Understanding Adversarial Imitation Learning in Small Sample Regime: A Stage-coupled Analysis

Tian Xu

Nanjing University, School of Artificial Intelligence

Based on the paper: Tian Xu*, Ziniu Li*, Yang Yu, Zhi-Quan Luo. Understanding AIL in Small Sample Regime: A Stage-coupled Analysis. TPAMI.




Tian Xu
NJU

Ziniu Li
CUHKSZ

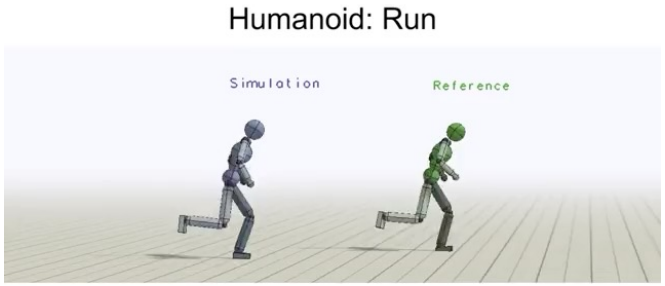
Yang Yu
NJU

Zhi-Quan Luo
CUHKSZ

What is Imitation Learning

Imitation Learning (IL) aims to learn effective policies by **mimicking expert demonstrations**

Imitate to run



Policy trained to imitate a running clip.

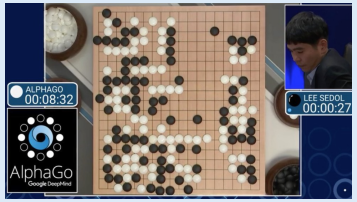
[https://gfycat.com/bowedinexperiencedhogget]

Imitate to take the banana

[https://www.comp.nus.edu.sg/~rishav1/blog/2018/Understanding-GAIL/]

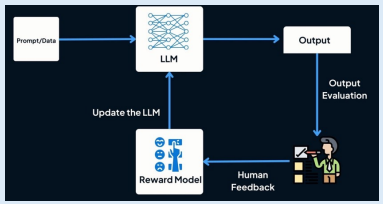
Applications of IL

DeepMind AlphaGo



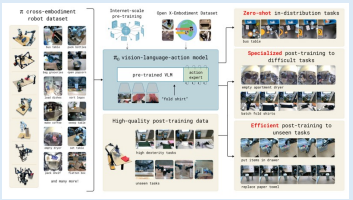
Imitate professional player for initial learning

OpenAI ChatGPT



Imitate internet corpora and high-quality human responses

Physical Intelligence π^0



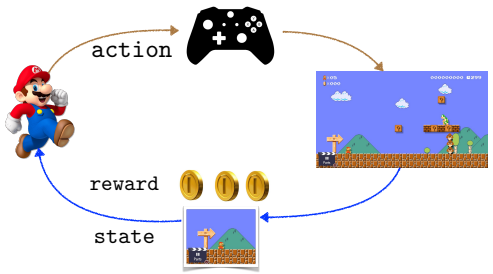
Imitate human demonstrations

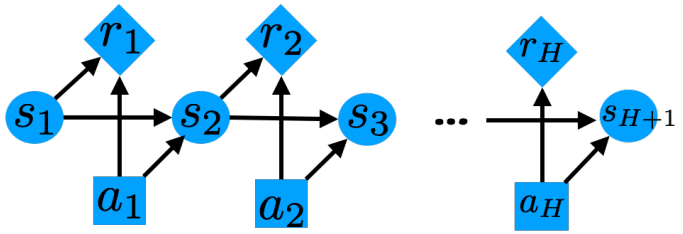
IL is a general paradigm underpins multiple breakthrough AI systems

IL Set-up: Markov Decision Process

Definition (Finite-horizon Markov Decision Process)**.** A finite-horizon Markov Decision Process is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, s_1)$, where:

- $\mathcal{S}$ is the state space;
- $\mathcal{A}$ is the action space;
- $P = \{P_h\}_{h=1}^H$, where $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability at step h;
- $r = \{r_h\}_{h=1}^H$, where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function at step h;
- H is the decision horizon;
- s_1 is the initial state.





IL Set-up: Markov Decision Process

For policy $\pi = \{\pi_h\}_{h=1}^{H}$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$

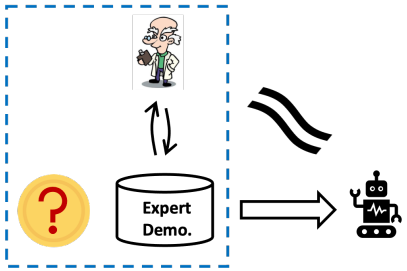
Policy Value V^π measures the quality of a policy:

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

State-action distribution $\{d_h^\pi(\cdot,\cdot)\}_{h=1}^{H}$:

$$d_h^\pi(s, a) = \mathbb{P}\left(s_h = s, a_h = a; \pi\right)$$

IL Set-up



IL Set-up

The learner has no knowledge of the reward function r, but has access to expert demonstrations $\mathcal{D}^{\mathrm{E}}$. The expert demonstrations consist of N expert trajectories collected by the deterministic expert policy π^{E}.

