Paper Sharing

Multi-Agent Adversarial Inverse Reinforcement Learning

Lecturer：Yuxin Wu

2022.10.27

Motivation

- Reinforcement Learning → Reward function

$$\max_\pi \mathbb{E}_\pi \left[\sum_{t=1}^{T} \gamma^t r(s_t, a_t) \right]$$

- Quite feasible in some simple scenarios

- Rather challenging in real world application

- Hand-tuning reward functions becomes increasingly more challenging

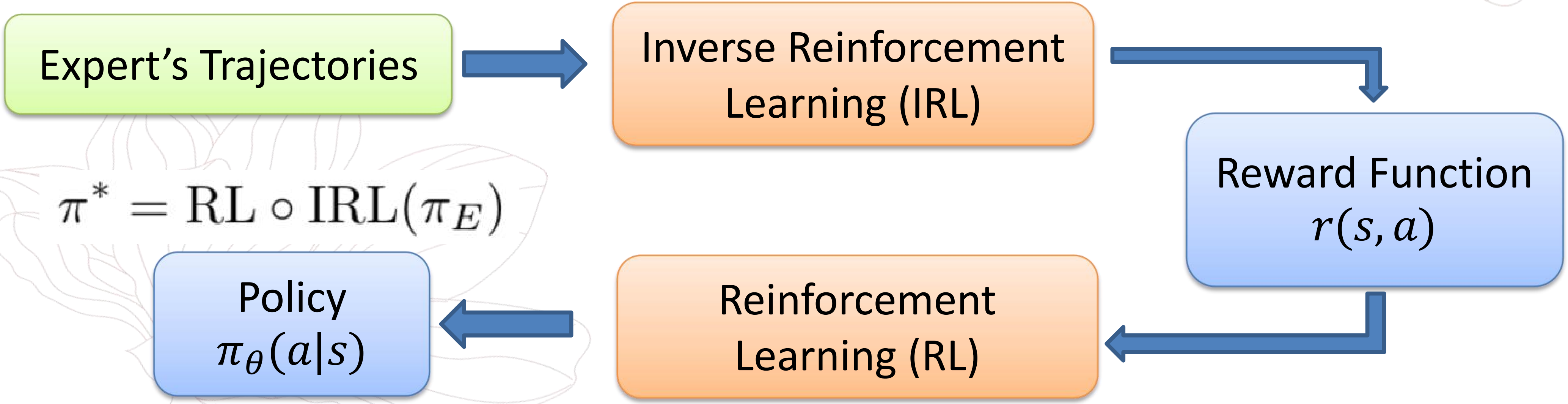
Motivation

- **Solution** ⇒ learning from expert demonstrations ⇒ imitation learning
  - **Behavior cloning, BC** $$\pi^* = \max_{\pi \in \Pi} \mathbb{E}_{\pi_E}[\log \pi(a|s)]$$
  - Inverse RL, IRL
  - Generative adversarial imitation learning, GAIL

- **However, BC does not recover any reward functions**

Motivation

- **Solution** ⇒ learning from expert demonstrations ⇒ imitation learning
  - Behavior cloning, BC $$\pi^* = \max_{\pi \in \Pi} \mathbb{E}_{\pi_E}[\log \pi(a|s)]$$
  - Inverse RL, IRL
  - Generative adversarial imitation learning, GAIL



$$\pi^* = \mathrm{RL} \circ \mathrm{IRL}(\pi_E)$$

Motivation

- **Solution** ⇨ learning from expert demonstrations ⇨ imitation learning
  - Behavior cloning, BC $\pi^* = \max_{\pi \in \Pi} \mathbb{E}_{\pi_E}[\log \pi(a|s)]$
  - Inverse RL, IRL
  - **Generative adversarial imitation learning, GAIL**

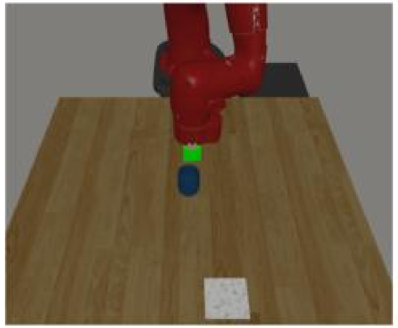
Expert's Trajectories ⇨ State-Action Distribution Matching with GAN ⇦ Policy $\pi_\theta(a|s)$

Can recover any reward function? X

R:at the optimality, the discriminator will converge to a non-informative uniform distribution.

