

# Imitation Learning from Observations by Minimizing Inverse Dynamics Disagreement

NeurIPS 2019

**Chao Yang**<sup>\*1</sup>, **Xiaojian Ma**<sup>\*1</sup>, **Wenbing Huang**<sup>\*1</sup>,  
**Fuchun Sun**<sup>1</sup>, **Huaping Liu**<sup>1</sup>, **Junzhou Huang**<sup>2</sup>, **Chuang Gan**<sup>3</sup>

*\*Denotes equal contribution*

<sup>1</sup>Tsinghua University, <sup>2</sup>Tencent AI Lab, <sup>3</sup>MIT-IBM Watson AI Lab



清华大学  
Tsinghua University



Tencent AI Lab



# Imitation learning (IL, LfD)

without reward

- MDP Formulation:  $\langle S, A, T(s'|s, a), \cancel{r(s, a)}, u, \gamma \rangle$
- An agent policy:  $\pi(a|s)$

# Imitation learning (IL, LfD)

without reward

- MDP Formulation:  $\langle S, A, T(s'|s, a), \cancel{r(s, a)}, u, \gamma \rangle$

- An agent policy:  $\pi(a|s)$

- Instead, a set of expert's demonstrations:

$$D = \{\tau_1, \dots, \tau_m\} = \{(s_0, a_0, s_1, a_1, \dots)\} \sim \pi_E(a|s)$$

- State-action or state-transition distribution of a policy  $\pi$ :

$$\rho_\pi(s, a) \text{ or } \rho_\pi(s, s')$$

- LfD goal: learning a policy of agent from expert demonstrations

# Imitation learning from observations (LfO)

- Given a set of expert's observations:

$$D = \{\tau_1, \dots, \tau_m\} = \{(s_0, \cancel{a_0}, s_1, \cancel{a_1}, \dots)\}$$

without actions

# Imitation learning from observations (LfO)

- Given a set of expert's observations:

$$D = \{\tau_1, \dots, \tau_m\} = \{(s_0, \cancel{a_0}, s_1, \cancel{a_1}, \dots)\}$$

without actions



# Imitation learning from observations (LfO)

- Given a set of **expert's observations**:

$$D = \{\tau_1, \dots, \tau_m\} = \{(s_0, \cancel{a_0}, s_1, \cancel{a_1}, \dots)\}$$

without actions

- **Advantage:**

- To save demo collection effort.
- Learning from internet videos.
- Human imitation never know what expert actions exactly are.



# Divergence minimization perspective on IL

GAIL or AIRL:

$$\min_{\pi} D_f(\rho_{\pi}(s, a) || \rho_E(s, a))$$

- $D_f$  could be KL or JS divergence.
- Adversarial training for divergence minimization.

Seyed & Zemel, CoRL'19  
Ho & Ermon, NIPS'16,  
Fu, Finn, ICLR'18, ICML'16

# Divergence minimization perspective on IL

GAIL or AIRL:

$$\min_{\pi} D_f(\rho_{\pi}(s, a) || \rho_E(s, a))$$

- $D_f$  could be KL or JS divergence.
- Adversarial training for divergence minimization.

Seyed & Zemel, CoRL'19  
Ho & Ermon, NIPS'16,  
Fu, Finn, ICLR'18, ICML'16



Intuitively generalize to LfO



# Divergence minimization perspective on IL

GAIL or AIRL:

$$\min_{\pi} D_f(\rho_{\pi}(s, a) || \rho_E(s, a))$$

- $D_f$  could be KL or JS divergence.
- Adversarial training for divergence minimization.

Seyed & Zemel, CoRL'19  
Ho & Ermon, NIPS'16,  
Fu, Finn, ICLR'18, ICML'16



Intuitively generalize to LfO

GAIfo:

$$\min_{\pi} D_f(\rho_{\pi}(s, s') || \rho_E(s, s'))$$

- Can bridging the state transition distribution of agent and expert sufficiently mimic expert behavior?

Torabi, et al. IJCAI'19

# Divergence minimization perspective on IL

GAIL or AIRL:

$$\min_{\pi} D_f(\rho_{\pi}(s, a) || \rho_E(s, a))$$

Saved & Zemel, CoRL'10

$$\text{IDD} = \text{GAIL} - \text{GAIfO}$$

$$D_f(\rho_{\pi}(a|s, s') || \rho_E(a|s, s')) = D_f(\rho_{\pi}(s, a) || \rho_E(s, a)) - D_f(\rho_{\pi}(s, s') || \rho_E(s, s'))$$

$\rho(a|s, s')$ : inverse dynamic model

$\pi$

- Can bridging the state transition distribution of agent and expert sufficiently mimic expert behavior?

