_1^k)\right)\\ &= \sum_{k=1}^{K} \frac{1}{-\beta}\left(\log\left(W_1^{\pi^k}(s_1^k)\right) - \log\left(W_1^k(s_1^k)\right)\right)\\ &\le \frac{1}{-\beta\exp(\beta H)}\sum_{k=1}^{K} W_1^{\pi^k}(s_1^k) - W_1^k(s_1^k),\end{aligned}$$

where the last inequality is due to that both $W_1^{\pi^k}(s_1^k)$ and $W_1^k(s_1^k)$ is larger than or equal to $\exp(\beta H)$. We can finally get

$$\operatorname{Regret}(K) \le \tilde{\mathcal{O}}\left(\frac{1-\exp(\beta H)}{-\beta\exp(\beta H)}\sqrt{HS^2AT}\right) = \tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H)-1}{|\beta|}\sqrt{HS^2AT}\right).$$

■

4.4 Computational inefficiency of `RODI-MF`

While `RODI-MF` enjoys near-optimal regret guarantee, it suffers from computational inefficiency, especially in contexts with a large number of states or a long horizon. For better illustration, let's consider a *Markov Reward Process* with S states at each step. In this setup, the transition kernel is uniform ($P_h(s'|s) = 1/S$) for any $(h, s') \in [H-1] \times \mathcal{S}$, and the reward function is bounded between 0 and 1 ($r_h(s) \in [0, 1]$). Starting from the final step H, the return distribution $\eta_H(s)$ is a Dirac delta function centered at $r_H(s)$. Applying the distributional Bellman equation at step at step $H-1$, we get

$$\eta_{H-1}(s) = \sum_{s'} p_{H-1}(s'|s)\delta_{r_H(s')+r_{H-1}(s)}.$$

Recall that $|\eta|$ represents the number of atoms (distinct elements) in a discrete distribution η, indicating the memory required to store this distribution. Since $|\eta_H(s)| = |\delta_{r_H(s)}| = 1$ for each $s \in \mathcal{S}$, and $\eta_{H-1}(s)$ is a uniform mixture of all $\eta_H(s)$ shifted by $r_{H-1}(s)$, we find

$$|\eta_{H-1}(s)| = \left|\left(r_{H-1}(s) + r_H(s'), \frac{1}{S}\right)_{s'\in\mathcal{S}}\right| = \mathcal{O}(S).$$

Continuing this process backwards through the time steps:

$$\begin{aligned}&|\eta_{H-2}(s)| = \mathcal{O}(S^2)\\ &\cdots\\ &|\eta_1(s)| = \mathcal{O}(S^{H-1}).\end{aligned}$$

This analysis shows that the number of atoms in the return distribution *exponentially* increases with the horizon H, scaled by the number of states S at each application of the distributional Bellman operator. As a result, the memory and computational requirements to implement an *exact* distributional RL algorithm like `RODI-MF` become prohibitive, particularly for problems with many states or a long horizon. This exponential growth in complexity highlights the computational challenges associated with `RODI-MF` and underscores the need for *approximations* for practical implementations.

5. DRL with Distribution Representation

To address the computational challenges in implementing the distributional Bellman equation, we introduce two versions of `RODI-MF` in the revised paper that utilize *distribution representation*. A widely used method of distribution representation is the *categorical representation*, as discussed in recent literature (Bellemare et al., 2023). This approach parameterizes the probability distribution at fixed locations. Specifically, we consider the simplest form of categorical representation that uses only two atoms. We refer to this as the *Bernoulli representation*. It represents the set of all discrete distributions with two distinct atoms, denoted as $\theta = (\theta_1, \theta_2)$. We refer to θ_1 and θ_2 as the left and right atom. The Bernoulli representation is formally defined as:

$$\mathscr{F}_{\mathrm{B}}(\theta) = \left\{(1-p)\delta_{\theta_1} + p\delta_{\theta_2} : p \in [0,1]\right\}.$$

With the Bernoulli representation in mind, let's consider distributional Bellman operator

$$\eta_h(s,a) = [\mathcal{T}_h \nu_{h+1}](s,a) = \sum_{s'} P_h(s'|s,a)\nu_{h+1}(s')(\cdot - r_h(s,a)).$$

This operator essentially performs two basic operations: shifting and mixing. Specifically, it shifts each next-step return distribution by the reward $r_h(s,a)$ and then takes a mixture of these shifted distributions with the mixture coefficients $P_h(s,a)$. However, these operations might change and expand the support of the distributions, leading to:

$$|\eta_h(s,a)| = \mathcal{O}(S)|\nu_{h+1}|.$$

To improve computational efficiency, we introduce the Bernoulli representation for $\mathcal{T}_h$. Let

$$\bar{\nu}_{h+1}(s) = (L_{h+1}(s), R_{h+1}(s); q_{h+1}(s)) \in \mathscr{F}_{\mathrm{B}}(L_{h+1}(s), R_{h+1}(s))$$

be a Bernoulli representation of the true return distribution $\nu_{h+1}(s)$, where $L_{h+1}(s)$ and $R_{h+1}(s)$ are the left and right atoms, and $q_{h+1}(s)$ is the probability at $R_{h+1}(s)$. Applying $\mathcal{T}_h$ to $\bar{\nu}_{h+1}$, we obtain

$$[\mathcal{T}_h \bar{\nu}_{h+1}](s,a) = \big(r_h(s,a) + L_{h+1}(s'), r_h(s,a) + R_{h+1}(s'); p_h(s'|s,a)q_{h+1}(s')\big)_{s' \in \mathcal{S}} \notin \mathscr{F}_{\mathrm{B}}.$$

The result is no longer a Bernoulli distribution but a categorical distribution with $2S$ atoms. This demonstrates that the Bernoulli representation is not *closed* under $\mathcal{T}_h$

$$\nu \in \mathscr{F}_{\mathrm{B}} \not\Longrightarrow \mathcal{T}_h \nu \in \mathscr{F}_{\mathrm{B}}.$$

To overcome this issue, we introduce the *Bernoulli projection operator*. This operator serves as a mapping from the space of probability distributions to $\mathscr{F}_{\mathrm{B}}$, and we denote it as Π :

$\mathscr{P}(\mathbb{R}) \mapsto \mathscr{F}_{\mathrm{B}}$. Algorithmically, we add a projection step immediately after the application of $\mathcal{T}$, resulting in a *projected distributional Bellman operator* $\Pi\mathcal{T}$. This projection ensures that each iteration of $\eta_h = \Pi\mathcal{T}_h\nu_{h+1}$ is representable using a limited amount of memory.

Note that the projection operator is not unique. Previous works (Bellemare et al., 2023) have developed projection operators aiming to find the best approximation to a given probability distribution, as measured by a specific probability metric. Our approach introduces a novel type of Bernoulli projection that *preserves the ERM value*, an essential aspect in risk-sensitive settings. Starting from a Dirac measure δ_c, we define the *value-equivalent Bernoulli projection operator* as:

$$\Pi\delta_c \triangleq (1 - q(c;\theta))\delta_{\theta_1} + q(c;\theta)\delta_{\theta_2} = (\theta_1, \theta_2; q(c;\theta)),$$

where the probability is defined as

$$q(c;\theta) = \frac{e^{\beta c} - e^{\beta\theta_1}}{e^{\beta\theta_2} - e^{\beta\theta_1}} \in [0,1]. \tag{5}$$

It is easy to verify that $U_\beta(\Pi\delta_c) = U_\beta(\delta_c) = c, \forall c \in [\theta_1, \theta_2]$. Now we extend the definition to the categorical distributions as:

$$\begin{aligned}\Pi\left(c_i, p_i\right)_{i\in[n]} = \Pi\left(\sum_{i\in[n]} p_i\delta_{c_i}\right) &\triangleq \sum_i p_i\Pi\delta_{c_i} = \sum_i p_i\left((1-q(c_i;\theta))\delta_{\theta_1} + q(c_i;\theta)\delta_{\theta_2}\right)\\ &= \left(\sum_i p_i(1-q(c_i;\theta))\right)\cdot\delta_{\theta_1} + \left(\sum_i p_i q(c_i;\theta)\right)\cdot\delta_{\theta_2}\\ &= \left(\theta_1, \theta_2; \sum_i p_i q(c_i;\theta)\right).\end{aligned}$$

Given that $\mathrm{EU}\left(\delta_{c_i}\right) = \mathrm{EU}\left(\Pi\delta_{c_i}\right), \forall i \in [n]$, the linearity of EU implies

$$\mathrm{EU}\left(\sum_i p_i\delta_{c_i}\right) = \sum_i p_i \mathrm{EU}\left(\delta_{c_i}\right) = \sum_i p_i\mathrm{EU}\left(\Pi\delta_{c_i}\right) = \mathrm{EU}\left(\sum_i p_i\Pi\delta_{c_i}\right) = \mathrm{EU}\left(\Pi\sum_i p_i\delta_{c_i}\right).$$

This verifies the value equivalence of Π.

To ensure the preservation of the value, the only requirement is that the interval $[\theta_1, \theta_2]$ covers the support of the input distribution, i.e., $\theta_1 \leq \min c_i \leq \max c_i \leq \theta_2$. The projection preserves the risk value of the original distribution, making it a powerful tool for efficient and accurate representation in DRL for RSRL.

Without the knowledge of MDP, `RODI-MF` deviates from the DDP in two crucial updates:

$$\begin{aligned}\hat{\eta}_h &\leftarrow \hat{\mathcal{T}}_h\nu_{h+1}\\ \tilde{\eta}_h &\leftarrow \mathrm{O}_c\hat{\eta}_h.\end{aligned}$$

In `RODI-MF`, the *approximate distributional Bellman operator* $\hat{\mathcal{T}}$ is applied first, which relies on the empirical transition $\hat{P}$ rather than the true transition P. Then, the *distributional optimism operator* O_c is used to generate an optimistic return distribution. Drawing from these observations, we propose two DRL algorithms with Bernoulli representation, differing in the order of projection and optimism operator. We term the two algorithms as `RODI-Rep`.