Motivation

- Why should we care about reward learning?
  - Scientific inquiry: human and animal behavioral study, inferring intentions, etc.
  - Presupposition: reward function is considered to be the most succinct, robust and transferable description of the task.



$$r^* = (\text{object\_pos} - \text{goal\_pos})^2$$

VS.

$$\pi^* : \mathcal{S} \to \mathcal{P}(\mathcal{A})$$

  - Re-optimizing policies in new environments, debugging and analyzing imitation learning algorithms, etc.
  - These properties are even more desirable in the multi-agent settings.

Preliminaries

- Single-Agent Inverse RL
  - Basic principle: find a reward function that explains the expert behaviors

    ill-defined ⟶ there can be many reward functions that can explain the same set of behaviors

  - Maximum Entropy Inverse RL (MaxEnt IRL) provides a general probabilistic framework to solve the ambiguity.

$$p_\omega(\tau) \propto \left[\eta(s^1)\prod_{t=1}^{T} P(s^{t+1}|s^t, a^t)\right] \exp\left(\sum_{t=1}^{T} r_\omega(s^t, a^t)\right)$$

$$\max_\omega \mathbb{E}_{\pi_E}\left[\log p_\omega(\tau)\right] = \mathbb{E}_{\tau\sim\pi_E}\left[\sum_{t=1}^{T} r_\omega(s^t, a^t)\right] - \log Z_\omega$$

  - where Z_ω is the partition function ⟶ Intractable

Preliminaries

- **Single-Agent Inverse RL**
  - Adversarial inverse reinforcement learning provides an efficient sampling-based approximations to MaxEnt IRL

Special discriminator structure:

$$D_{\omega,\phi}(s,a,s') = \frac{\exp(f_{\omega,\phi}(s,a,s'))}{\exp(f_{\omega,\phi}(s,a,s')) + \pi(a|s)}$$

$$f_{\omega,\phi}(s,a,s') = r_\omega(s,a) + \gamma h_\phi(s') - h_\phi(s)$$

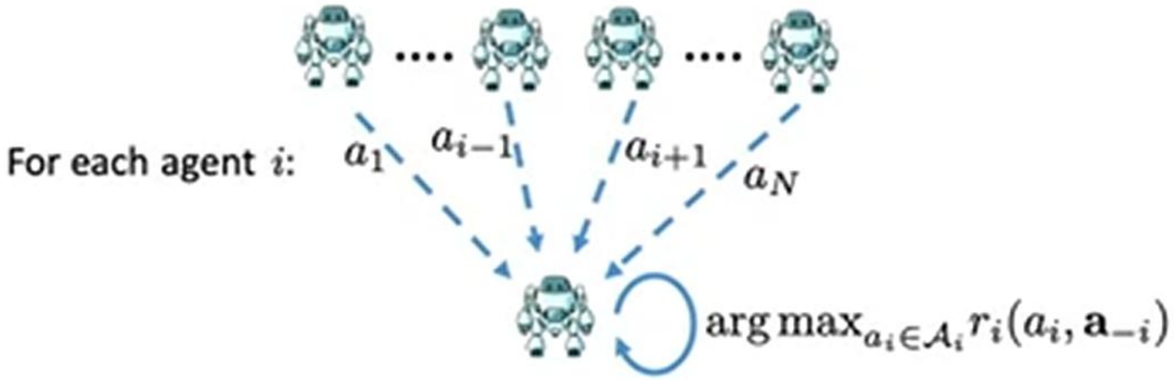
  - Train the policy (generator) with $\log D - \log(1-D)$
  - Under certain conditions, $r_\omega(s,a)$ is guaranteed to recover the ground-truth reward up to a constant.

