Provably Efficient Adversarial Imitation Learning with Unknown Transitions

Tian Xu∗ , Ziniu Li∗, Yang Yu‡, Zhi-Quan Luo‡

Nanjing University, China

UAI 2023

Contributors

Tian Xu
(NJU)

Ziniu Li
(CUHKSZ)

Yang Yu
(NJU)

Zhi-Quan Luo
(CUHKSZ)

What to expect from this talk?

- **For machine learning researchers:**
  - Key principles in imitation learning (IL) algorithms.
  - The error decomposition theory (foundation of machine learning theory) for IL.
- **For reinforcement learning/imitation learning researchers:**
  - Theoretical analysis framework for adversarial imitation learning (AIL).
  - A new AIL algorithm with better theoretical guarantee.

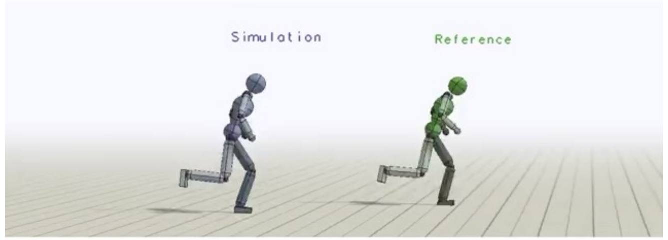
Contents

What is Imitation Learning?

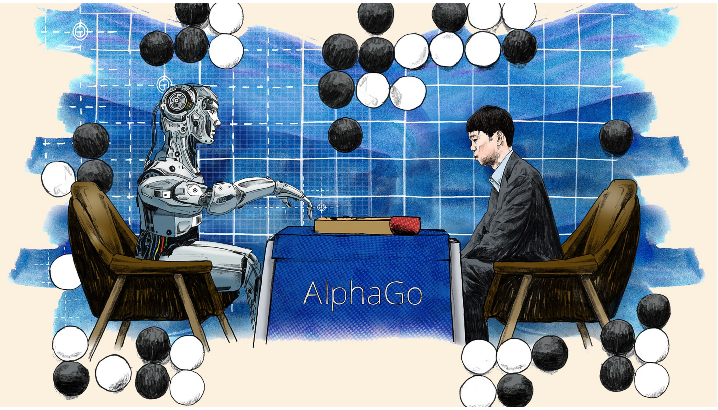
Imitation Learning (a.k.a., learning from demonstrations)

"Efficiently learn a desired behavior by imitating an expert's behavior"
[Takayuki et al., 2018]

Humanoid: Run



Policy trained to imitate a running clip.

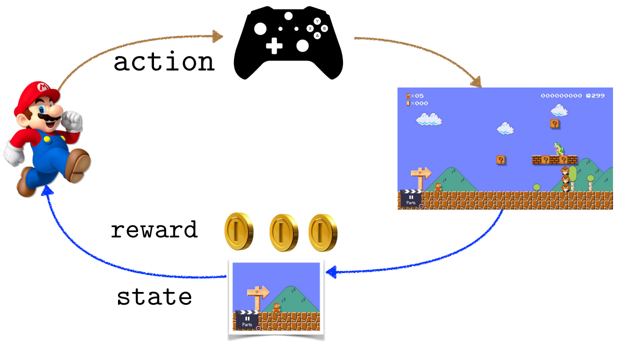


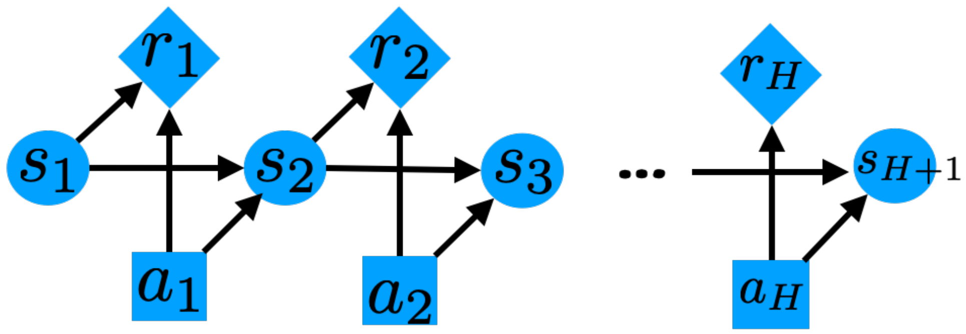
imitate to play the game Go [Silver et al., 2016]

imitate to follow instructions [OpenAI., 2023]