5.1 DRL with Bernoulli Representation

Given that $\eta_h \in D_{H+1-h}$, we set the *uniform* and predefined location parameters as

$$L_h \triangleq 0, \quad R_h \triangleq H+1-h,$$

which is independent of (s,a). We represent each iterate by a Bernoulli distribution

$$\eta_h^k(s,a) = (1-q_h^k(s,a))\delta_{L_h} + q_h^k(s,a)\delta_{R_h}, \nu_h^k(s) = (1-q_h^k(s))\delta_{L_h} + q_h^k(s)\delta_{R_h},$$

where we overload the notation for $q_h^k(s,a)$ and $q_h^k(s)$. Applying the approximate $\mathcal{T}_h$ to the Bernoulli represented $\nu_{h+1}^k \in \mathscr{F}_{\mathrm{B}}$ yields

$$\begin{aligned}
\eta_h^k(s,a) = [\hat{\mathcal{T}}_h \nu_{h+1}^k](s,a) &= \sum_{s' \in \mathcal{S}} \hat{P}_h^k(s'|s,a)\nu_{h+1}^k(s')(\cdot - r_h(s,a)) \\
&= \sum_{s' \in \mathcal{S}} \hat{P}_h^k(s'|s,a)((1-q_h^k(s'))\delta_{L_{h+1}} + q_h^k(s')\delta_{R_{h+1}})(\cdot - r_h(s,a)) \\
&= (1-[\hat{P}_h^k q_{h+1}^k](s,a)) \cdot \delta_{r_h(s,a)+L_{h+1}} + [\hat{P}_h^k q_{h+1}^k](s,a) \cdot \delta_{r_h(s,a)+R_{h+1}} \\
&= \Big(r_h(s,a) + L_{h+1}, r_h(s,a) + R_{h+1}; [\hat{P}_h^k q_{h+1}^k](s,a)\Big).
\end{aligned}$$

With slight abuse of notation, we let

$$L_h(s,a) \triangleq r_h(s,a) + L_{h+1}, \quad R_h(s,a) \triangleq r_h(s,a) + R_{h+1}.$$

$\eta_h^k(s,a) = \Big(L_h(s,a), R_h(s,a); [\hat{P}_h^k q_{h+1}^k](s,a)\Big)$ is a Bernoulli distribution with support not corresponding to L_h and R_h. Now, we propose two different algorithms differing in the order of projection and optimism operator.

Optimism-Then-Projection. `RODI-OTP` applies the optimism operator first, followed by the projection operator:

$$\eta_h^k \leftarrow \Pi \mathrm{O}_c \hat{\mathcal{T}}_h \nu_{h+1}^k.$$

Note that $\eta_h^k \leftarrow \hat{\mathcal{T}}_h \nu_{h+1}^k \in \mathscr{F}_{\mathrm{B}}(r_h(s,a)+L_{h+1}, r_h(s,a)+R_{h+1})$. For Bernoulli distribution, the optimism operator admits a simple form

$$\mathrm{O}_c\left(a,b;p\right) = \left(a,b;\min(p+c,1)\right).$$

Applying optimism operator to η_h^k yields

$$\begin{aligned}
\mathrm{O}_{c_h^k(s,a)}\left(\eta_h^k(s,a)\right) &= \mathrm{O}_{c_h^k(s,a)}\left(L_h(s,a), R_h(s,a); [\hat{P}_h^k q_{h+1}^k](s,a)\right) \\
&= \left(L_h(s,a), R_h(s,a); \min\left([\hat{P}_h^k q_{h+1}^k](s,a) + c_h^k(s,a), 1\right)\right).
\end{aligned}$$

We can simplify the update in a parametric form

$$\begin{aligned}
&q_h^k(s,a) \leftarrow [\hat{P}_h^k q_{h+1}^k](s,a) \quad \text{parametric Bellman update} \\
&q_h^k(s,a) \leftarrow \min(q_h^k(s,a) + c_h^k(s,a), 1) \quad \text{optimism operator.}
\end{aligned}$$

Finally, we apply the projection rule (cf. Equation 5) to obtain

$$\begin{aligned} q_h^k(s,a) &\leftarrow (1 - q_h^k(s,a)) q(L_h(s,a); L_h, R_h) + q_h^k(s,a) q(R_h(s,a); L_h, R_h) \\ &= (1 - q_h^k(s,a)) q_h^L(s,a) + q_h^k(s,a) q_h^R(s,a), \end{aligned}$$

where

$$\begin{aligned} q_h^R(s,a) &\triangleq q(L_h(s,a); L_h, R_h) = \frac{e^{\beta(r_h(s,a)+H-h)} - 1}{e^{\beta(H+1-h)} - 1}, \\ q_h^L(s,a) &\triangleq q(R_h(s,a); L_h, R_h) = \frac{e^{\beta r_h(s,a)} - 1}{e^{\beta(H+1-h)} - 1}. \end{aligned}$$

Remark 17 *$q_h^R(s,a)$ and $q_h^L(s,a)$ are* fixed *(independent of k) and* known*. Therefore we can compute their values for all (h,s,a) in advance.*

Projection-Then-Optimism. `RODI-PTO` applies the projection operator first, followed by the optimism operator:

$$\eta_h^k \leftarrow \mathrm{O}_c \Pi \hat{\mathcal{T}}_h \nu_{h+1}^k.$$

The update can be represented in a parametric form

$$\begin{aligned} &q_h^k(s,a) \leftarrow [\hat{P}_h^k q_{h+1}^k](s,a) \quad \text{parametric Bellman update} \\ &q_h^k(s,a) \leftarrow (1 - q_h^k(s,a)) q_h^L(s,a) + q_h^k(s,a) q_h^R(s,a) \quad \text{projection operator} \\ &q_h^k(s,a) \leftarrow \min(q_h^k(s,a) + c_h^k(s,a), 1) \quad \text{optimism operator.} \end{aligned}$$

After applying optimism operator and projection operator, both `RODI-OTP` and `RODI-PTO` update the value functions and policies accordingly

$$\begin{aligned} Q_h^k(s,a) &\leftarrow \frac{1}{\beta} \log\left(1 - q_h^k(s,a) + q_h^k(s,a) e^{\beta(H+1-h)}\right) \\ \pi_h^k(s) &\leftarrow \arg\max_a Q_h^k(s,a), V_h^k(s) \leftarrow Q_h^k(s, \pi_h^k(s)) \\ q_h^k(s) &\leftarrow q_h^k(s, \pi_h^k(s)). \end{aligned}$$

Computational complexity. The *time complexity* of `RODI-OTP` and `RODI-PTO` is given as follows: i) computation of q^L and q^R: $\mathcal{O}(HSA)$; ii) parametric Bellman update: $KHSA \cdot \mathcal{O}(S)$; iii) projection: $KHSA \cdot \mathcal{O}(1)$; iv) optimism operator: $KHSA \cdot \mathcal{O}(1)$; v) computation of Q-function: $KHSA \cdot \mathcal{O}(1)$; vi) greedy policy: $KHS \cdot \mathcal{O}(A \log A)$. Therefore, the total time complexity is given by

$$\mathcal{O}(KHSA(S + \log A),$$

which is the same as that of `RSVI2`. The *space complexity* of both algorithm is given as follows: i) q^L and q^R: $\mathcal{O}(HSA)$; ii) $N_h(s,a)$: $\mathcal{O}(HSA)$; iii) trajectory $(s_h^k, a_h^k)_{k,h}$: $\mathcal{O}(T)$; iv) probabilities $q_h(s,a)$: $\mathcal{O}(HSA)$; v) action-value function: $\mathcal{O}(HSA)$. Therefore, their total space complexity is $\mathcal{O}(HSA + T)$.

5.2 Optimism of DRL with Representation

While RODI-OTP and RODI-PTO adapts RODI-MF by Bernoulli representation, they maintain the optimism mainly due to the value-equivalence property of the projection operator. The optimism ensures that RODI-OTP and RODI-PTO enjoy the same regret bound as RODI-MF. By establishing the optimism of the value functions $V_1^* \leq V_1^k$, we have:

$$\text{Regret} = \sum_{k\in[K]} (V_1^* - V_1^{\pi^k}) \leq \sum_{k\in[K]} (V_1^k - V_1^{\pi^k}).$$

Thus, the regret of an algorithm can be bounded by the cumulative difference between the optimistic value function V^k and $V_1^{\pi^k}$. It is intuitive that a smaller V^k or less optimism leads to reduced regret. Furthermore, in Section 8.2, we provide a detailed analysis to theoretically demonstrate that RODI-OTP and RODI-PTO have smaller value functions compared to the non-distributional RL algorithm RSVI2, resulting in less regret than RSVI2.

Optimism of RODI-OTP. For simplicity, we rewrite the update formula of RODI-OTP as

$$\begin{aligned}
\hat{\eta}_h(s,a) &= [\hat{\mathcal{T}}_h \nu_{h+1}](s,a) = \left[\hat{P}_h \nu_{h+1}\right][s,a](\cdot - r_h(s,a)) \\
\tilde{\eta}_h(s,a) &= \mathrm{O}_{c_h(s,a)} \hat{\eta}_h(s,a) \\
\eta_h(s,a) &= \Pi \tilde{\eta}_h(s,a) \\
Q_h(s,a) &= U_\beta\left(\eta_h(s,a)\right), \pi_h(s) = \arg\max_a Q_h(s,a) \\
\nu_h(s) &= \eta_h(s, \pi_h(s)).
\end{aligned}$$

Define

$$\check{\eta}_h(s,a) \triangleq [\mathcal{T}_h \nu_{h+1}](s,a) = \left[P_h \nu_{h+1}\right][s,a](\cdot - r_h(s,a)),$$

which is the Bellman target that replaces $\hat{P}_h$ by the true model P_h. Note that $\nu_{h+1} \in \mathscr{F}_{\mathrm{B}}$ is the distribution generated by the algorithm, which is Bernoulli represented, rather than the optimal distribution ν_{h+1}^*. Since

$$\begin{aligned}
\left\|\check{\eta}_h(s,a) - \hat{\eta}_h(s,a)\right\|_\infty &= \left\|\left[\hat{P}_h \nu_{h+1}\right][s,a](\cdot - r_h(s,a)) - \left[P_h \nu_{h+1}\right][s,a](\cdot - r_h(s,a))\right\|_\infty \\
&= \left\|\left[\hat{P}_h \nu_{h+1}\right][s,a] - \left[P_h \nu_{h+1}\right][s,a]\right\|_\infty \\
&\leq \left\|\hat{P}_h(s,a) - P_h(s,a)\right\|_1 \leq c_h(s,a),
\end{aligned}$$

we have

$$\tilde{\eta}_h(s,a) = \mathrm{O}_{c_h(s,a)} \hat{\eta}_h(s,a) \succeq \check{\eta}_h(s,a).$$

We can prove the argument by induction. Fix $h+1 \in [2:H+1]$. Suppose $V_{h+1} = U_\beta\left(\eta_{h+1}\right) \geq U_\beta\left(\eta_{h+1}^*\right) = V_{h+1}^*$ for any s. It follows that

$$\begin{aligned}
Q_h(s,a) = U_\beta\left(\eta_h(s,a)\right) &= U_\beta\left(\Pi\tilde{\eta}_h(s,a)\right) = U_\beta\left(\tilde{\eta}_h(s,a)\right) = U_\beta\left(\mathrm{O}_{c_h(s,a)} \hat{\eta}_h(s,a)\right) \\
&\geq U_\beta\left(\check{\eta}_h(s,a)\right) = U_\beta\left(\mathcal{T}_h \nu_{h+1}\right) \\
&\geq U_\beta\left(\mathcal{T}_h \nu_{h+1}^*\right) = Q_h^*(s,a),
\end{aligned}$$

which implies $V_h(s) \geq V_h^*(s)$ for any s. The induction is completed.

