Entropic Risk Optimization in Discounted MDPs

Jia Lin Hau
University of New Hampshire
Durham, NH

Marek Petrik
University of New Hampshire and
Google Research

Mohammad Ghavamzadeh
Google Research
Mountain View, CA

Abstract

Risk-averse Markov Decision Processes (MDPs) have optimal policies that achieve high returns with low variability, but these MDPs are often difficult to solve. Only a few risk-averse objectives admit a dynamic programming (DP) formulation, which is the mainstay of most MDP and RL algorithms. We derive a new DP formulation for discounted risk-averse MDPs with Entropic Risk Measure (ERM) and Entropic Value at Risk (EVaR) objectives. Our DP formulation for ERM, which is possible because of our novel definition of value function with time-dependent risk levels, can approximate optimal policies in a time that is polynomial in the approximation error. We then use the ERM algorithm to optimize the EVaR objective in polynomial time using an optimized discretization scheme. Our numerical results show the viability of our formulations and algorithms in discounted MDPs.

1 INTRODUCTION

A major concern in high-stakes applications of reinforcement learning (RL), such as those in healthcare or finance, is to quantify the risk associated with the variability of returns. Since the standard expected objective does not capture the risk of random returns, *concave risk measures* have emerged as one of the most popular tools to quantify this risk in RL and beyond. They are sufficiently general to capture a wide range of stakeholder preferences and are more computationally convenient than many other alternatives (Follmer and Schied, 2016). Conditional value-at-risk (CVaR) is the best-known concave risk measure (Follmer and Schied, 2016; Shapiro et al., 2014) and the most commonly used to model risk aversion in MDPs (Angelotti et al., 2021; Bauerle and Glauner, 2022; Bäuerle and Ott, 2011; Bisi et al., 2022; Brown et al., 2020; Chow and Ghavamzadeh, 2014; Chow et al., 2015, 2018; Hiraoka et al., 2019; Lobo et al., 2021; Osogami, 2012; Santara et al., 2018; Tamar et al., 2014, 2015; Zhang et al., 2021).

The popularity of CVaR is mainly due to its intuitive interpretation as the expectation of the undesirable tail of the return random variable. Alas, solving risk-averse MDPs with the CVaR objective (CVaR-MDP) poses a difficult optimization problem. One can only formulate a dynamic program (DP) and a value function when the state space is augmented with an additional continuous variable (Bäuerle and Ott, 2011; Li et al., 2022; Pflug and Pichler, 2016a,b), which significantly complicates the computation of the value function and the implementation of the policy.

A popular remedy for the complexity of CVaR-MDPs is to use *nested*, also known as *iterated* or *Markov*, CVaR risk measure (Bauerle and Glauner, 2022; Defourny et al., 2008; Osogami, 2011). MDPs with a nested CVaR objective admit a value function that can be solved efficiently using DP. Unfortunately, nested CVaR approximates CVaR poorly and has several properties that make it impractical, e.g., it is difficult to interpret and is not law-invariant. The latter property is because the risk value also depends on the model dynamics and not only on the probability distribution of the returns (Follmer and Schied, 2016).

In this paper, we propose new algorithms for solving risk-averse discounted MDPs with two *entropic concave risk measures*: the entropic risk measure (ERM) (Follmer and Schied, 2016) and the entropic value-at-risk (EVaR) (Ahmadi-Javid, 2012; Follmer and Schied, 2016). Entropic risk measures are important alternatives to CVaR but their behavior in dynamic decision domains, like MDPs, is not yet well understood. Prior work on entropic risk measures in dynamic decision-making has been limited to ERM in undiscounted and average-reward MDPs (Borkar and Meyn, 2002; Neu et al., 2017), ERM for stochastic programs (Dowson et al., 2021), and nested EVaR-constrained models (Ahmadi et al., 2021a,b; Dixit et al., 2021). We believe this paper is the first work that investigates non-nested entropic risk measures in the *discounted* setting, which is the typical objective in RL.

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

We make two main contributions in this paper. As the first one, we show in Section 3 that in a discounted ERM-MDP, there exists an *optimal deterministic Markov policy* and an *optimal value function* that can be computed using *dynamic programming*. It is well-known that ERM is unique among law-invariant risk measures (Kupper and Schachermayer, 2006) in that it satisfies the tower property (see Theorem 2.1). However, the challenge with deriving DP equations with ERM in the discounted setting is that ERM is not positively-homogeneous, which makes it impossible to account for the discount factor. We hypothesize that this is the reason most prior work on ERM-MDPs have focused on undiscounted objectives (Borkar and Meyn, 2002; Neu et al., 2017), despite the popularity of discounting. Our main innovation in deriving the DP formulation is to compute a value function that uses *time-dependent risk levels* that decay exponentially over time to compensate for the discount factor. The DP is optimal for finite-horizon objectives and computes optimal infinite-horizon policies to a tolerance δ in $O(S^2 A \log(1/\delta))$ time, where S and A are the number of states and actions in the MDP.

As the second contribution, we show in Section 4 that in a discounted EVaR-MDP, there exists an *optimal deterministic Markov policy* and a policy that is arbitrarily close to optimal can be computed using a sequence of *dynamic programs*. This is particularly surprising because EVaR does not satisfy the tower property (Theorem 2.1), which is required for the existence of DP optimality equations. To show this result, we reduce solving the EVaR-MDP to solving a specific sequence of ERM-MDPs. In particular, our EVaR algorithm runs in $O(S^2 A(\frac{\log(1/\delta)}{\delta})^2)$ time. To the best of our knowledge, this is the first polynomial-time approximate algorithm for computing history-independent policies for coherent law-invariant risk measures in discounted MDPs.

Concurrently with our work, a state-augmentation approach has been proposed for solving discounted EVaR-MDPs (Ni and Lai, 2022a,b). This approach to EVaR-MDPs is inspired by a state augmentation method for solving CVaR-MDPs (Chow et al., 2015). Ni and Lai (2022a) states that the augmented EVaR-MDP Bellman operator is optimal with a sufficiently fine discretization of the augmented state. However, we show counterexamples to the optimality of these approaches in both EVaR-MDPs and CVaR-MDPs (Hau and Petrik, 2023).

Our empirical results in Section 5 confirm the efficacy of our algorithms. They outperform other techniques not only when evaluated in terms of ERM and EVaR metrics but also in terms of CVaR and VaR. This is not surprising because EVaR-MDP is easier to optimize than CVaR-MDP, and EVaR closely approximates CVaR and VaR (Ahmadi-Javid, 2012).

2 BACKGROUND

This section overviews the properties of MDPs and risk measures that we will need in the remainder of the paper.

2.1 Markov Decision Processes

We assume a problem formulated as a discounted MDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, r, p, s_0, \gamma)$, where $\mathcal{S} = 1{:}S$ and $\mathcal{A} = 1{:}A$ are the set of states and actions. The expression $a{:}b$ denotes a sequence, or a set, $a, a+1, \ldots, b$. The reward function $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the reward received in each state after taking an action. The transition probabilities are $p\colon \mathcal{S} \times \mathcal{A} \to \Delta^S$, where Δ^S is the probability simplex in $\mathbb{R}^S$. Finally, $s_0 \in \mathcal{S}$ is the initial state and $\gamma \in (0, 1]$ is the discount factor.

The most-general solution to an MDP is a randomized history-dependent policy that at each time-step prescribes a distribution over actions as a function of the history up to that step (Puterman, 2005). We use Π_{HR} and Π_{HD} to denote the sets of all history-dependent randomized and deterministic policies, respectively. A *randomized Markov* policy depends only on the time-step t and current state s_t as $\pi = (\pi_t)_{t=0}^{T-1}$, where $\pi_t\colon \mathcal{S} \to \Delta^A$. A policy π is *stationary* when it is time-independent (all π_t's are equal), in which case we omit the time subscript. We denote by Π_{MR} and Π_{SR}, the sets of Markov and stationary randomized policies, and by Π_{MD} and Π_{SD} their deterministic counterparts.

A common goal in an MDP is to maximize the discounted sum of rewards received by following a policy. We use $\mathfrak{R}^\pi_{t:T}(s)$ to denote the random *return* of a policy π from time-step t to T starting at state $s \in \mathcal{S}$, and define it as

$$\mathfrak{R}^\pi_{t:T}(s) \;=\; \sum_{t'=t}^{T-1} \gamma^{t'-t} \cdot \overbrace{r(S_{t'}, A_{t'})}^{R^\pi_{t'}} \mid S_t = s \;, \qquad (1)$$

where $S_{t'} \sim p(\cdot|S_{t'-1}, A_{t'-1})$, $A_{t'} \sim \pi_{t'}(\cdot|S_{t'})$, and $R^\pi_{t'}$ are the random variables of state visited, action taken, and reward received at a time-step $t' \in t{:}T{-}1$. We refer to $T \in \mathbb{N}^+ \cup \{\infty\}$ as the horizon with $T = \infty$ indicating an infinite-horizon objective. When $T = \infty$, we restrict the discount factor to $\gamma < 1$ to guarantee that $\mathfrak{R}^\pi_{t:T}$ is finite. While a discounted ($\gamma < 1$) finite-horizon objective is seldom used in practice, we use it later in the paper as an intermediate step for solving the infinite-horizon objective. Finally, we use $\Delta_{\mathfrak{R}} = (\max_{s\in\mathcal{S}, a\in\mathcal{A}} r(s,a) - \min_{s\in\mathcal{S}, a\in\mathcal{A}} r(s,a))/(1-\gamma)$ to denote the maximum range (span semi-norm) of the return random variable.

In standard risk-neutral MDPs, the objective is to maximize the *expected value* of the return random variable $\mathfrak{R}^\pi_T = \mathfrak{R}^\pi_{0:T}(s_0)$, that is,

$$\max_{\pi \in \Pi_{HR}} \mathbb{E}\left[\mathfrak{R}^\pi_T\right] \;. \qquad (2)$$

We denote the optimal policy in (2) by $\pi^\star$. Most MDP algorithms rely on the concept of a value function in one way or another. The value function $v^\pi = (v_t^\pi)_{t=0}^T$ for a policy $\pi \in \Pi_{MD}$ is a set of value functions $v_t^\pi \colon \mathcal{S} \to \mathbb{R},\ t \in 0{:}T$, each representing the expected return from a time-step t to the horizon T. For each $s \in \mathcal{S}$, we may write the value function v^π as

$$v_t^\pi(s) \ = \ \mathbb{E}\left[\mathfrak{R}_{t:T}^\pi(s)\right], \tag{3}$$

where $A \sim \pi(\cdot|s)$, $S' \sim p(\cdot|s, A)$, and $v_T(s) = 0$. The optimal value function $v^\star$ is simply the value function of the optimal policy $\pi^\star$: $v_t^\star = v_t^{\pi^\star},\ \forall t \in 0{:}T$. Both the value function of a policy π and the optimal value function satisfy Bellman equations (for all $s \in \mathcal{S}$ and all $t \in 0{:}T{-}1$):

$$\begin{aligned} v_t^\pi(s) \ &= \ \mathbb{E}\left[r(s, A) + \gamma \cdot v_{t+1}^\pi(S')\right], \\ v_t^\star(s) \ &= \ \max_{a \in \mathcal{A}} \mathbb{E}\left[r(s, a) + \gamma \cdot v_{t+1}^\star(S')\right], \end{aligned} \tag{4}$$

which allow us to compute them efficiently using DP. For the infinite-horizon setting ($T = \infty$ and $\gamma < 1$), one can show that there exists an optimal deterministic stationary policy $\pi^\star \in \Pi_{SD}$ and the value functions are also stationary: $v^\pi = v_t^\pi$ and $v^\star = v_t^\star$, for all $t = 0{:}T{-}1$.

Formulating the DP equations in (4) is only possible because of three important properties of the expectation operator (Puterman, 2005; Shapiro et al., 2014). In particular, the expectation operator $\mathbb{E}[\cdot]$ is monotone, positive homogeneous, and satisfies the tower property. It is *monotone* because $\mathbb{E}[X] \geq \mathbb{E}[Y]$ whenever $X \geq Y$, it is *positively homogeneous* because $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}[X]$, and it satisfies the *tower property* because $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$. In these equations, X and Y are any two random variables and c is a positive constant.

2.2 Concave Risk Measures

Concave risk measures are a generalization of the expectation operator $\mathbb{E}[\cdot]$ that can account for the variability of random variables. Formally, a concave risk measure $\psi[\cdot]$ is defined as a mapping $\psi \colon \mathbb{X} \to \mathbb{R}$ from the set of real-valued random variables $\mathbb{X}$ to real numbers that is *concave*, *monotone*, and *translation invariant* (Follmer and Schied, 2016). We summarize some of the most relevant risk measures and their properties in Appendix A.

Entropic Risk Measure (ERM) is the first risk measure we study in this paper. ERM is a *concave* risk measure with a parameter $\beta \in \mathbb{R}_+ \cup \{\infty\}$. It is defined for a random variable $X \in \mathbb{X}$ as (Follmer and Schied, 2016)[1]

$$\mathrm{ERM}_\beta[X] = -\beta^{-1} \cdot \log\Big(\mathbb{E}\left[e^{-\beta \cdot X}\right]\Big). \tag{5}$$

[1]Note that the ERM definition in (5) is for rewards; the definitions for cost does not negate the random variable X (Follmer and Schied, 2016; Shapiro et al., 2014).

For the risk level $\beta = 0$, ERM equals to the expectation: $\mathrm{ERM}_0[X] = \lim_{\beta \to 0^+} \mathrm{ERM}_\beta[X] = \mathbb{E}[X]$, while for $\beta \to \infty$, ERM equals to the minimum value of the random variable X: $\mathrm{ERM}_\infty[X] = \operatorname{ess\,inf}[X]$. ERM is *monotone* and satisfies the *tower property*. In fact, ERM is the only law-invariant risk measure that satisfies the tower property (Kupper and Schachermayer, 2006). Since we heavily use this property of ERM in our results, we state it in the following theorem and report its proof in Appendix B.

Theorem 2.1 (Tower Property)**.** *For any two random variables $X_1, X_2 \in \mathbb{X}$, we have*

$$\mathrm{ERM}_\beta[X_1] \ = \ \mathrm{ERM}_\beta\left[\mathrm{ERM}_\beta[X_1 \mid X_2]\right],$$

where the conditional ERM is defined analogously to a conditional expectation (see Definition A.4).

Despite the properties listed above, ERM is rarely employed in practice because it is *not positively homogeneous*, that is, $\mathrm{ERM}_\beta[c \cdot X] \neq c \cdot \mathrm{ERM}_\beta[X]$, which gives rise to undesirable risk preferences. For instance, a decision-maker guided by ERM may prefer an outcome X over Y when the profit is measured in dollars: $\mathrm{ERM}_\beta[X] > \mathrm{ERM}_\beta[Y]$, and yet the same decision-maker may prefer Y over X when the profit is measured in cents: $\mathrm{ERM}_\beta[100 \cdot X] < \mathrm{ERM}_\beta[100 \cdot Y]$. We analyze ERM primarily because it has favorable properties in dynamic decision-making, such as the tower property (mentioned above), and more importantly, we use it as a building block for our EVaR analysis.

Entropic Value-at-Risk (EVaR) is the second risk measure we study. EVaR was proposed as the tightest approximation of the popular value-at-risk (VaR) using the Chernoff inequality (Ahmadi-Javid, 2012). EVaR is *concave*, and unlike ERM, *positively homogeneous*, which makes it a *coherent risk measure*. EVaR with a confidence parameter $\alpha \in [0, 1)$ for a random variable $X \in \mathbb{X}$ is defined as (Ahmadi-Javid, 2012; Follmer and Schied, 2016)

$$\mathrm{EVaR}_\alpha[X] = \sup_{\beta > 0} \left(\mathrm{ERM}_\beta[X] + \frac{1}{\beta}\log(1 - \alpha)\right). \tag{6}$$

The meaning of EVaR's confidence level α is consistent with the level in value-at-risk (VaR) and conditional value-at-risk (CVaR), and we have $\mathrm{EVaR}_0[X] = \mathbb{E}[X]$ and $\lim_{\alpha \to 1} \mathrm{EVaR}_\alpha[X] = \operatorname{ess\,inf}[X]$. Computing the supremum in (6) is relatively easy because it involves maximizing a concave function over a single parameter (see proposition 2.11 in Ahmadi-Javid and Pichler 2017).

