Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, Igor Mordatch

Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)

1. Introduction

- Reinforcement Learning is often used in various single agent domain

- Traditional reinforcement (Q learning, Policy gradient) do not suit for multi-agent

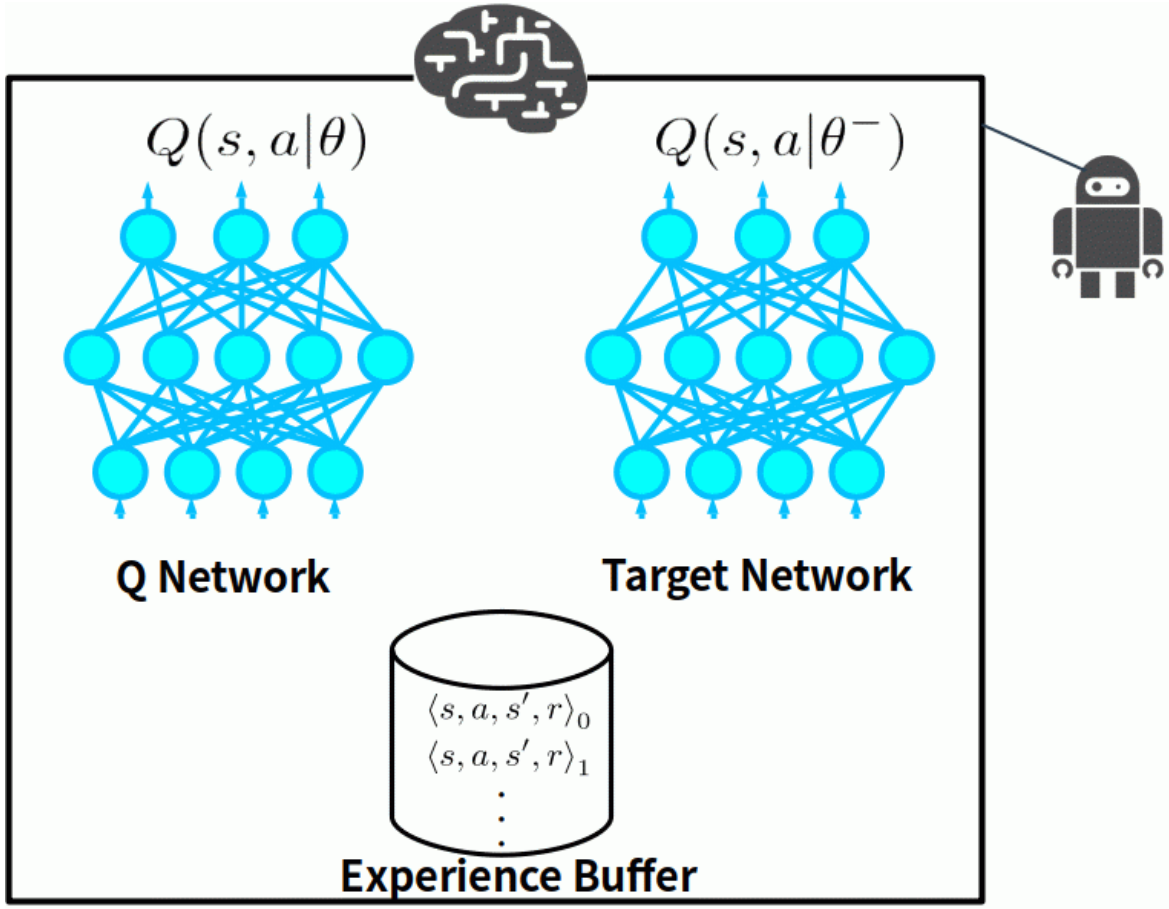
goal of algorithm

1. do not change model in different environment
2. no particular structure between agents
3. behave cooperative or competitive

2. method (related work)

DDPG(deep deterministic policy gradient)

