Adversarial Imitation Learning from State-only Demonstrations*

Extended Abstract

Faraz Torabi
The University of Texas at Austin
Austin, Texas
faraztrb@cs.utexas.edu

Garrett Warnell
Army Research Laboratory
Austin, Texas
garrett.a.warnell.civ@mail.mil

Peter Stone
The University of Texas at Austin
Austin, Texas
pstone@cs.utexas.edu

AAMAS-2019

Motivation

Traditional imitation learning requires demonstrations to contain actions for corresponding states, which makes a large number of valuable learning resources uselesss – e.g online videos.

Example.
We collect some videos of driving, and we would like to train an antonomous driving agent.

Can we learn from state-only demonstrations?
Yes, Imitation from observation (IfO) provides solution to such problem.

Approach

Two component
Discrimnator: try to distinguish data generated by expert's policy vs agent's policy.
Agent's policy: try to confuse discriminator by making data look like it was generated by expert.

Problem formulation

$$\min_{\pi} \max_{D} \mathbb{E}_{\pi}[\log(D(s, s'))] + \mathbb{E}_{\pi_E}\left[\log\left(1 - D(s, s')\right)\right]$$

$(\mathrm{s}, \mathrm{s}')$: state transition pair-data.
π: learned policy
π_E: expert's policy
D: 1- generated data; 0-real data.

Algorithm

Algorithm 1 *GAIfO*

1: Initialize parametric policy π_ϕ with random ϕ
2: Initialize parametric discriminator D_θ with random θ
3: Obtain state-only expert demonstration trajectories $\tau_E = \{(s, s')\}$
4: **while** Policy Improves **do**
5: Execute π_ϕ and store the resulting state transitions $\tau = \{(s, s')\}$
6: Update D_θ using loss

$$-\left(\mathbb{E}_\tau[\log(D_\theta(s,s'))] + \mathbb{E}_{\tau_E}[\log(1 - D_\theta(s,s'))]\right)$$

7: Update π_ϕ by performing *TRPO* updates with reward function

$$-\left(\mathbb{E}_{\tau_E}[\log(1 - D_\theta(s,s'))]\right)$$

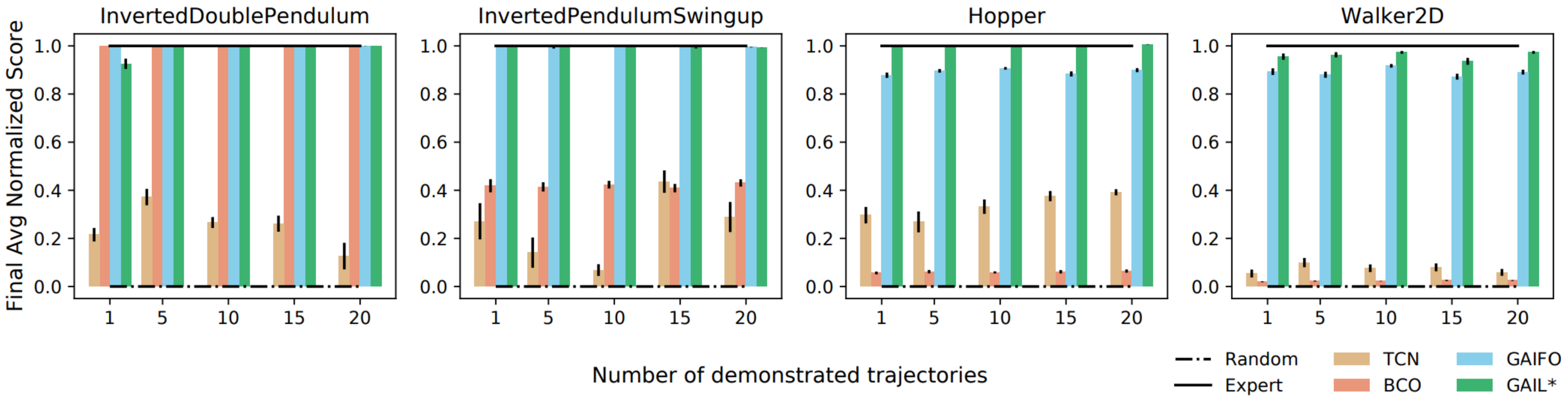
8: **end while**

$$\max_D \mathbb{E}_\pi[\log(D(s,s'))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,s'))]$$

$$\min_\pi \mathbb{E}_\pi[\log(D(s,s'))]$$

7: modify $\mathbb{E}_{\tau_E}\left[\log\left(1 - D(s,s')\right)\right]$ to $\mathbb{E}_\tau[\log(D(s,s'))]$

Experiment



Baseline
1. Behavioral Cloning from Observation
2. Time Contrastive Networks (TCN)
3. Generative Adversarial Imitation Learning (GAIL)