There are several ways to give an intuitive explanation of what EVaR measures. First, as mentioned above, EVaR can be seen as the tightest pessimistic approximation of both VaR and CVaR in the Chernoff bound sense (Ahmadi-Javid, 2012). In many settings, as the one depicted in Figure 1, EVaR approximates CVaR very closely. We are not

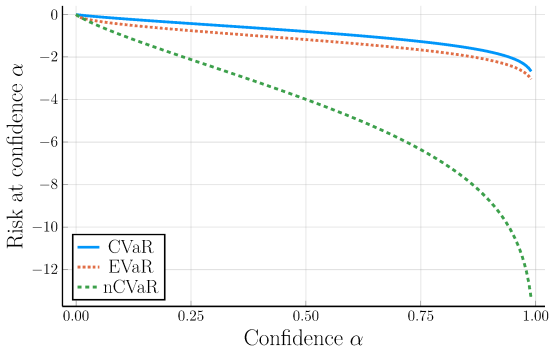

Figure 1: Comparison of CVaR, EVaR, and nCVaR (defined in Sec. 2.3) of the return random variable $\mathfrak{R}_T$ when all $T = 5$ immediate rewards are normally distributed.

Table 1: Properties of representative concave risk measures.

Risk measure	Law Inv.	Tower P.	Pos. Hom.
$\mathbb{E}$, Min	✓	✓	✓
ERM	✓	✓	✗
CVaR	✓	✗	✓
EVaR	✓	✗	✓
Nested CVaR	✗	✓	✓

aware, however, of a systematic analysis of this approximation error. Second, the *robust representation* of EVaR has a compelling interpretation in terms of a worst-case expectation to a distribution from a KL-divergence ball (Ahmadi-Javid, 2012):

$$\mathrm{EVaR}_\alpha[X] = \inf_{\xi \ll f} \left\{ \mathbb{E}_\xi[X] \mid \mathrm{KL}(\xi \| f) \leq \log\left(\frac{1}{1-\alpha}\right) \right\},$$

where $\ll$ denotes the absolute continuity of probability measures. The absolute continuity $\ll$ ensures that the KL divergence is well-defined.

2.3 Risk-averse MDPs

The objective in risk-averse MDPs is similar to the one in (2) for the risk-neutral case with the expectation operator $\mathbb{E}[\cdot]$ replaced with an appropriate risk measure $\psi[\cdot]$:

$$\max_{\pi \in \Pi_{HR}} \psi\left[\mathfrak{R}_T^\pi\right]. \tag{7}$$

Although this may appear as a small change compared to (2), it has profound implications on the complexity of the solution. Recall that the DP equations for MDPs exist because $\mathbb{E}[\cdot]$ is monotone, positively homogeneous, and satisfies the tower property. Alas, most interesting concave risk measures do not satisfy all these properties simultaneously, as shown in Table 1, which makes it difficult to solve the optimization problem in (7).

A common way to formulate DP equations for risk-averse MDPs is to use *nested* risk measures (Shapiro et al., 2014), also known as Markov (Ruszczynski, 2010), iterated (Osogami, 2011), or recursive (Bauerle and Glauner, 2022) risk measures. For instance, the nested CVaR is informally defined as

$$\mathrm{nCVaR}_\alpha[\mathfrak{R}_T^\pi] = \mathrm{CVaR}_\alpha\left[R_0^\pi + \gamma\, \mathrm{CVaR}_\alpha\left[R_1^\pi + \dots\right]\right].$$

When the nested risk measure is properly formalized, one can compute the optimal value function using DP equation

$$v_t^\star(s) = \max_{a \in \mathcal{A}} \mathrm{CVaR}_\alpha\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right],$$

which is similar to that in (4) for risk-neutral MDPs

Despite their favorable computational properties, nested risk measures suffer from an important drawback in that they are not *law-invariant*. Therefore, the risk value is not solely a function of the distribution of the return $\mathfrak{R}_T^\pi$, and is affected by the MDP's dynamics in a non-trivial way that is difficult to anticipate and interpret. The impact of law invariance on risk preferences is well-documented in the risk literature and can actually cause an agent to prefer returns with higher variability (Iancu et al., 2015).

Nested risk measures are poor approximations of static risk measures. To illustrate this fact, Figure 1 depicts the values of CVaR, EVaR, and nCVaR for $\mathfrak{R}_4 = R_0 + \dots + R_4$ with R_t's normally distributed with $\mu = 0$ and $\sigma = 1$.

Prior literature has explored several other approaches to optimizing risk in MDPs besides nested risk measures. As mentioned in the introduction, one can augment the state space in order to approximate the optimal policy in MDPs with VaR or CVaR objectives (Bäuerle and Ott, 2011; Li et al., 2022; Pflug and Pichler, 2016a,b). This is a powerful approach, but it can be very computationally intensive and return policies that depend on history. History-dependent policies can be complex and difficult to interpret and deploy. It is also possible to use gradient-based algorithms, such as policy gradient, to directly optimize objective (7) (e.g. Tamar et al. 2012, 2015). These algorithms often work well but lack any guarantees and usually converge to inferior local optima. Finally, many other notions of risk have been proposed and optimized, from utility-based risk (Ben-Tal, 2007), to variance-based risk (Prashanth and Ghavamzadeh, 2013, 2016; Tamar et al., 2013), to cautious RL (Zhang et al., 2021). These other notions of risk make very different modeling assumptions which makes it difficult to compare them with our framework that is based on coherent risk measures.

3 DISCOUNTED MDPs WITH ERM OBJECTIVE

In this section, we study the fundamental properties of discounted ERM-MDPs and describe a new DP formulation. In particular, we show that if one defines an optimal value

function for ERM-MDP with a specific time-dependent risk, then it can be computed using DP. We report the proofs for this section in Appendix C.

The objective in this section is to compute a policy that maximizes the ERM of the return random variable $\mathfrak{R}_T^\pi$ at some given risk level $\beta \geq 0$. That is, the objective is the optimization in (7) for risk-averse MDPs with the risk measure $\psi[\cdot]$ set to $\mathrm{ERM}_\beta[\cdot]$:

$$\max_{\pi \in \Pi_{HR}} \mathrm{ERM}_\beta\left[\mathfrak{R}_T^\pi\right]. \tag{8}$$

Although we formulate the objective in (8) in terms of history-dependent randomized policies, we will prove later in this section that there always exists a Markov (history-independent) deterministic policy for (8). In the remainder of this section, we first treat the finite-horizon case ($T < \infty$) and then extend the obtained results to the discounted infinite-horizon case ($T = \infty$ and $\gamma < 1$).

The closest objective to (8) studied in prior work is the ERM-MDP with an *average-reward* objective (Borkar and Meyn, 2002). Value iteration, policy iteration, and even q-learning (Borkar, 2002) have been studied for this objective. However, the average reward criterion is not as popular in RL as the discounted infinite horizon objective. In addition, the example below illustrates why the existing formulations do not readily extend to discounted ERM-MDPs.

Before defining the value function and deriving the corresponding DP equations for ERM-MDP, we describe a simple example that illustrates how one may use the tower property to derive such equations. The example also illustrates the challenge that discounting poses in ERM-MDP.

Example 1. Consider an MDP with a single action a, and thus, a single policy π. Assume that the horizon is $T = 2$ and the initial state S_0 is random. Recall that the return is defined as $\mathfrak{R}_2^\pi = r(S_0, a) + \gamma \cdot r(S_1, a)$. When $\gamma = 1$, one can directly use the tower property to decompose the return into value functions as

$$\begin{aligned}
&\mathrm{ERM}_\beta\left[\mathfrak{R}_2^\pi\right] = \mathrm{ERM}_\beta\left[r(S_0,a) + \gamma \cdot r(S_1,a)\right] \\
&\quad = \mathrm{ERM}_\beta\left[r(S_0,a) + \mathrm{ERM}_\beta\left[\gamma \cdot r(S_1,a) \mid S_0\right]\right] \\
&\quad = \mathrm{ERM}_\beta\left[r(S_0,a) + \gamma \cdot \mathrm{ERM}_\beta\left[r(S_1,a) \mid S_0\right]\right] \\
&\quad = \mathrm{ERM}_\beta\left[r(S_0,a) + \gamma \cdot v_1(S_1)\right],
\end{aligned} \tag{9}$$

where $v_1(S_0) = \mathrm{ERM}_\beta\left[r(S_1, a) \mid S_0\right]$. While the above derivation readily generalizes to the MDP with $\gamma = 1$, it is not valid when $\gamma < 1$, because the equality in (9) requires ERM to be positive homogeneous.

3.1 Finite Horizon ERM-MDP

Although ERM is not positively homogeneous, in the following new result we show that it has a similar property if we allow for a change in the risk level.

Theorem 3.1 (Positive Quasi-homogeneity)**.** *For any random variable $X \in \mathbb{X}$ and any constant $c \geq 0$, we have*

$$\mathrm{ERM}_\beta\left[c \cdot X\right] \;=\; c \cdot \mathrm{ERM}_{\beta \cdot c}\left[X\right]. \tag{10}$$

Theorem 3.1 indicates that we can propagate the discount factor γ out of ERM in (9), if we change the risk level of the inner ERM to $\beta\gamma$. In particular, if we define the value function as $v_1(S_0) = \mathrm{ERM}_{\beta\cdot\gamma}\left[r(S_1, a) \mid S_0\right]$, then the derivation in Example 1 works for any $\gamma \in (0, 1)$. Generalizing this intuition to the full ERM-MDP, we define the value function $v^\pi = (v_t^\pi)_{t=0}^T$ for a policy $\pi = (\pi_t)_{t=0}^{T-1} \in \Pi_{MR}$ and the optimal value function $v^\star = (v_t^\star)_{t=0}^T$ in a state $s \in \mathcal{S}$ as follows:

$$v_t^\pi(s) = \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathfrak{R}_{t:T}^\pi(s)\right], \tag{11}$$

$$v_t^\star(s) = \max_{\pi \in \Pi_{MR}^{t:T}} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathfrak{R}_{t:T}^\pi(s)\right], \tag{12}$$

where $\mathfrak{R}_{t:T}^\pi(s)$ is defined by (1) and $\Pi_{MR}^{t:T}$ is the set of randomized Markov policies for time-steps t:$T-1$.

As discussed above, it is important that the risk level in the definition of value function in (11) depends on the time-step t. As time progresses, the risk level $\beta\gamma^t$ decreases monotonically and the value function becomes less risk-averse. Recall that in the risk-neutral setting, the risk level is $\beta = 0$ and $\mathrm{ERM}_0\left[X\right] = \mathbb{E}[X]$. When we set $\beta = 0$ in (11), the value function becomes independent of t and the value function reduces to that in risk-neutral MDPs.

The following theorem states the main result of this section. It shows how the value functions defined in (11) and (12) can be efficiently computed by a DP when $T < \infty$.

Theorem 3.2 (Bellman Equations in ERM-MDP)**.** *For any policy $\pi \in \Pi_{MR}$, its value function $v^\pi = (v_t^\pi)_{t=0}^T$ defined in* (11) *is the unique solution to the following system of equations for all $s \in \mathcal{S}$,*

$$v_t^\pi(s) = \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s, A) + \gamma \cdot v_{t+1}^\pi(S')\right], \tag{13}$$

where $A \sim \pi_t(\cdot|s)$, $S' \sim p(\cdot|s, A)$, and $v_T^\pi(s) = 0$. Moreover, the optimal value function $v^\star = (v_t^\star)_{t=0}^T$ defined in (12) *is the unique solution to the following system of equations for all $s \in \mathcal{S}$,*

$$v_t^\star(s) = \max_{a \in \mathcal{A}} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s, a) + \gamma \cdot v_{t+1}^\star(S')\right]. \tag{14}$$

Theorem 3.2 suggests several new important and surprising properties for ERM-MDP. First, it shows the existence of value functions, both for any Markov policy and also the optimal value function. Unlike with other risk measures, these value functions do not require that the state space is augmented. Second, the theorem shows that the value function can be computed efficiently using a dynamic program. And finally, the next theorem built on Theorem 3.2 shows that there always exists an optimal Markov (as opposed to history-dependent) deterministic policy for the ERM-MDP, and this policy is greedy w.r.t. the optimal value function.

Theorem 3.3 (Optimal Policy in ERM-MDP)**.** *There exists a Markov deterministic optimal policy* $\pi^\star = (\pi_t^\star)_{t=0}^{T-1} \in \Pi_{MD}$ *for the optimization problem* (8)*, which is greedy w.r.t. the optimal value function* $v^\star$ *defined by* (14)*, that is,*

$$\pi_t^\star(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \mathrm{ERM}_{\beta \cdot \gamma^t} \left[r(s,a) + \gamma v_{t+1}^\star(S') \right], \quad (15)$$

for all $s \in \mathcal{S}$*. Moreover, the optimal value function satisfies that* $v^{\pi^\star} = v^\star$*.*

The existence of a deterministic optimal policy in ERM-MDP is surprising since many risk-averse formulations require randomization (Delage et al., 2019; Lobo et al., 2021; Steimle et al., 2021). Also surprisingly, ERM-MDP does not admit a stationary optimal policy ($\pi^\star$ in (15) is time-dependent) even when the horizon T is large or infinite. This is in contrast to risk-neutral discounted infinite-horizon MDPs which admit stationary optimal policies.

Given the results of Theorems 3.2 and 3.3, we can solve the ERM-MDP objective (8) when the horizon is finite ($T < \infty$) by adapting the standard value iteration (VI) algorithm to this setting. This algorithm, whose pseudo-code is shown in Algorithm 3 in Appendix C, computes the optimal value function $v_t^\star$ backwards in time ($t = T, T-1, \ldots, 0$) according to (14). The optimal policy is greedy w.r.t. $v^\star$ and can be computed by solving the optimization (15).

3.2 Infinite Horizon ERM-MDP

We now turn to deriving an algorithm that can solve the ERM-MDP objective (8) when the horizon T is large or infinite. Solving ERM-MDP in the *infinite-horizon* setting is considerably more challenging than in finite-horizon, because the risk level and optimal policy are both time dependent. The simplest way to address this issue is to truncate the horizon at some $T' < \infty$ and resort to an arbitrary policy for any $t > T'$. The main limitation of this approach is that T' may need to be very large to achieve a reasonably-small approximation error.

In Algorithm 1, we propose an approximation that is superior to the simple truncation of the planning horizon, described above. Algorithm 1 first computes the optimal risk-neutral value function $v_\infty^\star$ and (stationary) policy $\pi_\infty^\star$ using value or policy iteration algorithms (Puterman, 2005). It uses policy $\pi_\infty^\star$ to act for all time-steps $t > T'$ and value function $v_\infty^\star$ to approximate $v_{T'}^\star$. This approach takes advantage of the fact that the risk level $\beta \cdot \gamma^t$ in (11) and (12) approaches 0 as t gets larger, which means that the ERM value function becomes close to the risk-neutral $v_\infty^\star$.

To quantify the quality of a policy $\hat{\pi}^\star$ returned by Algorithm 1, we now derive a bound on its performance loss. In particular, we focus on how quickly the error decreases as a function of the planning horizon T'. This bound can be used both to determine the planning horizon and to quantify the improvement of Algorithm 1 over simple truncation.

Algorithm 1: VI for infinite-horizon ERM-MDP

Input: planning horizon $T' < \infty$, risk level $\beta > 0$
Output: policy $\hat{\pi}^\star = (\hat{\pi}_t^\star)_{t=0}^\infty$ and value function $\hat{v}^\star = (\hat{v}_t^\star)_{t=0}^\infty$
Compute $v_\infty^\star$ and $\pi_\infty^\star$ as the optimal solutions to the risk-neutral infinite-horizon discounted problem ;
Compute $(\tilde{v}_t^\star)_{t=0}^{T'}$ and $(\tilde{\pi}_t^\star)_{t=0}^{T'-1}$ using (14) and (15) with horizon T' and terminal value $\tilde{v}_{T'}^\star = v_\infty^\star$;
Construct a policy $(\hat{\pi}_t^\star)_{t=0}^\infty$, where $\hat{\pi}_t^\star = \pi_\infty^\star$ for $t \geq T'$ and $\hat{\pi}_t^\star = \tilde{\pi}_t^\star$, otherwise ;
Construct $\hat{v}^\star$ analogously to $\hat{\pi}^\star$;
return $\hat{\pi}^\star$, $\hat{v}^\star$

Theorem 3.4. *The performance loss of the policy* $\hat{\pi}^\star$ *returned by Algorithm 1 decreases with* T' *as*

$$\mathrm{ERM}_\beta \left[\mathfrak{R}_\infty^{\pi^\star} \right] - \mathrm{ERM}_\beta \left[\mathfrak{R}_\infty^{\hat{\pi}^\star} \right] \leq \frac{\beta \cdot \gamma^{2T'} \cdot \Delta_{\mathfrak{R}}^2}{8}, \quad (16)$$

where $\pi^\star$ *is optimal in* (8) *and* $\Delta_{\mathfrak{R}}$ *is the range of the return random variable* $\mathfrak{R}_\infty$*. Therefore, Algorithm 1 runs in* $O(S^2 A \log(1/\delta))$ *time to compute a* δ*-optimal policy.*

The proof of Theorem 3.4 reported in Appendix C uses the Hoeffding's lemma to bound the error between ERM and expectation, and then propagates it backwards using standard DP techniques. Analysis analogous to Theorem 3.4 shows that when we simply truncate the planning horizon at T' and follow an arbitrary policy thereafter, the performance loss decreases proportionally to $\gamma^{T'}$ as opposed to $\gamma^{2T'}$ in (16). As a result, simple truncation requires a planning horizon T' that is at least twice longer than the one used by Algorithm 1 to achieve the same performance.

