A Ranking Game for Imitation Learning

Harshit Sikchi, Akanksha Saran, Wonjoon Goo, Scott Niekum

The University of Texas at Austin

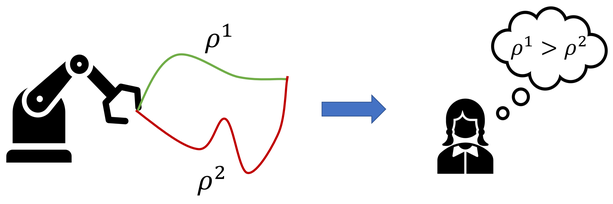
SCALAR LAB
Safe, Confident, and Aligned Learning + Robotics

TL;DR

- A novel algorithmic framework `rank-game` for imitation learning as a two-player ranking game.
- The framework provides a unified way to combine learning from preferences and demonstrations.
- We propose a new principled ranking loss that can incorporate preferences from diverse sources.
- `rank-game` outperforms state-of-the-art imitation learning methods in several MuJoCo environments and solve complex dextrous manipulation tasks that no prior methods could solve.

Motivation

Preferences are more informative about the reward function while being easy to obtain but…



Classical IRL methods provide no way to incorporate preferences among suboptimal trajectories. Prior learning from preferences method work in offline setting or use a restrictive reward class.

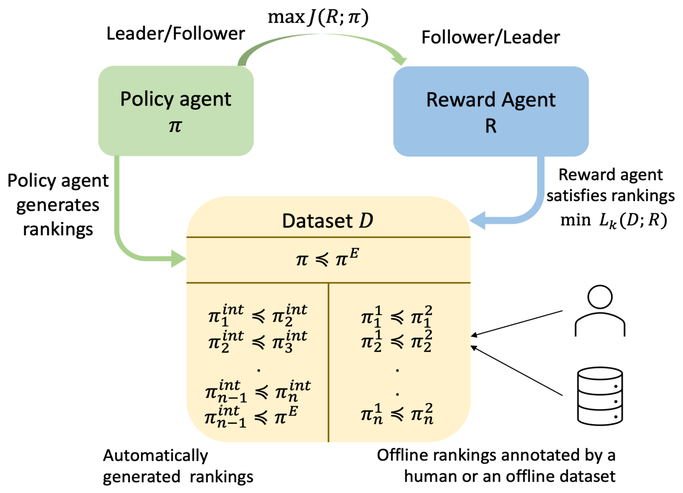
ρ denotes behavior represented by state-action or state only visitation distribution.

rank-game

$\underbrace{argmax_{\pi \in \Pi} J(R;\pi)}$ $\qquad$ $\underbrace{argmin_{R \in \mathcal{R}} L(D^p;R)}$

Policy Player
Maximizes the reward function

Reward Player
Learns a reward function satisfying all pairwise rankings in the dataset



L_k: A new ranking loss

A new class of ranking loss functions that attempts to induce a performance gap of k for all pairwise preferences.

$$L_k(D^p;R) = \mathbb{E}_{(\rho^i,\rho^j)\sim D^p}[\mathbb{E}_{s,a\sim\rho^i}[(R(s,a)-0)^2] + \mathbb{E}_{s,a\sim\rho^j}[(R(s,a)-k)^2]]$$

Theorem (Informal): Under finite samples representing preferences between behaviors and approximate policy optimization, `rank-game` has a bounded f-divergence with the expert at equilibrium of the game.

Solving the 2-player general-sum game

- Stackelberg formulation is a performant method for optimizing general sum games. One player is leader and updated slow and other is follower and updated fast. This formulation gives us two imitation algorithms:

Policy as Leader (PAL)

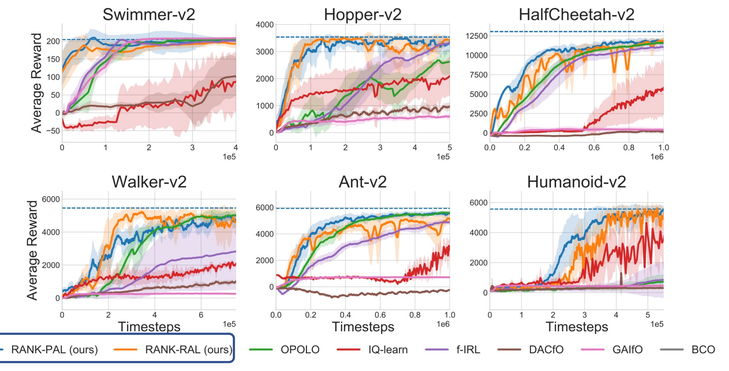
$$\max_{\pi} \{J(\hat{R};\pi) \;\; s.t \;\; \hat{R} = \arg\min_{R} L_k(D^{\pi};R)\}$$

Reward as Leader (RAL)

$$\min_{R} \{L_k(D^{\pi};\hat{R}) \;\; s.t \;\; \pi = \arg\max_{\pi} J(\hat{R};\pi)\}$$

Experiments

Online IL (LfO): Outperforms state of the art methods.



Online IL + preferences among suboptimal trajectories: Only method that solves the task.