Multi-Agent Generative Adversarial Imitation Learning

温昭晋 2021.02.20

Background

Most existing approaches are not applicable because of:

——multiple (Nash) equilibria

——non-stationary environments

Dataset:
——demonstrations of a set of experts interacting with each other within the same environment.

Preliminaries

1. Markov Games for N Agents

- $T : \mathbb{S} \times \mathbb{A}_1 \times \cdots \times \mathbb{A}_N \to \mathcal{P}(\mathbb{S})$ transition process between states
- $r_i : \mathbb{S} \times \mathbb{A}_1 \times \cdots \times \mathbb{A}_N \to \mathbb{R}$ bounded reward function
- $\pi_i : \mathbb{S} \times \mathbb{A}_i \to [0,1]$
- $\boldsymbol{\pi}(a|s) = \prod_{i=1}^{N} \pi_i(a_i|s)$
- $(a_i, a_{-i}) = (a_1, \cdots, a_N)$ –i denotes all agents expect for i

$$E_\pi[r(s,a)] \triangleq E_{s_t, a_t \sim \pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$$

$$E_\pi[r(s,a) + \sum_{s'} T(s'|s,a) v(s')] \triangleq E_{a \sim \pi(\cdot|s)}[r(s,a) + \sum_{s'} T(s'|s,a) v(s')]$$

Preliminaries

2. Reinforcement learning and Nash equilibrium

$$RL(r) = \arg\max_{\pi \in \Pi} H(\pi) + E_{\pi}[r(s,a)]$$

γ −discounted Causal Entropy

$$H(\pi) \triangleq E_{\pi}[-lg\ \pi(a|s)] = E_{s_t, a_t \sim \pi}[-\sum_{t=0}^{\infty} \gamma^t lg\ \pi(a_t|s_t)]$$

Nash equilibrium

$$\forall i \in [1, N],\ \forall \pi_i^{\square} \neq \pi_i,\ E_{\pi_i, \pi_{-i}}[r_i] \geq E_{\pi_i^{\square}, \pi_{-i}}[r_i]$$

Preliminaries

2. Reinforcement learning and Nash equilibrium

Nash equilibrium

$$\min_{\pi \in \Pi, v \in R^{S \times N}} f_r(\pi, v) = \sum_{i=1}^{N} \left(\sum_{s \in \mathbb{S}} v_i(s) - E_{a_i \sim \pi_i(\cdot|s)} q_i(s, a) \right)$$

$$v_i(s) \geq q_i(s, a_i) \triangleq E_{\pi_{-i}} [r_i(s, a) + \gamma \sum_{s' \in \mathbb{S}} T(s'|s, a) v_i(s')] \qquad \forall i \in [N], s \in \mathbb{S}, a_i \in \mathbb{A}$$

$$a \triangleq (a_i, a_{-i}) \triangleq (a_1, \cdots, a_N) \qquad v \triangleq [v_1; \cdots; v_N]$$

3. GAIL

$$RL(r) = \underset{\pi \in \Pi}{\operatorname{argmax}}\, H(\pi) + E_{\pi}[r(s,a)]$$

$$IB_{\psi}(\pi_E) = \underset{r \in R^{\mathcal{S}\times\mathcal{A}}}{\operatorname{argmax}} -\psi(r) + E_{\pi_E}[r(s,a)] - (\max_{\pi \in \Pi} H(\pi) + E_{\pi}[r(s,a)])$$

$$RL \circ IB = \underset{\pi \in \Pi}{\operatorname{argmin}} -H(\pi) + \psi^{*}(\rho_{\pi} - \rho_{\pi_E})$$

$$\psi^{*}_{GA}(\rho_{\pi} - \rho_{\pi_E}) = \max_{D \in (0,1)^{\mathcal{S}\times\mathcal{A}}} E_{\pi_E}[\log(D(s,a))] + E_{\pi}[\log(1 - D(s,a))]$$

Generalization

多智能体环境中每个智能体都有一个奖赏函数，需要寻找一个均衡策略，将MARL目标定义为寻找满足纳什均衡且熵最大的策略集合。

$$MARL(\boldsymbol{r}) = \underset{\boldsymbol{\pi}\in\Pi,\boldsymbol{v}\in R^{S\times N}}{\operatorname{argmin}} f_r(\boldsymbol{\pi},\boldsymbol{v}) - H(\boldsymbol{\pi})$$

$$v_i(s) \geq q_i(s, a_i), \forall i \in [N], s \in \mathbb{S}, a_i \in \mathbb{A}_i$$