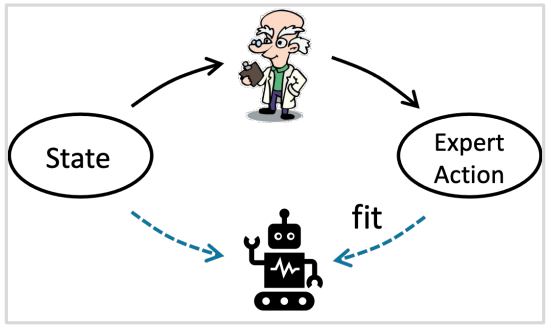
$$\mathcal{D}^{\mathrm{E}} = \left\{\mathrm{tr}^1, \ldots, \mathrm{tr}^N\right\}, \quad \mathrm{tr}^i = \left\{\left(s_1, a_1\right), \ldots, \left(s_H, a_H\right)\right\} \sim \pi^{\mathrm{E}}$$

IL Target

The learner aims to ouput a policy $\widehat{\pi}$ such that the policy value gap with π^{E} is small:

Imitation Gap: $V^{\pi^{\mathrm{E}}} - V^{\widehat{\pi}}$

Behavioral Cloning (BC)



Objective

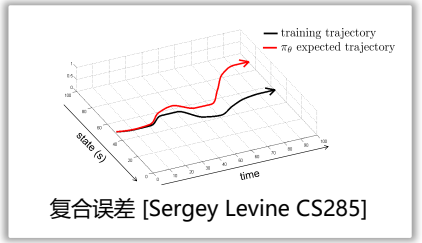
$$\max_{\pi} \mathbb{E}_{\tau \sim \pi^{\mathrm{E}}} \left[\sum_{h=1}^{H} \log \left(\pi_h \left(a_h | s_h \right) \right) \right]$$

Feature

- Easy to implement
- **Offline learning**: purely trained on offline demonstrations

BC Suffers from Compounding Errors

- Due to its offline learning manner, BC cannot recover expert actions on states **out of demonstrations**
- Single-step error **accumulates over the decision horizon**, resulting in huge compounding errors



复合误差 [Sergey Levine CS285]

- Theory: [Ross and Bagnell, 2012] proved that BC has an imitation gap bound with **quadratic dependence on horizon**

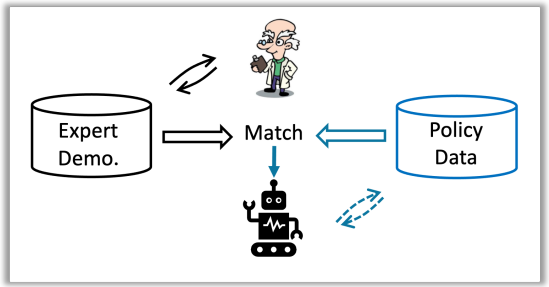
$$V^{\pi^{\mathrm{E}}} - V^{\pi^{\mathrm{BC}}} \leq \mathcal{O}\left(H^2 \varepsilon\right)$$

generalization error $\varepsilon = \frac{1}{H}\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\left[\sum_{h=1}^{H}\mathbb{E}_{a\sim\pi_h^{\mathrm{BC}}(\cdot|s_h)}\left[\mathbb{I}\left\{a \neq \pi_h^{\mathrm{E}}(s_h)\right\}\right]\right]$

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. AISTATS 2010.

Key difference from supervised learning (**H=1**)

Adversarial Imitation Learning



Objective

"state-action distribution matching"

$$\min_{\pi} \sum_{h=1}^{H} D(d_h^{\pi}(\cdot,\cdot), d_h^{\pi^{\mathrm{E}}}(\cdot,\cdot)) \quad , \quad D(\cdot,\cdot)\text{: divergence measure}$$

$$\min_{\pi} \max_{r} \underbrace{\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{\pi^{\mathrm{E}}}}[r(s,a)]}_{V^{\pi^{\mathrm{E}},r}} - \underbrace{\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{\pi}}[r(s,a)]}_{V^{\pi,r}}$$

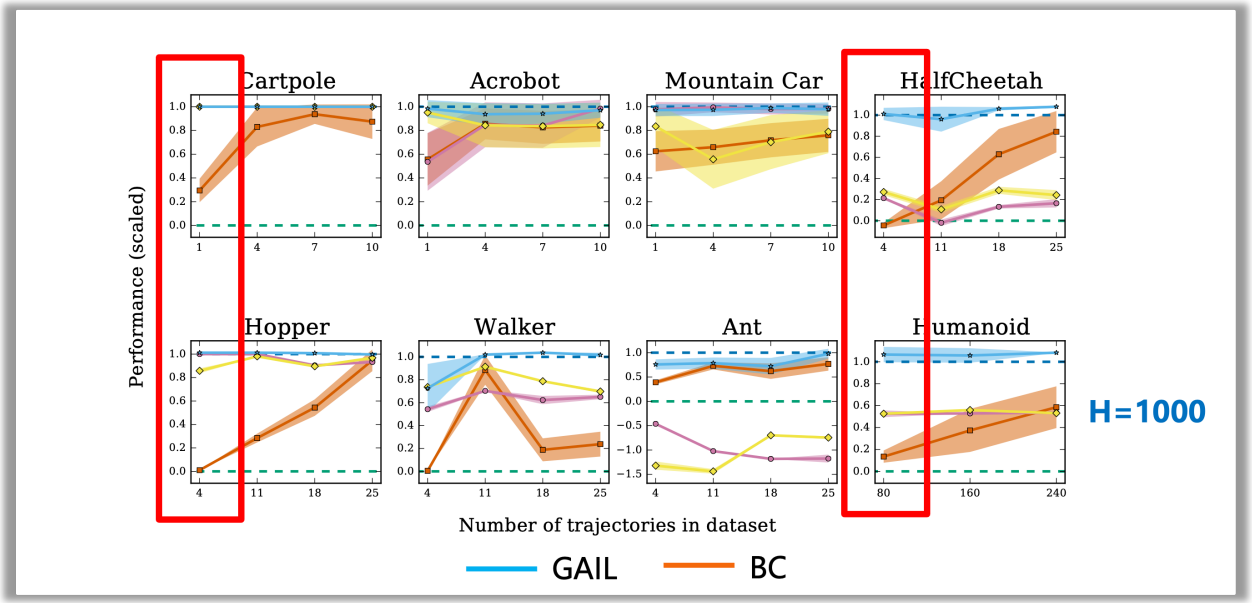
RL sub-problem

- inner: learn an adversarial reward to maximize the imitation gap
- outer: learn a policy to maximize the value w.r.t the adversarial reward

Feature

AIL involves solving RL problems, requiring **online interactions**

AIL Outperforms BC Significantly



Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. NeurIPS 2016.

