uai2023

Oral

Provably Efficient Adversarial Imitation Learning with Unknown Transitions

Tian Xu*[1,4], Ziniu Li*[2,3], Yang Yu‡[1,4], and Zhi-Quan Luo‡[2,3]

1: National Key Laboratory for Novel Software Technology, Nanjing University
2: The Chinese University of Hong Kong, Shenzhen
3: Shenzhen Research Institute of Big Data
4: Polixir.ai

Problem Formulation

➔ Imitation Learning (IL): mimic the expert π^{E} from demonstrations:

$$\mathcal{D} = \{\mathbf{tr} = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H);\, a_h \sim \pi_E(\cdot|s_h)\}.$$

➔ Adversarial imitation learning (AIL): imitate by state-action distribution matching

$$\min_{\pi\in\Pi} \sum_{h=1}^{H} \left\| d_h^{\pi,P} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 \tag{1}$$

- $d_h^{\pi,P}(s,a) = \mathbb{P}^{\pi}(s_h = s, a_h = a)$: true state-action distribution.
- $\widetilde{d}_h^{\pi^{\mathrm{E}}}$: estimation of $d_h^{\pi^{\mathrm{E}}}$ from the finite dataset $\mathcal{D}$.

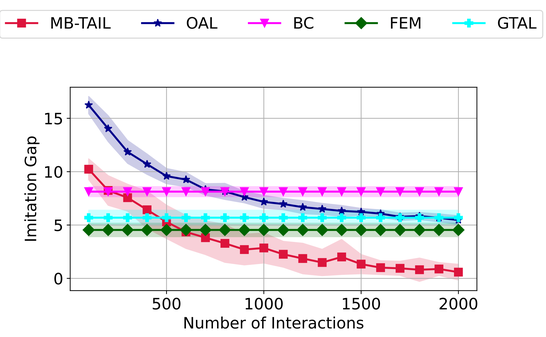
➔ Setting:

- Unknown π^{E}: m trajectories (of expert) are given.
- Unknown P: n trajectories (of learner) to interact with env.

Main Results

Theoretical guarantees of IL algorithms with unknown transitions

	Expert Sample Complexity (m)	Interaction Complexity (n)
BC[1]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	0
FEM[2]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2} + \frac{H^8\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{\varepsilon^5}\right)$	0
GTAL[3]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2} + \frac{H^6\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{\varepsilon^3}\right)$	0
OAL[4]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^4\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varepsilon^2}\right)$
MB-TAIL (this work)	$\widetilde{\mathcal{O}}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^3\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varepsilon^2}\right)$



Algorithm Performance

Remark

- Theoretically improvements on both m and n
- Better empirical performance

Algorithmic Framework

Algorithm 1 Meta-algorithm for AIL with Unknown Transitions

Require: Expert demonstrations $\mathcal{D}$.
1: $\widehat{P} \leftarrow$ Invoke a method to interact with the env and then learn a transition model.
2: $\widetilde{d}_h^{\pi^{\mathrm{E}}} \leftarrow$ Estimate the expert state-action distribution from $\mathcal{D}$.
3: $\overline{\pi} \leftarrow$ Apply AIL to perform distribution matching with $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ and $\widehat{P}$.
Ensure: Policy $\overline{\pi}$.

Proposition 1: Error Decomposition

(a) $\widehat{P}$ is $(\varepsilon_{\mathrm{RFE}}, \delta_{\mathrm{RFE}})$-PAC for uniform policy evaluation (UPE):

$$\mathbb{P}\left(\text{for any } r, \pi \in \Pi, |V^{\pi,P,r} - V^{\pi,\widehat{P},r}| \le \varepsilon\right) \ge 1 - \delta.$$

(b) $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ is $(\varepsilon_{\mathrm{EST}}, \delta_{\mathrm{EST}})$-PAC for estimating $d_h^{\pi^{\mathrm{E}}}$:

$$\mathbb{P}\left(\sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\| \le \varepsilon_{\mathrm{EST}}\right) \ge 1 - \delta_{\mathrm{EST}}.$$

(c) $\overline{\pi}$ is $\varepsilon_{\mathrm{OPT}}$-optimal solution:

$$\sum_{h=1}^{H} \left\| d_h^{\overline{\pi},\widehat{P}} - \widehat{d_h^{\pi^{\mathrm{E}}}} \right\|_1 \le \min_{\pi\in\Pi} \sum_{h=1}^{H} \left\| d_h^{\pi,\widehat{P}} - \widehat{d_h^{\pi^{\mathrm{E}}}} \right\|_1 + \varepsilon_{\mathrm{OPT}}.$$

Then we have

$$\mathbb{P}(V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} \le 2\varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}) \ge 1 - \delta_{\mathrm{EST}} - \delta_{\mathrm{RFE}}.$$