Preliminaries

- Markov Games [Littman, 1994]: A multi-agent generalization to markov decision process
  - Agent number N
  - State space $\mathcal{S}$
  - Action spaces $\{\mathcal{A}_i\}_{i=1}^{N}$
  - Transition dynamics $P : \mathcal{S} \times \mathcal{A}_1 \times \ldots \times \mathcal{A}_N \to \mathcal{P}(\mathcal{S})$
  - Initial state distribution $\eta \in \mathcal{P}(\mathcal{S})$

Preliminaries

- Solution Concepts to Markov Games
  - Nash equilibrium (NE) [Hu et al, 1998]: no agent can achieve higher expected reward through unilaterally changing its own policy.
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:



Preliminaries

- Solution Concepts to Markov Games
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:

	B1	B2	B3
A1	<u>1</u>,<u>1</u>	<u>2</u>,0	-1,-1
A2	0,<u>2</u>	-1,-1	<u>2</u>,1
A3	-1,1	1,<u>2</u>	1,0

Preliminaries

- Solution Concepts to Markov Games
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:

fix

	B1	B2	B3
A1	$\underline{1},\underline{1}$	$\underline{2},0$	-1,-1
A2	$0,\underline{2}$	-1,-1	$\underline{2},1$
A3	-1,1	$1,\underline{2}$	1,0

Preliminaries

- Solution Concepts to Markov Games
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:

fix

	B1	B2	B3
A1	1,1	2,0	-1,-1
A2	0,2	-1,-1	2,1
A3	-1,1	1,2	1,0

Preliminaries

- Solution Concepts to Markov Games
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:

fix

	B1	B2	B3
A1	<u>1</u>,<u>1</u>	<u>2</u>,0	-1,-1
A2	0,<u>2</u>	-1,-1	<u>2</u>,1
A3	-1,1	1,<u>2</u>	1,0

Preliminaries

- Solution Concepts to Markov Games
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:

fix

	B1	B2	B3
A1	1,1	2,0	-1,-1
A2	0,2	-1,-1	2,1
A3	-1,1	1,2	1,0

Preliminaries

- Solution Concepts to Markov Games
  - Best response dynamics [Nisan et al, 2011; Gandhi,2012]:

	B1	B2	B3
A1	1,1	2,0	-1,-1
A2	0,2	-1,-1	2,1
A3	-1,1	1,2	1,0

Preliminaries

- Solution Concepts to Markov Games
  - Nash equilibrium (NE) [Hu et al, 1998]: No agent can achieve higher expected reward through unilaterally changing its own policy.
  - Correlated equilibrium (CE) [Aumann, 1974]: A relaxation to NE, which allows extra coordination signals
- NE and CE are *incompatible* with MaxEnt IRL

Method

- Logistic Stochastic Best Response Equilibrium (LSBRE)

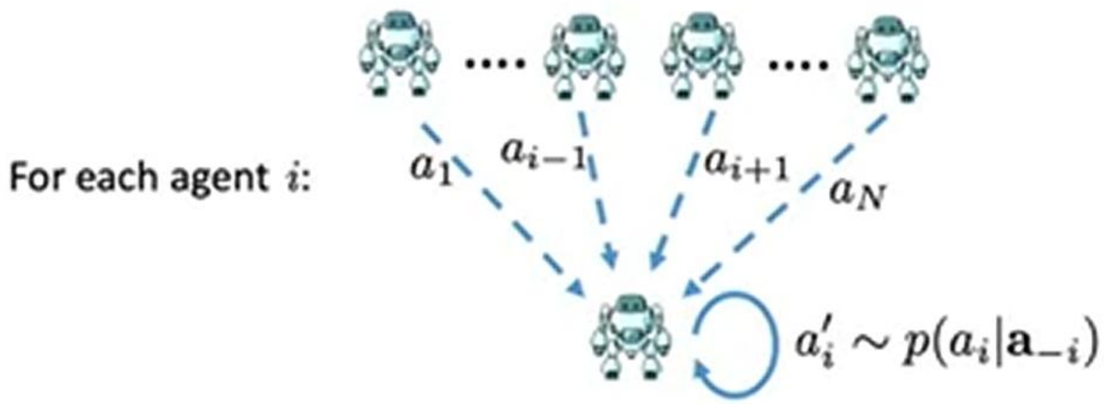
  -motivated by:

  - Logistic quantal response equilibrium (LQRE):A stochastic generalization to NE and CE [McKelvey & Palfrey, 1995; 1998]
  - Gibbs sampling [Hastings, 1970]
  - Dependency networks [Heckerman et al, 2000]
  - Best response dynamics [Nisan et al, 2011]

- LSBRE is *compatible* with MaxEnt IRL

Method

- **Logistic Stochastic Best Response Equilibrium (LSBRE)**
  - Single-shot normal-form game:



$$a_i' \sim p(a_i|\mathbf{a}_{-i}) = \frac{\exp(\lambda r_i(a_i, \mathbf{a}_{-i}))}{\sum_{a \in \mathcal{A}_i} \exp(\lambda r_i(a, \mathbf{a}_{-i}))}$$

Softmax action selection (MaxEnt RL)

Because the markov chain is ergodic, it admits a unique stationary joint policy

Method

- **Logistic Stochastic Best Response Equilibrium (LSBRE)**
  - Example (Single-shot normal-form game) :

	B1	B2	B3
A1	1,1	2,0	-1,-1
A2	0,2	-1,-1	2,1
A3	-1,1	1,2	1,0

$p\ (A|B_1)\ =\ [0.67,\ 0.24,\ 0.09]$

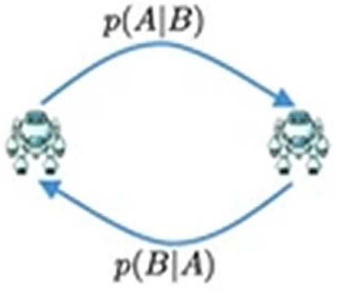
$p\ (A|B_2)\ =\ [0.71,\ 0.03,\ 0.26]$

$p\ (A|B_3)\ =\ [0.03,\ 0.71,\ 0.26]$

$p\ (B|A_1)\ =\ [0.67,\ 0.24,\ 0.09]$

$p\ (B|A_2)\ =\ [0.71,\ 0.03,\ 0.26]$

$p\ (B|A_3)\ =\ [0.24,\ 0.67,\ 0.09]$



Can also extend to Markov Games with a sequence of Markov chains and action-value functions!

Method

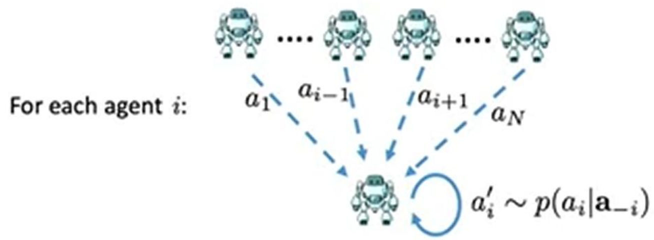
- MaxEnt with LSBRE
  - Multi-Agent Adversarial Inverse RL:
    - By parameterizing the reward functions with ω, the trajectory distribution under LSBRE is given by:

$$p(\tau) = \eta(s^1) \cdot \prod_{t=1}^{T} \boldsymbol{\pi}^t(\boldsymbol{a}^t|s^t;\boldsymbol{\omega}) \cdot \prod_{t=1}^{T} P(s^{t+1}|s^t,\boldsymbol{a}^t)$$

    - Maximizing the likelihood of expert demonstrations corresponds to:

$$\max_{\boldsymbol{\omega}} \mathbb{E}_{\tau\sim\boldsymbol{\pi}_E}\left[\sum_{t=1}^{T} \log \boldsymbol{\pi}^t(\boldsymbol{a}^t|s^t;\boldsymbol{\omega})\right]$$

  - However, maxmizing the joint likelihood is *intractable*



Method

- **Pseudolikelihood Maximization**
  - Multi-Agent Adversarial Inverse RL:
    - Bridging the optimization of joint likelihood and each conditional likelihood with maximum pseudolikelihood estimation (Theorem 2)

Theorem 2. *Let demonstrations $\tau_1, \ldots, \tau_M$ be independent and identically distributed (sampled from LSBRE induced by some unknown reward functions), and suppose that for all $t \in [1, \ldots, T], a_i^t \in \mathcal{A}_i$, $\pi_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t; \omega_i)$ is differentiable with respect to ω_i. Then, with probability tending to 1 as $M \to \infty$, the equation*

$$\frac{\partial}{\partial \omega} \sum_{m=1}^{M} \sum_{t=1}^{T} \sum_{i=1}^{N} \log \pi_i^t(a_i^{m,t}|\boldsymbol{a}_{-i}^{m,t}, s^{m,t}; \omega_i) = 0 \qquad (9)$$

has a root $\hat{\omega}_M$ such that $\hat{\omega}_M$ tends to the maximizer of the joint likelihood in Equation (8).

