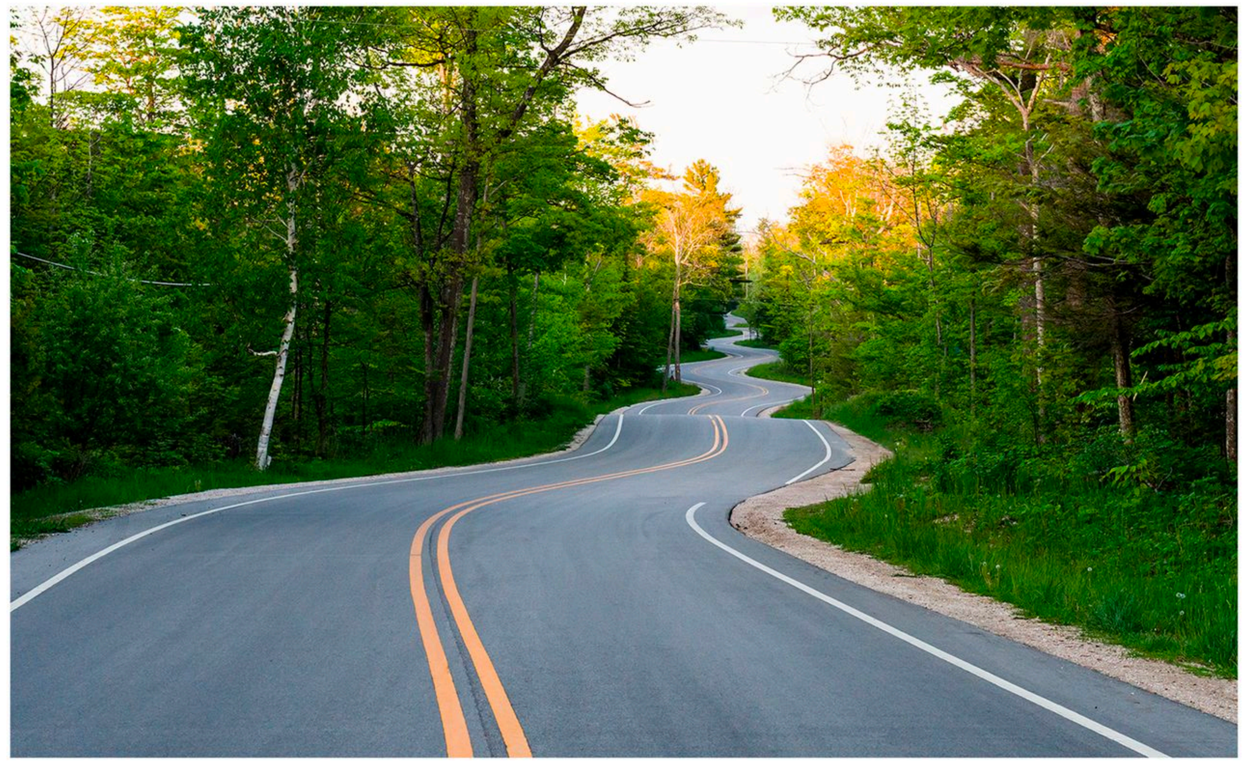
Accommodating Picky Customers

Regret Bound and Exploration Complexity for Multi-Objective RL

Jingfeng Wu, Vladimir Braverman, Lin Yang

Single-Objective vs. Multiple-Objective RL

Action



State

Reward:

$0.6 \times$ *Fast*

$0.4 \times$ *Smooth*

Faaaaster!

Smoooother~

Multiple Objectives?

Unknown Preferences?

