Provably Efficient Adversarial Imitation Learning with Unknown Transitions

Tian Xu[*1, 4] **Ziniu Li**[*2, 3] **Yang Yu**[†1, 4] **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

Abstract

Imitation learning (IL) has proven to be an effective method for learning good policies from expert demonstrations. Adversarial imitation learning (AIL), a subset of IL methods, is particularly promising, but its theoretical foundation in the presence of unknown transitions has yet to be fully developed. This paper explores the theoretical underpinnings of AIL in this context, where the stochastic and uncertain nature of environment transitions presents a challenge. We examine the expert sample complexity and interaction complexity required to recover good policies. To this end, we establish a framework connecting reward-free exploration and AIL, and propose an algorithm, MB-TAIL, that achieves the minimax optimal expert sample complexity of $\widetilde{\mathcal{O}}(H^{3/2}|\mathcal{S}|/\varepsilon)$ and interaction complexity of $\widetilde{\mathcal{O}}(H^3|\mathcal{S}|^2|\mathcal{A}|/\varepsilon^2)$. Here, H represents the planning horizon, $|\mathcal{S}|$ is the state space size, $|\mathcal{A}|$ is the action space size, and ε is the desired imitation gap. MB-TAIL is the first algorithm to achieve this level of expert sample complexity in the unknown transition setting and improves upon the interaction complexity of the best-known algorithm, OAL, by $\mathcal{O}(H)$. Additionally, we demonstrate the generalization ability of MB-TAIL by extending it to the function approximation setting and proving that it can achieve expert sample and interaction complexity independent of $|\mathcal{S}|$.

1 INTRODUCTION

In real-life scenarios, sequential decision-making tasks are ubiquitous, where agents devise policies to maximize the long-term return. Reinforcement learning (RL) [Sutton and Barto, 2018] is a popular paradigm for learning effective policies through trial and error in unknown environments. However, RL often requires a large amount of samples and laborious reward engineering to achieve satisfactory performance in practice. Alternatively, imitation learning (IL) [Argall et al., 2009, Osa et al., 2018] provides a more sample-efficient approach to policy optimization by directly learning from expert demonstrations, and has been proven successful in various applications [Levine et al., 2016, Shi et al., 2019, Jang et al., 2022]. By leveraging existing expert knowledge, IL methods enable efficient policy learning in situations where RL might be infeasible or expensive. Therefore, IL has become an increasingly popular and practical alternative for real-world applications.

Imitation learning (IL) is a framework that aims to minimize the difference between the expert policy and the imitated policy [Ross and Bagnell, 2010, Xu et al., 2020, Rajaraman et al., 2020]. The two prominent IL methods are behavioral cloning (BC) [Pomerleau, 1991, Ross and Bagnell, 2010] and adversarial imitation learning (AIL) [Abbeel and Ng, 2004, Syed and Schapire, 2007, Ziebart et al., 2008, Ho and Ermon, 2016]. BC employs supervised learning to minimize the discrepancy between the policy distribution of the imitated policy and the expert policy. On the other hand, AIL focuses on state-action distribution matching, where the learner estimates an adversarial reward function that maximizes the policy value gap and then learns a policy to minimize the gap with the inferred reward function through a min-max optimization. Practical algorithms that build upon these principles have been developed and applied to various domains [Torabi et al., 2018, Fu et al., 2018, Ke et al., 2019, Kostrikov et al., 2019, Brantley et al., 2020, Garg et al., 2021, Dadashi et al., 2021, Viano et al., 2022].

A remarkable observation from empirical studies [Ho and Ermon, 2016, Kostrikov et al., 2019, Ghasemipour et al., 2019] is that adversarial imitation learning (AIL) often outperforms behavioral cloning (BC) by a significant margin. This phenomenon has spurred numerous theoretical investi-

[*]Equal contribution. Author ordering is determined randomly using a coin flip.

[†]Corresponding author.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

gations [Zhang et al., 2020, Wang et al., 2020b, Rajaraman et al., 2020, 2021a, Xu et al., 2020, Liu et al., 2022, Xu et al., 2022] aimed at understanding the mechanisms of AIL. However, analyzing AIL is challenging because both the expert policy and environment transitions are unknown, making expert estimation and policy optimization/evaluation inaccurate. The complex min-max implementation of AIL further compounds the theoretical analysis difficulty. As a result, several prior works [Abbeel and Ng, 2004, Syed et al., 2008, Rajaraman et al., 2020, 2021a, Xu et al., 2022] have made the simplifying assumption of a known transition function to facilitate the analysis.

However, the characterization of environment transitions is often challenging in practical tasks, as noted in previous studies [Duan et al., 2016, Shi et al., 2019]. Therefore, there has been growing interest in investigating AIL with unknown transitions, where the learner does not have prior knowledge of the transition function but can collect trajectories by interacting with the environment. This setup is widely used in empirical studies [Ho and Ermon, 2016, Fu et al., 2018, Ke et al., 2019, Kostrikov et al., 2019, Brantley et al., 2020, Garg et al., 2021, Li et al., 2022]. From a theoretical perspective, it is important to understand both the expert sample complexity (i.e., the number of trajectories collected by the expert) and the interaction complexity (i.e., the number of trajectories collected by the online learner) to achieve good policies, as these are of practical interest. In this paper, we investigate AIL with unknown transitions and focus on analyzing the required expert sample and interaction complexity.

Compared with the progress made in IL with known transitions, AIL with unknown transitions still lacks a well-developed theoretical foundation. Earlier works, such as FEM [Abbeel and Ng, 2005] and GTAL [Syed and Schapire, 2007], estimated the transition function from expert demonstrations for imitation, rendering their algorithms impractical due to the prohibitively large expert sample complexity (as shown in Table 1). To the best of our knowledge, the online apprenticeship learning (OAL) algorithm in [Shani et al., 2022] is a promising approach that updates the policy and reward function using no-regret algorithms during environment interaction. In particular, OAL achieves an expert sample complexity $\widetilde{\mathcal{O}}(H^2|\mathcal{S}|/\varepsilon^2)$ and interaction complexity $\widetilde{\mathcal{O}}(H^4|\mathcal{S}|^2|\mathcal{A}|/\varepsilon^2)$[1], where $|\mathcal{S}|$ and $|\mathcal{A}|$ are the state and action space sizes, H is the planning horizon, and $\varepsilon = V^{\pi^{\mathrm{E}}} - V^{\pi}$ is the desired imitation gap. However, even with infinite environment interactions, OAL's expert sample complexity is sub-optimal, as the best expert sample complexity in the known transition setting is $\widetilde{\mathcal{O}}(H^{3/2}|\mathcal{S}|/\varepsilon)$ [Rajaraman et al., 2020]. Thus, improving AIL with unknown transitions is a significant area of research.

