A Regularized Opponent Model with Maximum Entropy Objective

Zheng Tian[1,*], Ying Wen[1,*], Zhicheng Gong[1], Faiz Punakkath[1], Shihao Zou[2], and Jun Wang[1,*]

[1]Computer Science, University College London [2]Electrical and Computer Engineering, University of Alberta [*]{zheng.tian, ying.wen, jun.wang}@cs.ucl.ac.uk

1 Stochastic Games

An n-agent stochastic game is a tuple $(\mathcal{S}, \mathcal{A}^1, \ldots, \mathcal{A}^n, R^1, \ldots, R^n, p, \mathcal{T}, \gamma)$, where

- $\mathcal{S}$ denotes the state space.
- $\mathcal{A}^i$ is set of action space for for agent $i \in \{1, \ldots, n\}$.
- $R^i = R^i(s, a^i, a^{-i})$ is set of reward function for agent $i \in \{1, \ldots, n\}$.
- $\mathcal{T} : \mathcal{S} \times \mathcal{A}$ is the transition function.

and p is the distribution of the initial state, γ is a discount factor. Agent i chooses its action $a^i \in \mathcal{A}^i$ according to the policy $\pi^i_{\theta^i}(a^i|s)$ parameterized by θ^i. From the perspective of agent i, it can interpret joint policy $\pi_\theta = (\pi^i_{\theta^i}(a^i|s), \pi^{-i}_{\theta^{-i}}(a^{-i}|s))$, where $a^{-i} = (a^j)_{j \neq i}$, $\theta^{-i} = (\theta^j)_{j \neq i}$, and $\pi^{-i}_{\theta^{-i}}(a^{-i}|s)$ is a compact representation of the joint policy of all complementary agents of i. In fully cooperative games, different agents have the same reward function $R^i(s, a^i, a^{-i}) = R^{-i}(s, a^i, a^{-i}), \forall i \in 1, \ldots, n$. Therefore, each agent's objective is to maximize the shared expected return:

$$\max \quad \eta^i(\pi_\theta) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R(s_t, a^i_t, a^{-i}_t)\right]. \tag{1}$$

2 A Variational Lower Bound for MARL

We transform the control problem into an inference problem by introducing a binary random variable o^i_t which serves as the indicator for "optimality" for each agent i at each time step t. Assume given other players actions a^{-i}_t, the posterior probability of agent i's optimality is proportional to its exponential reward:

$$P(o^i_t = 1|s_t, a^i_t, a^{-i}_t) \propto \exp(R(s_t, a^i_t, a^{-i}_t)). \tag{2}$$

Given the fact that other agents are playing their optimal policies $o^{-i} = 1$, the probability that agent i also plays its optimal policy $P(o^i = 1|o^{-i} = 1)$ is the probability of obtaining the maximum reward from agent i's perspective. Therefore, we define agent i's objective as:

$$\max \quad \mathcal{J} \triangleq \log P(o^i_{1:T} = 1|o^{-i}_{1:T} = 1) \tag{3}$$

To optimize the observed evidence defined in Eq. 3, therefore, we use variational inference (VI) with an auxiliary distribution over these latent variables $q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T} = 1, o^{-i}_{1:T} = 1)$. Without loss of generality, we here derive the solution for agent i. We factorize $q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T} = 1, o^{-i}_{1:T} = 1)$ so as to capture agent i's conditional policy on the current state and opponents actions, and beliefs regarding opponents actions. This way, agent i will learn optimal policy, while also possessing the capability to model opponents actions a^{-i}. Using all modelling assumptions, we may factorize $q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T} = 1, o^{-i}_{1:T} = 1)$ as:

$$\begin{aligned} q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T} = 1, o^{-i}_{1:T} = 1) &= P(s_1)\prod_t P(s_{t+1}|s_t, a_t) q(a^i_t|a^{-i}_t, s_t, o^i_t = o^{-i}_t = 1) \times q(a^{-i}_t|s_t, o^i_t = o^{-i}_t = 1) \\ &= P(s_1)\prod_t P(s_{t+1}|s_t, a_t)\pi(a^i_t|s_t, a^{-i}_t)\rho(a^{-i}_t|s_t), \end{aligned}$$

