

531 A Proofs of Section 2

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

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

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

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

536 which proves the desired result. \square

537 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 $(v_t^*)_{t=0}^T$ and policy $(\pi_t^*)_{t=0}^{T-1}$

- 1 Initialize $v_T^*(s) \leftarrow v'(s)$, $\forall s \in \mathcal{S}$;
 - 2 **for** $t = T - 1, \dots, 0$ **do**
 - 3 Update v_t^* using (11) and π_t^* using (12);
 - 4 **return** v^*, π^* ;
-

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

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

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

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

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

545 which is the definition of the value function. The equality (a) follows from the statement of the
 546 theorem, the equality (b) follows from the inductive assumption, and the equality marked by (c)
 547 follows by the translation equivariance of ERM (see Appendix E). The result readily generalizes to the
 548 infinite-horizon by considering the limit with $T \rightarrow \infty$ and using the fact that \mathfrak{R}_∞^π is bounded when
 549 $\gamma < 1$. The dynamic program representation for any fixed policy π follows analogously, replacing
 550 the maximization by a fixed policy. \square

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

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

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

$$\begin{aligned} \text{ERM}^\alpha[\text{ERM}^\alpha[\mathfrak{R}_T^\pi \mid P]] &= u^{-1}(\mathbb{E}[u(u^{-1}(\mathbb{E}[u(\mathfrak{R}_T^\pi) \mid P]))]) \\ &= u^{-1}(\mathbb{E}[\mathbb{E}[u(\mathfrak{R}_T^\pi) \mid P]]) \\ &\stackrel{(a)}{=} u^{-1}(\mathbb{E}[u(\mathfrak{R}_T^\pi) \mid \bar{P}]) \\ &= \text{ERM}^\alpha[\mathfrak{R}_T^\pi \mid \bar{P}] \end{aligned}$$

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

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

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

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

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

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

568 **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
 569 any risk-level $\alpha > 0$, we have $\mathbb{E}[X] - \epsilon(\alpha) \leq \text{ERM}^\alpha[X] \leq \mathbb{E}[X]$, where

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

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

571 *Proof of Lemma B.1.* To simplify the notation, let $X = \mathfrak{R}_T^\pi$ for any policy π which is bounded
 572 between x_{\min} and x_{\max} . We begin the our proof with the Hoeffding's lemma [13, 45]

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

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

$$\begin{aligned} \log(\mathbb{E}[e^{-\alpha X}]) &\leq -\alpha \cdot \mathbb{E}[x] + \frac{\alpha^2 \cdot (x_{\max} - x_{\min})^2}{8} \\ -\frac{1}{\alpha} \log(\mathbb{E}[e^{-\alpha X}]) &\geq \mathbb{E}[x] - \frac{\alpha(x_{\max} - x_{\min})^2}{8} \\ \mathbb{E}[x] - \frac{\alpha(x_{\max} - x_{\min})^2}{8} &\leq \text{ERM}^\alpha[X]. \end{aligned}$$

574 Now we have that $\mathbb{E}[X] - \epsilon(\alpha) \leq \text{ERM}^\alpha[X]$ where $\epsilon(\alpha) = 8^{-1}\alpha(x_{\max} - x_{\min})^2$, and $\text{ERM}^\alpha[X] \leq$
575 $\mathbb{E}[X]$ for $\alpha > 0$ is shown in Lemma D.1. Furthermore this upper bound vanishes as alpha decreases
576 to zero: $\lim_{\alpha \rightarrow 0} 8^{-1}\alpha(x_{\max} - x_{\min})^2 = 0$. \square

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

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

$$0 \leq v_\pi^\infty - v_t^\pi \leq \frac{\alpha \cdot \gamma^t \cdot (\Delta r)^2}{8 \cdot (1 - \gamma)^2}. \quad (15)$$

