Risk-averse Total-reward Reinforcement Learning

Xihong Su[1], Jia Lin Hau[1], Gersi Doko[1], Kishan Panaganti[2,3], Marek Petrik[1]

[1]University of New Hampshire,[2]Tencent AI Lab, [3]California Institute of Technology

Summary

Motivation

- Risk-averse objectives are important in critical applications
- The *total reward criterion* (TRC) 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
- Assume that the dynamic model is known

Contributions

- Stationary optimal policies for ERM-TRC and EVaR-TRC
- Algorithms: ERM-TRC Q-learning algorithm, EVaR-TRC Q-learning algorithm

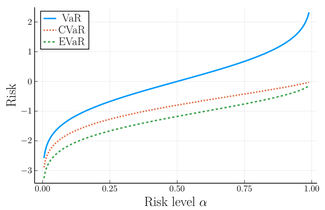
Monetary Risk Measures

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

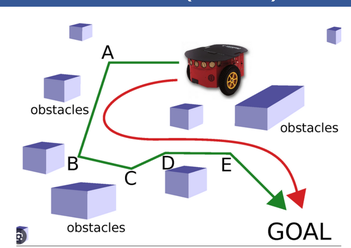
$$\mathrm{ERM}_\beta[\tilde{x}] = -\beta^{-1} \cdot \log \mathbb{E}[\exp(-\beta\tilde{x})]$$

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

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



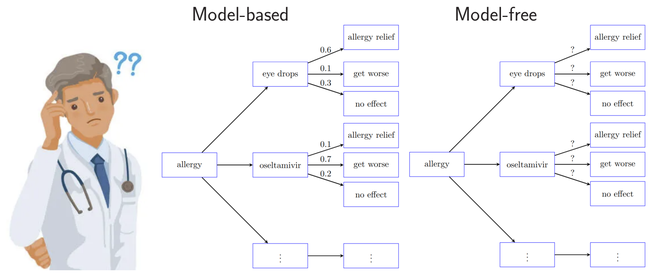
Total Reward Criterion (TRC)



- Capture the concept of stochastic termination
- Does not discount future rewards
- **Objective**:

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

Scenarios of Previous Work and This Work



ERM Bellman Operators and Fixed Point

- ERM Bellman operator

$$(B_\beta q)(s,a) := \mathrm{ERM}^{a,s}_\beta\left[r(s,a,\tilde{s}_1) + \max_{a'\in\mathcal{A}} q(\tilde{s}_1, a', \beta)\right], \qquad \forall s\in\mathcal{S}, a\in\mathcal{A}, \beta\in\mathcal{B}, q\in\mathcal{Q}.$$

- **Theorem**: assume some $\beta\in\mathcal{B}$ and suppose that $q^\star_\beta(s,a,\beta) > -\infty, \forall s\in\mathcal{S}, a\in\mathcal{A}$. Then $q^\star_\beta$ is the unique solution to

$$q^\star_\beta = B_\beta q^\star_\beta, \quad \text{where} \quad q^\star_\beta = q^\star(\cdot,\cdot,\beta).$$

- Use the elicitability property of ERM to define the Bellman operator

$$(\hat{B}_\beta q)(s,a) := \arg\min_{y\in\mathbb{R}} \mathbb{E}^{a,s}\left[\ell_\beta\left(r(s,a,\tilde{s}_1) + \max_{a'\in\mathcal{A}} q(\tilde{s}_1,a',\beta) - y\right)\right], \ \forall s\in\mathcal{S}, a\in\mathcal{A}.$$

- Equivalence of two Bellman operators: for each $\beta > 0$ and $q\in\mathcal{Q}$, we have that

$$B_\beta q = \hat{B}_\beta q.$$

ERM-TRC Q-learning Algorithm

- Standard Q-learning algorithm

$$\tilde{q}_{i+1}(\tilde{s}_i,\tilde{a}_i) = \tilde{q}_i(\tilde{s}_i,\tilde{a}_i) - \tilde{\eta}_i\tilde{z}_i, \quad \tilde{z}_i = r(\tilde{s}_i,\tilde{a}_i,\tilde{s}'_i) + \max_{a'\in\mathcal{A}} \tilde{q}_i(\tilde{s}'_i,a') - \tilde{q}_i(\tilde{s}_i,\tilde{a}_i).$$