Proposition 1 characterizes three types of errors in the training of AIL:

$$\texttt{Imitation Gap} \lesssim \texttt{Exp. Error} + \texttt{Est. Error} + \texttt{Opt. Error}$$

Theoretical Analysis

(Part a) Efficient Exploration for UPE

Definition 1: Reward-free Exploration (RFE)

An algorithm is (ε, δ)-PAC for RFE if

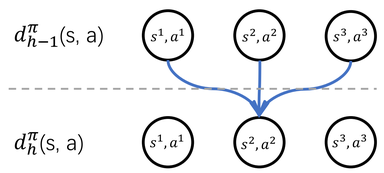
$$\mathbb{P}\left(\text{for any reward function } r, |V^{\pi_r^*} - V^{\widehat{\pi}_r^*}| \le \varepsilon\right) \ge 1 - \delta.$$

Lemma 1: RF-Express Solves RFE (and UPE)

RF-Express Algorithm in [5] ensures (ε, δ)-PAC for UPE if

$$n \gtrsim \frac{H^3|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(|\mathcal{S}| + \log\left(\frac{|\mathcal{S}|H}{\delta}\right)\right).$$

(Part b) Optimal State-action Distribution Estimation



Bellman-flow equation: $d_h^{\pi^{\mathrm{E}}}(s,a) = (\sum_{s',a'} d_{h-1}^{\pi^{\mathrm{E}}}(s',a') P_{h-1}(s|s',a')) \pi_h^{\mathrm{E}}(a|s)$.

Split $\mathcal{D}$ into two parts: $\mathcal{D} = \mathcal{D}^1 \cup \mathcal{D}_1^c$. Decompose $d_h^{\pi^{\mathrm{E}}}(s,a) =$

$$\underbrace{\sum_{\mathbf{tr}_h \in \mathbf{Tr}_h^{\mathcal{D}^1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot) = (s,a)\}}_{:=\clubsuit} + \underbrace{\sum_{\mathbf{tr}_h \notin \mathbf{Tr}_h^{\mathcal{D}^1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot) = (s,a)\}}_{:=\spadesuit}$$

Here $\mathbf{Tr}_h^{\mathcal{D}^1}$ is the set of sub-trajectories along which each state is covered in $\mathcal{D}^1$.
Estimate ♠ from the complementary set $\mathcal{D}_1^c$:

$$\frac{\sum_{\mathbf{tr}_h \in \mathcal{D}_1^c} \mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot)=(s,a), \mathbf{tr}_h \notin \mathbf{Tr}_h^{\mathcal{D}^1}\}}{|\mathcal{D}_1^c|}$$

Estimate ♣ from trajectories $\mathcal{D}'_{\mathrm{env}}$ collected by the BC's policy:

$$\frac{\sum_{\mathbf{tr}_h \in \mathcal{D}'_{\mathrm{env}}} \mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot)=(s,a), \mathbf{tr}_h \in \mathbf{Tr}_h^{\mathcal{D}^1}\}}{|\mathcal{D}'_{\mathrm{env}}|}$$

Lemma 2

The estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ is (ε, δ)-PAC for estimating $d_h^{\pi^{\mathrm{E}}}$ if

$$m \gtrsim \frac{H^{3/2}|\mathcal{S}|}{\varepsilon}\log\left(\frac{|\mathcal{S}|H}{\delta}\right),\ n' \gtrsim \frac{H^2|\mathcal{S}|}{\varepsilon^2}\log\left(\frac{|\mathcal{S}|H}{\delta}\right).$$

(Part c) Efficient Optimization

Minimax formulation of state-action distribution matching

$$\max_{w\in\mathcal{W}} \min_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)} w_h(s,a)(\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi,\widehat{P}}(s,a)).$$

Apply online optimization methods approximately to obtain the saddle point.

- $\pi^{(t)} \leftarrow$ Solve the RL optimal policy with $\widehat{P}$ and $w^{(t)}$.
- Update $w^{(t+1)} := \mathcal{P}_{\mathcal{W}}\left(w^{(t)} - \eta^{(t)}\nabla f^{(t)}(w^{(t)})\right)$.

Lemma 3

The optimization procedure can return an ε-accurate solution if $T \gtrsim \frac{H^2|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}$.

References:

[1] Rajaraman, N., et al. "Toward the fundamental limits of imitation learning." NeurIPS, 2020.
[2] Abbeel, P., et al. "Apprenticeship learning via inverse reinforcement learning." ICML, 2004.
[3] Syed, U., et al. "A game-theoretic approach to apprenticeship learning." NeurIPS, 2007.
[4] Shani, L., et al. "Online apprenticeship learning." AAAI, 2022.
[5] Menard, P., et al. "Fast active learning for pure exploration in reinforcement learning." ICML, 2021.

Code at GitHub:

Paper at arXiv: