Probabilistic recursive reasoning for multi-agent reinforcement learning

Wen, Ying; Yang, Yaodong; Luo, Rui; Wang, Jun; Pan, Wei

Publication date
2019

Document Version
Final published version

Citation (APA)
Wen, Y., Yang, Y., Luo, R., Wang, J., & Pan, W. (2019). *Probabilistic recursive reasoning for multi-agent reinforcement learning*. Poster session presented at 7th International Conference on Learning Representations, ICLR 2019, New Orleans, United States.

Probabilistic Recursive Reasoning for Multi-agent Reinforcement Learning

Ying Wen*, Yaodong Yang*, Rui Luo, Jun Wang and Wei Pan

University College London, TU Delft

Motivations

Similar to the way of thinking adopted by humans, **Recursive Reasoning** represents the belief reasoning process where each agent considers the reasoning process of other agents, based on which it expects to make better decisions. Importantly, it allows an opponent to reason about the modeling agent rather than being a fixed type; the process can therefore be nested in a form as:

"I believe that you believe that I believe ...".

there has been little work that tries to adopt this idea into the multi-agent deep reinforcement learning (DRL) setting.

Multi-agent Learning Objective

Each agent is presumed to pursue the maximal cumulative reward expressed as:

$$\max \ \eta^i(\pi_\theta) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r^i(s_t, a_t^i, a_t^{-i})\right], \qquad (1)$$

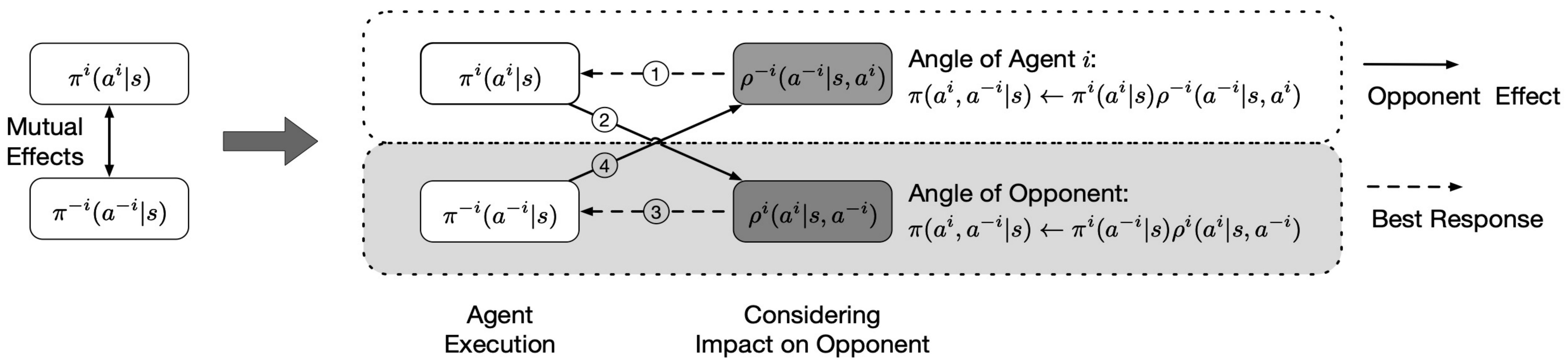
Non-correlated Joint Policy Factorization

One common approach is to decouple the joint policy assuming conditional independence of actions from different agents:

$$\pi_\theta(a^i, a^{-i}|s) = \pi^i_{\theta^i}(a^i|s)\pi^{-i}_{\theta^{-i}}(a^{-i}|s). \qquad (2)$$

But impacts of one agent's action on other agents, and the subsequent reactions from other agents are not molded. It gives **non-correlated multi-agent** learning objective:

$$\nabla_{\theta^i}\eta^i = \mathbb{E}_{s\sim p, a^i\sim\pi^i}[\nabla_{\theta^i}\log\pi^i(a^i|s)\int_{a^{-i}}\pi^{-i}(a^{-i}|s)Q^i(s, a^i, a^{-i})\,\mathrm{d}a^{-i}].$$



Probabilistic Recursive Reasoning Framework

PR2 decouples the connections between agents. **Step 1**: agent i takes the best response after considering all the potential consequences of opponents' actions given its own action a^i. **Step 2**: how agent i behaves in the environment serves as the prior for the opponents to learn how their actions would affect a^i. **Step 3**: similar to Step 1, opponents take the best response to agent i. **Step 4**: similar to Step 2, opponents' actions are the prior knowledge to agent i on estimating how a^i will affect the opponents. Looping from Step 1 to 4 forms recursive reasoning.