Remark 1 (Quadratic dependence on $\Delta_{\mathfrak{R}}$). An attentive reader may be puzzled by the fact that the bound in Theorem 3.4 scales quadratically with the range of the returns $\Delta_{\mathfrak{R}}$. Given the quadratic dependence, one can make the relative error arbitrarily small just by shrinking the rewards appropriately. This is indeed true but is less useful than it may seem at the first blush. Since ERM is not positively homogeneous, scaling the rewards can change the optimal policy, unlike in risk-neutral MDPs. To avoid changing the set of optimal policies when scaling the rewards, one also needs to scale the risk parameter β appropriately as dictated by Theorem 3.1. When both r and β are scaled simultaneously, the relative error in Theorem 3.4 does not change.

In practice, one can compute bounds that are tighter than the one in Theorem 3.4 by computing both an upper-bound on the optimal value function and a lower-bound on the value of the policy. It is easy to see that $v_\infty^\star$ is an upper-bound on $v^\star$, which can be used to compute an upper-bound on $v_0^\star$, and therefore, an upper-bound on the perfor-

mance loss. According to Theorem 3.4, given an arbitrary desired tolerance δ, one can select $T' \geq \frac{1}{2\log(\delta)} \log(\frac{8\delta}{\beta\Delta_{\mathfrak{R}}^2})$ to compute a δ-optimal policy. We give more details on this in Appendix C.

4 DISCOUNTED MDPs WITH EVaR OBJECTIVE

In this section, we analyze the EVaR-MDP objective and propose a new DP algorithm to solve it. As mentioned in Section 2, EVaR is preferable to ERM because it is coherent and approximates both VaR and CVaR well. We report the proofs of this section in Appendix D.

The objective in this section is to compute a policy that maximizes the EVaR of the return random variable $\mathfrak{R}_T^\pi$ at some given risk level $\alpha \in [0, 1]$. In other words, we are interested in solving the optimization problem in (7) with the risk measure $\psi[\cdot]$ set to $\mathrm{EVaR}_\alpha[\cdot]$:

$$\max_{\pi \in \Pi_{HR}} \mathrm{EVaR}_\alpha \left[\mathfrak{R}_T^\pi\right]. \tag{17}$$

It is important to note that the objective in (17) differs from prior work on EVaR in MDPs, which has focused on the nested EVaR objective (Ahmadi et al., 2021a,b; Dixit et al., 2021), and thus, does not approximate the static formulation in (17) well (Iancu et al., 2015).

The main challenge in solving (17) is that EVaR does not satisfy the tower property (or equivalently, it is *not* dynamically consistent) and cannot be directly optimized using a DP. Our main contribution in this section is to derive an algorithm that solves EVaR-MDP in time that is polynomial in the problem size and the desired accuracy. The main idea of our algorithm is to reduce EVaR-MDP to a specific sequence of ERM-MDP problems.

Using the definition of EVaR in (6), we may reformulate the EVaR-MDP objective (17) as

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha \left[\mathfrak{R}_T^\pi\right] = \sup_{\beta > 0} h(\beta) \tag{18}$$

where the objective function $h\colon \mathbb{R} \to \mathbb{R}$ is defined as

$$h(\beta) = \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_\beta \left[\mathfrak{R}_T^\pi\right] + \beta^{-1} \cdot \log(1-\alpha)\right).$$

We switch the notation to Markov policies, Π_{MR}, because we will show in Corollary 4.2 that an optimal policy for EVaR-MDP belongs to this class. The equality in (18) follows by swapping the order of max and sup operators. The connection that the reformulation in (18) establishes between the objectives of EVaR-MDP and ERM-MDP allows us to directly carry over the following properties from the ERM-MDP setting to EVaR-MDP.

Theorem 4.1. *Let $\pi^\star$ be an optimal solution to the EVaR-MDP in* (17) *and suppose that the supremum is attained. Then, there exists a risk level $\beta^\star \in (0, \infty]$ such that $\pi^\star$ is optimal for ERM-MDP in* (8) *with $\beta = \beta^\star$.*

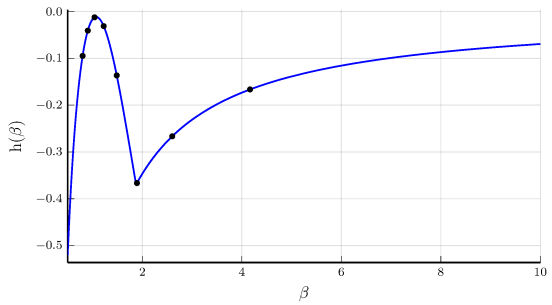

Figure 2: Function h for the EVaR-MDP described in Proposition D.1. The circles indicate the discretized $h(\beta_k)$ according to (19) with $\alpha = 0.5$, $\delta = 0.1$, and $\Delta_{\mathfrak{R}} = 1$.

A similar argument holds also when the supremum is not attained, but requires additional technical developments, which we leave it for a future extended version of this work.

Corollary 4.2. *There exists an optimal Markov deterministic policy for EVaR-MDP.*

We are now ready to describe our algorithm for solving the EVaR-MDP objective. Our algorithm, whose pseudo-code is shown in Algorithm 2, optimizes the single-dimensional objective function h in (18). Because the function h is not concave in general (see Proposition D.1 in Appendix D), we cannot use standard one-dimensional algorithms like Brent's method or BFGS. However, we leverage the fact that h is the sum of a non-increasing function and a concave function, and use a discrete grid $\{\beta_k\}_{k=1}^K$ to search over the risk level β that can approximate the optimal policy in polynomial time. We define the grid values for each $k \in 1{:}K{-}1$ as

$$\beta_1 = \frac{8\delta}{\Delta_{\mathfrak{R}}^2}, \quad \beta_{k+1} = \beta_k \cdot \frac{\log(1-\alpha)}{\beta_k \delta + \log(1-\alpha)}, \tag{19}$$

where $\delta > 0$ is the desired approximation error and $K \in \mathbb{N}$ is sufficiently large to ensure that

$$\beta_K \geq \frac{-\log(1-\alpha)}{\delta}. \tag{20}$$

We also assume that β_K is trimmed so that (20) holds with equality. The grid values in (19) are constructed to minimize the suboptimality bound δ in Theorem 4.3 below. Figure 2 depicts an example of a function h for an MDP described in the proof of Proposition D.1. The black circles are the values of h at the grid points $\{\beta_k\}_{k=1}^K$ constructed according to (19) with $\delta = 0.1$ and $\Delta_{\mathfrak{R}} = 1$.

The following theorem shows that Algorithm 2 runs in time that is polynomial in $1/\delta$ and computes a policy $\hat{\pi}^\star$ whose return has an EVaR that is δ-close to optimal.

Theorem 4.3. *For any $\delta > 0$, Algorithm 2 runs in $O(S^2 A(\frac{\log(1/\delta)}{\delta})^2)$ time and returns a policy $\hat{\pi}^\star$ such that*

$$\mathrm{EVaR}_\alpha \left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{EVaR}_\alpha \left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] \leq \delta,$$

Algorithm 2: Algorithm for EVaR-MDP

Input: Desired error tolerance δ
Output: EVaR-MDP optimized policy $\hat{\pi}^\star$
Let K be the smallest value that satisfies (20) ;
for $k = 1, \ldots, K$ **do**
  Compute v^k, π^k by solving ERM-MDP with risk level β_k defined in (19) ;
Let $k^\star \leftarrow \text{argmax}_{k=1:K}\ v_0^k(s_0) + \beta_k^{-1} \cdot \log(1-\alpha)$;
return *Policy* $\hat{\pi}^\star = \pi^{k^\star}$

where $\pi^\star$ is optimal for (17).

Theorem 4.3 establishes the time complexity that Algorithm 2 needs to compute a δ-optimal EVaR policy. Note that the bound in Theorem 4.3 takes into account both the errors due to the discretization in (19) and the truncated horizon in Theorem 3.4 when solving the ERM-MDPs. The proof of Theorem 4.3 is reported in Appendix D.

We report Algorithm 2 for solving EVaR-MDP because it is conceptually simple and relatively easy to analyze. However, significant computational improvements are possible in this setting. One approach to accelerate Algorithm 2 is by realizing that Algorithm 1 computes value functions for multiple risk levels $\beta, \gamma\beta, \gamma^2\beta, \ldots$. For example, running Algorithm 1 with $\beta = 0.5$ computes v_0 with risk level $\beta = 0.5$, v_1 with risk level $\beta = 0.5\gamma$, v_2 with risk level $\beta = 0.5\gamma^2$ and so on. This observation can significantly reduce the computational effort while introducing an additional small error due to the effective approximate horizon T' being different for different grids over the risk level β.

5 NUMERICAL EVALUATION

In this section, we evaluate our EVaR-MDP algorithm numerically on several tabular MDPs. We focus on the EVaR-MDP objective for two reasons. First, as discussed in Section 2, EVaR is a more practical risk measure than ERM because it is coherent and approximates the popular VaR and CVaR well. Second, the EVaR-MDP algorithm (Algorithm 2) also evaluates the ERM-MDP algorithm (Algorithm 1) since it uses it as a subroutine.

We assume that the objective is to solve an EVaR-MDP for a confidence level $\alpha = 0.9$. That is, we seek to find a policy π that maximizes $\text{EVaR}_{0.9}[\mathfrak{R}_T^\pi]$. The confidence level $\alpha = 0.9$ is a common choice in the risk-averse literature and the results are qualitatively insensitive to its choice. The numerical evaluation assumes a finite horizon $T = 100$, which makes it possible to evaluate the risk of $\mathfrak{R}_T^\pi$ by simulation. We sample 100,000 episodes of $\mathfrak{R}_T^\pi$ for this evaluation.

To understand how the components of Algorithm 2 contribute to the quality of its solution, we perform a small ablation study that compares it with two simplified algorithms: **1)** *naive grid* that uses a uniform grid of values $\beta_k,\ k = 1{:}K$, such that $\beta_1 = 0$ and $\beta_K = 10$, instead of what we propose in (19), and K is set to the same value as in (19), and **2)** *naive level* that uses the optimized grid but does not adjust the risk level with the time-step when it solves ERM-MDPs. Algorithm 2 uses the optimized grid in (19) with δ and $\Delta_{\mathfrak{R}}$ values given in Appendix E.

In addition to the ablation study, we also compare Algorithm 2 with several risk-averse algorithms that optimize objectives related to EVaR-MDPs. Specifically, we compare it with *risk-neutral* MDP, *nested CVaR* (Bauerle and Glauner, 2022), and *nested EVaR* (related to Ahmadi et al. 2021b), both with $\alpha = 0.9$, and finally *ERM* (Algorithm 3) and *nested ERM*, both with $\beta = 0.5$. The parameter α was chosen to match the EVaR objective, but the parameter $\beta = 0.5$ is chosen arbitrarily since no general method exists to find a β that matches a given α. All the above methods compute Markov policies. We also compare Algorithm 2 with *augmented CVaR* (Chow et al., 2015) that computes a history-dependent policy for CVaR-MDPs. We implemented the augmented CVaR method using the faster quantile-based approach described in section 4 of Li et al. (2022). All algorithms were implemented in Julia with the exception of augmented MDP which was implemented both in R and Julia. The R implementation using quantiles was significantly faster than the implementation of the original algorithm (Chow et al., 2015) in Julia 1.8.

As described in the introduction the augmented CVaR-MDP (Chow et al., 2015) may not compute an optimal policy (Hau and Petrik, 2023). While augmented CVaR-MDP is guaranteed to evaluate policies correctly, the dynamic program overestimates the true optimal value function and computes suboptimal policies. This is one possible reason for why EVaR-MDP achieves a better CVaR objective than the augmented CVaR-MDP. We do not compare with the augmented EVaR-MDP (Ni and Lai, 2022a,b) for two main reasons. First, this algorithm is even slower than the augmented CVaR-MDP because one needs to solve a conic optimization instead of a linear optimization in each time-step. Second, this augmented EVaR-MDP is not guaranteed to compute a correct (or even approximately correct) value function even when the policy is fixed (Hau and Petrik, 2023).

To obtain a holistic picture of the relative performance of the algorithms, we selected a diverse set of domains with varying numbers of actions, discount factors, and levels of uncertainty. These domains have all been used in risk-averse and robust RL literature and are as follows: *machine replacement* (MR) (Delage and Mannor, 2010), *gamblers ruin* (GR) (Bäuerle and Ott, 2011; Li et al., 2022), two classic *inventory* management problems (INV1) and (INV2) (Ho et al., 2021), and *river-swim* (RS) (Strehl and Littman, 2008).

Table 2: $\mathrm{EVaR}_{0.9}\left[\mathfrak{R}_T^\pi\right]$ for π returned by each method.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	**-6.73**	**5.34**	**67.4**	**189**	**303**
Naive grid	**-6.87**	**5.37**	43.2	**189**	**303**
Naive level	-10.00	4.17	64.6	**188**	217
Risk neutral	**-6.53**	2.29	40.6	**186**	**300**
Nested CVaR	-10.00	-0.02	-0.0	132	217
Nested EVaR	-10.00	4.61	-0.0	164	217
ERM	**-6.72**	5.19	50.7	178	217
Nested ERM	-10.00	4.76	24.9	150	217
Augmented CVaR	-7.06	3.64	49.0	82	93

Table 3: $\mathrm{CVaR}_{0.9}\left[\mathfrak{R}_T^\pi\right]$ for π returned by each method.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	**-4.62**	7.87	**76.6**	**195**	**382**
Naive grid	**-4.63**	7.91	47.8	**195**	**381**
Naive level	-10.00	7.41	73.1	**194**	217
Risk neutral	**-4.56**	5.47	52.3	**193**	**379**
Nested CVaR	-10.00	0.00	0.0	135	217
Nested EVaR	-10.00	7.12	0.0	169	217
ERM	**-4.58**	7.64	56.0	182	217
Nested ERM	-10.00	7.27	28.3	153	217
Augmented CVaR	**-4.83**	**8.27**	55.1	82	101

Table 2 summarizes $\mathrm{EVaR}_{0.9}\left[\mathfrak{R}_T^\pi\right]$ for policies π computed by the algorithms described above. Bold font indicates results within a 95% confidence interval of the best policy. The variation in these results is due to simulation used to estimate the risk. We can make the following observations from the results. First, the particular design of Algorithm 2 is important because it outperforms its ablated versions significantly on some domains. Second, the results confirm that none of the nested risk measures can optimize the static EVaR-MDP well. Even the risk-neutral policy often outperforms the nested risk measures. Finally in Table 3, we show that the results are similar when compared in terms of the $\mathrm{CVaR}_{0.9}\left[\mathfrak{R}_T^\pi\right]$ objective. This is not surprising since EVaR is often a good proxy for CVaR (Ahmadi-Javid, 2012). Note that *augmented CVaR* is guaranteed to be optimal for CVaR when the discretization is sufficiently fine, which significantly increases the computation time.

It is also important to discuss the run-time of the algorithms summarized in Table 4. We implemented all of them in Julia and ran each one in less than a 30 seconds on a laptop computer with the exception of *augmented CVaR* that we ran for up to 10 minutes. The difference in run-time between solving the nested risk measures and computing the ERM-MDP optimal value function (described in Theorem 3.2) is negligible since they all evaluate nearly identical dynamic programs. However, Algorithm 2 in our experiments typically needs to solve between 20 and 50 ERM-MDP problems, one for each $\beta_k,\ k = 1{:}K$. This additional computation is significant, but we believe it can be addressed. As described in Section 4, there are ways to significantly speed up Algorithm 2, but we decided to focus on algorithms that are conceptually simple and can be analyzed in this paper, and leave computational concerns for future work.

