Multi-Agent Adversarial Inverse Reinforcement Learning

Lantao Yu, Jiaming Song, Stefano Ermon
Department of Computer Science, Stanford University

Contact: lantaoyu@cs.stanford.edu

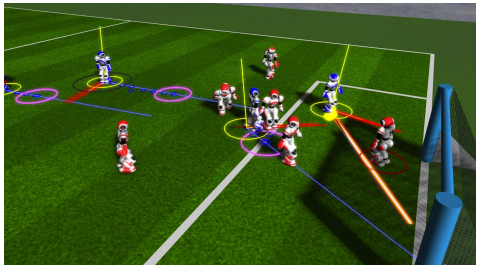
Motivation

- By definition, the performance of RL agents heavily relies on the quality of reward functions.

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

Computer Games

Dialogue



Multi-Agent System

- In many real-world scenarios, especially in multi-agent settings, hand-tuning informative reward functions can be very challenging.
- Solution: learning from expert demonstrations!

Motivation

- Imitation learning does not recover reward functions.
  - Behavior Cloning

$$\pi^* = \max_{\pi \in \Pi} \mathbb{E}_{\pi_E}[\log \pi(a|s)]$$

  - Generative Adversarial Imitation Learning [Ho & Ermon, 2016]

$$\mathrm{IRL}(\pi_E) = \arg\max_{r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} -\psi(r) + \mathbb{E}_{\pi_E}[r(s,a)] - \mathrm{RL}(r)$$

$$\mathrm{RL}(r) = \max_{\pi \in \Pi} \mathcal{H}(\pi) + \mathbb{E}_{\pi}[r(s,a)]$$

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

$$\pi_E(a|s) \xrightarrow{\text{IRL}} r(s,a) \xrightarrow{\text{RL}} \pi(a|s)$$

Motivation

- Imitation learning does not recover reward functions.
  - Behavior Cloning

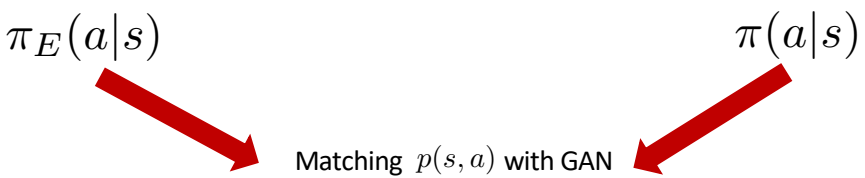
$$\pi^* = \max_{\pi \in \Pi} \mathbb{E}_{\pi_E}[\log \pi(a|s)]$$

  - Generative Adversarial Imitation Learning [Ho & Ermon, 2016]

$$\mathrm{IRL}(\pi_E) = \arg\max_{r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} -\psi(r) + \mathbb{E}_{\pi_E}[r(s,a)] - \mathrm{RL}(r)$$

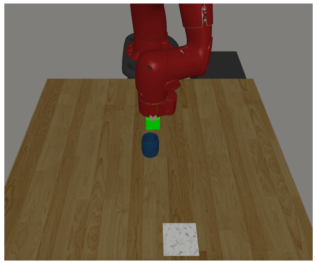
$$\mathrm{RL}(r) = \max_{\pi \in \Pi} \mathcal{H}(\pi) + \mathbb{E}_{\pi}[r(s,a)]$$

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



Motivation

- Why should we care 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. [Abbeel & Ng, 2014]



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

VS.

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

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

Preliminaries

- Single-Agent Inverse RL
  - Adversarial Inverse RL (AIRL) provides an efficient sampling-based approximation 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
  - *Correlated equilibrium* (CE) [Aumann, 1974]: A joint strategy profile, where no agent can achieve higher expected reward through unilaterally changing its own policy.
  - *Nash equilibrium* (NE) [Hu et al, 1998]: A more restrictive equilibrium which further requires agents' actions in each state to be independent.
  - *Incompatible* with MaxEnt IRL.

Preliminaries

- Solution Concepts to Markov Games
  - Logistic quantal response equilibrium (LQRE) [McKelvey & Palfrey, 1995; 1998]: A stochastic generalization to NE and CE.
  - LQRE is a joint strategy profile satisfying the set of constraints:

$$\pi_i(a_i|s) = \frac{\exp\left(\lambda \mathrm{ExpRet}_i^{\boldsymbol{\pi}}(s, a_i, \boldsymbol{a}_{-i})\right)}{\sum_{a_i'} \exp\left(\lambda \mathrm{ExpRet}_i^{\boldsymbol{\pi}}(s, a_i', \boldsymbol{a}_{-i})\right)}$$

$$\mathrm{ExpRet}_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) = \mathbb{E}_{s^{t+1:T}, \boldsymbol{a}^{t+1:T}}\left[\sum_{l \geq t} \gamma^{l-t} r_i(s^l, \boldsymbol{a}^l) | s_t, \boldsymbol{a}_t, \boldsymbol{\pi}\right]$$

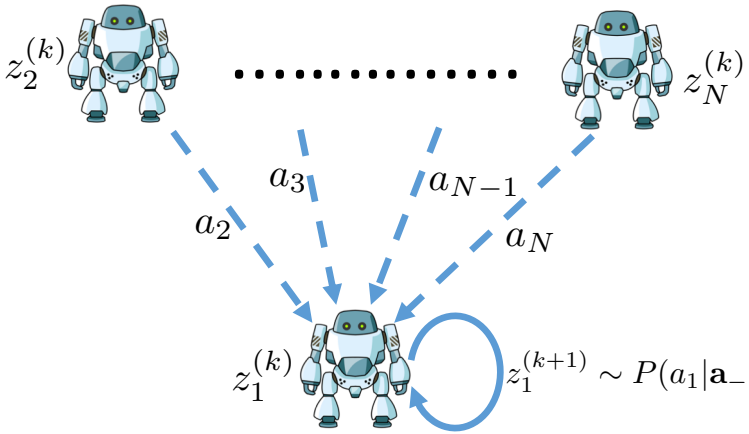
- Existing optimality notions do not *explicitly* define a tractable joint strategy profile, which we can use to maximize the likelihood of expert demonstrations.