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

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

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

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

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

589 where $S'_{t+1, a}$ denotes the random variable that represents the state that follows s at time $t + 1$ after
590 taking an action a . The last equality holds from the definition of $\hat{\pi}_t$ being greedy with respect to
591 u_t ; subtracting a constant from all states does not change the greedy policy. The function u_t is
592 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 $v_{T'}^{\pi}(s) \geq u_{T'}(s)$ for all $s \in \mathcal{S}$. Then, assuming $v_{t+1}^{\pi}(s) \geq u_{t+1}(s)$ for all $s \in \mathcal{S}$, we can use backward induction on t to show that

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

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

$$v_t^*(s) - u_t(s) \leq \gamma^{T'-t} \cdot \frac{\alpha \cdot \gamma^{T'} \cdot (\Delta r)^2}{8 \cdot (1 - \gamma)^2}. \quad (16)$$

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 $s \in \mathcal{S}$:

$$\begin{aligned}
v_t^*(s) - u_t(s) &\stackrel{(a)}{=} \text{ERM}^{t \cdot \gamma^t} \left[r(s, \pi_t^*(s)) + \gamma \cdot v_{t+1}^*(S'_{t+1, \pi_t^*(s)}) \right] - \\
&\quad - \text{ERM}^{t \cdot \gamma^t} \left[r(s, \hat{\pi}_t(s)) + \gamma \cdot u_{t+1}(S'_{t+1, \hat{\pi}_t(s)}) \right] \\
&\stackrel{(b)}{=} \text{ERM}^{t \cdot \gamma^t} \left[r(s, \pi_t^*(s)) + \gamma \cdot v_{t+1}^*(S'_{t+1, \pi_t^*(s)}) \right] - \\
&\quad - \text{ERM}^{t \cdot \gamma^t} \left[r(s, \pi_t^*(s)) + \gamma \cdot u_{t+1}(S'_{t+1, \pi_t^*(s)}) \right] \\
&\stackrel{(c)}{=} \text{ERM}^{t \cdot \gamma^t} \left[\gamma \cdot v_{t+1}^*(S'_{t+1, \pi_t^*(s)}) \right] - \text{ERM}^{t \cdot \gamma^t} \left[\gamma \cdot u_{t+1}(S'_{t+1, \pi_t^*(s)}) \right] \\
&\stackrel{(d)}{=} \gamma \cdot \left(\text{ERM}^{t \cdot \gamma^{t+1}} \left[v_{t+1}^*(S'_{t+1, \pi_t^*(s)}) \right] - \text{ERM}^{t \cdot \gamma^{t+1}} \left[u_{t+1}(S'_{t+1, \pi_t^*(s)}) \right] \right) \quad (17)
\end{aligned}$$

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:

$$\begin{aligned}
v_{t+1}^{\pi^*}(s) - u_{t+1}(s) &\leq \gamma^{T'-t-1} \cdot \frac{\alpha \cdot \gamma^{T'} \cdot (\Delta r)^2}{8 \cdot (1 - \gamma)^2} \quad \forall s \in \mathcal{S} \\
\text{ERM}^{t \cdot \gamma^{t+1}}[v_{t+1}^{\pi^*}(S)] - \text{ERM}^{t \cdot \gamma^{t+1}}[u_{t+1}(S)] &\leq \gamma^{T'-t-1} \cdot \frac{\alpha \cdot \gamma^{T'} \cdot (\Delta r)^2}{8 \cdot (1 - \gamma)^2} \\
\gamma \cdot (\text{ERM}^{t \cdot \gamma^{t+1}}[v_{t+1}^{\pi^*}(S)] - \text{ERM}^{t \cdot \gamma^{t+1}}[u_{t+1}(S)]) &\leq \gamma^{T'-t} \cdot \frac{\alpha \cdot \gamma^{T'} \cdot (\Delta r)^2}{8 \cdot (1 - \gamma)^2}.
\end{aligned}$$

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

The theorem then follows from the properties established above as

$$\text{ERM}^{\alpha} [\mathfrak{R}_{\infty}^{\pi^*} \mid \bar{P}] - \text{ERM}^{\alpha} [\mathfrak{R}_{\infty}^{\hat{\pi}} \mid \bar{P}] = v_0^*(s_0) - v_0^{\hat{\pi}}(s_0) \leq v_0^*(s_0) - u_0 \leq \frac{\alpha \cdot \gamma^{2 \cdot T'} \cdot (\Delta r)^2}{8 \cdot (1 - \gamma)^2}$$

