Provably Efficient Imitation Learning from Observations Alone

Wen Sun
CMU —> MSR NYC
[wensun@cs.cmu.edu]

Joint work with Anuridh Vemula, Byron Boots, and Drew Bagnell

Motivation

[Williams et.al, 17]

[Ross et al. 13]

[Mnih et al. 15]

[Sliver et al. 17]

Less data,
reward less clear

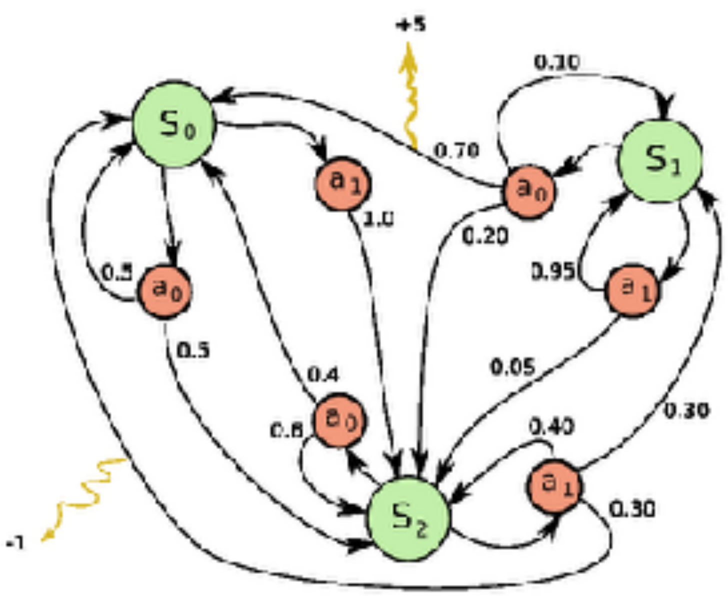
Huge data

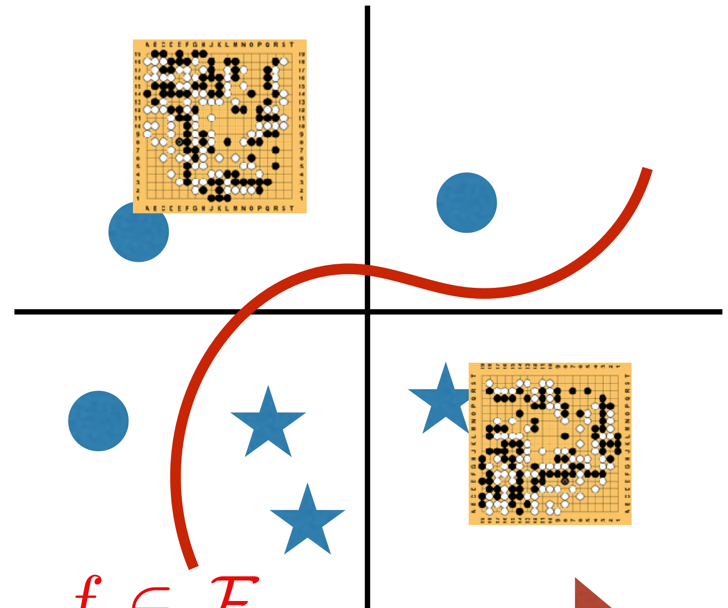
Leverage expert's demonstrations to learn efficiently,
even w/ unknown reward/cost

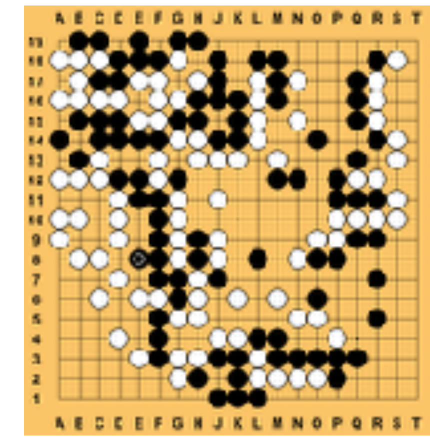
e.g., Apprenticeship Learning [Abbeel & Ng 05, Syed & Schapire 08]
Inverse Optimal Control [Ziebart & Bagnell, 10]
Interactive Imitation Learning [Ross& Bangell, 11, 14]
Generative Adversarial Imitation Learning [Ho & Ermon 16]