- With limited demonstrations, BC suffers from large performance gaps in long-horizon tasks
- AIL can achieve near-expert performance with limited demonstrations in long-horizon tasks

Understanding AIL in Small Sample Regime

CoRL 2019
Best Paper

A Divergence Minimization Perspective on Imitation Learning Methods

Seyed Kamyar Seyed Ghasemipour
University of Toronto, Vector Institute
kamyar@cs.toronto.edu

Richard Zemel
University of Toronto, Vector Institute
zemel@cs.toronto.edu

Shixiang Gu
Google Brain
shanegu@google.com

gence, whereas AIRL and GAIL use divergences that exhibit more mode-seeking behaviour. These observations allow us to generate the following two hypotheses about why IRL methods outperform BC, particularly in the low-data regime,

Hypothesis 1 *In common MDPs of interest, the reward function depends more on the state than the action. Hence encouraging policies to explicitly match expert state marginals is an important learning criterion.*

Hypothesis 2 *It is known that optimization using the forward KL divergence results in distributions with a mode-covering behaviour, whereas using the reverse KL results in mode-seeking behaviour [39]. In RL we care about the "quality of trajectories", as measured by the likelihood under the expert distribution. Therefore, being mode-seeking is more beneficial than mode-covering, particularly in the low-data regime.*

Open Question: "Why AIL methods outperform BC, particularly in the low-data regime?"

- Hypothesis 1: Matching state marginals is an important learning criterion (high level)
- Hypothesis 2: Divergence Choice: Forward KL V.S. Reverse KL (not tailored to IL)

Target: Develop **theory** to provide a **fundamental and concrete** understanding

Existing Theoretical Analysis for AIL

[Xu et al., 2020, Rajaraman et al., 2020] proved that AIL can achieve an imitation gap bound with **linear horizon dependence**:

$$V^{\pi^{\mathrm{E}}} - \mathbb{E}\left[V^{\pi^{\mathrm{AIL}}}\right] \leqslant \mathcal{O}\left(H\sqrt{\frac{|\mathcal{S}|}{N}}\right)$$

Tian Xu, et al. Error bounds of imitating policies and environments. NeurIPS 2020.
Nived Rajaraman, et al. Toward the fundamental limits of imitation learning. NeurIPS 2020.

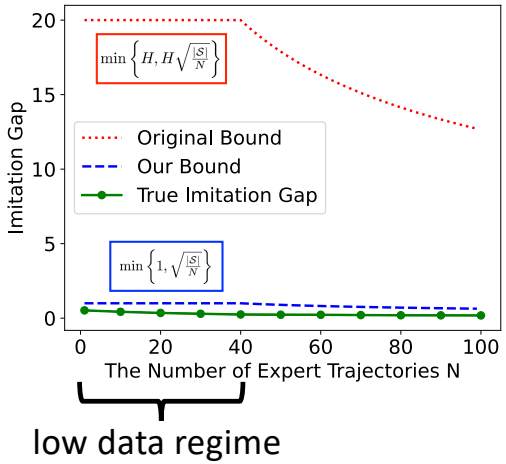
- Compared with the $\mathcal{O}(H^2\epsilon)$ bound in BC, AIL achieves an improved $\mathcal{O}(H\epsilon)$ bound, explaining why AIL performs good in long-horizon tasks
- In the small sample regime ($\boldsymbol{N \leqslant |\mathcal{S}|}$), this bound is meaningless because the maximum gap is H, contradicting with good practical performance of AIL 😕

Main Contributions

- The **first low data regime** imitation gap bound for AIL

Theorem. Under mild assumptions, AIL can achieve an imitation gap bound:

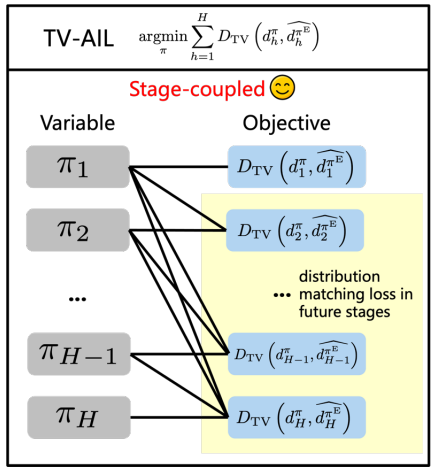
$$V\left(\pi^{\mathrm{E}}\right)-\mathbb{E}\left[V\left(\pi^{\mathrm{AIL}}\right)\right] \leqslant \mathcal{O}\left(\min\left\{1, \sqrt{\frac{|\mathcal{S}|}{N}}\right\}\right)$$



- A new stage-coupled analysis reveals why AIL has "good generalization" via the lens of "coupled optimization"

Key Mechanism

Future distribution matching loss guides AIL to recover expert actions on **preceding uncovered states**



Total-Variation Distance based AIL

Total-Variation Distance based AIL (TV-AIL)

$$\pi^{\mathrm{AIL}} \in \operatorname*{argmin}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d_h^{\pi}(s,a) - \widehat{d_h^{\pi^{\mathrm{E}}}}(s,a) \right|$$

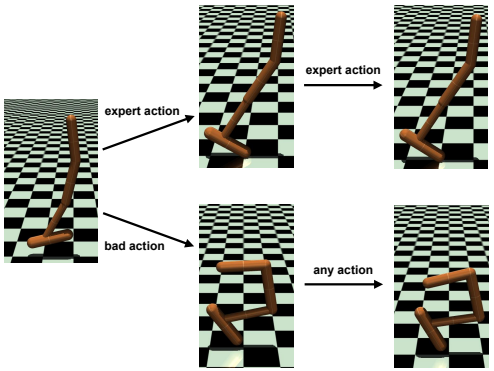
$\widehat{d_h^{\pi^{\mathrm{E}}}}$: empirical estimate of $d_h^{\pi^{\mathrm{E}}}$ from demos.