609

□

610 C Proofs of Section 4

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

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

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

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

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

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

624 \square

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

627 **Lemma C.1.** Suppose that the supremum of (14) is attained at α^* such that $\alpha_0 \geq \alpha^* \geq \alpha_K$, and
 628 $h(\hat{\alpha}) \geq h(\alpha_k)$ for $k = 0, \dots, K$ and some $\alpha_0 \geq \dots \geq \alpha_K$. Then

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

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

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

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

632 By the monotonicity property of ERM we get

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

633 where X_{α}^π refers to the total discounted reward distribution deploying the optimal policy of ERM^α .
 634 On the other hand,

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

635 We can conclude that

$$\text{ERM}^{\alpha_l}[X_{\alpha_l}^\pi] + \frac{\log(1-\beta)}{\alpha_{l+1}} \leq \text{ERM}^{\alpha^*}[X_{\alpha^*}^\pi] + \frac{\log(1-\beta)}{\alpha^*} \leq \text{ERM}^{\alpha_{l+1}}[X_{\alpha_{l+1}}^\pi] + \frac{\log(1-\beta)}{\alpha_l}$$

636 Therefore,

$$\begin{aligned} \text{EVaR}^\beta(X) - \max_{\alpha \in \Lambda} \left\{ \text{ERM}^\alpha[X] + \frac{\log(1-\beta)}{\alpha} \right\} \\ \leq \text{ERM}^{\alpha^*}[X_{\alpha^*}^\pi] + \frac{\log(1-\beta)}{\alpha^*} - \max_{\alpha \in \{\alpha_{l+1}\}} \left\{ \text{ERM}^\alpha[X_{\alpha}^\pi] + \frac{\log(1-\beta)}{\alpha} \right\} \\ \leq \frac{\log(1-\beta)}{\alpha_l} - \frac{\log(1-\beta)}{\alpha_{l+1}} \\ = \log(1-\beta) \left(\frac{1}{\alpha_l} - \frac{1}{\alpha_{l+1}} \right) \end{aligned}$$

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

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

638 The last part of the theorem can be proved as follows. Given an arbitrary error tolerance δ , β and
639 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 $h(\alpha^*) - h(\hat{\alpha}) \leq \delta$.

640 Moreover for $\alpha^* > \alpha_0$, given α_0 and β the error $h(\alpha^*) - h(\hat{\alpha}) \leq -\frac{\log(1-\beta)}{\alpha_0}$. \square

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

$$\begin{aligned} h(\alpha^*) &\geq h(\alpha^* + \epsilon) \\ \text{ERM}^{\alpha^*}[X] + \frac{\log(1-\beta)}{\alpha^*} &\geq \text{ERM}^{\alpha^* + \epsilon}[X] + \frac{\log(1-\beta)}{\alpha^* + \epsilon} \\ \text{ERM}^{\alpha^*}[X] - \text{ERM}^{\alpha^* + \epsilon}[X] &\geq \frac{\log(1-\beta)}{\alpha^* + \epsilon} - \frac{\log(1-\beta)}{\alpha^*} \\ \frac{(\Delta r)^2}{8(1-\gamma)^2} &\geq \frac{d(\text{ERM}^{\alpha^*}[X])}{d\alpha^*} \geq \log(1-\beta) \frac{d(\alpha^*)^{-1}}{d\alpha^*} \\ \frac{(\Delta r)^2}{8(1-\gamma)^2} &\geq -\log(1-\beta)(\alpha^*)^{-2} \\ (\alpha^*)^2 &\geq -\log(1-\beta) \frac{8(1-\gamma)^2}{(\Delta r)^2} \\ \alpha^* &\geq \sqrt{-8\log(1-\beta)} \frac{(1-\gamma)}{(\Delta r)} \end{aligned}$$