- Replay buffer
  (use past experience for learning)
- Actor-Critic

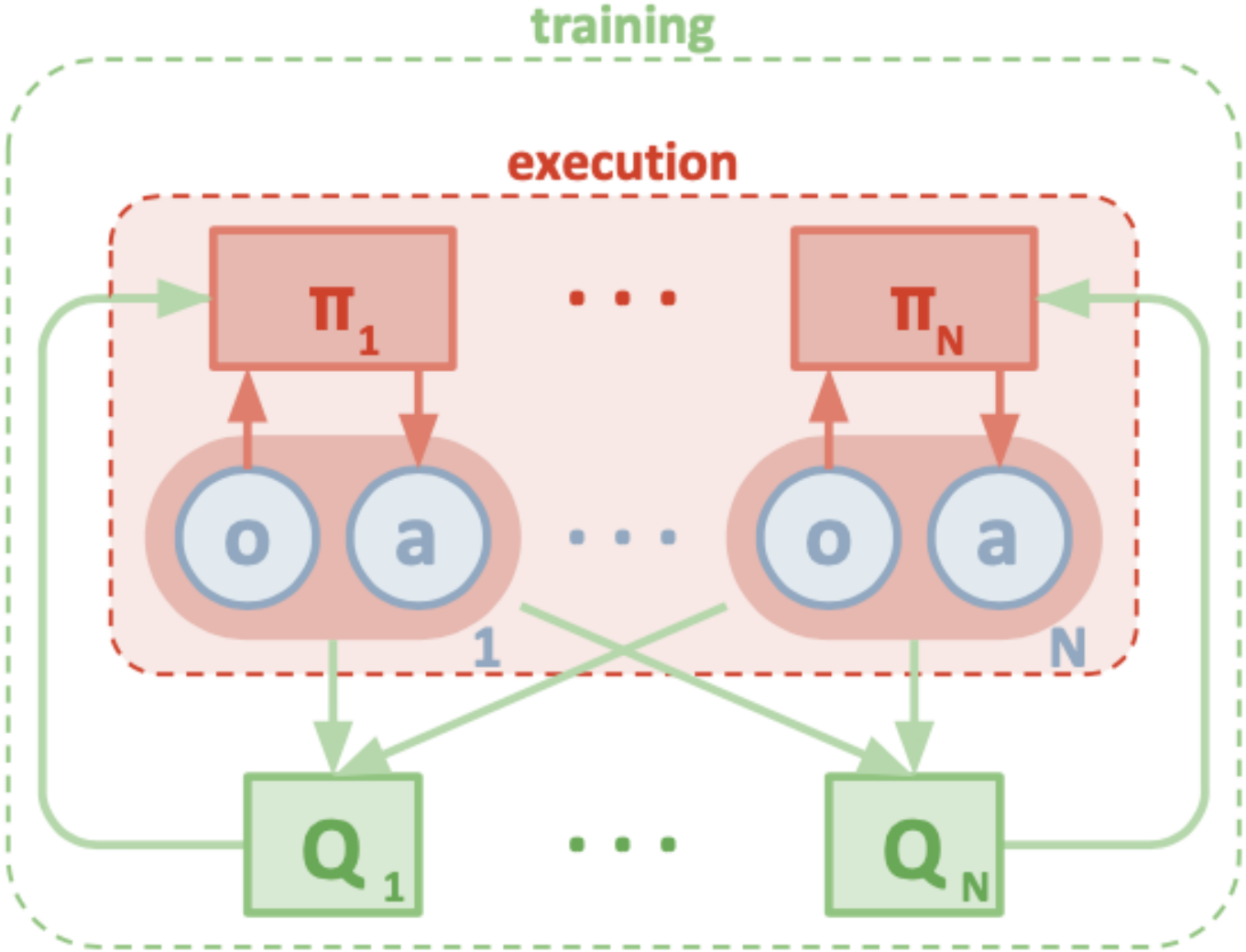


2.1. method

- Centralized training with Decentralized execution

MADDPG
(multi-agent DDPG)

agent information is
available for all critic



2.2 Inferring Policies of Other Agents

- each agent maintain approximation of agents' policy
- Maximize log probability of agents' actions