Optimism of `RODI-PTO`. We rewrite the update of $q_h(s,a)$ in `RODI-PTO` as:

$$\begin{aligned}
&\hat{q}_h(s,a) \leftarrow [\hat{P}_h q_{h+1}](s,a), \hat{\eta}_h(s,a) = (L_h(s,a), R_h(s,a); \hat{q}_h(s,a)) \\
&\bar{q}_h(s,a) \leftarrow (1-\hat{q}_h(s,a))q_h^L(s,a) + \hat{q}_h(s,a)q_h^R(s,a), \bar{\eta}_h(s,a) = (L_h, R_h; \bar{q}_h(s,a)) \\
&q_h(s,a) \leftarrow \min(\bar{q}_h(s,a) + c_h(s,a), 1), \eta_h(s,a) = (L_h, R_h; q_h(s,a)) .
\end{aligned}$$

Define

$$\check{q}_h(s,a) \triangleq \left[P_h q_{h+1}\right][s,a], \quad \check{\eta}_h(s,a) \triangleq (L_h(s,a), R_h(s,a); \check{q}_h(s,a)) ,$$

then we have

$$\Pi\check{\eta}_h(s,a) = \left(L_h, R_h; (1-\check{q}_h(s,a))q_h^L(s,a) + \check{q}_h(s,a)q_h^R(s,a)\right) .$$

$\check{\eta}_h(s,a)$ and $\hat{\eta}_h(s,a)$ are both Bernoulli distributions with the same support, thus

$$\|\check{\eta}_h(s,a) - \hat{\eta}_h(s,a)\|_\infty = |\check{q}_h(s,a) - \hat{q}_h(s,a)| = \left|\left[(\hat{P}_h - P_h)q_{h+1}\right](s,a)\right| \le \left\|(\hat{P}_h - P_h)(s,a)\right\|_1 .$$

We have

$$\begin{aligned}
&\|\Pi\check{\eta}_h(s,a) - \Pi\hat{\eta}_h(s,a)\|_\infty = \\
&\left|(1-\check{q}_h(s,a))q_h^L(s,a) + \check{q}_h(s,a)q_h^R(s,a) - (1-\hat{q}_h(s,a))q_h^L(s,a) - \hat{q}_h(s,a)q_h^R(s,a)\right| \\
&= \left|(\check{q}_h(s,a) - \hat{q}_h(s,a))(q_h^R(s,a) - q_h^L(s,a))\right| \\
&= \left|[(\hat{P}_h - P_h)q_{h+1}](s,a)(q_h^R(s,a) - q_h^L(s,a))\right| \\
&= (q_h^R(s,a) - q_h^L(s,a)) \|\check{\eta}_h(s,a) - \hat{\eta}_h(s,a)\|_\infty \\
&\le (q_h^R(s,a) - q_h^L(s,a)) \left\|\hat{P}_h(s,a) - P_h(s,a)\right\|_1 \le (q_h^R(s,a) - q_h^L(s,a))c_h(s,a) < c_h(s,a).
\end{aligned}$$

Suppose $V_{h+1} = U_\beta\left(\eta_{h+1}\right) \ge U_\beta\left(\eta_{h+1}^*\right) = V_{h+1}^*$ for any s. Since $\eta_h(s,a) = \mathrm{O}_{c_h(s,a)}\Pi\hat{\eta}_h(s,a) \succeq \Pi\check{\eta}_h(s,a)$, we have

$$\begin{aligned}
Q_h(s,a) = U_\beta\left(\eta_h(s,a)\right) &= U_\beta\left(\mathrm{O}_{c_h(s,a)}\bar{\eta}_h(s,a)\right) = U_\beta\left(\mathrm{O}_{c_h(s,a)}\Pi_h\hat{\eta}_h(s,a)\right) \ge U_\beta\left(\Pi_h\check{\eta}_h(s,a)\right) \\
&= U_\beta\left(\check{\eta}_h(s,a)\right) = U_\beta\left(\left[\mathcal{T}_h\nu_{h+1}\right](s,a)\right) \ge U_\beta\left(\left[\mathcal{T}_h\nu_{h+1}^*\right](s,a)\right) = Q_h^*(s,a).
\end{aligned}$$

which implies $V_h(s) \ge V_h^*(s)$ for any s. The induction is completed.

6. RODI-MB

We introduce the **M**odel-**B**ased **R**isk-sensitive **O**ptimistic **D**istribution **I**teration algorithm (`RODI-MB`, cf. Algorithm 2). Unlike its model-free counterpart, `RODI-MB` explicitly maintains and updates an empirical transition model within each episode, making it a model-based approach. However, `RODI-MB` also encounters issues with computational inefficiency. Remarkably, `RODI-MB` is equivalent to a non-distributional RL algorithm (Algorithm 3). This equivalence results in computational efficiency, as it operates on one-dimensional values rather than full distributions.

Planning phase (Line 5-14) Mirroring the structure of Algorithm 1, RODI-MB also employs approximate DDP in conjunction with the OFU principle. Initially, it applies the distributional optimism operator to the empirical transition model $\hat{P}_h^k$, resulting in an optimistic transition model $\tilde{P}_h^k$. The algorithm then utilizes this optimistic model for the Bellman update, generating optimistic return distributions η_h^k. The subsequent steps remain consistent with those outlined in Algorithm 1.

Algorithm 2 RODI-MB

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

Interaction phase (Line 16-19) During the interaction phase, the agent engages with the environment using the policy π^k and updates the counts N_h^{k+1} and the empirical transition model $\hat{P}_h^{k+1}$ based on newly acquired observations.

6.1 Distributional Optimism over the Model

In the RODI-MB algorithm, we introduce a nuanced approach to generating an optimistic transition model, $\tilde{P}_h^k(s,a)$, from the empirical transition model, $\widehat{P}_h^k(s,a)$. This approach is based on the concept of distributional optimism over the space of PMFs rather than CDFs. Specifically, the goal is to compute a return distribution, η_h^k, from $\tilde{P}_h^k(s,a)$ and the future return ν_{h+1}^k, such that $\eta_h^k \geq \eta_h^*$ with high probability.

The distributional optimism operator, O_c^1, is defined for PMFs over the space $\mathscr{D}(\mathcal{S})$ with a level c, and it operates differently from O_c^∞ by also considering the future return distribution ν:

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

This operator selects a model from within the ℓ_1 norm ball, $B_1(\widehat{P}(s,a), c)$, that yields the largest EntRM value, $U_\beta([P\nu])$. This approach ensures that O_c^1 generates a model with optimistically biased estimates of the future returns, and it leverages an efficient method to achieve this (as detailed in Appendix C).

Given that Lemma 10 assures the high-probability event $\mathcal{G}_\delta$, the analysis primarily focuses on scenarios conditioned on $\mathcal{G}_\delta$. Additionally, due to the equivalence between EntRM and EU, the verification of optimism is conducted in terms of EU for $\beta > 0$.

Lemma 18 (Optimistic model) *For any (s,a,k,h) and $P \in B_1(\hat{P}_h^k(s,a), c_h^k(s,a))$, we have*

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

Proof Use the definition of O_c^1 and the equivalence between EntRM and EU. ■

Lemma 19 (Optimism) *Conditioned on event $\mathcal{G}_\delta$, the sequence $\{W_1^k(s_1^k)\}_{k\in[K]}$ produced by Algorithm 2 are all greater than or equal to $W_1^*(s_1^k)$, i.e.,*

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

The proof uses induction, paralleling the methodology in `RODI-MF`, and leverages Lemma 18 to ensure that the return distributions are optimistically biased.

Proof The induction begins with the terminal stage, H, and progresses backwards. For the visited (s,a), we have

$$J_H^k(s,a) = E_\beta(\eta_H^k(s,a)) = \exp(\beta r_H(s,a)) = J_H^*(s,a).$$

For the unvisited (s,a), it holds that $J_H^k(s,a) = \exp(\beta) \geq J_H^*(s,a)$. Thus $W_H^k(s) = \max_a J_H^k(s,a) \geq \max_a J_H^*(s,a) = W_H^*(s)$ for any s. Assuming $W_{h+1}^k(s) \geq W_{h+1}^*(s), \forall s$ for $h \in [H-1]$. It follows that for the (s,a) with $N_h^k(s,a) > 0$

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

The first inequality is due to Lemma 18. The second inequality follows from the induction assumption. For the unvisited (s,a), we have $J_h^k(s,a) = \exp(\beta(H+1-h)) \geq J_h^*(s,a)$. Since $W_h^k(s) = \max_a J_h^k(s,a) \geq \max_a J_h^*(s,a) = W_h^*(s)$ for any s, the induction is completed. ■

6.2 Equivalence to `ROVI`

The **R**isk-sensitive **O**ptimistic **V**alue **I**teration (`ROVI`) algorithm, as outlined in Algorithm 3, is a non-distributional approach that processes value functions directly, as opposed to handling return distributions. The `RODI-MB` algorithm, however, can be demonstrated to be equivalent to `ROVI`. This equivalence signifies that both algorithms generate the same policy sequence, implying that their resulting trajectories, denoted as $\mathcal{F}_{K+1}$, follow the same distribution. This relationship is grounded in the connection between the EntRM and

EU, coupled with the linearity property of EU. To formalize this concept of algorithmic equivalence, we define:

Algorithm 3 ROVI

```
Input: $T$ and $\delta$
$N_h^1(\cdot,\cdot) \leftarrow 0$; $\hat{P}_h^1(\cdot,\cdot) \leftarrow \frac{1}{S}\mathbf{1}\ \forall h \in [H]$
for $k = 1 : K$ do
    $W_{H+1}^k(\cdot) \leftarrow 1$
    for $h = H : 1$ do
        if $N_h^k(\cdot,\cdot) > 0$ then
            $\tilde{P}_h^k(\cdot,\cdot) \leftarrow \mathrm{O}^1_{c_h^k(\cdot,\cdot)}\left(\hat{P}_h^k(\cdot,\cdot), W_{h+1}^k\right)$
            $J_h^k(\cdot,\cdot) \leftarrow e^{\beta r_h(\cdot,\cdot)}\left[\tilde{P}_h^k W_{h+1}^k\right](\cdot,\cdot)$
        else
            $J_h^k(\cdot,\cdot) \leftarrow \exp(\beta(H+1-h))$
        end if
        $W_h^k(\cdot) \leftarrow \max_a J_h^k(\cdot,a)$
    end for
    Receive $s_1^k$
    for $h = 1 : H$ do
        $a_h^k \leftarrow \arg\max_a \mathrm{sign}(\beta) J_h^k(s_h^k, a)$ 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
```

Definition 20 *Recall that for an algorithm $\mathscr{A}_k$, $\mathscr{A}(\mathcal{F}_k) \in \Pi$ denotes the policy to be deployed in episode k. 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 $\mathscr{A}(\mathcal{F}_k) = \tilde{\mathscr{A}}(\mathcal{F}_k)$.*

Under this definition, if two algorithms are equivalent, the trajectories or histories generated by their interactions with any MDP instance will follow the same distribution throughout the episodes. Consequently, these algorithms will enjoy the same regret.

