Accommodating Picky Customers
Regret Bound & Exploration Complexity for **Multi-Objective RL**

Jingfeng Wu, Vladimir Braverman, Lin Yang

Multi-Objective Reinforcement Learning

- State S
- Action A
- Horizon H
- Transition $\mathbb{P}$

Vector Reward

$\mathbf{r} : [H] \times S \times A \to [0,1]^d$

Preferences

$\{w \in [0,1]^d : \|w\|_1 = 1\}$

Scalarization

$$Q_h^\pi(x, a; w) := \mathbb{E}\langle w, \mathbf{r}_h(x_h, a_h)\rangle + \cdots + \langle w, \mathbf{r}_H(x_H, a_H)\rangle$$

$$V_h^\pi(x; w) := Q_h^\pi(x, \pi_h(x); w) \qquad V_1^*(x_1; w) = \max_\pi V_1^\pi(x_1; w)$$

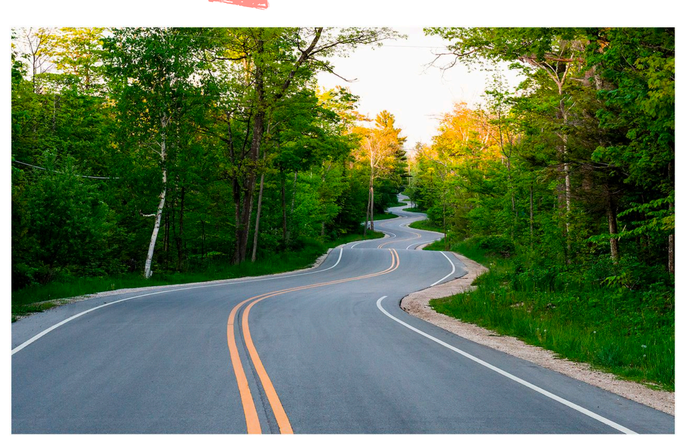
π^* depends on w

Multiple Objectives and Unknown Preferences

Faaaaster!

Action

State



Smooooother~

Reward:
$0.6 \times$ Fast
$0.4 \times$ Smooth

Multiple Objectives?
Unknown Preferences?

Online MORL

agent observes trajectory $\{(x_h, a_h, x_{h+1})\}_{h=1}^H$, collects reward $V_1^\pi(x_1; w)$

adversary chooses preference w

agent chooses policy π

$$\mathtt{regret}(K) := \sum_{k=1}^K V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k)$$

[MO-UCBVI]

$$\widehat{Q}_h(x, a; w) \leftarrow \langle w, \mathbf{r}_h(x, a)\rangle + \widehat{\mathbb{P}}\,\widehat{V}_{h+1}(x, a; w) + b(x, a)$$

UCB Bonus

- Model-based
- Optimistic estimation
- Planning based on preference

$$\frac{\#(x, a, y)}{\#(x, a)} \qquad \approx \sqrt{\frac{\min\{d, S\} H^2 \log}{\#(x, a)}}$$

can be improved to Bernstein version

[Upper Bound] For any $\{w^1, \ldots, w^K\}$ and with high prob., MO-UCBVI (Bernstein ver.) satisfies:

$$\mathtt{regret}(K) \le \mathcal{O}\left(\sqrt{\min\{d, S\} \cdot H^2 SAK \cdot \log}\right)$$

matching single-obj. RL when $d = 1$

[Lower Bound] For every MORL algorithm, there is a distribution of MOMDPs and a (necessarily adversarial) sequence $\{w^1, \ldots, w^K\}$ such that:

$$\mathbb{E}[\mathtt{regret}(K)] \ge \Omega\left(\sqrt{\min\{d, S\} \cdot H^2 SAK}\right)$$

tight up to log factors

MORL is statistically harder than single-objective RL

Preference-Free Exploration

agent collects K trajectories

agent computes PAC-policy π

$$\mathbb{P}\{\forall w, V_1^*(x_1; w) - V_1^\pi(x_1; w) \le \epsilon\} \ge 1 - \delta$$

Reward-Free Exploration [Jin et al. 2020]

$d = SA$

Task-Agnostic Exploration [Zhou et al, 2020]

$d = 1$

$w \in \mathbb{R}^d$

[Exploration] MO-UCBVI (Hoeffding ver.) with 0 reward

[Planning] Typical UCBVI with input preference/reward

[Upper Bound] For our algorithm to be (ϵ, δ)-PAC, it suffices to have

$$K = \mathcal{O}\left(\min\{d, S\} \cdot H^3 SA \cdot \log / \epsilon^2\right)$$

nearly tight except for H

[Lower Bound] There is a distribution of MOMDPs such that for every $(\epsilon, \delta = 0.1)$-PAC algorithm, there is a (necessarily adversarial) w such that:

$$\mathbb{E}[K] \ge \Omega\left(\min\{d, S\} \cdot H^2 SA / \epsilon^2\right)$$

$\min\{d, S\}$ *vs.* S*:*
exploration is easier when rewards are structured

Numerical Simulations