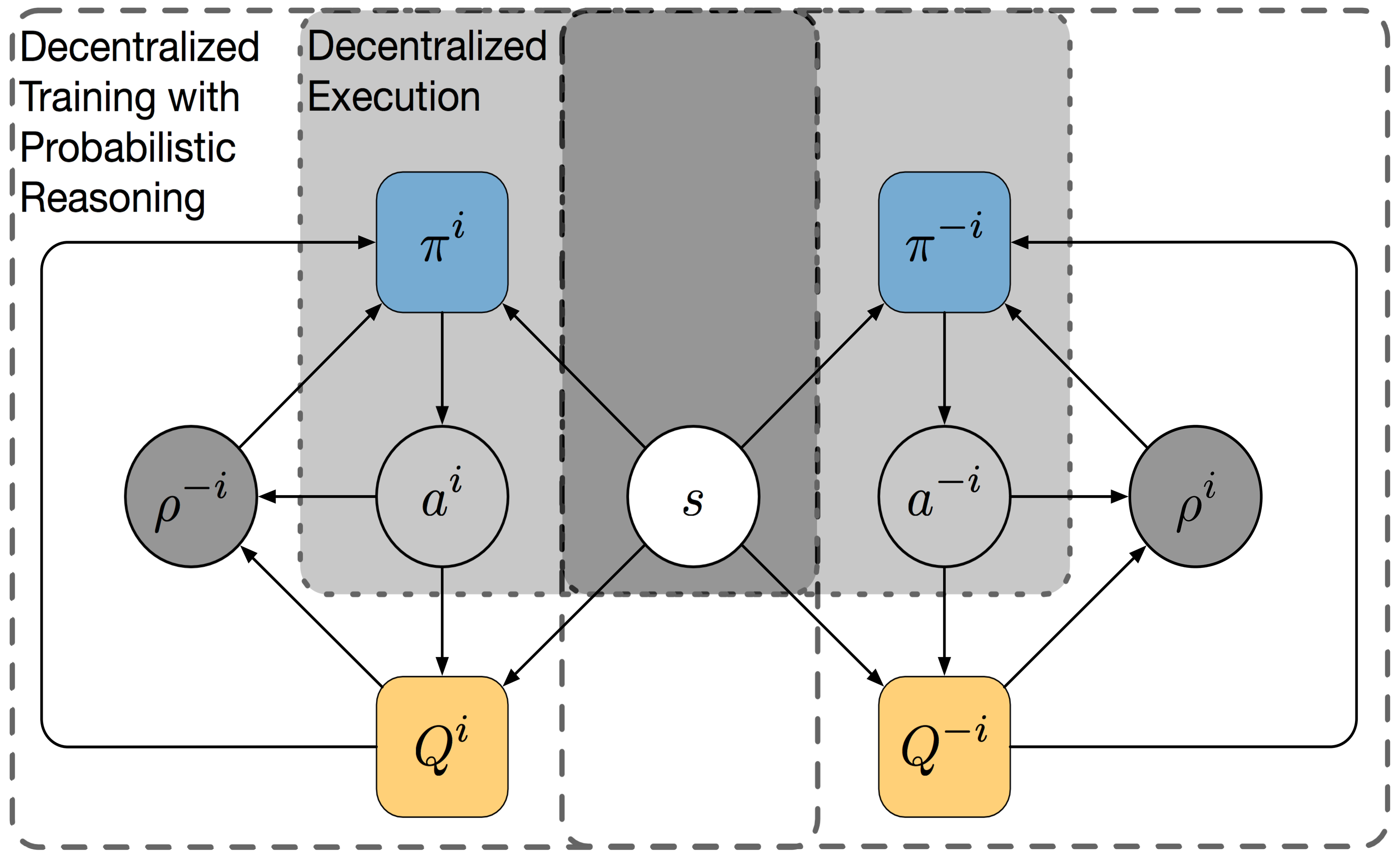


Diagram of multi-agent PR2 learning algorithms. It conducts decentralized training with decentralized execution. The light grey panels on two sides indicate decentralized execution for each agent whereas the white counterpart shows the decentralized learning procedure. All agents share the interaction experiences in the environment inside the dark rectangle in the middle.

