RASR: Risk-Averse Soft-Robust MDPs with EVaR and Entropic Risk

Anonymous Author(s)
Affiliation
Address
email

Abstract

Prior work on safe Reinforcement Learning (RL) has studied risk-aversion to randomness in dynamics (aleatory) and to model uncertainty (epistemic) in isolation. We propose and analyze a new framework to jointly model the risk associated with epistemic and aleatory uncertainties in finite-horizon and discounted infinite-horizon MDPs. We call this framework that combines Risk-Averse and Soft-Robust methods RASR. We show that when the risk-aversion is defined using either EVaR or the entropic risk, the optimal policy in RASR can be computed efficiently using a new dynamic program formulation with a time-dependent risk level. As a result, the optimal risk-averse policies are deterministic but time-dependent, even in the infinite-horizon discounted setting. We also show that particular RASR objectives reduce to risk-averse RL with mean posterior transition probabilities. Our empirical results show that our new algorithms consistently mitigate uncertainty as measured by EVaR and other standard risk measures.

1 Introduction

A major concern in high-stakes applications of reinforcement learning (RL), such as those in healthcare and finance, is to quantify the risk associated with the variability of returns. This variability is a form of *aleatory* uncertainty that arises from the inherent randomness in system dynamics. Since the risk of random returns cannot be captured by the standard *expected* objective, *convex risk measures* have emerged as perhaps 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 [33]. Conditional value-at-risk (CVaR), entropic value-at-risk (EVaR) [1, 31], and entropic risk measure (ERM) [33] are common examples of convex risk measures.

The goal in robust Markov decision process (MDP) is to mitigate performance loss due to uncertainty in modeling the system dynamics [37, 38, 40]. This uncertainty, often caused by limited or noisy data, is a form of *epistemic* uncertainty. *Soft-robust* formulations refine robust optimization by assuming a Bayesian distribution over plausible models (of the system dynamics) and then quantify the risk of model errors using convex risk measures [26, 43]. These formulations have close connections to distributional robustness [61]. While being risk-averse to epistemic uncertainty, existing soft-robust RL formulations are risk-neutral when it comes to the aleatory uncertainty that arises from the randomness in the system dynamics. This combination of risk-aversion to epistemic uncertainty with risk-neutrality to aleatory uncertainty can be problematic from the modeling perspective [19], and as we show below, may introduce unnecessary computational complexity.

The overarching objective of our work is to compute policies for MDPs that *jointly* mitigate the risk associated with epistemic (model) and aleatory (random dynamics) uncertainties. We call this objective RASR as it combines Risk Averse (aleatory) and Soft-Robust (epistemic) methods. This is

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

in contrast to the existing soft-robust MDP algorithms that are risk-neutral to the aleatory uncertainty. In this paper, we study RASR with two popular risk measures: ERM and EVaR.

As our first contribution, in Section 3, we introduce our RASR-ERM framework and propose new algorithms and analysis for it. ERM is unique among law-invariant risk measures in being dynamically consistent [42], which makes it compatible with dynamic programming (DP). Unfortunately, ERM is *not* positively homogeneous, which makes it incompatible with the use of discount factors. As a result, ERM has only been solved exactly in average-reward MDPs [11] and *undiscounted* stochastic programs [27]. Our main innovation is to use time-dependent risk-levels to precisely solve ERM in *discounted finite-horizon* MDPs and to employ new bounds to tightly approximate it in *discounted infinite-horizon* MDPs (Section 3.2). We build on the DP decomposition of the RASR-ERM objective to show that there exists an optimal value function and (surprisingly) a *deterministic* Markovian optimal policy for this problem. This is unusual because most other risk-averse formulations require randomized optimal policies. We also show that under an assumption of a dynamic model of epistemic uncertainty [26, 28], the RASR-ERM objective reduces to a risk-averse MDP with the mean posterior transition model (Section 3.1).

As our second contribution, we formulate and study the RASR-EVaR framework in Section 4. Although ERM is computationally convenient, it is often an impractical method to measure risk since the result is scale-dependent. EVaR is preferable to ERM because it is coherent, positive-definite, interpretable, and comparable with VaR and CVaR. However, EVaR is not dynamically consistent and cannot be directly optimized using a DP. Our main contribution here is to reduce the RASR-EVaR optimization to multiple RASR-ERM problems that each can be solved by DP. Our theoretical analysis shows that the RASR-EVaR properties mirror those for RASR-ERM and that the proposed algorithm can compute a solution arbitrarily close to the optimum. We empirically evaluate our RASR algorithms in Section 5 and show their benefits over prior robust, soft-robust, and risk-averse MDP algorithms. Finally, in Section 6, we position our RASR framework in the context of the literature on soft-robust and risk-averse MDPs.

2 Preliminaries

We assume the general problem can be formulated as an MDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, r, p, s_0, \gamma)$. The state and action sets $\mathcal{S}$ and $\mathcal{A}$ are finite with cardinality S and A. The reward function $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the reward received in each state after taking an action. We use $\triangle r = \max_{s\in\mathcal{S}, a\in\mathcal{A}} r(s,a) - \min_{s\in\mathcal{S}, a\in\mathcal{A}} r(s,a)$ to refer to the span semi-norm of the rewards. The transition probabilities (dynamics) are shown as $p\colon \mathcal{S} \times \mathcal{A} \to \Delta^S$, where Δ^S is the probability simplex in $\mathbb{R}^S$. The initial state is denoted by $s_0 \in \mathcal{S}$. Finally, $\gamma \in (0, 1]$ is the discount factor. We assume a fixed horizon $T \in \mathbb{N}^+ \cup \{\infty\}$ with $T = \infty$ indicating an infinite-horizon objective.

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 [54]. A *randomized Markovian* policy depends only on the time-step and current state 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 Markovian and stationary randomized policies, and by Π_{MD} and Π_{SD}, the corresponding sets of deterministic policies.

We define $\mathfrak{R}_T^\pi$, the random variable of the return of a policy π after T time steps as

$$\mathfrak{R}_T^\pi \;=\; \sum_{t=0}^{T} \gamma^t \cdot R_t^\pi = \sum_{t=0}^{T} \gamma^t \cdot r(S_t^\pi, A_t^\pi)\,, \tag{1}$$

where S_t^π, $A_t^\pi \sim \pi_t(\cdot|S_t^\pi)$ and R_t^π are the random variables of state visited, action taken, and reward received at time $t \in 0, \ldots, T$, when following policy π. The objective in the standard risk-neutral MDP is to maximize the *expectation* of the return random variable,

$$\max_\pi \; \mathbb{E}\big[\mathfrak{R}_T^\pi\big]\;. \tag{2}$$

In finite-horizon MDPs, $T < \infty$ and (usually) $\gamma = 1$. In infinite-horizon discounted MDPs, we use $T = \infty$ (as a shorthand for $T \to \infty$) and restrict the discount factor to $\gamma \in (0, 1)$. It is known that the finite and infinite horizon discounted settings have optimal policies in Π_{MD} and Π_{SD}, respectively.

Risk-averse MDP A risk measure $\psi\colon \mathbb{X} \to \mathbb{R}$ assigns a scalar risk value to a random variable $X \in \mathbb{X}$, where $\mathbb{X}$ denotes the set of real-valued random variables. Convex risk measures are an axiomatic generalization of the expectation operator $\mathbb{E}[\cdot]$ that capture a wide range of risk-aversion preferences [30, 32, 35]. We describe *coherent* and *convex* risk measures, and summarize their properties that are desirable in studying risk-averse MDPs in Appendix E. The objective in risk-averse MDP is defined by replacing the expectation in (2) with an appropriate risk measure

$$\max_{\pi} \psi\big[\mathfrak{R}_T^{\pi}\big]. \tag{3}$$

Soft-robust MDP The soft-robust setting makes the Bayesian assumption that the transition model P is a random variable with a distribution that can be computed, for instance, using Bayesian inference [26, 28, 43]. In this paper, we assume a *dynamic* model of uncertainty [26, 28]. In the dynamic model, the transition probability is not only unknown, but can also change during the execution. This is in contrast to the *static* model [22, 43], in which it is uncertain but does not change throughout an episode. We target dynamic uncertainty because it is easier to optimize and our results lay down the foundations necessary to tackle static models in future. In the dynamic model, the transition probability is defined as $P = (P_t)_{t=0}^{T-1}$, where each model P_t is a random variable distributed as $P_t \sim f_t$, and f_t's are derived from Bayesian inference methods.

Prior work on soft-robust RL (e.g., [26, 28, 43]) has focused on the following objective:

$$\max_{\pi \in \Pi_{SR}} \psi\Big[\mathbb{E}\big[\mathfrak{R}_T^{\pi} \mid P\big]\Big]. \tag{4}$$

In (4), the risk measure ψ is applied only to the epistemic uncertainty over P, and the optimization is risk-neutral (uses $\mathbb{E}[\cdot]$) to the randomness in $\mathfrak{R}_T^{\pi} \mid P$ (aleatory uncertainty). In some instances, the optimization in (4) reduces to a form of distributionally-robust MDP [37, 43, 61].

RASR Our RASR formulation, introduced formally below, takes into account both the epistemic uncertainty in the transition model P and the aleatory uncertainty in $\mathfrak{R}_T^{\pi} \mid P$, and optimizes the objective

$$\max_{\pi \in \Pi_{SR}} \psi\Big[\psi\big[\mathfrak{R}_T^{\pi} \mid P\big]\Big]. \tag{5}$$

Risk Measures We study two convex risk measures in our RASR formulation: *entropic risk measure* (ERM) and *entropic value-at-risk* (EVaR). ERM with a risk-aversion parameter $\alpha \in \mathbb{R}_+ \cup \{\infty\}$, for a random variable $X \in \mathbb{X}$, is defined as [33]

$$\mathrm{ERM}^{\alpha}[X] = -\alpha^{-1} \cdot \log\Big(\mathbb{E}\left[e^{-\alpha \cdot X}\right]\Big). \tag{6}$$

For the risk level $\alpha = 0$, ERM of a random variable equals to its expectation, $\mathrm{ERM}^0[X] = \lim_{\alpha \to 0^+} \mathrm{ERM}^{\alpha}[X] = \mathbb{E}[X]$. Similarly, $\mathrm{ERM}^{\infty}[X] = \operatorname{ess\,inf}[X]$ is the minimum value of X.

ERM is the only law-invariant convex risk measure that is dynamically-consistent [42] (see Appendix E.5). This is an important property for a risk measure in multi-stage decision problems, because it allows defining a dynamic program (DP) for the risk measure and optimizing it. The following theorem is crucial for deriving our results. It has been proved in earlier work, but for completeness, we report its proof in Appendix A.

Theorem 2.1 (Tower Property). *Any two random variables $X_1, X_2 \in \mathbb{X}$ satisfy that*

$$\mathrm{ERM}^{\alpha}[X_1] \;=\; \mathrm{ERM}^{\alpha}\left[\mathrm{ERM}^{\alpha}[X_1 \mid X_2]\right].$$

Note that the tower property also holds for the expectation operator $\mathbb{E}[\cdot]$, but is violated by most common risk measures, including VaR, CVaR, and EVaR. Despite its many nice features, ERM also has several undesirable properties. It is not positively-homogeneous: $\mathrm{ERM}^{\alpha}[c \cdot X] \neq c \cdot \mathrm{ERM}^{\alpha}[X]$, for $c \geq 0$, which means that $\mathrm{ERM}^{\alpha}[X]$ does not scale linearly with X. Moreover, ERM is difficult to interpret and its risk-level α is not readily comparable to the risk-levels of VaR and CVaR.