Table 1: Expert sample complexity and interaction complexity of BC [Rajaraman et al., 2020], FEM [Abbeel and Ng, 2004], GTAL [Syed and Schapire, 2007], OAL [Shani et al., 2022], and MB-TAIL (ours) with unknown expert and transitions. We use $\widetilde{\mathcal{O}}$ to hide logarithmic factors.

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

Contribution. This paper presents a new and general framework (Algorithm 1) that overcomes the challenge of unknown transitions and unknown expert policies. At a high level, our framework establishes a connection between AIL and reward-free exploration (RFE) [Jin et al., 2020, Ménard et al., 2021, Chen et al., 2022], which is an emerging topic in online RL. We prove that any effective AIL algorithm that works with known transitions can be transferred to the unknown transition setting using an efficient RFE method, as shown in Proposition 1.

Further, we also introduce a new algorithm called MB-TAIL[2], which incorporates recent advances in AIL with known transitions and RFE. MB-TAIL builds on MIMIC-MD [Rajaraman et al., 2020] and RF-Express [Ménard et al., 2021] but requires new designs to apply their main ideas in the unknown transition setting. Notably, MB-TAIL achieves an expert sample complexity of $\widetilde{\mathcal{O}}(H^{3/2}|\mathcal{S}|/\varepsilon)$, meeting the lower bound $\Omega(H^{3/2}/\varepsilon)$ [Rajaraman et al., 2021b] in H and ε. This sample complexity is nearly minimax optimal and the first to be achieved in the unknown transition setting. Additionally, MB-TAIL has an interaction complexity of $\widetilde{\mathcal{O}}(H^3|\mathcal{S}|^2|\mathcal{A}|/\varepsilon^2)$, which improves upon the best-known OAL algorithm by a factor of $\mathcal{O}(H)$.

Finally, we extend the MB-TAIL algorithm to the function approximation setting and demonstrate its ability to achieve the expert sample and interaction complexity independent of the state space size $|\mathcal{S}|$. Specifically, we investigate the case of state abstraction [Li et al., 2006], which involves approximating functions using piecewise constant functions. By employing appropriate state abstractions, MB-TAIL can estimate the abstract state-action distribution instead of the tabular counterpart, which is crucial for generalization.

[1] In [Shani et al., 2022], a regret $\widetilde{\mathcal{O}}(\sqrt{H^4|\mathcal{S}|^2|\mathcal{A}|K} + \sqrt{H^3|\mathcal{S}||\mathcal{A}|K^2/m})$ is proved, where K is the number of interaction episodes and m is the number of expert trajectories. We convert this regret guarantee to the sample complexity guarantee (see the Appendix).

[2] MB-TAIL stands for model-based transition-aware adversarial imitation learning.

2 RELATED WORK

In the realm of AIL with known transitions, there have been numerous theoretical investigations into expert sample complexity [Abbeel and Ng, 2004, Syed and Schapire, 2007, Zahavy et al., 2020, Rajaraman et al., 2020, Swamy et al., 2022, Xu et al., 2021, 2022]. For example, FEM and GTAL, which are traditional AIL algorithms, have expert sample complexity of $\widetilde{\mathcal{O}}(H^2|\mathcal{S}|/\varepsilon^2)$ [3]. This upper bound is proven to be tight in the worst-case [Xu et al., 2022, Swamy et al., 2022]. Additionally, Rajaraman et al. [2020] proposed a novel AIL technique, MIMIC-MD, which leverages the transition function to obtain an enhanced expert sample complexity of $\widetilde{\mathcal{O}}(H^{3/2}|\mathcal{S}|/\varepsilon)$. MIMIC-MD meets the information-theoretic lower bound of expert sample complexity with known transitions, which is $\widetilde{\Omega}(H^{3/2}/\varepsilon)$ [Rajaraman et al., 2021b], in terms of both H and ε. Recently, horizon-free expert sample complexity was studied in [Xu et al., 2022], which explains the superior performance of AIL with known transitions. However, there are only a limited number of theoretical investigations into AIL with unknown transitions. We have already discussed these in the previous section and thus will not repeat them here.

Our research establishes a connection between adversarial imitation learning and reward-free exploration, which is an emerging area of interest in online reinforcement learning. The reward-free exploration framework was introduced in [Jin et al., 2020] with two primary goals: 1) isolating the exploration and planning problems within a standard RL framework and 2) learning an environment that is robust enough to cover all possible training scenarios. Since then, several advances have been made in this field [Kaufmann et al., 2021, Wang et al., 2020a, Zhang et al., 2021, Chen et al., 2022]. Specifically, [Ménard et al., 2021] achieved the minimax rate in the tabular setting.

It is worth noting that AIL is closely related to inverse reinforcement learning (IRL) [Ng and Russell, 2000], which aims to infer the ground truth reward function from expert demonstrations. Recent works in IRL include [Metelli et al., 2021], which studied the error propagation of the obtained policy's performance when transferring the reward function to a new environment, and [Zeng et al., 2022], which developed a single-loop algorithm to recover the reward function under the maximum entropy IRL formulation. Additionally, [Lindner et al., 2022] proposed an upper confidence approach that actively explores the environment and expert policy to learn the reward function. However, our focus differs from these studies as our goal is to solve the imitation learning problem by learning a high-quality policy, rather than inferring the reward function.

[3] Results from [Abbeel and Ng, 2004, Syed and Schapire, 2007] are transformed from the infinite-horizon setting to the episodic setting by 1) substituting the effective planning horizon $1/(1-\gamma)$ with the finite planning horizon H; 2) instantiating the linear feature with the one-hot feature under the tabular setting.

3 BACKGROUND

Episodic Markov Decision Process. In this paper, we consider episodic Markov decision process (MDP), which can be described by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, \rho)$. Here $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively. H is the planning horizon and ρ is the initial state distribution. $P = \{P_1, \cdots, P_H\}$ specifies the non-stationary transition function of this MDP; concretely, $P_h(s_{h+1}|s_h, a_h)$ determines the probability of transiting to state s_{h+1} conditioned on state s_h and action a_h at time step h, for $h \in [H]$, where $[x]$ denotes the set of integers from 1 to x. Similarly, $r = \{r_1, \cdots, r_H\}$ specifies the reward function of this MDP; without loss of generality, we assume that $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$, for $h \in [H]$. A non-stationary policy $\pi = \{\pi_1, \cdots, \pi_h\}$ with $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the probability simplex and $\pi_h(a|s)$ gives the probability of selecting action a on state s at time step h, for $h \in [H]$.

The sequential decision process runs as follows: at the beginning of an episode, the environment is reset to an initial state according to ρ; then the agent observes a state s_h and takes an action a_h based on $\pi_h(a_h|s_h)$; consequently, the environment makes a transition to the next state s_{h+1} according to $P_h(s_{h+1}|s_h, a_h)$ and sends a reward $r_h(s_h, a_h)$ to the agent. This episode ends after H repeats.