where we assumed same initial and states transitions as in the original model. With this factorization, lower bound is derived on the likelihood of optimality of agent i:

$$\begin{aligned} &\log P(o^i_{1:T} = 1|o^{-i}_{1:T} = 1) \\ &\geq \mathcal{J}(\pi, \rho) \triangleq \sum_t \mathbb{E}_{(s_t, a^i_t, a^{-i}_t)\sim q}[R(s_t, a^i_t, a^{-i}_t) + H(\pi(a^i_t|s_t, a^{-i}_t)) - D_{\mathrm{KL}}(\rho(a^{-i}_t|s_t)||P(a^{-i}_t|s_t))] \\ &= \sum_t \mathbb{E}_{s_t}[\underbrace{\mathbb{E}_{a^i_t\sim\pi, a^{-i}_t\sim\rho}[R(s_t, a^i_t, a^{-i}_t) + H(\pi(a^i_t|s_t, a^{-i}_t))]}_{\text{MEO}} - \underbrace{\mathbb{E}_{a^{-i}_t\sim\rho}[D_{\mathrm{KL}}(\rho(a^{-i}_t|s_t)||P(a^{-i}_t|s_t))]}_{\text{Regularizer of } \rho}]. \end{aligned} \tag{4}$$

3 Multi-Agent Soft Actor Critic

By defining multi-agent soft Q-function and V-function at first, we show that the conditional policy and opponent model defined in Eq. 7 and 8 below are optimal solutions with respect to the objective defined in Eq. 4:

Theorem 1. We define the soft state-action value function of agent i as

$$Q^{\pi^*, \rho^*}_{soft}(s_t, a^i_t, a^{-i}_t) = r_t + \mathbb{E}_{(s_{t+l}, a^i_{t+l}, a^{-i}_{t+l}, \ldots)\sim q}[\sum_{l=1}^{\infty} \gamma^l (r_{t+l} + \alpha H(\pi^*(a^i_{t+l}|a^{-i}_{t+l}, s_{t+l})) - D_{\mathrm{KL}}(\rho^*(a^{-i}_{t+l}|s_{t+l})||P(a^{-i}_{t+l}|s_{t+l}))], \tag{5}$$

and soft state value function as

$$V^*(s) = \log \sum_{a^{-i}} P(a^{-i}|s)\left(\sum_{a^i} \exp(\frac{1}{\alpha} Q^*_{soft}(s, a^i, a^{-i}))\right)^\alpha, \tag{6}$$

Then the optimal conditional policy and opponent model for Eq. 4 are

$$\pi^*(a^i|s, a^{-i}) = \frac{\exp(\frac{1}{\alpha} Q^{\pi^*, \rho^*}_{soft}(s, a^i, a^{-i}))}{\sum_{a^i} \exp(\frac{1}{\alpha} Q^{\pi^*, \rho^*}_{soft}(s, a^i, a^{-i}))}, \tag{7}$$

and

$$\rho^*(a^{-i}|s) = \frac{P(a^{-i}|s)\left(\sum_{a^i} \exp(\frac{1}{\alpha} Q^*_{soft}(s, a^i, a^{-i}))\right)^\alpha}{\exp(V^*(s))}. \tag{8}$$

Following from Theorem 1, multi-agent soft Bellman equation is defined:

Theorem 2. We define the soft multi-agent Bellman equation for the soft state-action value function $Q^{\pi, \rho}_{soft}(s, a^i, a^{-i})$ of agent i as

$$Q^{\pi^*, \rho^*}_{soft}(s, a^i, a^{-i}) = r_t + \gamma \mathbb{E}_{(s_{t+1})}[V^*_{soft}(s_{t+1})]. \tag{9}$$

With this Bellman equation defined above, the solution to Eq. 9 is derived with a fixed point iteration, which we call ROMMEO Q-iteration (ROMMEO-Q):