Table 4: Run-time for the algorithms in second.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	2.70	6.35	1.14	0.96	3.87
Naive grid	2.64	6.30	1.05	0.88	3.81
Naive level	2.79	6.38	1.19	0.92	3.95
Risk neutral	0.00	0.00	0.18	0.20	0.00
Nested CVaR	0.01	0.01	0.26	0.16	0.01
Nested EVaR	0.01	0.03	0.66	0.06	0.01
ERM	0.00	0.00	0.24	0.16	0.00
Nested ERM	0.01	0.01	0.10	0.02	0.01
Augmented CVaR	14.8	29.01	780	120	22.9

6 CONCLUSION

We analyzed discounted MDPs with two risk measures: ERM and EVaR that had not been studied in discounted multi-stage decision-making literature. This lack of interest is surprising because their properties make them especially suitable for dynamic decision-making. We derived the first exact DP formulation for ERM in discounted MDPs. We also showed that the optimal value function and an optimal deterministic Markov policy exist for ERM-MDP, and can be computed using value iteration. We showed that EVaR-MDP also has deterministic optimal policies, proposed a new polynomial-time algorithm for computing them, and demonstrated the algorithms numerically. Our numerical results showed that our EVaR algorithm performs consistently well across several domains and risk measures.

Acknowledgments

We would like to thank Reazul Russel, Erick Delage, Julien Grand-Clément, and Yinlam Chow for their comments that helped to improve the presentation and correctness of the paper. We also thank the anonymous reviewers for their comments. This work was supported, in part, by NSF grants 2144601 and 1815275.

References

M. Ahmadi, U. Rosolia, M. D. Ingham, R. M. Murray, and A. D. Ames. Constrained risk-averse Markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11718–11725, 2021a.

M. Ahmadi, U. Rosolia, M. D. Ingham, R. M. Murray, and A. D. Ames. Risk-averse decision making under uncertainty, 2021b.

A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 2012.

A. Ahmadi-Javid and A. Pichler. An analytical study of norms and Banach spaces induced by the entropic value-at-risk. *Mathematics and Financial Economics*, 11(4): 527–550, 2017.

G. Angelotti, N. Drougard, and C. P. C. Chanel. Exploitation vs caution: Risk-sensitive policies for offline learning. *arXiv:2105.13431 [cs, eess]*, 2021.

P. Artzner, F. Delbaen, J. M. Eber, D. Heath, and H. Ku. Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, 2004.

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

N. Bäuerle and J. Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

A. Ben-Tal. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17:449–476, 2007.

L. Bisi, D. Santambrogio, F. Sandrelli, A. Tirinzoni, B. D. Ziebart, and M. Restelli. Risk-averse policy optimization via risk-neutral policy optimization. *Artificial Intelligence*, 311:103765, 2022.

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

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

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

D. S. Brown, S. Niekum, and M. Petrik. Bayesian robust optimization for imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3509–3517, 2014.

Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making : A CVaR optimization approach. In *Neural Information Processing Systems (NIPS)*, 2015.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18: 1–51, 2018.

J. Cvitanić and I. Karatzas. On dynamic measures of risk. *Finance and Stochastics*, 1999.

B. Defourny, D. Ernst, and L. Wehenkel. Risk-aware decision making and dynamic programming. In *NIPS Workshop on Model Uncertainty and Risk in Reinforcement Learning*, 2008.

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

E. Delage, D. Kuhn, and W. Wiesemann. "Dice"-sion-making under uncertainty: When can a random decision reduce risk? *Management Science*, 65(7):3282–3301, 2019.

F. Delbaen. The structure of m–stable sets and in particular of the set of the risk neutral measures. *In Memoriam Paul-André Meyer*, 2006.

A. Dixit, M. Ahmadi, and J. W. Burdick. Risk-sensitive motion planning using entropic value-at-risk. In *European Control Conference (ECC)*, pages 1726–1732, 2021.

O. Dowson, D. P. Morton, and B. K. Pagnoncelli. Multistage stochastic programs with the entropic risk measure. *Preprint in Optimization Online*, 2021.

H. Follmer and A. Schied. *Stochastic Finance: Introduction in Discrete Time*. De Gruyter Graduate, fourth edition, 2016.

M. Frittelli and E. R. Gianin. Dynamic convex risk measure. *Risk measures for the 21st century*, 2004.

J. L. Hau and M. Petrik. Counterexamples to risk-averse dynamic program decompositions. *Arxiv*, 2023.

T. Hiraoka, T. Imagawa, T. Mori, T. Onishi, and Y. Tsuruoka. Learning robust options by conditional value at risk optimization. In *Neural Information Processing Systems (NeurIPS)*, 2019.

C. P. Ho, M. Petrik, and W. Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

D. A. Iancu, M. Petrik, and D. Subramanian. Tight approximations of dynamic risk measures. *Mathematics of Operations Research*, 40(3):655–682, 2015.

M. Kupper and W. Schachermayer. Representation results for law invariant time consistent functions. *Mathematics and Financial Economics*, 16(2):419–441, 2006.

X. Li, H. Zhong, and M. L. Brandeau. Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447, 2022.

E. A. Lobo, M. Ghavamzadeh, and M. Petrik. Soft-robust algorithms for batch reinforcement learning. *Arxiv*, 2021.

P. Massart. *Concentration Inequalities and Model Selection*. Springer, 2003.

G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *Arxiv*, 2017.

X. Ni and L. Lai. EVaR optimization for risk-sensitive reinforcement learning, 2022a.

X. Ni and L. Lai. Policy gradient based entropic-var optimization in risk-sensitive reinforcement learning. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2022b.

T. Osogami. Iterated risk measures for risk-sensitive Markov decision processes with discounted cost. In *Conference on Uncertainty in Artificial Intelligence*, pages 573–580, 2011.

T. Osogami. Robustness and risk-sensitivity in Markov decision processes. In *Advances in Neural Information Processing Systems*, 2012.

G. C. Pflug and A. Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. *Mathematics of Operations Research*, 41(2):682–699, 2016a.

G. C. Pflug and A. Pichler. Time-inconsistent multistage stochastic programs: Martingale bounds. *European Journal of Operational Research*, 249(1):155–163, 2016b.

G. C. Pflug and A. Ruszczyński. Measuring risk for income streams. *Computational Optimization and Applications*, 2005.

L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 252–260, 2013.

L. Prashanth and M. Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward mdps. *Machine Learning Journal*, 105(3):367–417, 2016.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

F. Riedel. Dynamic coherent risk measures. *Stochastic processes and their applications*, 2004.

S. M. Ross and E. A. Peköz. *A Second Course in Probability*. ProbabilityBookstore.com, 2007.

A. Ruszczynski. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming B*, 125(2):235–261, July 2010.

A. Santara, A. Naik, B. Ravindran, D. Das, D. Mudigere, S. Avancha, and B. Kaul. RAIL: Risk-averse imitation learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 2, 2018.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

L. N. Steimle, D. L. Kaufman, and B. T. Denton. Multi-model Markov decision processes. *IISE Transactions*, Forthcoming, 2021.

A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Elsevier*, 2008.

A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. *International Conference on Machine Learning*, 2012.

A. Tamar, D. D. Castro, and S. Mannor. Temporal difference methods for the variance of the reward to go. *International Conference on Machine Learning (ICML)*, 28: 495–503, 2013.

A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via Sampling. In *AAAI Conference on Artificial Intelligence*, pages 2993–2999, 2014.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy Gradient for Coherent Risk Measures. In *Neural Information Processing Systems (NIPS)*, 2015.

J. Zhang, A. S. Bedi, M. Wang, and A. Koppel. Cautious Reinforcement Learning via Distributional Risk in the Dual Domain. *IEEE Journal on Selected Areas in Information Theory*, 2(2):611–626, 2021.

A MONETARY RISK MEASURES

Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathbb{X}\colon \Omega \to \mathbb{R}$ be a space of $\mathcal{F}$-measurable functions (space of $\mathcal{F}$-measurable random variables).

A.1 Basic Definitions

Definition A.1 (Monetary Risk Measure)**.** A monetary risk measure is a function $\psi\colon \mathbb{X} \to \mathbb{R}$ that maps a random variable $X \in \mathbb{X}$ to real numbers and satisfies the following properties:

A1. Monotonicity:

$$X_1 \leq X_2 \text{ (a.s.)} \quad \Longrightarrow \quad \psi[X_1] \leq \psi[X_2], \quad \forall X_1, X_2 \in \mathbb{X},$$

A2. Translation invariance:

$$\psi[c + X] = c + \psi[X], \quad \forall c \in \mathbb{R},\ \forall X \in \mathbb{X}.$$

Monetary risk measures are called coherent when they satisfy consistency and concavity properties as defined below. Well known risk measures, like CVaR and EVaR are coherent.

Definition A.2 (Coherent risk measure)**.** A monetary risk measure $\psi\colon \mathbb{X} \to \mathbb{R}$ is *coherent* if it satisfies the following properties:

A3. Super-additivity:

$$\psi[X_1 + X_2] \geq \psi[X_1] + \psi[X_2], \ \ \forall X_1, X_2 \in \mathbb{X},$$

A4. Positive homogeneity:

$$\psi[c \cdot X] = c \cdot \psi[X], \ \ \forall c \in \mathbb{R}_+,\ \forall X \in \mathbb{X}$$

Concave risk measures, defined below, generalize the class of coherent risk measures by dropping the positive homogeneity requirement and replacing it with concavity.

Definition A.3 (Concave risk measure)**.** A monetary risk measure $\psi[:]\mathbb{X} \to \mathbb{R}$ is concave if it satisfies the following properties:

A5. Concavity:

$$\psi[c \cdot X_1 + (1-c)X_2] \geq c \cdot \psi[X_1] + (1-c) \cdot \psi[X_2], \quad \forall c \in [0,1],\ \forall X_1, X_2 \in \mathbb{X}.$$

Every coherent risk measure is a concave risk measure but a concave risk measure may not be coherent. For instance, the Entropic Risk Measure (ERM), defined below, is concave but not incoherent.

Next, we summarize some other important properties of monetary risk measures that are relevant to our work. A risk measure is *law invariant* if its value depends only on the probability distribution of the random variable as opposed also on the values the random variable assigns to particular elements of the probability space (Shapiro et al., 2014). A risk measure is *dynamically consistent* if it satisfies the tower property (Shapiro et al., 2014) and can be optimized using a dynamic program (Artzner et al., 2004; Cvitanić and Karatzas, 1999; Delbaen, 2006; Dowson et al., 2021; Frittelli and Gianin, 2004; Pflug and Ruszczyński, 2005; Riedel, 2004). Unfortunately, expectation and the minimum (Min) are the only coherent risk measures that are law invariant, dynamically consistent.

A.2 Value-at-Risk

For a random variable $X \in \mathbb{X}$, its value-at-risk with a confidence level $\alpha \in (0,1)$, denoted by $\mathrm{VaR}_\alpha[X]$, is the $(1-\alpha)$-quantile of X:

$$\begin{aligned}\mathrm{VaR}_\alpha[X] &= \inf\{x \in \mathbb{R} \mid \mathbb{P}[X \leq x] > 1-\alpha\} \\ &= \sup\{x \in \mathbb{R} \mid \mathbb{P}[X < x] \leq 1-\alpha\} \\ &= F_X^{-1}(x),\end{aligned}$$

where F_X is the cumulative distribution function (cdf) of X. The last equality holds only when F_X^{-1} exists.

A.3 Conditional Value-at-Risk

For a random variable $X \in \mathbb{X}$, its conditional value-at-risk $\mathrm{CVaR}_\alpha[X]$ with a confidence level $\alpha \in (0,1)$ is defined as the expectation of the worst $(1-\alpha)$-fraction of X, and can be computed as the solution of the following optimization problem:

$$\mathrm{CVaR}_\alpha[X] = \sup_{\zeta \in \mathbb{R}} \left(\zeta - \frac{1}{1-\alpha} \cdot \mathbb{E}\big[(\zeta - X)_+\big] \right).$$

It is easy to see that $\mathrm{CVaR}_0[X] = \mathbb{E}[X]$ and $\lim_{\alpha \to 1} \mathrm{CVaR}_\alpha[X] = \operatorname{ess\,inf}[X]$, where the *essential infimum* of X is defined as $\operatorname{ess\,inf}[X] = \sup_{\zeta \in \mathbb{R}} \mathbb{P}[X < \zeta] = 0$.

A.4 Entropic Risk Measure

For a random variable $X \in \mathbb{X}$, its entropic risk measure $\mathrm{ERM}_\beta[X]$ with the risk parameter $\beta \in (0, \infty)$ is defined as

$$\mathrm{ERM}_\beta[X] \;=\; -\frac{1}{\beta} \log\left(\mathbb{E}[e^{-\beta X}]\right), \quad \beta > 0.$$

The definition is extended to the interval $[0,\infty) \cup \{\infty\}$ as

$$\mathrm{ERM}_0[X] = \mathbb{E}[X]$$
$$\mathrm{ERM}_\infty[X] = \operatorname{ess\,inf}[X].$$

We also need a conditional ERM to construct the dynamic programs. This is defined as follows.

Definition A.4. The conditional ERM is defined for $X_1, X_2 \in \mathbb{X}$ as

$$\mathrm{ERM}_\beta[X_1 \mid X_2] \;=\; -\frac{1}{\beta} \log\left(\mathbb{E}[e^{-\beta X_1} \mid X_2]\right).$$

The following proposition shows that ERM indeed is not a coherent risk measure because it violates the assumption A4 in Definition A.2.

Proposition A.5. *There exists a random variable X such that* $\mathrm{ERM}_\beta[c \cdot X] \neq c \cdot \mathrm{ERM}_\beta[X]$.

The following lemma plays a crucial role in efficiently computing EVaR, defined below, which can be expressed in terms of ERM.

Lemma A.6 ((Ahmadi-Javid, 2012))**.** *The function $t \mapsto \mathrm{ERM}_{t^{-1}}[X]$ for any random variable $X \in \mathbb{X}$ and $t > 0$ is concave and non-decreasing.*

We use the following lemma in the analysis of EVaR solution approximation by discretization in this paper.

Lemma A.7. *The function $\beta \mapsto \mathrm{ERM}_\beta[X]$ for any random variable $X \in \mathbb{X}$ and $\beta > 0$ is continuous and non-increasing.*

The following lemma, which represents a new result to the best of our knowledge, plays an important role in bounding the difference between ERM and the expectation. This result serves to bound the error of replacing the risk-averse value function by a risk-neutral value function in Algorithm 1.

Lemma A.8. *Let $X \in \mathbb{X}$ be a bounded random variable such that $x_{\min} \leq X \leq x_{\max}$ a.s. Then, for any risk level $\beta > 0$,* $\mathrm{ERM}_\beta[\cdot]$ *can be bounded as*

$$\mathbb{E}[X] - \frac{\beta(x_{\max} - x_{\min})^2}{8} \leq \mathrm{ERM}_\beta[X] \leq \mathbb{E}[X].$$

Proof. Recall that the Hoeffding's lemma shows that for any $\forall \lambda \in \mathbb{R}$, we may write (Boucheron et al., 2013; Massart, 2003)

$$0 < \mathbb{E}[e^{\lambda X}] \leq \exp\left(\lambda \cdot \mathbb{E}[X] + \frac{\lambda^2 (x_{\max} - x_{\min})^2}{8}\right).$$

Applying log to both sides of the inequality above gives

$$\log\left(\mathbb{E}[e^{\lambda X}]\right) \leq \lambda \mathbb{E}[X] + \frac{\lambda^2 (x_{\max} - x_{\min})^2}{8}.$$

Then, variable substitution $\lambda = -\beta$ and algebraic manipulation shows that

$$\log\left(\mathbb{E}[e^{-\beta X}]\right) \leq -\beta \cdot \mathbb{E}[X] + \frac{\beta^2 (x_{\max} - x_{\min})^2}{8}$$
$$-\frac{1}{\beta}\log\left(\mathbb{E}[e^{-\beta X}]\right) \geq \mathbb{E}[X] - \frac{\beta (x_{\max} - x_{\min})^2}{8}$$

Substituting the definition of ERM into the inequality above yields then the first desired inequality:

$$\mathbb{E}[X] - \frac{\beta (x_{\max} - x_{\min})^2}{8} \leq \mathrm{ERM}_\beta\left[X\right] \ .$$

The second inequality in the lemma's statement, $\mathrm{ERM}_\beta\left[X\right] \leq \mathbb{E}[X]$, follows immediately from the Donsker-Varadhan's Variational Formula. □

The following lemma helps to show that a deterministic policy can attain the same return as a randomized policy when the objective is an ERM. This result is not surprising and derives from the fact that the $\mathrm{ERM}_\beta\left[X\right] \leq \max_{\omega\in\Omega} X$ for any random variable $X \in \mathbb{X}$ defined over a finite probability space.