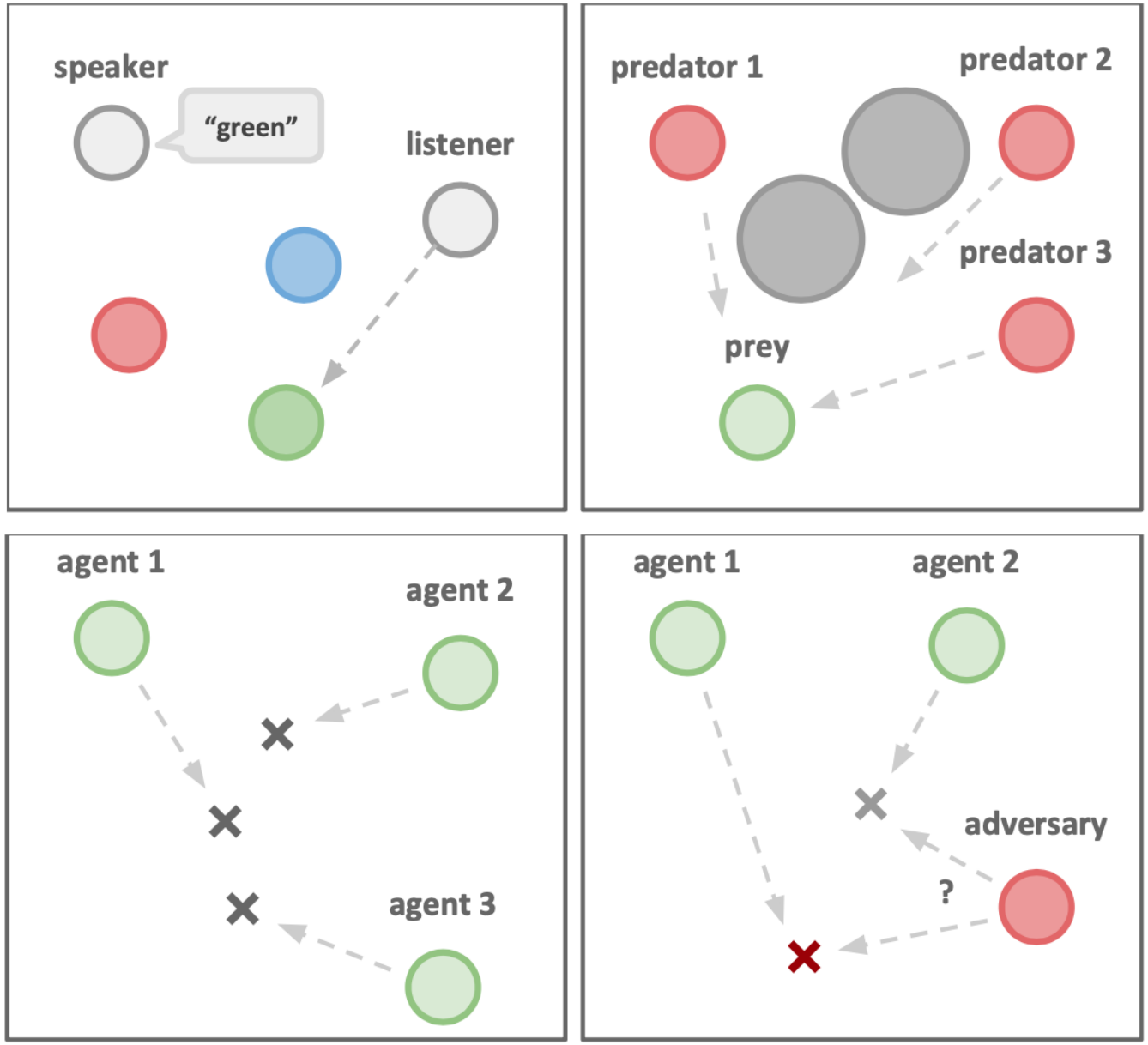
$$\mathcal{L}(\phi_i^j) = -\mathbb{E}_{o_j, a_j}\left[\log \hat{\boldsymbol{\mu}}_i^j(a_j|o_j) + \lambda H(\hat{\boldsymbol{\mu}}_i^j)\right]$$

