Entropic Risk Optimization in Discounted MDPs

Mohammad Ghavamzadeh

Amazon

joint work with Jia Lin Hau & Marek Petrik at UNH
Erick Delage at HEC Montreal

January 23, 2025

Outline

Outline

Outline

Markov Decision Process

MDP: $\langle \mathcal{S}, \mathcal{A}, r, p, s_0, \gamma \rangle$

- $\mathcal{S}$ and $\mathcal{A}$: state and action spaces with cardinality S and A
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function *($\Delta_r$: range of reward)*
- $p : \mathcal{S} \times \mathcal{A} \to \Delta^S$: transition probability (dynamics)
- $s_0 \in \mathcal{S}$: initial state
- $\gamma \in (0, 1]$: discount factor

$T \in \mathbb{N}^+ \cup \{\infty\}$: fixed horizon of control

- $T < \infty$: finite-horizon MDP *(usually $\gamma = 1$)*
- $T = \infty$: infinite-horizon MDP $\gamma \in (0, 1)$

Markov Decision Process

Policy:

- $\pi = \{\pi_t\}_{t=0}^{T-1}$, $\pi_t : \mathcal{S} \to \Delta^A$: Markovian randomized policy Π_{MR}
- π_t's are all equal: stationary randomized policy Π_{SR}
- Π_{MD} and Π_{SD}: their deterministic counterparts

Return: random variable (RV) of the return of a policy π after T steps

$$\mathfrak{R}_T^\pi := \mathfrak{R}_{0:T}^\pi(s_0)$$

$$\mathfrak{R}_{t:T}^\pi(s) = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \cdot \overbrace{r(S_\tau, A_\tau)}^{R_\tau^\pi} \mid S_t = s$$

Risk-neutral MDP

Objective: maximize the ***expectation*** of the return RV $\mathfrak{R}_T^\pi$

$$\max_\pi \mathbb{E}\left[\mathfrak{R}_T^\pi\right]$$

- optimal policy in *finite horizon* setting is in Π_{MD}
- optimal policy in *infinite horizon discounted* setting is in Π_{SD}

Risk-neutral MDP

Value Function: $v^\pi = (v_t^\pi)_{t=0}^{T-1}$

$$v_t^\pi(s) = \mathbb{E}\left[\mathfrak{R}_{t:T}^\pi(s)\right], \qquad A \sim \pi(\cdot|s), \quad S' \sim p(\cdot|s,A), \quad v_T^\pi(s) = 0$$

value function of π: $v_t^\pi(s) = \mathbb{E}\left[r(s,A) + \gamma \cdot v_{t+1}^\pi(S')\right]$

optimal value function: $v_t^\star(s) = \max_{a\in\mathcal{A}} \ \mathbb{E}\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right]$

Outline

Risk Measure

Risk Measure: $\psi : \mathbb{X} \to \mathbb{R}$ $\mathbb{X}$ is the space of random variables (RVs)

Coherent Risk Measure:

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

Convex Risk Measure: satisfies A1 and A2 — replaces A3 and A4 with

A5. (a) Convexity
(b) Concavity $\psi\big[cX_1 + (1-c)X_2\big] \geq c\psi[X_1] + (1-c)\psi[X_2], \ \ \forall c \in [0, 1]$

- every coherent risk measure is convex but not the other way around

Popular Risk Measures

Value-at-Risk: with confidence level α

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

$(1-\alpha)$-quantile of X ***or*** the worst $(1-\alpha)$-fraction of X

Popular Risk Measures

Value-at-Risk: with confidence level α

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

$(1-\alpha)$-quantile of X ***or*** the worst $(1-\alpha)$-fraction of X

Conditional Value-at-Risk: expectation of the worst $(1-\alpha)$-fraction of X

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

- CVaR_α is a ***coherent*** risk measure
- $\mathrm{CVaR}_0[X] = \mathbb{E}[X]$
- $\lim_{\alpha \to 1} \mathrm{CVaR}_\alpha[X] = \operatorname{ess\,inf}[X] = \sup_{\zeta \in \mathbb{R}} \left\{ \mathbb{P}(X < \zeta) = 0 \right\}$