The quality of a policy is measured by its *policy value* (i.e., the expected long-term return):

$$V^\pi = \mathbb{E}\Big[\sum_{h=1}^{H} r_h(s_h, a_h)|s_1 \sim \rho; a_h \sim \pi_h(\cdot|s_h), \\ s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]\Big].$$

To facilitate later analysis, we introduce the state-action distribution induced by a policy π:

$$d_h^\pi(s, a) = \mathbb{P}(s_h = s, a_h = a|s_1 \sim \rho; a_\ell \sim \pi_h(\cdot|s_\ell), \\ s_{\ell+1} \sim P_\ell(\cdot|s_\ell, a_\ell),\ \forall \ell \in [h]).$$

In other words, $d_h^\pi(s, a)$ qualifies the visitation probability of state-action pair (s, a) at time step h. In this way, we get an equivalent dual form of the policy value [Puterman, 2014]:

$$V^\pi = \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\pi(s, a) r_h(s, a), \tag{1}$$

which will be used in later analysis.

Imitation Learning. The goal of IL is to learn a high quality policy *without* the environment reward function. To this end, we often assume there is a nearly optimal expert policy π^{E}

that could interact with the environment to generate a dataset (i.e., m trajectories of length H):

$$\mathcal{D} = \{\mathbf{tr} = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H); s_1 \sim \rho; \\ a_h \sim \pi_h^{\mathrm{E}}(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]\}.$$

Then, the learner can use the dataset $\mathcal{D}$ to mimic the expert and to obtain a good policy. The quality of imitation is measured by the *imitation gap* [Abbeel and Ng, 2004, Ross and Bagnell, 2010, Rajaraman et al., 2020]: $V^{\pi^{\mathrm{E}}} - V^{\pi}$, where π is the learned policy. That is, we hope the learned policy can perfectly imitate the expert such that the imitation gap is small. In this paper, we assume the expert policy is deterministic, which is common in the literature [Rajaraman et al., 2020, Swamy et al., 2022, Xu et al., 2022].

Notation. We denote Π as the set of all stochastic policies for the learner. Furthermore, $|\mathcal{D}|$ is the number of trajectories in $\mathcal{D}$. We reserve the symbol m to denote the number of expert trajectories. We write $a(n) \gtrsim b(n)$ if there exist constants $C > 0, n_0 \geq 1$ such that $a(n) \geq Cb(n)$ for $n \geq n_0$.

4 WARM-UP: AIL WITH KNOWN TRANSITIONS

To imitate the expert policy, AIL methods solve the state-action distribution matching problem [Ho and Ermon, 2016, Ke et al., 2019, Xu et al., 2020]. As an introduction to general readers, we consider the known transition setting in this section. Our starting point is the following state-action distribution matching problem:

$$\min_{\pi \in \Pi} \sum_{h=1}^{H} \left\| d_h^{\pi} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 . \tag{2}$$

where $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ is an estimation of the expert state-action distribution $d_h^{\pi^{\mathrm{E}}}$. We can explain why Equation (2) is a good learning objective with the following two definitions.

Definition 1. *An estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ is said to be $\varepsilon_{\mathrm{EST}}$-accurate for $d_h^{\pi^{\mathrm{E}}}$ if $\sum_{h=1}^{H} \|\widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}}\|_1 \leq \varepsilon_{\mathrm{EST}}$.*

Definition 2. *For optimization problem (2), a policy $\overline{\pi}$ is said to be $\varepsilon_{\mathrm{OPT}}$-optimal if $\sum_{h=1}^{H} \|d_h^{\overline{\pi}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\|_1 \leq \min_{\pi \in \Pi} \sum_{h=1}^{H} \|d_h^{\pi} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\|_1 + \varepsilon_{\mathrm{OPT}}$.*

Lemma 1. *Given an $\varepsilon_{\mathrm{EST}}$-accurate estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}$, suppose that $\overline{\pi}$ is $\varepsilon_{\mathrm{OPT}}$-optimal for problem (2), then we have that $V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} \leq \varepsilon_{\mathrm{OPT}} + 2\varepsilon_{\mathrm{EST}}$.*

Proof of Lemma 1 can be found in the Appendix along with other theoretical results. This lemma establishes a strong theoretical foundation for state-action distribution matching. It is worth noting that similar versions of this lemma have been presented in prior works such as [Syed and Schapire, 2007, Rajaraman et al., 2020]. We will discuss how to control estimation and optimization errors in the next section.

While significant theoretical progress has been made in the known transition setting, this assumption is not always practical in real-world applications where the transition function is unknown. In such cases, empirical studies have been carried out under the unknown transition setting, where the interaction with environments is allowed but the analytic form of transition function is not available. In addition to expert sample complexity, the interaction complexity is also of great interest in this scenario, which we will explore in the next section.

5 MAIN RESULTS: AIL WITH UNKNOWN TRANSITIONS

In this section, we consider the unknown transition setting where d_h^{π} is not accessible, rendering the learning objective in Equation (2) inapplicable. A sound solution is to replace d_h^{π} with its estimated version $\widehat{d}_h^{\pi}$ in Equation (2). We highlight that the unknown transition leads to the exploration-and-exploitation trade-off, which is shared with online RL [Agarwal et al., 2022]. The prior work OAL addresses this challenge by an optimistic estimation of the value function [Shani et al., 2022].

In this paper, we explore an alternative model-based approach: we first learn the transition function from collected trajectories and subsequently estimate d_h^{π} based on the recovered transition model. The key challenge is how to recover a good transition model such that policy evaluation/optimization can be conducted accurately. To this end, we propose a general algorithmic framework, which connects AIL with reward-free exploration (or RFE for short) [Jin et al., 2020, Ménard et al., 2021], which is an emerging topic in online RL. Under this framework, a proper AIL algorithm that works under the known transition setting could be transferred to the unknown transition setting by leveraging an efficient RFE method. Before presenting the details of our framework, we formally introduce RFE.

Definition 3 ([Ménard et al., 2021])**.** *Given an MDP $\mathcal{M}$ without reward function r, an algorithm is said to be (ε, δ)-PAC for reward-free exploration (RFE) if*

$$\mathbb{P}\big(\textit{for any reward function } r, |V^{\pi_r^*} - V^{\widehat{\pi}_r^*}| \leq \varepsilon\big) \geq 1 - \delta,$$

where π_r^* is the optimal policy in the MDP with the reward function r, and $\widehat{\pi}_r^*$ is the optimal policy in the MDP with the learned transition model $\widehat{P}$ by RFE and reward function r.

By algorithmic designs, RFE methods usually satisfy the so-called uniform policy evaluation property, which is crucial for the discussion of AIL.