2.3 Agents with Policy Ensembles

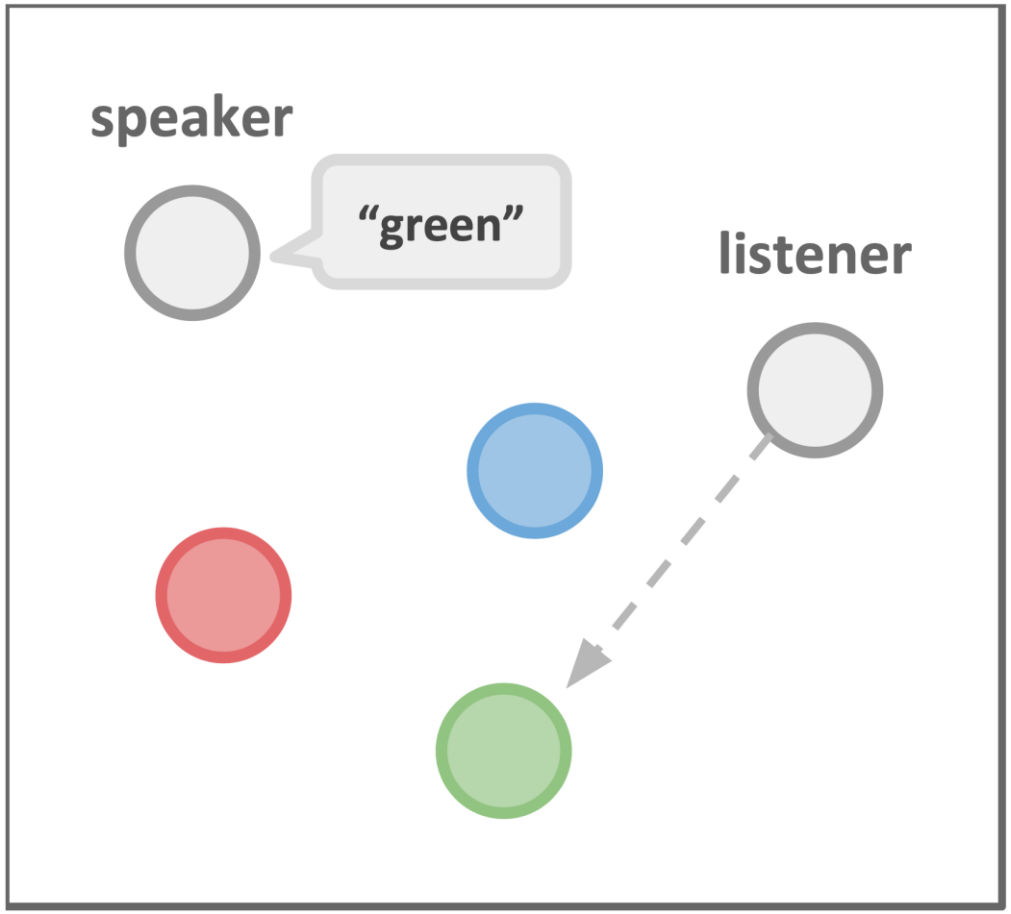
- Solve overfitting problem
- train a collection of different sub-policies

3. Environments

- Cooperative communication
- Cooperative navigation
- Keep-away
- Physical deception
- Predator-prey
- Covert communication