Markov Decision Process

- Consider a finite-horizon Markov Decision Process $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, \rho\right)$
- Policy $\pi = \{\pi_1, \cdots, \pi_H\}$ with π_h: $\mathcal{S} \to \Delta(\mathcal{A})$.
- Policy value: $V^\pi = \mathbb{E}\left[\sum_{h=1}^{H} r_h\left(s_h, a_h\right) \middle| s_1 \sim \rho; a_h \sim \pi_h\left(\cdot|s_h\right), s_{h+1} \sim P_h\left(\cdot|s_h, a_h\right)\right]$
- State (-action) distributions: $d_h^\pi(s) := \mathbb{P}(s_h = s|\pi)$, $d_h^\pi(s, a) := \mathbb{P}(s_h = s, a_h = a|\pi)$





Imitation Learning Set-up

- **Task:** Given a dataset that contains expert trajectories, the learner aims to learn a policy that matches the expert performance.
    - Expert trajectory (H state-action pairs) collected by a deterministic expert policy π^{E} :
    $$\mathrm{tr} = \{(s_1, a_1), \ldots, (s_H, a_H)\} \sim \pi^{\mathrm{E}}$$
    - Expert dataset (n expert trajectories):
    $$\mathcal{D}^{\mathrm{E}} = \{\mathrm{tr}^1, \ldots, \mathrm{tr}^n\}$$
- **Criterion (Imitation Gap):** the policy value gap between the learner $\widehat{\pi}$ and expert π^{E}.
$$V^{\pi^{\mathrm{E}}} - V^{\widehat{\pi}}$$

Adversarial Imitation Learning (AIL)

AIL mimics the expert policy via **state-action distribution matching** [Abbeel and Ng, 2004; Syed and Schapire, 2007; Ho and Ermon, 2016].

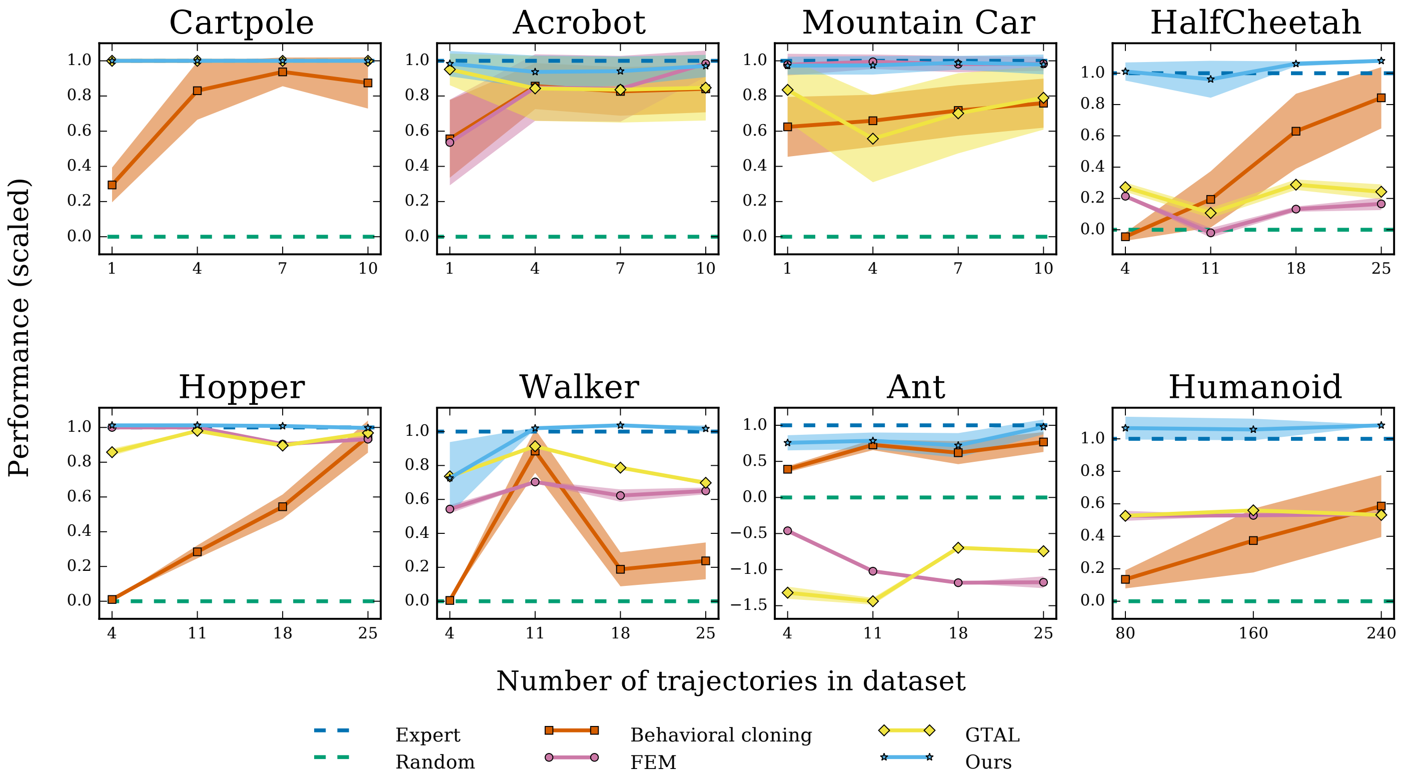
$$\min_{\pi} \sum_{h=1}^{H} \phi \left(d_h^{\pi}, \widehat{d_h^{\pi^{\mathrm{E}}}} \right)$$