Proposition 21 *Algorithm 2 is equivalent to Algorithm 3.*

Proof We focus on the case where $\beta > 0$, noting that the case for $\beta < 0$ can be argued in a similar manner. Fix an arbitrary $k \in [K]$ and $\mathcal{F}_k = \{s_1^1, a_1^1, \cdots, s_H^{k-1}, a_H^{k-1}\}$. Let $\mathscr{A}$ (and π_h^k) represent Algorithm 3 (and its corresponding policy sequence), while $\tilde{\mathscr{A}}$ (and $\tilde{\pi}_h^k$) denote Algorithm 2 (and its respective policy sequence). To establish equivalence, we need to show that π^k aligns with $\tilde{\pi}^k$ for the given history $\mathcal{F}_k$. By the definition of the two algorithms

$$\tilde{\pi}_h^k(s) = \arg\max_a Q_h^k(s,a) = U_\beta(\eta_h^k(s,a)),\ \pi_h^k(s) = \arg\max_a J_h^k(s,a).$$

If $J_h^k(s,a) = E_\beta(\eta_h^k(s,a)) = \exp(\beta Q_h^k(s,a))$ for any (s,a), then $\pi_h^k = \tilde{\pi}_h^k$ due to the monotonicity of the exponential function. We will prove that $J_h^k(s,a) = E_\beta(\eta_h^k(s,a))$ for any

(s,a) by the induction. The base case is evident as $J_H^k(s,a) = E_\beta(\eta_H^k(s,a))$. Assuming $J_h^k(s,a) = E_\beta(\eta_h^k(s,a))$ for all (s,a) for some $h \in [H]$, we have $\pi_h^k = \tilde{\pi}_h^k$ and

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

Given the same history $\mathcal{F}_k$, both algorithms share the empirical transition model $\hat{P}_{h-1}^k$, the count N_{h-1}^k, and the optimism constants c_{h-1}^k. Therefore, they also share the optimistic transition model $\tilde{P}_{h-1}^k$. According to the update formula of Algorithm 3, for any (s,a) with $N_h^k(s,a) > 0$, we have

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

This equality also holds for the unvisited state-action pairs, thereby completing the proof of equivalence between Algorithm 2 and Algorithm 3. ■

6.3 Regret Upper Bound of `RODI-MB`/`ROVI`

Theorem 22 (Regret upper bound of `RODI-MB`/`ROVI`) *For any $\delta \in (0,1)$, with probability $1-\delta$, the regret of Algorithm 1 or Algorithm 3 is bounded as*

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

Proof The regret can be bounded as

$$\text{Regret}(K) \le \frac{1}{\beta}\sum_{k=1}^K W_1^k(s_1^k) - W_1^{\pi^k}(s_1^k) = \frac{1}{\beta}\sum_{k=1}^K \delta_1^k.$$

We can decompose δ_h^k as follows

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

Using the Lipschitz property of EU

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

where the second inequality is due to Lemma 11. Term (b) is bounded as

$$\begin{aligned}
(b) &= \exp(\beta r_h^k)\left[P_h^{\pi^k}(W_{h+1}^k - W_{h+1}^{\pi^k})\right](s_h^k) = \exp(\beta r_h^k)\left[P_h^{\pi^k}\Delta_{h+1}^k\right](s_h^k) \\
&= \exp(\beta r_h^k)(\epsilon_h^k + \delta_{h+1}^k),
\end{aligned}$$

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 $(k,h) \in [K] \times [H]$. In summary, we can bound δ_h^k recursively as

$$\delta_h^k \le L_{H+1-h} c_h^k + \exp(\beta r_h^k)(\epsilon_h^k + \delta_{h+1}^k).$$

Repeating the procedure, we can get

$$\begin{aligned}
\delta_1^k &\le \sum_{h=1}^{H-1} L_{H+1-h}\prod_{i=1}^{h-1}\exp(\beta r_i^k) c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h}\exp(\beta r_i^k)\epsilon_h^k + \prod_{i=1}^{H-1}\exp(\beta r_i^k)\delta_H^k \\
&\le \sum_{h=1}^{H-1}(\exp(\beta(H+1-h)) - 1)\exp(\beta(h-1)) c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h}\exp(\beta r_i^k)\epsilon_h^k + \exp(\beta(H-1))\delta_H^k \\
&\le \sum_{h=1}^{H-1}(\exp(\beta H) - 1) c_h^k + \sum_{h=1}^{H-1}\prod_{i=1}^{h}\exp(\beta r_i^k)\epsilon_h^k + \exp(\beta(H-1))\delta_H^k.
\end{aligned}$$

It follows that

$$\sum_{k=1}^{K} \delta_1^k \leq (\exp(\beta H) - 1) \sum_{k=1}^{K} \sum_{h=1}^{H-1} c_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H-1} \prod_{i=1}^{h} \exp(\beta r_i^k) \epsilon_h^k + \sum_{k=1}^{K} \exp(\beta(H-1)) \delta_H^k.$$

The following follows analogously: with probability at least $1-\delta$,

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

where $\iota \triangleq \log(2SAT/\delta)$. ■

Remark 23 *Compared to the traditional/non-distributional analysis dealing with scalars, our analysis is distribution-centered, and we call it the* distributional analysis*. The distributional analysis deals with the distributions of the return rather than the risk measure values of the return. In particular, 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 scalars (value functions).*

7. Regret Lower Bound

The section establishes a regret lower bound for EntRM-MDP, serving to understand the fundamental limitations of any learning algorithm in such settings. While previous works, like Fei et al. (2020), have approached this problem by drawing parallels to simpler models like the two-armed bandit, leading to lower bounds that are independent of S, A, and H, this approach does not capture the full complexity of MDPs.

In contrast, the approach motivated by Domingues et al. (2021) aims to derive a more comprehensive and tight minimax lower bound that incorporates these factors. For risk-neutral MDPs, the tight minimax lower bound has been established as $H\sqrt{SAT}$, but extending this to the risk-sensitive domain is challenging due to the non-linearity of EntRM. In risk-neutral scenarios, the linearity of expectation allows for the interchange of the risk measure (expectation) and summation, simplifying the analysis. However, in the risk-sensitive setting, the non-linear nature of EntRM precludes such straightforward manipulations, necessitating novel 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 24 (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*

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

Remark 25 *As $\beta \to 0$, it recovers the* tight *lower bound for risk-neutral episodic MDP $\Omega(H\sqrt{SAT})$ (see Domingues et al., 2021).*

Remark 26 *The two conditions in Assumption 1 are used in the our paper and Domingues et al. (2021) to simply the proof. Technically, we can relax these conditions to any MDP with $S \geq 11, A \geq 4$ and $H \geq 6$, which is modestly large. In particular, condition (i) allows us to consider a full A-ary tree with $S-3$ nodes, which implies that all the leaves are at the same level $d-1$ in the tree. The proof can be generalized to any $S \geq 6$ by arranging the states in a balanced, but not necessarily full, A-ary tree. We can also technically relax condition (ii) to the case $H < 3d$. In this case, the resulting bounds will replace S by $\left[A^{\frac{H}{3}-2}\right]$.*

Before presenting the proof of Theorem 24, we first fix the lower bound in Fei et al. (2020).

7.1 Fixing Lower Bound

Fei et al. (2020) presents the following lower bound.

Proposition 27 (Theorem 3, Fei et al. (2020)) *For sufficiently large K and H, the regret of any algorithm obeys*

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

However, a critical reassessment of the proof reveals inaccuracies that necessitate a revision of the lower bound. The main issue lies in the derivation of the second inequality in the proof provided by Fei et al. (2020), specifically:

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

The authors establish the second inequality based on the following fact

Fact 3 (Fact 5, Fei et al. (2020)) *For any $\alpha > 0$, the function $f_\alpha \triangleq \frac{e^{\alpha x}-1}{x}, x > 0$ is increasing and satisfies $\lim_{x\to 0} f_\alpha = \alpha$.*

In fact, we can only use Fact 3 to derive $\frac{\exp(\beta H/2)-1}{\beta} \gtrsim H$, which combined with the first inequality yields

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

which, notably, does not capture the dependency on β and H as the original lower bound suggested. Consequently, the corrected version of the lower bound is more conservative and does not reflect the exponential influence of the risk factor and the horizon on the regret. This corrected proposition reads:

Proposition 28 (Correction of Theorem 3, Fei et al. (2020)) *For sufficiently large K and H, the regret of any algorithm obeys*

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

7.2 Proof of Theorem 24

We define $\mathrm{kl}(p,q) \triangleq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ as the KL divergence between two Bernoulli distributions with parameters p and q. We define the probability measure induced by an algorithm $\mathscr{A}$ and an MDP instance $\mathcal{M}$ as

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

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

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

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 $\mathcal{M}$. We omit $\mathscr{A}$ and $\mathcal{M}$ if it is clear in the context.

Proof 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 $\tilde{H} \triangleq \bar{H}+d+1$ and $H' \triangleq H+1-\tilde{H}$. We denote by $\mathcal{L} \triangleq \{s_1, s_2, ..., s_{\bar{L}}\}$ the set of $\bar{L}$ leaves of the tree. For each $u^* \triangleq (h^*, \ell^*, a^*) \in [d+1:\bar{H}+d] \times \mathcal{L} \times \mathcal{A}$, 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

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

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

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

where $\Delta_{u^*}(h, s_i, a) \triangleq \epsilon\mathbb{I}\{(h, s_i, a) = (h^*, s_{\ell^*}, a^*)\}$ 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 $\forall a, P_h(s_g \mid s_g, a) \triangleq P_h(s_b \mid s_b, a) \triangleq 1$. The reward is a deterministic function of the state

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

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

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

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 $x_h^k\triangleq(s_h^k,a_h^k)$ for each (k,h) and $x^*\triangleq(s_{\ell^*},a^*)$, then it is not hard to obtain that

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

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 $p_{u^*}^k\triangleq\mathbb{P}_{\pi^k,u^*}(x_{h^*}^k=x^*)$, then the expected regret of $\mathscr{A}$ in $\mathcal{M}_{u^*}$ can be bounded below as

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

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 $N_K(u^*)\triangleq\sum_{k=1}^{K}\mathbb{I}\{x_{h^*}^k=x^*\}$.