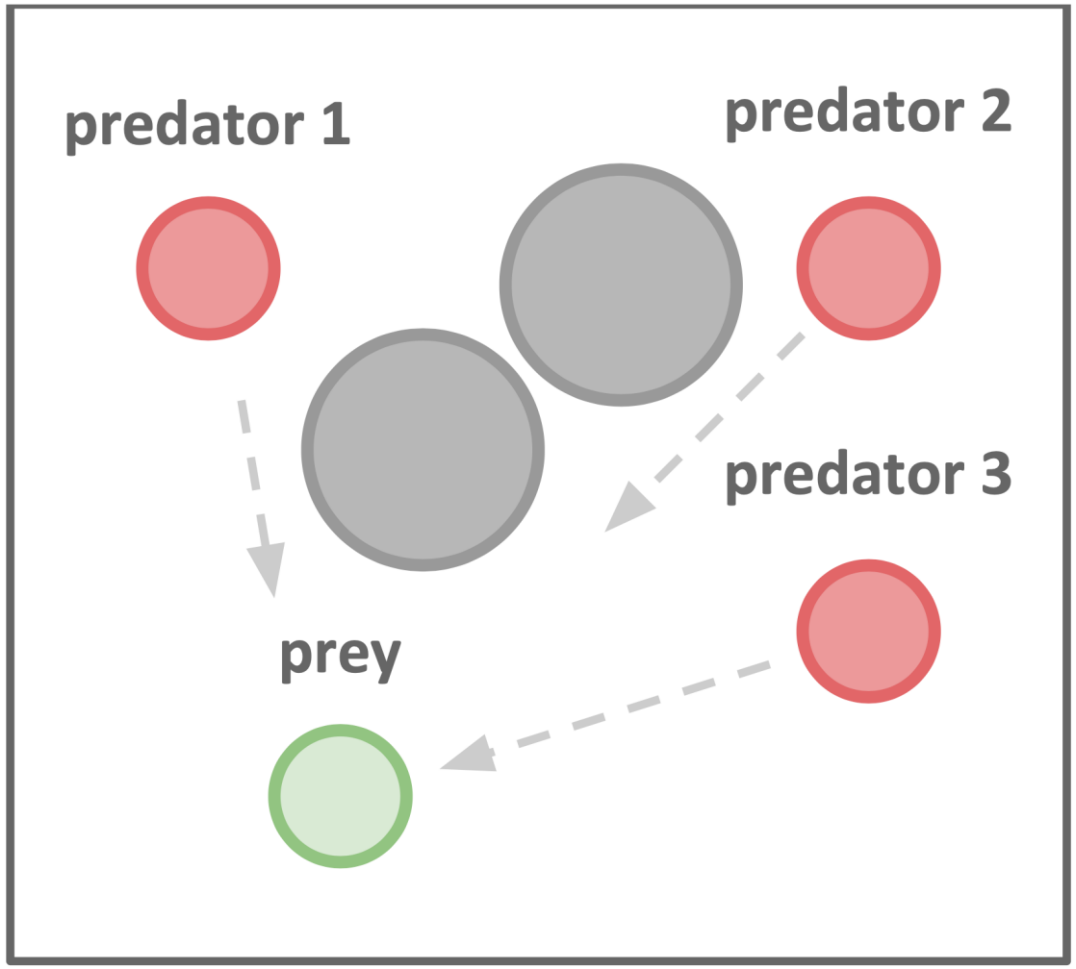


Cooperative Communication



- 2 cooperative agent
- Speaker agent teach Listener correct landmark
- goal:
  listener reach to true landmark
- Reward:
  distance from true landmark