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

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

647 We conclude the proof with Lemma C.1 since $\alpha_0 \geq \alpha^* \geq \alpha_K$. \square

648 D Technical Lemmas

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

$$\max_{a \in \mathcal{A}} g(a) \geq \max_{\pi \in \Delta^\Omega} \text{ERM}_{A \sim \pi}^\alpha [g(A)] .$$

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

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

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

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

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

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

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

656 **Corollary D.2.** *Given an arbitrary error tolerance δ , β and α_k we construct α_{k+1} as $\alpha_{k+1} =$
 657 $(\frac{1}{\alpha_k} - \frac{\delta}{\log(1-\beta)})^{-1}$ such that $\alpha_k \geq \alpha_{k+1} > 0$ and*

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

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

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

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

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

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

661 and conclude that

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

662 \square

E Risk Measures

Consider a probability space (Ω, \mathcal{F}, P) . Let $\mathbb{X} : \Omega \rightarrow \mathbb{R}$ 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 $\psi : \mathbb{X} \rightarrow \mathbb{R}$ 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]:

$$\text{A1. Monotonicity:} \quad X_1 \leq X_2 \text{ (a.s.)} \implies \psi[X_1] \leq \psi[X_2], \quad \forall X_1, X_2 \in \mathbb{X}.$$

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

$$\begin{aligned} \text{A3. (a) Sub-Additivity:} \quad & \psi[X_1 + X_2] \leq \psi[X_1] + \psi[X_2], \quad \forall X_1, X_2 \in \mathbb{X}. \\ \text{(b) Super-Additivity:} \quad & \psi[X_1 + X_2] \geq \psi[X_1] + \psi[X_2], \quad \forall X_1, X_2 \in \mathbb{X}. \end{aligned}$$

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

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:

$$\begin{aligned} \text{A5. (a) Convexity:} \quad & \psi[cX_1 + (1-c)X_2] \leq c\psi[X_1] + (1-c)\psi[X_2], \quad \forall c \in [0, 1], \forall X_1, X_2 \in \mathbb{X}. \\ \text{(b) Concavity:} \quad & \psi[cX_1 + (1-c)X_2] \geq c\psi[X_1] + (1-c)\psi[X_2], \quad \forall c \in [0, 1], \forall X_1, X_2 \in \mathbb{X}. \end{aligned}$$

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 $\text{VaR}^\beta[X]$, is the $(1 - \beta)$ -quantile of X , i.e.,

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

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 $\text{CVaR}^\beta[X]$, 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:

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

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

698 E.3 Entropic Risk Measure

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

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

701 Properties of ERM:

- 702 1. It is easy to see that $\lim_{\alpha \rightarrow 0} \text{ERM}^\alpha[X] = \mathbb{E}[X]$ and $\lim_{\alpha \rightarrow \infty} \text{ERM}^\alpha[X] = \text{ess inf}[X]$.
- 703 2. For any random variable $X \in \mathbb{X}$, we have $\text{ERM}^\alpha[X] = \mathbb{E}[X] - \frac{\alpha}{2} \text{VaR}[X] + o(\alpha)$.
- 704 3. If X is a Gaussian random variable, we have $\text{ERM}^\alpha[X] = \mathbb{E}[X] - \frac{\alpha}{2} \text{VaR}[X]$.
- 705 4. For any two random variables $X_1, X_2 \in \mathbb{X}$, we have $\text{ERM}^\alpha[X_2|X_1] =$
706 $-\frac{1}{\alpha} \log(\mathbb{E}[e^{-\alpha X_2}|X_1])$.
- 707 5. Since ERM does not satisfy the axiom A4 (positive homogeneity), we have $\text{ERM}^\alpha[cX] \neq$
708 $c \text{ERM}^\alpha[X]$.