- **Linear programming formulation** (over d_h^{π}): the exact solution can be obtained in polynomial time
- **Minimax formulation**: standard gradient-descent-ascent yields an approximate optimal solution in polynomial time

$$\min_{\pi \in \Pi} \max_{c \in \mathcal{C}_{\mathrm{TV}}} \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d_h^{\pi}} \left[c_h(s,a) \right] - \mathbb{E}_{(s,a) \sim \widehat{d_h^{\pi^{\mathrm{E}}}}} \left[c_h(s,a) \right],$$

$$\mathcal{C}_{\mathrm{TV}} = \{ c = (c_1, \ldots, c_H) : c_h : \mathcal{S} \times \mathcal{A} \to [-1,1], \forall h \in [H] \}$$

Reachable Bad Absorbing States



The robot takes an expert action and jumps forward, resulting in a positive reward

The robot takes an non-expert action, results in a fall and a zero reward on the **absorbing** state

Assumption (RBAS MDPs)**.** For a tabular and episodic MDP and an expert policy, we define $\mathcal{S}^{\mathrm{G}} = \{s : \exists h \in [H], d_h^{\pi^{\mathrm{E}}}(s) > 0\}$ as expert states and $\mathcal{S}^{\mathrm{B}} = \mathcal{S} \setminus \mathcal{S}^{\mathrm{G}}$ as non-expert states. Assume that

- Non-expert states only have transitions to themseleves: $\forall h \in [H], b \in \mathcal{S}^{\mathrm{B}}, a \in \mathcal{A}, \sum_{b' \in \mathcal{S}^{\mathrm{B}}} P_h(b'|b,a) = 1.$
- Expert states are reachable via expert actions and are non-reachable via non-expert actions: $\forall s, s' \in \mathcal{S}^{\mathrm{G}}, h \in [H],\ P_h(s'|s, \pi_h^{\mathrm{E}}(s)) > 0,\ \forall a \neq \pi_h^{\mathrm{E}}(s), P_h(s'|s,a) = 0$

If taking a non-expert action, the agent transits into **absorbing non-expert states**

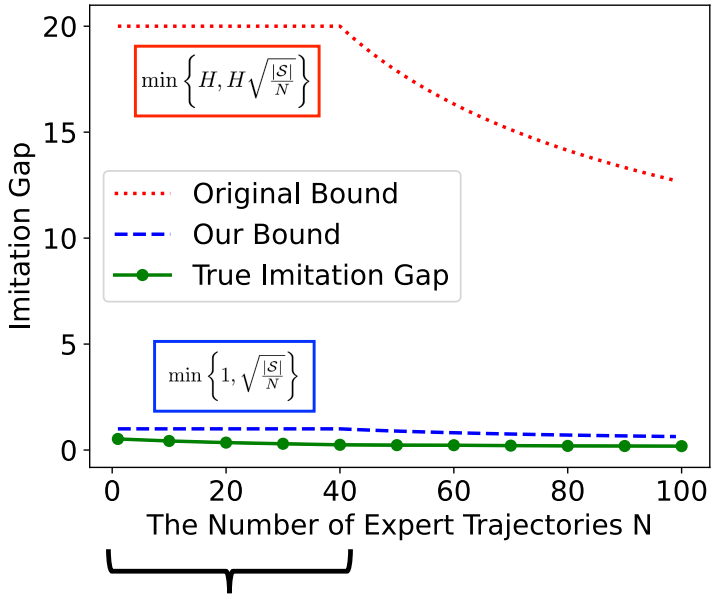
Low-Data-Regime Imitation Gap of TV-AIL

Theorem (Imitation Gap of TV-AIL). For any RBAS MDPs, suppose that π^{AIL} is the policy recovered by TV-AIL, we have that

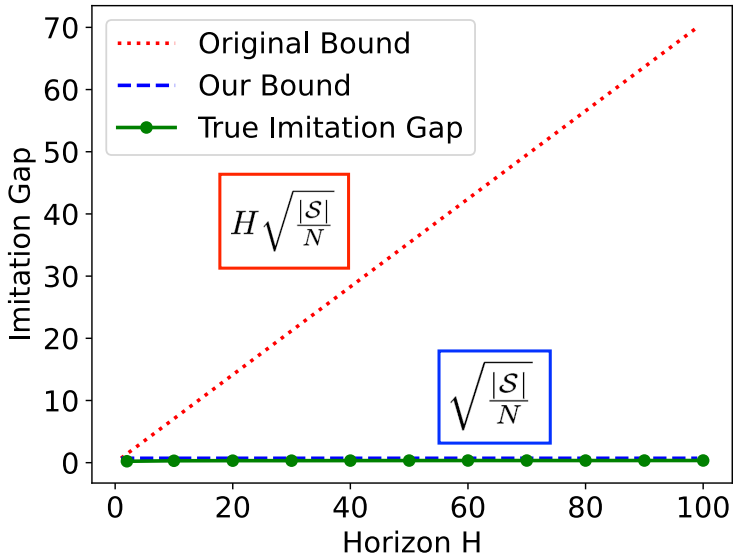
$$V\left(\pi^{\mathrm{E}}\right)-\mathbb{E}\left[V\left(\pi^{\mathrm{AIL}}\right)\right] \leqslant \mathcal{O}\left(\min\left\{1, \sqrt{\frac{|\mathcal{S}|}{N}}\right\}\right)$$

- The **first low data regime bound** to our best knowledge
  - Small sample regime ($N \preccurlyeq |\mathcal{S}|$): the imitation gap of TV-AIL is at most 1, which is significantly smaller than H
  - Large sample regime ($N \succcurlyeq |\mathcal{S}|$): the imitation gap of TV-AIL is $\sqrt{|\mathcal{S}|/N}$, which diminishes to 0 as $N \to \infty$
- The bound is also **horizon-free** in the whole sample regime

New Theory Predicts Practice Well



low data regime



- **Our bound predicts the true imitation gap accurately** with varying numbers of expert trajectories and horizons
- The original bound provides a **loose prediction**