Predator-Prey



- 1 prey, N predator, Obstacles

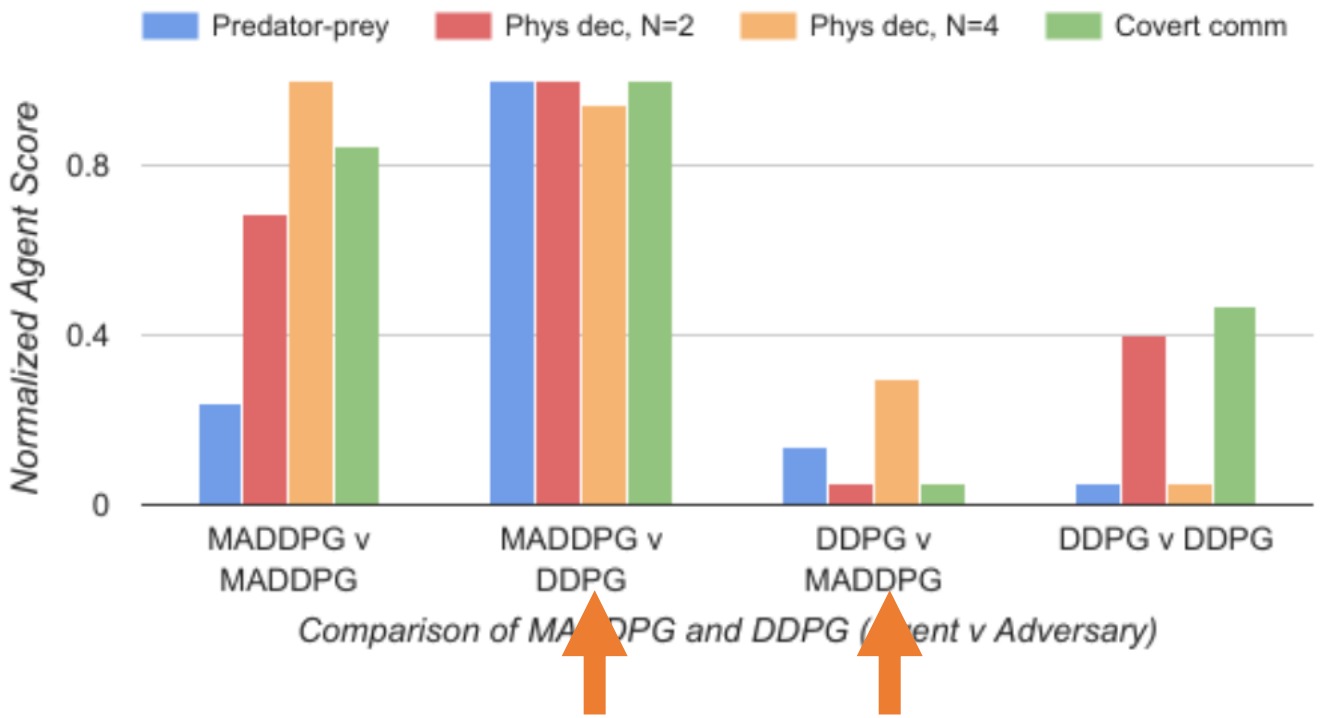
- goal, reward

prey: run away
predator: catch prey

4.Result

- MADDPG
- DDPG
- DQN
- Actor-Critic