Probabilistic Recursive Reasoning Policy Gradient

By considering the level-1 recursion, we re-formulate the joint policy:

$$\pi_\theta(a^i, a^{-i}|s) = \underbrace{\pi^i_{\theta^i}(a^i|s)\pi^{-i}_{\theta^{-i}}(a^{-i}|s, a^i)}_{\text{Agent } i\text{'s perspective}} = \underbrace{\pi^{-i}_{\theta^{-i}}(a^{-i}|s)\pi^i_{\theta^i}(a^i|s, a^{-i})}_{\text{The opponents' perspective}}. \qquad (3)$$

Given the opponent policy $\pi^{-i}_{\theta^{-i}}$, and that each agent tries to maximize its objective defined in Eq. 1, we establish the policy gradient theorem by accounting for the PR2 joint policy decomposition in Eq. 3:

Proposition 1. *In a stochastic game, under the recursive reasoning framework defined by Eq. 3, the update for the multi-agent recursive reasoning policy gradient method can be derived as follows:*

$$\nabla_{\theta^i}\eta^i = \mathbb{E}_{s\sim p, a^i\sim\pi^i}\left[\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s)\int_{a^{-i}}\pi^{-i}_{\theta^{-i}}(a^{-i}|s, a^i)Q^i(s, a^i, a^{-i})\,\mathrm{d}a^{-i}\right].$$

Variational Inference on Opponent Conditional Policy

Optimization-based approximation to infer the unobservable $\rho^{-i}_{\phi^{-i}}(a^{-i}|\, s, a^i)$ via variational inference with soft RL formulation:

Theorem 1. *The optimal Q-function for agent i that satisfies minimizing KL-divergence in soft RL is formulated as:*

$$Q^i_{\pi_\theta}(s, a^i) = \log\int_{a^{-i}}\exp(Q^i_{\pi_\theta}(s, a^i, a^{-i}))\,\mathrm{d}a^{-i}.$$

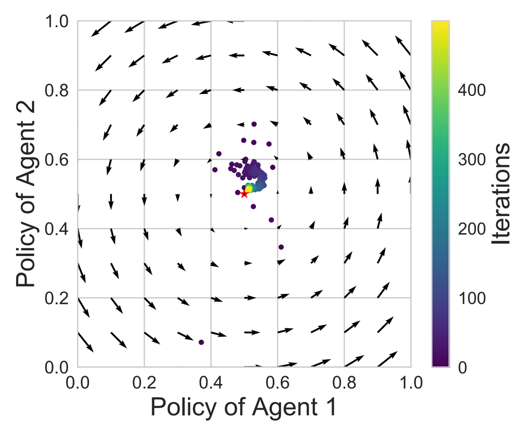
And the corresponding optimal opponent conditional policy reads:

$$\rho^{-i}_{\phi^{-i}}(a^{-i}|s, a^i) = \frac{1}{Z}\exp(Q^i_{\pi_\theta}(s, a^i, a^{-i}) - Q^i_{\pi_\theta}(s, a^i))$$

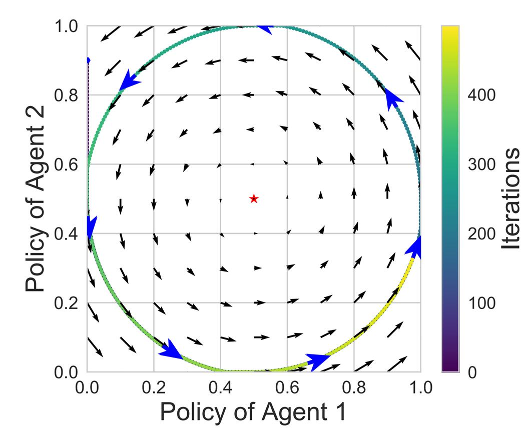
Experiments

- **Iterated Matrix Game**

IGA fails to converge to the equilibrium but rotate around the equilibrium point. On the contrary, PR2-Q can find precisely the central equilibrium with a fully distributed fashion.



