Entropic Risk Optimization in Discounted MDPs

Jia Lin Hau[1], Marek Petrik[1,2], Mohammad Ghavamzadeh[2]

[1]UNH, [2]Google

Summary

Motivation

- Risk avoidance is very important many domains, like health care, or autonomous driving.
- Stake holders seek policies that minimize risk while maximizing return.

Limitations of existing methods

- Compute complex history-dependent policies: difficult to deploy and analyze.
- Often lack practical optimality guarantees.
- Usually only optimize VaR and CVaR risk measures.

Our contributions

- New algorithms for optimizing entropic risk (EVaR and ERM) objectives in MDPs.
- History-independent policies are optimal in ERM/EVaR MDPs.
- Guarantee δ-optimal policy in poly-time, $\log(1/\delta)$ for ERM and $(\frac{\log(1/\delta)}{\delta})^2$ for EVaR.

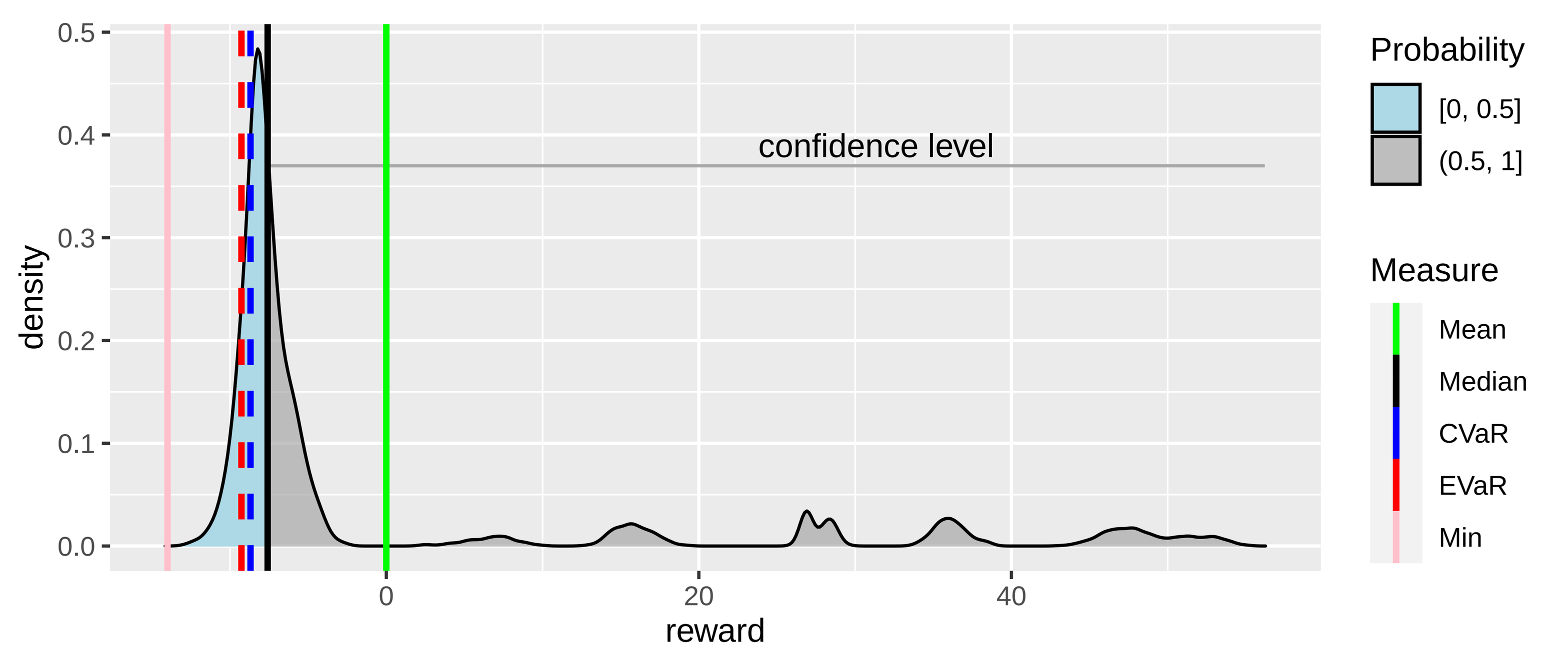
Risk Averse MDPs

- Maximizes the risk measure $\psi[\cdot]$ of the total γ-discounted reward in a Markov decision process (MDP) for finite and inifnite horizon T

$$\max_{\pi\in\Pi} \psi\left[\sum_{t=0}^{T} \gamma^t r^\pi(S_t, A_t, S_{t+1})\right] \quad = \quad \max_{\pi\in\Pi} \psi\left[\mathfrak{R}_T^\pi\right]$$