New Theory Predicts Practice Well

Performance of TV-AIL and BC on MuJoCo benchmarks

Varying N

		$N = 1$	$N = 4$	$N = 7$	$N = 10$
Hopper (scale = 3.2)	BC	$784.97_{\pm 28.09}$	$887.04_{\pm 31.26}$	$666.44_{\pm 106.58}$	$460.72_{\pm 74.95}$
	TV-AIL	$10.38_{\pm 11.25}$	$1.81_{\pm 2.88}$	$-4.96_{\pm 10.88}$	$2.33_{\pm 10.59}$
HalfCheetah (scale = 7.7)	BC	$1058.48_{\pm 8.27}$	$1066.21_{\pm 22.76}$	$988.07_{\pm 35.52}$	$579.53_{\pm 171.28}$
	TV-AIL	$-22.45_{\pm 101.65}$	$-84.96_{\pm 16.84}$	$-78.69_{\pm 6.98}$	$-79.29_{\pm 7.95}$
Walker2d (scale = 5.0)	BC	$1002.13_{\pm 9.68}$	$939.27_{\pm 19.10}$	$528.91_{\pm 181.46}$	$222.98_{\pm 52.97}$
	TV-AIL	$12.89_{\pm 19.61}$	$9.04_{\pm 17.99}$	$-4.91_{\pm 8.14}$	$14.36_{\pm 12.48}$

TV-AIL matches the expert performance with limited demos. (N=1) while BC suffers a large imitation gap, aligning with the low-data property of the new bound

Varying H

		$H = 100$	$H = 500$	$H = 1000$	$H = 2000$
Hopper (scale=3.2)	BC	$0.80_{\pm 1.72}$	$178.56_{\pm 79.27}$	$784.97_{\pm 28.09}$	$1950.42_{\pm 36.20}$
	TV-AIL	$4.96_{\pm 4.56}$	$-1.73_{\pm 6.77}$	$10.38_{\pm 11.25}$	$-9.92_{\pm 37.15}$
HalfCheetah (scale=7.7)	BC	$56.44_{\pm 9.23}$	$491.91_{\pm 21.01}$	$1058.48_{\pm 8.27}$	$2198.61_{\pm 12.93}$
	TV-AIL	$8.49_{\pm 8.26}$	$-24.73_{\pm 11.46}$	$-22.45_{\pm 101.65}$	$-169.83_{\pm 16.06}$
Walker2d (scale=5.0)	BC	$10.25_{\pm 7.89}$	$413.02_{\pm 13.65}$	$1002.13_{\pm 9.68}$	$2158.84_{\pm 2.05}$
	TV-AIL	$-0.18_{\pm 1.04}$	$16.26_{\pm 22.64}$	$12.93_{\pm 19.61}$	$71.69_{\pm 66.30}$

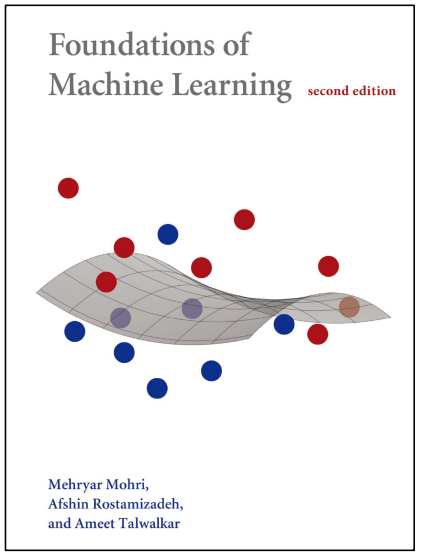
When increasing H, TV-AIL maintains small imitation gaps while BC's imitation gaps blow up, aligning with the horizon-free property of the new bound

An Illustration: Classical Machine Learning Theory

Classical machine learning theory often provides such a generalization error bound

$$R(h) \leq \hat{R}(h) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\log|\mathcal{H}|}{N}}\right)$$

- $R(h) = \mathbb{E}_{(x,y)\sim P}[\mathbb{I}\{h(x) \neq y\})]$ is the expected risk
- $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{h(x_i) \neq y_i\}$ is the empirical risk
- $\log|\mathcal{H}|$ is the complexity of the hypothesis class



This bound is meaningless in the small sample regime $N \preccurlyeq \log|\mathcal{H}|$ since the generalization error is 1

Classical Reduction-and-Estimation Analysis

Existing AIL analysis follows classical ml theory and decomposes the imitation gap into statistical error and optimization error

$$
\begin{aligned}
&\left|V\left(\pi^{\mathrm{E}}\right)-V\left(\pi^{\mathrm{AIL}}\right)\right| \\
\leq & \cdots \\
\leq & \underbrace{\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_h^{\pi^{\mathrm{E}}}(s,a)-\widehat{d_h^{\pi^{\mathrm{E}}}}(s,a)\right|}_{\text{statistical error}} + \underbrace{\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\widehat{d_h^{\pi^{\mathrm{E}}}}(s,a)-d_h^{\pi^{\mathrm{AIL}}}(s,a)\right|}_{\text{optimization error}} \\
\leq & \underbrace{2\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\widehat{d_h^{\pi^{\mathrm{E}}}}(s,a)-d_h^{\pi^{\mathrm{E}}}(s,a)\right|}_{\text{statistical error}} \\
\leq & \mathcal{O}\left(H\sqrt{\frac{|\mathcal{S}|}{N}}\right)
\end{aligned}
$$

Classical Reduction-and-Estimation Analysis

Claim: The reduction to statistical error is very loose for AIL **in the low data regime**

TABLE 5: Imitation gap and estimation error of TV-AIL on RBAS MDPs with $N = 1$.