EVaR was proposed to address some of the shortcomings of ERM. EVaR with confidence parameter $\beta \in [0, 1)$, for a random variable $X \in \mathbb{X}$, is defined as [1, 31]

$$\mathrm{EVaR}^{\beta}[X] \;=\; \sup_{\alpha > 0}\left(\mathrm{ERM}^{\alpha}[X] + \alpha^{-1} \cdot \log(1 - \beta)\right). \tag{7}$$

Although EVaR is not dynamically consistent, we show in Section 4 that it can be optimized using a DP by representing it in terms of ERM. Unlike ERM, EVaR is positively-homogeneous, and thus, coherent, which makes its riskiness independent of the scale of the random variable. Moreover, the meaning of its risk-level β is consistent with those used in VaR and CVaR, with $\mathrm{EVaR}^0[X] = \mathbb{E}[X]$ and $\lim_{\beta\to 1}\mathrm{EVaR}^\beta[X] = \operatorname{ess\,inf}[X]$. Finally $\mathrm{EVaR}^\beta[X] \le \mathrm{CVaR}^\beta[X] \le \mathrm{VaR}^\beta[X]$, EVaR can be interpreted as the tightest conservative approximation that can be obtained from the Chernoff inequality for VaR and CVaR [1] .

3 RASR-ERM Framework

In this section, we describe our RASR formulation with the entropic risk measure (ERM), which we refer to as RASR-ERM. In particular, we show that the RASR-ERM objective can be optimized using a novel DP formulation with time-dependent risk. We also establish fundamental properties for the optimal policies of this formulation. The proofs of all the results of this section are in Appendix B.

We adopt the soft-robust RL model with dynamic uncertainty. Thus, we assume that the transition model $P = (P_t)_{t=0}^{T-1}$ is a collection of random variables as described in Section 2. Following the RASR objective in (5), the RASR-ERM objective is to maximize the ERM of the total return with respect to *both* model uncertainty (epistemic) and random dynamics (aleatory), and is formally defined as

$$\max_{\pi\in\Pi_{MR}} \mathrm{ERM}^\alpha\left[\mathfrak{R}_T^\pi\right] \;=\; \mathrm{ERM}^\alpha\left[\mathrm{ERM}^\alpha\left[\mathfrak{R}_T^\pi \mid P\right]\right]. \tag{8}$$

The ERM on the LHS of (8) is applied simultaneously to epistemic and aleatory uncertainties and equals to the nested ERM formula on the RHS of Theorem 2.1. Compared with (3), the optimization in (8) involves risk-aversion to the model (epistemic) uncertainty. Compared to (4), the aleatory uncertainty in the return random variable, $\mathfrak{R}_T^\pi \mid P$, is modeled by the same risk measure (ERM in place of $\mathbb{E}[\cdot]$) as the one used to model the risk associated with the epistemic (model) uncertainty. We refer to an optimal solution to (8) as an *optimal policy* $\pi^\star = (\pi_t^\star)_{t=0}^{T-1}$. To simplify the exposition, we restrict our attention in (8) to Markov policies, because the DP formulation that we derive in Section 3.1 shows that history-dependent policies offer no advantage in RASR-ERM.

3.1 Dynamic Program Formulation for RASR-ERM

Before deriving DP equations for the value function in RASR-ERM, we show a simple, but critical, property of ERM. While ERM is known not to be positively homogeneous, the following new result shows that it has a similar property, if we allow for a change in the risk level.

Theorem 3.1 (Positive Quasi-homogeneity)**.** *Let $X \in \mathbb{X}$ be a random variable. Then, for any constant $c \ge 0$, we have*

$$\mathrm{ERM}^\alpha[c\cdot X] \;=\; c\cdot \mathrm{ERM}^{\alpha\cdot c}[X]\,.$$

With the two ERM properties stated in Theorems 2.1 and 3.1, we are now ready to propose the value function and DP (Bellman) equations for RASR-ERM. The value function for a policy π is the collection $v^\pi = (v_t^\pi)_{t=0}^T$, where $v_t^\pi : \mathcal{S} \to \mathbb{R}$ is the value at time-step t and is defined as

$$v_t^\pi(s) \;=\; \mathrm{ERM}^{\alpha\cdot\gamma^t}\left[\sum_{t'=t}^{T}\gamma^{t'-t}\cdot R_{t'}^\pi \mid S_t = s\right], \quad \forall s\in\mathcal{S}. \tag{9}$$

We define the *optimal value function* $v^* = (v_t^*)_{t=0}^T$ as the value function of an optimal policy $\pi^\star$, i.e., $v^\star = v^{\pi^\star}$, and let the terminal value function equal to $v_T^\pi(s) = 0$. It can be readily seen from (9) that the value function of any policy π at the initial state $v_0^\pi(s_0)$ is equal to the objective in (8), i.e., $v_0^\pi(s_0) = \mathrm{ERM}^\alpha\left[\mathfrak{R}_T^\pi\right]$.

The dependence of risk-level on time-step t in the value function definition (9) is quite important in deriving our DP formulation for RASR-ERM below. As time progresses, the risk level $\alpha\gamma^t$ decreases monotonically, and the value function in (9) becomes less risk-averse. Recall that in the risk-neutral setting, the risk-level is $\alpha = 0$ and $\mathrm{ERM}^0[X] = \mathbb{E}[X]$. Similarly, when we set $\alpha = 0$ in (9), the risk-level becomes 0 and is independent of t, and thus, can be replaced with an expectation. In this case, when there is no model (epistemic) uncertainty, the value function in (9) coincides with the one in standard risk-neutral MDPs.

The next result states the Bellman equations for RASR-ERM value functions.

Theorem 3.2 (Bellman Equations)**.** *For any policy $\pi \in \Pi_{MR}$, its value function $v^\pi = (v_t^\pi)_{t=0}^T$ defined in (9) is the unique solution to the following system of equations:*

$$v_t^\pi(s) \;=\; \mathrm{ERM}^{\alpha\cdot\gamma^t}\left[r(s,A) + \gamma \cdot v_{t+1}^\pi(S')\right], \quad \forall s \in \mathcal{S}, \tag{10}$$

where $A \sim \pi_t(\cdot|s)$ and $S' \sim \bar{P}_t(\cdot|s,A)$, $\bar{P}_t(s'|s,a) = \mathbb{E}[P_t(s'|s,a)]$, and $v_T^\pi(s) = 0$ for each $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, and $t = 0, \ldots, T-1$. Moreover, the optimal value function $v^\star = (v_t^\star)_{t=0}^T$ (defined previously) is the unique solution to

$$v_t^\star(s) \;=\; \max_{a\in\mathcal{A}} \mathrm{ERM}^{\alpha\cdot\gamma^t}\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right], \quad \forall s \in \mathcal{S},\; S' \sim \bar{P}_t(\cdot|s,a). \tag{11}$$

Note that the ERM operator in (10) and (11) applies to the random variables A and S'.

Theorem 3.2 suggests several new important and surprising properties for the RASR-ERM objective (8). The first property that follows from the DP equations in Theorem 3.2 is that the RASR-ERM objective (8) is equivalent to a risk-averse RL problem with the mean posterior transition model $\bar{P}$ defined in Theorem 3.2.

Corollary 3.3. *For any policy $\pi \in \Pi_{MR}$, we have that*

$$\mathrm{ERM}^\alpha\left[\mathrm{ERM}^\alpha[\mathfrak{R}_T^\pi \mid P]\right] = \mathrm{ERM}^\alpha\left[\mathfrak{R}_T^\pi \mid \bar{P}\right].$$

The second important result that follows from Theorem 3.2 is that there exists an optimal Markovian deterministic policy for the RASR-ERM objective (8), which is greedy to the optimal value function $v^\star$ defined by (11).

Theorem 3.4. *There exists a deterministic time-dependent optimal policy $\pi^\star = (\pi_t^\star)_{t=0}^{T-1} \in \Pi_{MD}$ for (8), which is greedy to the optimal value function $v^\star$ in (11), i.e., for any $t = 0, \ldots, T-1$,*

$$\pi_t^\star(s) \in \operatorname*{argmax}_{a\in\mathcal{A}} \mathrm{ERM}^{\alpha\cdot\gamma^t}\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right], \quad \forall s \in \mathcal{S},\; S' \sim \bar{P}_t(\cdot|s,a). \tag{12}$$

The fact that RASR-ERM may have a deterministic optimal policy is especially surprising because optimizing most risk-averse formulations often requires randomization [23]. Another surprising observation is that, unlike the risk-neutral formulation, RASR-ERM does not admit a stationary optimal policy in the infinite-horizon discounted setting. This is mainly due to the fact that the risk-level is time-dependent in the DP equations of RASR-ERM. Finally, note that the above results provide stronger guarantees than the DP equations for the existing soft-robust MDP formulations [28, 43]. By adjusting the risk-level with time, our DP formulation in Theorem 3.2 guarantees that the optimal value function solves the soft-robust objective (8). This is in contrast to the DP in other soft-robust formulations, whose optimal value function is not an exact solution to the corresponding soft-robust objective.

3.2 Algorithms for Optimizing RASR-ERM

We now turn to algorithms that can compute RASR-ERM value functions and policies. With the *finite-horizon* objective ($T < \infty$), the optimal value function can be computed by adapting the standard value iteration (VI) to this setting. This algorithm computes the optimal value function $v_t^\star$ backwards in time $t = T, T-1, \ldots, 0$ according to (11). The optimal policy is greedy with respect $v^\star$ and can be computed by solving the discrete optimization problem in (12). We include the full algorithms in the appendix in Appendix B.

Solving the *infinite-horizon* problem is considerably more challenging than the finite-horizon problem, because the risk level α and the optimal policy are in general time dependent. The simplest way to address this issue is to simply truncate the horizon to some $T' < \infty$ and resort to an arbitrary policy for any $t > T'$. The significant limitation to truncating the horizon 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 a truncated planning horizon. The algorithms works as follows. First, it computes the optimal stationary risk-neutral value function v^∞ and policy π^∞ using value iteration or policy iteration [54]. The policy π^∞ is used for all time steps

Algorithm 1: VI for infinite-horizon RASR-ERM

Input: Planning horizon $T' < \infty$, risk level $\alpha > 0$

Output: Optimal 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 optimal v^∞ and π^∞ as a solution to the infinite-horizon discounted MDP with $\bar{P}$;

Compute $(\tilde{v}_t^\star)_{t=0}^{T'}$ and $(\tilde{\pi}_t^\star)_{t=0}^{T'-1}$ using (11) and (12) with horizon T' and terminal value $\tilde{v}_{T'}^\star = v^\infty$;

Construct a policy $(\hat{\pi}_t^\star)_{t=0}^\infty$, where $\hat{\pi}_t^\star = \pi^\infty$ when $t \geq T'$ and $\hat{\pi}_t^\star = \tilde{\pi}_t^\star$, otherwise ;

Construct $\hat{v}^\star$ analogously to $\hat{\pi}^\star$;

return $\hat{\pi}^\star$, $\hat{v}^\star$

$t > T'$ and the value function v^∞ is used to approximate $v_{T'}^\star$. This approach takes an advantage of the fact that the risk level $\alpha \cdot \gamma^t$ in (11) approaches 0 as $t \to \infty$. This means that the ERM value function becomes ever closer to the optimal risk-neutral discounted value function v^∞.

To quantify the quality of the 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 simply truncating the planning horizon. [Note: we managed to tighten this bound when revising the appendix]

Theorem 3.5. *The performance loss of the policy $\hat{\pi}^\star$ returned by Algorithm 1 for a discount factor $\gamma < 1$ decreases with T' as*

$$\mathrm{ERM}^\alpha\left[\mathfrak{R}_\infty^{\pi^\star} \mid \bar{P}\right] - \mathrm{ERM}^\alpha\left[\mathfrak{R}_\infty^{\hat{\pi}^\star} \mid \bar{P}\right] \;\leq\; c \cdot \gamma^{2T'}$$

where $\pi^\star$ is optimal in (5) *and $c = 8^{-1}\alpha \cdot (\triangle r)^2 (1-\gamma)^{-2}$.*

The proof of Theorem 3.5 uses the Hoeffding's lemma to bound the error between ERM and the expectation and propagates the error using standard dynamic programming techniques.

Analysis analogous to Theorem 3.5 shows that when ones truncates the horizon at T' and follows an arbitrary policy thereafter, the performance loss decreases proportionally to $\gamma^{T'}$ as opposed to $\gamma^{2T'}$. As a result, truncating a policy requires more than double the planning horizon T' to achieve the same approximation guarantee as Algorithm 1.