Lemma A.9. *Let $X\colon \Omega \to \mathcal{A}$ be a random variable defined over a finite action set $\mathcal{A}$ and let $g\colon \mathcal{A} \to \mathbb{R}$ be a function defined for each action. Then, for any $\beta \geq 0$, we have*

$$\max_{a\in\mathcal{A}} g(a) \ = \ \max_{d\in\Delta^A} \mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right] \ .$$

Proof. We first prove that $\max_{a\in\mathcal{A}} g(a) \leq \max_{d\in\Delta^A} \mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right]$. Let $a^\star \in \arg\max_{a\in\mathcal{A}} g(a)$ be an optimal action. We now construct a policy $\bar{d} \in \Delta^A$ as $\bar{d}(a^\star) = 1$ and $\bar{d}(a) = 0,\ \forall a \in \mathcal{A} \setminus \{a^\star\}$. Substituting $\bar{d}$ in the definition of ERM yields that

$$\begin{aligned}
\mathrm{ERM}_\beta\left[g(X) \mid X \sim \bar{d}\right] &= -\beta^{-1} \cdot \log\Big(\mathbb{E}\left[\exp\left(-\beta \cdot g(X)\right) \mid X \sim \bar{d}\right]\Big) \\
&= -\beta^{-1} \cdot \log\Big(\exp\left(-\beta \cdot g(a^\star)\right)\Big) \\
&= g(a^\star) \ .
\end{aligned} \tag{21}$$

Using (21) and the fact that $\bar{d}$ is a valid probability distribution in Δ^A, we obtain the desired inequality as

$$\max_{a\in\mathcal{A}}\ g(a) = g(a^\star) = \mathrm{ERM}_\beta\left[g(X) \mid X \sim \bar{d}\right] \leq \max_{d\in\Delta^A}\ \mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right] \ .$$

To prove the converse inequality $\max_{a\in\mathcal{A}} g(a) \ \geq \ \max_{d\in\Delta^A} \mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right]$, we define $d^\star$ as an optimal distribution $d^\star \ \in \ \arg\max_{d\in\Delta^A} \mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right]$. It will be convenient to use the dual representation of $\mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right]$, which for for any $d \in \Delta^A$ is defined as (see e.g., (Ahmadi-Javid, 2012))

$$\mathrm{ERM}_\beta\left[g(X) \mid X \sim d\right] \ = \ \inf_{\bar{d}\in\Delta^A,\, \bar{d}\ll d} \left\{\mathbb{E}[g(X) \mid X \sim \bar{d}] + \frac{1}{\beta}\,\mathrm{KL}(\bar{d}\|d)\right\} \ ,$$

where KL is the KL-divergence and $\ll$ denotes the absolute continuity of probability measures. Using this dual representation, we get the following upper-bound on $\mathrm{ERM}_\beta\left[g(X) \mid X \sim d^\star\right]$:

$$\begin{aligned}
\mathrm{ERM}_\beta\left[g(X) \mid X \sim d^\star\right] &= \inf_{\bar{d}\in\Delta^A} \left\{\mathbb{E}[g(X) \mid X \sim \bar{d}] + \frac{1}{\beta}\,\mathrm{KL}(\bar{d}\|d^\star) \mid \bar{d} \ll d^\star\right\} \\
&\leq \mathbb{E}[g(X) \mid X \sim d^\star] + \frac{1}{\beta}\,\mathrm{KL}(d^\star\|d^\star) \\
&\overset{\text{(a)}}{=} \mathbb{E}[g(X) \mid X \sim d^\star] \\
&\overset{\text{(b)}}{\leq} \max_{a\in\mathcal{A}} g(a) \ ,
\end{aligned}$$

where **(a)** holds because $\mathrm{KL}(d\|d) = 0$, and **(b)** follows because A is finite, and thus, for each $d \in \Delta^A$, we have

$$\max_{a\in\mathcal{A}} g(a) \geq \mathbb{E}\left[g(X) \mid X \sim d\right] \ .$$

This proves the second desired inequality since $d^\star \in \Delta^A$ and concludes the proof. □

The result in Lemma A.9 can be further generalized to a broader class of risk measures (Delage et al., 2019).

A.5 Entropic Value-at-Risk

For a random variable $X \in \mathbb{X}$, its entropic value-at-risk with $\mathrm{EVaR}_\alpha[X]$ confidence level $\alpha \in (0,1)$ is defined as

$$\mathrm{EVaR}_\alpha[X] = \sup_{\beta>0} \left(\mathrm{ERM}_\beta[X] + \frac{\log(1-\alpha)}{\beta} \right). \tag{22}$$

It is easy to see that $\mathrm{EVaR}_0[X] = \mathbb{E}[X]$ and $\lim_{\alpha\to 1} \mathrm{EVaR}_\alpha[X] = \operatorname{ess\,inf}[X]$. In addition, EVaR is a non-increasing function in α and is bounded as:

$$\operatorname{ess\,inf}[X] \le \mathrm{EVaR}_\alpha[X] \le \mathbb{E}[X].$$

EVaR was proposed as a tightest Chernoff-style lower bound on the popular VaR risk measure with the same confidence level α. It is also a lower bound CVaR as the following lemma shows.

Lemma A.10 (proposition 3.2 in (Ahmadi-Javid, 2012))**.** *The following inequalities hold for any $\alpha \in (0,1)$ and a random variable $X \in \mathbb{X}$:*

$$\mathrm{EVaR}_\alpha[X] \le \mathrm{CVaR}_\alpha[X] \le \mathrm{VaR}_\alpha[X].$$

The following lemma, which shows how the optimal solution of (22) scales with the scale of the random variable is necessary when analyzing the properties of EVaR solutions.

Lemma A.11. *Suppose that the supremum in* (22) *is attained by some $\beta^\star > 0$ for a random variable X. Then, the supremum in* (22) *is attained at $c^{-1}\cdot\beta^\star$ for any random variable $c\cdot X$ and a constant $c > 0$.*

Proof. Using positive quasi-homogeneity (Theorem 3.1) of ERM and algebraic manipulation, we can reformulate (22) for $c\cdot X$ as

$$\begin{aligned}
\mathrm{EVaR}_\alpha[c\cdot X] &= \sup_{\beta>0}\left(\mathrm{ERM}_\beta[c\cdot X] + \frac{\log(1-\alpha)}{\beta}\right)\\
&= \sup_{\beta>0}\left(c\cdot\mathrm{ERM}_{c\cdot\beta}[X] + \frac{\log(1-\alpha)}{\beta}\right)\\
&= c\cdot\sup_{\beta>0}\left(\mathrm{ERM}_{c\cdot\beta}[X] + \frac{\log(1-\alpha)}{c\cdot\beta}\right)\\
&= c\cdot\sup_{\beta>0}\left(\mathrm{ERM}_{c\cdot\beta}[X] + \frac{\log(1-\alpha)}{c\cdot\beta}\right)\\
&= c\cdot\sup_{\tau>0}\left(\mathrm{ERM}_{\tau}[X] + \frac{\log(1-\alpha)}{\tau}\right)\\
&= c\cdot\mathrm{EVaR}_\alpha[X]
\end{aligned}$$

We used the variable substitution $\tau = c\cdot\beta$. Therefore, if the supremum is attained at $\tau^\star$ for $c\cdot X$, it is attained at $\beta^\star = c^{-1}\cdot\tau^\star$ for $c\cdot X$.

Note that the derivation above also confirms that EVaR is positively homogeneous. □

B PROOFS OF SECTION 2

The following proposition states a simple, but important property of the expectation operator which plays a crucial role in formulating the dynamic programs. The property is known under several different names, including *the tower property*, *the law of total expectation*, and *the law of iterated expectations*.

Proposition B.1 (Tower Property for Expectation (e.g., Proposition 3.4 in (Ross and Peköz, 2007))). *Any two random variables $X_1, X_2 \in \mathbb{X}$ satisfy that*

$$\mathbb{E}[X_1] \;=\; \mathbb{E}\left[\mathbb{E}[X_1 \mid X_2]\right] \,.$$

A convenient way to represent ERM is to use its *certainty equivalent* form. This form relates the risk measure to the popular expected utility framework for decision-making (Ben-Tal, 2007). In the expected utility framework, one prefers a lottery (or a random reward) $X_1 \in \mathbb{X}$ over $X_2 \in \mathbb{X}$ if and only if

$$\mathbb{E}[u(X_1)] \geq \mathbb{E}[u(X_2)] \,,$$

for some increasing *utility function* $u\colon \mathbb{R} \to \mathbb{R}$.

The expected utility $\mathbb{E}[u(X)]$ is difficult to interpret because its units are incompatible with X. A more interpretable characterization of the expected utility is to use the *certainty equivalent* $z \in \mathbb{R}$, which is defined as the certain quantity that achieves the same expected utility as X:

$$\mathbb{E}[u(z)] = \mathbb{E}[u(X)], \quad \text{and therefore,} \quad z = u^{-1}(\mathbb{E}[u(X)]) \,. \tag{23}$$

Algebraic manipulation from (23) then shows that ERM for any $X \in \mathbb{X}$ can be represented as the certainty equivalent

$$\mathrm{ERM}_\beta\,[X] \;=\; u^{-1}(\mathbb{E}[u(X)]) \,, \tag{24}$$

for the utility function $u\colon \mathbb{R} \to \mathbb{R}$ (see definition 2.1 in (Ben-Tal, 2007)) defined as

$$u(x) = \beta^{-1} - \beta^{-1} \cdot \exp\left(-\beta \cdot x\right) \,.$$

Because the function u is strictly increasing, its inverse $u^{-1}\colon \mathbb{R} \to \mathbb{R}$ exists and equals to

$$u^{-1}(z) = -\beta^{-1} \cdot \log(1 - \beta \cdot z) \,.$$

Proof of Theorem 2.1. The property is trivially true when $\beta = 0$ from Proposition B.1 since $\mathrm{ERM}_0\,[=]\,\mathbb{E}$. The property then follows by algebraic manipulation for $\beta > 0$ using the certainty equivalent representation in (24) as

$$\begin{aligned}
\mathrm{ERM}_\beta\left[\mathrm{ERM}_\beta\left[X_1 \mid X_2\right]\right] &= \mathrm{ERM}_\beta\left[u^{-1}\left(\mathbb{E}[u(X_1) \mid X_2]\right)\right] \\
&= u^{-1}\left(\mathbb{E}\left[u\left(u^{-1}\left(\mathbb{E}[u(X_1) \mid X_2]\right)\right)\right]\right) \\
&= u^{-1}(\mathbb{E}\left[\mathbb{E}[u(X_1) \mid X_2]\right]) \\
&= u^{-1}(\mathbb{E}[u(X_1)]) && \text{Proposition B.1} \\
&= \mathrm{ERM}_\beta\left[X_1\right] \,.
\end{aligned}$$

□

C PROOFS OF SECTION 3

We start this section by reporting the pseudo-code for the value iteration (VI) algorithm in finite-horizon ERM-MDP (Algorithm 3). This algorithm is an adaptation of the standard VI algorithm to the finite-horizon ($T < \infty$) setting. It uses the results of Theorems 3.2 and 3.3, and first computes the optimal value function $v_t^\star$ backwards in time ($t = T, T-1, \ldots, 0$) according to (14), and then obtains the optimal policy as a policy greedy to $v^\star$ by solving the optimization (15).

Algorithm 3: VI for finite-horizon ERM-MDP

Input: Horizon $T < \infty$, risk level $\beta > 0$, terminal value $v_T(s),\ \forall s \in \mathcal{S}$
Output: Optimal value $(v_t^\star)_{t=0}^T$ and policy $(\pi_t^\star)_{t=0}^{T-1}$
Initialize $v_T^\star(s) \leftarrow v'(s),\ \forall s \in \mathcal{S}$;
for $t = T{-}1{:}0$ **do**
  Update $v_t^\star$ using (14) and $\pi_t^\star$ using (15);
return $v^\star, \pi^\star$;

Proof of Theorem 3.1. The property is trivially true for $c = 0$ or $\beta = 0$ because $\mathrm{ERM}_\beta[0] = 0$ and $\mathrm{ERM}_0[\cdot] = \mathbb{E}[\cdot]$. For $c > 0$ and $\beta > 0$, the property follows by rearranging the terms as

$$\mathrm{ERM}_{\beta\cdot c}[X] = -\frac{1}{\beta c}\log\left(\mathbb{E}[e^{-\beta\cdot c\cdot X}]\right) \implies c\cdot\mathrm{ERM}_{\beta\cdot c}[X] = -\frac{1}{\beta}\log\left(\mathbb{E}[e^{-\beta\cdot c\cdot X}]\right) = \mathrm{ERM}_\beta[c\cdot X].$$

□

Proof of Theorem 3.2. The proof is divided into two parts: the proof for v^π (Eq. 13) and a proof for $v^\star$ (Eq. 14).

Proof for v^π: For any fixed $\pi \in \Pi_{MR}$, we prove the claim for v^π by backward induction on t from $t = T$ to $t = 0$. The base case of the induction with $t = T$ is trivial because by definition $v_T(s) = 0,\ \forall s \in \mathcal{S}$. To prove the inductive step, we first assume that any function $v_{t'},\ t' = t+1{:}T$ defined by (11) satisfies (13), and then show that the same is true for v_t^π. By the induction hypothesis we can substitute the definition of $v_{t+1}^\pi(S')$ from (11) into (13) and write

$$\begin{aligned}
v_t^\pi(s) &= \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma\cdot v_{t+1}^\pi(S')\right] \\
&= \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma\cdot\mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[\mathfrak{R}_{t+1:T}^\pi(S')\right]\right] \\
&= \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma\cdot\mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[\sum_{t'=t+1}^{T-1}\gamma^{t'-t-1}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S'\right]\right] \\
&\overset{(a)}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t+1}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S'\right]\right] \\
&= \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(S_t,A_t) + \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t+1}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S', A_t\right] \mid S_t = s\right] \\
&\overset{(b)}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(S_t,A_t) + \sum_{t'=t+1}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S', A_t\right] \mid S_t = s\right] \\
&= \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathrm{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S', A_t\right] \mid S_t = s\right] = \\
&\overset{(c)}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_t = s\right] = \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathfrak{R}_{t:T}^\pi(s)\right]
\end{aligned}$$

where **(a)**, **(b)**, and **(c)** come from the positive quasi-homogeneity (Theorem 3.1), translation invariance (A2 in Definition A.1), and tower (Theorem 2.1) properties of ERM. This derivation proves the inductive step and shows that any function v^π that satisfies the Bellman equation in (13) satisfies the definition of value function in (11), and thus, is unique.

Proof for $v^\star$: The proof of the Bellman equation for the optimal value function $v^\star$ proceeds by backward induction analogously to the proof of (13) with the difference that it incorporates the optimization over actions. As before, the base case with $t = T$ is trivial because $v_T^\star(s) = 0,\ \forall s \in \mathcal{S}$ by definition. To prove the inductive step, we first assume that any function $v_{t'}^\star,\ t' : t+1{:}T$ defined by (12) satisfies (14), and then show that the same is true for $v_t^\star$.

In the proof of the inductive step, we use Lemma A.9, which shows how the maximization over actions can be replaced by a maximization over randomized policies that are distributions over actions as

$$\begin{aligned} v_t^\star(s) &= \max_{a\in\mathcal{A}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma\cdot v_{t+1}^\star(S') \mid A_t = a\right] \\ &= \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma\cdot v_{t+1}^\star(S') \mid A_t \sim d\right]\ . \end{aligned} \tag{25}$$

By the induction hypothesis we can substitute the definition of $v_{t+1}^\star(S')$ from (12) into (25) and write

$$\begin{aligned} v_t^\star(s) &= \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma \max_{\pi\in\Pi_{MR}^{t+1:T}} \operatorname{ERM}_{\beta\cdot\gamma^{t+1}}\left[\mathfrak{R}_{t:T}^\pi(s)\right]\right] \\ &= \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \gamma \max_{\pi\in\Pi_{MR}^{t+1:T}} \operatorname{ERM}_{\beta\cdot\gamma^{t+1}}\left[\sum_{t'=t+1}^{T-1}\gamma^{t'-t-1} r(S_{t'},A_{t'}) \mid S_{t+1} = S'\right]\right] \\ &\overset{\text{(a)}}{=} \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(s,A_t) + \max_{\pi\in\Pi_{MR}^{t+1:T}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t+1}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S'\right]\right] \\ &= \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(S_t,A_t) + \max_{\pi\in\Pi_{MR}^{t+1:T}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t+1}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S', A_t\right] \mid S_t = s\right] \\ &\overset{\text{(b)}}{=} \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\max_{\pi\in\Pi_{MR}^{t+1:T}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[r(S_t,A_t) + \sum_{t'=t+1}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S', A_t\right] \mid S_t = s\right] \\ &= \max_{d\in\Delta^A} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\max_{\pi\in\Pi_{MR}^{t+1:T}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_{t+1} = S', A_t\right] \mid S_t = s\right] \\ &\overset{\text{(c)}}{=} \max_{\pi\in\Pi_{MR}^{t:T}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\sum_{t'=t}^{T-1}\gamma^{t'-t}\cdot r(S_{t'},A_{t'}) \mid S_t = s\right] = \max_{\pi\in\Pi_{MR}^{t:T}} \operatorname{ERM}_{\beta\cdot\gamma^t}\left[\mathfrak{R}_{t:T}^\pi(s)\right], \end{aligned}$$

where **(a)** comes from the positive quasi-homogeneity (Theorem 3.1) property of ERM, **(b)** comes from the translation invariance (A2 in Definition A.1) property of ERM, and **(c)** comes from the monotonicity (A1 in Definition A.1) and tower (Theorem 2.1) properties of ERM. This derivation proves the inductive step and shows that any function $v^\star$ that satisfies the Bellman equation in (14) satisfies the definition of value function in (12), and thus, is unique. □

Proof of Theorem 3.3. Following the notation of chapter 4 in (Puterman, 2005), let $\mathcal{H}_t$ be the set of all histories up to time t inclusively. Let the optimal history-dependent value function be $u_t^\star\colon \mathcal{H}_t \to \mathbb{R}, t = 0{:}T{-}1$. The value function $u^\star = (u_t^\star)_{t=0}^T$ is achieved by the optimal history-dependent policy because the state and actions are finite, and thus, the space of randomized history-dependent policies is compact.