Method

- **Multi-Agent Adversarial Inverse RL:**

  - Maximizing the pseudolikelihood objective:

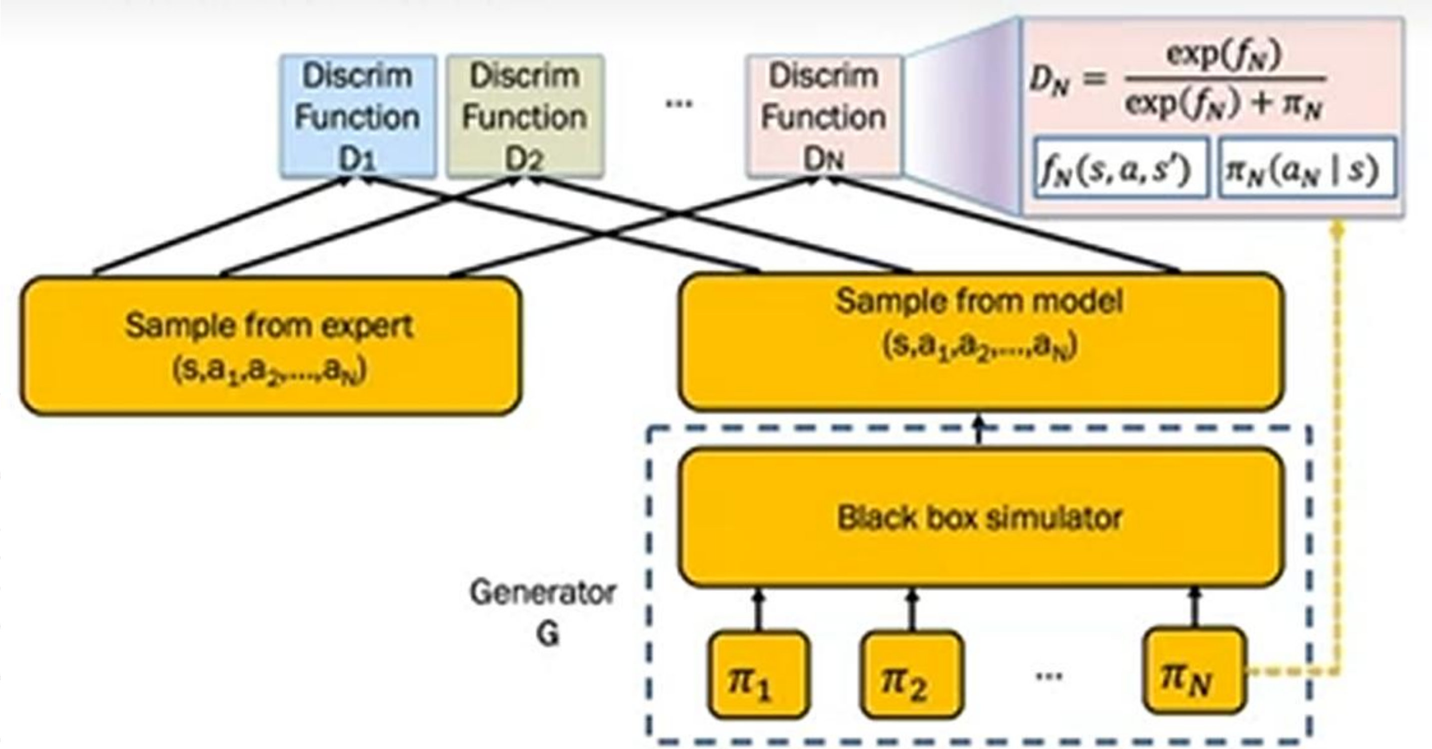
$$\mathbb{E}_{\boldsymbol{\pi}_E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial}{\partial\omega}\log\pi_i^t(a_i^t|\boldsymbol{a}_{-i}^t,s^t;\omega_i)\right]$$