Theorem 3. ROMMEO Q-iteration. In a symmetric game with only one global optimum, i.e. $\mathbb{E}_{\pi^*}\left[Q^i_t(s)\right] \geq \mathbb{E}_{\pi}\left[Q^i_t(s)\right]$, where π^* is the optimal strategy profile. Let $Q_{soft}(\cdot, \cdot, \cdot)$ and $V_{soft}(\cdot)$ be bounded and assume

$$\sum_{a^{-i}} P(a^{-i}|s)\left(\sum_{a^i} \exp(\frac{1}{\alpha} Q^*_{soft}(s, a^i, a^{-i}))\right)^\alpha < \infty$$

and that $Q^*_{soft} < \infty$ exists. Then the fixed-point iteration

$$Q_{soft}(s_t, a^i_t, a^{-i}_t) \leftarrow r_t + \gamma \mathbb{E}_{(s_{t+1})}[V_{soft}(s_{t+1})], \tag{10}$$

where $V_{soft}(s_t) \leftarrow \log \sum_{a^{-i}} P(a^{-i}_t|s_t) \times \left(\sum_{a^i_t} \exp(\frac{1}{\alpha} Q_{soft}(s_t, a^i_t, a^{-i}_t))\right)^\alpha \forall s_t, a^i_t, a^{-i}_t$, converges to Q^*_{soft} and V^*_{soft} respectively.

Finally, to recover the optimal conditional policy and opponent model and avoid intractable inference steps defined in Eq. 7 and 8 in complex problems, we minimize the KL-divergence between functions of Q values and parameterized opponent model and conditional policy. By using the reparameterization trick: $\hat{a}^{-i}_t = g_\phi(\epsilon^{-i}_t; s_t)$ and $a^i_t = f_\theta(\epsilon^i_t; s_t, \hat{a}^{-i}_t)$, we can rewrite the objectives above as

$$\mathcal{J}_\pi(\theta) = \mathbb{E}_{s_t\sim D, \epsilon^i_t\sim N, \hat{a}^{-i}_t\sim\rho}[\alpha \log \pi_\theta(f_\theta(\epsilon^i_t; s_t, \hat{a}^{-i}_t)) - Q_\omega(s_t, f_\theta(\epsilon^i_t; s_t, \hat{a}^{-i}_t), \hat{a}^{-i}_t)], \tag{11}$$

$$\begin{aligned} \mathcal{J}_\rho(\phi) = \mathbb{E}_{(s_t, a_t)\sim D, \epsilon^{-i}_t\sim N}[&\log \rho_\phi(g_\phi(\epsilon^{-i}_t; s_t)|s_t) \\ &- logP(\hat{a}^{-i}_t|s_t) - Q(s_t, a^i_t, g_\phi(\epsilon^{-i}_t; s_t)) + \alpha \log \pi_\theta(a^i_t|s_t, g_\phi(\epsilon^{-i}_t; s_t))]. \end{aligned} \tag{12}$$

4 Experiments

4.1 Iterated Matrix Games

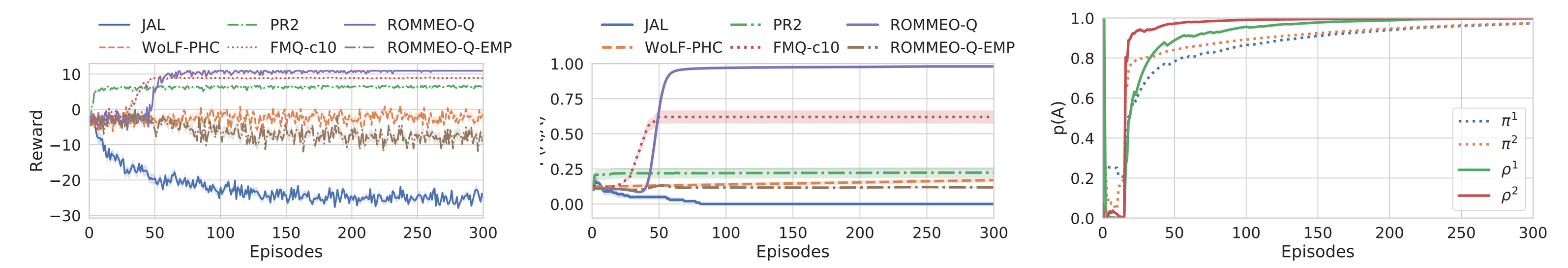

Figure 1: (Left): Learning curves of ROMMEO and baselines on ICG over 100 episodes. (Middle): Probability of convergence to the global optimum for ROMMEO and baselines on ICG over 100 episodes. The vertical axis is the joint probability of taking actions A for both agents. (Right): Probability of taking A estimated by agent i's opponent model ρ^i and observed empirical frequency P^i in one trail of training, $i \in \{1, 2\}$