- Here $\phi(\cdot,\cdot)$ is a divergence measure and $\widehat{d_h^{\pi^E}}$ is the empirical version of $d_h^{\pi^E}$.
- As $d_h^{\pi^E}$ is unknown, AIL needs to establish the empirical distribution from the **expert dataset**.
- As d_h^{π} is unknown, AIL needs to evaluate it from **environment interactions**.

Expert Sample Efficiency

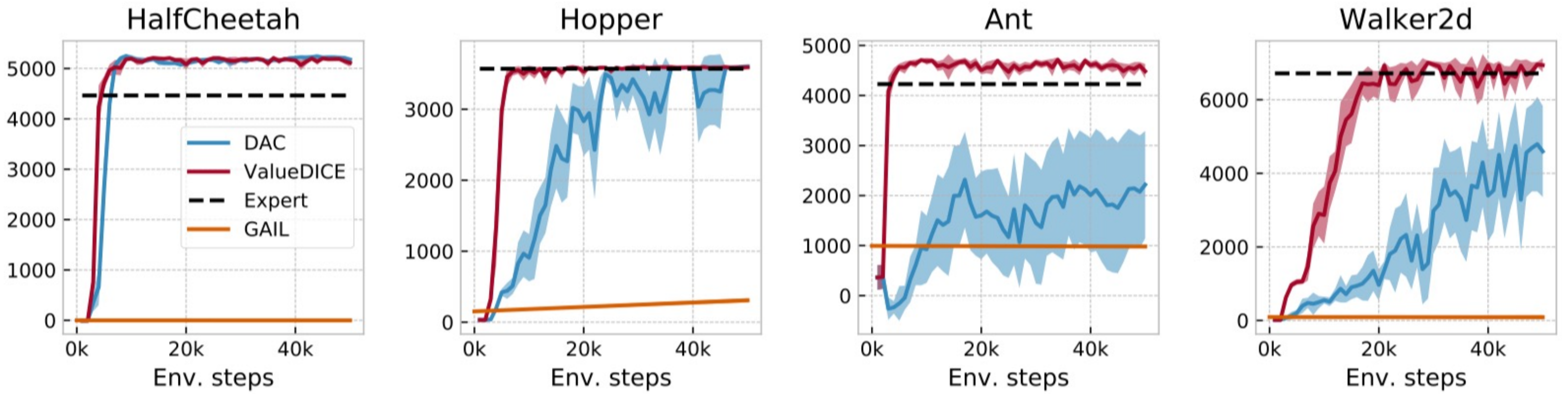
Interaction Efficiency

Empirical Observation: Expert Dataset



AIL methods (e.g., GAIL, FEM, GTAL) outperforms BC significantly in terms of **expert sample complexity**.
Figure is from [Ho and Ermon, 2016].