	$H = 100$	$H = 500$	$H = 1000$	$H = 2000$
Imitation Gap	$0.69_{\pm 0.00}$	$0.70_{\pm 0.00}$	$0.71_{\pm 0.00}$	$0.71_{\pm 0.00}$
Estimation Error	$189.47_{\pm 0.00}$	$947.37_{\pm 0.00}$	$1894.74_{\pm 0.00}$	$3789.47_{\pm 0.00}$

In the low data regime, despite a large estimation error, TV-AIL can generalize well and still achieve a small imitation gap

Sharp Characterization of TV-AIL

🤔 Target: Establish a sharp characterization of π^{AIL} instead of reducing to estimation error coarsely

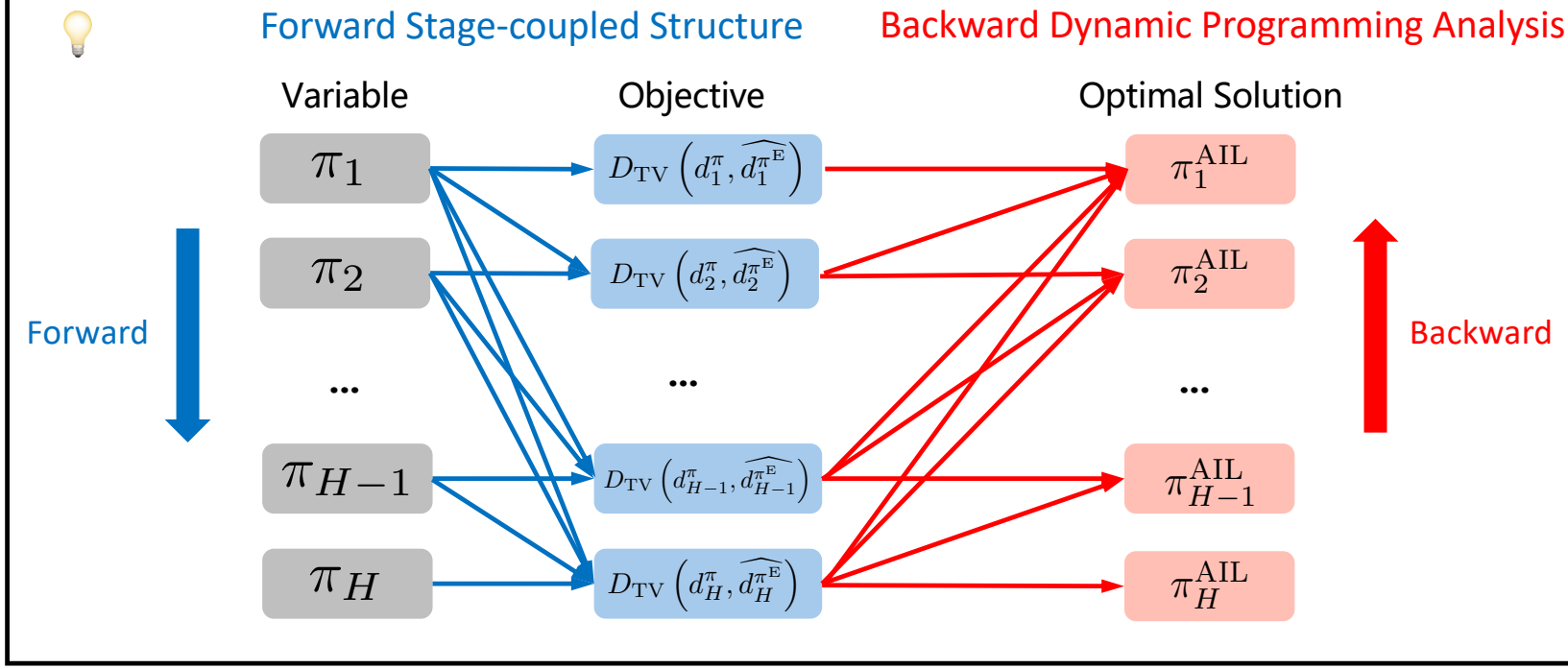
$$\pi^{\mathrm{AIL}} \in \operatorname*{argmin}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d_h^{\pi}(s,a) - \widehat{d_h^{\pi^{\mathrm{E}}}}(s,a) \right|$$ $\widehat{d_h^{\pi^{\mathrm{E}}}}$: empirical estimate of $d_h^{\pi^{\mathrm{E}}}$ from demos.

⛰️ Difficulty: This is a non-smooth and non-convex problem, making classical optimality conditions unapplicable

Proposition. There exist tabular and episodic MDPs such that the objective of TV-AIL is non-convex.

Stage-Coupled Analysis

$$\pi^{\mathrm{AIL}} \in \operatorname*{argmin}_{\pi} \sum_{h=1}^{H} D_{\mathrm{TV}}\left(d_h^{\pi}, \widehat{d_h^{\pi^{\mathrm{E}}}}\right)$$



Optimality Condition of TV-AIL

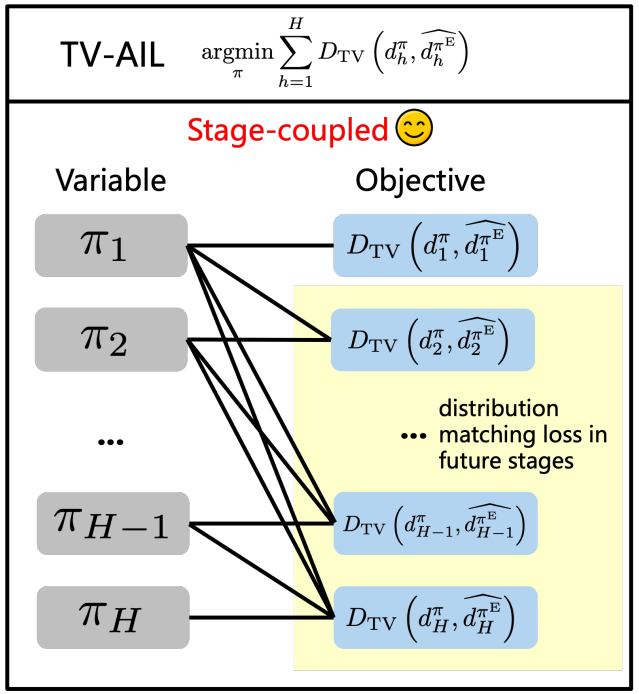
Proposition. For any RBAS MDP, suppose that π^{AIL} is the policy recovered by TV-AIL. When $N \geq 1$, the following optimality conditoin holds almost surely:

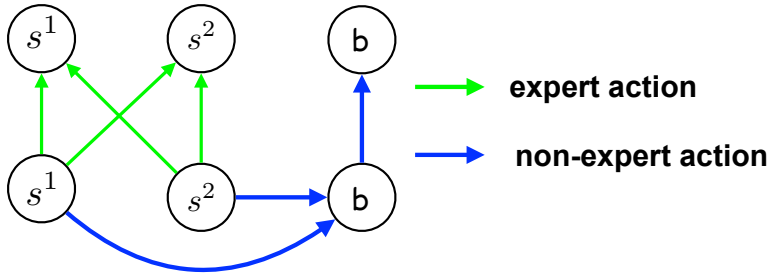
$$\pi_h^{\mathrm{AIL}}\left(\pi_h^{\mathrm{E}}(s)|s\right) = 1, \forall s \in \mathcal{S}^{\mathrm{G}}, h \in [H-1].$$

- In the first H-1 stages, AIL can recover expert actions **even on states uncovered by demonstrations**, implying its **generalization ability beyond demonstrations**

- With additional analysis on stage H, we can obtain the final bound $\mathcal{O}\left(\min\left\{1, \sqrt{|\mathcal{S}|/N}\right\}\right)$

Mechanism Under the Generalization of TV-AIL



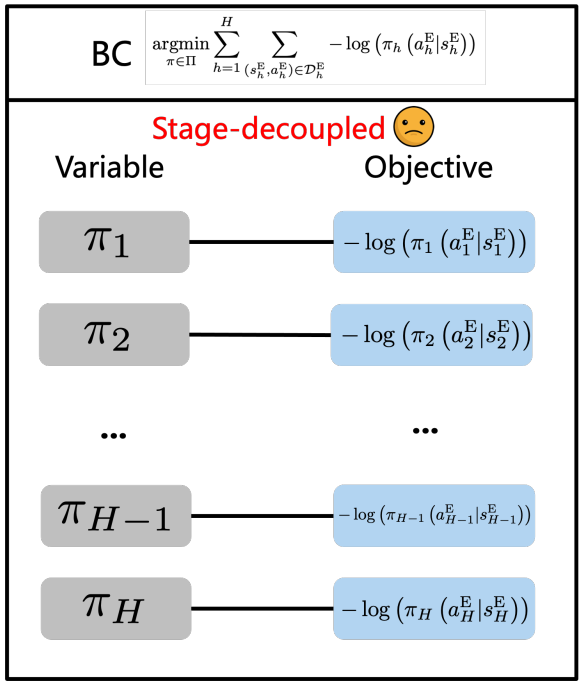


- demonstrations=$\{s^1 \rightarrow s^1\}$
- uncovered state s^2

$$\pi_1^{\mathrm{AIL}}(a \mid s^2) = 1$$

- Key insight: Due to the **coupling structure**, future distribution loss provably guides AIL to recover expert actions on preceding uncovered states
- The RL procedure for multi-step distribution matching is the key for the generalization in AIL

Stage-decoupled Objective in BC



- The BC objective is **stage-decoupled:** each policy π_h is optimized solely against expert data at step h, with no signal from other time steps
- Thus BC cannot generalize beyond demonstrations

Extension for Approximate Solutions

Exact solution:

$$\pi^{\mathrm{AIL}} \in \operatorname*{argmin}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d_h^{\pi}(s,a) - \widehat{d_h^{\pi^{\mathrm{E}}}}(s,a) \right|$$

Approximate solution:

Definition (ε-optimal solution)**.** A policy $\bar{\pi}$ is an ε-optimal solution for TV-AIL if

$$\sum_{h=1}^{H} \left\| d_h^{\bar{\pi}} - \widehat{d_h^{\pi^{\mathrm{E}}}} \right\|_1 \leqslant \min_{\pi \in \Pi} \sum_{h=1}^{H} \left\| d_h^{\pi} - \widehat{d_h^{\pi^{\mathrm{E}}}} \right\|_1 + \varepsilon.$$

Extension for Approximate Solutions

Theorem. For any RBAS MDP, the candidate policy set is defined as $\Pi^{\mathrm{OPT}} = \left\{\pi \in \Pi : \forall h \in [H], \exists s \in \mathcal{S}^{\mathrm{G}}, \pi_h\left(\pi_h^{\mathrm{E}}(s)|s\right) > 0\right\}$. Suppose that $\bar{\pi} \in \Pi^{\mathrm{OPT}}$ is an ε-optimal solution of TV-AIL. Then we have

$$V\left(\pi^{\mathrm{E}}\right) - \mathbb{E}[V(\bar{\pi})] \lesssim \min\left\{1 + \frac{8\varepsilon}{c(\bar{\pi})}, \sqrt{\frac{|\mathcal{S}|}{N}} + \frac{8\varepsilon}{c(\bar{\pi})}\right\},$$

where $c(\bar{\pi}) > 0$ is defined as $c(\bar{\pi}) := \min_{1 \leqslant \ell < h \leqslant H, s, s' \in \mathcal{S}^{\mathrm{G}}} \left\{\mathbb{P}^{\bar{\pi}}\left(s_h = s \mid s_\ell = s', a_\ell = \pi_h^{\mathrm{E}}(s)\right)\right\}$. Here $\mathbb{P}^{\bar{\pi}}\left(s_h = s \mid s_\ell = s', a_\ell = \pi_h^{\mathrm{E}}(s)\right)$ is the visitation probability of s in time step h by starting from $s', \pi_h^{\mathrm{E}}(s)$ in time step ℓ, which is jointly determined by the transition function and policy $\bar{\pi}$.