709 E.4 Entropic Value-at-Risk

710 For a random variable $X \in \mathbb{X}$, its entropic value-at-risk with confidence level β , denoted by
711 $\text{EVaR}^\beta[X]$, is defined as

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

712 Properties of EVaR:

- 713 1. The EVaR with confidence level β is the tightest possible lower-bound that can be obtained
714 from the Chernoff inequality for VaR and CVaR with confidence level β , i.e.,

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

- 715 2. The following inequality also holds for the EVaR:

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

- 716 3. It is easy to see that $\text{EVaR}^0[X] = \mathbb{E}[X]$ and $\lim_{\beta \rightarrow 1} \text{EVaR}^\beta[X] = \text{ess inf}[X]$.

717 E.5 Properties of Risk Measures

718 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.

719 A *law-invariant* (LI) risk measure depends only on the total return and not on the particular sequence
720 of individual rewards [30]. A *dynamically-consistent* (DC), or time-consistent, risk measure satisfies
721 the tower property [57] and can be optimized using a dynamic program [4, 21, 24, 27, 34, 53, 55].
722 Finally, a positively-homogeneous (PH) risk measure satisfies $\psi(c \cdot X) = c \cdot \psi(X)$, for any $c \geq 0$,
723 which is an important property in the risk-averse parameter selection and discounted setting [3, 30, 31].
724 Unfortunately, expectation ($\mathbb{E}[\cdot]$) and minimum (Min) are the only convex risk measures that satisfy
725 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

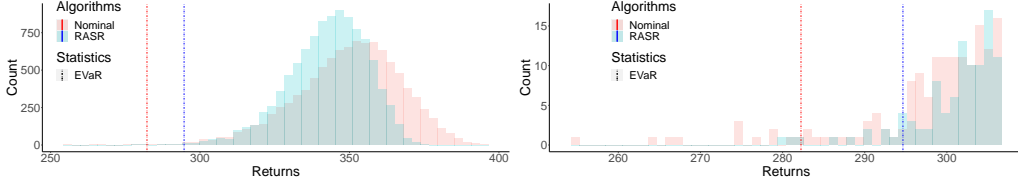


Figure 2: Full (left) and tail (right) histogram of return \mathcal{R}_∞^π in the inventory domain.

F Additional Experimental Results and Details

Figure 2 compares the distribution of returns \mathcal{R}_∞^π 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 $\alpha_{0:(T'+1)} = \{e^{10}, e^{10}\gamma, e^{10}\gamma^2, \dots, e^{10}\gamma^{T'}, 0\} = \Lambda$ 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 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 both CVaR and EVaR for all domains.

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.

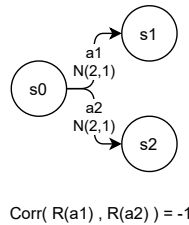


Figure 3: Example used to illustrate the difference between diversification and randomization.

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		Complexity
			Epistemic	Aleatory	
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.

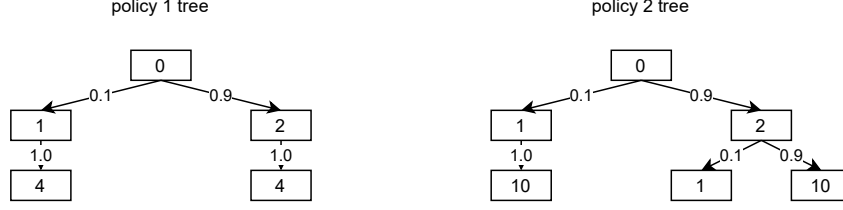


Figure 4: Example policy trees.

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

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

762 In Figure 3, given initial state s_0 the agent have two option for (actions/assets) a_1, a_2 , which
763 provide a randomize reward $R \sim N(2, 1)$ that is distributed normally with mean of 2 and standard
764 deviation of 1, $r(a_1)$ and $r(a_2)$ are perfectly inverse correlated. In portfolio diversification, agent can
765 simultaneously own multiple assets, a_1, a_2 are consider as assets. The delta neutral portfolio consist
766 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
767 randomization, at each instance only one action is selected. Therefore, regarding the distribution
768 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
769 example above explains the idea of diversification differs from randomization, thus does not contradict
770 with optimal risk averse policy being deterministic in uncertain non-adversarial domain.

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