Popular Risk Measures

Entropic Risk Measure: with risk parameter $\beta > 0$

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

Properties of ERM:

1. $\lim_{\beta\to 0}\mathrm{ERM}_\beta[X] = \mathbb{E}[X]$ $\lim_{\beta\to\infty}\mathrm{ERM}_\beta[X] = \operatorname{ess\,inf}[X]$
2. $\mathrm{ERM}_\beta[X] = \mathbb{E}[X] - \frac{\beta}{2}\mathrm{var}[X] + o(\beta)$ for Gaussian $\mathrm{ERM}_\beta[X] = \mathbb{E}[X] - \frac{\beta}{2}\mathrm{var}[X]$
3. $\mathrm{ERM}_\beta[cX] \neq c\,\mathrm{ERM}_\beta[X]$ (*not **coherent** but **concave***)
4. $\mathrm{ERM}_\beta[X_1] = \mathrm{ERM}_\beta\left[\mathrm{ERM}_\beta[X_1 \mid X_2]\right]$ (***Tower Property***)

Popular Risk Measures

Entropic Value-at-Risk: with confidence level $\alpha \in [0,1)$

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

Properties of EVaR:

1. $\mathrm{EVaR}_0[X] = \mathbb{E}[X]$ $\lim_{\alpha\to 1}\mathrm{EVaR}_\alpha[X] = \operatorname{ess\,inf}[X]$
2. $\mathrm{EVaR}_\alpha[X] \le \mathrm{CVaR}_\alpha[X] \le \mathrm{VaR}_\alpha[X]$ *(tightest possible lower-bound for CVaR)*
3. a ***coherent*** risk measure

Risk-averse MDP

Objective: maximize the ***risk measure*** ψ of the return RV $\mathfrak{R}_T^\pi$

$$\max_\pi \psi\left[\mathfrak{R}_T^\pi\right]$$

- replacing $\mathbb{E}[\cdot]$ in *risk-neutral* with risk measure $\psi[\cdot]$
- risk measure ψ is applied to the *aleatory* uncertainty over $\mathfrak{R}_T^\pi$

Nested Risk Measures

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

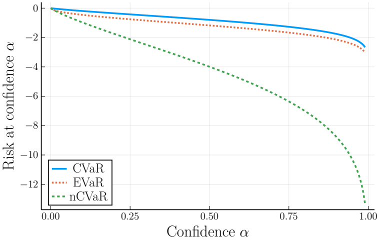
$$v_t^\star(s) = \max_{a \in \mathcal{A}} \mathrm{CVaR}_\alpha\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right]$$

Nested Risk Measures

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

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

- (+) favorable computational properties
- (−) poor approximation of static risk measures



Properties of Concave Risk Measures

Risk measure	Tower Property	Positive Homogeneity
$\mathbb{E}$, Min	✓	✓
ERM	✓	✗
CVaR	✗	✓
EVaR	✗	✓
Nested CVaR	✓	✓

Outline

Discounted MDP with ERM Objective

ERM-MDP

Objective: maximize

$$\max_{\pi \in \Pi_{HR}} \mathrm{ERM}_\beta \left[\mathfrak{R}_T^\pi \right] \tag{1}$$

optimal policy $\pi^\star = (\pi^\star)_{t=0}^{T-1}$

ERM-MDP

Objective: maximize

$$\max_{\pi \in \Pi_{HR}} \mathrm{ERM}_\beta \left[\mathfrak{R}_T^\pi\right] \tag{1}$$

optimal policy $\pi^\star = (\pi^\star)_{t=0}^{T-1}$

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

ERM is ***not*** positively homogeneous $\mathrm{ERM}_\beta[c \cdot X] \neq c \cdot \mathrm{ERM}_\beta[X]$