Method

- Logistic Stochastic Best Response Equilibrium
  - Motivated by LQRE, Gibbs sampling [Hastings, 1970], dependency networks [Heckerman et al, 2000] and best response dynamics [Nisan et al, 2011].

Method

- Logistic Stochastic Best Response Equilibrium
  - Single-shot normal-form game: Consider a *Markov chain* over $\mathcal{A}_1 \times \ldots \times \mathcal{A}_N$, where the state of the markov chain at step k is denoted $\mathbf{z}^{(k)} = (z_1, \cdots, z_N)^{(k)}$.



  - Because the markov chain is ergodic, it admits a unique stationary joint policy, which we call a *LSBRE for normal-form game*.

Method

- Logistic Stochastic Best Response Equilibrium
  - Markov game: Consider T markov chains over $(\mathcal{A}_1 \times \cdots \mathcal{A}_N)^{|\mathcal{S}|}$, where the state of the $t\text{-}th$ markov chain at step k is $\{z_i^{t,(k)} : \mathcal{S} \to \mathcal{A}_i\}_{i=1}^N$.
  - For $t \in [T, \ldots, 1]$, we recursively define the $t\text{-}th$ markov chain with the following update rule:

$$z_i^{t,(k+1)}(s^t) \sim P_i^t(a_i^t | \boldsymbol{a}_{-i}^t = \boldsymbol{z}_{-i}^{t,(k)}(s^t), s^t) = \frac{\exp\left(\lambda Q_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, a_i^t, \boldsymbol{z}_{-i}^{t,(k)}(s^t))\right)}{\sum_{a_i'} \exp\left(\lambda Q_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, a_i', \boldsymbol{z}_{-i}^{t,(k)}(s^t)\right)}$$

  - We define the unique stationary joint distribution of the markov chains as LSBRE strategy profiles:

$$\boldsymbol{\pi}^t(a_1, \cdots, a_N | s^t) = P\left(\bigcap_i \{z_i^{t,(\infty)}(s^t) = a_i\}\right)$$

Method

- 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:

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

Method

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

    Let $\tau_1, \ldots, \tau_M$ be i.i.d. sampled from LSBRE induced by some unknown reward function.
    Suppose that $\pi_i^t(a_i^t|\mathbf{a}_{-i}^t, s^t; \omega_i)$ is differentiable *w.r.t.* ω_i.
    Then as $M \to \infty$, with probability tending to 1, the equation

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

    has a root that tends to be the maximizer of joint likelihood.

Method

- Multi-Agent Adversarial Inverse RL
  - Maximizing the pseudolikelihood objective:

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

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

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

Method

- Multi-Agent Adversarial Inverse RL
  - Practical MA-AIRL Framework
    - Train the ω-parameterized discriminators as:

$$\max_{\omega} \ \mathbb{E}_{\pi_E}\left[\sum_{i=1}^{N}\log\frac{\exp(f_{\omega_i}(s,\mathbf{a}))}{\exp(f_{\omega_i}(s,\mathbf{a}))+q_{\theta_i}(a_i|s)}\right]+\mathbb{E}_{\mathbf{q}_\theta}\left[\sum_{i=1}^{N}\log\frac{q_{\theta_i}(a_i|s)}{\exp(f_{\omega_i}(s,\mathbf{a}))+q_{\theta_i}(a_i|s)}\right]$$

    - Train the θ-parameterized generators (policies) as:

$$\begin{aligned}\max_{\theta} \ \mathbb{E}_{\mathbf{q}_\theta}&\left[\sum_{i=1}^{N}\log(D_{\omega_i}(s,\mathbf{a}))-\log(1-D_{\omega_i}(s,\mathbf{a}))\right]\\ &=\mathbb{E}_{\mathbf{q}_\theta}\left[\sum_{i=1}^{N}f_{\omega_i}(s,\mathbf{a})-\log(q_{\theta_i}(a_i|s))\right]\end{aligned}$$

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: we place the experts and learned policies in the same environment; 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 $\pm$ 0.316
MA-GAIL	Expert	-6.978 $\pm$ 0.305
MA-AIRL	Expert	**-6.785** $\pm$ 0.312
Expert	MA-GAIL	-6.919 $\pm$ 0.298
Expert	MA-AIRL	**-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.
  - *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	$\mathbf{0.848} \pm 0.099$

Experiments

- Reward recovery
  - Competitive task

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

- We proposed a new solution concept for Markov games, which allows us to characterize the trajectory distribution induced by parameterized rewards.
- We propose the first multi-agent MaxEnt IRL framework, which is effective and scalable to Markov games with continuous state-action space and unknown dynamics.
- We 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!

Lantao Yu, Jiaming Song, Stefano Ermon
Department of Computer Science, Stanford University

Poster: 06:30 -- 09:00 PM @ Pacific Ballroom #36

Lantao Yu, Jiaming Song, Stefano Ermon. Multi-Agent Adversarial Inverse Reinforcement Learning. ICML 2019.