Problem Setup

State S

Action A

Horizon H

Transition $\mathbb{P}$

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

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

$$V_h^\pi(x; w) := Q_h^\pi(x, \pi_h(x); w)$$

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

Scalarization

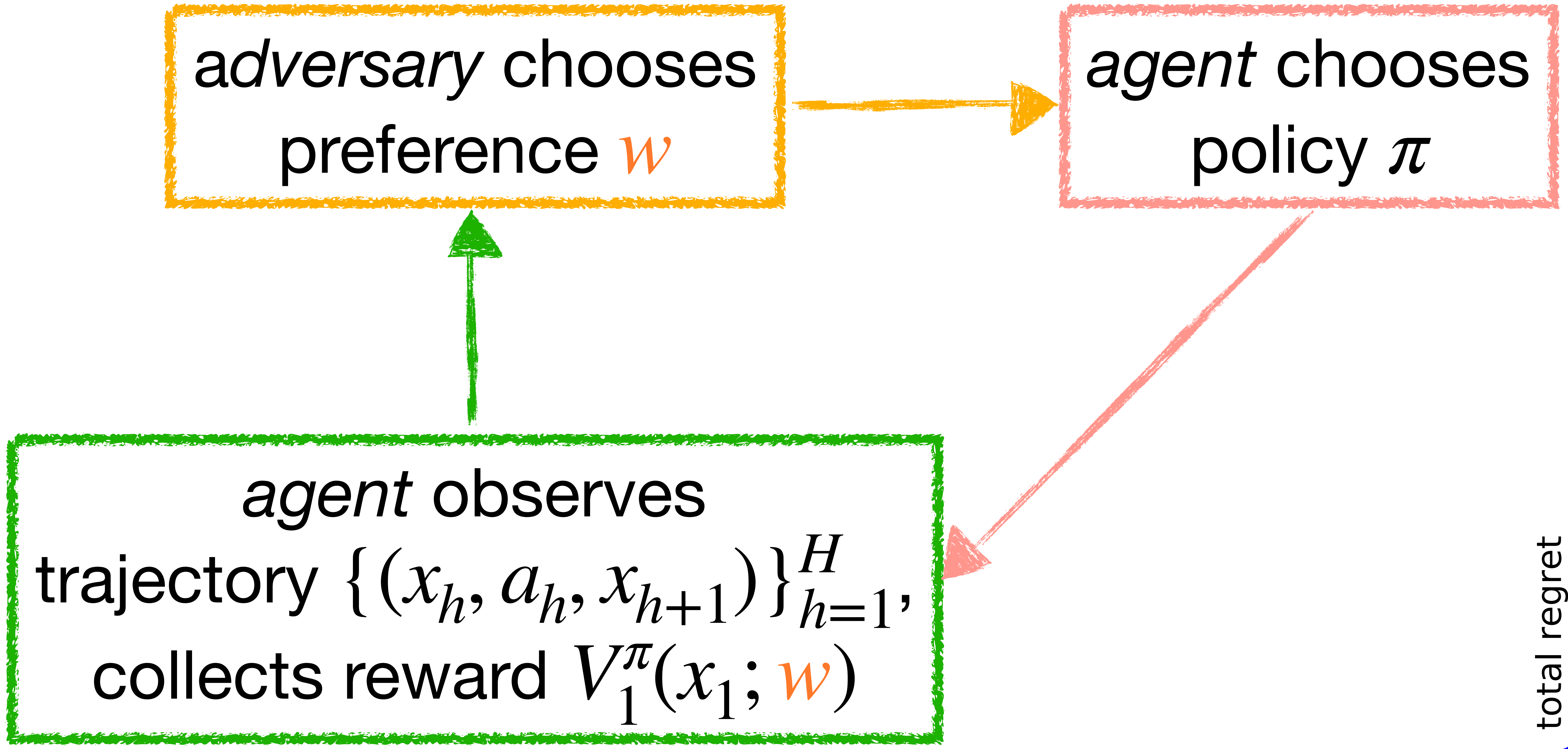
$$V_1^*(x_1; w) = \max_\pi V_1^\pi(x_1; w)$$

