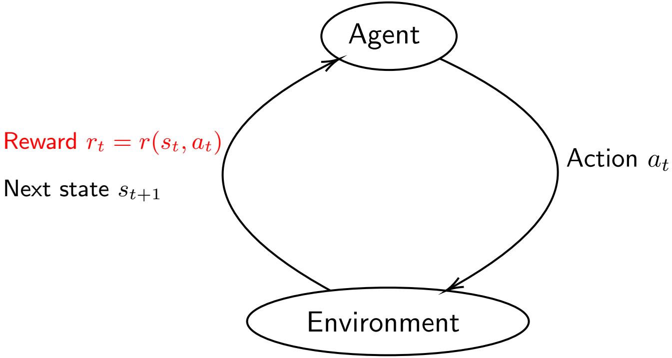
Learning from Demonstration: Provably Efficient Adversarial Policy Imitation with Linear Function Approximation

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

[1]Northwestern University [2]Yale University

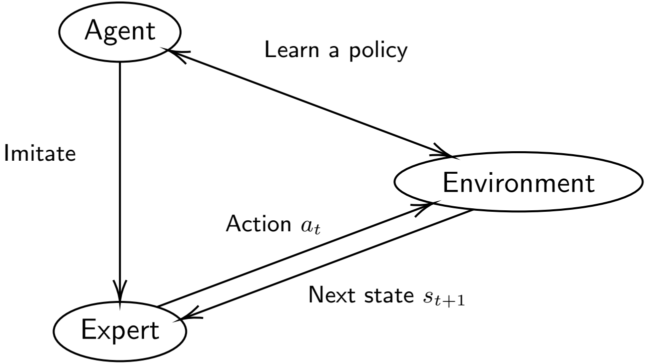
July 14, 2022

Reinforcement Learning (RL)



The agent aims to learn a policy by interacting with the environment. However, in real-world tasks, the reward function may not be available.

Imitation Learning (IL)



- The agent aims to learn a policy that has similar performance to the expert policy.
- Different from RL, the agent has *no access to reward information but an expert demonstration* $\mathbb{D}^{\mathrm{E}}$ that stores a finite number of expert trajectories.

Generative Adversarial Imitation Learning (GAIL)

Online GAIL (Goodfellow et al., 2014; Arjovsky et al., 2017)

- *Minimax optimization problem*:

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

- Available: Expert demonstration $\mathbb{D}^{\mathrm{E}}$ and online interaction.
- Lack of theoretical study with linear function approximation on both transition kernels and reward functions.

Offline GAIL

- Scenario: Online interaction is expensive but a historical dataset is available.
- Available: Expert demonstration $\mathbb{D}^{\mathrm{E}}$ and *an additional dataset* $\mathbb{D}^{\mathrm{A}}$ *collected a priori*.

Challenges

- *Minimax optimization* problems with respect to 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* (Both the transition kernels $\mathcal{P}_h$ and reward set $\mathcal{R}$ is linear).

Main Contribution

- For online GAIL with linear function approximation, we propose OGAPI and prove its online regret, showing that OGAPI is provably efficient.
- For offline GAIL with linear function approximation, we design PGAPI and obtain its optimality gap in the general case.
- If we further assume that the additional dataset has sufficient coverage on the expert policy, we prove that PGAPI achieves global convergence.

Optimistic Generative Adversarial Policy Imitation (OGAPI)

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

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

  - 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 $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)$ is defined as

$$\underbrace{\nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^{\mu})|_{\mu=\mu^k}}_{\text{Monte Carlo (MC) estimation on } \mathbb{D}^{\mathrm{E}}} - \underbrace{\widehat{\nabla}_{\mu_h} J(\pi^k, r^{\mu})|_{\mu=\mu^k}}_{\text{Evaluated on the trajectory induced by } \pi^k.}.$$

Analysis of OGAPI

- Online regret for K episodes:

$$\mathrm{Regret}(K) = \max_{r \in \mathcal{R}} \sum_{k=1}^{K} \left[J(\pi^{\mathrm{E}}, r) - J(\pi^{k}, r) \right]$$

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

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

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

- When $K, N_1 \to \infty$, average regret $\mathrm{Regret}(K)/K$ shrinks to zero, meaning that the output policy has the same performance on average with π^{E} w.r.t. the reward set $\mathcal{R}$.

Pessimistic Generative Adversarial Policy Imitation (PGAPI)

- 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:**
  - Policy improvement: Same as OGAPI.
  - Policy evaluation: Based on Bellman equation and constructed $\widehat{\mathcal{P}}_h$ and uncertainty qualifiers Γ_h, we update $\widehat{Q}_h^{k-1}$. We incorporate pessimism principle by subtraction of Γ_h.
- **Reward update stage:**
  - Project 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 $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)$ is defined as

$$\underbrace{\nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^\mu)|_{\mu=\mu^k}}_{\text{Monte Carlo (MC) estimation on } \mathbb{D}^{\mathrm{E}}} - \underbrace{\nabla_{\mu_h} \widehat{J}(\pi^k, r^\mu)|_{\mu=\mu^k}}_{\text{Term}(\star)}.$$

  - Based on $\widehat{Q}_h^{k-1}$, term $(\star)$ is calculated in Proposition D.1.

Analysis of PGAPI

- Optimality gap:

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \pi) = \max_{r \in \mathcal{R}} \big[J(\pi^{\mathrm{E}}, r) - J(\pi, r)\big].$$

- In the general case, Theorem 4.2 characterizes the optimality gap of PGAPI by

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

  where the MC estimation error Δ_{N_1} also appears in Theorem 4.1, and intrinsic error $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$ is defined as $2 \sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}[\Gamma_h(s_h, a_h) \mid s_1 = x]$.
- Proposition F.1 provides a lower bound, showing that PGAPI achieves minimax optimality in utilizing $\mathbb{D}^{\mathrm{A}}$.

Analysis of PGAPI

- Sufficient Coverage: A weak assumption, which only involves policy π^{E} and the dataset, and does NOT restrict the distribution of the dataset or assume the dataset is well-explored.
- Assuming that $\mathbb{D}^{\mathrm{A}}$ has sufficient coverage, Corollary 4.4 proves that PGAPI attains global convergence at a rate of negative square-root,

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

Thank You!