Empirical Observation: Environment Interaction



AIL methods (e.g., GAIL, DAC, ValueDICE) require substantial **environment interactions**. Figure is from [Kostrikov et al., 2010].

Theoretical Study of AIL with Unknown Transitions

Problem set-up:

The expert policy π^E is unknown. → The learner has access to m expert trajectories.

The transition function P is unknown. → The learner can interact with the env. for n trajectories.

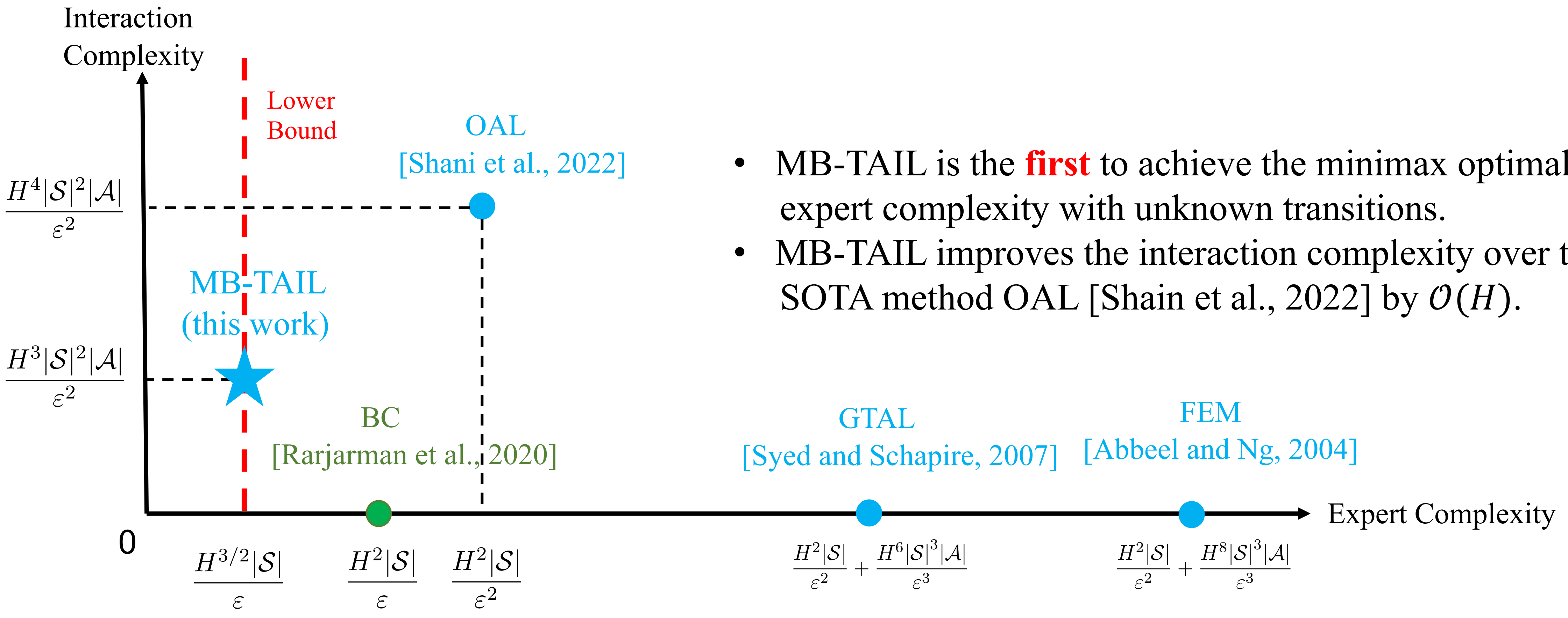
The learner aims to recover a policy $\hat{\pi}$ with small imitation gap $V^{\pi^E} - V^{\hat{\pi}}$.

