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. □

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(s) ← v'(s), ∀s ∈ S ;
for t = T − 1, . . . , 0 do
    Update v*_t using (11) and π*_t using (12);
return v*, π* ;
```

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 $c \geq 0$. □

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. □

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. □

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. □

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 $t = 0, \ldots$:

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

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:

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

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 $v^{\hat{\pi}}_t$ for $t = 0, \ldots, T'-1$ and all $s \in \mathcal{S}$ as:

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

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 $v^{\hat{\pi}}_t$ and at the same time be a value such that $\hat{\pi}$ is greedy to it.

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

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

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

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

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 $s \in \mathcal{S}$:

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

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:

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

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

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

□

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 $\alpha = \alpha^\star$. $\square$

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

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

$\square$

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 $\alpha_0 \ge \alpha^\star \ge \alpha_K$, and $h(\hat{\alpha}) \ge h(\alpha_k)$ for $k = 0, \ldots, K$ and some $\alpha_0 \ge \cdots \ge \alpha_K$. Then*

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

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

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

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

By the monotonicity property of ERM we get

$$\begin{aligned} \mathrm{ERM}^{\alpha_l}[X^\pi_{\alpha_{l+1}}] &\le \mathrm{ERM}^{\alpha_l}[X^\pi_{\alpha_l}] \le \mathrm{ERM}^{\alpha^\star}[X^\pi_{\alpha_l}] \\ &\le \mathrm{ERM}^{\alpha^\star}[X^\pi_{\alpha^\star}] \le \mathrm{ERM}^{\alpha_{l+1}}[X^\pi_{\alpha^\star}] \le \mathrm{ERM}^{\alpha_{l+1}}[X^\pi_{\alpha_{l+1}}] \end{aligned}$$

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

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

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

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

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

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

A3. (a) Sub-Additivity: $\psi[X_1 + X_2] \leq \psi[X_1] + \psi[X_2], \ \ \forall X_1, X_2 \in \mathbb{X}.$
(b) 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: $\psi[cX] = c\psi[X], \ \ \forall c \in \mathbb{R}_+, \ \forall X \in \mathbb{X}.$

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: $\psi\big[cX_1 + (1-c)X_2\big] \leq c\psi[X_1] + (1-c)\psi[X_2], \ \ \forall c \in [0,1], \ \forall X_1, X_2 \in \mathbb{X}.$
(b) Concavity: $\psi\big[cX_1 + (1-c)X_2\big] \geq c\psi[X_1] + (1-c)\psi[X_2], \ \ \forall c \in [0,1], \ \forall X_1, X_2 \in \mathbb{X}.$

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

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.,

$$\mathrm{EVaR}^{\beta}[X] \leq \mathrm{CVaR}^{\beta}[X] \leq \mathrm{VaR}^{\beta}[X].$$

2. The following inequality also holds for the EVaR:

$$\operatorname{ess\,inf}[X] \leq \mathrm{EVaR}^{\beta}[X] \leq \mathbb{E}[X].$$

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 $c \geq 0$, 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

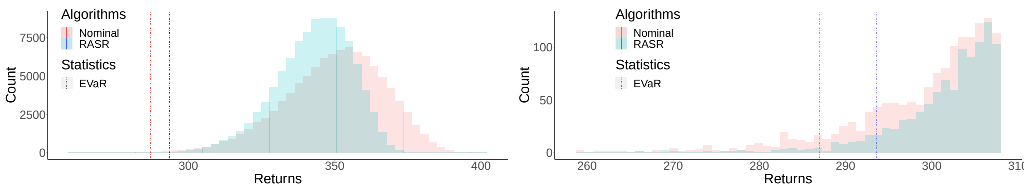

Figure 2: Full (left) and tail (right) histogram of return $\mathfrak{R}^{\pi}_{\infty}$ in the inventory domain.

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.

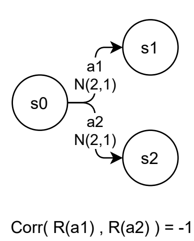

Figure 3: Example used to illustrate the difference between diversification and randomization.

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	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.

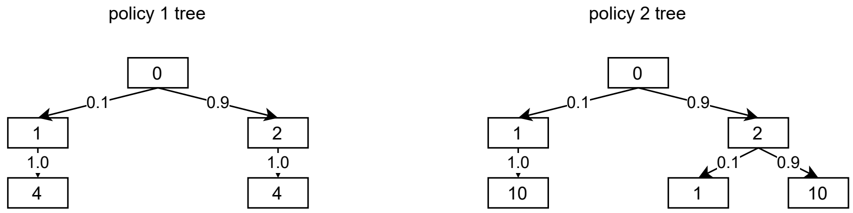

Figure 4: Example policy trees.

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.