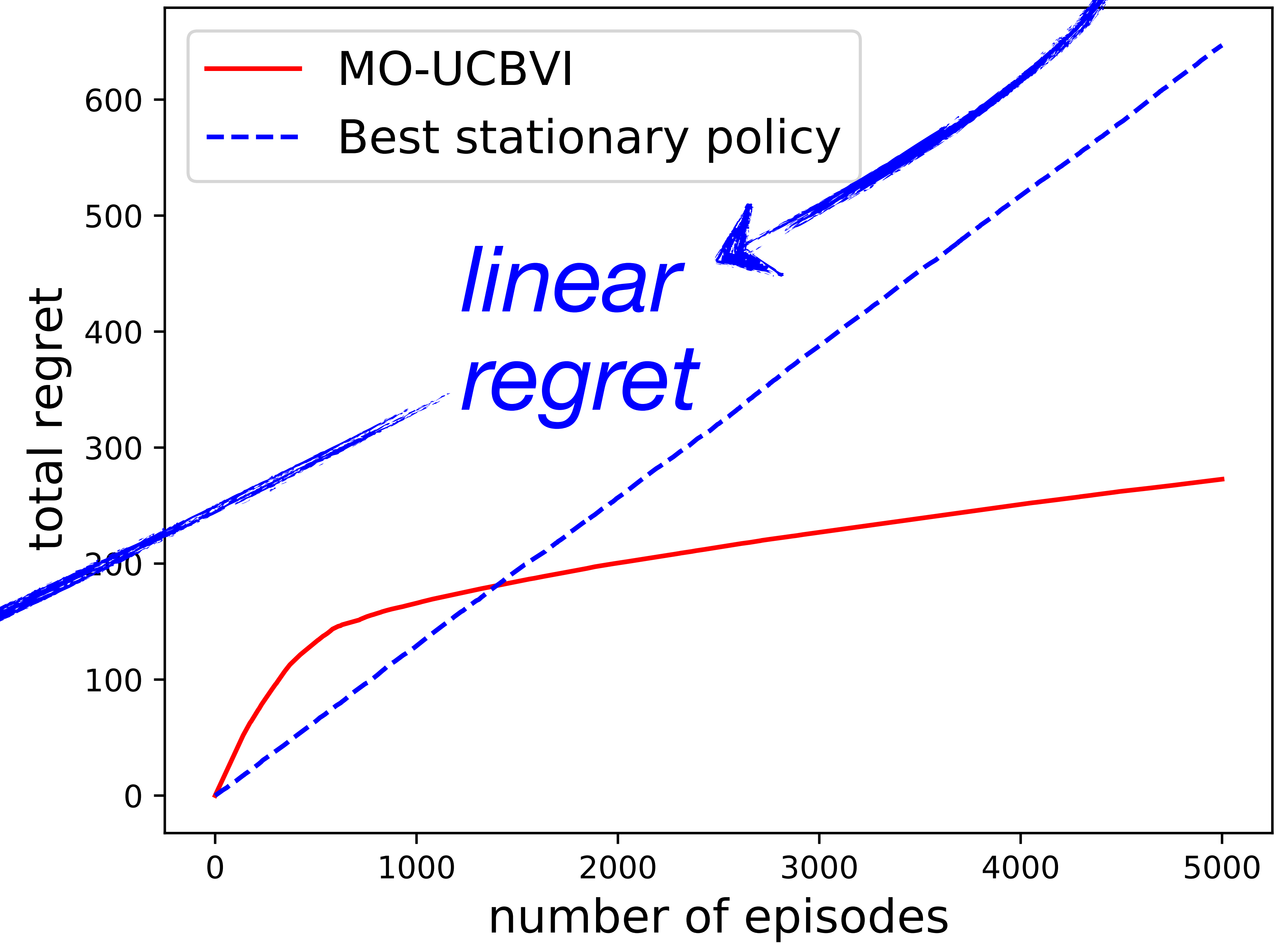
π^* depends on w

Online MORL

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

$\pi^{*,k}$ varies according to w^k





Single-obj. / adv. RL methods fail to apply

Multi-Objective UCB Value Iteration

adversary chooses preference w

agent chooses greedy policy π according to $\widehat{Q}$

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

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

$$\frac{\#(x, a, y)}{\#(x, a)}$$

UCB Bonus

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

Lemma [optimistic estimation]: with high prob. $Q_h^*(x, a; w) \leq \widehat{Q}_h(x, a; w)$ for every h, x, a, w

can be improved to Bernstein version

Regret Analysis

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

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

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

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

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

tight up to log factors

MORL is statistically harder than single-objective RL

Preference-Free Exploration

How large K is sufficient / necessary?

adversary chooses preference w

agent computes PAC-policy π

agent collects K trajectories

unsupervised exploration

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

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

$d = 1$ *Task-Agnostic Exploration*

[X. Zhou, Y. Ma, A. Singla, NeurIPS 2020]

$d = SA$ *Reward-Free Exploration*

[C. Jin, A. Krishnamurthy, M. Simchowitz, T. Yu, ICML 2020]

Algorithm & Sample Complexity

[Exploration] Set preference/reward to zero, and run MO-UCBVI (Hoeffding ver.)