Research Goal:

The expert sample complexity (m) and interaction complexity (n) required to ensure $V^{\pi^E} - V^{\hat{\pi}} \leq \varepsilon$ for a small tolerated error ε.

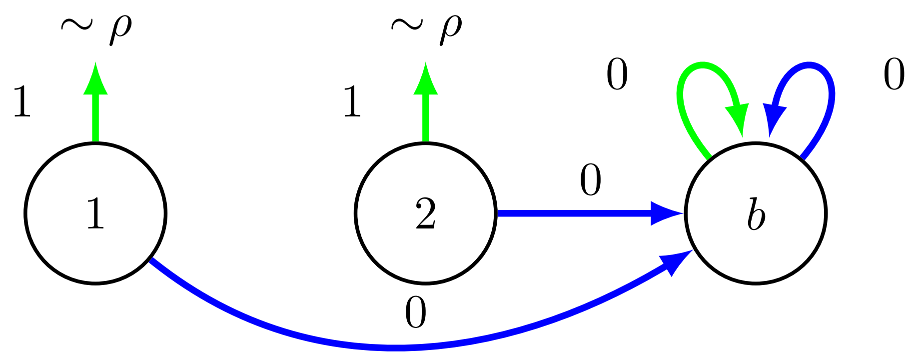
Contents

Main Results

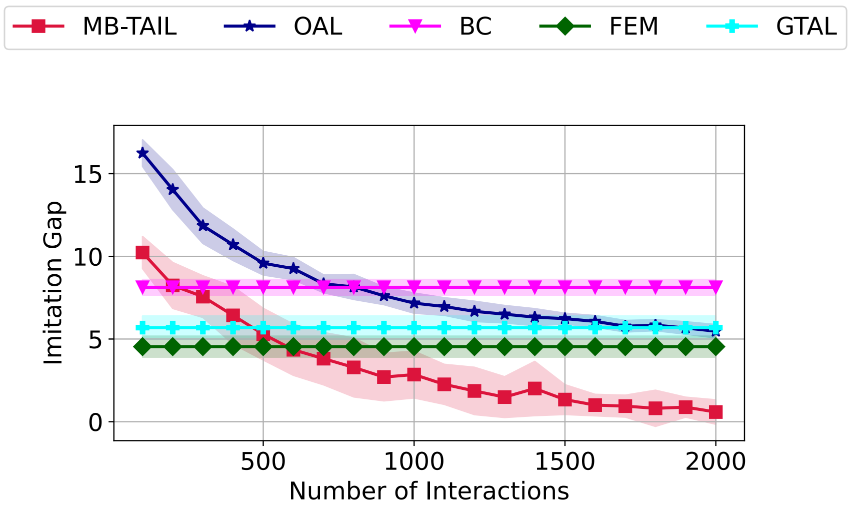


- MB-TAIL is the **first** to achieve the minimax optimal expert complexity with unknown transitions.
- MB-TAIL improves the interaction complexity over the SOTA method OAL [Shain et al., 2022] by $\mathcal{O}(H)$.

Simulation Study



Reset Cliff MDP [Rajaraman et al., 2020]



Policy value gaps of $V^{\pi^E} - V^{\hat{\pi}}$

MB-TAIL outperforms the other methods when the number of interactions exceeds 500.

Algorithmic Framework with Unknown Transitions

State-action distribution matching with total variation distance:

$$\min_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d_h^{\pi, P}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) \right|.$$

Model-based AIL:

$$\min_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d_h^{\pi, \widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) \right|.$$

Algorithm 1 Meta-algorithm for AIL with Unknown Transitions

Require: Expert demonstrations $\mathcal{D}$.

1: $\widehat{P} \leftarrow$ Invoke an exploration algorithm A to collect n trajectories and learn a transition model.

2: $\widetilde{d}_h^{\pi^{\mathrm{E}}} \leftarrow$ Apply an algorithm B to estimate the expert state-action distribution.

3: $\bar{\pi} \leftarrow$ Apply an optimization algorithm C to solve the distribution matching problem with the expert estimation $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ under transition model $\widehat{P}$.

Ensure: Policy $\bar{\pi}$.