Outline

Preliminaries

MDP with ERM Objective

Value Function, DP Formulation, Policy Class

Algorithms for ERM-MDP *(finite & infinite horizon)*

MDP with EVaR Objective

Numerical Evaluation

Risk-Averse Soft-Robust (RASR) MDP

Value Function for ERM-MDP

Theorem *(Positive Quasi-homogeneity)*

Let $X \in \mathbb{X}$ be a random variable. Then, for any constant $c \geq 0$, we have

$$\mathrm{ERM}_\beta[c \cdot X] = c \cdot \mathrm{ERM}_{\beta . c}[X]$$

Value Function: for a policy π is the collection $v^\pi = (v_t^\pi)_{t=0}^T, \;\; v_t^\pi : \mathcal{S} \to \mathbb{R}$

$$\begin{aligned} v_t^\pi(s) &= \mathrm{ERM}_{\beta . \gamma^t}\left[\sum_{t'=t}^{T} \gamma^{t'-t} \cdot R_{t'}^\pi \mid S_t = s\right] \\ &= \mathrm{ERM}_{\beta . \gamma^t}\left[\mathfrak{R}_{t:T}^\pi(s)\right], \qquad \forall s \in \mathcal{S} \end{aligned} \tag{2}$$

$$v_T^\pi(s) = 0, \quad \forall s \in \mathcal{S} \qquad , \qquad v_0^\pi(s_0) = \mathrm{ERM}_\beta[\mathfrak{R}_T^\pi]$$

Note: $\lim_{t\to\infty} \mathrm{ERM}_{\beta . \gamma^t}[\cdot] = \mathbb{E}[\cdot]$

Optimal Value Function for ERM-MDP

Optimal Value Function: $v^\star = (v_t^\star)_{t=0}^T$ VF of an optimal policy $\pi^\star$

$$v^\star = v^{\pi^\star}$$

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

Bellman Equations for ERM-MDP

Theorem *(Bellman Equations)*

For any policy $\pi \in \Pi_{MR}$, its value function $v^\pi = (v_t^\pi)_{t=0}^T$ defined in (2) is the unique solution to the following system of equations

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

where $A \sim \pi_t(\cdot|s)$, $S' \sim p(\cdot|s,A)$, and $v_T^\pi(s) = 0$.

Moreover, the optimal value function $v^\star = (v_t^\star)_{t=0}^T$ is the unique solution to

$$v_t^\star(s) = \max_{a \in \mathcal{A}} \mathrm{ERM}_{\beta \cdot \gamma^t}\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right] \;. \tag{3}$$

Optimal Policy of ERM-MDP

ERM-MDP has a Markovian ***deterministic*** optimal policy

Theorem ***(Optimal Policy)***

There exists a ***deterministic*** time-dependent optimal policy $\pi^\star = (\pi_t^\star)_{t=0}^{T-1} \in \Pi_{MD}$ for (1), which is greedy w.r.t. the optimal value function $v^\star$ in (3), i.e.,

$$\pi_t^\star(s) \in \arg\max_{a\in\mathcal{A}} \; \mathrm{ERM}_{\beta.\gamma^t}\left[r(s,a) + \gamma \cdot v_{t+1}^\star(S')\right], \quad \forall s \in \mathcal{S}, \;\; S' \sim p(\cdot|s,a) \; . \quad (4)$$

Outline

Preliminaries

MDP with ERM Objective

- Value Function, DP Formulation, Policy Class
- Algorithms for ERM-MDP *(finite & infinite horizon)*

MDP with EVaR Objective

Numerical Evaluation

Risk-Averse Soft-Robust (RASR) MDP

Optimizing ERM-MDP ***(finite horizon)***

Algorithm ***(VI for finite-horizon ERM-MDP)***

Input: Horizon $T < \infty$, risk level $\beta > 0$, terminal value $v_T(s)$, $\forall s \in \mathcal{S}$
Output: Optimal value $(v_t^\star)_{t=0}^{T}$ and policy $(\pi_t^\star)_{t=0}^{T-1}$