In practice, one can compute bounds that are tighter than Theorem 3.5 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^∞ 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 performance loss. This bound does not have an analytical form, but our anecdotal experimental results shows that it converges to 0 with an increasing T' even more rapidly than Theorem 3.5. We give more details in Appendix B.

4 RASR-EVaR Framework

In this section, we introduce and analyze RASR with the EVaR objective, which we refer to as the RASR-EVaR framework. As mentioned in Section 2, EVaR is preferable to ERM because it is coherent, positive-definite, interpretable, and comparable with VaR and CVaR. The main challenge with RASR-EVaR is that EVaR is "not" dynamically consistent, and thus, cannot be directly optimized using a DP. Our main contribution here is to show that despite this issue, it is possible to solve RASR-EVaR by extending the algorithms developed for RASR-ERM in Section 3. The detailed proofs of all the results of this section are reported in Appendix C.

The RASR-EVaR formulation assumes the same soft-robust setting as in (8) with the following modified objective:

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}^\beta\left[\mathfrak{R}_T^\pi\right] . \tag{13}$$

The EVaR operator in (13) applies simultaneously to both epistemic and aleatory uncertainties over returns. Note that because EVaR does not satisfy the tower property, it is impossible to rewrite (13) using separate risk for the aleatory and epistemic uncertainty, similarly to (8). We use $\pi^\star$ throughout this section to denote an optimal policy in (13).

We propose to tractably approximate the optimal RASR-EVaR policies using the dual formulation of the EVaR [1]. Our reformulation makes it possible to reduce RASR-EVaR to a sequence of tractable RASR-ERM problems. In particular, define a function $h : \mathbb{R} \to \mathbb{R}$ as

$$h(\alpha) \;=\; \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}^{\alpha}[\mathfrak{R}_T^{\pi}] + \frac{\log(1-\beta)}{\alpha} \right). \tag{14}$$

This function represents the RASR-ERM return of the best policy for any level $\alpha > 0$ and can be computed using the methods developed in Section 3. The dual representation of EVaR immediately shows that $\max_{\alpha \geq 0} h(\alpha) = \max_{\pi \in \Pi_{MR}} \mathrm{EVaR}^{\beta}[\mathfrak{R}_T^{\pi}]$ for any $\beta \in (0, 1)$.

The reformulation of RASR-EVaR in terms of ERM can be used to establish the following results.

Theorem 4.1. *Let $\pi^\star$ be an optimal solution to RASR-EVaR (Eq. 13). Then, there exists a risk-level $\alpha^\star$, such that $\pi^\star$ is optimal in RASR-ERM (Eq. 8) with $\alpha = \alpha^\star$.*

Theorem 4.1 combined with the properties of RASR-ERM, shown in Section 3, can be used to establish the following properties for RASR-EVaR.

Corollary 4.2. *There exists an optimal policy $\pi^\star$ in (13) that is Markovian and deterministic ($\pi^\star \in \Pi_{MD}$). In addition, for any policy $\pi \in \Pi_{MR}$, the RASR-EVaR objective in (13) equals to*

$$\mathrm{EVaR}^{\beta}[\mathfrak{R}_T^{\pi}] = \mathrm{EVaR}^{\beta}\left[\mathfrak{R}_T^{\pi} \mid \bar{P}\right] \;,$$

where $\bar{P}$ is defined as in Theorem 3.2.

We are now ready to describe our algorithms for solving the RASR-EVaR objective given in Algorithm 2. The algorithm takes advantage of the fact that the optimization problem $\max_{\alpha \geq 0} h(\alpha)$ is single-dimensional. The algorithm searches a grid of candidate α values. Each $h(\alpha)$ is computed the RASR-ERM algorithms described in Section 3.

Algorithm 2: Algorithm for RASR-EVaR

Input: Discretized risk-levels $\alpha_0 \geq \cdots \geq \alpha_K > 0$
Output: RASR-EVaR optimized policy $\hat{\pi}^\star$
1 Compute policy π^k and value function v^k to solve RASR-ERM for risk-level α_k, for $k = 1, \ldots, K$;
2 Let $k^\star \leftarrow \mathrm{argmax}_{k=1,\ldots,K}\, v_0^k(s_0) + \alpha_k^{-1}\log(1-\beta)$;
3 **return** *Policy* $\hat{\pi}^\star = \pi^{k^\star}$

Algorithm 2 resorts to discretizing α values because $h(\alpha)$ may be non-concave and, therefore, cannot be maximized using more efficient algorithms. Although the EVaR objective in (7) is concave, it is the maximization over π in (14) that may make h non-concave. Our key contribution is that we use beneficial properties of h to show that the discrete grid of points can be constructed to compute a close-to-optimal solution without an excessive computational burden, as summarized in the theorem below.

Theorem 4.3. *Given an error tolerance $\delta > 0$, construct a discretization $(\alpha_k)_{k=0}^K$ such that $\alpha_0 = \infty$ and for $k > 0$*

$$\alpha_k = \frac{-\log(1-\beta)}{k \cdot \delta}, \qquad K \geq \sqrt{\frac{-\log(1-\beta)}{8}}\,\frac{\triangle r}{(1-\gamma)\cdot\delta}\,.$$

Then, the performance loss of the policy $\hat{\pi}^\star$ returned by Algorithm 2 with $(\alpha_i)_{i=0}^n$ compared with the optimal $\pi^\star$ is bounded as $\mathrm{EVaR}^{\beta}\left[\mathfrak{R}_\infty^{\pi^\star} \mid \bar{P}\right] - \mathrm{EVaR}^{\beta}\left[\mathfrak{R}_\infty^{\hat{\pi}^\star} \mid \bar{P}\right] \leq \delta$.

One can make Algorithm 2 more efficient by realizing that Algorithm 1 computes value functions for multiple risk levels $\alpha, \gamma\alpha, \gamma^2\alpha, \ldots$. For instance, running Algorithm 1 with $\alpha = 0.5$ computes v_0 with a risk $\alpha = 0.5$, v_1 with a risk $\alpha = 0.5\gamma$, v_2 with a risk level $\alpha = 0.5\gamma^2$ 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 risk levels α. Given that this is the first work proposing and optimizing RASR-EVaR, we focus on the conceptually simple Algorithm 2 and leave computational improvements for future work.

Method	RS	POP	INV
RASR	**50**	**-7020**	**294**
Naive	**50**	-8291	290
Erik	45	-8628	290
Derman	7	-7259	287
RSVF	45	-8874	257
BCR	34	-8731	281
SRVI	34	-8714	280
Chow	23	-7238	290

Table 1: Risk $\mathrm{EVaR}_{\pi}^{0.99}[\mathfrak{R}_{\infty}^{\pi}]$ for π computed by each method.

		Risk Measure	
Method	Object.	Epistemic	Aleatory
RASR	Disc.	EVaR	EVaR
Erik [28]	Disc.	ERM	E
Derman [26]	Aver.	E	E
RSVF [56]	Disc.	VaR	E
BCR [6]	Disc.	VaR	E
SRVI [43]	Disc.	CVaR	E
Chow [20]	Disc.	–	CVaR

Table 2: Summary of the soft-robust and risk-averse models in the MDP/RL literature.

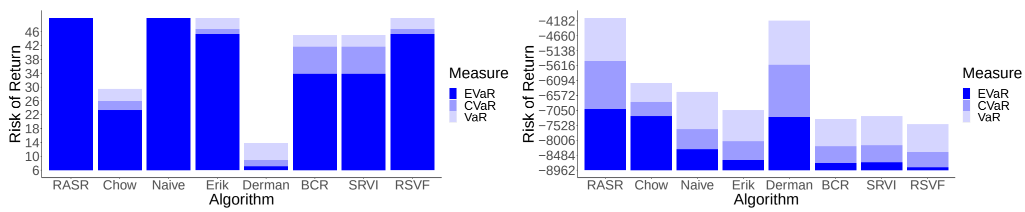

Figure 1: $\psi^{0.99}[\mathfrak{R}^{\pi}]$ of return in riverswim (left) and population (right) .

5 Empirical Evaluation

In this section, we evaluate our RASR framework empirically on several MDPs used previously to evaluate soft-robust and risk-averse algorithms. The empirical evaluation focuses on RASR-EVaR for two reasons. First, as discussed in Section 2, EVaR is a more practical risk measure than ERM because it is closely related to the popular VaR and CVaR. Second, any RASR-EVaR optimal policy is also a RASR-ERM policy for some α optimal in (14). We provide additional results, information, and details in Appendix F.

We now describe the experimental setup. As the primary metric for the comparison, we use $\mathrm{EVaR}^{0.99}[\mathfrak{R}_{\infty}^{\pi}]$ for a policy π computed by RASR-EVaR or another baseline algorithm. For the sake of completeness, we also compare the risk computed using VaR and CVaR, two common risk measures. The epistemic uncertainty in our experiments follows the dynamic model described in Section 2.

The following three domains from the robust RL literature are used to evaluate the algorithms: river-swim [6], population [56], inventory [6]. The *river-swim* (RS) problem tests whether algorithms are sufficiently risk-averse. It involves small epistemic uncertainty that, nevertheless, impacts the return significantly. In contrast, the *population* (POP) problem tests whether the algorithms are overly risk-averse. The epistemic uncertainty is large but makes a small difference in the overall return. Finally, the *inventory* (INV) domain combines the characteristics of the other two domains.

To understand how well RASR-EVaR performs, we compare the policy it computes with several related methods. Even though these baselines were designed to be risk-averse with respect to the epistemic risk, comparing RASR-EVaR with them helps us understand the importance of optimizing jointly for epistemic and aleatory uncertainties. The *naive* algorithm computes the ERM value function by solving a dynamic program akin to Theorem 3.2, but with risk α that is constant across time. Algorithms *Erik* [28], *Derman* [26], *BCR* [6], *RSVF* [56], *SRVI* [43] originate in robust RL literature and their objectives are summarized in Section 6. BCR and RSVF represent two recent algorithms proposed to optimize the percentile objective which maps to VaR. SRVI optimizes a CVaR objective. Finally, we also compare with a risk-averse MDP algorithm *Chow* [20], which is related to RASR-ERM. It augments the state space is a way that is superficially similar to our time-dependent value functions. We use risk-averse methods with the average model which is possible thanks to our results in Corollaries 3.3 and 4.2. The downside of Chow is that augmented state space is infinite and policies are history dependent.

Table 1 summarizes the risk $\mathrm{EVaR}_{\pi}^{0.99}[\mathfrak{R}_{\infty}^{\pi}]$ for policies π computed by RASR-EVaR and the baseline algorithms described above. The results show that RASR-EVaR chooses the *appropriate* level of risk-aversion across all domains. The plots in Figure 1 help to visualize the situation for two of the domains. Derman et al., which is risk neutral, performs particularly poorly in *Riverswim*, which has small but impactful epistemic risk. Risk averse algorithms, like RSVF and Erik, perform well in the domain. In contrast, Derman et al., performs well in *Population*, which involves large but inconsequential epistemic uncertainty. The risk-averse algorithms, RSVF, BCR, put too much emphasis on the epistemic uncertainty in this domain and compute policies that are too conservative.

Examining the results in Figure 1 closer leads one to several other important conclusions. First, the figures show that RASR-EVaR outperforms other algorithms even when the risk is evaluated using CVaR or VaR and may be a viable approximate approach optimizing these other risk measures. Second, the results in Figure 1 point to the importance of using the time dependent risk in the dynamic program equations. The *Naive* algorithm performs poorly compared with RASR-EVaR.

6 Related Work

In this section, we discuss how the RASR models and algorithms proposed in this paper are related to the existing results in soft-robust and risk averse decision-making.

Our RASR framework falls under the broader umbrella of robust and soft-robust MDP and RL. Robust optimization is a methodology that reduces the sensitivity of the solution to model errors [8] and has been extensively studied in MDP [38, 40, 48, 60] and RL [37, 51, 56, 61]. Since robust MDPs often compute policies that are overly conservative, soft-robust (also known as Bayesian robust, light robust, or multi-model objectives) formulations were proposed as an alternative that can better balance robustness and the quality of an average solution (e.g., [7, 16, 26, 44]). Soft-robust algorithms replace the worst-case objective of robust optimization with risk-aversion to some distribution over uncertain models. Table 2 summarizes representative soft-robust and risk-averse algorithms studied in the MDP/RL literature which we use for the empirical comparison. We defer a more comprehensive overview of related work to Appendix G.