The proof proceeds in three steps.

(i) First, we show that $v^\star$ attains the return of the optimal history-dependent value function:

$$u_t^\star(h_t) = v_t^\star(s_t) \quad \forall h_t \in \mathcal{H}_t,\ t = 0{:}T{-}1\ ,$$

where s_t is the t-th and final state in the history h_t. This result is a consequence of the dynamic programming formulation in Theorem 3.2. An argument analogous to the proof of Theorem 4.4.2(a) in (Puterman, 2005) shows that $u_t^\star(h_t)$ depends only on s_t, which is the final state in the history h_t.

Using the standard backward-induction argument on t, we assume that $u_{t+1}^\star(h_{t+1}) = v_{t+1}^\star(s_{t+1})$ holds and then prove that $u_t^\star(h_t) = v_t^\star(s_t)$. Let $d^\star \in \Delta^A$ be the part of a decision rule achieving $u^\star$ that decides about the actions that should

be taken at time-step t. Applying Theorem 3.2 to the optimal history-dependent value function $u^\star$, we may write

$$\begin{aligned} u_t^\star(h_t) &\overset{\text{(a)}}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s_t,A_t)+\gamma\cdot u_{t+1}^\star((h_t,A,S')) \mid A_t\sim d^\star\right] \\ &\overset{\text{(b)}}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s_t,A_t)+\gamma\cdot v_{t+1}^\star(S') \mid A_t\sim d^\star\right] \\ &= \max_{d\in\Delta^{\mathcal{A}}} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s_t,A_t)+\gamma\cdot v_{t+1}^\star(S') \mid A_t\sim d\right] \\ &\overset{\text{(c)}}{=} \max_{a\in\mathcal{A}} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s_t,a)+\gamma\cdot v_{t+1}^\star(S')\right] \overset{\text{(d)}}{=} v_t^\star(s_t)\ , \end{aligned}$$

where **(a)** follows from the fact that (i) the reward function depends only on the current state and not the full history and (ii) the history at time-step $t+1$, h_{t+1} is constructed by appending action A_t and state S' to the history at time-step t: $h_{t+1}=(h_t,A_t,S')$, **(b)** comes from the inductive hypothesis, **(c)** is the result of Lemma A.9, and **(d)** comes from (14) in Theorem 3.2. This result shows that the optimal history-dependent randomized $u^\star$ and the optimal Markov deterministic $v^\star$ value functions are equal in ERM-MDP.

(ii) The second part of the proof is to show that the value function of any (optimal) policy $\pi^\star\in\Pi_{MD}$ that is a solution to the optimization problem (8) is equal to the optimal value function $v^\star$:

$$v_t^{\pi^\star}(s)=v_t^\star(s),\quad \forall s\in\mathcal{S},\ t=0{:}T\ .$$

This result follows using the standard backward induction argument and algebraic manipulation from Theorem 3.2. The derivation relies on the fact that $\mathcal{A}$ is finite and the maximum in (14) exists and is achievable.

(iii) Here we show that any greedy policy to the optimal value function $v^\star$ is an optimal policy, that is, a solution to the optimization problem (8). This is trivial from parts (i) and (ii) because the value function of the greedy policy to $v^\star$ is $v^\star$, and we know from part (ii) that any policy that solves (8) also has value $v^\star$. Thus, the greedy policy is in fact optimal. □

Proof of Theorem 3.4. It is important to reiterate the following definitions:

1. $\pi^\star=\{\pi_t^\star\}_{t=0}^\infty$ and $v^\star=v^{\pi^\star}=\{v_t^\star\}_{t=0}^\infty$ are the optimal policy and optimal value function of the ERM-MDP in the infinite-horizon discounted setting with $T=\infty$ and $\gamma\in(0,1)$. In other words, $\pi^\star$ is a solution to the ERM-MDP optimization (8).
2. $\pi_\infty^\star$ and $v_\infty^\star=v_\infty^{\pi_\infty^\star}$ are the optimal policy and value function of the MDP in the risk-neutral infinite-horizon discounted setting.
3. $\tilde\pi^\star=\{\tilde\pi_t^\star\}_{t=0}^{T'}$ and $\tilde v^\star=v^{\tilde\pi^\star}=\{\tilde v_t^\star\}_{t=0}^{T'}$ are the optimal policy and value function of the finite-horizon discounted ERM-MDP with risk level β, discount factor γ, horizon T', and the value function at horizon T' set to $\tilde v_{T'}^\star=v_\infty^\star$. In other words, $\tilde\pi_t^\star$ and $\tilde v^\star$ are the outputs of Algorithm 3 with $T=T'$ and $v'=v_\infty^\star$.
4. $\hat\pi^\star=\{\hat\pi_t^\star\}_{t=0}^\infty$ and its value $\hat v^\star=\{\hat v_t^\star\}_{t=0}^\infty$ are the policy and value function returned by Algorithm 1 that are constructed as follows:

$$\hat\pi_t^\star=\begin{cases}\tilde\pi_t^\star & \text{if } 0\le t<T',\\ \pi_\infty^\star & \text{if } t\ge T'.\end{cases}\qquad\qquad \hat v_t^\star=v_t^{\hat\pi^\star}=\begin{cases}\tilde v_t^\star=v_t^{\tilde\pi^\star} & \text{if } 0\le t<T',\\ v_\infty^\star=v_\infty^{\pi_\infty^\star} & \text{if } t\ge T'.\end{cases}$$

Our goal is to prove an upper-bound on the difference between the $\mathrm{ERM}_\beta[\cdot]$ of the returns of policies $\pi^\star$ and $\hat\pi^\star$:

$$\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\pi^\star}\right]-\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\hat\pi^\star}\right]\qquad\text{or equivalently}\qquad v_0^{\pi^\star}(s_0)-v_0^{\hat\pi^\star}(s_0)\ .$$

As the first step of the proof, we bound the difference between the infinite-horizon γ-discounted risk-neutral and ERM value functions of any policy π, at any time-step $t=0{:}\infty$, and any state $s\in\mathcal{S}$ as

$$\begin{aligned} 0\overset{\text{(a)}}{\le} v_\pi^\infty(s)-v_t^\pi(s) &= \mathbb{E}[\mathfrak{R}_{0:\infty}^\pi(s)]-\mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathfrak{R}_{t:\infty}^\pi(s)\right] \\ &= \mathbb{E}[\mathfrak{R}_{0:\infty}^\pi(s)]-\mathrm{ERM}_{\beta\cdot\gamma^t}\left[\mathfrak{R}_{0:\infty}^\pi(s)\right]\overset{\text{(b)}}{\le}\frac{\beta\cdot\gamma^t\cdot\Delta_{\mathfrak{R}}^2}{8}\ . \end{aligned} \tag{26}$$

(a) holds because $\mathbb{E}$ is an upper-bound on the ERM, and **(b)** follows from the result of Lemma A.8.

We now bound the difference between the optimal value function of the infinite-horizon discounted ERM-MDP, $v^\star_{T'}$, and the value function returned by Algorithm 1, $v^{\hat{\pi}^\star}_{T'}$, at the planning horizon T' and for any state $s \in \mathcal{S}$ as follows:

$$\begin{aligned} v^\star_{T'}(s) - v^{\hat{\pi}^\star}_{T'}(s) &\overset{\text{(a)}}{\leq} v^\star_{T'}(s) - v^\infty_{\hat{\pi}^\star_{T'}}(s) + \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8} \\ &\overset{\text{(b)}}{\leq} v^\infty_{\pi^\star}(s) - v^\infty_{\hat{\pi}^\star_{T'}}(s) + \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8} \\ &\overset{\text{(c)}}{\leq} \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8} \,. \end{aligned}$$

(a) follows from the RHS of (26) and the fact that $\hat{\pi}^\star_t = \pi^\star_\infty,\ \forall t \geq T'$, **(b)** comes from the fact that using the LHS of (26), we have $0 \leq v^\star_\infty - v^{\pi^\star}_{T'}$, and **(c)** is true because $\hat{\pi}^\star_{T'}(s) = \pi^\star_\infty(s) \in \arg\max_{\pi \in \Pi_{MR}}\ v^\pi_\infty(s), \forall s \in \mathcal{S}$, and thus, $v^{\pi^\star}_\infty(s) - v^{\hat{\pi}^\star_{T'}}_\infty(s)$ is negative for all s.

As the second step of the proof, we construct the value function $u_t \in \mathbb{R}^S$ from $\hat{\pi}^\star$ for all $t \in 0{:}T'$ and $s \in \mathcal{S}$ as

$$\begin{aligned} u_{T'}(s) &= v^{\hat{\pi}^\star_{T'}}_\infty(s) - \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8} = v^{\pi^\star_\infty}_\infty(s) - \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8} = v^\star_\infty(s) - \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8}, \\ u_t(s) &= \max_{a \in \mathcal{A}} \ \mathrm{ERM}_{\beta \cdot \gamma^t}\left[r(s,a) + \gamma \cdot u_{t+1}(S'_{t+1,a})\right] \\ &= \mathrm{ERM}_{\beta \cdot \gamma^t}\left[r(s, \hat{\pi}^\star_t(s)) + \gamma \cdot u_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] \,, \end{aligned} \tag{27}$$

where $S'_{t+1,a}$ denotes the random variable representing the state at time $t+1$ that follows by taking action $a \in \mathcal{A}$ in state s at time t.

Note that u_t has been constructed such that **(i)** it is a lower-bound on $\hat{v}^\star_t = v^{\hat{\pi}^\star}_t$ and **(ii)** $\hat{\pi}^\star$ is its greedy policy.

(i) is true because $v^{\hat{\pi}^\star}_t = v^{\tilde{\pi}^\star}_t = \tilde{v}^\star_t$, which is the optimal finite-horizon ERM value function when we set $\tilde{v}^\star_{T'} = v^\star_\infty$, and u_t has been constructed as the optimal finite-horizon ERM value function when we set its value at the planning horizon T' by a lower-bound of $v^\star_\infty$ from (26): $u_{T'} = v^\star_\infty - \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8}$. We now provide a fomal proof for this. From (26), we have $v^{\hat{\pi}^\star}_{T'}(s) \geq u_{T'}(s)$ for all $s \in \mathcal{S}$. Then, assuming $v^{\hat{\pi}^\star}_{t+1}(s) \geq u_{t+1}(s)$ for all $s \in \mathcal{S}$ (inductive hypothesis), we can use backward induction on t to show that for all $s \in \mathcal{S}$, we have

$$\begin{aligned} v^{\hat{\pi}^\star}_t(s) - u_t(s) &= \mathrm{ERM}_{\beta \cdot \gamma^t}\left[r(s, \hat{\pi}^\star_t(s)) + \gamma \cdot v^{\hat{\pi}^\star}_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] - \mathrm{ERM}_{\beta \cdot \gamma^t}\left[r(s, \hat{\pi}^\star_t(s)) + \gamma \cdot u_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] \\ &\overset{\text{(a)}}{=} \mathrm{ERM}_{\beta \cdot \gamma^t}\left[\gamma \cdot v^{\hat{\pi}^\star}_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] - \mathrm{ERM}_{\beta \cdot \gamma^t}\left[\gamma \cdot u_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] \\ &\overset{\text{(b)}}{=} \gamma \cdot \left(\mathrm{ERM}_{\beta \cdot \gamma^{t+1}}\left[v^{\hat{\pi}^\star}_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] - \mathrm{ERM}_{\beta \cdot \gamma^{t+1}}\left[u_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right]\right) \\ &\overset{\text{(c)}}{\geq} 0 \,. \end{aligned} \tag{28}$$

(a) is by subtracting the constant reward from both terms. This can be done because ERM is translation invariant. **(b)** follows from the positive quasi-homogeneity of ERM (see Theorem 3.1). **(c)** follows from the monotonicity of ERM and the inductive hypothesis.

(ii) is true because $\hat{\pi}^\star_t$ is a greedy policy to $\hat{v}^\star_t$, and since subtracting a constant from all states does not change the greedy policy, it is also a greedy policy to u_t. The last equality in (27) is the result of this fact.

As the third step of the proof, we show that for each $s \in \mathcal{S}$ and $t = 0{:}T'$, we have

$$v^\star_t(s) - u_t(s) \leq \gamma^{T'-t} \frac{\beta \cdot \gamma^{T'} \cdot \Delta^2_{\mathfrak{R}}}{8} \,. \tag{29}$$

To prove (29) by induction, we first show that the inequality (29) holds for $t = T'$, that is, $v^\star_{T'}(s) - u_{T'}(s) \leq$

$\frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8}$, $\forall s \in \mathcal{S}$, as follows:

$$\begin{aligned} v^\star_{T'}(s) - u_{T'}(s) &= v^\star_{T'}(s) - v^\star_\infty(s) + \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8} \\ &= v^{\pi^\star}_{T'}(s) - v^\star_\infty(s) + \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8} \\ &\overset{\text{(a)}}{\leq} v^{\pi^\star}_\infty(s) - v^\star_\infty(s) + \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8} \overset{\text{(b)}}{\leq} \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8}. \end{aligned} \tag{30}$$

(a) follows from the LHS of (26) that $0 \leq v^{\pi^\star}_\infty(s) - v^{\pi^\star}_{T'}(s)$, and **(b)** comes from the fact that $v^\star_\infty$ is the optimal value function of the infinite-horizon discounted risk-neutral MDP, and thus, $v^\star_\infty(s) \leq v^{\pi^\star}_\infty(s)$.

Now assuming that (29) holds for $t+1$ for each $s \in \mathcal{S}$ (inductive hypothesis), we use backward induction on t and show that for all $s \in \mathcal{S}$, we have

$$\begin{aligned} v^\star_t(s) - u_t(s) &\overset{\text{(a)}}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,\pi^\star_t(s)) + \gamma \cdot v^\star_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] - \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,\hat{\pi}^\star_t(s)) + \gamma \cdot u_{t+1}(S'_{t+1,\hat{\pi}^\star_t(s)})\right] \\ &\overset{\text{(b)}}{\leq} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,\pi^\star_t(s)) + \gamma \cdot v^\star_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] - \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s,\pi^\star_t(s)) + \gamma \cdot u_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] \\ &\overset{\text{(c)}}{=} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\gamma \cdot v^{\pi^\star}_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] - \mathrm{ERM}_{\beta\cdot\gamma^t}\left[\gamma \cdot u_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] \\ &\overset{\text{(d)}}{=} \gamma \cdot \left(\mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[v^{\pi^\star}_{t+1}(S'_{t+1,\pi^\star(s)})\right] - \mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[u_{t+1}(S'_{t+1,\pi^\star(s)})\right]\right). \end{aligned} \tag{31}$$

(a) holds by the definition of $v^\star_t$ and u_t, **(b)** follows from $\hat{\pi}^\star$ being greedy to u, **(c)** is by subtracting the constant reward from both terms which can be done because ERM is translation invariant, and finally **(d)** follows from the positive quasi-homogeneity of ERM (see Theorem 3.1).