Theoretical Motivation:
Scale Provably Efficient RL to Large Scale MDPs

Sample Efficiency







Bridge

Sample Complexity:

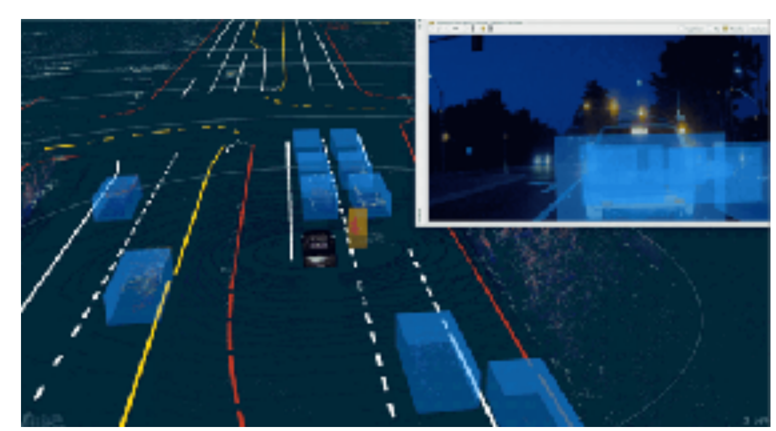
To achieve ϵ near-optimal policy,

we need at most

$$\boxed{\text{poly}(\#\text{ of states},}\ \#\text{ of actions, Horizon}, 1/\epsilon)$$

e.g. VC-dim

many interactions



[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]

Previous Works that Can Achieve:

$$\text{poly}(\text{Horizon}, \# \text{ of actions}, 1/\epsilon, \text{complexity of function classes})$$

1. Reactive POMDP (small # of hidden state)

(Krishnamurthy et al., NeurIPS 16, Dann et al, NeurIPS 18, Du et al, ICML 19)

2. Decision Process w/ Low Bellman Rank

(Jiang et al., ICML 17)

3. Markov Decision Process w/ Low Witness Rank

(Sun et al., COLT 19)

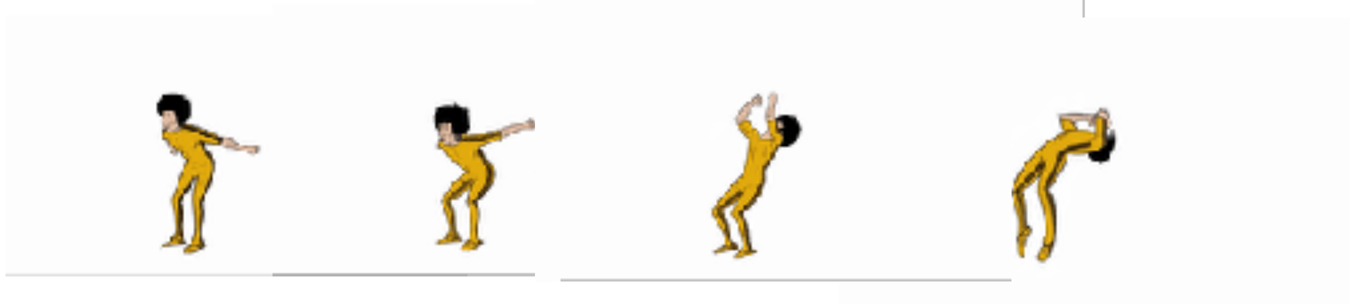
…and a lot of works on Contextual Bandits (horizon=1)

(e.g., Agarwal et al., ICML 14)