因此，我们的目标是定义一个类似于IRL的MAIRL算子，在专家策略和其他策略中构建margin。

用拉格朗日方法将约束转移到目标函数中

Generalization

Step1: 将约束中的1步约束替换为t+1步约束

定理：对固定的策略π和奖赏r, $v_i(s;\pi,r)$为唯一的bellman equation

$$v_i(s;\pi,r) = E_\pi[r_i(s,a) + \gamma \sum_{s' \in S} T(s'|s,a)\, v_i(s';\pi,r)] \qquad \forall s \in S$$

定义：

$$q_i^{(t)}(\{s^{(j)}, a^{(j)}\}_{j=0}^{t-1}, s^{(t)}, a_i^{(t)};\pi,r) =$$
$$\sum_{j=0}^{t-1} \gamma^j r_i(s^{(j)}, a^{(j)}) + \gamma^t E_{\pi_{-i}}[r_i(s^{(t)}, a^{(t)}) + \gamma \sum_{s' \in S} T(s'|s, a^{(t)}) v_i(s';\pi,r)]$$

则π为纳什均衡策略当且仅当：

$$v_i(s^{(0)};\pi,r) \geq E_{\pi_{-i}}[q_i^{(t)}(\{s^{(j)}, a^{(j)}\}_{j=0}^{t-1}, s^{(t)}, a_i^{(t)};\pi,r)] \triangleq Q_i^{(t)}(\{s^{(j)}, a^{(j)}\}_{j=0}^{t};\pi,r)$$

$$\forall t \in \mathbb{N}^+, i \in [N], j \in [t], s^{(j)} \in S, a^{(j)} \in \mathcal{A}$$

Generalization

Step2: 考虑原问题的拉格朗日对偶问题

$$\max_{\lambda \geq 0} \min_{\pi} L_{\boldsymbol{r}}^{(t+1)}(\boldsymbol{\pi}, \lambda) \triangleq \sum_{i=1}^{N} \sum_{\tau_i \in \Gamma_i^t} \lambda(\tau_i)(Q_i^{(t)}(\tau_i; \boldsymbol{\pi}, \boldsymbol{r}) - v_i(s^{(0)}; \boldsymbol{\pi}, \boldsymbol{r}))$$

接下来给出了特定的λ，将该目标表达式转换为
与GAIL类似衡量两个策略之间距离的形式

Generalization

Step3:MAIRL

$$\lambda_{\pi}^{\star}(\tau_i) = \eta(s^{(0)})\pi_i(a_i^{(0)}|s^{(0)}) \prod_{j=1}^{t} \pi_i(a_i^{(j)}|s^{(j)}) \sum_{a_{-i}^{(j-1)}} T(s^{(j)}|s^{(j-1)}, a^{(j-1)})\pi_{-i}^{\star}(a_{-i}^{(j)}|s^{(j)})$$

λ是agent i 使用策略π_i,其他agent使用策略π^*_{-i}时生成轨迹τ_i的概率

$$\lim_{t\to\infty} L_r^{(t+1)}(\boldsymbol{\pi}^{\star}, \lambda_{\boldsymbol{\pi}}^{\star}) = \sum_{i=1}^{N} \mathbb{E}_{\pi_i,\pi_{-i}^{\star}}[r_i(s,a)] - \sum_{i=1}^{N} \mathbb{E}_{\pi_i^{\star},\pi_{-i}^{\star}}[r_i(s,a)]$$

$$\text{MAIRL} \quad \psi(\boldsymbol{\pi}_E) = \arg\max_{\boldsymbol{r}} -\psi(\boldsymbol{r}) + \sum_{i=1}^{N} (E_{\pi_E}[r_i]) - (\max_{\boldsymbol{\pi}} \sum_{i=1}^{N} \beta H_i(\pi_i) + E_{\pi_i,\pi_{E_{-i}}}[r_i])$$

Generalization

Step4:MAGAIL

设$\psi(\boldsymbol{r}) = \sum_{i=1}^{N} \psi_i(r_i), \psi_i$对$i \in [N]$均为凸函数，且假设MARL(r)对所有的$r \in MAB_{\psi}(\pi_E)$有唯一解，则

$$MAB_{\psi} \circ MAB_{\psi}(\pi_E) = \underset{\pi \in \Pi}{\operatorname{argmin}} \sum_{i=1}^{N} -\beta H_i(\pi_i) + \psi_i^*(\rho_{\pi_i, E_{-i}} - \rho_{\pi_E})$$