Definition 4. *Given an MDP $\mathcal{M}$ without reward function r, an algorithm is said to be (ε, δ)-PAC for uniform policy evaluation if*

$$\mathbb{P}\big(\text{for any reward function } r \text{ and policy } \pi, \\ |V^{\pi,P,r} - V^{\pi,\widehat{P},r}| \le \varepsilon\big) \ge 1-\delta,$$

where $V^{\pi,P,r}$ and $V^{\pi,\widehat{P},r}$ are the policy values of policy π with reward function r under the real transition model P and recovered transition model $\widehat{P}$, respectively.

Examples of algorithms that satisfy Definition 4 include RF-RL-Explore [Jin et al., 2020] (see their Lemma 3.6), RF-UCRL [Kaufmann et al., 2021] (see their Lemma 1 and the stopping rule) and RF-Express in [Ménard et al., 2021] (see their Lemma 1 and the stopping rule).

Definition 4 is connected with AIL in the following way:

$$\begin{aligned}
&\sum_{h=1}^{H} \|\widehat{d}_h^{\pi} - d_h^{\pi}\|_1 \\
&= \max_{w\in\mathcal{W}} \sum_{h=1}^{H} \sum_{(s,a)} w_h(s,a)(\widehat{d}_h^{\pi}(s,a) - d_h^{\pi}(s,a)) \\
&= \max_{w\in\mathcal{W}} V^{\pi,\widehat{P},w} - V^{\pi,P,w} \le \varepsilon.
\end{aligned}$$

Here the first equality follows the dual representation of ℓ_1-norm, and $\mathcal{W} = \{w : \|w\|_\infty \le 1\}$ is the unit ball. The second equality follows Equation (1). The last inequality follows Definition 4. In plain language, the above formula shows that we can get an accurate estimation of d_h^{π}, based on the recovered model by RFE.

Based on the above relation, with a transition model learned by RFE, AIL can be implemented as if this empirical transition function were the same as the true transition function. More specifically, the state-action distribution matching problem Equation (2) becomes

$$\min_{\pi\in\Pi} \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi,\widehat{P}} \right\|_1 \tag{3}$$

where $d_h^{\pi,\widehat{P}}$ is the state-action distribution of policy π with the transition model $\widehat{P}$. We outline the whole procedure in Algorithm 1 and the theoretical guarantee is provided below.

Proposition 1. *Suppose that*

(a) a reward-free exploration algorithm A satisfies the uniform policy evaluation property (see Definition 4) up to an error $\varepsilon_{\mathrm{RFE}}$ with probability at least $1-\delta_{\mathrm{RFE}}$;

(b) an algorithm B has a state-action distribution estimator for $d_h^{\pi^{\mathrm{E}}}$, which satisfies $\sum_{h=1}^{H} \|\widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}}\|_1 \le \varepsilon_{\mathrm{EST}}$, with probability at least $1-\delta_{\mathrm{EST}}$;

(c) with the transition model in (a) and the estimator in (b), an algorithm C solves the optimization problem in Equation (3) up to an error $\varepsilon_{\mathrm{OPT}}$.

Then applying algorithms A, B and C under the framework in Algorithm 1 could return a policy $\overline{\pi}$, which has a policy value gap (i.e., $V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}}$) at most $2\varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}$, with probability at least $1-\delta_{\mathrm{EST}}-\delta_{\mathrm{RFE}}$.

Algorithm 1 Meta-algorithm for AIL with Unknown Transitions

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

1: $\widehat{P} \leftarrow$ Invoke a reward-free exploration method to collect n trajectories and learn a transition model.

2: $\widetilde{d}_h^{\pi^{\mathrm{E}}} \leftarrow$ Estimate the expert state-action distribution.

3: $\overline{\pi} \leftarrow$ Apply an AIL approach to perform imitation with the expert estimation $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ under transition model $\widehat{P}$.

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

Next, we show how to substantiate the framework in Algorithm 1 with detailed procedures. We will consider the tabular formulation, where the space of parameterized value functions spans all possible functions. In this scenario, expert policies and reward functions are realizable. We discuss how to control $\varepsilon_{\mathrm{RFE}}$, $\varepsilon_{\mathrm{EST}}$, and $\varepsilon_{\mathrm{OPT}}$ in a sequential order.

5.1 CONTROLLING REWARD-FREE EXPLORATION ERROR

To ensure that condition (a) in Proposition 1 is satisfied, we make use of the RF-Express algorithm, as described in [Ménard et al., 2021]. This advanced algorithm allows us to control $\varepsilon_{\mathrm{RFE}}$ effectively. Below, we provide the theoretical property of RF-Express.

Lemma 2 (Theorem 1 in [Ménard et al., 2021])**.** *Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Consider the* RF-Express *algorithm (see Algorithm 1 in Appendix) and $\widehat{P}$ is the empirical transition function built on the collected trajectories, if the number of trajectories collected by* RF-Express *satisfies*

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

Then with probability at least $1-\delta$, for any policy π and any bounded reward function r between $[-1,1]$, we have $|V^{\pi,P,r} - V^{\pi,\widehat{P},r}| \le \varepsilon/2$; furthermore, for any bounded reward function r between $[-1,1]$, we have $\max_{\pi\in\Pi} V^{\pi,P,r} \le V^{\widehat{\pi}_r^*,P,r} + \varepsilon$, where $\widehat{\pi}_r^*$ is the optimal policy under the empirical transition function $\widehat{P}$ with reward function r.

5.2 CONTROLLING EXPERT STATE-ACTION DISTRIBUTION ESTIMATION ERROR

In this part, we talk about how to control the expert state-action distribution estimation error. Quite often, the maximum likelihood estimator (MLE) is considered in the literature [Abbeel and Ng, 2004, Syed and Schapire, 2007, Shani et al., 2022]. Mathematically, MLE counts how frequently a state-action pair appears in the observed expert trajectories:

$$\widehat{d}_h^{\pi^{\mathrm{E}}}(s,a) = \frac{\sum_{\mathbf{tr}\in\mathcal{D}} \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot) = (s,a)\right\}}{|\mathcal{D}|}, \tag{4}$$

where $\mathbf{tr}_h(\cdot,\cdot)$ indicates the specific state-action pair of trajectory $\mathbf{tr}$ in time step h. The sample complexity of MLE is well-known.

Lemma 3 (Rajaraman et al. [2020]). *Fix $\varepsilon \in (0, H)$ and $\delta \in (0,1)$, if the number of expert trajectories in $\mathcal{D}$ satisfies*

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

then with probability at least $1-\delta$, we have $\sum_{h=1}^{H}\|\widehat{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}}\|_1 \le \varepsilon$.

The above sample complexity of MLE is tight in the worst case; see, e.g., [Kamath et al., 2015, Lemma 8]. Though MLE can be implemented under our framework, this estimator cannot lead to the minimax optimal expert sample complexity $\Theta(H^{3/2}|\mathcal{S}|/\varepsilon)$. To address this issue, in light of [Rajaraman et al., 2020], we develop a new estimator. For a better presentation, let us introduce the following notations.

