Learning from Demonstration: Provably Efficient Adversarial Policy Imitation with Linear Function Approximation

Zhihan Liu[1] Yufeng Zhang[1] Zuyue Fu[1] Zhuoran Yang[2] Zhaoran Wang[1]

[1]Northwestern University [2]Yale University

Overview

In this paper, we propose two algorithms for online and offline generative adversary imitation learning (GAIL) with linear function approximation, respectively. We also analyze their theoretical properties, showing that they are both provably efficient.

Online and Offline GAIL

- **Online GAIL.** Online GAIL is proposed to solve Imitation Learning, where the agent cannot access the reward information but an expert demonstration $\mathbb{D}^{\mathrm{E}}$ is available. Online GAIL can be formulated by the following minimax optmiziation problem:

$$\min_{\pi\in\Delta(\mathcal{S}|\mathcal{A},H)}\max_{r\in\mathcal{R}} J(\pi^{\mathrm{E}},r)-J(\pi,r).$$

- **Offline GAIL.** When online interaction is expensive but a historical dataset is available, we may turn to consider the offline GAIL, where expert demonstration $\mathbb{D}^{\mathrm{E}}$ and an additional dataset $\mathbb{D}^{\mathrm{A}}$ are both available.

Challenges

- *Minimax optimization* problems w.r.t. the policy and reward function.
- Exploration-exploitation tradeoff in online GAIL and distribution shift in offline GAIL.
- For offline GAIL, we are incapable to update the reward function based on the trajectory of present policy.
- Adoption of *linear function approximation*.

OGAPI for Online GAIL

- **Policy update stage:**
(i) Policy improvement: We apply *mirror descent* to update policy,

$$\pi_h^k(\cdot\,|\,s)\propto\pi_h^{k-1}(\cdot\,|\,s)\cdot\exp\{\alpha\cdot\widehat{Q}_h^{k-1}(s,\cdot)\}.$$

(ii) Policy evaluation: Based on *Bellman equation* and regression on the finite historical data, we update $\widehat{Q}_h^{k-1}$. *Optimistic bonus is also incorporated here to enhance exploration.*

- **Reward update stage:** *Projected gradient ascent* on the reward parameter,

$$\mu_h^{k+1}=\mathrm{Proj}_B\{\mu_h^k+\eta\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\},$$

where the gradient estimator $\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)$ is chosen as $\nabla_{\mu_h}\tilde{J}(\pi^{\mathrm{E}},r^\mu)|_{\mu=\mu^k}-\widehat{\nabla}_{\mu_h}J(\pi^k,r^\mu)|_{\mu=\mu^k}$. Here the first term is derived from Monte Carlo (MC) estimation based on $\mathbb{D}^{\mathrm{E}}$ and the last term can be evaluated on the trajectory induced by π^k.

Analysis of OGAPI

- **Sublinear Regret.** Theorem 4.1 shows the online regret of OGAPI for K episodes can be bounded by:

$$\mathrm{Regret}(K)\le\mathcal{O}\big(\sqrt{H^4d^3K}\log(HdK/\xi)\big)+K\Delta_{N_1},$$

where $\Delta_{N_1}=\mathcal{O}(\sqrt{H^3d^2/N_1}\log(N_1/\xi))$ is an inevitable statistical error from the MC estimation on $\mathbb{D}^{\mathrm{E}}$. Here N_1 is the size of $\mathbb{D}^{\mathrm{E}}$. Thus OGAPI is provably efficient.

PGAPI for Offline GAIL

- Based on $\mathbb{D}^{\mathrm{A}}$, we construct the estimated kernels $\widehat{\mathcal{P}}_h$ and uncertainty qualifiers Γ_h (Jin et al., 2021).
- **Policy update stage:**
(i) Policy improvement: Same as OGAPI.
(ii) Policy evaluation: Based on Bellman equation and constructed $\widehat{\mathcal{P}}_h$ and Γ_h, we update $\widehat{Q}_h^{k-1}$ *by the pessimism principle.*
- **Reward update stage:** We estimate the gradient w.r.t. the reward parameter by, $\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)=\nabla_{\mu_h}\tilde{J}(\pi^{\mathrm{E}},r^\mu)|_{\mu=\mu^k}-\nabla_{\mu_h}\widehat{J}(\pi^k,r^\mu)|_{\mu=\mu^k}$, where the last term is calculated in Proposition D.1.

Analysis of PGAPI

- **General Conclusion.** Theorem 4.2 characterizes the optimality gap of PGAPI by

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\widehat{\pi})\le\mathcal{O}\big(\sqrt{H^4d^2/K}\big)+\Delta_{N_1}+\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}},$$

where $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$ is the intrinsic error determined by the overlap of $\mathbb{D}^{\mathrm{A}}$ and π^{E}.
- **Minimax Optimality.** Proposition F.1 provides a lower bound, showing that PGAPI achieves minimax optimality in utilizing $\mathbb{D}^{\mathrm{A}}$.
- **Global Convergence.** Assuming that $\mathbb{D}^{\mathrm{A}}$ has sufficient coverage, Corollary 4.4 proves that PGAPI attains global convergence:

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\widehat{\pi})\le\tilde{\mathcal{O}}\big(\sqrt{H^4d^2/K}+\sqrt{H^4d^3/N_2}+\sqrt{H^3d^2/N_1}\big).$$