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

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

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

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

By Pinsker's inequality, it implies that

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

Since $\mathcal{M}_0$ and $\mathcal{M}_{u^*}$ only differs at stage h^* when $(s,a) = x^*$, it follows from Fact 6 that

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

By Lemma 33, we have $\operatorname{kl}(p, p+\epsilon) \leq \frac{\epsilon^2}{p}$ 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 the fact that $\sum_{u^*\in[d+1:\bar{H}+d]\times\mathcal{L}\times\mathcal{A}} N_K(u^*) = K$.
It follows that

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

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

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

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 $\epsilon = \sqrt{\frac{p}{2}}(1-\frac{1}{\bar{L}A\bar{H}})\sqrt{\frac{LA\bar{H}}{K}} \leq \frac{1}{2\sqrt{2}}\exp(-\beta H'/2)\sqrt{\frac{\bar{L}A\bar{H}}{K}} \leq \frac{1}{4}\exp(-\beta H')$ if $K \geq 2\exp(\beta H')\bar{L}A\bar{H}$. Such choice of p and ϵ guarantees the assumption of Lemma 33 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

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

■

Theorem 24 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.

8. Discussion

In this section, we provide a comprehensive comparison of DRL algorithms (RODI-MB, RODI-MF), DRL with distribution representation (RODI-OTP, RODI-PTO), RSVI2 (Fei et al., 2021), RSVI (Fei et al., 2020), and UCBVI (Azar et al., 2017) in terms of regret guarantees and computational complexity. The comparison is also succinctly encapsulated in Table 1.

8.1 Numerical Results

To validate the empirical performance of our algorithms, we conducted numerical experiments comparing RODI-MB, RODI-MF, and RODI-Rep with the risk-neutral algorithm UCBVI (Azar et al., 2017), RSVI in Fei et al. (2020), and RSVI2 in Fei et al. (2021).

The experimental setup involved an MDP with $S = 5$ states, $A = 5$ actions, and a horizon $H = 5$, mirroring the setup in Du et al. (2022). The MDP consists of a fixed initial state denoted as state 0, and S additional states. The agent started in state 0 and could take actions from the set $[A]$, transitioning to one of the states in $[S]$ in the next step. The transition probabilities and reward functions were defined as follows for $2 \leq h \leq H$:

$$\forall a \in [A-1] : P_h(s'|s,a) = \frac{0.5}{S-2} \ \forall s' \in [2 : S-1], P_h(1|s,a) = 0.5,$$
$$P_h(s'|s,A) = \frac{0.001}{S-1}, \forall s' \in [S-1], P_h(S|s,A) = 0.999$$
$$\forall a \in [A] : r_h(1,a) = 1, r_h(S,a) = 0.4, r_h(s,a) = 0 \ \forall s \in [2 : S-1].$$

This MDP was designed to be highly risky, with the risk-neutral optimal policy leading to a mean reward of 0.5 but with a chance of receiving no reward. A risk-sensitive policy might prefer the last action A, which offers slightly less mean reward but a more consistent return, indicating lower risk.

We set $\delta = 0.005$ and $\beta = -1.1$. The results, as illustrated in Figure 1, demonstrates the regret ranking of these algorithms :

$$\underbrace{\texttt{RODI-MB} < \texttt{RODI-MF}}_{\texttt{RODI}} < \underbrace{\texttt{RODI-OTP} < \texttt{RODI-PTO}}_{\texttt{RODI-Rep}} \lesssim \texttt{RSVI2} < \texttt{RSVI} < \texttt{UCBVI}.$$

Figure 1 includes the following key observations:
(i) Advantage of distributional over non-Distributional algorithms: DRL algorithms (RODI and RODI-Rep) outperforms non-distributional algorithms, demonstrating the effectiveness of distributional optimism over bonus-based optimism.
(ii) Performance of RODI vs. RODI-Rep: While RODI shows better performance than RODI-Rep, the latter offers a balance between statistical and computational efficiency.
(iii) Comparison of RODI-Rep with RSVI2: RODI-Rep demonstrates advantages over RSVI2 in terms of sample efficiency, while also maintaining computational efficiency.

8.2 Theoretical Comparisons

8.2.1 RODI vs. RSVI2

We first provide theoretical justifications regarding the regret ranking of RSVI (Fei et al., 2020), RSVI2 (Fei et al., 2021), RODI-MF, and RODI-MB, which demonstrates the advantage

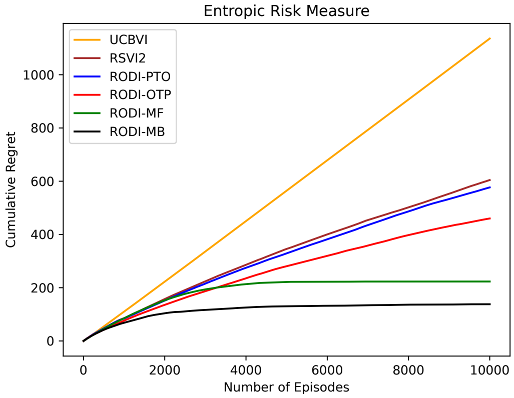

Figure 1: Comparison of regret for different algorithms.

of distributional optimism over bonus-based optimism used in $\texttt{RSVI}$ and $\texttt{RSVI2}$. A key observation regarding the ranking of their value functions V^k is that:

$$\text{value functions} : \texttt{RSVI} > \texttt{RSVI2} > \texttt{RODI-MF} > \texttt{RODI-MB} \geq V^*.$$

This ordering will be formally presented in Equation 6. The last part of this inequality sequence indicates that all these value functions are indeed optimistic. Given that the level of optimism is mirrored in the value functions, we can deduce:

$$\text{optimism level} : \texttt{RSVI} > \texttt{RSVI2} > \texttt{RODI-MF} > \texttt{RODI-MB}.$$

Considering the relationship between regret and the optimistic value function V^k

$$\text{Regret} = \sum_{k\in[K]} V_1^* - V_1^{\pi^k} \leq \sum_{k\in[K]} V_1^k - V_1^{\pi^k},$$

it is intuitive that a smaller V^k or less optimism induces reduced regret. Consequently, their regret can be ranked as:

$$\text{regret} : \texttt{RSVI} > \texttt{RSVI2} > \texttt{RODI-MF} > \texttt{RODI-MB},$$

which explains Figure 1. The regret bounds of $\texttt{RODI}$ should at least match those of $\texttt{RSVI2}$, explaining the ranking of their regret bounds reported in Table 1:

$$\text{regret bound} : \texttt{RSVI} > \texttt{RSVI2} = \texttt{RODI-MF} = \texttt{RODI-MB}.$$

Despite sharing same regret bounds with $\texttt{RSVI2}$, $\texttt{RODI}$ outperforms $\texttt{RSVI2}$ both theoretically and empirically. Formally speaking, let V', V'', V denote the value functions generated by $\texttt{RSVI}$, $\texttt{RSVI2}$, and $\texttt{RODI-MB}$ respectively. Let $\tilde{\eta}$ denote the distribution generated by $\texttt{RODI-MF}$. We omit k for simplicity.

Proposition 29 *Fix (s, a, k, h). The comparison of their values is as follows:*

$$\begin{aligned}
\texttt{RSVI} \quad \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V'_{h+1}}\right] + b'_h\right) &\overset{(a)}{>} \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b'_h\right) \\
&\overset{(b)}{>} \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b''_h\right) \quad \texttt{RSVI2} \\
&\overset{(c)}{>} U_\beta\left(\tilde{\eta}_h\right) \quad \texttt{RODI-MF} \qquad (6) \\
&\overset{(d)}{>} \frac{1}{\beta}\log\left(\left[\tilde{P}_h e^{\beta V_{h+1}}\right]\right) \quad \texttt{RODI-MB} \\
&\overset{(e)}{>} \frac{1}{\beta}\log\left(\left[P_h e^{\beta V^*_{h+1}}\right]\right).
\end{aligned}$$

The proof is detailed in Appendix B. Both `RSVI` and `RSVI2` use exploration bonuses, defined as $b'_h = |e^{\beta H} - 1|c_h$ and $b''_h = |e^{\beta(H+1-h)} - 1|c_h$ respectively, where $c_h(s, a)$ represents the model estimation error

$$\left\|\hat{P}_h(s, a) - P_h(s, a)\right\|_1 \le c_h(s, a) = \sqrt{\frac{S\iota}{N_h(s, a)}}.$$

Both b''_h and b'_h are formulated as a multiplier times c_h. Notably, b''_h, referred to as the the *doubly decaying bonus* (Fei et al., 2021), decreases its multiplier exponentially across stages h, contrasting with b'_h in `RSVI`. In comparison, `RODI` directly incorporates optimism into the return distribution using an optimism constant c_h. Our *distributional analysis* establishes a connection between c_h and the bonus via the Lipschitz constant of EU:

$$b''_h = L(E_\beta, H - h)c_h < L(E_\beta, H)c_h = b'_h,$$

where $L(E_\beta, M)$ denotes the Lipschitz constant of EU over the distributions supported in $[0, M]$. This distributional perspective posits that `RSVI` and `RSVI2` design bonuses to offset the error in value estimates, which is bounded by the product of the Lipschitz constant of EU and the error in the return distribution:

$$V^k_h - V_h \le L(E_\beta, H - h)\left\|\eta^k_h - \eta_h\right\| \le L(E_\beta, H - h)\left\|P^k_h - P_h\right\| \le L(E_\beta, H - h)c^k_h.$$

Under the distributional perspective, the multiplier in the bonus b''_h is interpreted as the Lipschitz constant that links the return estimation error c_h to the value estimation error b''_h. The Lipschitz constant decreases exponentially in h as the range $[0, H - h]$ of the return distribution narrows. Furthermore, b''_h used in `RSVI2` is not improvable in the sense that its corresponding Lipschitz constant is proven to be tight, as shown in Lemma 6.

In conclusion, bonus-based optimism requires an exponentially decaying multiplier or Lipschitz constant, whereas distributional optimism functions directly at the distributional level, obviating the need for a multiplier. Next, we theoretically justify the regret ranking of `RODI-OTP` and `RODI-PTO`, which interpolates between `RODI` and `RSVI2`.

8.2.2 `RODI-Rep` vs. `RSVI2`

We delve into the analysis by first explaining why `RODI-PTO` achieves marginally lower regret compared to `RSVI2`, and subsequently, we justify the advantage of `RODI-OTP` over `RODI-PTO`.