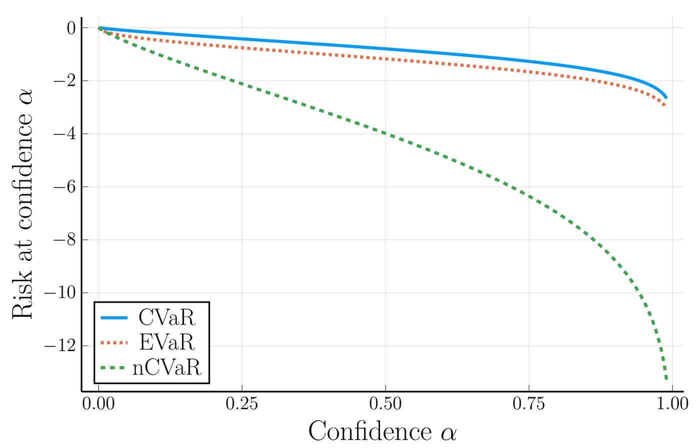
- Known rewards $r(s, a, s') \in \mathbb{R}$ and transition probabilities $P(s, a) \in \triangle^S$ and a tabular state and action spaces.

Risk Measures



- **Challenges**: Common risk measures, like VaR and CVaR, do not admit direct dynamic program representations. Nested risk measures, like nCVaR, are difficult to interpret and result in loose approximations.

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



- *Law invariant*: Identically distributed random variables have identical risk values.
- *Tower property*: Allows one to nest the risk measure: $\psi[X] = \psi[\psi[X \mid Y]]$.
- *Positively homogeneous*: The risk scale equals to the scale of the distribution.

Entropic Risk Measure (ERM-MDP)

Objective for a risk parameter $\beta \in (0, \infty)$ ($\mathrm{ERM}_0[X] = \mathbb{E}[X]$ and $\mathrm{ERM}_\infty[X] = \min X$):

$$\max_{\pi\in\Pi} \mathrm{ERM}_\beta\left[\mathfrak{R}_T^\pi\right] \quad = \quad \max_{\pi\in\Pi} -\beta^{-1} \cdot \log\left(\mathbb{E}\left[e^{-\beta\cdot\mathfrak{R}_T^\pi}\right]\right),$$

- Challenge: ERM struggles with discounting because it lacks positive homogeneity.
- Main idea: Use time-dependent risk level the Bellman equation.

- **Theorem 3.1: ERM is Positive Quasi homogeneous**:

$$\mathrm{ERM}_\beta\left[c\cdot\mathfrak{R}_T^\pi\right] \quad = \quad c\cdot\mathrm{ERM}_{c\cdot\beta}\left[\mathfrak{R}_T^\pi\right].$$

- **Theorem 3.2: Bellman equations for ERM-MDP**:

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

- Risk level $\beta_t = \beta\cdot\gamma^t$ decreases with time t and decisions become less risk-averse.
- **Theorem 3.4: Infinite horizon approximation error** / convergence rate (w.r.t) T':

$$\mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\pi^\star}\right] - \mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^{\hat{\pi}^\star}\right] \;\le\; \frac{\beta\cdot\gamma^{2T'}\cdot\Delta_{\mathfrak{R}}^2}{8}.$$

- Select $T'(\delta) = \lceil\frac{1}{2\log(\delta)}\log(\frac{8\delta}{\beta\Delta_{\mathfrak{R}}^2})\rceil$ for δ-optimal policy $\hat{\pi}^\star$

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

- Total run-time of our ERM MDP algorithm $O(S^2 A\log(1/\delta))$.
- Main limitation: Risk parameter $\beta \in \mathbb{R}_+$ is difficult to interpret

Entropic Value at Risk (EVaR-MDP)