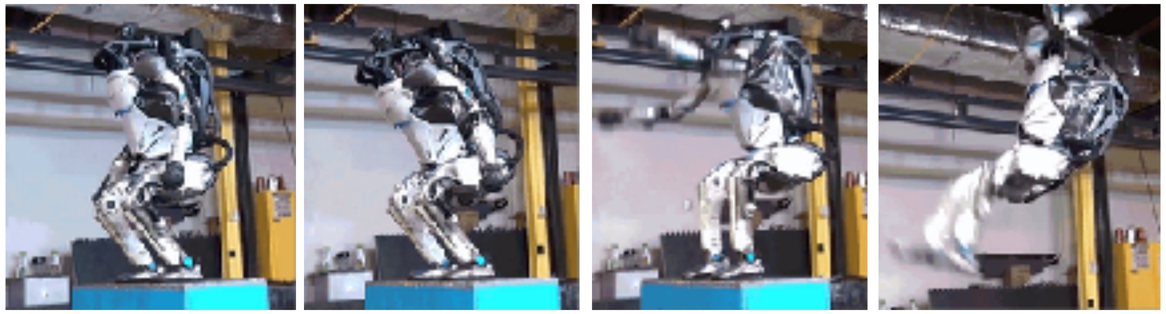
Imitation Learning from Observations

[e.g., Torabi et.al 18, Edwards et.al, 18, Liu et.al, 17, Peng et.al,18]

Trajectories of Observations



Learning From Observations



No interactive expert, no expert action, no reset, no cost signals.

Finite Horizon (T-step) Episodic MDP
Ground-Truth cost function $c_T(x) \in [0, 1]$.

Different from RL:
Unknown cost, but
we have state-only demonstrations from expert $\pi^\star$

Model-Free Algorithm:
Forward Adversarial Imitation Learning (FAIL):

$$\pi_1, \pi_2, \pi_3, \ldots, \pi_{T-1}$$

Two-player game

Reduce Sequential Problem into H many min-max games

Min-Max Games: Minimizing Integral Probability Metrics (IPM)

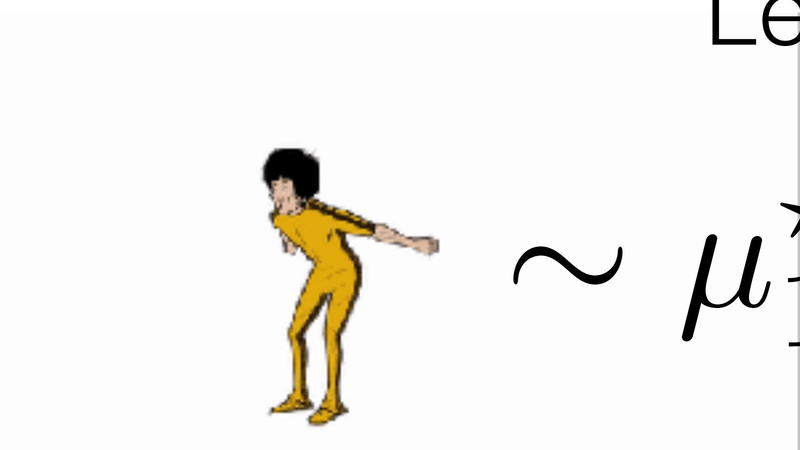
Distinguish 2 distributions:

$$\max_{f \in \mathcal{F}} \left(\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)] \right)$$

Set of discriminators

Total-variation
Wasserstein Distance
Max Mean Discrepancy
...

Learning the first policy:



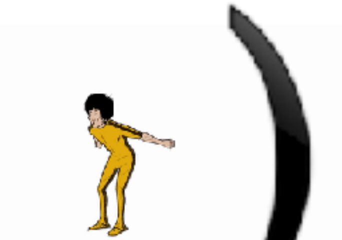
$\sim \mu_1^\star$

Expert distribution

$\sim P(\cdot | x_0, \pi_0(x_0))$

π_0 and Dynamics:
Generator

$$\min_{\pi_0 \in \Pi} \max_{f \in \mathcal{F}} f\Big($$



$$\Big) - f\Big($$

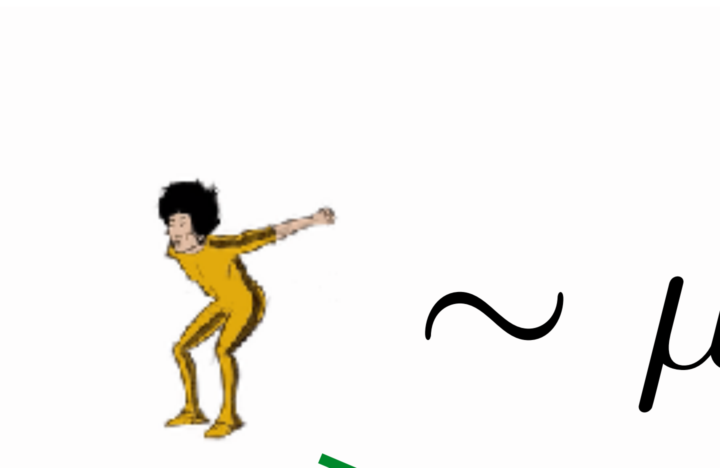
$$\Big)$$

IPM

Learning the *Second* Policy

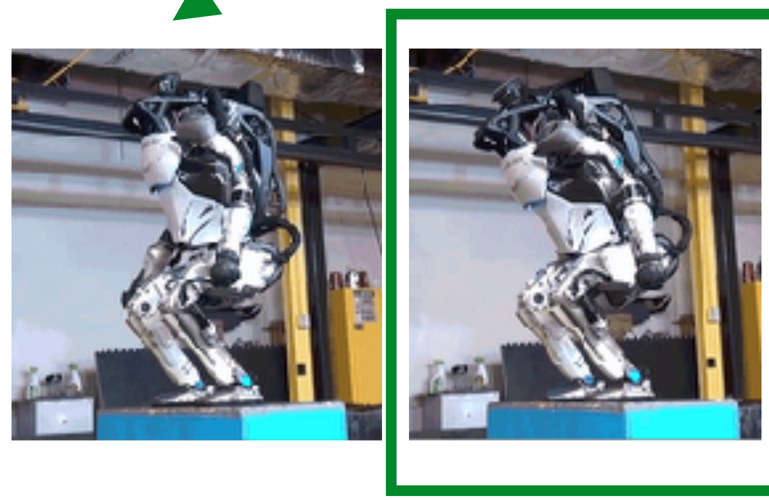
Now we have **already learned** π_0

Roll in



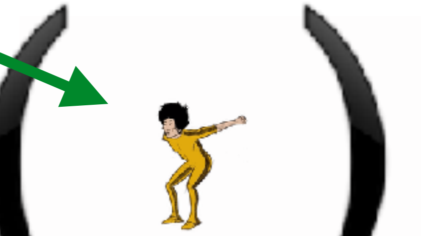
$\sim \mu_2^\star$

Expert distribution

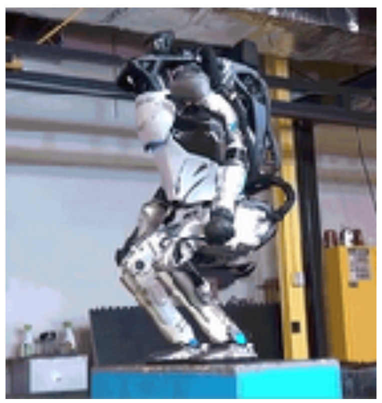


$\sim P(\cdot|x_1, \pi_1(x_1))$

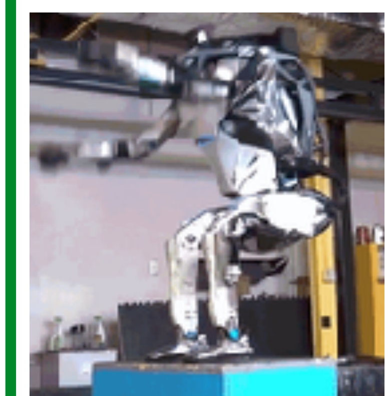
π_1 and Dynamics:
Generator

$\min_{\pi_1 \in \Pi} \max_{f \in \mathcal{F}} f\Big($  $\Big) - f\Big($  $\Big)$

Keep Forward Training....