- TRPO
- REINFORCE

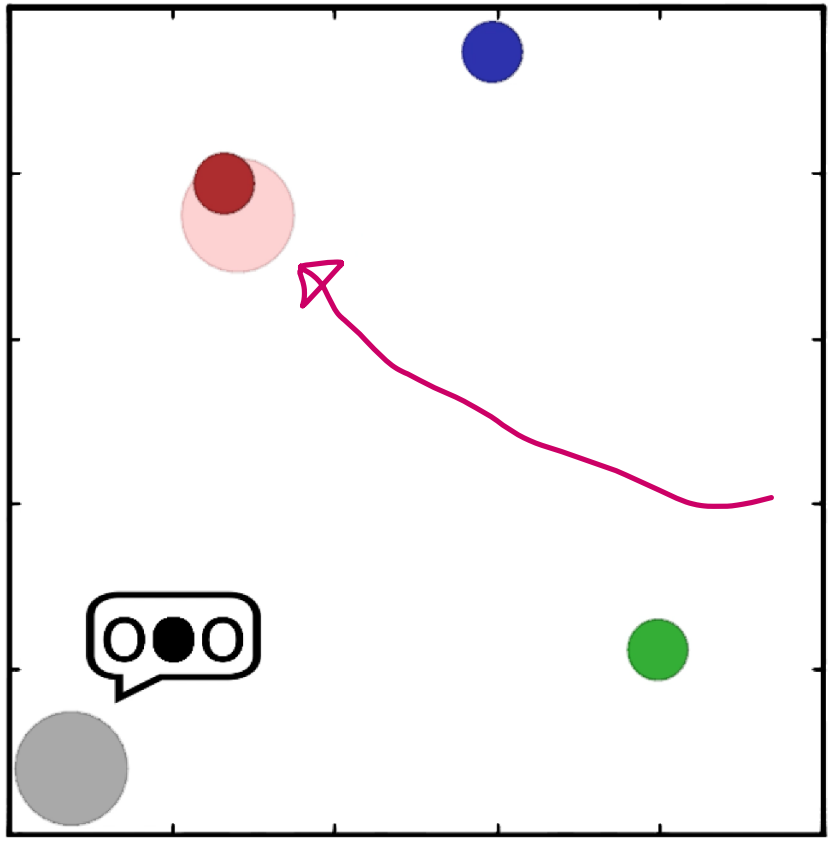
MADDPG scored highest



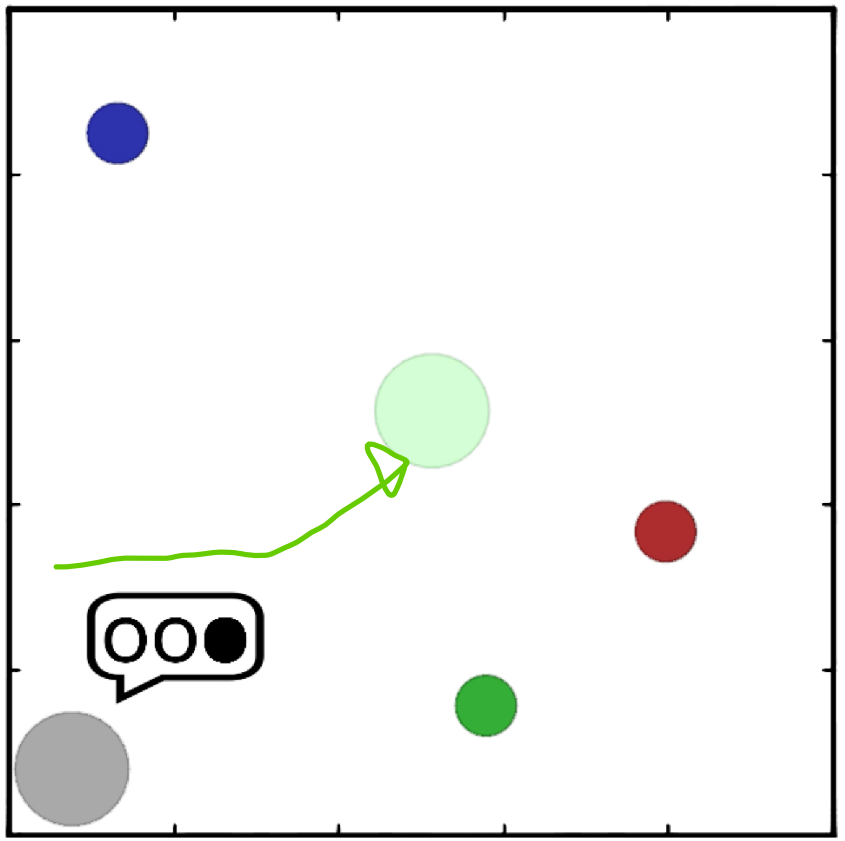
MADDPG vs DDPG

Cooperative communication

MADDPG(84%)