The risk-averse MDP methods account only for the aleatory uncertainty in the return random variable and do not explicitly consider the error in the model. The risk-averse formulations that are most closely related to our work use ERM. This list includes the results in the average reward [10–12] and those in the undiscounted finite-horizon settings [27, 29, 46]. Note that some of these papers address risk-aversion in stochastic programming and not in MDPs [27]. To the best of our knowledge, none of the prior work has studied ERM in the discounted case. We believe this is because ERM is not positive-homogeneous, which complicates using it with a discount factor, as shown in Theorem 3.1. Moreover, we are unaware of any EVaR adaptation of these earlier ERM algorithms. Most other formulations for risk-averse RL are based on VaR and CVaR [12, 18, 19, 59], which are not dynamically consistent and generally NP hard to optimize. To build a DP in these formulations, one must augment the state space and optimize over a continuously infinite variable [5, 18, 20, 52], which significantly complicates the computation in comparison with the time-dependent value functions in RASR-ERM.

7 Conclusion and Future Work

We proposed a framework, called RASR, that can mitigate the risk associated with both model uncertainty(epistemic) and random dynamics (aleatory) in MDPs. We studied RASR with two separate risk measures: ERM and EVaR. In RASR-ERM, we derived the first exact DP formulation for ERM in discounted MDPs. We also showed that the optimal value function exists, the optimal policy is time-dependent and deterministic, and we proposed VI algorithms. For RASR-EVaR, we show that RASR-EVaR optimization can be optimized by reducing it to multiple RASR-ERM problems. Our empirical results highlight the utility of our RASR algorithms.

Future directions include scaling our RASR algorithms beyond tabular MDPs and dynamic epistemic uncertainty. It is also essential to better understand the relation between RASR and regularized (robust) MDPs [25, 36, 47].

References

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

[2] Giorgio Angelotti, Nicolas Drougard, and Caroline Ponzoni Carvalho Chanel. Exploitation vs caution: Risk-sensitive policies for offline learning. *arXiv:2105.13431 [cs, eess]*, May 2021.

[3] Philippe Artzner, Freddy Delbaen, Jean-marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

[4] Philippe Artzner, Freddy Delbaen, Jean Marc Eber, David Heath, and Hyejin Ku. Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, 2004.

[5] Nicole Bauerle and Jonathan Ott. Markov Decision Processes with Average-Value-at-Risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

[6] Bahram Behzadian, Reazul Russel, Chin Pang Ho, and Marek Petrik. Optimizing percentile criterion using robust MDPs. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, 2021.

[7] Aharon Ben-Tal, Dmitris Bertsimas, and David B. Brown. A soft robust model for optimization under ambiguity. *Operations Research*, 2010.

[8] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[9] Aharon Ben-Tal and Marc Teboulle. An Old-New Concept of Convex Risk Measures: The Optimized Certainty Equivalent. *Mathematical Finance*, 17:449–476, 2007.

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

[11] 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, February 2002.

[12] Vivek Borkar and Rahul Jain. Risk-constrained Markov decision processes. *IEEE Transactions on Automatic Control*, 2014.

[13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013.

[14] Daniel S. Brown, Scott Niekum, and Marek Petrik. Bayesian robust optimization for imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] Peter Buchholz and Dimitri Scheftelowitsch. Light robustness in the optimization of Markov decision processes with uncertain parameters. *Computers and Operations Research*, 108:69–81, 2019.

[16] Peter Buchholz and Dimitri Scheftelowitsch. Concurrent MDPs with finite Markovian policies. In *Measurement, Modeling, and Evaluation of Computing*, pages 37–53, 2020.

[17] Katherine Chen and Michael Bowling. Tractable objectives for robust policy optimization. *Advances in Neural Information Processing Systems*, 3:2069–2077, 2012.

[18] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. *Advances in Neural Information Processing Systems*, 2014.

[19] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 2018.

[20] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making : A CVaR optimization approach. In *Neural Information Processing Systems (NIPS)*, 2015.

[21] Jakša Cvitanić and Ioannis Karatzas. On dynamic measures of risk. *Finance and Stochastics*, 1999.

[22] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 2009.

[23] Erick Delage, Daniel Kuhn, and Wolfram Wiesemann. "Dice"-sion-making under uncertainty: When can a random decision reduce risk? *Management Science*, 65(7):3282–3301, July 2019.

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

[25] Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized MDPs and the equivalence between robustness and regularization. *arXiv:2110.06267 [cs, math]*, October 2021.

[26] Esther Derman, Daniel J. Mankowitz, Timothy A. Mann, and Shie Mannor. Soft-robust actor-critic policy-gradient. *Conference on Uncertainty in Artificial Intelligence*, 2018.

[27] Oscar Dowson, David P Morton, and Bernardo K Pagnoncelli. Multistage stochastic programs with the entropic risk measure. *Preprint in Optimization Online*, 2021.

[28] Hannes Eriksson and Christos Dimitrakakis. Epistemic risk-sensitive reinforcement learning. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2020.

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

[30] H. Follmer and A. Schied. Convex and coherent risk measures. *Preprint*, 2008.

[31] Hans Föllmer and Thomas Knispel. Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 2011.

[32] Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 2002.

[33] Hans Follmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, 2004.

[34] Marco Frittelli and Emanuela Rosazza Gianin. Dynamic convex risk measure. *Risk measures for the 21st century*, 2004.

[35] Marco Frittelli and Emanuela Rosazza Gianin. Putting order in risk measures. *Journal of Banking and Finance (JBF)*, 2002.

[36] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning (ICML)*, 2019.

[37] Julien Grand-Clement and Christian Kroer. First-order methods for Wasserstein distributionally robust MDPs. In *International Conference of Machine Learning (ICML)*, 2021.

[38] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

[39] Dan A. Iancu, Marek Petrik, and Dharmashankar Subramanian. Tight approximations of dynamic risk measures. *Mathematics of Operations Research*, 40(3):655–682, 2015.

[40] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 2005.

[41] Zaynah Javed, Daniel Brown, Satvik Sharma, Jerry Zhu, Ashwin Balakrishna, Marek Petrik, Anca Dragan, and Ken Goldberg. Policy gradient Bayesian robust optimization for imitation learning. In *International Conference on Machine Learning (ICML)*, 2021.

[42] Michael Kupper and Walter Schachermayer. Representation results for law invariant time consistent functions. *Mathematics and Financial Economics*, 16(2):419–441, 2006.

[43] Elita A. Lobo, Mohammad Ghavamzadeh, and Marek Petrik. Soft-robust algorithms for batch reinforcement learning. *Arxiv*, 2021.

[44] Daniel J. Mankowitz, Nir Levine, Rae Jeong, Yuanyuan Shi, Jackie Kay, Abbas Abdolmaleki, Jost Tobias Springenberg, Timothy Mann, Todd Hester, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification, 2019.

[45] Pascal Massart. *Concentration Inequalities and Model Selection*. Springer, 2003.

[46] David Nass, Boris Belousov, and Jan Peters. Entropic risk measure in policy search. *Investment Management and Financial Innovations*, 2020.

[47] Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *Arxiv*, 2017.

[48] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, sep 2005.

[49] Takayuki Osogami. Iterated risk measures for risk-sensitive Markov decision processes with discounted. In *Uncertainty in Artificial Intelligence*, 2011.

[50] Marek Petrik and Dharmashankar Subramanian. An approximate solution method for large risk-averse Markov decision processes. In *Uncertainty in Artificial Intelligence (UAI)*, 2012.

[51] Marek Petrik and Dharmashankar Subramanian. RAAM : The benefits of robustness in approximating aggregated MDPs in reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2014.

[52] Georg Ch Pflug and Alois Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. *Mathematics of Operations Research*, 41(2):682–699, 2016.

[53] Georg Ch Pflug and Andrzej Ruszczyński. Measuring risk for income streams. *Computational Optimization and Applications*, 2005.

[54] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

[55] Frank Riedel. Dynamic coherent risk measures. *Stochastic processes and their applications*, 2004.

[56] Reazul Hasan Russel and Marek Petrik. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in Neural Information Processing Systems*, 2019.

[57] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: Modeling and theory*. 2014.

[58] Lauren N Steimle, David L Kaufman, and Brian T Denton. Multi-model Markov decision processes. *IISE Transactions*, Forthcoming, 2021.

[59] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Neural Information Processing Systems*, 2015.

[60] Wolfram Wiesemann, Daniel Kuhn, and Berc Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 2013.

[61] Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 2012.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We foresee no immediate societal impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The confidence intervals are negligible and clutter the figure.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] But we do not point to the exact location of the assets and code because this would reveal the authors of the paper.
   (b) Did you mention the license of the assets? [N/A] There are no new significant assets
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] There are no new significant assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] There is no real-world data used in this work.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No personal information is used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No humans were involved in this research (besides the authors, of course).
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs of Section 2

Proof of Theorem 2.1. To prove this property, we use the certainty equivalence representation of ERM (e.g., [9]):

$$\mathrm{ERM}^\alpha[X] = u^{-1}(\mathbb{E}[u(X)])\ ,$$

where $u(X) = e^{-\alpha X}$ is a utility function. Then, since u is invertible, we obtain the following by algebraic manipulation and basic properties of the expectation:

$$\begin{aligned}
\mathrm{ERM}^\alpha\left[\mathrm{ERM}^\alpha[X_1 \mid X_2]\right] &= u^{-1}\left(\mathbb{E}\left[u\left(u^{-1}\left(\mathbb{E}[u(X_1) \mid X_2]\right)\right)\right]\right) \\
&= u^{-1}(\mathbb{E}\left[\mathbb{E}[u(X_1) \mid X_2]\right]) \\
&= u^{-1}(\mathbb{E}[u(X_1)]) \\
&= \mathrm{ERM}^\alpha[X_1]\ ,
\end{aligned}$$

which proves the desired result. $\square$

B Proofs of Section 3

Algorithm 3: VI for finite-horizon RASR-ERM

Input: Horizon $T < \infty$, risk level $\alpha > 0$, 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 $t = T-1, \ldots, 0$ **do**
  Update $v_t^\star$ using (11) and $\pi_t^\star$ using (12);
return $v^\star, \pi^\star$;

Proof of Theorem 3.1. The proof of the result follows by algebraic manipulation using the definition of ERM and the fact that $\alpha > 0$. First, assume that $c > 0$. Then:

$$\begin{aligned}
\mathrm{ERM}^{\alpha\cdot c}[X] &= -\frac{1}{\alpha c}\log\left(\mathbb{E}[e^{-\alpha\cdot c\cdot X}]\right) \\
c\cdot\mathrm{ERM}^{\alpha\cdot c}[X] &= -\frac{1}{\alpha}\log\left(\mathbb{E}[e^{-\alpha\cdot c\cdot X}]\right) && \text{Multiply by } c \\
c\cdot\mathrm{ERM}^{\alpha\cdot c}[X] &= \mathrm{ERM}^\alpha[c\cdot X].
\end{aligned}$$

The desired equality is trivially true for $c = 0$ and, therefore, the result holds for any $c \geq 0$. $\square$