  - By characterizing the trajectory distribution of LSBRE (Theorem 1), we can optimize the following surrogate loss:

$$\mathbb{E}_{\boldsymbol{\pi}_E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial}{\partial\omega}r_i(s^t,\boldsymbol{a}^t;\omega_i)\right]-\sum_{i=1}^{N}\frac{\partial}{\partial\omega}\log Z_{\omega_i}$$

Method

- **Multi-Agent Adversarial Inverse RL:**
  - Practical MA-AIRL Framework:



Experiments

- **Policy imitation performance:**
  - Cooperative tasks: cooperative navigation & cooperative communication,
  - Use the ground-truth reward as the oracle evaluation metric.

Table 1. Expected returns in cooperative tasks. Mean and variance are taken across different random seeds used to train the policies.

Algorithm	Nav. ExpRet	Comm. ExpRet
Expert	-43.195 $\pm$ 2.659	-12.712 $\pm$ 1.613
Random	-391.314 $\pm$ 10.092	-125.825 $\pm$ 3.4906
MA-GAIL	-52.810 $\pm$ 2.981	-12.811 $\pm$ 1.604
MA-AIRL	**-47.515** $\pm$ 2.549	**-12.727** $\pm$ 1.557

Experiments

- Policy imitation performance:
  - Competitive task (competitive keep-away)
  - “Battle” evaluation: let the modle play against the experts:a learned policy is considered better if it receives a higher expected return than its opponent.

Table 2. Expected returns of the agents in competitive task. Agent #1 represents the agent trying to reach the target and Agent #2 represents the adversary. Mean and variance are taken across different random seeds.

Agent #1	Agent #2	Agent #1 ExpRet
Expert	Expert	-6.804 ± 0.316
MA-GAIL	Expert	-6.978 ± 0.305
MA-AIRL	Expert	$\mathbf{-6.785} \pm 0.312$
Expert	MA-GAIL	-6.919 ± 0.298
Expert	MA-AIRL	$\mathbf{-7.367} \pm 0.311$

Experiments

- Reward recovery:
  - Measuring the statistical correlation between the learned reward and the ground-truth.
  - A more direct evaluation in multi-agent system.

- Two Examples:
  - Pearson 's correlation coefficient (PCC): measures the linear correlation between two random variables.
  - Spearman 's rank correlation coefficient (SCC): measures the statistical dependence between the rankings of two random variables.

Experiments

- **Reward recovery:**

  - Cooperative tasks

Table 3. Statistical correlations between the learned reward functions and the ground-truth rewards in cooperative tasks. Mean and variance are taken across N independently learned reward functions for N agents.

Task	Metric	MA-GAIL	MA-AIRL
Nav.	SCC	0.792 ± 0.085	$\mathbf{0.934} \pm 0.015$
	PCC	0.556 ± 0.081	$\mathbf{0.882} \pm 0.028$
Comm.	SCC	0.879 ± 0.059	$\mathbf{0.936} \pm 0.080$
	PCC	0.612 ± 0.093	$\mathbf{0.848} \pm 0.099$

  - Competitive tasks

Table 4. Statistical correlations between the learned reward functions and the ground-truth rewards in competitive task.

Algorithm	MA-GAIL	MA-AIRL
SCC #1	0.424	**0.534**
SCC #2	0.653	**0.907**
Average SCC	0.538	**0.721**
PCC #1	0.497	**0.720**
PCC #2	0.392	**0.667**
Average PCC	0.445	**0.694**

Summary

- The paper proposed a new solution concept for Markov games, which allows us to characterize the trajectory distribution induced by parameterized rewards.
- The paper propose the first multi-agent MaxEnt IRL framework, which is effective and scalable to Markov games with continuous state-action space and unknown dynamics.
- The paper employ maximum pseudolikelihood estimation and adversarial reward learning to achieve tractability.
- Experimental results demonstrate that MA-AIRL can recover both policy and reward function that is highly correlated with the ground-truth.

Thank you for listening

主講人：吳雨欣

2022.10.27