Now we can write the following sequence of inequalities:

$$\begin{aligned} v^{\pi^\star}_{t+1}(s) - u_{t+1}(s) &\overset{\text{(a)}}{\leq} \gamma^{T'-t-1} \cdot \frac{\beta \cdot \gamma^{T'} \cdot V^2}{8}, \qquad \forall s \in \mathcal{S} \\ \mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[v^{\pi^\star}_{t+1}(S)\right] - \mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[u_{t+1}(S)\right] &\overset{(b)}{\leq} \gamma^{T'-t-1} \cdot \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8} \\ \gamma \cdot \left(\mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[v^{\pi^\star}_{t+1}(S)\right] - \mathrm{ERM}_{\beta\cdot\gamma^{t+1}}\left[u_{t+1}(S)\right]\right) &\leq \gamma^{T'-t} \cdot \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8}. \end{aligned} \tag{32}$$

(a) follows from the inductive hypothesis and **(b)** comes from the monotonicity and translation invariance of ERM.

We can conclude the induction by combining (31) and (32).

Now that we proved (29), we can set $t = 0$ in it and use the fact that for all $t \in 0{:}T'$, we have $u_t(s) \leq v^{\hat{\pi}^\star}_t(s)$, to write

$$\begin{aligned} v^\star_0(s_0) - u_0(s_0) \leq \gamma^{T'} \cdot \frac{\beta \cdot \gamma^{T'} \cdot \Delta_{\mathfrak{R}}^2}{8} \implies \\ v^{\pi^\star}_0(s_0) - v^{\hat{\pi}^\star}_0(s_0) \leq \frac{\beta \cdot \gamma^{2T'} \cdot \Delta_{\mathfrak{R}}^2}{8}, \end{aligned}$$

which concludes the proof. □

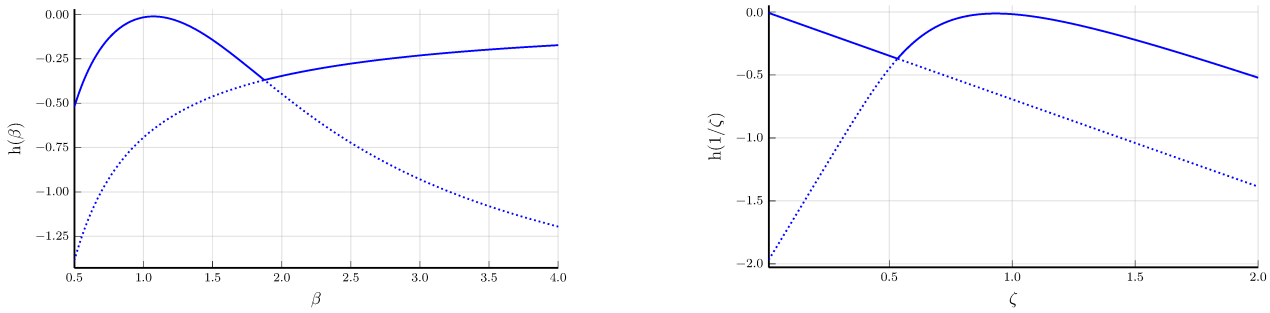

Figure 3: Plots of $h(\beta)$ (left) and $\zeta \mapsto h(\zeta^{-1})$ (right) for $\alpha = 0.5$, which are used in the proof of Proposition D.1.

D PROOFS OF SECTION 4

Before reporting the proofs for the theorems in Section 4, we state some results that highlight certain properties of EVaR. Recall that the function $h\colon \mathbb{R} \to \mathbb{R}$ is defined in (18) as

$$h(\beta) \;=\; \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_\beta\left[\mathfrak{R}_T^\pi\right] + \beta^{-1} \cdot \log(1-\alpha)\right) .$$

Because solving EVaR-MDP reduces to computing $\max_{\beta \geq 0} h(\beta)$, it would be ideal if it were concave or at least quasi-concave. Without the maximum in the definition of h (e.g., when $|\Pi_{MR}| = 1$), it is easy to see that the function $\zeta \mapsto h(\zeta^{-1})$ is concave (Ahmadi-Javid, 2012), and thus, h is quasi-concave. Unfortunately, the following proposition shows that h is not necessarily quasi-concave when $|\Pi_{MR}| > 1$ (with the maximum), which precludes the use of more efficient optimization techniques for solving $\max_{\beta \geq 0} h(\beta)$.

Proposition D.1. *There exists an MDP and $\alpha \in [0,1)$ such that the function $h\colon \mathbb{R} \to \mathbb{R}$ defined in* (18) *is neither concave nor convex either in β or β^{-1}.*

Proof. We show the property by constructing a counter-example for which the function h is not concave. Consider an EVaR-MDP with states $\mathcal{S} = \{s_0, s_1, s_2, s_3\}$, actions $\mathcal{A} = \{a_1, a_2\}$, a finite-horizon objective with $T = 2$ and $\gamma = 1$, and the following parameters:

$$p(\cdot \mid s_0, a_1) = [0, 0, 1, 0], \qquad p(\cdot \mid s_0, a_2) = [0, 0.02, 0, 0.98], \qquad r(\cdot, a_i) = [0, -2, 0, 1],\ \forall i \in \{1, 2\} .$$

The transition probabilities from s_1, s_2, s_3 are irrelevant because the horizon is $T = 2$. Since the reward is independent of the action and only depends on the state, the returns of the policies depend only on the action they take at state s_0. Thus, we only have two returns for all the policies in this MDP. Setting the confidence parameter to $\alpha = 0.5$, the plots in Appendix D show that neither $h(\beta)$ *(left)* nor $\zeta \mapsto h(\zeta^{-1})$ *(right)* is concave. In each plot, the functions for the two individual returns are indicated by dotted lines and are concave, but their maximum, shown by the solid line, is not. □

The following technical lemma will be useful when analyzing the optimal EVaR solution.

Lemma D.2. *Assume a fixed $\alpha \in (0,1)$ and a random variable $X \in \mathbb{X}$. Then, either the supremum in* (22) *is attained at some $\beta^\star > 0$ or*

$$\mathrm{EVaR}_\alpha\left[X\right] = \lim_{\beta \to \infty} \left(\mathrm{ERM}_\beta\left[X\right] + \beta^{-1} \cdot \log(1-\alpha)\right) .$$

Proof. The lemma follows directly from Proposition 2.11 in (Ahmadi-Javid and Pichler, 2017). □

Proof of Theorem 4.1. Recall that we assume that the supremum in (18) is attained for some $\beta^\star$. Using the existence of an optimal $\beta^\star$ we get by algebraic manipulation that

$$\begin{aligned}
\pi^\star &\in \arg\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha\left[\mathfrak{R}_T^\pi\right] \\
&= \arg\max_{\pi \in \Pi_{MR}} \max_{\beta > 0} \left(\mathrm{ERM}_\beta\left[\mathfrak{R}_T^\pi\right] + \beta^{-1} \cdot \log(1-\alpha)\right) \\
&\supseteq \arg\max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_{\beta^\star}\left[\mathfrak{R}_T^\pi\right] + (\beta^\star)^{-1} \cdot \log(1-\alpha)\right) \\
&= \arg\max_{\pi \in \Pi_{MR}} \mathrm{ERM}_{\beta^\star}\left[\mathfrak{R}_T^\pi\right] .
\end{aligned} \tag{33}$$

□

Proof of Corollary 4.2. The result follows directly from Theorem 4.1. □

Before stating the proof of Theorem 4.3, we report some results that we use there. The following lemma shows how to decompose the approximation error of Algorithm 2.

Lemma D.3. *Let $\pi^\star$ be the optimal solution to* (17) *and $\hat{\pi}^\star$ be the policy returned by Algorithm 2 when it is executed with a grid $\beta_1 < \ldots < \beta_K$ and calls to Algorithm 1 with horizon T'. Then, for any $\alpha \in (0,1)$, the approximation error of Algorithm 2 can be bounded as*

$$\mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] \leq \tag{34}$$
$$\max\left\{\sup_{\beta\in(0,\beta_1)} h(\beta) - h(\beta_1),\ \max_{k\in 1:K-1}\left(\sup_{\beta\in[\beta_k,\beta_{k+1})} h(\beta) - h(\beta_k)\right),\ \sup_{\beta\in[\beta_K,\infty)} h(\beta) - h(\beta_K)\right\} + \frac{\beta_K\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8}\,,$$

where the function h is defined by (18)*. Moreover, the bound for the finite-horizon objective is the same except the last term that depends on T' is zero.*

Proof. First, recall that Algorithm 2 calls Algorithm 1 for each value $\beta_1,\ldots,\beta_K$ in the grid and Algorithm 1 returns an approximately optimal policy $\hat{\pi}^\star$ for the ERM-MDP with the corresponding risk level β_k and its corresponding value function $\hat{v}^k$. In the following derivation, we use $\hat{\pi}_t^k$ and $\hat{v}_t^k$ for $t = 0{:}T, k = 1{:}K$ for the policy and value function, respectively, computed for β_k, by Algorithm 1. That is, the value function $\hat{v}_t^k$ uses ERM in the time steps $0{:}T'-1$ and the standard risk-neutral value function $v_\infty^\star$ thereafter. In contrast, the value $v^{\hat{\pi}^k}$ refers to the true ERM value function of the policy $\hat{\pi}^k$.

Using arguments analogous to (28), one can show for each $k = 1{:}K$ that

$$v_0^{\hat{\pi}^k}(s) \leq \hat{v}_0^k(s), \tag{35}$$

for each $s\in\mathcal{S}$. Similarly, using arguments analogous to (30), one can show that

$$\hat{v}_0^k(s) \leq v_0^{\hat{\pi}^k}(s) + \frac{\beta_k\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8}, \tag{36}$$

for each $s\in\mathcal{S}$.

Given the definition of $\hat{v}$, we can also define the EVaR objective function $h\colon\mathbb{R}\to\mathbb{R}$ and its approximation in the discrete points $(\beta_k)_{k=1}^K$ as

$$\begin{aligned} h(\beta) &= \max_{\pi\in\Pi_{MR}}\left(\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^\pi\right] + \beta^{-1}\cdot\log(1-\alpha)\right) \\ &= \max_{\pi\in\Pi_{MR}}\left(v_0^\pi(s_0) + \beta^{-1}\cdot\log(1-\alpha)\right) \\ \tilde{h}_k &= \hat{v}_0^k(s_0) + \beta_k^{-1}\cdot\log(1-\alpha)\,. \end{aligned}$$

The bound in (35) then implies for $k = 1{:}K$ that

$$h(\beta_k) \leq \tilde{h}_k. \tag{37}$$

Note that $\hat{\pi}^\star$ refers to the EVAR-MDP policy computed by Algorithm 2 and, therefore, $\hat{\pi}^\star = \hat{\pi}^{k^\star}$ for the optimal $k^\star$. Then,

assuming that $k^\star \in \arg\max_{k=1:K} \tilde{h}_k$ in Algorithm 2, we get that $\mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] = \delta$ for

$$
\begin{aligned}
\delta &= \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \sup_{\beta>0}\left(\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] + \beta^{-1}\cdot\log(1-\alpha)\right) \\
&\leq \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - v_0^{\hat{\pi}^\star}(s_0) - \beta_{k^\star}^{-1}\cdot\log(1-\alpha) && \text{Substitute feasible } \beta_{k^\star} \\
&= \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - v_0^{\pi^{k^\star}}(s_0) - \beta_{k^\star}^{-1}\cdot\log(1-\alpha) && \text{Choice of } \hat{\pi}^\star \\
&\leq \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \hat{v}_0^{k^\star}(s_0) - \beta_{k^\star}^{-1}\cdot\log(1-\alpha) + \frac{\beta_{k^\star}\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8} && \text{From (36)} \\
&\leq \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \hat{v}_0^{k^\star}(s_0) - \beta_{k^\star}^{-1}\cdot\log(1-\alpha) + \frac{\beta_K\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8} && \text{Because } \beta_K \geq \beta_{k^\star} \\
&= \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \max_{k=1:K}\tilde{h}_k + \frac{\beta_K\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8} && \text{From the optimality of } k^\star \\
&\leq \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \max_{k=1:K} h(\beta_k) + \frac{\beta_K\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8} && \text{From (37)} \\
&\leq \sup_{\beta>0} h(\beta) - \max_{k=1:K} h(\beta_k) + \frac{\beta_K\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8}. && \text{From the definition of } \pi^\star
\end{aligned}
$$

The lemma then follows by decomposing the supremum above as

$$
\sup_{\beta>0} h(\beta) = \max\left\{\sup_{\beta\in(0,\beta_1)} h(\beta), \sup_{\beta\in[\beta_k,\beta_{k+1})} h(\beta), \sup_{\beta\in[\beta_k,\infty)} h(\beta)\right\}.
$$

□

The following three lemmas now bound each one of terms in the maximum in (34).

Lemma D.4. *The function* $h\colon \mathbb{R}\to\mathbb{R}$ *defined in* (18) *satisfies that*

$$
\sup_{\beta\in(0,\beta_1)} h(\beta) - h(\beta_1) \;\leq\; \frac{\beta_1\cdot\Delta_{\mathfrak{R}}^2}{8}.
$$

Therefore, for any $\delta>0$, $\sup_{\beta\in(0,\beta_1)} h(\beta) - h(\beta_1) \leq \delta$ *when*

$$
\beta_1 \;\leq\; \frac{8\delta}{\Delta_{\mathfrak{R}}^2}.
$$

Proof. Because the function $\beta\mapsto \mathrm{ERM}_\beta[X]$ is non-increasing as shown in Lemma A.7 and $\beta^{-1}\cdot\log(1-\alpha)$ is increasing for $\beta>0$, we derive the bound as

$$
\begin{aligned}
\sup_{\beta\in(0,\beta_1)} h(\beta) - h(\beta_1) &= \sup_{\beta\in(0,\beta_1)} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^\pi\right] + \beta^{-1}\cdot\log(1-\alpha)\right) - \\
&\qquad \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_{\beta_1}\left[\mathfrak{R}_\infty^\pi\right] + \beta_1^{-1}\cdot\log(1-\alpha)\right) \\
&\leq \sup_{\beta\in(0,\beta_1)} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_0\left[\mathfrak{R}_\infty^\pi\right] + \beta^{-1}\cdot\log(1-\alpha)\right) - \\
&\qquad \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_{\beta_1}\left[\mathfrak{R}_\infty^\pi\right] + \beta_1^{-1}\cdot\log(1-\alpha)\right) \\
&\leq \max_{\pi\in\Pi_{MR}} \mathrm{ERM}_0\left[\mathfrak{R}_\infty^\pi\right] - \max_{\pi\in\Pi_{MR}} \mathrm{ERM}_{\beta_1}\left[\mathfrak{R}_\infty^\pi\right] \\
&\leq \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_0\left[\mathfrak{R}_\infty^\pi\right] - \mathrm{ERM}_{\beta_1}\left[\mathfrak{R}_\infty^\pi\right]\right).
\end{aligned}
$$

The lemma then follows readily by algebraic manipulation from Lemma A.8 because $\mathrm{ERM}_0\left[\mathfrak{R}_\infty^\pi\right] = \mathbb{E}[\mathfrak{R}_\infty^\pi]$. □

Lemma D.5. *The function $h\colon \mathbb{R} \to \mathbb{R}$ defined in* (18) *satisfies that*

$$\sup_{\beta\in[\beta_k,\beta_{k+1})} h(\beta) - h(\beta_k) \;\le\; (\beta_{k+1}^{-1} - \beta_k^{-1}) \cdot \log(1-\alpha)$$

for each $k \in 1{:}K$. Therefore, for any $\delta > 0$, $\sup_{\beta\in[\beta_k,\beta_{k+1})} h(\beta) - h(\beta_k) \le \delta$ when

$$\beta_{k+1} \;\le\; \frac{\beta_k \cdot \log(1-\alpha)}{\beta_k\delta + \log(1-\alpha)} \tag{38}$$

and $\beta_k\delta + \log(1-\alpha) < 0$. If $\beta_k\delta + \log(1-\alpha) \ge 0$, then β_{k+1} is not bounded from above.