由于训练时无法得知专家策略，$\rho_{\pi_i, E_{-i}}$未知，因此采用GAIL中类似的ψ正则项，并摒弃熵项。

Generalization

Step4:MAGAIL

若$\beta = 0$，且$\psi(\boldsymbol{r}) = \sum_{i=1}^{N} \psi_i(r_i)$, 其中$\psi_i(r_i) = E_{\pi_E}[g(r_i)]$, $if\ r_i > 0; +\infty\ otherwise$

$$g(x) = \begin{cases} -x - \log(1 - e^x) & if\ r_i > 0 \\ +\infty & otherwise \end{cases}$$

则，

$$\underset{\pi \in \Pi}{\operatorname{argmin}} \sum_{i=1}^{N} \psi_i^*(\rho_{\pi_i, \pi_{E_{-i}}} - \rho_{\pi_E}) = \underset{\pi \in \Pi}{\operatorname{argmin}} \sum_{i=1}^{N} \psi_i^*(\rho_{\pi_i, \pi_{-i}} - \rho_{\pi_E}) = \pi_E$$

Generalization

$$\min_{\theta}\max_{\omega}\mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\sum_{i=1}^{N}\log D_{\omega_i}(s,a_i)\right]+\mathbb{E}_{\boldsymbol{\pi}_E}\left[\sum_{i=1}^{N}\log(1-D_{\omega_i}(s,a_i))\right]$$

Algorithm 1 Multi-Agent GAIL (MAGAIL)

Input: Initial parameters of policies, discriminators and value (baseline) estimators θ_0, ω_0, ϕ_0; expert trajectories $\mathcal{D}=\{(s_j,a_j)\}_{j=0}^{M}$; batch size B; Markov game as a black box $(N,\mathcal{S},\mathcal{A},\eta,T,r,\mathbf{o},\gamma)$.
Output: Learned policies π_θ and reward functions D_ω.
for $u=0,1,2,\ldots$ **do**
  Obtain trajectories of size B from π by the process: $s_0\sim\eta(s), a_t\sim\pi_{\theta_u}(a_t|s_t), s_{t+1}\sim T(s_t|a_t)$.
  Sample state-action pairs from $\mathcal{D}$ with batch size B.
  Denote state-action pairs from π and $\mathcal{D}$ as χ and χ_E.
  for $i=1,\ldots,n$ **do**
    Update ω_i to increase the objective

$$\mathbb{E}_{\chi}[\log D_{\omega_i}(s,a_i)]+\mathbb{E}_{\chi_E}[\log(1-D_{\omega_i}(s,a_i))]$$

  end for
  for $i=1,\ldots,n$ **do**
    Compute value estimate $V^\star$ and advantage estimate A_i for $(s,a)\in\chi$.
    Update ϕ_i to decrease the objective

$$\mathbb{E}_{\chi}[(V_\phi(s,a_{-i})-V^\star(s,a_{-i}))^2]$$

    Update θ_i by policy gradient with small step sizes:

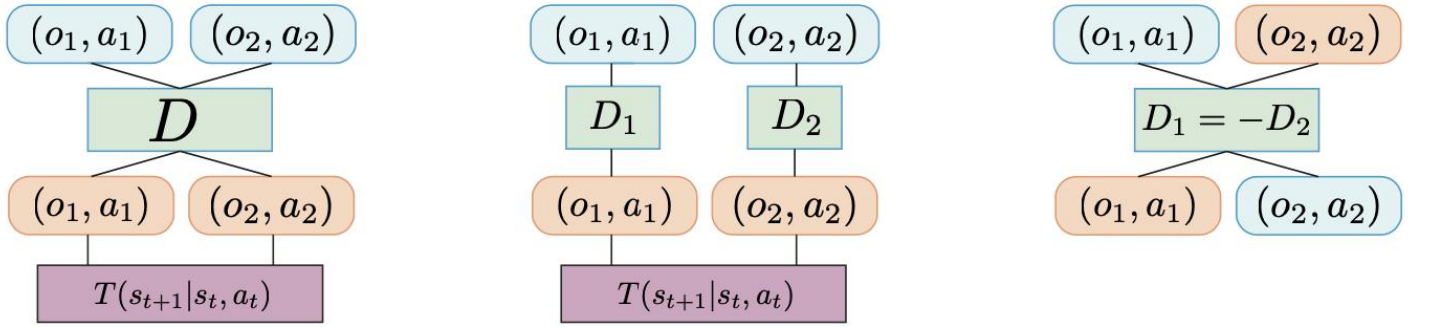
$$\mathbb{E}_{\chi}[\nabla_{\theta_i}\pi_{\theta_i}(a_i|o_i)A_i(s,a)]$$