Near-equivalence between `RSVI2` and `RODI-PTO`. We can show the near-equivalence between `RSVI2` and `RODI-PTO` using induction. Let V and V' denote the value functions generated by `RODI-PTO` and `RSVI2` respectively. We start with the base case that $h = H$. By the construction of `RODI-PTO`, we have

$$
\begin{aligned}
& q_H(s,a) = q(r_H(s,a); 0, 1) \Longrightarrow \\
& Q_H(s,a) = \frac{1}{\beta}\log\left((1-q_H(s,a))e^0 + q_H(s,a)e^{\beta}\right) = r_H(s,a) = Q'_H(s,a) \Longrightarrow \\
& V_H(s) = \max_a Q_H(s,a) = \max_a Q'_H(s,a) = V'_H(s),
\end{aligned}
$$

verifying the equivalence at step H. Now fix $h \in [H-1]$. Suppose the following holds

$$
\begin{aligned}
V_{h+1}(s) = \frac{1}{\beta}\log\left(1 - q_{h+1}(s) + q_{h+1}(s)e^{\beta(H-h)}\right) \leq V'_{h+1}(s), \forall s \in \mathcal{S} \Longrightarrow \\
1 - q_{h+1}(s) + q_{h+1}(s)e^{\beta(H-h)} \leq e^{\beta V'_{h+1}(s)}, \forall s \in \mathcal{S}.
\end{aligned}
$$

Recall the recursion of $q_h(s,a)$ in `RODI-PTO`

$$
\begin{aligned}
& \hat{q}_h(s,a) \leftarrow [\hat{P}_h q_{h+1}](s,a) \\
& \bar{q}_h(s,a) \leftarrow (1-\hat{q}_h(s,a))q_h^L(s,a) + \hat{q}_h(s,a)q_h^R(s,a) \\
& q_h(s,a) \leftarrow \min(\bar{q}_h(s,a) + c_h(s,a), 1).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
Q_h(s,a) &= \frac{1}{\beta}\log\left((1-q_h(s,a))e^0 + q_h(s,a)e^{\beta(H+1-h)}\right) \\
&= \frac{1}{\beta}\log\left(1 + q_h(s,a)(e^{\beta(H+1-h)} - 1)\right) \\
&\leq \frac{1}{\beta}\log\left(1 + \bar{q}_h(s,a)(e^{\beta(H+1-h)} - 1) + c_h(s,a)(e^{\beta(H+1-h)} - 1)\right),
\end{aligned}
$$

where the last inequality becomes equality if $\bar{q}_h(s,a) + c_h(s,a) \leq 1$. By the definition of projection, we obtain

$$
\begin{aligned}
1 + \bar{q}_h(s,a)(e^{\beta(H+1-h)} - 1) &= 1 - \bar{q}_h(s,a) + \bar{q}_h(s,a)e^{\beta(H+1-h)} \\
&= (1-\hat{q}_h(s,a))e^{\beta r_h(s,a)} + \hat{q}_h(s,a)e^{\beta(r_h(s,a)+H-h)} \\
&= [\hat{P}_h(1-q_{h+1})](s,a)e^{\beta r_h(s,a)} + [\hat{P}_h q_{h+1}](s,a)e^{\beta(r_h(s,a)+H-h)} \\
&= \sum_{s'} \hat{P}_h(s'|s,a)\left((1-q_{h+1}(s'))e^{\beta r_h(s,a)} + q_{h+1}(s')e^{\beta(r_h(s,a)+H-h)}\right) \\
&= e^{\beta r_h(s,a)} \sum_{s'} \hat{P}_h(s'|s,a)\left((1-q_{h+1}(s')) + q_{h+1}(s')e^{\beta(H-h)}\right) \\
&= e^{\beta r_h(s,a)} \sum_{s'} \hat{P}_h(s'|s,a)e^{\beta V_{h+1}(s')} \\
&\leq e^{\beta r_h(s,a)} \sum_{s'} \hat{P}_h(s'|s,a)e^{\beta V'_{h+1}(s')},
\end{aligned}
$$

which implies

$$Q_h(s,a) \leq \frac{1}{\beta}\log\left(e^{\beta r_h(s,a)}\sum_{s'}\hat{P}_h(s'|s,a)e^{\beta V'_{h+1}(s')} + c_h(s,a)(e^{\beta(H+1-h)}-1)\right) = Q'_h(s,a).$$

Then we have $V_h(s) = \max_a Q_h(s,a) \leq \max_a Q'_h(s,a) = V'_h(s)$. The induction is completed. Moreover, it holds that $V_h = V'_h$ for every $h \in [H]$ if $\bar{q}_h(s,a) + c_h(s,a) \leq 1$ for every (h,s,a). This condition is likely to be met for large values of k, considering that

$$k \uparrow \Longrightarrow N_h^k \downarrow \Longrightarrow c_h^k \propto 1/\sqrt{N_h^k} \downarrow .$$

Advantage of `RODI-OTP` over `RSVI2` Let V and V' denote the value functions generated by `RODI-OTP` and `RSVI2` respectively. The recursion of $q_h(s,a)$ in `RODI-OTP` writes

$$\begin{aligned}
&\hat{q}_h(s,a) \leftarrow [\hat{P}_h q_{h+1}](s,a)\\
&\tilde{q}_h(s,a) \leftarrow \min(\hat{q}_h(s,a) + c_h(s,a), 1)\\
&q_h(s,a) \leftarrow (1-\tilde{q}_h(s,a))q_h^L(s,a) + \tilde{q}_h(s,a)q_h^R(s,a).
\end{aligned}$$

Fix $(h,s,a) \in [H-1]\times\mathcal{S}\times\mathcal{A}$. Note that

$$\begin{aligned}
V_{h+1}(s) &= \frac{1}{\beta}\log\left(1 - q_{h+1}(s) + q_{h+1}(s)e^{\beta(H-h)}\right), \forall s \in \mathcal{S},\\
\Longrightarrow [\hat{P}_h e^{\beta V_{h+1}}](s,a) &= (1-\hat{q}_h(s,a)) + \hat{q}_h(s,a)e^{\beta(H-h)}, \forall (s,a),
\end{aligned}$$

then we have

$$\begin{aligned}
Q_h(s,a) &= \frac{1}{\beta}\log\left(1 - q_h(s,a) + q_h(s,a)e^{\beta(H+1-h)}\right)\\
&= \frac{1}{\beta}\log\left((1-\tilde{q}_h(s,a))e^{\beta r_h(s,a)} + \tilde{q}_h(s,a)e^{\beta(r_h(s,a)+H-h)}\right)\\
&\leq \frac{1}{\beta}\log\left((1-\hat{q}_h(s,a))e^{\beta r_h(s,a)} + \hat{q}_h(s,a)e^{\beta(r_h(s,a)+H-h)} + c_h(s,a)(e^{\beta(r_h(s,a)+H-h)} - e^{\beta r_h(s,a)})\right)\\
&= \frac{1}{\beta}\log\left(e^{\beta r_h(s,a)}[\hat{P}_h e^{\beta V_{h+1}}](s,a) + c_h(s,a)e^{\beta r_h(s,a)}(e^{\beta(H-h)}-1)\right)\\
&< \frac{1}{\beta}\log\left(e^{\beta r_h(s,a)}[\hat{P}_h e^{\beta V'_{h+1}}](s,a) + c_h(s,a)e^{\beta r_h(s,a)}(e^{\beta(H-h)}-1)\right)\\
&< \frac{1}{\beta}\log\left(e^{\beta r_h(s,a)}[\hat{P}_h e^{\beta V'_{h+1}}](s,a) + c_h(s,a)(e^{\beta(H+1-h)}-1)\right) = Q'_h(s,a).
\end{aligned}$$

Remark 30 *This explains why `RODI-OTP` achieves an order of magnitude improvement in regret compared with `RSVI2` as well as `RODI-PTO`, as the "optimism level ratio" of `RODI-OTP` to `RSVI2` at step h is quantifiable by*

$$\frac{e^{\beta(r_h(s,a)+H-h)} - e^{\beta r_h(s,a)}}{e^{\beta(H+1-h)}-1} < 1.$$

Remark 31 *The difference in the optimism level between the two algorithms stems from originates from their respective approaches to bounding the estimation error:*

$$\sum_{s'} \hat{P}_h(s'|s,a) e^{\beta(r_h(s,a)+V'_{h+1}(s'))}.$$

Specifically, `RSVI2` *treats* $e^{\beta(r_h(s,a)+V'_{h+1})}$ *as a variable within the range* $[1, e^{\beta(H+1-h)}]$*. However, since* $e^{\beta r_h(s,a)}$ *is deterministic and known, the bonus can be refined by acknowledging*

$$\sum_{s'} \hat{P}_h(s'|s,a) e^{\beta(r_h(s,a)+V'_{h+1}(s'))} = e^{\beta(r_h(s,a)}[\hat{P}_h e^{\beta V_{h+1}}](s,a),$$

where $e^{\beta V_{h+1}} \in [1, e^{\beta(H-h)}]$.

Why `OTP` is better than `PTO`. The superiority of `OTP` over `PTO` can be substantiated through an insightful observation about the optimization problem:

$$\begin{aligned} \min_q \quad & U_\beta\left(L, R; q\right) \\ \text{s.t.} \quad & U_\beta\left(L, R; q\right) \geq U_\beta\left(\eta\right) \\ & \left\|\eta - \hat{\eta}\right\|_\infty \leq c \\ & \eta = D(\mathrm{Supp}(\hat{\eta})) \end{aligned} \tag{7}$$

Let $(L, R; \tilde{q})$ be the optimal solution to this problem. It turns out that the optimal solution is given by $(L, R; \tilde{q}) = \Pi \mathrm{O}_c \hat{\eta}$, aligning with the `OTP` principle. Fixing (h, s, a), we interpret $\hat{\eta} \triangleq [\hat{\mathcal{T}}_h \nu_{h+1}](s,a)$ as the empirical Bellman operator applied to ν_{h+1}. Suppose ν_{h+1} is optimistic relative to the true distribution ν^*_{h+1}, i.e., $U_\beta\left(\nu_{h+1}\right) \geq \nu^*_{h+1}$. Define $\check{\eta} \triangleq [\mathcal{T}_h \nu_{h+1}](s,a)$, which is the exact Bellman operator applied to ν_{h+1}. Given that

$$\left\|\hat{\eta} - \check{\eta}\right\|_\infty = \left\|[(\hat{\mathcal{T}}_h - \mathcal{T}_h)\nu_{h+1}](s,a)\right\|_\infty \leq c_h(s,a),$$

the optimal solution satisfies

$$U_\beta\left(L, R; \tilde{q}\right) \geq U_\beta\left(\check{\eta}\right) = U_\beta\left([\hat{\mathcal{T}}_h \nu_{h+1}](s,a)\right) \geq U_\beta\left([\hat{\mathcal{T}}_h \nu^*_{h+1}](s,a)\right) = U_\beta\left(\eta^*_h(s,a)\right) = Q^*_h(s,a).$$

Hence, the optimal solution $(L, R; \tilde{q})$ is optimistic over $\eta^*_h(s,a)$. The nature of the optimization problem compels $(L, R; \tilde{q})$ to be the Bernoulli distribution with support (L_h, R_h) that necessitates minimal optimism over $\eta^*_h(s,a)$. Notably, the `PTO` solution $\mathrm{O}_c \Pi \hat{\eta}$ is also a feasible solution. Consequently, `OTP` induces less optimism than `PTO`:

$$U_\beta\left(\Pi \mathrm{O}_c \hat{\eta}\right) < U_\beta\left(\mathrm{O}_c \Pi \hat{\eta}\right).$$

This analysis elucidates the inherent advantage of the `OTP` approach over `PTO`. By inverting the order of the projection and optimism operators, `OTP` not only ensures an optimism over the true distribution but also guarantees that the induced optimism is minimal and necessary.