Proof. From the definition of h and because the function $\beta \mapsto \mathrm{ERM}_\beta[X]$ is non-increasing as shown in Lemma A.7 and $\beta^{-1}\cdot\log(1-\alpha)$ is increasing for $\beta > 0$, we have

$$\begin{aligned}
\sup_{\beta\in[\beta_k,\beta_{k+1})} h(\beta) - h(\beta_k) &= \sup_{\beta\in[\beta_k,\beta_{k+1})} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_\infty^\pi] + \beta^{-1}\cdot\log(1-\alpha)\right) - \\
&\qquad \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_{\beta_k}[\mathfrak{R}_\infty^\pi] + \beta_k^{-1}\cdot\log(1-\alpha)\right) \\
&\le \sup_{\beta\in[\beta_k,\beta_{k+1})} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_\infty^\pi] + \beta_{k+1}^{-1}\cdot\log(1-\alpha)\right) - \\
&\qquad \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_{\beta_k}[\mathfrak{R}_\infty^\pi] + \beta_k^{-1}\cdot\log(1-\alpha)\right) \\
&\le \sup_{\beta\in[\beta_k,\beta_{k+1})} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_\infty^\pi] - \mathrm{ERM}_{\beta_k}[\mathfrak{R}_\infty^\pi]\right) + \\
&\qquad \left(\beta_{k+1}^{-1}\cdot\log(1-\alpha) - \beta_k^{-1}\cdot\log(1-\alpha)\right) \\
&\le \beta_{k+1}^{-1}\cdot\log(1-\alpha) - \beta_k^{-1}\cdot\log(1-\alpha).
\end{aligned}$$

The lemma then follows readily by algebraic manipulation. □

It is important to note that the multiplicative steps in Lemma D.5 increase with an increasing k. In particular, when $\delta\beta_k = -\log(1-\alpha)$, the constraint on β_{k+1} becomes vacuous with $\beta_{k+1} \le \infty$. At this point, we know that β_k is the last grid point that needs to be evaluated in order to guarantee an error of δ.

Lemma D.6. *The function $h\colon \mathbb{R} \to \mathbb{R}$ defined in* (18) *satisfies that*

$$\sup_{\beta\in[\beta_K,\infty)} h(\beta) - h(\beta_K) \le \frac{-\log(1-\alpha)}{\beta_K}.$$

Therefore, for any $\delta > 0$, $\sup_{\beta\in[\beta_K,\infty)} h(\beta) - h(\beta_K) \le \delta$ when

$$\beta_K \;\ge\; \frac{-\log(1-\alpha)}{\delta}.$$

Proof. From the definition of h and because $\beta \mapsto \mathrm{ERM}_\beta[X]$ is non-increasing and $\beta^{-1}\cdot\log(1-\alpha)$ is increasing for $\beta > 0$, we have

$$\begin{aligned}
\sup_{\beta\in[\beta_K,\infty)} h(\beta) - h(\beta_K) &\le \sup_{\beta\in[\beta_K,\infty)} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_\infty^\pi] + \beta^{-1}\cdot\log(1-\alpha)\right) - \\
&\qquad \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_{\beta_K}[\mathfrak{R}_\infty^\pi] + \beta_K^{-1}\cdot\log(1-\alpha)\right) \\
&\overset{\text{(a)}}{\le} \sup_{\beta\in[\beta_K,\infty)} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_\infty^\pi]\right) - \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_{\beta_K}[\mathfrak{R}_\infty^\pi] + \beta_K^{-1}\cdot\log(1-\alpha)\right) \\
&\le \sup_{\beta\in[\beta_K,\infty)} \max_{\pi\in\Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_\infty^\pi] - \mathrm{ERM}_{\beta_K}[\mathfrak{R}_\infty^\pi]\right) - \beta_K^{-1}\cdot\log(1-\alpha) \\
&\le -\beta_K^{-1}\cdot\log(1-\alpha).
\end{aligned}$$

(a) follows because $\beta^{-1}\cdot\log(1-\alpha)$ is negative for all $\beta \in [\beta_K, \infty)$. The lemma then follows readily by algebraic manipulation. □

Equipped with the above lemmas, we are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. Suppose that Algorithm 2 is executed with the grid defined by (19) and (20), and with T' set as

$$T' = \frac{\log(8\delta) - \log(\beta_K \Delta_{\mathfrak{R}}^2)}{2\log\gamma}. \tag{39}$$

Then, Lemmas D.3 to D.6 show that

$$\mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] \leq 2\delta.$$

It remains to show that the Algorithm 2 runs in time that is polynomial in $1/\delta$.

First, note that Algorithm 1 runs in time that is $O(S^2AT')$, assuming that v^∞ is computed using value iteration for some fixed $\gamma < 1$. Then using the choice of T' in (39), we have that evaluating a single β_k takes $O(S^2A\log(1/\delta))$ time.

Second, we need to upper-bound the value K since Algorithm 2 examines each one of these values. To emphasize that K is a function of δ, we denote it as K_δ in the remainder of the proof. To upper-bound K_δ, we first construct a lower-bound on each $\beta_{k+1},\ \forall k \in 1{:}K_\delta{-}1$ using definition (19) and the fact that $\beta_1 \leq \beta_k$ as

$$\begin{aligned}
\beta_{k+1} &= \beta_k \cdot \frac{\log(1-\alpha)}{\beta_k \cdot \delta + \log(1-\alpha)} \\
&\geq \beta_k \cdot \frac{\log(1-\alpha)}{\beta_1 \cdot \delta + \log(1-\alpha)} \\
&\geq \beta_1 \cdot \left(\frac{\log(1-\alpha)}{\beta_1 \cdot \delta + \log(1-\alpha)}\right)^k.
\end{aligned} \tag{40}$$

Here, we assume that β_1 is sufficiently small such that $\beta_k \cdot \delta + \log(1-\alpha) < 0$. Otherwise, we can use $K_\delta = 1$ to achieve the desired approximation error δ.

Recall that K_δ is chosen such that (20) is satisfied:

$$\beta_{K_\delta} \geq \frac{-\log(1-\alpha)}{\delta}.$$

Substituting the lower-bound on β_{K_δ} from (40) we get that the sufficient condition for K_δ is that

$$\beta_1 \cdot \left(\frac{\log(1-\alpha)}{\beta_1 \cdot \delta + \log(1-\alpha)}\right)^{K_\delta - 1} \geq \frac{-\log(1-\alpha)}{\delta}. \tag{41}$$

Next, define a variable z as follows and substitute the value for β_1 from (19) to get

$$z = \frac{\beta_1 \cdot \delta}{-\log(1-\alpha)} = \frac{8\delta^2}{-\Delta_{\mathfrak{R}}^2 \cdot \log(1-\alpha)}. \tag{42}$$

From the assumption that $\beta_k \cdot \delta + \log(1-\alpha) < 0$ and the fact that $\alpha \in (0,1)$, we get that $-\log(1-\alpha) \in (0,\infty)$ and, thus, $z \in (0,1)$. Substituting the variable z into (41) yields that the sufficient condition for K_δ is that

$$\left(\frac{1}{1-z}\right)^{K_\delta - 1} \geq \frac{1}{z}.$$

Taking the log of both sides and algebraic manipulation realizing that $(1-z)^{-1} \in (0,\infty)$ gives that it is sufficient to choose K_δ such that

$$K_\delta = \frac{\log\frac{1}{z}}{\log\frac{1}{1-z}} + 1 = \frac{\log\frac{1}{z}}{\sum_{n=1}^{\infty}\frac{z^n}{n}} + 1 \leq \frac{1}{z}\cdot\log\frac{1}{z} + 1.$$

The derivation above follows by a MacLaurin expansion of the denominator which is valid because $z \in (0,1)$. The last inequality follows because $z > 0$. Then, substituting the expression for z from (42), we get that

$$K_\delta \in O\left(\frac{\log(1/\delta)}{\delta^2}\right).$$

The complexity statement in the theorem then follows from the fact that running Algorithm 1 for each K_δ takes $O(S^2A\log(1/\delta))$ time. □

Table 5: $\mathrm{VaR}_{0.9}\left[\mathfrak{R}_T^\pi\right]$ for π returned by each method.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	-2.82	10.80	87.80	202	500
Naive grid	-2.90	10.80	52.60	202	501
Naive level	-10.00	11.40	83.30	201	217
Risk neutral	-2.90	12.60	67.50	202	499
Nested CVaR	-10.00	0.00	0.00	138	217
Nested EVaR	-10.00	10.30	0.00	173	217
ERM	-3.00	9.75	62.40	187	217
Nested ERM	-10.00	10.30	32.20	157	217
Augmented CVaR	-3.18	12.56	55.80	82	110

Table 6: $\mathbb{E}\left[\mathfrak{R}_T^\pi\right]$ for π returned by each method.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	-1.01	14.30	114.00	218	873
Naive grid	-1.01	14.30	63.20	219	873
Naive level	-10.00	15.80	107.00	217	217
Risk neutral	-0.98	17.10	128.00	219	871
Nested CVaR	-10.00	0.00	0.00	142	217
Nested EVaR	-10.00	14.60	0.00	182	217
ERM	-0.99	14.20	76.40	197	217
Nested ERM	-10.00	14.60	39.70	163	217
Augmented CVaR	-2.36	14.55	69.68	135	101

E NUMERICAL RESULTS: DETAILS

E.1 Domain Details

For each domain, we also provide CSV files in supplementary material with the exact specifications of the domains we use. The states in our tabular domains are identified with integer values $0, \ldots, S-1$, and actions for each state s are identified also with integer values $0, \ldots, A_s - 1$. Note that the action counts may be state-dependent. In our experiments, we assume that the reward $r\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ depends both on the originating and the destination state. Each CSV file has the following columns: "idstatefrom", "idaction", "idstateto", "probability", and "reward". Each row entry specifies a transition from "idstatefrom" after taking an ction "idaction" to state "idstateto" with the associated probability and reward. It is important to note that each combination ("idstatefrom", "idaction", "idstateto") is not necessarily unique; repeated combinations indicate different transitions to the same state. These transitions need to be properly accounted for when computing the risk of $r(S, A, S')$ since the associated rewards may be different.

E.1.1 Machine Replacement

This is the domain with the nominal transition probabilities described in (Delage and Mannor, 2010). We use the same discount factor $\gamma = 0.9$ and the time horizon $T = 100$. The initial state s_0 is that the machine is in the repair state R_1 indexed as "idstate" = 1. The exact definition of the problem is given in `machine.csv`.

E.1.2 Gambler's Ruin

This domain is based on a problem given in (Bäuerle and Ott, 2011). In this problem, a gambler starts with an initial capital c_0 can invest some of it in each time period. This investment doubles with a probability p and is lost with a probability $1 - p$. The reward is zero until a target wealth level c_f is achieved. The reward in the absorbing state c_f is 1.0. The state in this problem is the current, and the action is the investment. We use the initial capital $c_0 = 7$, the target capital $c_f = 10$, the probability of win $p = 0.7$, and the discount factor $\gamma = 0.95$. For this domain, we use a longer horizon $T = 200$. The

Table 7: Parameters of the inventory management problems.

Parameter	INV1	INV2
γ	0.9	0.9
$S_{\max}$	100	40
$A_{\max}$	50	10
Distribution D	Categorical	Poisson, $\lambda = 30$
p	16	4.99
c^h	0.3	0.05
c^f	2	0.49
c^v	5	2.49

precise definition of the problem is given in `ruin.csv`.

E.1.3 Inventory Management

This is a classical single-product stochastic inventory control problem (Puterman, 2005). The states $\mathcal{S} = \{0, \dots, S_{\max}\}$ represent the current stock of the product. The actions $\mathcal{A}_s = \{0, \dots, \min\{A_{\max}, (S_{\max} - s)\}\}$ for each state $s \in \mathcal{S}$ represent the amount of product ordered. The integer-valued random variable D represents the random demand. At any time step t, the next state S_{t+1} is a random variable computed as

$$S_{t+1} = [s_t + a_t - D]_+ .$$

The amount of product sold computed in a time step is

$$l_t = s_t - S_{t+1} + a_t.$$

The revenue u and expenses e are computed as

$$u_t = l_t \cdot p$$
$$x_t = \begin{cases} c^h \cdot s_t \cdot & \text{if } a_t = 0 \\ c^h \cdot s_t + c^f + c^v \cdot a_t & \text{otherwise} \end{cases}.$$

Here, p is the purchase price and c^h, c^f, and c^v are holding, fixed, and variable costs, respectively. The reward is then $r_t = u_t - x_t$.

The specific parameters that we use for the two inventory problems are summarized in Table 7. The time horizon for both problems is $T = 100$. The exact specifications of the two inventory domains are given in `inventory1.csv` and `inventory2.csv`.

E.1.4 Riverswim

This is an adapted version of the riverswim problem described in (Strehl and Littman, 2008). The discount factor in this problem is $\gamma = 0.9$ and the horizon is $T = 100$.

E.2 Algorithms

E.2.1 Algorithm 2

We implemented the algorithm in Julia, closely following the pseudo-code in Algorithm 2. The grid of values β_k are selected according to (19) with the parameters δ and $\Delta_{\mathfrak{R}}$ given in Table 8. We chose the value $\Delta_{\mathfrak{R}}$ based on the reward function structure of the problem and chose the tolerance value δ accordingly to be about 10% of the value function span. For small problems, like MR, we reduced δ even further. Anecdotally, δ has a smaller impact on the solution's quality than $\Delta_{\mathfrak{R}}$, but can significantly increase the computation time. The ERM-MDP sub-problem is solved exactly, which is possible because the horizon is finite.

Table 8: Parameters of Algorithm 2 for each benchmark problem.

Domain	Tolerance δ	Scale $(1-\gamma)\cdot\Delta_{\mathfrak{R}}$
MR	2	20
GR	0.5	1
INV1	5	5
INV2	1	1
RS	1	1

E.2.2 Naive Grid

Follows the same approach as Algorithm 2, but uses $\beta_k, k = 1{:}K$ computed as

$$\beta_1 = 10^5, \quad \beta_k = \frac{10 - \beta_1}{K - 1}.$$

The value K is chosen to be the same for each domain as the optimized K in (19).

E.2.3 Naive Level

Follows the same approach as Algorithm 2, but computes the value v^k for β^k by solving the following dynamic program for each $s \in \mathcal{S}$ and $t = 0{:}T{-}1$ as

$$v_t^k(s) = \max_{a\in\mathcal{A}} \mathrm{ERM}_{\beta_k}\left[r_{sa} + \gamma \cdot v_{t+1}^k(S'_{sa})\right],$$

where S'_{sa} is the random variable that represents the state that follows after taking an action a in state s.

E.2.4 Nested EVaR, CVaR, ERM

For any risk measure $\psi\colon \mathbb{X} \to \mathbb{R}$, like CVaR and EVaR, solve computes the value $v^\star$ by solving the following dynamic program for each $s \in \mathcal{S}$ and $t = 0{:}T{-}1$ as

$$v_t^\star(s) = \max_{a\in\mathcal{A}} \psi\left[r_{sa} + \gamma \cdot v_{t+1}^\star(S'_{sa})\right],$$

where S'_{sa} is the random variable that represents the state that follows after taking an action a in state s. Then, we evaluate a greedy policy $\pi_t^\star\colon \mathcal{S} \to \mathcal{A}, t = 0{:}T{-}1$ constructed to satisfy

$$\pi_t^\star(s) \in \max_{a\in\mathcal{A}} \psi\left[r_{sa} + \gamma \cdot v_{t+1}(S'_{sa})\right].$$

For EVaR and CVaR, we use $\alpha = 0.9$ and for ERM, we use $\beta = 0.5$.

E.2.5 ERM

We solve the optimal ERM value function and policy as described in Section 3 using $\beta = 0.5$

E.2.6 Augmented CVaR

We implemented the tabular version of the algorithm described in (Chow et al., 2015). We chose the discretization as recommended in (Chow et al., 2015) with the maximum number of points so that the computation finished in at most 10 minutes (about 20 times longer than the computation of other methods). One of the sources of complexity in this algorithm is that one needs to solve a linear program for every evaluation of the Bellman operator.

E.3 Results

Tables 9 and 10 show additional results for the algorithms and domains that we have compared. The results are broadly consistent with the results for other risk measures, and we include them for the sake of completeness.

Table 9: $\mathrm{VaR}_{0.9}\left[\mathfrak{R}_T^\pi\right]$ for π returned by each method.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	-2.82	10.80	87.80	202	500
Naive grid	-2.90	10.80	52.60	202	501
Naive level	-10.00	11.40	83.30	201	217
Risk neutral	-2.90	12.60	67.50	202	499
Nested CVaR	-10.00	0.00	0.00	138	217
Nested EVaR	-10.00	10.30	0.00	173	217
ERM	-3.00	9.75	62.40	187	217
Nested ERM	-10.00	10.30	32.20	157	217
Augmented CVaR	-3.18	12.56	55.80	82	110

Table 10: $\mathbb{E}\left[\mathfrak{R}_T^\pi\right]$ for π returned by each method.

Method	MR	GR	INV1	INV2	RS
Algorithm 2	-1.01	14.30	114.00	218	873
Naive grid	-1.01	14.30	63.20	219	873
Naive level	-10.00	15.80	107.00	217	217
Risk neutral	-0.98	17.10	128.00	219	871
Nested CVaR	-10.00	0.00	0.00	142	217
Nested EVaR	-10.00	14.60	0.00	182	217
ERM	-0.99	14.20	76.40	197	217
Nested ERM	-10.00	14.60	39.70	163	217
Augmented CVaR	-2.36	14.55	69.68	135	101