[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

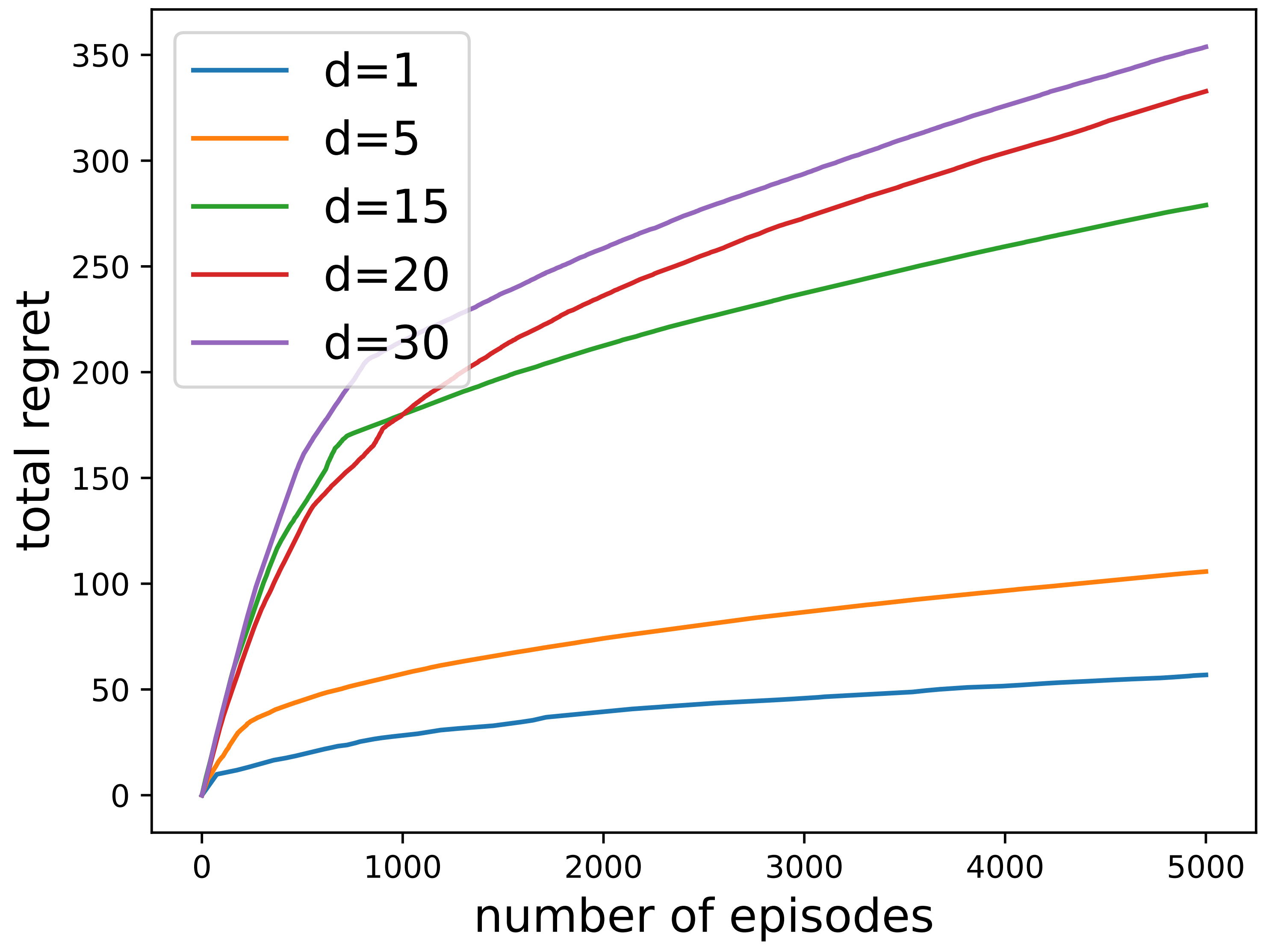
[Lower Bound] There is a distribution of MOMDPs such that for every $(\epsilon, \delta = 0.1)$-PAC algorithm, there is a (necessarily adversarial) w such that:

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

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

Numerical Simulations

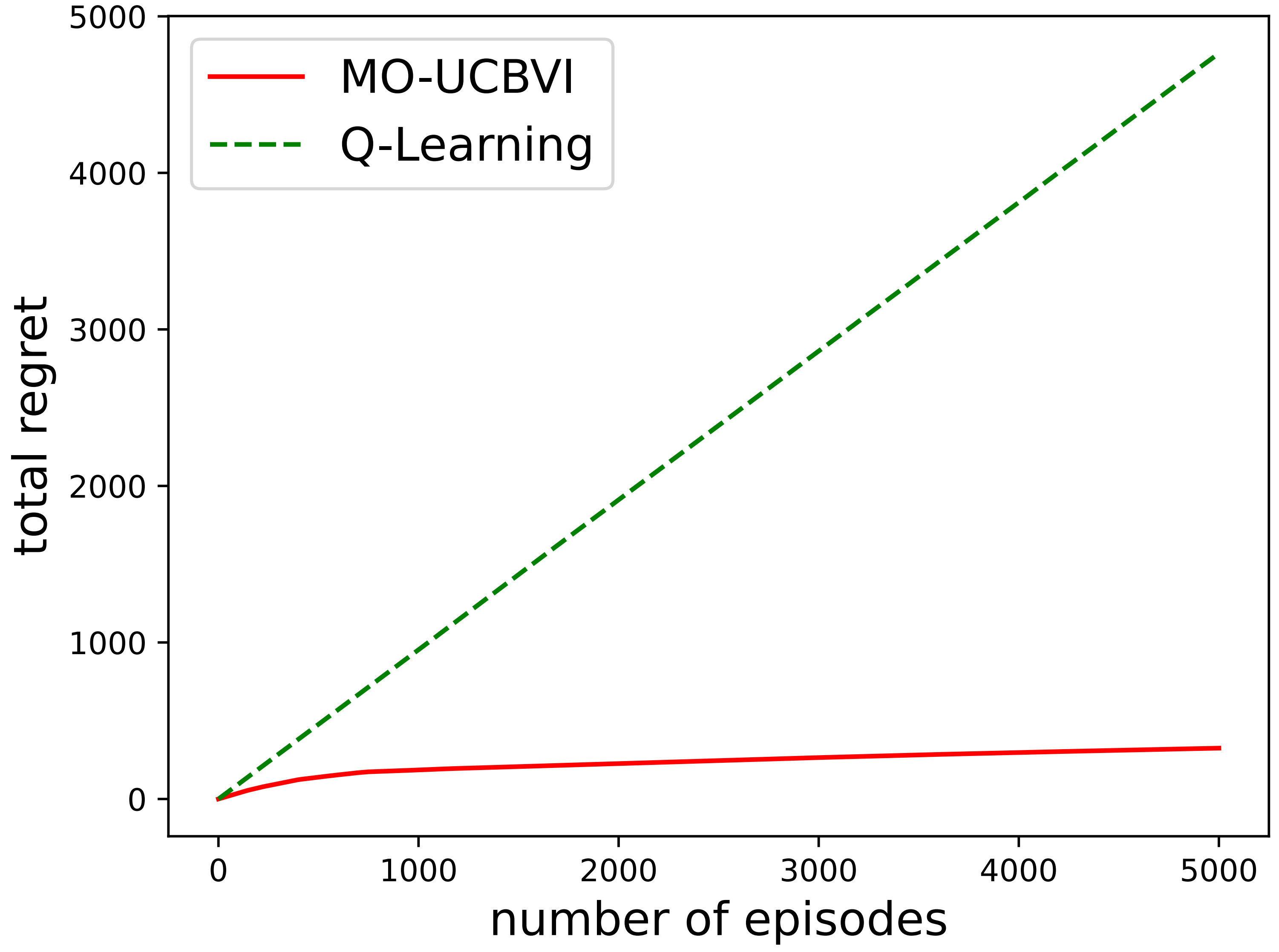
Effect of Number of Objectives



Performance of MO-UCBVI in simulated MOMDPs with different number of objectives $d \in \{1, 5, 15, 20, 30\}$

more objectives, larger regret

Single-Objective RL Method Fail to Apply



MO-UCBVI vs. Q-Learning in a simulated MOMDP with $d = 15$

sublinear regret for MO-UCBVI

Where the $\min\{d, S\}$ Stems from?

Lemma [optimistic estimation]: With high probability,

$$Q_h^*(x, a; w) \leq \widehat{Q}_h(x, a; w) \text{ for every } h, x, a, w \text{ and in every episode.}$$

[Proof]: Use induction. The key is to show:

for every h, x, a, w and in every episode,

$$|(\widehat{\mathbb{P}} - \mathbb{P})V_h^*(x, a; w)| \lesssim b(x, a) \approx \sqrt{\min\{d, S\} \cdot H^2 \cdot \log(\cdot) \ / \ \#(x, a)}$$

covering number for value function set $\approx (1/\epsilon)^S$

covering number for preference set $\approx (1/\epsilon)^d$

two union bounds + Hoeffding's ineq.

Take-Home

- RL with *multiple objectives* and *adversarial preferences*
    - upper + lower bounds
- [Online Setting] multi-objective >> single-objective
- [Unsupervised Setting] structured rewards << arbitrary rewards
- Generalize existing settings:
    - $d = 1$: *single-objective RL, task-agnostic exploration*
    - $d = SA$: *reward-free exploration*

get the paper