Torabi, et al. IJCAI'19

$$\text{IDD} = \text{GAIL} - \text{GAIfO}$$

- **Inverse dynamic model disagreement (IDD):**
  - The gap between GAIL and GAIfO
  - GAIL is a upper-bound of GAIfO

$$\text{IDD} = \text{GAIL} - \text{GAIfO}$$

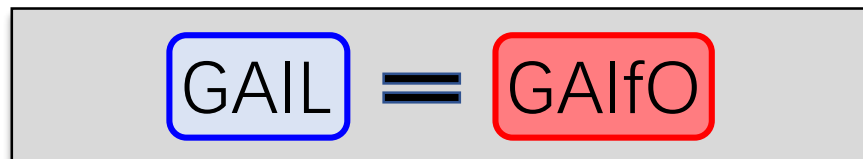
- **Inverse dynamic model disagreement (IDD):**

- The gap between GAIL and GAIfO
- GAIL is a upper-bound of GAIfO

- **IDD vanishment:**

- If the dynamics model  $T(s'|s, a)$  is an **injective** mapping,

$$D_f(\rho_\pi(s, a) \parallel \rho_E(s, a)) = D_f(\rho_\pi(s, s') \parallel \rho_E(s, s')).$$


$$\text{GAIL} = \text{GAIfO}$$

# Inverse dynamic disagreement minimization(IDDM)

- The overall objective is:

GAIfO

+

IDD :=  $D_f(\rho_\pi(a|s, s') || \rho_E(a|s, s')) \leq -\mathcal{H}_\pi(s, a) + \text{const.}$

# Inverse dynamic disagreement minimization (IDDM)

- The overall objective is:

$$\boxed{\text{GAIfO}} + \boxed{\text{IDD} := D_f(\rho_\pi(a|s, s') || \rho_E(a|s, s')) \leq \underline{-\mathcal{H}_\pi(s, a)} + \text{const.}}$$

Mutual information neural estimation  
[Belghazi, et al. ICML'18]

$$\min_{\pi} D_f(\rho_\pi(s, s') || \rho_E(s, s')) - \lambda_p \mathcal{H}_\pi(a|s) - \lambda_s I_\pi(s; (s', a))$$

# Inverse dynamic disagreement minimization (IDDM)

- The overall objective is:

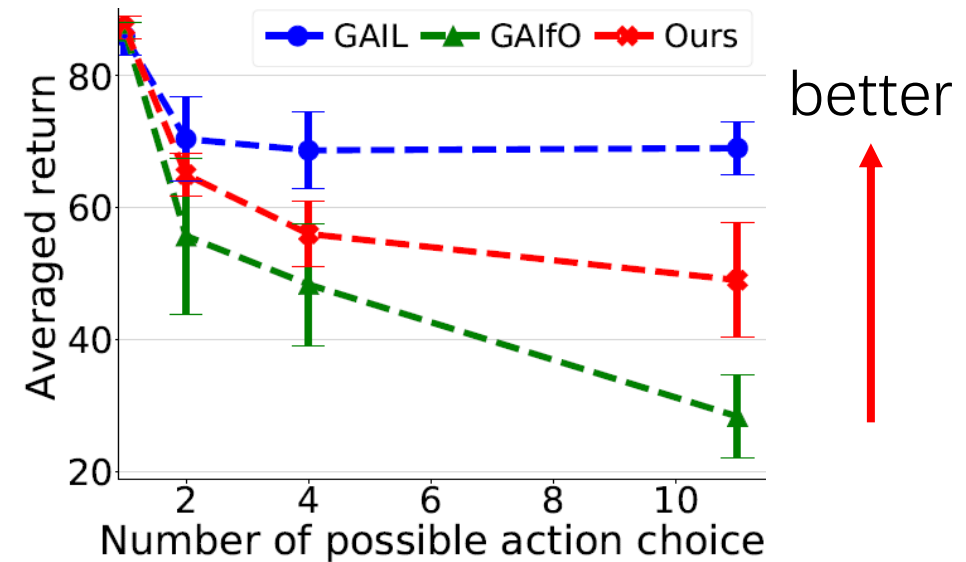
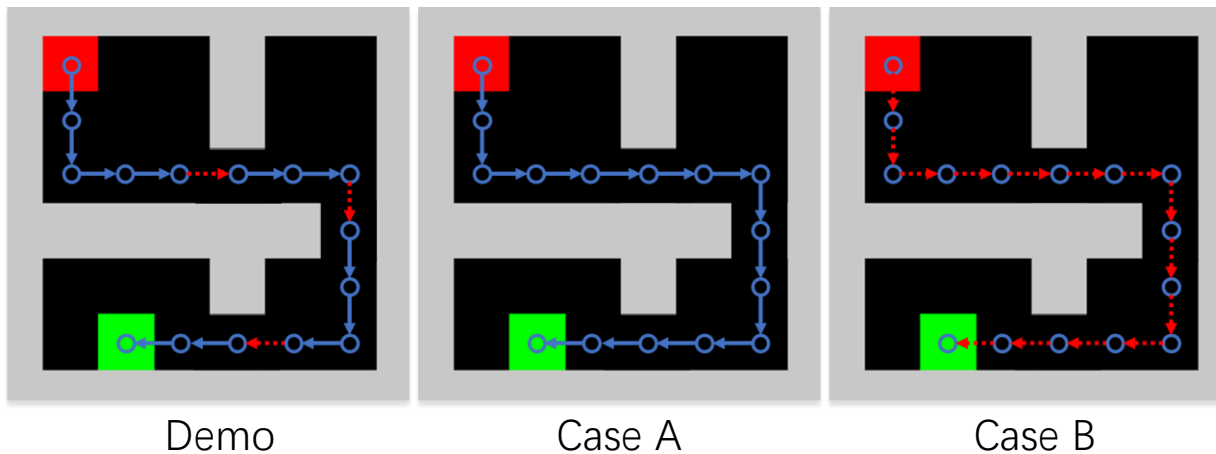
$$\boxed{\text{GAIfO}} + \boxed{\text{IDD} := D_f(\rho_\pi(a|s, s') || \rho_E(a|s, s')) \leq \underline{-\mathcal{H}_\pi(s, a)} + \text{const.}}$$