Theoretical Analysis for Model-based AIL

Definition 1 (Uniform Policy Evaluation, UPE)

A learned transition model $\widehat{P}$ is (ε, δ)-PAC for UPE if

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

Definition 2 ($\varepsilon_{\mathrm{EST}}$-accurate Estimation)

An estimation $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ is $\varepsilon_{\mathrm{EST}}$-accurate for $d_h^{\pi^{\mathrm{E}}}$ if $\sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}} \right\|_1 \leq \varepsilon_{\mathrm{EST}}$

Definition 3 ($\varepsilon_{\mathrm{OPT}}$-optimal Policy)

A policy $\bar{\pi}$ is $\varepsilon_{\mathrm{OPT}}$-optimal for the distribution matching problem with $\widehat{P}$ and $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ if $\sum_{h=1}^{H} \left\| d_h^{\bar{\pi}, \widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 \leq \min_{\pi \in \Pi} \sum_{h=1}^{H} \left\| d_h^{\pi, \widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 + \varepsilon_{\mathrm{OPT}}$.

Error Decomposition Theory

Proposition 1 (Error Decomposition in AIL)

Suppose that

(a) an exploration algorithm A can interact with the env. and output a learned transition model $\widehat{P}$ that is $(\varepsilon_{\mathrm{EXP}}, \delta_{\mathrm{EXP}})$-PAC for UPE;

(b) an algorithm B can establish $\varepsilon_{\mathrm{EST}}$-accurate estimation $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ with probability at least $\geq 1-\delta_{\mathrm{EST}}$;

(c) with $\widehat{P}$ in (a) and $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ in (b), an algorithmic C returns an $\varepsilon_{\mathrm{OPT}}$-optimal policy $\bar{\pi}$.

Then applying algorithms A, B and C under the above framework could return a policy $\bar{\pi}$ with

$$\mathbb{P}\left(V^{\pi^{\mathrm{E}}} - V^{\bar{\pi}} \leq 2\varepsilon_{\mathrm{EXP}} + 2\varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{OPT}}\right) \geq 1 - \delta_{\mathrm{EXP}} - \delta_{\mathrm{EST}}$$

Three types of errors in AIL's training: exploration error, estimation error and optimization error.

Contents

1. Background and Problem Setup
2. Main Results
3. **Algorithmic Designs and Analysis**
4. Summary

Part (a): Controlling the Exploration Error

Applying reward-free exploration methods [Chi et al., 2021] can learn the desired transition model.

Lemma 1 (Theorem 1 of [Ménard et al., 2021])

The reward-free exploration algorithm RF-Express can learn a transition model $\widehat{P}$ that is (ε, δ)-PAC for uniform policy evaluation, if the number of trajectories collected by RF-Express satisfies

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

Connection with AIL: both RFE and AIL needs to solve RL problems with different rewards.

Part (b): Controlling the Estimation Error

- The maximum likelihood estimator (MLE) is considered in the literature [Abbeel and Ng, 2004; Syed and Schapire, 2007; Shani et al., 2022].

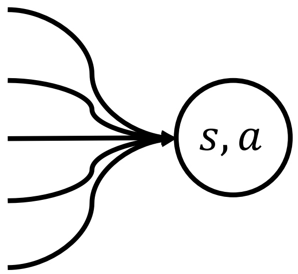
$$\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) = \frac{\sum_{\mathrm{tr}\in\mathcal{D}} \mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}}{|\mathcal{D}|}$$

where $tr_h(\cdot,\cdot)$ indicates the specific state-action pair of trajectory tr in time step h.

- To obtain an ε-accurate estimation, the expert sample complexity required by the MLE is $\widetilde{\mathcal{O}}(\frac{H^2|\mathcal{S}|}{\varepsilon^2})$ [Xu et al., 2022] while the minimax optimal one is $\widetilde{\mathcal{O}}(\frac{H^{3/2}|\mathcal{S}|}{\varepsilon})$ with known transitions [Rajaraman et al., 2020].

