Risk-averse Total-reward MDPs with ERM and EVaR

Xihong Su[1], Marek Petrik[1],Julien Grand-Clément[2]

[1]University of New Hampshire,[2]HEC Paris

Summary

Motivation

- Risk-averse objectives are important in critical applications
- The *total reward criterion* (TRC), also known as the *stochastic shortest path*, is a natural and popular objective
- When risk averse: discounted and total return objectives differ!

Limitations of existing methods

- The optimal policy of some common risk-averse discounted objectives must be history-dependent or at least time-dependent
- No existing algorithms for law-invariant coherent risk measures

Contributions

1. Stationary optimal policies for ERM-TRC
2. Stationary optimal policies for EVaR-TRC
3. Algorithms: value iteration, policy iteration, linear programming

Monetary Risk Measures

- *Nested Risk Measures* for Markov decision processes

$$r_0 + \mathrm{CVaR}_\alpha\left[\tilde{r}_1 + \mathrm{CVaR}_\alpha\left[\tilde{r}_2 + \mathrm{CVaR}_\alpha\left[\ldots\right] \mid \tilde{s}_1\right]\right]$$

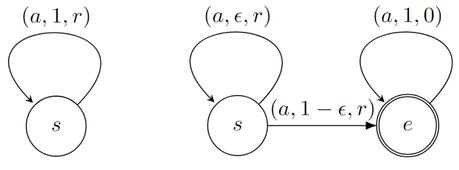
- *Entropic Risk Measure* (ERM) is a popular risk measure, with good dynamic properties, for a risk level $\beta > 0$,

$$\mathrm{ERM}_\beta\left[\tilde{x}\right] \;=\; -\beta^{-1} \cdot \log \mathbb{E}[\exp\left(-\beta \tilde{x}\right)]$$

- *Entropic Value at Risk* (EVaR), defined as, for a given risk level $\alpha \in (0, 1)$,

$$\mathrm{EVaR}_\alpha\left[\tilde{x}\right] \;=\; \sup_{\beta>0} -\beta^{-1} \log\left(\alpha^{-1}\mathbb{E}[\exp\left(-\beta\tilde{x}\right)]\right) \;=\; \sup_{\beta>0} \mathrm{ERM}_\beta\left[\tilde{x}\right] + \beta^{-1}\log\alpha$$

Total Reward Criterion



$\gamma-$discounted MDP Transient MDP

- **Transient MDPs** terminate in a sink state e eventually

$$\sum_{t=0}^{\infty} \mathbb{P}^{\pi,s}\left[\tilde{s}_t = s'\right] \;<\; \infty, \qquad \forall s, s' \in \mathcal{S} \setminus \{e\}, \pi \in \Pi_{HR}.$$

- **Objective**: no discounting of future rewards

$$\sup_{\pi\in\Pi_{HR}} \liminf_{t\to\infty} \mathbb{E}^{\pi,\mu}\left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right]$$

- Risk neutral: discount equivalent to probability of non-termination ($\gamma = \epsilon$)
- Expected return guaranteed to be bounded
- Nested CVaR, CVaR, ERM return unbounded for some ϵ

Contribution 1: ERM-TRC

$$\sup_{\pi\in\Pi_{HR}} \liminf_{t\to\infty} \mathrm{ERM}_\beta^{\pi,\mu}\left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right]$$

- Exponential value function $w^\infty(\pi, z)$

$$w_s^t(\pi) \;:=\; -\exp\left(-\beta \cdot v_s^t(\pi)\right)$$

- Exponential Bellman operator (also optimization) L^d

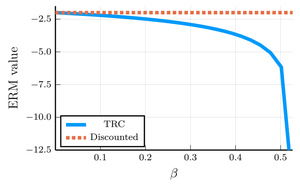
$$L^d w \;:=\; B^d w - b^d$$

- **Corollary**: There exists optimal ERM stationary policy
- Linear programming

$$\min\left\{\mathbf{1}^\mathsf{T} w \mid w \in \mathbb{R}^S, w \geq -b^a + B^a w, \forall a \in \mathcal{A}\right\}$$

Risk-averse Values

- $r = -0.2, \epsilon = 0.9$. β is risk level. ρ is spectral radius.
- **Lemma**: $w^\infty(\pi, z) > -\infty \quad \Leftrightarrow \quad \rho(B^d) < 1.$



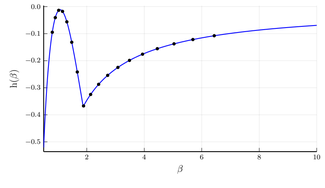
Contribution 2: EVaR-TRC

$$\sup_{\pi\in\Pi_{HR}} \liminf_{t\to\infty} \mathrm{EVaR}_\alpha^{\pi,\mu}\left[\sum_{k=0}^{t} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right]$$

$$\approx \max_{\beta>0} \max_{\pi\in\Pi_{HR}} \liminf_{t\to\infty} \underbrace{\left(\mathrm{ERM}_\beta^\pi\left[\sum_{k=0}^{t} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right] + \beta^{-1}\log\alpha\right)}_{h_t(\beta)}$$

- Idea: reduce the EVaR-TRC problem to a specific sequence of ERM-TRC problems
- **Theorem**: There exists a $\beta^\star$ for which the ERM optimal policy is also EVaR optimal
- **Corollary**: There exists an optimal EVaR stationary policy

Computing EVaR Policy



Summary of Results

	Risk properties		Optimal policy	
Risk measure	Coherent	Law invariant	Discounted	TRC
$\mathbb{E}$	✓	✓	S	S
EVaR	✓	✓	M	S
ERM	✗	✓	M	S
NCVaR	✓	✗	S	S
VaR	✓	✓	H	H
CVaR	✓	✓	H	H

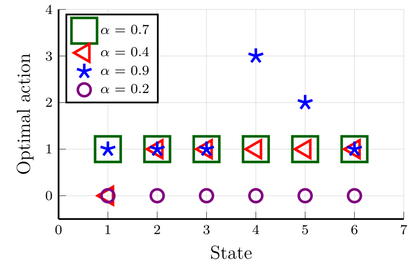
S: stationary policy M: Markov policy H: history-dependent policy

Parameters of Gambler's ruin

- Initial capital sampled from a uniform distribution$[1, 7]$
- Winning probability 0.68; losing probability 0.32
- Reward $r \in \{-1, 1, 2, 3, 4, 5, 6, 7\}$
- Game ends when the gambler either loses all capital or achieves the cap 7

Optimal EVaR Policy

- State i: how much capital the gambler holds
- Action 0 : quit the game and keep the current wealth
- Action $i, i > 0$: how much capital the gambler bets



Distribution of the Final Capital