- Similar to $\mathbf{tr}_h(\cdot,\cdot)$, $\mathbf{tr}_h(\cdot)$ indicates the specific state of trajectory $\mathbf{tr}$ in time step h.
- Without $(\cdot)$ or $(\cdot,\cdot)$, $\mathbf{tr}_h$ is the truncated version of trajectory $\mathbf{tr}$ up to time step h, i.e., $\mathbf{tr}_h = (s_1, a_1, \cdots, s_h, a_h)$.
- $\mathcal{S}_h(\mathcal{D}) = \{s : \exists \mathbf{tr} \in \mathcal{D} \text{ such that } s = \mathbf{tr}_h(\cdot)\}$ is the set of states visited at time step h in $\mathcal{D}$.
- $\mathrm{Tr}_h^{\mathcal{D}} = \{\mathbf{tr}_h = (s_1, a_1, \ldots, s_h, a_h) : s_\ell \in \mathcal{S}_\ell(\mathcal{D}), \forall \ell \in [h]\}$ is the set of truncated trajectories (that may not appear in $\mathcal{D}$), along which each state has been visited in $\mathcal{D}$ up to time step h.

From the definition of state-action distribution, we have

$$\begin{aligned} d_h^{\pi}(s,a) &= d_h^{\pi}(s)\pi_h(a|s) \\ &= \Big[\sum_{s',a'} d_{h-1}^{\pi}(s',a')P_{h-1}(s|s',a')\Big]\pi_h(a|s) \end{aligned} \tag{5}$$

This equation offers another perspective on visitation probability: $d_h^{\pi}(s,a)$ represents the weighted average of flows. Specifically, each flow path is determined by ancestral state-action sequences that lead to the target state-action pair (s,a), and the weight of this flow is influenced by both the transition probability and the policy distribution.

However, when dealing with a finite sample regime, only a subset of trajectories executed by the expert policy is observed, while others remain unobserved. We can use the transition function to calculate the visitation probability for the observed trajectories, but we require statistical estimation for the non-observed ones. This idea has been exploited in [Rajaraman et al., 2020] in the known transition setting.

Now, consider the dataset $\mathcal{D}$ is randomly divided into two equal parts, i.e., $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$ and $\mathcal{D}_1 \cap \mathcal{D}_1^c = \emptyset$ with $|\mathcal{D}_1| = |\mathcal{D}_1^c| = m/2$. We have the following decomposition:

$$\begin{aligned} d_h^{\pi^{\mathrm{E}}}(s,a) = &\underbrace{\sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot) = (s,a)\right\}}_{:=\clubsuit} \\ &+ \underbrace{\sum_{\mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot) = (s,a)\right\}}_{:=\spadesuit}, \end{aligned} \tag{6}$$

where $\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)$ is the probability of the truncated trajectory $\mathbf{tr}_h$ induced by the deterministic expert policy π^{E}. As we have mentioned, if the transition function is known, we can calculate $\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)$ directly: $\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h) = \rho(s_1)\prod_{\ell=1}^{h-1} P_\ell(s_{\ell+1}|s_\ell, a_\ell)$ with $\mathbf{tr}_h = (s_1, a_1, \cdots, s_h, a_h)$

We explain two terms in Equation (6) separately. On the one hand, term $\clubsuit$ can be calculated exactly if we know both the transition function and $\mathcal{D}_1$, as explained previously. However, this is not applicable in our case as the transition function is unknown. We will discuss how to deal with this trouble later. On the other hand, term $\spadesuit$ accounts for non-observed trajectories, which is not easy to compute (because we have no clue about expert actions on non-observed states). To address this issue, Rajaraman et al. [2020] proposed to use trajectories in $\mathcal{D}_1^c$ to make a maximum likelihood estimation. This is because, $\mathcal{D}_1^c$ is statistically independent of $\mathcal{D}_1$ and therefore can be viewed as a new dataset. We follow the approach in [Rajaraman et al., 2020] to estimate term $\spadesuit$.

Now, we explain how to estimate term $\clubsuit$ in the unknown transition setting. Our solution has two steps. The first step is to apply BC on $\mathcal{D}_1$ to learn policy π':

$$\pi'_h(a|s) = \begin{cases} \frac{n_h^1(s,a)}{n_h^1(s)} & \text{if } n_h^1(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases}$$

Here $n_h^1(s,a)$ ($n_h^1(s)$) is the number of state-action (state) pairs that appeared in $\mathcal{D}_1$ in step h. This step recovers the expert behaviors on visited states in $\mathcal{D}_1$. The second step is to let π' interact with the environment to collect a new dataset $\mathcal{D}'_{\mathrm{env}}$, from which we can estimate term $\clubsuit$ by MLE.

To get a better sense, we mention that the uncertainty of estimating term ♣ comes from the transition function, rather than the expert policy. Furthermore, by our design, trajectories in $\mathcal{D}'_{\text{env}}$ are collected as if the expert policy were roll-out (because π' can perfectly match π^{E} on $\mathcal{S}(\mathcal{D}_1)$, so the randomness of MLE is only caused by the stochastic transitions.

In summary, we arrive at the following estimator:

$$\widetilde{d}_h^{\pi^{\text{E}}}(s,a) = \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}'_{\text{env}}} \mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot) = (s,a), \mathbf{tr}_h \in \text{Tr}_h^{\mathcal{D}_1}\}}{|\mathcal{D}'_{\text{env}}|} + \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}_1^c} \mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot) = (s,a), \mathbf{tr}_h \notin \text{Tr}_h^{\mathcal{D}_1}\}}{|\mathcal{D}_1^c|}. \quad (7)$$

Two terms in Equation (7) give estimation for terms ♣ and ♠ in Equation (6), respectively. It is important to note that the state-action distribution largely depends on the transition probability, as shown in Equation (5). In contrast to the MLE in Equation (4), our proposed estimator additionally leverages the transition information from the online interactions; see the first term in RHS in Equation (7). This advancement leads to a more accurate estimation of the expert's state-action distribution.