Proof of Theorem 3.2. We prove the result only for $v_t^\star$; the result for v_t^π follows analogously. The proof proceeds by induction on the time step t for all risk-levels α assuming a discount rate γ. The base case with $t = T$ follows trivially. For the inductive step, assume the claim holds for $t+1$ and we show that it also holds for $t \geq 0$:

$$\begin{aligned}
v_t^\star(s) &\overset{\text{(a)}}{=} \max_{a\in\mathcal{A}}\left\{\mathrm{ERM}^\alpha\left[r(s,a) + \gamma\cdot v_{t+1}^\star(S')\right]\right\} \\
&\overset{\text{(b)}}{=} \max_{a\in\mathcal{A}}\left\{\mathrm{ERM}^\alpha\left[r(s,A) + \gamma\max_{\pi\in\Pi_{\mathrm{MR}}}\mathrm{ERM}^{\alpha\gamma}\left[\sum_{t=1}^{n+1}\gamma^{t-1}\cdot r(S_t,A_t) \mid S',\pi\right]\right]\right\} \\
&\overset{\text{Lem D.1}}{=} \max_{\pi\in\Pi_{\mathrm{MR}}}\left\{\mathrm{ERM}^\alpha\left[r(s,A) + \gamma\,\mathrm{ERM}^{\alpha\gamma}\left[\sum_{t=1}^{n+1}\gamma^{t-1}\cdot r(S_t,A_t) \mid S',\pi\right]\right]\right\} \\
&\overset{\text{Thm 3.1}}{=} \max_{\pi\in\Pi_{\mathrm{MR}}}\left\{\mathrm{ERM}^\alpha\left[r(s,A) + \mathrm{ERM}^{\alpha}\left[\sum_{t=1}^{n+1}\gamma^{t}\cdot r(S_t,A_t) \mid S',\pi\right]\right]\right\} \\
&\overset{\text{(c)}}{=} \max_{\pi\in\Pi_{\mathrm{MR}}}\left\{\mathrm{ERM}^\alpha\left[\sum_{t=0}^{n+1}\gamma^{t}\cdot r(S_t,A_t) \mid S_0 = s,\pi\right]\right\} \\
&= \max_{\pi\in\Pi_{\mathrm{MR}}}\left\{\mathrm{ERM}^\alpha\left[\mathfrak{R}_{n+1} \mid S_0 = s,\pi\right]\right\}\ ,
\end{aligned}$$

which is the definition of the value function. The equality (a) follows from the statement of the theorem, the equality (b) follows from the inductive assumption, and the equality marked by (c) follows by the translation equivariance of ERM (see Appendix E). The result readily generalizes to the infinite-horizon by considering the limit with $T \to \infty$ and using the fact that $\mathfrak{R}^{\pi}_{\infty}$ is bounded when $\gamma < 1$. The dynamic program representation for any fixed policy π follows analogously, replacing the maximization by a fixed policy. $\square$

Proof of Corollary 3.3. This result builds on the tower property in Theorem 2.1. To prove it, we use the certainty equivalence representation of ERM (e.g. [9]):

$$\mathrm{ERM}^{\alpha}[X] = u^{-1}(\mathbb{E}[u(X)]) ,$$

where $u(X) = e^{-\alpha X}$ is a utility function. Using this representation we can derive the desired equality as

$$\begin{aligned}
\mathrm{ERM}^{\alpha}\left[\mathrm{ERM}^{\alpha}[\mathfrak{R}^{\pi}_{T} \mid P]\right] &= u^{-1}\left(\mathbb{E}\left[u\left(u^{-1}\left(\mathbb{E}\left[u(\mathfrak{R}^{\pi}_{T}) \mid P\right]\right)\right)\right]\right) \\
&= u^{-1}\left(\mathbb{E}\left[\mathbb{E}\left[u(\mathfrak{R}^{\pi}_{T}) \mid P\right]\right]\right) \\
&\overset{\text{(a)}}{=} u^{-1}(\mathbb{E}[u(\mathfrak{R}^{\pi}_{T}) \mid \bar{P}]) \\
&= \mathrm{ERM}^{\alpha}[\mathfrak{R}^{\pi}_{T} \mid \bar{P}]
\end{aligned}$$

The step (a) follows from the tower property of the expectation operator using the fact that P_t random variables are independent because of dynamic uncertainty assumption described in Section 2. $\square$

Proof of Theorem 3.4. The existence of an optimal deterministic policy follows directly from the dynamic program formulation in Theorem 3.2 which uses the technical result in Lemma D.1. Here, we prove that an optimal RASR-ERM policy can be chosen to be greedy to the value function. The proof proceeds by mathematical induction. The base case follows from the statement of Lemma D.1 as

$$\max_{a\in\mathcal{A}} \mathrm{ERM}^{\alpha}[r(s,a)] \geq \mathrm{ERM}^{\alpha}_{A\sim\pi}\left[\mathrm{ERM}^{\alpha}[r(s,A) \mid A]\right]$$

Next, given $v^{\star}_{t+1}(\alpha\gamma, s')$ is achieved by the greedy policy, then also $v^{\star}_{t}(s)$ is achieved using the greedy policy. The proof of the inductive step proceeds by deriving a contradiction. Assume that there exist a $\pi' \in \Pi_{MR}$ such that $v^{\pi'}_{t}(s) > v^{\star}_{t}(s)$.

$$\begin{aligned}
v^{\star}_{t}(s) &= \max_{a\in\mathcal{A}} \mathrm{ERM}^{\alpha}\left[r(s,a,S') + \gamma \cdot v^{\star}_{t+1}(S')\right] \\
&\geq \mathrm{ERM}^{\alpha}_{A\sim\pi'(s)}\left[\mathrm{ERM}^{\alpha}\left[r(s,A) + \gamma \cdot v^{\star}_{t+1}(S') \mid A\right]\right] \\
&\geq \mathrm{ERM}^{\alpha}_{A\sim\pi'(s)}\left[\mathrm{ERM}^{\alpha}\left[r(s,A) + \gamma \cdot v^{\pi'}_{t+1}(S')\right]\right] \\
&= v^{\pi'}_{t+1}(s') .
\end{aligned}$$

The last statement follows because $v^{\star}_{t+1}(s') \geq v^{\pi'}_{t+1}(s')$ by the inductive assumption. Since this derives a contradiction with the optimality of $v^{\star}$, there exist no π' such that $v^{\pi'}(\alpha, s) > v^{\star}(\alpha, s)$ given that $v^{\star}(\alpha\gamma, s)$ is selected greedily. $\square$

Lemma B.1. *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 $\alpha > 0$, we have $\mathbb{E}[X] - \epsilon(\alpha) \leq \mathrm{ERM}^{\alpha}[X] \leq \mathbb{E}[X]$, where*

$$\epsilon(\alpha) \;=\; 8^{-1} \cdot \alpha \cdot (x_{\max} - x_{\min})^2 .$$

The gap vanishes with a decreasing risk: $\lim_{\alpha\to 0} \epsilon(\alpha) = 0$.

Proof of Lemma B.1. To simplify the notation, let $X = \mathfrak{R}^{\pi}_{T}$ for any policy π which is bounded between $x_{\min}$ and $x_{\max}$. We begin the our proof with the Hoeffding's lemma [13, 45]

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &\leq e^{\lambda\mathbb{E}[x] + \frac{\lambda^2 (x_{\max} - x_{\min})^2}{8}} \qquad , \forall \lambda \in \mathbb{R} \\
\log\left(\mathbb{E}[e^{\lambda X}]\right) &\leq \lambda\mathbb{E}[x] + \frac{\lambda^2 (x_{\max} - x_{\min})^2}{8} .
\end{aligned}$$

Then, substitute $\lambda = -\alpha$ into the equation above to get

$$\begin{aligned} \log\left(\mathbb{E}[e^{-\alpha X}]\right) &\le -\alpha \cdot \mathbb{E}[x] + \frac{\alpha^2 \cdot (x_{\max} - x_{\min})^2}{8} \\ -\frac{1}{\alpha}\log\left(\mathbb{E}[e^{-\alpha X}]\right) &\ge \mathbb{E}[x] - \frac{\alpha(x_{\max} - x_{\min})^2}{8} \\ \mathbb{E}[x] - \frac{\alpha(x_{\max} - x_{\min})^2}{8} &\le \mathrm{ERM}^{\alpha}[X] \,. \end{aligned}$$

Now we have that $\mathbb{E}[X] - \epsilon(\alpha) \le \mathrm{ERM}^{\alpha}[X]$ where $\epsilon(\alpha) = 8^{-1}\alpha(x_{\max} - x_{\min})^2$, and $\mathrm{ERM}^{\alpha}[X] \le \mathbb{E}[X]$ for $\alpha > 0$ is shown in Lemma D.1. Furthermore this upper bound vanishes as alpha decreases to zero: $\lim_{\alpha\to 0} 8^{-1}\alpha(x_{\max} - x_{\min})^2 = 0$. □

Proof of Theorem 3.5. To simplify the notation in the proof we use $\hat{\pi}$ in place of $\hat{\pi}^\star$ throughout the proof.

The main idea of the proof is to lower-bound the value function $v^{\hat{\pi}}$ of the policy $\hat{\pi}$ using the value function v^{∞} of the optimal risk-neutral policy. Recall that Lemma B.1 bounds the error between the risk-neutral and ERM value function of any policy π and any $t = 0, \ldots$:

$$0 \le v_{\pi}^{\infty} - v_t^{\pi} \le \frac{\alpha \cdot \gamma^t \cdot (\triangle r)^2}{8 \cdot (1-\gamma)^2} \,. \tag{15}$$

The symbol v_{π}^{∞} denotes the ordinary risk-neutral (ERM^0) γ-discounted infinite-horizon value function of the policy π. Note that this value function is stationary. The left-hand side of the equation above holds because $\mathbb{E}$ is an upper bound on the ERM.

As the first step of the proof, we bound the error at time T' as follows. Consider any state $s \in \mathcal{S}$, then:

$$\begin{aligned} v_{T'}^{\star}(s) - v_{T'}^{\hat{\pi}}(s) &\le v_{T'}^{\star}(s) - v_{\hat{\pi}}^{\infty}(s) + \frac{\alpha \cdot \gamma^{T'} \cdot (\triangle r)^2}{8 \cdot (1-\gamma)^2} && \text{from r.h.s of (15)} \\ &\le v_{\pi^\star}^{\infty}(s) - v_{\hat{\pi}}^{\infty}(s) + \frac{\alpha \cdot \gamma^{T'} \cdot (\triangle r)^2}{8 \cdot (1-\gamma)^2} && \text{from l.h.s. of (15)} \\ &\le \frac{\alpha \cdot \gamma^{T'} \cdot (\triangle r)^2}{8 \cdot (1-\gamma)^2} && \text{from } \hat{\pi} \in \arg\max_{\pi\in\Pi} v_{\pi}^{\infty}(s) \,. \end{aligned}$$

As the second step of the proof, we construct an approximation $u_t \in \mathbb{R}^S, t = 0, \ldots, T'$ of the value function $v_t^{\hat{\pi}}$ for $t = 0, \ldots, T'-1$ and all $s \in \mathcal{S}$ as:

$$\begin{aligned} u_{T'}(s) &= v_{\hat{\pi}}^{\infty} - \frac{\alpha \cdot \gamma^{T'} \cdot (\triangle r)^2}{8 \cdot (1-\gamma)^2} \\ u_t(s) &= \max_{a\in\mathcal{A}} \mathrm{ERM}^{t\cdot\gamma^t}\left[r(s,a) + \gamma \cdot u_{t+1}(S'_{t+1,a})\right] \\ &= \mathrm{ERM}^{t\cdot\gamma^t}\left[r(s,\hat{\pi}(s)) + \gamma \cdot u_{t+1}(S'_{t+1,\hat{\pi}(s)})\right] \,, \end{aligned}$$

where $S'_{t+1,a}$ denotes the random variable that represents the state that follows s at time $t+1$ after taking an action a. The last equality holds from the definition of $\hat{\pi}_t$ being greedy with respect to u_t; subtracting a constant from all states does not change the greedy policy. The function u_t is constructed to be a lower bound on $v_t^{\hat{\pi}}$ and at the same time be a value such that $\hat{\pi}$ is greedy to it.

From (15), we have that $v^{\pi}_{T'}(s) \geq u_{T'}(s)$ for all $s \in \mathcal{S}$. Then, assuming $v^{\pi}_{t+1}(s) \geq u_{t+1}(s)$ for all $s \in \mathcal{S}$, we can use backward induction on t to show that

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

The equality (a) is shown by subtracting the constant reward from both terms which can be done because ERM is translation equivariant. The equality (b) follows from the positive quasi-homogeneity in Theorem 3.1, and (c) follows from the monotonicity of ERM.