Objective for risk level $\alpha \in [0, 1]$

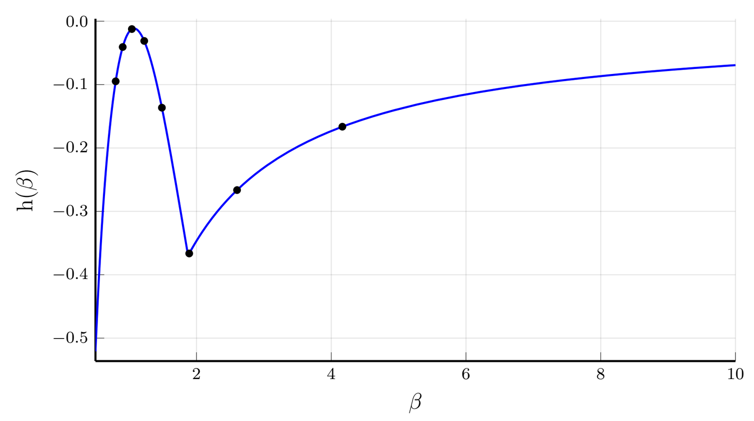
$$\max_{\pi\in\Pi} \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^\pi\right] = \max_{\pi\in\Pi} \inf_{\xi\ll f}\left\{\mathbb{E}_\xi[\mathfrak{R}_\infty^\pi] \mid \mathrm{KL}(\xi\|f) \le \log\left(\frac{1}{1-\alpha}\right)\right\}$$
$$= \sup_{\beta>0}\left(\max_{\pi\in\Pi} \mathrm{ERM}_\beta\left[\mathfrak{R}_\infty^\pi\right] + \frac{\log(1-\alpha)}{\beta}\right).$$

- Challenge: EVaR does not satisfy the tower property.
- Main idea: Reduce EVaR optimization to a sequence of ERM optimizations.

- **Theorem 4.1: Reduce EVaR-MDP to ERM-MDP**

$$\max_{\pi\in\Pi} \mathrm{EVaR}_\alpha\left[\mathfrak{R}_\infty^\pi\right] \quad = \quad \sup_{\beta>0} h(\beta).$$

- Function $h(\beta)$ is neither convex nor concave in β.



- **Theorem 4.3:** Our algorithm computes δ-optimal EVaR-MDP policy $\hat{\pi}^\star$ in $O(S^2 A(\frac{\log(1/\delta)}{\delta})^2)$ time when using a grid $B = \{\beta_k\}_{k=1}^K$ is constructed (for $K(\delta) \in O\left(\frac{\log(1/\delta)}{\delta^2}\right)$) as

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

Algorithms for ERM-MDP and EVaR-MDP

Algorithm 1: VI for ERM-MDP

Input: planning horizon $T' < \infty$, risk level $\beta > 0$

1. $\hat{v}_{T':\infty}^\star \leftarrow 0$ for finite horizon, otherwise $\hat{v}_{T':\infty}^\star \leftarrow \bar{v}^\star$ value function of standard infinite-horizon MDP
2. $\hat{v}_t^\star(s) \leftarrow \max_{a\in\mathcal{A}} \mathrm{ERM}_{\beta\cdot\gamma^t}\left[r(s, a) + \gamma\cdot\hat{v}_{t+1}^\star(S')\right]$, for $t \in \{T'-1, \cdots, 0\}$
3. Construct $\hat{\pi}^\star$ analogously to $\hat{v}^\star$