Initialize $v_T^\star(s) \leftarrow v'(s)$, $\forall s \in \mathcal{S}$

for $t = T-1, \ldots, 0$ **do**
  Update $v_t^\star$ using (3) and $\pi_t^\star$ using (4)

Return $v^\star, \pi^\star$

Optimizing ERM-MDP *(infinite horizon)*

Algorithm *(VI for infinite-horizon ERM-MDP)*

Input: Planning horizon $T' < \infty$, risk level $\beta > 0$
Output: policy $\hat{\pi}^\star = (\hat{\pi}^\star_t)_{t=0}^{\infty}$ and value $\hat{v}^\star = (\hat{v}^\star_t)_{t=0}^{\infty}$

Compute $(v^\star_\infty, \pi^\star_\infty)$ solution to the risk-neutral ∞-horizon discounted MDP

Compute $(\tilde{v}^\star_t)_{t=0}^{T'}$ and $(\tilde{\pi}^\star_t)_{t=0}^{T'-1}$ using (3) and (4) with horizon T' and $\tilde{v}^\star_{T'} = v^\star_\infty$

Construct a policy $\hat{\pi}^\star = (\hat{\pi}^\star_t)_{t=0}^{\infty}$, where $\hat{\pi}^\star_t = \begin{cases} \tilde{\pi}^\star_t & \text{if } t < T', \\ \pi^\star_\infty & \text{otherwise .} \end{cases}$

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

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

$\lim_{t\to\infty} \mathrm{ERM}_{\beta\cdot\gamma^t}[\cdot] = \mathbb{E}[\cdot]$ *(risk-neutral)*

Optimizing ERM-MDP ***(infinite horizon)***

Theorem ***(sub-optimality of $\hat{\pi}^\star$)***

The performance loss of the policy $\hat{\pi}^\star$ is bounded as

$$\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] \leq c \cdot \gamma^{2T'},$$

where $\pi^\star$ is the optimal ERM-MDP policy and $c = \beta \cdot \Delta_r^2 \,/\, 8\,(1-\gamma)^2$.

truncating at horizon T' and following with an arbitrary policy thereafter, the performance loss decreases proportionally to $\gamma^{T'}$

Optimizing ERM-MDP *(infinite horizon)*

Theorem *(sub-optimality of $\hat{\pi}^\star$)*

The performance loss of the policy $\hat{\pi}^\star$ is bounded as

$$\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] \le c \cdot \gamma^{2T'},$$

where $\pi^\star$ is the optimal ERM-MDP policy and $c = \beta \cdot \Delta_r^2 \;/\; 8\,(1-\gamma)^2$.

truncating at horizon T' and following with an arbitrary policy thereafter, the performance loss decreases proportionally to $\gamma^{T'}$

Algorithm runs in $O(S^2 A \log(1/\delta))$ time to compute a δ-optimal policy

Outline

Discounted MDP with EVaR Objective

EVaR-MDP

Objective: maximize

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha[\mathfrak{R}_T^\pi]$$

EVaR Properties

(+) coherent

(+) interpretable $\mathrm{EVaR}_\alpha[X] \leq \mathrm{CVaR}_\alpha[X] \leq \mathrm{VaR}_\alpha[X]$

(+) its confidence level α is readily comparable to that of VaR and CVaR

(−) **no** Tower Property

Outline

Reducing EVaR-MDP to ERM-MDP

EVaR: with confidence level $\alpha \in [0, 1)$

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

EVaR-MDP Objective: maximize

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha[\mathfrak{R}_T^\pi] = \sup_{\beta > 0} \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_T^\pi] + \frac{\log(1-\alpha)}{\beta} \right) \tag{5}$$

Reducing EVaR-MDP to ERM-MDP

EVaR: with confidence level $\alpha \in [0, 1)$

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

EVaR-MDP Objective: maximize

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha[\mathfrak{R}_T^\pi] = \sup_{\beta > 0} \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_T^\pi] + \frac{\log(1-\alpha)}{\beta} \right) \qquad (5)$$

Theorem

Let $\pi^\star$ be an optimal solution to EVaR-MDP (5). Then, there exists a risk-level $\beta^\star$ such that $\pi^\star$ is optimal for ERM-MDP with $\beta = \beta^\star$.