- The low data regime imitation gap bound **still holds for approximate solutions** with $8\varepsilon/c(\bar{\pi})$ being the induced error
- We apply the dynamic programming analysis to establish the **growth condition of TV-AIL** $\mathrm{dist}\left(\bar{\pi}, \pi^{\mathrm{AIL}}\right) \leq f(\bar{\pi}) - f(\pi^{\mathrm{AIL}})$, which is highly technical and of independent interests

Imitation Gap Bound for Any MDP

🤔 Q: Can TV-AIL achieve the desired low data regime bound in all MDPs?

Proposition. Suppose that π^{AIL} is the policy output by TV-AIL, there exists an MDP such that

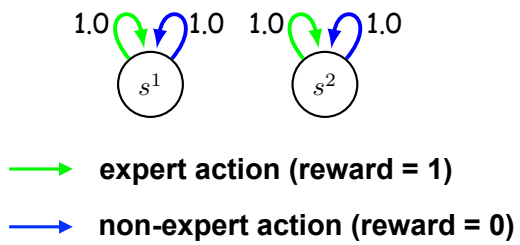
$$V\left(\pi^{\mathrm{E}}\right)-\mathbb{E}\left[V\left(\pi^{\mathrm{AIL}}\right)\right] \geqslant \Omega\left(\min\left\{H, H\sqrt{\frac{|\mathcal{S}|}{N}}\right\}\right)$$

A: **In the worst case**, TV-AIL must suffer an imitation gap of $\min\left\{H, H\sqrt{|\mathcal{S}|/N}\right\}$, which equals to H in the low data regime

Hard Instance for AIL

Assumption. For a tabular and episodic MDP and an expert policy, we assume that

- Each state is absorbing and each action has the same transitions, i.e., $\forall(s,a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$, we have $P_h(s|s,a) = 1$.
- For any state, a^1 is the expert action with a reward 1 and the others are non-expert actions with a reward 0.



- Why this instance is hard for AIL?

  1. Each action has the same absorbing transitions

  2. Key insight: state-action distributions becomes decoupled across stages

$$\forall(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], d_h^\pi(s,a) = \rho(s)\pi_h(a|s)$$

  3. Due to this decoupling, TV-AIL cannot leverage multi-step distribution matching to generalize beyond demonstrations

- Remark: This is instance is very artificial and rarely occur in practice where AIL exhibits excellent performance

Summary

Why AIL outperforms BC in the low-data regime?

- The **first** low-data regime imitation gap bound

Theorem. Under mild assumptions, AIL can achieve an imitation gap bound:

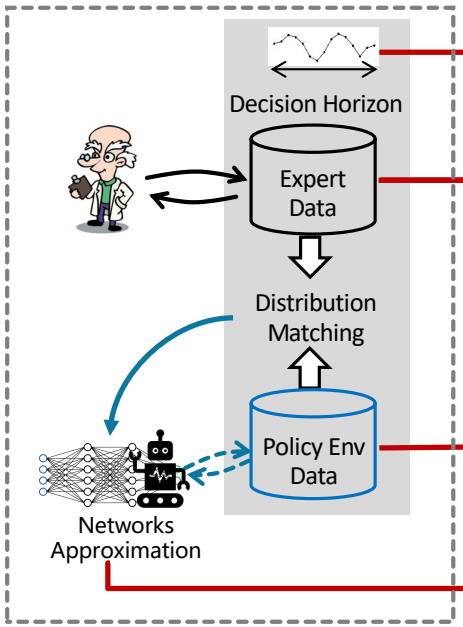
$$V\left(\pi^{\mathrm{E}}\right)-\mathbb{E}\left[V\left(\pi^{\mathrm{AIL}}\right)\right] \leqslant \mathcal{O}\left(\min\left\{1, \sqrt{\frac{|\mathcal{S}|}{N}}\right\}\right)$$

- A stage-coupled analysis

Reveal why AIL has good “generalization” through the lens of coupled “optimization”

- New insights
  - Algorithmic mechanism: **Future** distribution loss provably guides AIL to recover expert actions on **uncovered** states
  - Task understanding: **The coupling in state-action distributions** is the key for generalization

Our Exploration on Imitation Learning Theory



Long Decision Horizon	Prove AIL achieves a $\mathcal{O}(H\varepsilon)$ bound with **linear dependence on horizon**, provably mitigating BC's $\mathcal{O}(H^2\varepsilon)$ compounding errors [1, 2].
Limited Expert Demonstrations	Prove AIL achieves a **low data regime bound** $\mathcal{O}\left(\min\left\{1, \sqrt{\lvert\mathcal{S}\rvert/N}\right\}\right)$, theoretically explaining its superior performance with limited demos [3].
Imperfect Expert Demonstrations	Develop **data-section** and **data-augmentation** approaches that utilize imperfect data to improve performance theoretically and empirically [4,5].
Finite Env. Interactions	Develop a AIL method that achieves the **minimax optimal expert complexity and improved environment complexity** [6].
General Function Approximation	Develop a provably and practically efficient AIL approach with **general function approximation** [7].

[1] Tian Xu, Ziniu Li, Yang Yu. Error bounds of imitating policies and environments for reinforcement learning. TPAMI.
[2] Yi-Chen Li*, Tian Xu*, Yang Yu*, et al. Generalist reward models: found inside large language models. arXiv. 2025.
[3] Tian Xu*, Ziniu Li*, Yang Yu, Zhi-Quan Luo. Understanding AIL in Small Sample Regime: A Stage-coupled Analysis. TPAMI.
[4] Ziniu Li*, Tian Xu*, Yang Yu, Zhi-Quan Luo. Imitation Learning from Imperfection: Theoretical Justifications and Algorithms. NeurIPS 2023.
[5] Zhilong Zhang*, Tian Xu*, Xinghao Du*, et al. Improving Reward Model Generalization from Adversarial Process Enhanced Preferences. ICML 2025.
[6] Tian Xu*, Ziniu Li*, Yang Yu, Zhi-Quan Luo. Provably Efficient Adversarial Imitation Learning with Unknown Transitions. UAI 2023.
[7] Tian Xu*, Zhilong Zhang*, et al. Adversarial Imitation Learning with General Function Approximation: Theoretical Analysis and Practical Algorithms. TPAMI.