8.3 Distributional Perspective

The distributional perspective is crucial in both the algorithm design and the regret analysis of RODI, offering advantages and novel approaches.

8.3.1 Algorithm design

Revisiting RSVI2: RSVI2 effectively operates as a model-based algorithm, implicitly maintaining an empirical model through visiting counts. We rewrite the key step in RSVI2 as:

$$Q''_h = \min\left\{H+1-h, r_h + \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b''_h\right)\right\}.$$

Here, b''_h is chosen to ensure optimism:

$$\left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b''_h \geq \left[P_h e^{\beta V''_{h+1}}\right] \geq \left[P_h e^{\beta V^*_{h+1}}\right] \Longleftarrow b''_h \geq \left[(P_h - \hat{P}_h) e^{\beta V''_{h+1}}\right]$$

Distributional perspective: In contrast, the distributional perspective leads to a fundamentally different algorithm design. The primary distinction of RODI is its implementation of return distribution iterations based on approximate distributional Bellman equation. When $\beta \to 0$, RODI transitions to a risk-neutral algorithm, unlike RSVI2, where the log term becomes constant. RODI also introduces *distributional optimism*, yielding optimistic return distributions without needing a multiplier, unlike bonus-based optimism. This approach not only contrasts sharply with bonus-based methods but also demonstrates improved theoretical and empirical performance.

8.3.2 Regret analysis

Our regret analysis, which we term *distributional analysis*, stands apart from traditional scalar-focused approaches. This analysis is centered around the distributions of returns rather than the risk values of these returns. It involves various distributional operations, including understanding the optimism between different distributions and the errors caused by distribution estimation. These elements fundamentally differ from classical analysis methods that focus on scalars (value functions). Let's highlight some novel aspects of our distributional analysis compared to traditional approaches (Fei et al., 2020, 2021).

(i) Distributional optimism. Traditional analysis typically employs OFU to construct a series of optimistic value functions. In contrast, our distributional approach implements optimism directly at the distribution level, leading to a sequence of optimistic return distributions. This involves defining a high probability event under which the true return distribution is close to the estimated one within a certain confidence radius, followed by the application of a distributional optimism operator.

(ii) Lipschitz continuity and linearity in EU. We leverage key properties of EU, such as Lipschitz continuity and linearity, that are crucial in establishing regret upper bounds. The Lipschitz continuity of EU relates the distance between distributions to their EU values' difference. In contrast, EntRM is non-linear w.r.t. the distribution, potentially introducing a factor of $\exp(|\beta|H)$ in error propagation across time steps, leading to a compounded factor

of $\exp(|\beta|H^2)$ in the regret bound.
(iii) Better interpretability. Both `RODI` and `RSVI2` share a same regret bound of

$$\tilde{\mathcal{O}}\left(\frac{\exp(|\beta|H)-1}{|\beta|}H\sqrt{S^2AK}\right).$$

From the distributional perspective, the exponential term $\frac{\exp(|\beta|H)-1}{|\beta|}$ is interpreted as the Lipschitz constant of EntRM, highlighting the impact of EntRM's nonlinearity on sample complexity. A larger Lipschitz constant implies a greater estimation error in values, thus leading to a more unfavorable regret bound.

8.3.3 Applicability of General Risk Measures

Our decision to focus on EntRM is primarily driven by its computational tractability and its effectiveness in representing risk preferences within decision-making frameworks. A recent study (Marthe et al., 2023), which postdates our work, establishes **(MP)** as a necessary condition for DDP. They further ascertain that EntRM is the only continuous risk measure that facilitates DDP, making it the optimal choice in terms of computational feasibility.

Furthermore, EntRM's ability to balance the mean and variance of returns provides a nuanced approach to risk that is especially relevant in environments where understanding the trade-offs between violation and return is critical. In addition, it aligns well with the *exponential utility functions* used in economic theory, providing a foundation in established risk-sensitive models.

We note a recent study (Chen et al., 2024) on applying DRL for general Lipschitz risk measures, which proposes DRL algorithms with sublinear regret bounds. While inspired by our distributional perspective, these algorithms still face challenges regarding computational tractability. In future work, we aim to explore how to design computationally efficient and statistically optimal DRL algorithms for general risk measures.

9. Closing Remarks

In this paper, we present a distributional dynamic programming framework for RSRL. We then introduce two types of computationally efficient DRL algorithms, which implement the OFU principle at the distributional level to strike a balance between exploration and exploitation under the risk-sensitive setting. We provide theoretical justification and numerical results demonstrating that these algorithms outperforms existing methods while maintaining computational efficiency compared. Furthermore, we prove that DRL can attain near-optimal regret upper bounds compared with our improved lower bound.

Looking forward, there are several promising avenues for future research. Our current regret upper bound has an additional factor of $\sqrt{HS}$ compared to the lower bound, and it may be possible to eliminate this factor through further algorithmic improvements or refined analysis techniques. Additionally, extending the DRL algorithm from tabular MDP to function approximation settings would be an interesting and valuable direction for future investigation. Lastly, it would be worthwhile to explore how to design computationally efficient and statistically optimal DRL algorithms for general risk measures.

Appendix A. Table of Notation

Symbol	Explanation
$\mathscr{D}$	The space of all CDFs
$\mathscr{D}(a,b)$	The space of all CDFs supported on $[a,b]$
$\mathscr{D}_M$	The space of all CDFs supported on $[0,M]$
$B_\infty(F,c)$	The $\|\cdot\|_\infty$ norm ball centered at F with radius c
δ_c	the step function with parameter c
L_M	The Lipschitz constant of EntRM w.r.t. $\infty\cdot_\infty$ over $\mathscr{D}_M$
$\mathbf{O}_{\boldsymbol{c}}^{\boldsymbol{\infty}}$	The optimism operator w.r.t. $\|\cdot\|_\infty$ with coefficient c
$\mathcal{M}$	MDP instance
$\mathcal{S}$	finite state space
$\mathcal{A}$	finite action space
r_h	deterministic function at step h
S	number of states
A	number of actions
H	Number of time-steps per episode
K	Number of episodes
$Z_h^\pi(s,a)$	return of (s,a) at step h with policy π
$Y_h^\pi(s)$	return of s at step h with policy π
$Z_h^*(s,a)$	optimal return of (s,a) at step h
$Y_h^*(s)$	optimal return of s at step h
$\eta_h^\pi(s,a)$	distribution of $Z_h^\pi(s,a)$
$\nu_h^\pi(s)$	distribution of $Y_h^\pi(s)$
$\eta_h^*(s,a)$	distribution of $Z_h^*(s,a)$
$\nu_h^*(s)$	distribution of $Y_h^*(s)$
$Q_h^\pi(s,a)$	EntRM value of $Z_h^\pi(s,a)$
$V_h^\pi(s)$	EntRM value of $Y_h^\pi(s)$
$Q_h^*(s,a)$	EntRM value of $Z_h^*(s,a)$
$V_h^*(s)$	EntRM value of $Y_h^*(s)$
$J_h^\pi(s,a)$	EU value of $Z_h^\pi(s,a)$
$W_h^\pi(s)$	EU value of $Y_h^\pi(s)$
$J_h^*(s,a)$	EU value of $Z_h^*(s,a)$
$W_h^*(s)$	EU value of $Y_h^*(s)$
$\mathcal{H}_h^k$	history up to step h of episode k
$\mathcal{F}_k$	history up to episode $k-1$
$\mathscr{A}$	RL algorithm
π	policy
N^k	visiting count
$\hat{P}^k$	empirical transition function in episode k
$\mathcal{T}$	distributional Bellman operator
Π	projection operator
$(x_1,x_2;p)$	a distribution taking values x_1,x_2 with probability $1-p$ and p
$(x;p)$	a discrete distribution with $\mathbb{P}(X=x_i)=p_i$.
$\|\eta\|$	the number of atoms of the distribution η

Appendix B. Missing Proofs

B.1 Missing Proofs in Section 4

Proof of Lemma 6

Proof We first provide the proof for the case $\beta > 0$. 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.

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

For the case $\beta < 0$, we assume $\int_0^M G(x)d\exp(\beta x) - \int_0^M F(x)d\exp(\beta x) \geq 0$.

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

Thus $L_M = |\exp(\beta M) - 1|$ for EU. To show the tightness of the constant, consider two scaled Bernoulli distributions $F = (1 - \mu_1)\psi_0 + \mu_1\psi_M$ and $G = (1 - \mu_2)\psi_0 + \mu_2\psi_M$, where $\mu_1, \mu_2 \in (0, 1)$ are some constants. It holds that

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

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

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

■

Proof of Lemma 10

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

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

Lemma 10 does not directly follow from a union bound together with Fact 4 since the case $N_h^k(s,a)=0$ need to be checked.

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

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

It follows that

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

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

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

Thus,

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

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

Lemma 32 *Let* $0<m\leq a<b$, *it holds that* $\log(b)-\log(a)\leq\frac{1}{m}(b-a)$.

B.2 Missing Proofs in Section 6

B.3 Missing Proofs in Section 7

Proof of Lemma 33

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

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

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 $1 - p \geq \frac{1}{2}$. ■

Lemma 33 *If $\epsilon \geq 0$, $p \geq 0$ and $p + \epsilon \in [0, \frac{1}{2}]$, then $\mathrm{kl}(p, p+\epsilon) \leq \frac{\epsilon^2}{2p(1-p)} \leq \frac{\epsilon^2}{p}$.*

Fact 5 (Lemma 1, Garivier et al. (2019)) *Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0, 1]$, we have*

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

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

Fact 6 (Lemma 5, Domingues et al. (2021)) *Let $\mathcal{M}$ and $\mathcal{M}'$ be two MDPs that are identical except for their transition probabilities, denoted by P_h and P'_h, respectively. Assume that we have $\forall(s, a)$, $P_h(\cdot \mid s, a) \ll P'_h(\cdot \mid s, a)$. Then, for any stopping time τ with respect to $(I_k)_{k\geq 1}$ that satisfies $\mathbb{P}_{\mathcal{M}}[\tau < \infty] = 1$*

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

B.4 Missing Proofs in Section 8

Proof of Proposition 29

Proof Recall that

$$\mathrm{EU}\left(x, P\right) = \sum_{i\in[n]} e^{\beta x_i} P_i = \left[P \circ e^{\beta x}\right]$$

We can prove the above inequalities by induction. We only show the proof for $\beta > 0$. Assume that

$$V'_{h+1} \geq V''_{h+1} \geq U_\beta\left(\tilde{\nu}_{h+1}\right) \geq V_{h+1} \geq V^*_{h+1}.$$

$(a) \Longleftarrow$ induction $V'_{h+1} > V''_{h+1}$

$(b) \Longleftarrow b''_h = |e^{\beta(H+1-h)} - 1|c_h < |e^{\beta H} - 1|c_h = b'_h$