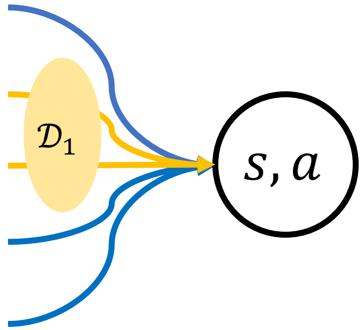
Part (b): Transition-aware Estimator

- Start with the marginal formulation:



$$d_h^{\pi^{\mathrm{E}}}(s,a) = \sum_{\mathrm{tr}_h} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\, \mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}$$

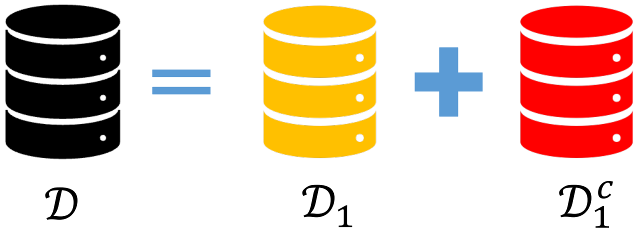
- Split the expert dataset $\mathcal{D}$ into two equal parts $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^C$ and define $\mathrm{Tr}_h^{\mathcal{D}_1}$ as the set of sub-trajectories covered in $\mathcal{D}_1$.



- Then we have the following decomposition:

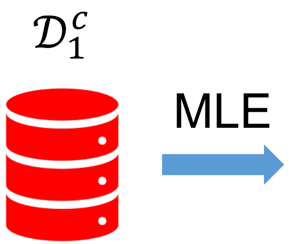
$$d_h^{\pi^{\mathrm{E}}}(s,a) = \underbrace{\sum_{\mathrm{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\, \mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}}_{:=\clubsuit} + \underbrace{\sum_{\mathrm{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\, \mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}}_{:=\spadesuit}$$

Part (b): Transition-aware Estimator



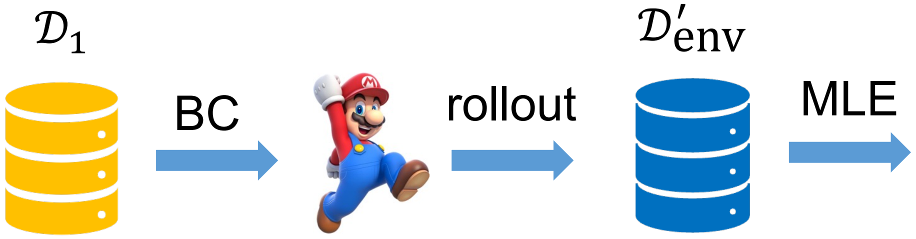
$$d_h^{\pi^{\mathrm{E}}}(s,a) = \underbrace{\sum_{\mathrm{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\, \mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}}_{:=\clubsuit} + \underbrace{\sum_{\mathrm{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\, \mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}}_{:=\spadesuit}$$

- To estimate the term $\spadesuit$, we can establish the MLE using $\mathcal{D}_1^c$.



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

- To estimate the term $\clubsuit$:



$$\frac{\sum_{\mathrm{tr}_h \in \mathcal{D}'_{\mathrm{env}}} \mathbb{I}\{\mathrm{tr}_h(\cdot,\cdot) = (s,a), \mathrm{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}\}}{|\mathcal{D}'_{\mathrm{env}}|}$$

Part (b): Transition-aware Estimator

$$d_h^{\pi^{\mathrm{E}}}(s,a) = \boxed{\sum_{\mathrm{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}} + \boxed{\sum_{\mathrm{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathrm{tr}_h)\mathbb{I}\left\{\mathrm{tr}_h(\cdot,\cdot) = (s,a)\right\}}$$

$$\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) = \boxed{\frac{\sum_{\mathrm{tr}_h \in \mathcal{D}'_{\mathrm{env}}} \mathbb{I}\{\mathrm{tr}_h(\cdot,\cdot)=(s,a), \mathrm{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}\}}{|\mathcal{D}'_{\mathrm{env}}|}} + \boxed{\frac{\sum_{\mathrm{tr}_h \in \mathcal{D}_1^c} \mathbb{I}\{\mathrm{tr}_h(\cdot,\cdot)=(s,a), \mathrm{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}\}}{|\mathcal{D}_1^c|}}$$

