A Regularized Opponent Model with Maximum Entropy Objective

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

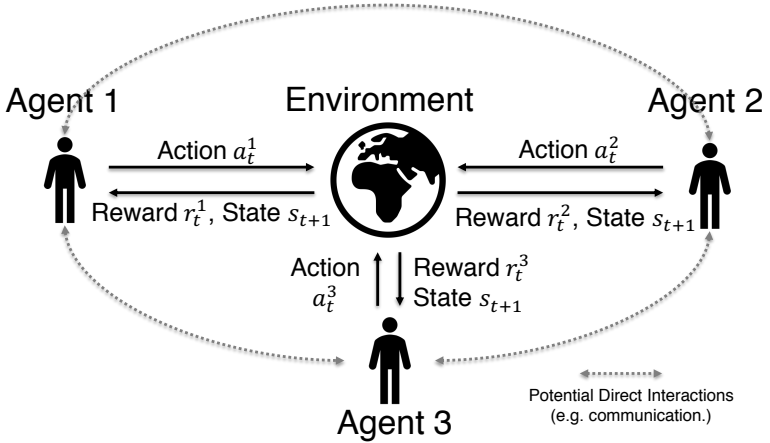
[1] University College London
[2] University of Alberta

August 19, 2019

[*] The first two authors contributed equally.

Overview

Multi Agent Reinforcement Learning



Where MDP Fails in MARL?

Your opponent is learning and updating its policy over time: $P_t(a^{-i}|s) \neq P_{t+\delta t}(a^{-i}|s)$.

Therefore, for a single naive agent i, the environment it perceives is non-stationary:

$$\begin{aligned} P_t(s'|s,a^i) &= \sum_{a^{-i}} P(s'|s,a^i,a^{-i}) P_t(a^{-i}|s) \\ &\neq \sum_{a^{-i}} P(s'|s,a^i,a^{-i}) P_{t+\delta t}(a^{-i}|s) \\ &= P_{t+\delta t}(s'|s,a^i) \end{aligned} \tag{1}$$

An Oracle Augmented MDP

An oracle to foretell agent i its opponents' actions a^{-i} at each time step.

We can have an augmented environment state $\bar{s}$ by combining the orginal state s with the foretold a^{-i} and we define the augmented state space as:

$$\bar{S} = S \times A^{-i}.$$

Then the oracle augmented MDP is stationary again:

$$P_t(\bar{s}'|\bar{s}, a^i) = P_{t+\delta t}(\bar{s}'|\bar{s}, a^i).$$

Opponent Modelling (OM)

- The powerful oracle rarely exists in any situations;
- But we can try to learn a model which can predict opponents' actions.

Opponent Modelling in Cooperative Game

- Modelling opponents' changing behaviour without any prior knowledge is generally hard.
- We define cooperative game as an environment where all agents receive their maximum long term return by cooperation.
- Then we can safely assume reasonable opponents update their policies towards the optimum.
- In this work, we propose a framework where we integrate this assumption into an agent's learning algorithm by variational inference.

Optimum and optimal policy in Cooperative Game

Definition

In cooperative multi-agent reinforcement learning, optimum is a strategy profile $(\pi^{1*}, \ldots, \pi^{n*})$ such that:

$$\begin{aligned} &\mathbb{E}_{s\sim p_s, a_t^{i*}\sim\pi^{i*}, a_t^{-i*}\sim\pi^{-i*}}\left[\sum_{t=0}^{\infty}\gamma^t R^i(s_t, a_t^{i*}, a_t^{-i*})\right] \\ &\geq \mathbb{E}_{s\sim p_s, a_t^{i}\sim\pi^{i}, a_t^{-i}\sim\pi^{-i}}\left[\sum_{t=0}^{\infty}\gamma^t R^i(s_t, a_t^{i}, a_t^{-i})\right] \forall\pi\in\Pi, i\in(1\ldots n) \end{aligned} \tag{2}$$

where $\pi = (\pi^i, \pi^{-i})$ and Agent i's optimal policy is π^{i*}.

A Variational Lower Bound

- We define a random variable $o_t^i = 1$ indicating agent i's policy at time step t is optimal.
- We define agent i's objective as:

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

- In cooperative game, we can assume:

$$P(o_t^i = 1 | o_t^{-i} = 1, s_t, a_t^i, a_t^{-i}) \propto \exp(R^i(s_t, a_t^i, a_t^{-i})).$$

A Variational Lower Bound

- We factorise the auxiliary distribution as:

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

- Finally, we have:

$$
\log P(o^i_{1:T} = 1 | o^{-i}_{1:T} = 1) \geq \sum_t \mathbb{E}_{s_t} [\underbrace{\mathbb{E}_{a^i_t \sim \pi, a^{-i}_t \sim \rho} [R^i(s_t, a^i_t, a^{-i}_t) + H(\pi(a^i_t | s_t, a^{-i}_t))]}_{\text{MEO}}
$$

$$
\underbrace{- \mathbb{E}_{a^{-i}_t \sim \rho} [D_{\mathrm{KL}}(\rho(a^{-i}_t | s_t) || P(a^{-i}_t | s_t, o^{-i}_t))]}_{\text{Regularizer of } \rho}]. \tag{4}
$$

An Analysis of the Learning of the Opponent Model ρ

- Recall $\rho(a_t^{-i}|s_t) = q(a_t^{-i}|s_t, o_t^i = o_t^{-i} = 1)$ models opponent policy at optimum.
- We use observed opponent historical behaviour as the prior $P(a_t^{-i}|s_t, o_t^{-i})$.
- The optimisation of $\underbrace{\mathbb{E}_{a_t^i \sim \pi, a_t^{-i} \sim \rho}[R^i(s_t, a_t^i, a_t^{-i}) + H(\pi(a_t^i|s_t, a_t^{-i}))]}_{\text{MEO}}$ with respect to ρ.
  - Updates opponent model towards the optimum from agent i's perspective.
- The optimisation of $\underbrace{\mathbb{E}_{a_t^{-i} \sim \rho}[D_{\mathrm{KL}}(\rho(a_t^{-i}|s_t)||P(a_t^{-i}|s_t, o_t^{-i}))]}_{\text{Regularizer of } \rho}$ with respect to ρ.
  - To avoid building an unrealistically optimistic opponent model, regularise it by punishment of deviation from observed opponent behavior.

Soft Value Functions

Definition

We define the soft state-action value function of agent i and soft state value function as:

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

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

We also define the soft multi-agent Bellman equation as

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

Soft Policy Iteration[†]

Theorem

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

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

and

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

[†]The experiment code and appendix are available at https://github.com/rommeoijcai2019/rommeo.

Iterated Climbing Games

	A	B	C
A	(11, 11)	(-30, -30)	(0 , 0)
B	(-30, -30)	(7, 7)	(6, 6)
C	(0, 0)	(0, 0)	*(5, 3)*

Table: Payoff matrix of Climbing Game

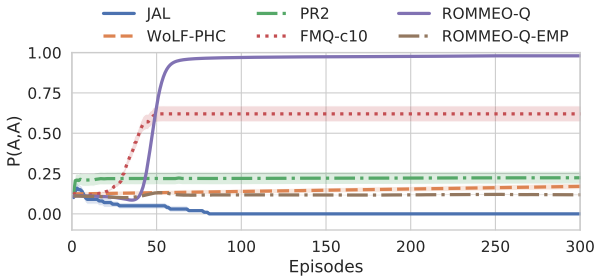


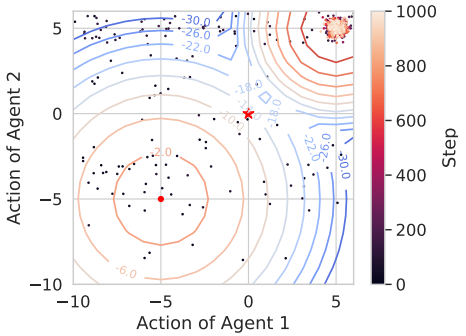
Figure: Probability of convergence to the global optimum for ROMMEO and baselines on ICG over 100 episodes.

Results: Differential Games

The agents have continuous action space of $[-10, 10]$.

$r^1\left(a^1, a^2\right) = r^2\left(a^1, a^2\right) = \max\left(f_1, f_2\right)$, where

$f_1 = 0.8 \times [-(\frac{a^1+5}{3})^2 - (\frac{a^2+5}{3})^2], f_2 = 1.0 \times [-(\frac{a^1-5}{1})^2 - (\frac{a^2-5}{1})^2] + 10.$



Results: Differential Games

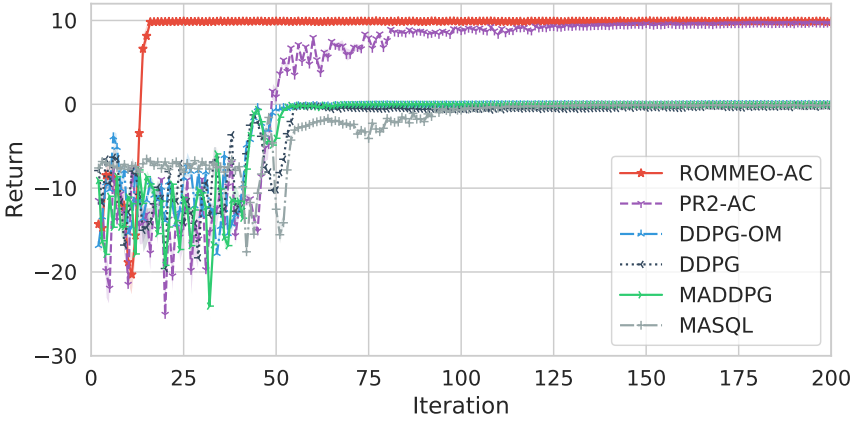


Figure: Learning curve of ROMMEO and baselines.

Conclusion

- We use variantional inference to formulate MARL problem and derive a novel objective ROMMEO which gives rise to a new perspective on opponent modelling.
- We design an off-policy algorithm ROMMEO-Q with complete convergence proof for optimising ROMMEO. For better generality, we also propose ROMMEO-AC, an actor critic algorithm powered by NNs to solve complex and continuous problems.
- Theorems in our paper only guarantee the convergence to optimal solutions with respect to ROMMEO objective but not the optimum in the game. The achievement of the optimum in the game also relies on the opponent learning algorithm.