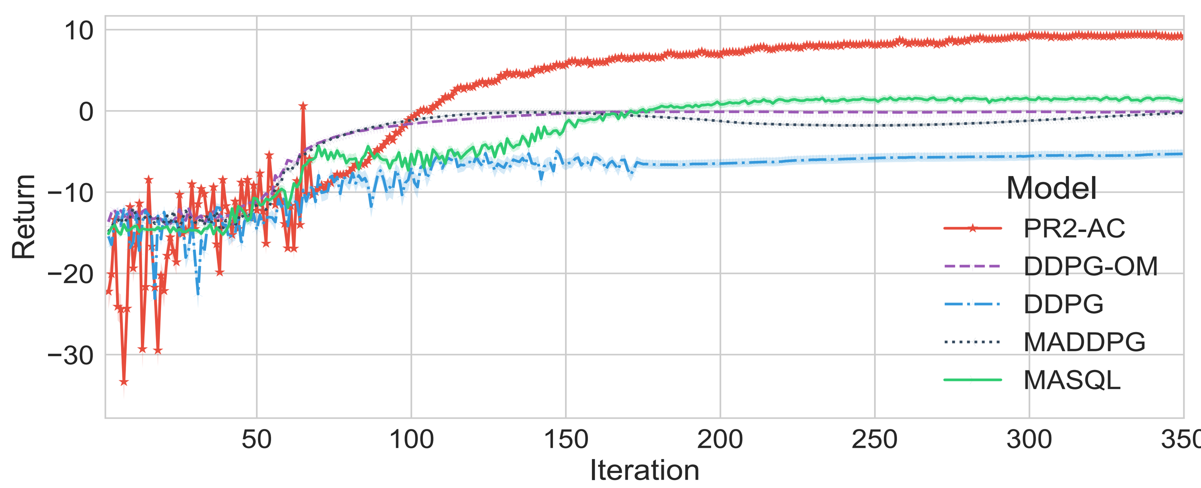
PR2-Q learning dynamics on matrix game



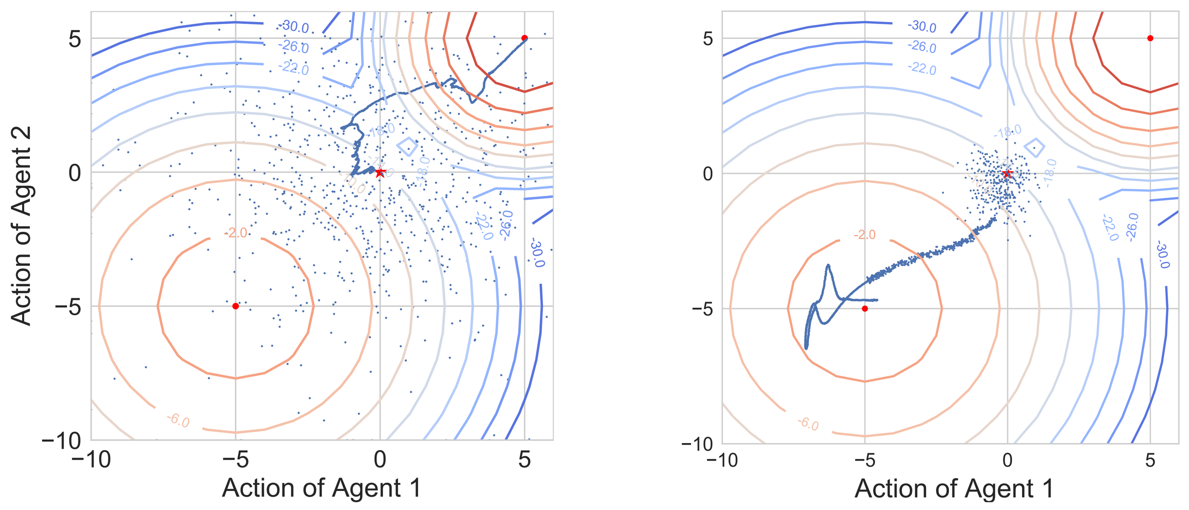
IGA learning dynamics on matrix game

- **Differential Game**

PR2-AC model finds the peak point in joint action space, the agents can quickly go through the shortcut out of the local basin in a clever way, while other algorithms just converge to the local equilibrium.



Learning Curve on Quadratic Game



Learning Dynamics on Quadratic Game, Left: PR2-AC, Right: MADDPG.

TU Delft MediaGamma