Reducing EVaR-MDP to ERM-MDP

EVaR: with confidence level $\alpha \in [0, 1)$

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

EVaR-MDP Objective: maximize

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha[\mathfrak{R}_T^\pi] = \sup_{\beta>0} \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_T^\pi] + \frac{\log(1-\alpha)}{\beta} \right) \tag{5}$$

Theorem

Let $\pi^\star$ be an optimal solution to EVaR-MDP (5). Then, there exists a risk-level $\beta^\star$ such that $\pi^\star$ is optimal for ERM-MDP with $\beta = \beta^\star$.

Corollary

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

Outline

Optimizing EVaR-MDP

Objective: maximize

$$\max_{\pi \in \Pi_{MR}} \mathrm{EVaR}_\alpha[\mathfrak{R}_T^\pi] = \sup_{\beta > 0} \max_{\pi \in \Pi_{MR}} \left(\mathrm{ERM}_\beta[\mathfrak{R}_T^\pi] + \frac{\log(1-\alpha)}{\beta} \right)$$

- Unlike the EVaR objective, the one with max over π is not ***concave***

Algorithm

Input: Discretized risk-levels $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_K > 0$
Output: EVaR-MDP optimal policy $\hat{\pi}^\star$

for $k = 1, \ldots, K$ **do**
Compute $v^{(k)}$ and $\pi^{(k)}$ by solving ERM-MDP at risk-level β_k

Let $k^\star \leftarrow \arg\max_{k=1:K} \; v_0^{(k)}(s_0) + \frac{\log(1-\alpha)}{\beta_k}$

Return $\hat{\pi}^\star = \pi^{(k^\star)}$

Optimizing EVaR-MDP

Theorem *(sub-optimality of $\hat{\pi}^\star$)*

Given the error tolerance $\delta > 0$ and the discretization

$$\beta_1 = \frac{8\delta(1-\gamma)^2}{\Delta_r^2}, \qquad \beta_{k+1} = \beta_k \cdot \frac{\log(1-\alpha)}{\beta_k\delta + \log(1-\alpha)}, \qquad \beta_K \geq \frac{-\log(1-\alpha)}{\delta},$$

algorithm runs in $\mathcal{O}\left(S^2 A \left(\frac{\log(1/\delta)}{\delta}\right)^2\right)$ time and returns a policy $\hat{\pi}^\star$, whose performance loss is bounded as

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

- solving ERM-MDP computes VF for multiple risk levels $\beta, \beta\gamma, \beta\gamma^2, \ldots$
- a branch-and-bound algorithm

Outline

Experimental Setup

Setup

- Optimization Criterion: $\mathrm{EVaR}_{0.9}[\mathfrak{R}^{\pi}_{100}], \quad \alpha = 0.9, \quad T = 100$
- Our Algorithm: EVaR-MDP Algo
- Evaluation: 100,000 episodes of $\mathfrak{R}^{\pi}_{T}$

Ablation Study

- **naive grid:** uniform grid over β_k such that $\beta_1 = 0$ and $\beta_K = 10$
- **naive level:** optimized grid – doesn't adjust risk-level with time in ERM-MDP

Baselines Computing Markov Policies

- **risk-neutral MDP**
- **nested** $\mathrm{CVaR}_{0.9}$ ***(Bauerle and Glauner, 2022)***
- **nested** $\mathrm{EVaR}_{0.9}$ ***(Ahmadi et al. 2021)***
- $\mathrm{ERM}_{0.5}$