Climbing game (CG) is a fully cooperative game whose payoff matrix is summarized as follows: $R = \begin{array}{c|ccc} & A & B & C \\ A & (11, 11) & (-30, -30) & (0, 0) \\ B & (-30, -30) & (7, 7) & (6, 6) \\ C & (0, 0) & (0, 0) & (5, 3) \end{array}$. It is a challenging benchmark because of the difficulty of convergence to its global optimum. There are two Nash equilibrium (A, A) and (B, B) but one global optimal (A, A). The punishment of miscoordination by choosing a certain action increases in the order of $C \to B \to A$. The safest action is C and the miscoordination punishment is the most severe for A. Therefore it is very difficult for agents to converge to the global optimum in ICG.

4.2 Differential Games

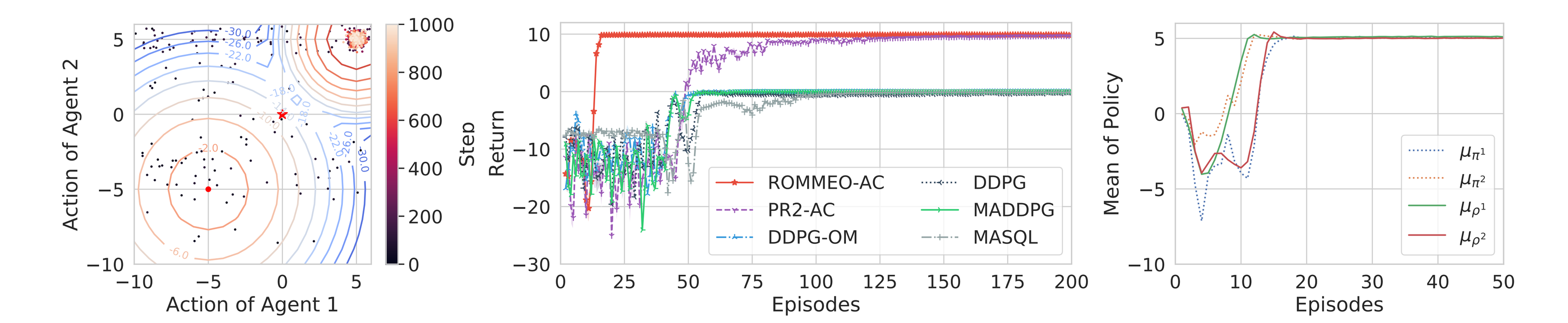

Figure 2: Experiment on Max of Two Quadratic Game. (Left): Reward surface and learning path of agents. Scattered points are actions taken at each step; (Middle): Learning curve of ROMMEO and baselines. (Right): Mean of agents' policies π and opponent models ρ

The Max of Two Quadratic is Differential Game for continuous case. The agents have continuous action space of $[-10, 10]$. Each agent's reward depends on the joint action following the equations: $r^1(a^1, a^2) = r^2(a^1, a^2) = \max(f_1, f_2)$, where $f_1 = 0.8 \times [-(\frac{a^1+5}{3})^2 - (\frac{a^2+5}{3})^2]$, $f_2 = 1.0 \times [-(\frac{a^1-5}{1})^2 - (\frac{a^2-5}{1})^2] + 10$. The reward surface is provided in Fig. 2 ; there is a local maximum 0 at $(-5, -5)$ and a global maximum 10 at $(5, 5)$, with a deep valley staying in the middle. If the agents' policies are initialized to $(0, 0)$ (the red starred point) that lies within the basin of the left local maximum, the gradient-based methods would tend to fail to find the global maximum equilibrium point due to the valley blocking the upper right area.