Lemma 4. *Given the expert dataset $\mathcal{D}$, let $\mathcal{D}$ be divided into two equal subsets, i.e., $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$ and $\mathcal{D}_1 \cap \mathcal{D}_1^c = \emptyset$ with $|\mathcal{D}_1| = |\mathcal{D}_1^c| = m/2$. Fix $\pi' \in \Pi_{BC}(\mathcal{D}_1)$, let $\mathcal{D}'_{\text{env}}$ be the dataset collected by π' and $|\mathcal{D}'_{\text{env}}| = n'$. Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$; suppose $H \geq 5$. Consider the estimator $\widetilde{d}_h^{\pi^{\text{E}}}$ shown in (7), if the expert sample complexity (m) and the interaction complexity (n') satisfy*

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

then with probability at least $1-\delta$, we have

$$\sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\text{E}}} - d_h^{\pi^{\text{E}}} \right\|_1 \leq \varepsilon.$$

To our best knowledge, the estimator (7) is the first to enjoy a better expert sample complexity than MLE in the unknown transition setting. The nature of unknown transitions raises a technical difficulty in analyzing the estimation error of two sub-estimators in (7). We highlight that the classical concentration inequality, used to analyze the MLE estimator in Lemma 3, cannot be used to upper bound this estimation error, as the distributions involved are not valid. To overcome this obstacle, we employ Chernoff's bound and additional statistical arguments.

5.3 CONTROLLING OPTIMIZATION ERROR

We now consider the optimization issue. Again, we utilize the dual representation of ℓ_1-norm and the min-max theorem [Bertsekas, 2016] to obtain the following max-min optimization problem:

$$\max_{w \in \mathcal{W}} \min_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a)} w_h(s,a)(\widetilde{d}_h^{\pi^{\text{E}}}(s,a) - d_h^{\pi,\widehat{P}}(s,a)). \quad (8)$$

where $\mathcal{W} = \{w : \|w\|_\infty \leq 1\}$ is the unit ball. We see that the inner problem in (8) is to maximize the policy value of π given the reward function $w_h(s,a)$ (see Equation (1) for the dual form of policy value). For the outer optimization problem, we can use online gradient descent methods [Shalev-Shwartz, 2012] so that the overall objective can finally reach an approximate saddle point. Formally, let us define the objective $f^{(t)}(w)$:

$$\underbrace{\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S}\times\mathcal{A}} w_h(s,a) \left(d_h^{\pi^{(t)},\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\text{E}}}(s,a) \right)}_{:= f^{(t)}(w)}, \quad (9)$$

where $\pi^{(t)}$ is the optimized policy in iteration t. Then the update rule for w is:

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

where $\eta^{(t)} > 0$ is the stepsize to be chosen later, and $\mathcal{P}_{\mathcal{W}}$ is the Euclidean projection on the unit ball $\mathcal{W}$, i.e., $\mathcal{P}_{\mathcal{W}}(w) := \text{argmin}_{z \in \mathcal{W}} \|z - w\|_2$. The procedure for solving (8) is outlined in Algorithm 2.

Algorithm 2 Gradient-based Optimization

Input: Transition model $\widehat{P}$, and expert state-action distribution estimator $\widetilde{d}_h^{\pi^{\text{E}}}$.

1: **for** $t = 1, 2, \cdots, T$ **do**
2: $\pi^{(t)} \leftarrow$ Solve the optimal policy with the transition model $\widehat{P}$ and reward function $w^{(t)}$ up to an error ε_{RL}.
3: Compute the state-action distribution $d_h^{\pi^{(t)},\widehat{P}}$ for $\pi^{(t)}$.
4: Update $w^{(t+1)} := \mathcal{P}_{\mathcal{W}}\left(w^{(t)} - \eta^{(t)} \nabla f^{(t)}(w^{(t)})\right)$ with $f^{(t)}(w)$ defined in Equation (9).
5: **end for**
6: Compute the mean state-action distribution $\bar{d}_h(s,a) = \sum_{t=1}^{T} d_h^{\pi^{(t)},\widehat{P}}(s,a)/T$.
7: Derive $\overline{\pi}_h(a|s) \leftarrow \bar{d}_h(s,a) / \sum_a \bar{d}_h(s,a)$.

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

Line 2 in Algorithm 2 formulates a typical reinforcement learning (RL) optimization problem. We allow $\pi^{(t)}$ to be ε_{RL}-optimal with respect to the optimal policy with reward function $w^{(t)}$, i.e., $V^{\pi^{(t)},\widehat{P},w^{(t)}} \geq V^{\pi^*_{w^{(t)}},\widehat{P},w^{(t)}} - \varepsilon_{\text{RL}}$. In the tabular case, $\varepsilon_{\text{RL}} = 0$ by value iteration with finite and polynomial computation steps. For approximate methods such as policy gradient ascent, we require that they can guarantee ε_{RL} is small with low computational cost.

Lemma 5. *Fix $\varepsilon > 0$. Consider the gradient-based optimization procedure in Algorithm 2 with $\varepsilon_{\mathrm{RL}} \leq \varepsilon/2$. If we take $T \gtrsim H^2|\mathcal{S}||\mathcal{A}|/\varepsilon^2$ and $\eta^{(t)} := \sqrt{|\mathcal{S}||\mathcal{A}|/(8T)}$, then we have*

$$\sum_{h=1}^{H} \left\| d_h^{\overline{\pi},\widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 \leq \min_{\pi\in\Pi} \sum_{h=1}^{H} \left\| d_h^{\pi,\widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 + \varepsilon.$$

5.4 MB-TAIL: COMBING ALL TOGETHER

Combing the above all pieces together, we obtain the final approach called MB-TAIL presented in Algorithm 3. Here "MB-TAIL" stands for model-based transition-aware adversarial imitation learning.

Algorithm 3 Model-based Transition-aware AIL

Input: 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 estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ in (7) with $\mathcal{D}$ and $\mathcal{D}'_{\mathrm{env}}$.
5: $\overline{\pi} \leftarrow$ Apply Algorithm 2 with the estimation $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ under transition model $\widehat{P}$.

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

Theorem 1. *Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$; suppose $H \geq 5$. Under the unknown transition setting, consider MB-TAIL displayed in Algorithm 3 and $\overline{\pi}$ is output policy, assume that the RL error $\varepsilon_{\mathrm{RL}} \leq \varepsilon/2$, the number of iterations and the step size are the same as in Lemma 5, if the expert sample complexity and the interaction complexity satisfy*

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

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

Remark 1. *Our MB-TAIL algorithm achieves expert sample complexity $m = \widetilde{\mathcal{O}}(H^{3/2}|\mathcal{S}|/\varepsilon)$ and total interaction complexity $n + n' = \widetilde{\mathcal{O}}(H^3|\mathcal{S}|^2|\mathcal{A}|/\varepsilon^2)$, even in the case of unknown transitions. In comparison, the OAL algorithm in [Shani et al., 2022] has expert sample complexity $\widetilde{\mathcal{O}}(H^2|\mathcal{S}|/\varepsilon^2)$ and interaction complexity $\widetilde{\mathcal{O}}(H^4|\mathcal{S}|^2|\mathcal{A}|/\varepsilon^2)$ in the same scenario. Theorem 1 validates that our approach provides significant improvements over OAL in terms of both expert sample complexity and interaction complexity.*

The success of this improvement hinges on the design of our algorithm. Unlike OAL, which uses a maximum likelihood estimate of the expert's state-action distribution for imitation, MB-TAIL leverages transition information to construct a more accurate estimator. In addition, OAL uses a tailored optimistic value function in a model-free manner for exploration, but MB-TAIL employs a model-based, reward-free exploration method to efficiently explore the state-action space. These algorithmic designs yield substantial enhancements in both expert sample complexity and interaction complexity.