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

$$v^{\star}_t(s) - u_t(s) \leq \gamma^{T'-t}\cdot\frac{\alpha\cdot\gamma^{T'}\cdot(\triangle r)^2}{8\cdot(1-\gamma)^2}\,. \tag{16}$$

The inequality (16) holds for $t = T'$ by (15) and the construction of $u_{T'}$. To prove (16) by induction, assume it holds for $t+1$. Then for each $s \in \mathcal{S}$:

$$\begin{aligned}
v^{\star}_t(s) - u_t(s) &\overset{\text{(a)}}{=} \mathrm{ERM}^{t\cdot\gamma^t}\left[r(s,\pi^\star_t(s)) + \gamma\cdot v^{\star}_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] - \\
&\quad - \mathrm{ERM}^{t\cdot\gamma^t}\left[r(s,\hat{\pi}_t(s)) + \gamma\cdot u_{t+1}(S'_{t+1,\hat{\pi}_t(s)})\right] \\
&\overset{\text{(b)}}{\leq} \mathrm{ERM}^{t\cdot\gamma^t}\left[r(s,\pi^\star_t(s)) + \gamma\cdot v^{\star}_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] - \\
&\quad - \mathrm{ERM}^{t\cdot\gamma^t}\left[r(s,\pi^\star_t(s)) + \gamma\cdot u_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] \\
&\overset{\text{(c)}}{=} \mathrm{ERM}^{t\cdot\gamma^t}\left[\gamma\cdot v^{\hat{\pi}}_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] - \mathrm{ERM}^{t\cdot\gamma^t}\left[\gamma\cdot u_{t+1}(S'_{t+1,\pi^\star_t(s)})\right] \\
&\overset{\text{(d)}}{=} \gamma\cdot\left(\mathrm{ERM}^{t\cdot\gamma^{t+1}}\left[v^{\pi^\star}_{t+1}(S'_{t+1,\pi^\star(s)})\right] - \mathrm{ERM}^{t\cdot\gamma^{t+1}}\left[u_{t+1}(S'_{t+1,\pi^\star(s)})\right]\right)
\end{aligned} \tag{17}$$

The equality (a) is derived from the definition, (b) follows from $\hat{\pi}$ being greedy with respect to u, (c) follows by subtracting the constant reward from both terms which can be done because ERM is translation equivariant. Finally, the equality (d) follows from the positive quasi-homogeneity in Theorem 3.1. Then, from the inductive assumption we get the desired inequality from the monotonicity and translation equivariance of ERM by bounding the terms in (17) above as:

$$\begin{aligned}
v^{\pi^\star}_{t+1}(s) - u_{t+1}(s) &\leq \gamma^{T'-t-1}\cdot\frac{\alpha\cdot\gamma^{T'}\cdot(\triangle r)^2}{8\cdot(1-\gamma)^2} \qquad \forall s\in\mathcal{S} \\
\mathrm{ERM}^{t\cdot\gamma^{t+1}}[v^{\pi^\star}_{t+1}(S)] - \mathrm{ERM}^{t\cdot\gamma^{t+1}}[u_{t+1}(S)] &\leq \gamma^{T'-t-1}\cdot\frac{\alpha\cdot\gamma^{T'}\cdot(\triangle r)^2}{8\cdot(1-\gamma)^2} \\
\gamma\cdot(\mathrm{ERM}^{t\cdot\gamma^{t+1}}[v^{\pi^\star}_{t+1}(S)] - \mathrm{ERM}^{t\cdot\gamma^{t+1}}[u_{t+1}(S)]) &\leq \gamma^{T'-t}\cdot\frac{\alpha\cdot\gamma^{T'}\cdot(\triangle r)^2}{8\cdot(1-\gamma)^2}\,.
\end{aligned}$$

The second line holds for S distributed arbitrarily and substituting $S = S'_{t+1,\pi^\star_{t+1}(s)}$ from (17) proves the bound on u_t.

The theorem then follows form the properties established above as

$$\mathrm{ERM}^{\alpha}\left[\mathfrak{R}^{\pi^\star}_{\infty} \mid \bar{P}\right] - \mathrm{ERM}^{\alpha}\left[\mathfrak{R}^{\hat{\pi}^\star}_{\infty} \mid \bar{P}\right] = v^\star_0(s_0) - v^{\hat{\pi}}_0(s_0) \leq v^\star_0(s_0) - u_0 \leq \frac{\alpha\cdot\gamma^{2\cdot T'}\cdot(\triangle r)^2}{8\cdot(1-\gamma)^2}$$

□

C Proofs of Section 4

Proof of Theorem 4.1. We prove the contra-positive: If $\pi^\star$ is not optimal policy in RASR-ERM for all $\alpha > 0$, then $\pi^\star$ is not an optimal solution to RASR-EVaR. Assume $\pi^\star$ is not an optimal policy for all $\alpha > 0$, and π_α is an optimal policy for RASR-ERM$^\alpha$,

$$\begin{aligned} \mathrm{ERM}^\alpha\left[X^{\pi^\star}\right] &< \mathrm{ERM}^\alpha\left[X^{\pi_\alpha}\right] \qquad ,\forall\, \alpha > 0 \\ \sup_{\alpha>0}\left\{\mathrm{ERM}^\alpha[X^{\pi^\star}] + \frac{\log(1-\beta)}{\alpha}\right\} &< \sup_{\alpha>0}\left\{\mathrm{ERM}^\alpha[X^{\pi_\alpha}] + \frac{\log(1-\beta)}{\alpha}\right\} \\ \mathrm{EVaR}^\beta[X] &< \sup_{\alpha>0}\left\{\mathrm{ERM}^\alpha\left[X^{\pi_\alpha}\right] + \frac{\log(1-\beta)}{\alpha}\right\} \end{aligned}$$

We prove that if $\pi^\star$ is not optimal policy in RASR-ERM for all $\alpha > 0$, then $\pi^\star$ is not an optimal solution to RASR-EVaR. With contra-positive we prove that if $\pi^\star$ is an optimal solution to RASR-EVaR$^\beta$ in (13) then there exists $\alpha^\star$ such that $\pi^\star$ is optimal in RASR-ERM with risk level $\alpha = \alpha^\star$. □

Proof of Corollary 4.2. Theorem 4.1 shows that the optimal policy $\pi^\star$ for $\mathrm{EVaR}^\beta[X^{\pi^\star}]$ implies there exists $\alpha^\star$ such that $\mathrm{ERM}^{\alpha^\star}[X^{\pi^\star}]$ is optimal in RASR-ERM and Theorem 3.4 shows that there exists a markovian deterministic time-dependent optimal policy $\pi^\star = (\pi_t^\star)_{t=0}^{T-1} \in \Pi_{MD}$ for (8). Therefore there exists a markovian deterministic time-dependent optimal policy $\pi^\star$ which optimizes the EVaR objectives $\mathrm{EVaR}^\beta[X^{\pi^\star}]$.

The second part of the corollary can be shown as follows. For any policy $\pi \in \Pi_{MR}$, the RASR-EVaR objective in (13) can be written as

$$\begin{aligned} \mathrm{EVaR}^\beta\left[\mathfrak{R}_T^\pi\right] &= \sup_{\alpha>0}\left(\mathrm{ERM}^\alpha[\mathfrak{R}_T^\pi] + \frac{\log(1-\beta)}{\alpha}\right) \\ &= \sup_{\alpha>0}\left(\mathrm{ERM}^\alpha[\mathfrak{R}_T^\pi \mid \bar{P}] + \frac{\log(1-\beta)}{\alpha}\right) \\ &= \mathrm{EVaR}^\beta\left[\mathfrak{R}_T^\pi \mid \bar{P}\right]\ . \end{aligned}$$

□

The following lemma plays an important role in bounding the error introduced by discretizing the risk-level α in Algorithm 2.

Lemma C.1. *Suppose that the supremum of (14) is attained at $\alpha^\star$ such that $\alpha_0 \geq \alpha^\star \geq \alpha_K$, and $h(\hat{\alpha}) \geq h(\alpha_k)$ for $k = 0, \ldots, K$ and some $\alpha_0 \geq \cdots \geq \alpha_K$. Then*

$$h(\alpha^\star) - h(\hat{\alpha}) \leq \log(1-\beta) \max_{k\in 0,\ldots,K-1}\left(\alpha_k^{-1} - \alpha_{k+1}^{-1}\right)\ .$$

Also, $h(\alpha^\star) - h(\hat{\alpha}) \leq -\log(1-\beta)\alpha_0^{-1}$ when $\alpha^\star > \alpha_0$.

Proof. Given that the optimal risk $\alpha_{l+1} \leq \alpha^\star \leq \alpha_l$, where α_l and α_{l+1} are in the set of ERM levels Λ we have computed. We can bound

$$\mathrm{EVaR}^\beta(X) - \max_{\alpha\in\Lambda}\left\{\mathrm{ERM}^\alpha[X] + \frac{\log(1-\beta)}{\alpha}\right\} \leq \log(1-\beta)\left(\frac{1}{\alpha_l} - \frac{1}{\alpha_{l+1}}\right)$$

By the monotonicity property of ERM we get

$$\begin{aligned} \mathrm{ERM}^{\alpha_l}[X_{\alpha_{l+1}}^\pi] &\leq \mathrm{ERM}^{\alpha_l}[X_{\alpha_l}^\pi] \leq \mathrm{ERM}^{\alpha^\star}[X_{\alpha_l}^\pi] \\ &\leq \mathrm{ERM}^{\alpha^\star}[X_{\alpha^\star}^\pi] \leq \mathrm{ERM}^{\alpha_{l+1}}[X_{\alpha^\star}^\pi] \leq \mathrm{ERM}^{\alpha_{l+1}}[X_{\alpha_{l+1}}^\pi] \end{aligned}$$

where X_α^π refers to the total discounted reward distribution deploying the optimal policy of ERM$^\alpha$. On the other hand,

$$\frac{\log(1-\beta)}{\alpha_{l+1}} \leq \frac{\log(1-\beta)}{\alpha^\star} \leq \frac{\log(1-\beta)}{\alpha_l}$$

We can conclude that

$$\mathrm{ERM}^{\alpha_l}[X^\pi_{\alpha_l}] + \frac{\log(1-\beta)}{\alpha_{l+1}} \le \mathrm{ERM}^{\alpha^\star}[X^\pi_{\alpha^\star}] + \frac{\log(1-\beta)}{\alpha^\star} \le \mathrm{ERM}^{\alpha_{l+1}}[X^\pi_{\alpha_{l+1}}] + \frac{\log(1-\beta)}{\alpha_l}$$

Therefore,

$$\begin{aligned}
&\mathrm{EVaR}^\beta(X) - \max_{\alpha\in\Lambda}\left\{\mathrm{ERM}^\alpha[X] + \frac{\log(1-\beta)}{\alpha}\right\} \\
&\le \mathrm{ERM}^{\alpha^\star}[X^\pi_{\alpha^\star}] + \frac{\log(1-\beta)}{\alpha^\star} - \max_{\alpha\in\{\alpha_{l+1}\}}\left\{\mathrm{ERM}^\alpha[X^\pi_\alpha] + \frac{\log(1-\beta)}{\alpha}\right\} \\
&\le \frac{\log(1-\beta)}{\alpha_l} - \frac{\log(1-\beta)}{\alpha_{l+1}} \\
&= \log(1-\beta)\left(\frac{1}{\alpha_l} - \frac{1}{\alpha_{l+1}}\right)
\end{aligned}$$

Now we relax the assumption to $\alpha^\star \in [\alpha_0, \alpha_K]$, and conclude that

$$h(\alpha^\star) - h(\hat\alpha) \le \max_{k=0,\dots,K-1}\left\{\log(1-\beta)\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k+1}}\right)\right\}$$

The last part of the theorem can be proved as follows. Given an arbitrary error tolerance δ, β and α_k Corollary D.2 shows that we can set $\alpha_{k+1} = (\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1}$ such that $h(\alpha^\star) - h(\hat\alpha) \le \delta$. Moreover for $\alpha^\star > \alpha_0$, given α_0 and β the error $h(\alpha^\star) - h(\hat\alpha) \le -\frac{\log(1-\beta)}{\alpha_0}$. □

