Nearly Minimax Optimal Adversarial Imitation Learning with Known and Unknown Transitions

Tian Xu[1], Ziniu Li[2], and Yang Yu[1]

1: National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
2: Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China
xut@lamda.nju.edu.cn, ziniuli@link.cuhk.edu.cn, yuy@nju.edu.cn

Download Paper

Problem Formulation

Problem: consider the following imitation learning problem [6, 3]:

$$\min_{\pi} V^{\pi^{\mathrm{E}}} - V^{\pi}, \tag{1}$$

where $V^{\pi^{\mathrm{E}}}$ and V^{π} are the policy values (i.e, cumulative reward) of the expert and the learner, respectively.

Setting: solve (1) with the given expert demonstrations $\mathcal{D}$ of trajectories by π^{E}:

$$\mathcal{D} = \left\{\mathbf{tr} = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H);\, a_h \sim \pi_h^{\mathrm{E}}(\cdot|s_h)\right\}.$$

Method: mimic expert behaviors from $\mathcal{D}$:

$$\min_{\pi} \psi(\pi, \pi^{\mathrm{E}}; \mathcal{D}),$$

where ψ is a certain distance measure.

Our Contribution

Table 1. Sample complexity (i.e., the number of expert trajectories) and interaction complexity (i.e., the number of interactions with environments) to achieve an ε-optimal policy value gap (i.e., $V^{\pi^{\mathrm{E}}} - V^{\pi} \le \varepsilon$). $|\mathcal{S}|$ is the state space size, $|\mathcal{A}|$ is the action space size, and H is the planning horizon. We use $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ to hide logarithmic factors.

	Known Transition Setting	Unknown Transition Setting	
	Sample Complexity	Sample Complexity	Interaction Complexity
BC [3]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	0
FEM [1]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2} + \frac{H^8\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{\varepsilon^5}\right)$	0
GTAL [5]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2} + \frac{H^6\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{\varepsilon^3}\right)$	0
OAL [4]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\varepsilon^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^4\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varepsilon^2}\right)$
MIMIC-MD [3]	$\widetilde{\mathcal{O}}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	-	-
TAIL (Ours)	$\widetilde{\mathcal{O}}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	-	-
MB-TAIL (Ours)	-	$\widetilde{\mathcal{O}}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^3\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varepsilon^2}\right)$
Lower Bound [2]	$\widetilde{\Omega}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	$\widetilde{\Omega}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\varepsilon}\right)$	-

Remark

- The proposed algorithms are the first to match the lower bound under both the known and unknown transition settings w.r.t. sample complexity.
- MB-TAIL improves the interaction complexity under the unknown transition setting.
- We deny the conjecture in [3] that "the conventional minimum distance functional (AIL) approach, $\cdots$, does not achieve the (minimax) rate".

Proposed Algorithm

Main Ingredient

A **transition-aware** estimator for the state-action distribution of the exert policy.

$$\left(\widetilde{P}_h^{\pi^{\mathrm{E}}} \circ \mathcal{D}\right)(s,a) := \underbrace{\sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{D_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(s_h,a_h) = (s,a)\right\}}_{\text{exact computation}} + \underbrace{\left(\widehat{P}_h^{\pi^{\mathrm{E}}} \circ \mathcal{D}_1^c\right)(s,a)}_{\text{maximum likelihood estimation}}. \tag{2}$$

With the estimator in (2), Transition-aware Adversarial Imitation (**TAIL**) aims to solve

$$\max_{w \in \mathcal{W}} \min_{\pi \in \Pi} f := \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S}\times\mathcal{A}} w_h(s,a)\left(\widetilde{P}_h^{\pi^{\mathrm{E}}}(s,a) - P_h^{\pi}(s,a)\right). \tag{3}$$

Algorithm 1 Transition-aware Adversarial Imitation Learning (TAIL)

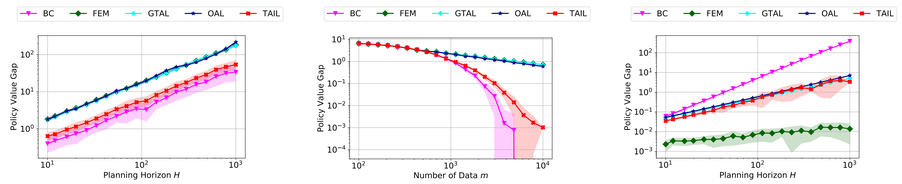
Input: expert demonstrations $\mathcal{D}$, number of iterations T, step size $\eta^{(t)}$, and initialization $w^{(1)}$.

1: Randomly split $\mathcal{D}$ into two equal parts: $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$ and obtain the estimation $\widetilde{P}_h^{\pi^{\mathrm{E}}}$ in (2).
2: **for** $t = 1, 2, \cdots, T$ **do**
3: $\pi^{(t)} \leftarrow$ solve the optimal policy with the reward function $w^{(t)}$ up to an error of $\varepsilon_{\mathrm{opt}}$.
4: Compute the state-action distribution $P_h^{\pi^{(t)}}$ for $\pi^{(t)}$.
5: Update $w^{(t+1)} := \mathcal{P}_{\mathcal{W}}\left(w^{(t)} + \eta^{(t)}\nabla f(w^{(t)})\right)$ with f defined in (3).
6: **end for**
7: Compute the mean state-action distribution $\overline{P}_h(s,a) = \sum_{t=1}^{T} P_h^{\pi^{(t)}}(s,a)/T$.
8: Derive $\overline{\pi}_h(a|s) \leftarrow \overline{P}_h(s,a)/\sum_a \overline{P}_h(s,a)$.

Output: policy $\overline{\pi}$.

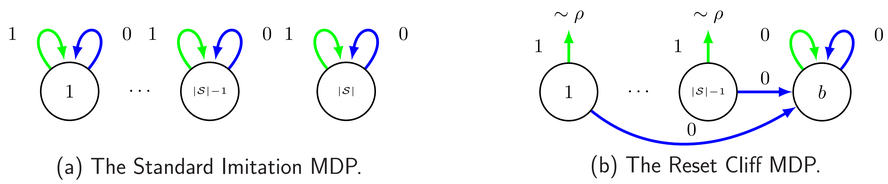
Model-based TAIL (**MB-TAIL**) further relaxes the assumption that transition function is known: MB-TAIL builds an empirical transition model and then performs the adversarial imitation learning; see the paper.

Empirical Performance



(a) On planning horizon in Standard Imitation.
(b) On sample size in Standard Imitation.
(c) On planning horizon in Reset Cliff.

Figure 1. The policy value gap (i.e., $V^{\pi^{\mathrm{E}}} - V^{\pi}$) in Standard Imitation and Reset Cliff with different number of expert demonstrations or horizons.



(a) The Standard Imitation MDP.
(b) The Reset Cliff MDP.

Figure 2. Two MDPs [3] used for comparison. Green and blue arrows indicate state transitions under the expert and non-expert actions, respectively. Digits on arrows mean reward values.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.
[2] Nived Rajaraman, Yanjun Han, Lin F. Yang, Kannan Ramchandran, and Jiantao Jiao. Provably breaking the quadratic error compounding barrier in imitation learning, optimally. *arXiv*, 2102.12948, 2021.
[3] Nived Rajaraman, Lin F. Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. In *NeurIPS*, 2020.
[4] Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. *arXiv*, 2102.06924, 2021.
[5] Umar Syed and Robert E. Schapire. A game-theoretic approach to apprenticeship learning. In *NeurIPS*, 2007.
[6] Tian Xu, Ziniu Li, and Yang Yu. Error bounds of imitating policies and environments. In *NeurIPS*, 2020.