**SELF-INERENCE OF OTHERS’ POLICIES FOR HOMOGENEOUS AGENTS IN COOPERATIVE MULTI-AGENT REINFORCEMENT LEARNING**

Qifeng Lin  
School of Computer Science and Engineering, Sun Yat-Sen University

Qing Ling  
School of Computer Science and Engineering, Sun Yat-Sen University

### Background
- Multi-agent reinforcement learning (MARL) remains to be a challenging research field and has various applications in multi-robot control, multi-player games, etc.
- In cooperative MARL, agents are trained to cooperatively achieve a global goal.
- Partial observability: only local observation available rather than global states
  - is one of the critical challenges in MARL
  - motivates a training paradigm named centralized training and decentralized execution (CTDE)

- Policy inference: infer policies of other agents
  - plays an important role in MARL
  - is helpful to improve coordination efficiency

### Related Works
- Fully observable scenarios
  - AMS-A3C and AMF-A3C share learned parameters and add extra policy features, respectively
  - Attention Multi-agent DDPG (ATT-MADDPG) introduces attention mechanism
  - Hard to access global states in real world
- Partially observable scenarios
  - Extra hidden representations of other agents’ policies are required
    - Deep reinforcement opponent network (DRON)
    - Deep policy inference Q-network (DPIQ) and deep policy inference recurrent Q-network (DPIRQN)
    - Multi-agent DDPG with policy inference (MADDPG-PI)
  - Massive resource consumption

⇒ A self-inference approach to infer other agents’ policies under
- cooperative MARL
- partially observable
- CTDE
- homogeneous agents

### Method
Partially observable Markov decision process (POMDP) with N agents

\[ <S, O, A, P, R> \]

- \( S \): the sets of state space
- \( O \): the joint observation spaces \( \{O_1, \ldots, O_N\} \)
- \( A \): the joint action spaces \( \{A_1, \ldots, A_N\} \)
- \( R \): the joint rewards \( \{R_1, \ldots, R_N\} \)

The long-term return of agent \( i \) is \( R_i = \sum_{t=0}^{T} (\gamma)^t r_i^t \)

\( \gamma \) is a discount factor
\( T \) is the time horizon
\( r_i^t \in R_i \) is the instantaneous reward at time \( t \)

Goal: find optimal policies \( \mu_i; \mathcal{O}_i; A_i \rightarrow [0, 1] \) to maximize \( R = \sum_{t=1}^{N} R_i \).

Further, we parameterize the policy \( \mu_i \) with \( \theta_i \).

For critic part in MADDPG-SI, the update rule of each agent \( i \) is by minimizing the critic loss:

\[
L(\phi_i) = \mathbb{E}_{x,a,s \sim D}[Q_{\phi_i}^\mu_i(x, a, s) - y_i]^2,
\]

\[ y_i = r_i + \gamma Q_{\phi_i}(s', a', y_i = \mu_{y_i}(a_i)) \]

(1)

\( D \) is the experience replay buffer.

Only use agent \( i \)'s model to infer policies of other agents

For actor part in MADDPG-SI, the update rule of each agent \( i \) is by minimizing the actor loss

\[
L(\theta_i) - \beta P_{Si}(\theta_i) \]

(2)

where

\[
L(\theta_i) = -Q_{\phi_i}^\mu_i(x, a_1, \ldots, a_N) |_{a_i = \mu_{\theta_i}(a_i)}
\]

\[
P_{Si}(\theta_i) = \frac{1}{N} \sum_{j \neq i} E_{\phi_j, \theta_j} [\mathbb{E}_{a_j} Q_{\phi_j}(s_j, a_j, \theta_j)]
\]

(3)

and \( \beta \) is a positive scale factor that balances learning from its own experience and learning from other agents’ experience.

Compared with MADDPG-PI [1], MADDPG-SI requires less deep neural networks as given by

\[
f_{\phi_i} = 2N(N + 1),
\]

\[
f_{\theta_i} = 4N.
\]

Therefore, the space complexities for MADDPG-PI and MADDPG-SI are \( O(N^2) \) and \( O(N) \), respectively.

### Experimental Results
- Environment: cooperative navigation with \( N \) agents and \( L \) landmarks [1]
  - Task: agents occupy all the landmarks cooperatively
  - A shared reward:
    - sums up negative distances between agents and landmarks
    - every collision between the agents contributes -1
- Four settings: \( (N=3, L=3), (N=4, L=4), (N=5, L=5), \) and \( (N=6, L=6) \).

As shown in Fig. 1, MADDPG-SI can achieve almost equivalent performance and even outperform MADDPG and MADDPG-PI in some cases. Fig. 2 shows that the agent of MADPG-SI can be closer to landmarks compared with the one of MADDPG.

### References