Simulation Studies. Finally, we conclude by validating our theoretical results through experiments, where we compare the performance of MB-TAIL with four other state-of-the-art algorithms: BC [Pomerleau, 1991], FEM [Abbeel and Ng, 2005], GTAL [Syed and Schapire, 2007], and OAL [Shani et al., 2022]. All algorithms are given 100 expert trajectories, and we evaluate their performance on the Reset Cliff MDP (shown in Figure 1 in Appendix), which is known to be challenging for imitation learning algorithms [Rajaraman et al., 2020, Xu et al., 2021]. We conduct experiments with 20 random seeds, and provide more experimental details in the Appendix. The code to reproduce our results is available at our GitHub repository [4].

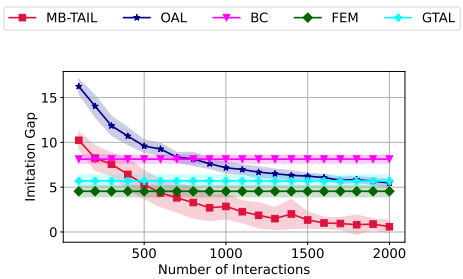

Figure 1: The imitation gap (i.e., $V^{\pi^{\mathrm{E}}} - V^{\pi}$) in Reset Cliff.

Figure 1 shows the imitation gap for each algorithm. Note that BC, FEM, and GTAL do not leverage environment interactions. Our results show that MB-TAIL outperforms FEM and GTAL when the number of interactions exceeds 500. Additionally, we observe that MB-TAIL outperforms OAL with the same number of interactions, which confirms the superior theoretical bounds of MB-TAIL.

6 MB-TAIL WITH STATE ABSTRACTION

Previously, we considered the tabular representation, which leads to theoretical bounds that depend on the size of the problem $|\mathcal{S}|$. However, as suggested by the lower bounds in [Rajaraman et al., 2020, Theorem 6.1, 6.2], this dependence is inevitable and could be unacceptable when $|\mathcal{S}|$ is huge. In this section, we investigate the use of state abstractions [Li et al., 2006] within MB-TAIL, so the dependence on $|\mathcal{S}|$ can be eliminated.

[4] `https://github.com/tianxusky/tabular-ail`

Specifically, we assume that we have a set of state abstractions $\{\phi_h\}_{h=1}^H$, where $\phi_h : \mathcal{S} \to \Phi$ and Φ is the abstract state space. State abstractions correspond to function approximations using a series of piecewise constant functions [Chen and Jiang, 2019]. The abstract state space is much smaller than the original state space, i.e., $|\Phi| \ll |\mathcal{S}|$. We also assume that $\{\phi_h\}_{h=1}^H$ satisfies a condition that is common in the literature [Li et al., 2006, Jiang et al., 2015].

Assumption 1. *There exists a set of known state abstractions $\{\phi_h\}_{h=1}^H$, which satisfies $\forall h \in [H]$, for any $s^1, s^2 \in \mathcal{S}$ such that $\phi_h(s^1) = \phi_h(s^2)$,*

$$\textit{bisimulation} : \forall a \in \mathcal{A}, x' \in \Phi, r_h(s^1, a) = r_h(s^2, a) \tag{10}$$

$$\sum_{s' \in \phi_h^{-1}(x')} P_h(s'|s^1, a) = \sum_{s' \in \phi_h^{-1}(x')} P_h(s'|s^2, a); \tag{11}$$

$$\pi^{\mathrm{E}}\textit{-irrelevant} : \ \pi_h^{\mathrm{E}}(s^1) = \pi_h^{\mathrm{E}}(s^2), \tag{12}$$

where $\phi_h^{-1}(x') = \{s' \in \mathcal{S} : \phi_h(s) = x'\}$.

In bisimulation, the reward-consistent condition in (10) ensures that two different states mapped to the same abstract state share an identical reward. We highlight that this condition is important for MB-TAIL to avoid the dependence of *expert sample complexity* on $|\mathcal{S}|$. In particular, the bottleneck of the sample complexity of AIL methods is the estimation of $d_h^{\pi^{\mathrm{E}}}(s, a)$. Under the reward-consistent condition, we can calculate the expert policy value as

$$\begin{aligned} V^{\pi^{\mathrm{E}}} &= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} r_h(s, a) d_h^{\pi^{\mathrm{E}}}(s, a) \\ &= \sum_{h=1}^{H} \sum_{(x,a) \in \Phi \times \mathcal{A}} r_h^{\phi}(x, a) d_h^{\pi^{\mathrm{E}}, \phi}(x, a), \end{aligned}$$

where $r_h^{\phi}(x, a) = r_h(s, a)$ for an arbitrary $s \in \phi_h^{-1}(x)$ and $d_h^{\pi,\phi}(x, a) = \mathbb{P}^{\pi^{\mathrm{E}}}(\phi_h(s_h) = x, a_h = a) = \sum_{s \in \phi_h^{-1}(x)} d_h^{\pi}(s, a)$ is the abstract state-action distribution. With the above formulation, to estimate the expert policy value, we can estimate the *abstract* state-action distribution rather than the tabular counterpart, which can remove the dependence on $|\mathcal{S}|$. Analogously, the transition-consistent condition in (11) guarantees that two different states mapped to the same abstract state share an identical transition. This condition is crucial for removing the dependence of *interaction complexity* on $|\mathcal{S}|$ since it allows estimating the "abstract transition function".

We now extend MB-TAIL to the state abstraction setting, which we describe in detail in the Appendix due to space limitations. We prove that under Assumption 1, MB-TAIL achieves expert sample and interaction complexities that are independent of $|\mathcal{S}|$. However, the proof is not straightforward, and the primary challenge is to connect the state-action distributions in the original and abstract MDPs. We provide a detailed discussion of the specialized analysis tools in the Appendix.

Theorem 2. *Under Assumption 1, fix $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$; suppose $H \geq 5$. Under the unknown transition setting, consider Algorithm 2 in Appendix and $[\overline{\pi}^{\phi}]^M$ is output policy. Assume that the RL error $\varepsilon_{\mathrm{RL}} \leq \varepsilon/2$, the number of iterations $T \gtrsim H^2|\Phi||\mathcal{A}|/\varepsilon^2$, and the step size $\eta^{(t)} := \sqrt{|\Phi||\mathcal{A}|/(8T)}$. If the number of expert trajectories (m), the number of interaction trajectories for estimation (n'), and the number of interaction trajectories for reward-free exploration (n) satisfy*

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

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

then with probability at least $1 - \delta$, we have the imitation gap $V^{\pi^{\mathrm{E}}} - V^{[\overline{\pi}^{\phi}]^M} \leq \varepsilon$.