Output: policy $\hat{\pi}^\star$ and value function $\hat{v}^\star$

Algorithm 2: Algorithm for EVaR-MDP

Input: Desired error tolerance δ, confidence level $\alpha \in [0, 1]$

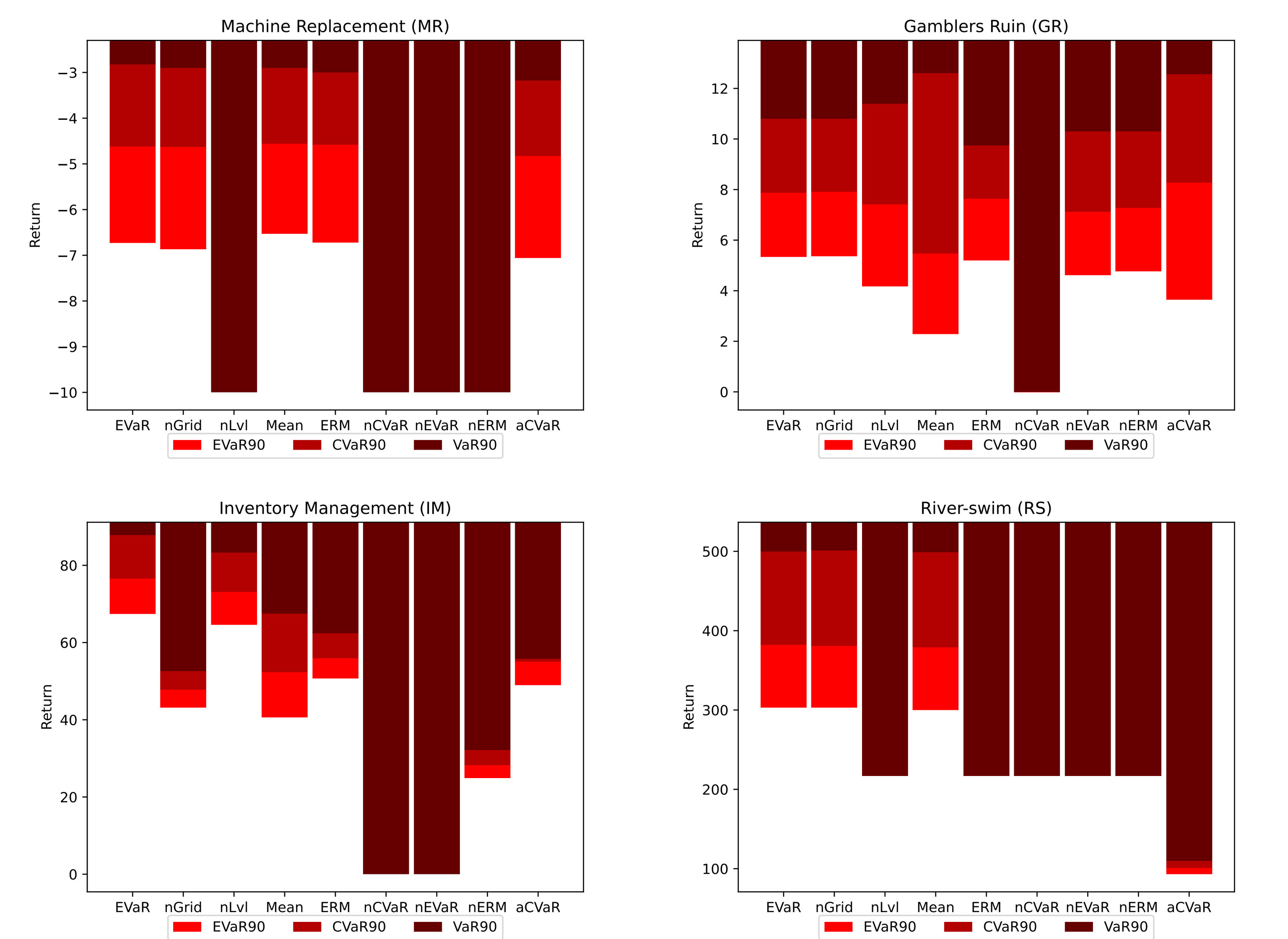
1. $T \leftarrow \lceil\frac{1}{2\log(\delta/2)}\log(\frac{4\delta}{\beta\Delta_{\mathfrak{R}}^2})\rceil$ for infinite horizon.
2. Let K be the smallest value that satisfies $\beta_K \ge \frac{-\log(1-\alpha)}{\delta/2}$.
3. $v^k, \pi^k \leftarrow \mathit{ErmVI}(T, \beta_k)$ for $k = 1, \cdots, K$
4. Let $k^\star \leftarrow \arg\max_{k=1:K} v_0^k(s_0) + \beta_k^{-1}\cdot\log(1-\alpha)$

Output: policy $\hat{\pi}^\star \leftarrow \pi^{k^\star}$ and value function $\hat{v}_0^\star \leftarrow v_0^{k^\star}(s_0) + \beta_{k^\star}^{-1}\cdot\log(1-\alpha)$

Simulation Results

Time horizon $T = 100$, number of episodes $= 100,000$, risk level: $\alpha = 0.9 = 90\%$ confidence.

Tail risk performance measured in VaR (dark red), CVaR (medium red), and EVaR (light red)



Higher (shorter) the better

- EVaR-MDP algorithms perform well across all domains for both CVaR and EVaR.
- Naive algorithms ("Naive grid" or "Naive level") exhibit inconsistent performance.
- Risk-neutral "$\mathbb{E}$" and "ERM" optimize different also exhibit inconsistent performance.
- Nested risk measures ("nCVaR", "nEVaR", "nERM") perform poorly across all domains.
- Quantile augmentation "Aug CVaR" is slow, computes history-dependent policies, and fails in larger domains.