Proof of Theorem 4.3. Assume $\alpha^\star \in \arg\max_{\alpha>0} h(\alpha)$ be the α that achieves the optimality in the definition $\mathrm{EVaR}^\beta\left[X\right] = \sup_{\alpha>0} h(\alpha)$. The supremum is achieved whenever $\beta > 0$ since then there exists an optimal $\alpha^\star > 0$. Then, $h(\alpha^\star) \ge h(\alpha^\star + \epsilon)$ for any $\epsilon > 0$

$$\begin{aligned}
h(\alpha^\star) &\ge h(\alpha^\star + \epsilon) \\
\mathrm{ERM}^{\alpha^\star}[X] + \frac{\log(1-\beta)}{\alpha^\star} &\ge \mathrm{ERM}^{\alpha^\star+\epsilon}[X] + \frac{\log(1-\beta)}{\alpha^\star+\epsilon} \\
\mathrm{ERM}^{\alpha^\star}[X] - \mathrm{ERM}^{\alpha^\star+\epsilon}[X] &\ge \frac{\log(1-\beta)}{\alpha^\star+\epsilon} - \frac{\log(1-\beta)}{\alpha^\star} \\
\frac{(\triangle r)^2}{8(1-\gamma)^2} \ge \frac{d(\mathrm{ERM}^{\alpha^\star}[X])}{d\alpha^\star} &\ge \log(1-\beta)\frac{d(\alpha^\star)^{-1}}{d\alpha^\star} \\
\frac{(\triangle r)^2}{8(1-\gamma)^2} &\ge -\log(1-\beta)(\alpha^\star)^{-2} \\
(\alpha^\star)^2 &\ge -\log(1-\beta)\frac{8(1-\gamma)^2}{(\triangle r)^2} \\
\alpha^\star &\ge \sqrt{-8\log(1-\beta)}\frac{(1-\gamma)}{(\triangle r)}
\end{aligned}$$

We let $\alpha_0 \to \infty$. Then, to achieve the desired bound, we need to choose the number of points K such that $\sqrt{-8\log(1-\beta)}\frac{1-\gamma}{\triangle r} \ge \alpha_K$. Then, following the construction in Corollary D.2, we get that $\alpha_K = \frac{-\log(1-\beta)}{K\delta}$ and

$$\begin{aligned}
\sqrt{-8\log(1-\beta)}\frac{1-\gamma}{\triangle r} &\ge \frac{-\log(1-\beta)}{K\delta} \\
K &\ge \sqrt{\frac{-\log(1-\beta)}{8}}\frac{\triangle r}{(1-\gamma)\delta}\,.
\end{aligned}$$

We conclude the proof with Lemma C.1 since $\alpha_0 \ge \alpha^\star \ge \alpha_K$. □

D Technical Lemmas

Lemma D.1 (Deterministic action)**.** *Let $A \colon \Omega \to \mathcal{A}$ be a random variable and $g : \mathcal{A} \to \mathbb{R}$ by any function. Then for any $\alpha \geq 0$:*

$$\max_{a\in\mathcal{A}} g(a) \;\geq\; \max_{\pi\in\Delta^\Omega} \mathrm{ERM}^\alpha_{A\sim\pi}\left[g(A)\right] \;.$$

Proof. To prove the lemma, use the well-known dual representation of $\mathrm{ERM}^\alpha_{A\sim\pi}[g(A)]$ [9]

$$\mathrm{ERM}^\alpha_{A\sim\pi}[g(A)] \;=\; \inf_{\bar\pi\in\Delta^\pi} \left\{ \mathbb{E}_{A\sim\bar\pi}[g(A)] + \frac{1}{\alpha} D_{\mathrm{KL}}(\bar\pi\|\pi) \right\} \;,$$

where D_{KL} refers to the KL-divergence metric. Because Ω is finite, we have for any $\pi \in \Delta^\Omega$ that

$$\max_{a\in\mathcal{A}} g(a) \geq \mathbb{E}_{A\sim\pi}\left[g(A)\right] \;.$$

Next, we use the dual representation of ERM to show that for any $\pi \in \Delta^\Omega$ that

$$\begin{aligned}
\mathrm{ERM}^\alpha_{A\sim\pi}[g(A)] &= \inf_{\bar\pi\in\Delta^\pi} \left\{ \mathbb{E}_{A\sim\bar\pi}[g(A)] + \frac{1}{\alpha} D_{\mathrm{KL}}(\bar\pi\|\pi) \right\} \\
&\leq \mathbb{E}_{A\sim\pi}[g(A)] + \frac{1}{\alpha} D_{\mathrm{KL}}(\pi\|\pi) \\
&= \mathbb{E}_{A\sim\pi}[g(a)] \;.
\end{aligned}$$

We used the fact that $D_{\mathrm{KL}}(\pi\|\pi) = 0$. The combination of the inequalities above proves the lemma. ☐

Corollary D.2. *Given an arbitrary error tolerance δ, β and α_k we construct α_{k+1} as $\alpha_{k+1} = (\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1}$ such that $\alpha_k \geq \alpha_{k+1} > 0$ and*

$$\log(1-\beta)\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k+1}}\right) = \delta \;.$$

Moreover, given α_{k+1} and β the error $\delta \leq -\frac{\log(1-\beta)}{\alpha_{k+1}}$.

Proof. Let $\alpha_{k+1} = c \cdot \alpha_k$ for $c \in (0,1)$, we can derive the following

$$\begin{aligned}
\log(1-\beta)(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k+1}}) &= \delta \\
\log(1-\beta)(\frac{c-1}{c\cdot\alpha_k}) &= \delta \\
c - 1 &= \frac{\delta\cdot c\cdot\alpha_k}{\log(1-\beta)} \\
c\cdot\alpha_k(\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)}) &= 1 \\
c\cdot\alpha_k &= (\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1} \\
\alpha_{k+1} &= (\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1}
\end{aligned}$$

Let α_k approach ∞, the reverse implication of α_{k+1} to the error δ can be evaluate as

$$\alpha_{k+1} = (\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1} \leq \lim_{\alpha_k\to\infty} (\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1} = -\frac{\log(1-\beta)}{\delta}$$

and conclude that

$$\delta \leq -\frac{\log(1-\beta)}{\alpha_{k+1}}$$

☐

E Risk Measures

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

Definition E.1 (Risk Measure)**.** A risk measure is a function $\psi : \mathbb{X} \to \mathbb{R}$ that maps a random variable $X \in \mathbb{X}$ to real numbers.

Definition E.2 (Coherent Risk Measure)**.** A risk measure ψ is *coherent* if it satisfies the following four axioms [3]:

A1. Monotonicity: $X_1 \le X_2 \ (a.s.) \Longrightarrow \psi[X_1] \le \psi[X_2], \ \ \forall X_1, X_2 \in \mathbb{X}.$

A2. Translation Equivariance: $\psi[c + X] = c + \psi[X], \ \ \forall c \in \mathbb{R}, \ \forall X \in \mathbb{X}.$

A3. (a) Sub-Additivity: $\psi[X_1 + X_2] \le \psi[X_1] + \psi[X_2], \ \ \forall X_1, X_2 \in \mathbb{X}.$
(b) Super-Additivity: $\psi[X_1 + X_2] \ge \psi[X_1] + \psi[X_2], \ \ \forall X_1, X_2 \in \mathbb{X}.$

A4. Positive Homogeneity: $\psi[cX] = c\psi[X], \ \ \forall c \in \mathbb{R}_+, \ \forall X \in \mathbb{X}.$

Axioms A3(a) and A3(b) are used for cost minimization and reward maximization, respectively.

Common coherent risk measures include CVaR^β, and EVaR^β that we define them below. Convex risk measures are a more general class of risk measures (than coherent risk measures) and are defined as

Definition E.3 (Convex Risk Measure)**.** A *convex* risk measure ψ satisfies axioms A1 and A2 (in Definition E.2) and replaces axioms A3 and A4 with the following axiom:

A5. (a) Convexity: $\psi\big[cX_1 + (1-c)X_2\big] \le c\psi[X_1] + (1-c)\psi[X_2], \ \ \forall c \in [0,1], \ \forall X_1, X_2 \in \mathbb{X}.$
(b) Concavity: $\psi\big[cX_1 + (1-c)X_2\big] \ge c\psi[X_1] + (1-c)\psi[X_2], \ \ \forall c \in [0,1], \ \forall X_1, X_2 \in \mathbb{X}.$

Axioms A5(a) and A5(b) are used for cost minimization and reward maximization, respectively.

Every coherent risk measure is a convex risk measure but the other way is not always true. In other words, if a risk measure satisfies A3 (sub or super additivity) and A4 (positive homogeneity), then it satisfies A5 (convexity), but the reverse is not always true. Entropic risk measure (ERM) is a common convex, but not coherent, risk measure.

E.1 Value-at-Risk

For a random variable $X \in \mathbb{X}$, its value-at-risk with confidence level β, denoted by $\mathrm{VaR}^\beta[X]$, is the $(1-\beta)$-quantile of X, i.e.,

$$\mathrm{VaR}^\beta[X] = \inf_{x \in \mathbb{R}} \big\{ F_X(x) > 1 - \beta \big\} = F_X^{-1}(1 - \beta), \quad \beta \in [0, 1),$$

where F_X is the cumulative distribution function of X.

E.2 Conditional Value-at-Risk

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

$$\mathrm{CVaR}^\beta[X] = \inf_{\zeta \in \mathbb{R}} \left(\zeta - \frac{1}{1-\beta} \cdot \mathbb{E}\big[(\zeta - X)_+\big] \right), \quad \beta \in [0, 1).$$

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

E.3 Entropic Risk Measure

For a random variable $X \in \mathbb{X}$, its entropic risk measure with risk parameter α, denoted by $\mathrm{ERM}^\alpha[X]$, is defined as

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

Properties of ERM:

1. It is easy to see that $\lim_{\alpha\to 0}\mathrm{ERM}^\alpha[X] = \mathbb{E}[X]$ and $\lim_{\alpha\to\infty}\mathrm{ERM}^\alpha[X] = \operatorname{ess\,inf}[X]$.
2. For any random variable $X \in \mathbb{X}$, we have $\mathrm{ERM}^\alpha[X] = \mathbb{E}[X] - \frac{\alpha}{2}\mathrm{VaR}[X] + o(\alpha)$.
3. If X is a Gaussian random variable, we have $\mathrm{ERM}^\alpha[X] = \mathbb{E}[X] - \frac{\alpha}{2}\mathrm{VaR}[X]$.
4. For any two random variables $X_1, X_2 \in \mathbb{X}$, we have $\mathrm{ERM}^\alpha[X_2|X_1] = -\frac{1}{\alpha}\log\left(\mathbb{E}[e^{-\alpha X_2}|X_1]\right)$.
5. Since ERM does not satisfy the axiom A4 (positive homogeneity), we have $\mathrm{ERM}^\alpha[cX] \neq c\,\mathrm{ERM}^\alpha[X]$.

E.4 Entropic Value-at-Risk

For a random variable $X \in \mathbb{X}$, its entropic value-at-risk with confidence level β, denoted by $\mathrm{EVaR}^\beta[X]$, is defined as

$$\mathrm{EVaR}^\beta[X] = \sup_{\alpha>0}\left(\mathrm{ERM}^\alpha[X] + \frac{\log(1-\beta)}{\alpha}\right), \quad \beta \in [0,1).$$

Properties of EVaR**:**

1. The EVaR with confidence level β is the tightest possible lower-bound that can be obtained from the Chernoff ineqaulity for VaR and CVaR with confidence level β, i.e.,
$$\mathrm{EVaR}^\beta[X] \le \mathrm{CVaR}^\beta[X] \le \mathrm{VaR}^\beta[X].$$
2. The following inequality also holds for the EVaR:
$$\operatorname{ess\,inf}[X] \le \mathrm{EVaR}^\beta[X] \le \mathbb{E}[X].$$
3. It is easy to see that $\mathrm{EVaR}^0[X] = \mathbb{E}[X]$ and $\lim_{\beta\to 1}\mathrm{EVaR}^\beta[X] = \operatorname{ess\,inf}[X]$.