Baseline Computing History-dependent Policies

- **augmented** $\mathrm{CVaR}_{0.9}$ ***(Chow et al., 2015)***

EVaR Results

$\mathrm{EVaR}_{0.9}[\mathfrak{R}_T^\pi]$ for π returned by each method

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

bold: *results within a 95% confidence interval of the best policy*

- machine replacement **(MR)** ***(Delage and Mannor, 2010)***
- gamblers ruin **(GR)** ***(Bauerle and Ott, 2011; Li et al., 2022)***
- two classic inventory management problems **(INV1)** and **(INV2)** ***(Ho et al., 2021)***
- river-swim **(RS)** ***(Strehl and Littman, 2008)***

CVaR Results

$\mathrm{CVaR}_{0.9}[\mathfrak{R}_T^\pi]$ for π returned by each method

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

bold: *results within a 95% confidence interval of the best policy*

Run-Time Results

Run-time for the algorithms in second

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

Summary

- **ERM-MDP:**

  first exact DP formulation for ERM in discounted MDPs

  showed optimal VF exists – optimal policy is time-dependent and deterministic

  proposed a VI algorithm *(exact in finite-horizon – approximate in infinite-horizon)*

- **EVaR-MDP:**
  - *(approximately)* optimized it by reducing it to multiple ERM-MDPs
  - empirical results highlighting the utility of our EVaR-MDP algorithm

Outline

Risk-averse MDP

Objective: maximize the ***risk measure*** ψ of the return RV $\mathfrak{R}_T^\pi$

$$\max_\pi \ \psi \left[\mathfrak{R}_T^\pi\right]$$

- replacing $\mathbb{E}[\cdot]$ in *risk-neutral* with risk measure $\psi[\cdot]$
- risk measure ψ is applied to the *aleatory* uncertainty over $\mathfrak{R}_T^\pi$

Soft-robust MDP

Objective: maximize the ***risk measure*** ψ of the RV $\mathbb{E}\big[\mathfrak{R}_T^\pi \mid P\big]$

$$\max_\pi \ \psi_{P\sim f}\Big[\mathbb{E}\big[\mathfrak{R}_T^\pi \mid P\big]\Big]$$

optimization is

- ***risk-averse*** to the *epistemic* uncertainty in P (uses $\psi_{P\sim f}[\cdot]$)
- ***risk-neutral*** to the *aleatory* uncertainty in $\mathfrak{R}_T^\pi \mid P$ (uses $\mathbb{E}[\cdot]$)

dynamic vs. ***static*** model of uncertainty

- we use the ***dynamic*** model *(easier to optimize)* $\quad P = (P_t)_{t=0}^{T-1}, \ \ P_t \sim f_t$

Risk-Averse Soft-Robust (RASR) MDP

Objective: maximize two ***risk measures***

$$\max_{\pi} \psi_{P\sim f}\left[\psi\left[\mathfrak{R}_T^\pi \mid P\right]\right]$$

over the ***epistemic*** uncertainty in P and ***aleatory*** uncertainty in $\mathfrak{R}_T^\pi \mid P$

Risk-Averse Soft-Robust (RASR) MDP

Objective: maximize two ***risk measures***

$$\max_{\pi} \ \psi_{P\sim f}\left[\psi\left[\mathfrak{R}^{\pi}_{T} \mid P\right]\right]$$

over the ***epistemic*** uncertainty in P and ***aleatory*** uncertainty in $\mathfrak{R}^{\pi}_{T} \mid P$

RASR-ERM objective is equivalent to a risk-averse RL problem with $\bar{P}$

Corollary

For any policy $\pi \in \Pi_{MR}$, we have

$$\underbrace{\mathrm{ERM}_{\beta}\left[\mathrm{ERM}_{\beta}[\mathfrak{R}^{\pi}_{T} \mid P]\right]}_{\text{RASR-ERM objective}} = \overbrace{\mathrm{ERM}_{\beta}\left[\mathfrak{R}^{\pi}_{T} \mid \bar{P}\right]}^{\text{risk-averse RL}}.$$

Thank you!!

Mohammad Ghavamzadeh

ghavamza@amazon.com OR
mohammad.ghavamzadeh51@gmail.com