7 CONCLUSION

This paper contributes to the establishment of theoretical foundations for AIL with unknown transitions. We propose a new and general framework that enables AIL to explore and imitate efficiently. As mentioned, AIL methods can have much better theoretical guarantees on structured instances, such as horizon-free bounds suggested in [Xu et al., 2022]. Thus, we believe that investigating AIL with unknown transitions on structured instances is an interesting and valuable direction for future research.

ACKNOWLEDGMENT

Tian Xu would like to thank Zhilong Zhang, Fanming Luo, and Jingcheng Pang for reading the manuscript and providing helpful comments. The work of Yang Yu is supported by National Key Research and Development Program of China (2020AAA0107200), NSFC(61876077), and Collaborative Innovation Center of Novel Software Technology and Industrialization. The work of Zhi-Quan Luo is supported in part by the National Key Research and Development Project under grant 2022YFA1003900, and in part by the Guangdong Provincial Key Laboratory of Big Data Computing.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, pages 1–8, 2004.

P. Abbeel and A. Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1–8, 2005.

A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. *Reinforcement Learning: Theory and Algorithms*. https://rltheorybook.github.io/, 2022.

B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

K. Brantley, W. Sun, and M. Henaff. Disagreement-regularized imitation learning. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1042–1051, 2019.

X. Chen, J. Hu, L. Yang, and L. Wang. Near-optimal reward-free exploration for linear mixture mdps with plug-in solver. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin. Primal wasserstein imitation learning. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1329–1338, 2016.

J. Fu, K. Luo, and S. Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon. Iq-learn: Inverse soft-q learning for imitation. In *Advances in Neural Information Processing Systems 34*, pages 4028–4039, 2021.

S. K. S. Ghasemipour, R. S. Zemel, and S. Gu. A divergence minimization perspective on imitation learning methods. In *Proceedings of the 3rd Conference on Robot Learning*, pages 1259–1277, 2019.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, pages 4565–4573, 2016.

E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Proceedings of the 6th Conference on Robot Learning*, pages 991–1002, 2022.

N. Jiang, A. Kulesza, and S. Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 179–188, 2015.

C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4870–4879, 2020.

S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh. On learning distributions from their samples. In *Proceedings of the 28th Conference on Learning Theory*, pages 1066–1100, 2015.

E. Kaufmann, P. Ménard, O. D. Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pages 865–891, 2021.

L. Ke, M. Barnes, W. Sun, G. Lee, S. Choudhury, and S. S. Srinivasa. Imitation learning as f-divergence minimization. *arXiv preprint arXiv:1905.12888*, 2019.

I. Kostrikov, K. K. Agrawal, D. Dwibedi, S. Levine, and J. Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for mdps. *ISAIM*, 4:5, 2006.

Z. Li, T. Xu, Y. Yu, and Z.-Q. Luo. Rethinking valuedice: Does it really improve performance? *arXiv preprint arXiv:2202.02468*, 2022.

D. Lindner, A. Krause, and G. Ramponi. Active exploration for inverse reinforcement learning. In *Advances in Neural Information Processing Systems 35*, pages 5843–5853, 2022.

Z. Liu, Y. Zhang, Z. Fu, Z. Yang, and Z. Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 14094–14138, 2022.

P. Ménard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7599–7608, 2021.

A. M. Metelli, G. Ramponi, A. Concetti, and M. Restelli. Provably efficient learning of transferable rewards. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7665–7676, 2021.

A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 663–670, 2000.

T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotic*, 7(1-2): 1–179, 2018.

D. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1): 88–97, 1991.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

N. Rajaraman, L. F. Yang, J. Jiao, and K. Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems 33*, pages 2914–2924, 2020.

N. Rajaraman, Y. Han, L. Yang, J. Liu, J. Jiao, and K. Ramchandran. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems 34*, pages 1325–1336, 2021a.

N. Rajaraman, Y. Han, L. F. Yang, K. Ramchandran, and J. Jiao. Provably breaking the quadratic error compounding barrier in imitation learning, optimally. *arXiv preprint arXiv:2102.12948*, 2021b.

S. Ross and D. Bagnell. Efficient reductions for imitation learning. In *Proceedings of the 13rd International Conference on Artificial Intelligence and Statistics*, pages 661–668, 2010.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

L. Shani, T. Zahavy, and S. Mannor. Online apprenticeship learning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 8240–8248, 2022.

J. Shi, Y. Yu, Q. Da, S. Chen, and A. Zeng. Virtual-taobao: virtualizing real-world online retail environment for reinforcement learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 4902–4909, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

G. Swamy, N. Rajaraman, M. Peng, S. Choudhury, J. Bagnell, S. Z. Wu, J. Jiao, and K. Ramchandran. Minimax optimal online imitation learning via replay estimation. In *Advances in Neural Information Processing Systems 35*, pages 7077–7088, 2022.

U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems 20*, pages 1449–1456, 2007.

U. Syed, M. H. Bowling, and R. E. Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1032–1039, 2008.

F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957, 2018.

L. Viano, A. Kamoutsi, G. Neu, I. Krawczuk, and V. Cevher. Proximal point imitation learning. In *Advances in Neural Information Processing Systems 35*, pages 24309–24326, 2022.

R. Wang, S. S. Du, L. F. Yang, and R. R. Salakhutdinov. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems 33*, pages 17816–17826, 2020a.

Y. Wang, T. Liu, Z. Yang, X. Li, Z. Wang, and T. Zhao. On computation and generalization of generative adversarial imitation learning. In *Proceedings of the 8th International Conference on Learning Representations*, 2020b.

T. Xu, Z. Li, and Y. Yu. Error bounds of imitating policies and environments. In *Advances in Neural Information Processing Systems 33*, pages 15737–15749, 2020.

T. Xu, Z. Li, and Y. Yu. Error bounds of imitating policies and environments for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6968–6980, 2021.

T. Xu, Z. Li, Y. Yu, and Z.-Q. Luo. Understanding adversarial imitation learning in small sample regime: A stage-coupled analysis. *arXiv*, 2208.01899, 2022.

T. Zahavy, A. Cohen, H. Kaplan, and Y. Mansour. Apprenticeship learning via frank-wolfe. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6720–6728, 2020.

S. Zeng, C. Li, A. Garcia, and M. Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems 35*, pages 10122–10135, 2022.

W. Zhang, D. Zhou, and Q. Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems 34*, 2021.

Y. Zhang, Q. Cai, Z. Yang, and Z. Wang. Generative adversarial imitation learning with neural network parameterization: Global optimality and convergence rate. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11044–11054, 2020.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.