Mutual information neural estimation  
[Belghazi, et al. ICML'18]

$$\min_{\pi} D_f(\rho_\pi(s, s') || \rho_E(s, s')) - \lambda_p \mathcal{H}_\pi(a|s) - \lambda_s I_\pi(s; (s', a))$$

- Toy navigation

■ start state    ○→○ walk action    ■ end state    ○⋯○ jump action



# OpenAI-Gym tasks

- Quantitative performance (original reward)

	CartPole	Pendulum	DoublePendulum	Hopper	HalfCheetah	Ant
DeepMimic	-	731.0±19.0	454.4±154.0	2292.6±1068.9	202.6±4.4	-985.3±13.6
BCO	200.0±0.0	24.9±0.8	80.3±13.1	1266.2±1062.8	4557.2±90.0	562.5±384.1
GAIfo	197.5±7.3	980.2±3.0	4240.6±4525.6	1021.4±0.6	3955.1±22.1	-1415.0±161.1
GAIfo-s*	200.0±0.0	952.1±23.0	1089.2±51.4	1022.5±0.40	2896.5±53.8	-5062.3±56.9
<b>Ours</b>	<b>200.0±0.0</b>	<b>1000.0±0.0</b>	<b>9359.7±0.2</b>	<b>3300.9±52.1</b>	<b>5699.3±51.8</b>	<b>2800.4±14.0</b>
GAIL	200.0±0.0	1000.0±0.0	9174.8±1292.5	3249.9±34.0	6279.0±56.5	5508.8±791.5
Expert	200.0±0.0	1000.0±0.0	9318.8±8.5	3645.7±181.8	5988.7±61.8	5746.8±117.5

\*GAIfo with single state only.



# OpenAI-Gym tasks

- Quantitative performance (original reward)

	CartPole	Pendulum	DoublePendulum	Hopper	HalfCheetah	Ant
DeepMimic	-	731.0±19.0	454.4±154.0	2292.6±1068.9	202.6±4.4	-985.3±13.6
BCO	200.0±0.0	24.9±0.8	80.3±13.1	1266.2±1062.8	4557.2±90.0	562.5±384.1
GAIfo	197.5±7.3	980.2±3.0	4240.6±4525.6	1021.4±0.6	3955.1±22.1	-1415.0±161.1
GAIfo-s*	200.0±0.0	952.1±23.0	1089.2±51.4	1022.5±0.40	2896.5±53.8	-5062.3±56.9
<b>Ours</b>	<b>200.0±0.0</b>	<b>1000.0±0.0</b>	<b>9359.7±0.2</b>	<b>3300.9±52.1</b>	<b>5699.3±51.8</b>	<b>2800.4±14.0</b>
GAIL	200.0±0.0	1000.0±0.0	9174.8±1292.5	3249.9±34.0	6279.0±56.5	5508.8±791.5
Expert	200.0±0.0	1000.0±0.0	9318.8±8.5	3645.7±181.8	5988.7±61.8	5746.8±117.5

\*GAIfo with single state only.

- More learning curve, num of demos and ablation experiments can be found in our paper and supplementary.

# OpenAI-Gym tasks

- Quantitative performance (original reward)

**For more information, please come to our poster session!**

**Tue Dec 10th 05:30 -- 07:30 PM @ East Exhibition Hall B + C #205**

**Thanks**

- More learning curve, num of demos and ablation experiments can be found in our paper and supplementary.