  end for
end for

Practical multi-agent Imitation Learning

针对不同任务可以分为三种情况：

1. Centralized: 所有agent的reward function一致并且纯合作，用一个D就够，这样的流程和单agent的GAIL是一样的，只不过policy是联合策略。
2. Decentralized：agent的reward不一致，有合作有竞争，每个agent有一个D，但这些D不是独立地通过和环境交互来学习的。
3. Zero Sum：零和，r1=-r2，可以对D设立对抗目标，最大化一个agent的r同时最小化另一个agent的r。



(a) Centralized (Cooperative) (b) Decentralized (Mixed) (c) Zero-sum (Competitive)

Experiments

Cooperative Communication:

两个agent合作到达三个landmarks，一个agent(speaker)知道目标但无法移动，传递信息给另一个agent(listener)：能移动但无法看到目标.

Cooperative Navigation:

三个agents合作到达三个landmarks，理想状态是三个agents分别到达不同的landmark.

Keep-Away:

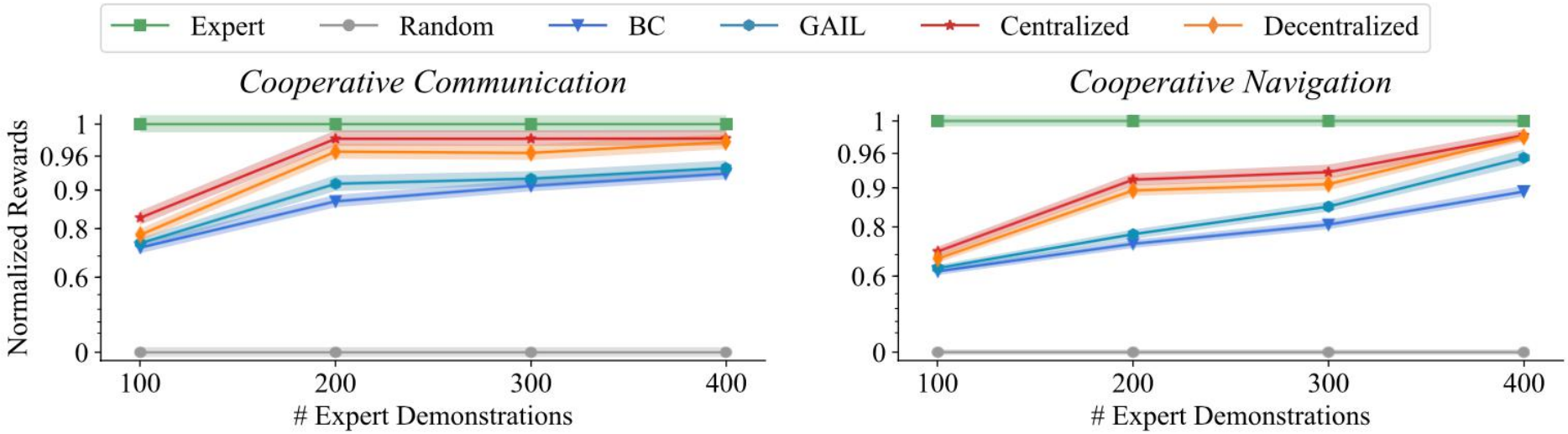
两个agents有相反的目标，agent1要到达两个landmarks中的一个，agent2要阻止agent1的到达，agent2看不到目标，只能基于agent1的动作行动.

Predator-Prey:

三个较慢的adversaries追击一个较快的agent.

Experiments

Cooperative Tasks



Experiments

Competitive Tasks

Task	Predator-Prey								
Agent	Behavior Cloning					G	C	D	ZS
Adversary	BC	G	C	D	ZS	Behavior Cloning			
Rewards	-93.20	-93.71	-93.75	-95.22	**-95.48**	-90.55	-91.36	**-85.00**	-89.4
Task	Keep-Away								
Agent	Behavior Cloning					G	C	D	ZS
Adversary	BC	G	C	D	ZS	Behavior Cloning			
Rewards	24.22	24.04	23.28	23.56	**23.19**	26.22	26.61	**28.73**	27.80

Thanks!