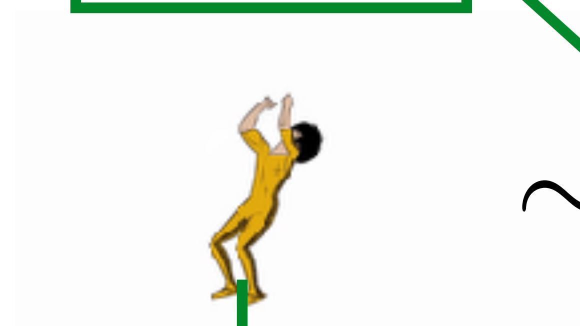
π_0  π_1 

π_2

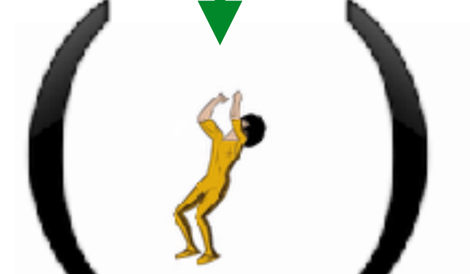


$\sim P(\cdot|x_2, \pi_2(x_2))$

Rolling in...



$\sim \mu_3^\star$

$\min_{\pi_2 \in \Pi} \max_{f \in \mathcal{F}} f($  $) - f($  $)$

Capacity of Discriminators

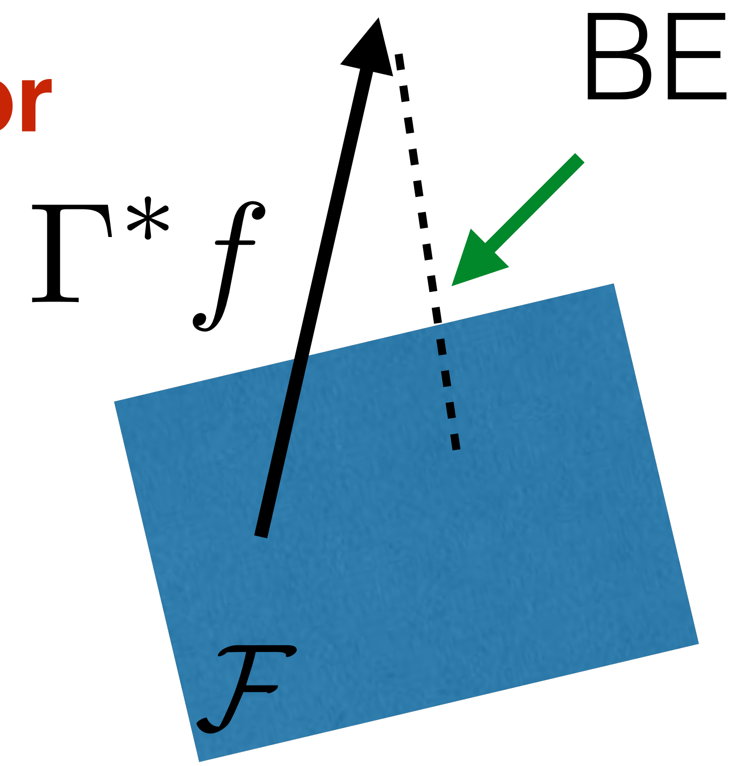
$$\min_{\pi_2\in\Pi}\max_{f\in\mathcal{F}} f\Big($$  $$\Big) - f\Big($$  $$\Big)$$

Strong Discriminator => Overfitting (need a lot of samples)
Weak Discriminator => Unable to distinguish

Inherent Bellman Error

$$(\Gamma^\star f)(x) \triangleq \mathbb{E}_{a\sim\pi^\star(x),x'\sim P_{x,a}} f(x')$$

$$\mathrm{BE} = \min_{f'}\max_{f} \|f' - \Gamma^\star f\|_\infty$$



Analysis of FAIL

Realizability Assumption:

$$\pi^\star \in \Pi, V^\star \in \mathcal{F}$$

($\pi^\star, V^\star$: expert's policy & value function)

To learn a near-optimal policy:

$$J(\pi) - J(\pi^\star) \leq T^2(\mathrm{BE} + \epsilon)$$

we need samples:

$$\mathrm{poly}(T, A, 1/\epsilon, \mathrm{SC}(\Pi), \mathrm{SC}(\mathcal{F}))$$

Statistical Complexity of
Policy & Discriminator classes

Discriminators $\approx$ expert's value functions

Approximate Policy Improvement over expert

Is Inherent Bellman Error Avoidable in the IL from Observation Setting?

Yes in model-based setting…

Start with a realizable **model class** $\mathcal{P}$ & discriminator class $\mathcal{F}$

$$P \in \mathcal{P}, V^\star \in \mathcal{F}$$

There exists an algorithm that takes $\{\mathcal{P}, \mathcal{F}\}$ as input, outputs an ϵ optimal policy, with # of samples:

$$\mathrm{Poly}\left(H, A, 1/\epsilon, \mathrm{SC}(\mathcal{P}), \mathrm{SC}(\mathcal{F})\right)$$

(Note such result is not possible in RL setting [1])

but, model-free IL from Observation setting?

[1] Model-based RL in CDPs: PAC bounds and Exponential Improvements over Model-free Approaches, W Sun, N Jiang, A Krishnamurthy, A Agarwal, J Langford, COLT 19

A Simpler Baseline…

Minimizing some divergence between avg state distributions
(e.g., Generative Adversarial Imitation Learning (GAIL))

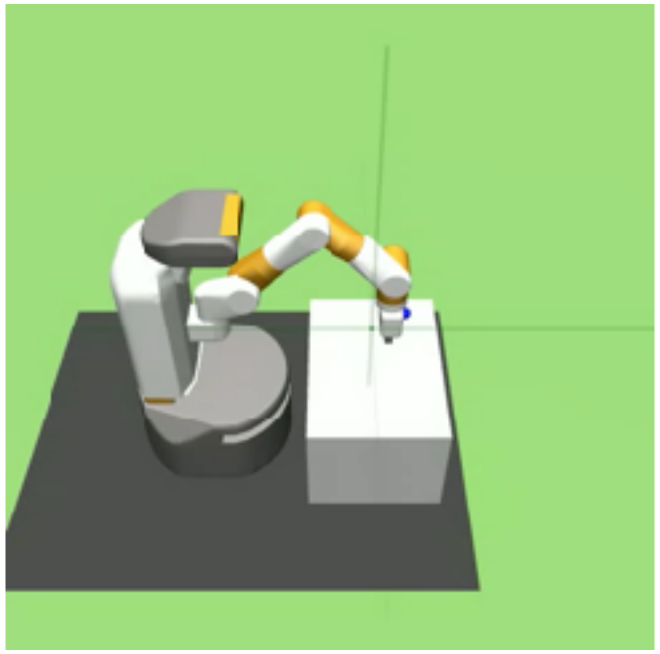
[Ho & Ermon 16]

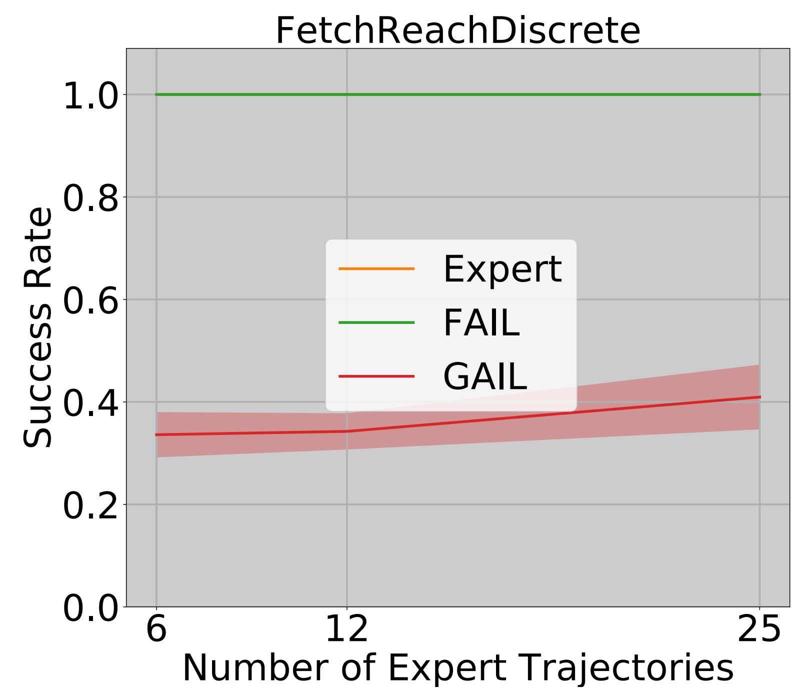
d_π average state distribution over horizon of π

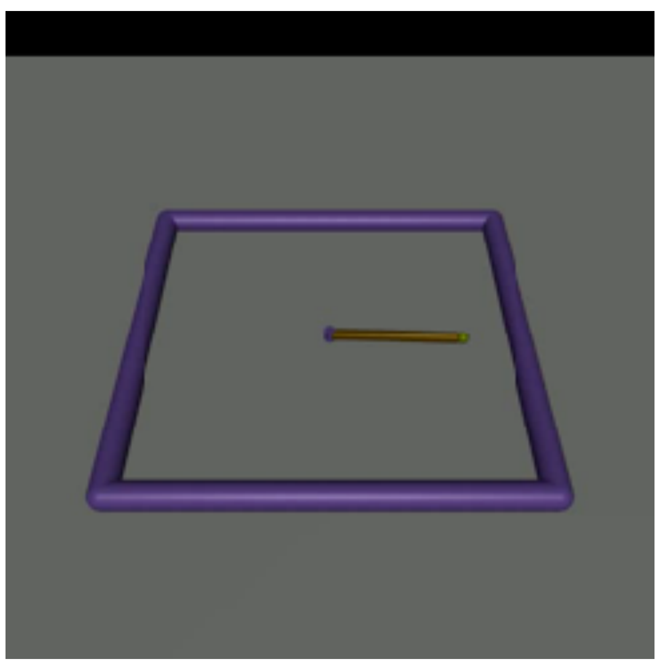
$$\min_{\pi \in \Pi} \boxed{\max_{f \in \mathcal{F}} \mathbb{E}_{x \sim d_\pi}[f(x)] - \mathbb{E}_{x \sim d_{\pi^\star}}[f(x)]}$$

new RL objective function,

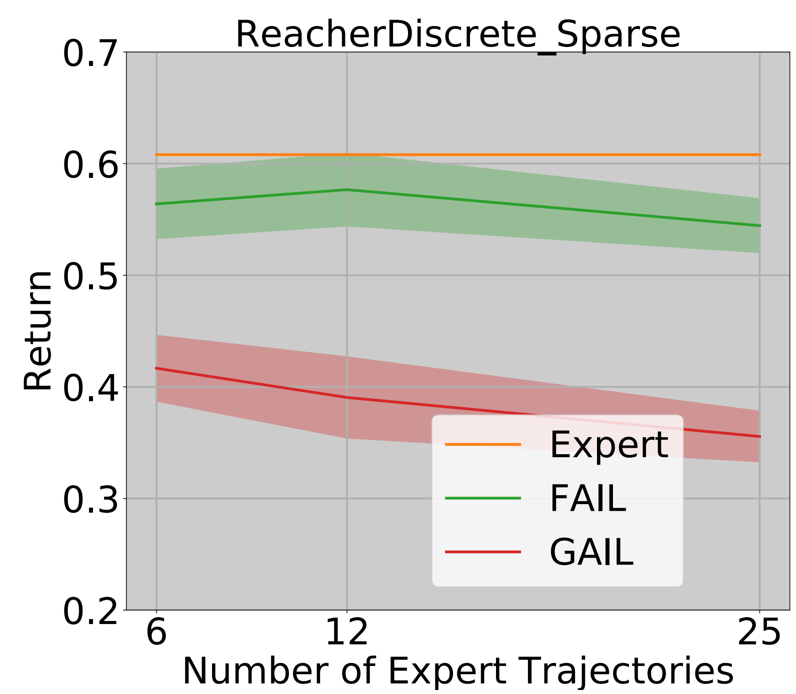
Simulation Results







Success rate



# of expert trajectories

FAIL code: https://github.com/wensun/Imitation-Learning-from-Observation
GAIL (without actions) is adopted from existing code in OpenAI baselines

Take Away Messages

With Observations alone from experts, we can learn near optimal policies:

- Near-Optimal Guarantee
- Supervised Learning type sample complexity
- Out-of-box performance is pretty good

Future Work

- Get rid of inherent Bellman error in model-free IL setting?
- A computationally efficient model-based algorithm?

Thanks!

wensun@cs.cmu.edu

https://github.com/wensun/Imitation-Learning-from-Observation