$(c) \Longleftarrow U_\beta\left(\tilde{\nu}_{h+1}(s)\right) \leq V''_{h+1}(s)$ for all $s \in \mathcal{S}$ is equivalent to $\mathrm{EU}\left(\tilde{\nu}_{h+1}(s)\right) \leq e^{\beta V''_{h+1}(s)}$. Given the linearity of EU, we have

$$\mathrm{EU}\left(\left[\hat{P}_h \tilde{\nu}_{h+1}\right]\right) - \left[\hat{P}_h e^{\beta V''_{h+1}}\right] = \left[\hat{P}_h \mathrm{EU}\left(\tilde{\nu}_{h+1}\right)\right] - \left[\hat{P}_h e^{\beta V''_{h+1}}\right] \leq 0.$$

On the other hand,

$$\begin{aligned}\mathrm{EU}\left(O_{c_h}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right)\right) - \mathrm{EU}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right) &\le L(\mathrm{EU}, H-h)\left\|O_{c_h}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right) - \left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right\|_\infty \\ &\le L(\mathrm{EU}, H-h)c_h \le b''_h.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathrm{EU}\left(O_{c_h}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right)\right) \le \mathrm{EU}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right) + b''_h \le \left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b''_h \\ &\Longrightarrow U_\beta\left(\tilde{\eta}_h\right) \le \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b''_h\right)\end{aligned}$$

$(d) \Longleftarrow$ Since $\left\|\left[\tilde{P}_h\tilde{\nu}_{h+1}\right] - \left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right\|_\infty \le \left\|\tilde{P}_h - \hat{P}_h\right\|_1 \le c_h$, we have $O_{c_h}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right) \succeq \left[\tilde{P}_h\tilde{\nu}_{h+1}\right]$. $U_\beta\left(\tilde{\nu}_{h+1}(s)\right) \ge V_{h+1}(s)$ for all $s \in \mathcal{S}$ implies $\mathrm{EU}\left(\tilde{\nu}_{h+1}(s)\right) \ge e^{\beta V_{h+1}(s)}$.

$$\begin{aligned}\mathrm{EU}\left(\tilde{\eta}_h\right) &= \mathrm{EU}\left(O_{c_h}\left(\left[\hat{P}_h\tilde{\nu}_{h+1}\right]\right)\right) \ge \mathrm{EU}\left(\left[\tilde{P}_h\tilde{\nu}_{h+1}\right]\right) \\ &\ge \left[\tilde{P}_h \circ \mathrm{EU}\left(\tilde{\nu}_{h+1}\right)\right] \ge \left[\tilde{P}_h \circ e^{\beta V_{h+1}}\right] \\ &= \mathrm{EU}\left(V_{h+1}, \tilde{P}_h\right).\end{aligned}$$

$(c+d) \Longleftarrow$

$$\begin{aligned}\left[\left(\tilde{P}_h - \hat{P}_h\right)e^{\beta V''_{h+1}}\right] &= \mathrm{EU}\left(V''_{h+1}, \tilde{P}_h\right) - \mathrm{EU}\left(V''_{h+1}, \hat{P}_h\right) \\ &\le L(\mathrm{EU}, H-h)\left\|\left(V''_{h+1}, \tilde{P}_h\right) - \left(V''_{h+1}, \hat{P}_h\right)\right\|_\infty \\ &\le L(\mathrm{EU}, H-h)\left\|\tilde{P}_h - \hat{P}_h\right\|_1 \\ &\le |e^{\beta(H+1-h)} - 1|c_h\end{aligned}$$

$(e) \Longleftarrow$

$$\left[\tilde{P}_h e^{\beta V_{h+1}}\right] \ge \left[P_h e^{\beta V_{h+1}}\right] \ge \left[P_h e^{\beta V^*_{h+1}}\right]$$

Observe that

$$\begin{aligned}Q'_h &= \min\left\{H+1-h, r_h + \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V'_{h+1}}\right] + b'_h\right)\right\} \\ &\ge \min\left\{H+1-h, r_h + \frac{1}{\beta}\log\left(\left[\hat{P}_h e^{\beta V''_{h+1}}\right] + b''_h\right)\right\} = Q''_h \\ &\ge U_\beta\left(\tilde{\eta}_h\right) \\ &\ge \min\left\{H+1-h, r_h + \frac{1}{\beta}\log\left(\left[\tilde{P}_h e^{\beta V_{h+1}}\right]\right)\right\} = Q_h,\end{aligned}$$

which implies that

$$V'_h = \max_a Q'_h(\cdot, a) \ge \max_a Q''_h(\cdot, a) = V''_h \ge \max_a U_\beta\left(\tilde{\eta}_h(\cdot, a)\right) = U_\beta\left(\tilde{\nu}_h\right) \ge \max_a Q_h(\cdot, a) = V_h.$$

■

Appendix C. Additional Property of EntRM

We state some lemmas about the monotonicity-preserving property and their proofs here. The results hold for general risk measures satisfying the monotonicity-preserving property.

Lemma 34 *Let ρ be a risk measure satisfying **(I)**. For any F and G such that $\rho(F) < \rho(G)$ and $0 \leq \theta' < \theta \leq 1$,*

$$\rho(\theta F + (1-\theta)G) < \rho(\theta' F + (1-\theta')G).$$

Proof Let $\tilde{\theta} = \frac{\theta'}{\theta'+1-\theta} \in [\theta', \theta]$ and $\bar{\theta} = \theta - \theta' \in (0,1]$. It holds that

$$\begin{aligned} \theta F + (1-\theta)G &= \bar{\theta} F + (1-\bar{\theta})(\tilde{\theta} F + (1-\tilde{\theta})G) \\ \theta' F + (1-\theta')G &= \bar{\theta} G + (1-\bar{\theta})(\tilde{\theta} F + (1-\tilde{\theta})G). \end{aligned}$$

The result follows from **(I)**

$$\rho(\bar{\theta} F + (1-\bar{\theta})(\tilde{\theta} F + (1-\tilde{\theta})G)) < \rho(\bar{\theta} G + (1-\bar{\theta})(\tilde{\theta} F + (1-\tilde{\theta})G)).$$

■

Lemma 35 *Let ρ be a risk measure satisfying **(I)** and $n \geq 2$ be an arbitrary integer. If $\rho(F_i) \geq \rho(G_i), \forall i \in [n]$ (and $\rho(F_j) \neq \rho(G_j)$ for some $j \in [n]$) then $\rho\left(\sum_{i=1}^n \theta_i F_i\right) \geq (>)\rho(\sum_{i=1}^n \theta_i G_i)$ 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 Lemma 34 we have $\rho(\sum_{i=1}^n \theta_i F_i) \geq \rho(\theta_1 G_1 + \sum_{i=2}^n \theta_i F_i)$. Suppose that for some $k \in [n-1]$ it holds that $\rho(\sum_{i=1}^n \theta_i F_i) \geq \rho(\sum_{i=1}^k \theta_i G_i + \sum_{i=k+1}^n \theta_i F_i)$. 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

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

The induction is completed. If in addition $\rho(F_j) > \rho(G_j)$ for some $j \in [n]$, the proof follows analogously by replacing the inequality to the strict inequality and the fact that $\theta_j > 0$. ■

Lemma 36 (Monotonicity-preserving under pairwise transport) *Let ρ be a risk measure satisfying the monotonicity-preserving property. Suppose $n \geq 2$ and $(F_i)_{i\in[n]}$ satisfies $\rho(F_1) \leq \rho(F_2)... \leq \rho(F_n)$. 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 $\rho(\sum_{i=1}^n \theta_i F_i) \leq \rho(\sum_{i=1}^n \theta'_i F_i)$.

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 Lemma 34, it suffices to prove $\rho(\frac{1}{\theta_i+\theta_j}(\theta'_i F_i + \theta'_j F_j)) \geq \rho(\frac{1}{\theta_i+\theta_j}(\theta_i F_i + \theta_j F_j))$. The result follows from the definition and the fact that $\rho(F_i) \leq \rho(F_j)$ and $\theta'_i \leq \theta_i$. ■

Lemma 37 (Monotonicity-preserving under block-wise transport) *Suppose $n \geq 2$ and $(F_i)_{i\in[n]}$ satisfies $\rho(F_1) \leq \rho(F_2)... \leq \rho(F_n)$. It holds that $\rho(\sum_{i=1}^n \theta_i F_i) \leq \rho(\sum_{i=1}^n \theta'_i F_i)$ 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 $\rho(\theta^l) \leq \rho(\theta^{l+1})$, then the proof shall be completed.

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 $\rho(\theta^l) \leq \rho(\theta^{l+1})$ for each iteration follows from Lemma 36. ■

Recall that the distributional optimism operator $\mathrm{O}^1_c : \mathscr{D}(\mathcal{S}) \mapsto \mathscr{D}(\mathcal{S})$ over space of PMFs with level c and future return $\nu \in \mathscr{D}^{\mathcal{S}}$ as

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

By Lemma 37, $\mathrm{O}^1_c\left(\widehat{P}, \nu\right)$ can be computed as follows

- sort ν in the ascending order such that $U_\beta(\nu^1) \leq U_\beta(\nu^2) \cdots \leq U_\beta(\nu^S)$
- permute $\hat{P}$ in the order of ν

- move probability mass $\frac{c}{2}$ of the first $S-1$ states sequentially to the S-th state

The computational complexity of the three steps are $O(S\log(S))$, $O(S)$, and $O(S)$. Therefore the computational complexity of applying O_c^1 in Line 6 of Algorithm 2 is only $O(S\log(S))$.

References

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

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

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.

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

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.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

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

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

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

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

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

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

Rolando Cavazos-Cadena and 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.

Yu Chen, Xiangcheng Zhang, Siwei Wang, and Longbo Huang. Provable risk-sensitive distributional reinforcement learning with general function approximation. *arXiv preprint arXiv:2402.18159*, 2024.

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, 2018a.

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

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

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

Darinka Dentcheva and Andrzej Ruszczynski. Common mathematical foundations of expected utility and dual utility theories. *SIAM Journal on Optimization*, 23(1):381–405, 2013.

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.

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.

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.

Yihan Du, Siwei Wang, and Longbo Huang. Provably efficient risk-sensitive reinforcement learning: Iterated cvar and worst path. In *The Eleventh International Conference on Learning Representations*, 2022.

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.

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.

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.

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

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.

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

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

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

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.

Michael Kupper and Walter Schachermayer. Representation results for law invariant time consistent functions. *Mathematics and Financial Economics*, 2:189–210, 2009.

Hao Liang and Zhi-Quan Luo. A distribution optimization framework for confidence bounds of risk measures. In *International Conference on Machine Learning*, pages 20677–20705. PMLR, 2023.

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.

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.

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

Alexandre Marthe, Aurélien Garivier, and Claire Vernade. Beyond average return in markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

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.

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

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

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.

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

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

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

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.

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

John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

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.

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

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34:1519–1529, 2021.