E.5 Properties of Risk Measures

Table 3 summarizes some properties of convex risk measures that are desirable in RL and MDP.

Risk measure	LI	DC	PH
$\mathbb{E}$, Min	✓	✓	✓
CVaR	✓	·	✓
EVaR	✓	·	✓
ICVaR	·	✓	✓
ERM	✓	✓	·

Table 3: Properties of representative risk measures.

A *law-invariant* (LI) risk measure depends only on the total return and not on the particular sequence of individual rewards [30]. A *dynamically-consistent* (DC), or time-consistent, risk measure satisfies the tower property [57] and can be optimized using a dynamic program [4, 21, 24, 27, 34, 53, 55]. Finally, a positively-homogeneous (PH) risk measure satisfies $\psi(c \cdot X) = c \cdot \psi(X)$, for any $c \ge 0$, which is an important property in the risk-averse parameter selection and discounted setting [3, 30, 31]. Unfortunately, expectation ($\mathbb{E}[\cdot]$) and minimum (Min) are the only convex risk measures that satisfy all the desirable conditions. In Table 3, ICVaR is an iterated version of CVaR [39, 50].

Method	RS	POP	INV
RASR	< 2	24	< 7
Naive	27	175	186
Erik	1117	110306	9977
Chow	69	861	572

Table 4: Time (sec) to compute each algorithm

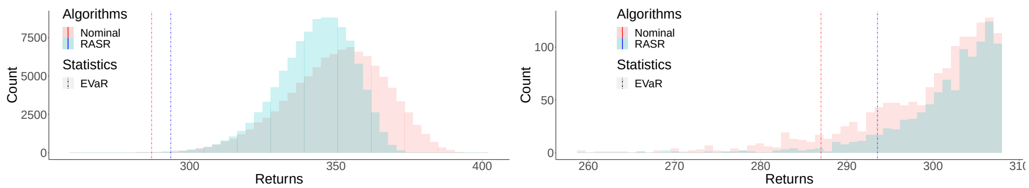

Figure 2: Full (left) and tail (right) histogram of return $\mathfrak{R}^{\pi}_{\infty}$ in the inventory domain.

F Additional Experimental Results and Details

Figure 2 compares the distribution of returns $\mathfrak{R}^{\pi}_{\infty}$ for a policy computed by RASR-EVaR with $\beta = 0.99$ with a policy computed by the *nominal* algorithm, which solves a regular MDP with $\bar{P}$. The histogram and the vertical lines that indicate EVaR values shows that the RASR policy significantly reduces the tail risk and improves the EVaR value at some cost to the average returns.

To remove bias and hyperparameter (Λ) tuning for our algorithm, we use the same set Λ for all domains. In all the numerical results of our paper, we only call Algorithm 1 once ($K = 1$) by using $\alpha = e^{10}$, $T' = (10 + 15)/(1 - \gamma)$ without discarding any intermediate α_t. By doing so, we have $\alpha_{0:(T'+1)} = \{e^{10}, e^{10}\gamma, e^{10}\gamma^2, ..., e^{10}\gamma^{T'}, 0\} = \Lambda$ for EVaR where $e^{-15} > e^{10}\gamma^{T'} \approx 0$. This method allow us to generate each α_t beyond 0 in one single value iteration.

Furthermore, for the Table 1 and Figure 1 in the main body of the paper, we sample 100,000 Monte-Carlo instances with 1,000 time horizon for each instance which take days to compute.

In the appendix and code for the supplementary material, to reduce time consumption and for reproducible purposes. We set an arbitrary seed (1), sample only 10,000 Monte-Carlo instances, and uses only 500 time horizon for each instance. The risk of return in the appendix are consistent with the paper despite generated with different Monte Carlo samples. In Table 5, all other benchmarks except Derman perform badly in population, and Derman perform poorly in riverswim. However, RASR is able to consistently mitigate risk of return when measured in all VaR, CVaR and EVaR for all domains. Moreover, RASR was able to be computed in polynomial-time and outperform the other benchmark algorithms in computation time [see Table 4] makes it the most practical method available for risk averse soft robust RL.

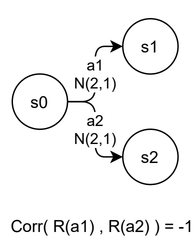

Figure 3: Example used to illustrate the difference between diversification and randomization.

90% Risk of return

domain	riverswim			inventory			population		
	VaR	CVaR	EVaR	VaR	CVaR	EVaR	VaR	CVaR	EVaR
RASR	**50**	**50**	**50**	327	**319**	**310**	-623	**-1954**	**-3920**
Naive	**50**	**50**	**50**	325	317	**310**	**-566**	-2014	-4378
Erik	**50**	**50**	47	327	317	307	-1916	-4090	-5792
Derman	**50**	36	24	327	316	305	-625	-2082	-4364
RSVF	**50**	49	42	304	298	292	-2807	-4881	-6204
BCR	**50**	49	42	307	301	295	-2969	-4985	-6282
RSVI	**50**	49	41	306	300	294	-2646	-4702	-6104
Chow	**50**	46	34	**328**	**319**	307	-914	-2126	-4517

95% Risk of return

domain	riverswim			inventory			population		
	VaR	CVaR	EVaR	VaR	CVaR	EVaR	VaR	CVaR	EVaR
RASR	**50**	**50**	**50**	320	312	**305**	-1531	**-2948**	**-4735**
Naive	**50**	**50**	**50**	318	311	304	**-1525**	-3052	-5285
Erik	**50**	49	46	320	310	301	-3620	-5553	-6739
Derman	39	26	18	318	309	297	-1626	-3117	-5277
RSVF	**50**	48	40	272	268	263	-4950	-6465	-7292
BCR	**50**	48	40	302	296	291	-4640	-6258	-7177
RSVI	**50**	48	40	301	296	291	-4314	-6042	-7000
Chow	**50**	33	29	**321**	**313**	301	-2305	-3428	-5557

99% Risk of return

domain	riverswim			inventory			population		
	VaR	CVaR	EVaR	VaR	CVaR	EVaR	VaR	CVaR	EVaR
RASR	**50**	**50**	**50**	307	**301**	295	-4059	**-5349**	**-6387**
Naive	**50**	**50**	**50**	306	300	295	-6397	-7534	-8127
Erik	**50**	46	45	306	300	**296**	-6978	-7956	-8474
Derman	17	11	9	303	294	282	**-3976**	-5450	-7197
RSVF	**50**	46	45	266	262	258	-7465	-8262	-8722
BCR	45	43	36	293	288	284	-7400	-8212	-8650
RSVI	45	43	36	291	285	281	-7215	-8087	-8560
Chow	30	26	23	**308**	300	289	-6131	-6822	-7489

Table 5: Risk of Return for 10,000 Monte Carlo instances

Name / author	Horizon	Uncertainty	Risk Measure Epistemic	Risk Measure Aleatory	Complexity
RASR-ERM	Discounted ∞	Dynamic	ERM	ERM	P
RASR-EVaR	Discounted ∞	Dynamic	EVaR	EVaR	P
Iyengar et al. [40, 44]	Discounted ∞	Dynamic	Min	E	P
Xu et al. [37, 60, 61]	Discounted ∞	Dynamic	CVaR	E	NP-Hard
Eriksson et al. [28]	Discounted ∞	Dynamic	ERM	E	–
Delage et al. [6, 22, 56]	Discounted ∞	Static	VaR	E	NP-Hard
Lobo et al. [2, 14, 41, 43]	Discounted ∞	Static	CVaR	E	NP-Hard
Derman et al. [26]	Average ∞	Dynamic	E	E	P
Steimle et al. [15, 58]	Finite	Static	E	E	NP-Hard
Chen et al. [17]	Finite	Static	CVaR	E	NP-Hard
Chow et al. [20]	Discounted ∞	–	–	CVaR	NP-Hard
Osogami et al. [49]	Discounted ∞	–	–	I-CVaR/I-ERM	P
Borkar et al. [11]	Average ∞	–	–	ERM	

Table 6: Summary of the soft-robust and risk-averse models in the MDP/RL literature.

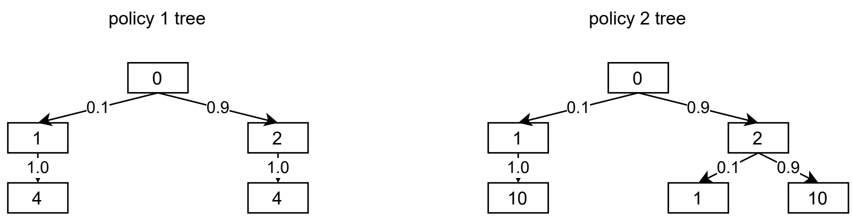

Figure 4: Example policy trees.

G Additional Related Work

Table 6 summarizes soft-robust and risk-averse results studied in the MDP/RL literature, together with the properties of their proposed formulations and algorithms. Other than the two RASR results presented in this paper: RASR-ERM and RASR-EVaR, we used the name of a representative author to refer to all results in each category.

The description of the rest of the columns is as follows: "horizon" indicates the considered MDP setting; "uncertainty" shows whether the uncertainty is static or dynamic as discussed in Section 3; "risk measure" contains the risk measure used by the work for epistemic and aleatory uncertainties (with E being the expectation or risk-neutral), and finally, "complexity" indicates the complexity of the proposed algorithm(s), if known. Algorithms 1 and 2 are marked as "P" because they can compute an ϵ-optimal policy in polynomial time for any fixed $\epsilon > 0$, $\gamma < 1$, $r_{\min}$, and $r_{\max}$ as shown in Theorems 3.5 and 4.3.

Theorem 3.4 shows that there exists an optimal deterministic policy for RASR-ERM. It may sound counter-intuitive because ERM is a convex risk measure, and the convexity axiom says that diversification reduces-risk/increases-profit. Action randomization is useful under adversarial settings and exploration, but portfolio diversification benefits from mitigating risk via negatively correlated assets. In this section, we provide an example as support to show that action randomization differs from portfolio diversification.

In Figure 3, given initial state s_0 the agent have two option for (actions/assets) a_1, a_2, which provide a randomize reward $R \sim N(2, 1)$ that is distributed normally with mean of 2 and standard deviation of 1, $r(a_1)$ and $r(a_2)$ are perfectly inverse correlated. In portfolio diversification, agent can simultaneously own multiple assets, a_1, a_2 are consider as assets. The delta neutral portfolio consist of 50% of each asset a_1, a_2 which results in a reward distribution of $\hat{R} \sim N(2, 0)$. However in action randomization, at each instance only one action is selected. Therefore, regarding the distribution of action selection $\pi(a_1|s_0), \pi(a_2|s_0)$ the agent receives a reward distribution $\hat{R} \sim N(2, 1)$. The example above explains the idea of diversification differs from randomization, thus does not contradict with optimal risk averse policy being deterministic in uncertain non-adversarial domain.

Theorem 3.1 shows that ERM is positive quasi-homogeneous, the risk level has to be discounted every time step. Here, we provide two policy trees with discount factor $\gamma = 0.9$ and initial risk-averse parameter $\alpha_0 = 1$ as an example to show the suboptimality of ERM bellman operator without discounting the risk (Naive Bellman). Figure 4 shows two policy trees both with only two time-horizon. Policy 1 has a deterministic reward at the second horizon, therefore both bellman operators yield a value of 4.90 for policy 1. However, for policy 2 the Exact Bellman (11) operator yields a value of 5.01 while the Naive Bellman operator yields a value of 4.78. Note that the Exact Bellman operator will prefer policy 2 over 1 while the Naive Bellman operator will prefer policy 1 over 2. The unchanged risk-averse parameter α of the Naive Bellman operator causes it to behave more pessimistically compared to the Exact Bellman operator. It is possible to use a smaller α_0 to negate the pessimism of the Naive Bellman operator, but the selection of α_0 to negate the pessimism in the Naive ERM is generally unclear because of the inaccurate approximate of the naive value function. For example, if we use $\alpha_0 = 0.9$ for the Naive Bellman operator, then 4.91 and 5.02 will be the value referring to policies 1 and 2 respectively which provide the same preference to the Exact Bellman operator with $\alpha_0 = 1$ in this example.