- ERM-TRC Q-learning algorithm

Algorithm 1: ERM-TRC Q-learning algorithm

Input: Risk levels $\mathcal{B}\subseteq\mathbb{R}_{++}$, samples: $(\tilde{s}_i,\tilde{a}_i,\tilde{s}'_i)$, step sizes $\tilde{\eta}_i, i\in\mathbb{N}$, bounds $z_{\min}, z_{\max}$

Output: Estimate state-action value function $\tilde{q}_i$

$\tilde{q}_0(s,a,\beta) \leftarrow 0, \quad \forall s\in\mathcal{S}, a\in\mathcal{A}$;

for $i\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}, \beta\in\mathcal{B}$ **do**

  if $s=\tilde{s}_i \wedge a=\tilde{a}_i$ **then**

    $\tilde{z}_i(\beta) \leftarrow r(s,a,\tilde{s}'_i) + \max_{a'\in\mathcal{A}} \tilde{q}_i(\tilde{s}'_i,a',\beta) - \tilde{q}_i(s,a,\beta)$;

    if $\neg(z_{\min} \le \tilde{z}_i(\beta) \le z_{\max})$ **then return** $\tilde{q}_{i+1}(s,a,\beta) = -\infty$;

    $\tilde{q}_{i+1}(s,a,\beta) \leftarrow \tilde{q}_i(s,a,\beta) - \tilde{\eta}_i \cdot (\exp(-\beta\cdot\tilde{z}_i(\beta)) - 1)$;

  else $\tilde{q}_{i+1}(s,a,\beta) \leftarrow \tilde{q}_i(s,a,\beta)$;

- **Convergence theorem**: for $\beta\in\mathcal{B}$, assume that the sequence $(\tilde{\eta}_i)_{i=0}^\infty$ and $(\tilde{s}_i,\tilde{a}_i,\tilde{s}'_i)_{i=0}^\infty$ in the above algorithm satisfies Assumption 4.1 and step size condition

$$\sum_{i=0}^\infty \tilde{\eta}_i = \infty, \quad \sum_{i=0}^\infty \tilde{\eta}_i^2 < \infty,$$

where $i\in\{i\in\mathbb{N} \mid (\tilde{s}_i,\tilde{a}_i) = (s,a)\}$, if $\tilde{z}_i\in[z_{\min}, z_{\max}]$ almost surely, then the sequence $(\tilde{q}_i)_{i=0}^\infty$ produced by the above algorithm convergences almost surely to q_∞ such that $q_\infty = \hat{B}_\beta q_\infty$.

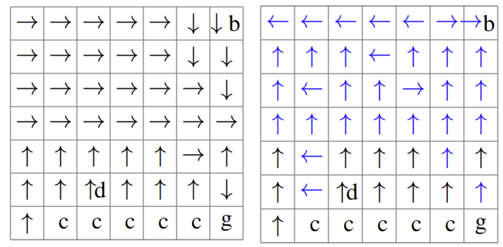
Previous Work and Risk-averse Q-learning

	Previous work [su2025]	Risk-averse Q-learning
Scenario	Model-based	Model-free
TRC	✓	✓
Objective	ERM, EVaR	ERM, EVaR
Model-free approach	✗	✓
Convergence proof of proposed algorithms	weighted-norm contraction	monotonicity, boundedness condition

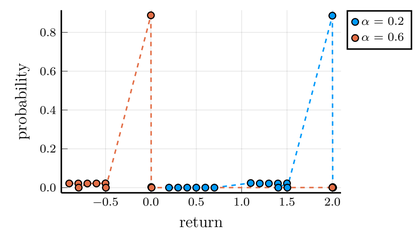
Optimal Policies and Simulated Returns

- EVaR risk level $\alpha = 0.2$ and $\alpha = 0.6$ on cliff walking domain

Optimal policy ($\alpha = 0.2$) Optimal policy ($\alpha = 0.6$)



Simulated returns



Convergence Result

- EVaR risk level $\alpha = 0.2$ on cliff walking domain