Lemma 2

Let $\mathcal{D}$ be the expert dataset. Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$; suppose $H \geq 5$. The transition-aware estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ is ε-accurate with probability at least $1-\delta$, if

$$|\mathcal{D}| \gtrsim \frac{H^{3/2}|\mathcal{S}|}{\varepsilon} \log\left(\frac{|\mathcal{S}|H}{\delta}\right), \; |\mathcal{D}'_{\mathrm{env}}| \gtrsim \frac{H^2|\mathcal{S}|}{\varepsilon^2} \log\left(\frac{|\mathcal{S}|H}{\delta}\right).$$

- At a high level, the new estimator utilizes the transition information from environment interactions to improve the estimation.
- The expert sample complexity matches the lower bound [Rajaraman et al., 2021] in the known transition setting in terms of H and ε.

Part (c): Controlling the Optimization Error

- Transform the original minimization problem into a minimax one via the dual form of l_1-norm.

$$\min_{\pi \in \Pi} \sum_{h=1}^{H} \left\| d_h^{\pi, \widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 \Longleftrightarrow \max_{w \in \mathcal{W}} \min_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a)} w_h(s,a) \left(\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi, \widehat{P}}(s,a) \right).$$

- Applying online gradient descent [Shalev-Shwartz et al., 2014] solves this minimax problem.
  - For w, apply online projected gradient descent with objective $\underbrace{\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} w_h(s,a) \left(d_h^{\pi^{(t)}, \widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) \right)}_{:= f^{(t)}(w)}$

    $$w^{(t+1)} := \mathcal{P}_{\mathcal{W}}(w^{(t)} - \eta^{(t)} \nabla f^{(t)}(w^{(t)}))$$

  - For π, solve the RL problem with the reward function $w^{(t+1)}$ and $\hat{P}$.

Lemma 3

The gradient-based optimization procedure can return an ε-optimal policy, if

$$T \gtrsim \frac{H^2 |\mathcal{S}||\mathcal{A}|}{\varepsilon^2}, \quad \eta^{(t)} := \sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{8T}}.$$

MB-TAIL: Putting All Together

Algorithm 2 Model-based Transition-aware AIL

Require: Expert demonstrations $\mathcal{D}$.

1: Invoke RF-Express to collect n trajectories and learn an empirical transition function $\widehat{P}$.
2: Randomly split $\mathcal{D}$ into two equal parts: $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$.
3: Learn $\pi' \in \Pi_{\mathrm{BC}}(\mathcal{D}_1)$ by BC and roll out π' to obtain dataset $\mathcal{D}'_{\mathrm{env}}$ with $|\mathcal{D}'_{\mathrm{env}}| = n'$.
4: Obtain the transition-aware estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ with $\mathcal{D}$ and $\mathcal{D}'_{\mathrm{env}}$.
5: $\bar{\pi} \leftarrow$ Apply the gradient-based optimization method with the estimation $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ under transition model $\widehat{P}$.

Ensure: Policy $\bar{\pi}$.

Theorem 1

Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Under the unknown transition setting, consider `MB-TAIL` and $\bar{\pi}$ is the output policy, if the expert sample complexity and the interaction complexity satisfy

$$m = \widetilde{\mathcal{O}}\left(\frac{H^{3/2}|\mathcal{S}|}{\varepsilon}\right), \; n' + n = \widetilde{\mathcal{O}}\left(\frac{H^3|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, we have $V^{\pi^{\mathrm{E}}} - V^{\bar{\pi}} \le \varepsilon$.

Contents

Summary

- This paper proposes a provably efficient AIL method with minimax optimal expert sample complexity and improved interaction complexity.
  - An algorithmic framework, which establishes a connection between AIL and reward-free exploration.
  - A better expert state-action distribution estimator with unknown transitions.
  - A provably efficient optimization procedure for AIL.
- We also extend MB-TAIL to the function approximation setting and prove that it can achieve expert sample and interaction complexity free of $|\mathcal{S}|$, showing its generalization ability.

Paper: https://arxiv.org/abs/2306.06563

Code: https://github.com/tianxusky/tabular-ail