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

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 & Bagnell, 11, 14]
Generative Adversarial Imitation Learning [Ho & Ermon 16]
Theoretical Motivation: Scale Provably Efficient RL to Large Scale MDPs

Sample Efficiency

Sample Complexity:
To achieve $\epsilon$ near-optimal policy, we need at most
poly(# of states, # of actions, Horizon, $1/\epsilon$)
many interactions

e.g. VC-dim

[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]
Previous Works that Can Achieve:

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

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 $\pi^*$. 
Model-Free Algorithm: Forward Adversarial Imitation Learning (FAIL):

Reduce Sequential Problem into $H$ many min-max games

Two-player game

$\pi_1, \pi_2, \pi_3, \ldots, \pi_{T-1}$
Min-Max Games: Minimizing Integral Probability Metrics (IPM)

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

Distinguish 2 distributions:

Learning the first policy:

\[ \sim \mu_1^{*} \]

Expert distribution

\[ \sim P(\cdot | x_0, \pi_0(x_0)) \]

\[ \pi_0 \text{ and Dynamics: Generator} \]

\[ \min_{\pi_0 \in \Pi} \max_{f \in \mathcal{F}} f(\cdot) - f(\cdot) \]

Total-variation Wasserstein Distance
Max Mean Discrepancy

\[ \ldots \]
Learning the *Second* Policy

Now we have *already learned* $\pi_0$

Roll in

$p \sim \mu^*_{2}$

Expert distribution

$\pi_1$ and Dynamics: Generator

$\min_{\pi_1 \in \Pi} \max_{f \in F} f(\cdot) - f(\cdot)$
Keep Forward Training....

\[ \pi_0 \quad \pi_1 \quad \pi_2 \]

Rolling in...

\[ \min_{\pi_2 \in \Pi} \max_{f \in \mathcal{F}} \left( f(\cdot) - f(\cdot) \right) \]

\[ \sim P(\cdot | x_2, \pi_2(x_2)) \]

\[ \sim \mu_3^* \]
Capacity of Discriminators

\[
\min_{\pi_2 \in \Pi} \max_{f \in \mathcal{F}} \ f(\cdot) - f(\cdot)
\]

Strong Discriminator \(\Rightarrow\) Overfitting (need a lot of samples)
Weak Discriminator \(\Rightarrow\) Unable to distinguish

**Inherent Bellman Error**

\[
(\Gamma^* f)(x) \triangleq \mathbb{E}_{a \sim \pi^*(x), x' \sim P_{x,a}} f(x')
\]

\[
\text{BE} = \min_{f'} \max_f \| f' - \Gamma^* f \|_\infty
\]
Analysis of FAIL

Realizability Assumption:
\[ \pi^* \in \Pi, \quad V^* \in \mathcal{F} \]

\((\pi^*, V^* : \text{expert's policy & value function})\)

To learn a near-optimal policy:
\[ J(\pi) - J(\pi^*) \leq T^2 (BE + \epsilon) \]

we need samples:
\[ \text{poly}(T, A, 1/\epsilon, \text{SC}(\Pi), \text{SC}(\mathcal{F})) \]

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 $\mathcal{F}$

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

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

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

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

but, model-free IL from Observation setting?

A Simpler Baseline…

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

\[ d_{\pi} \text{ average state distribution over horizon of } \pi \]

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

new RL objective function,
Simulation Results

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