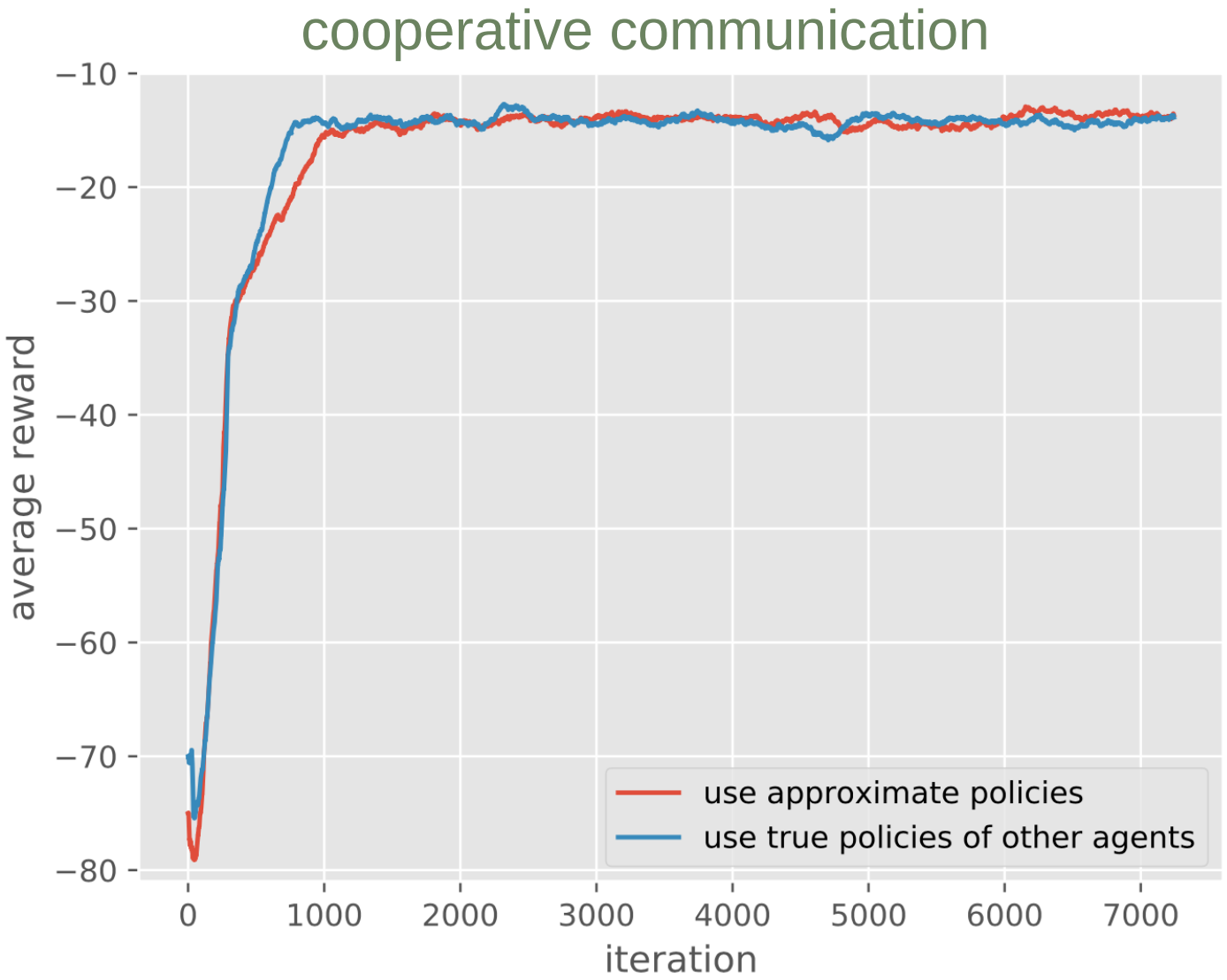


DDPG(32%)



Lack of consistent gradient signal

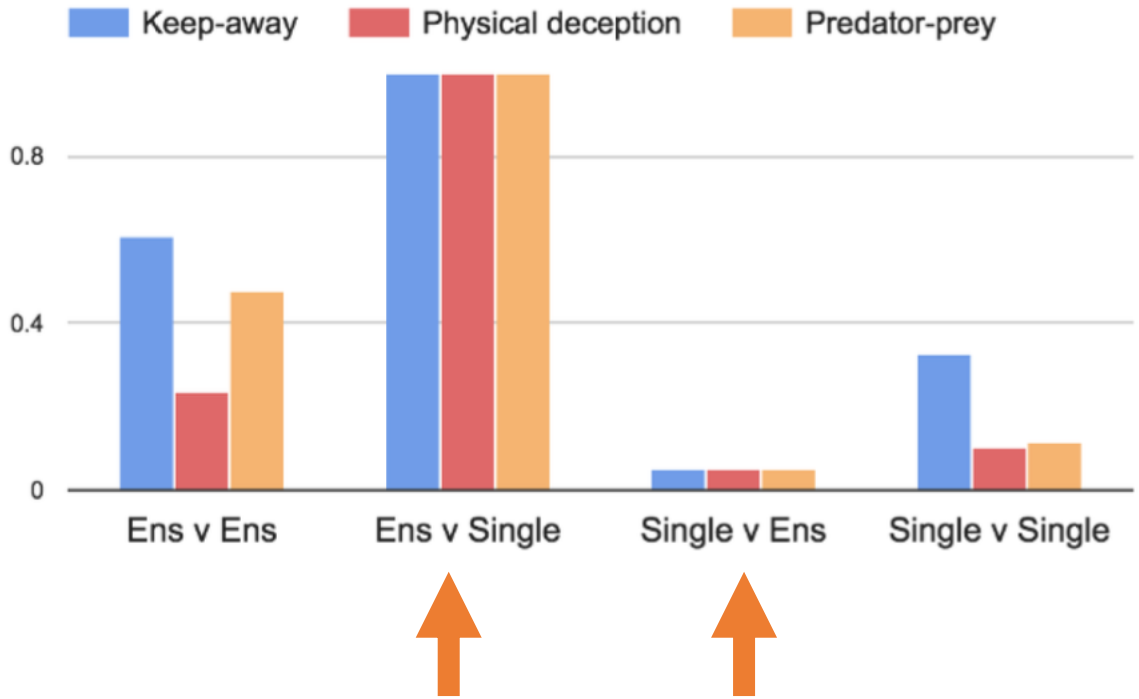
Result: Learning Policies of Other Agents



- Same success rate as using true policy

Result: Training with Policy Ensembles

- Effective in competitive environments
  - keep-away
  - cooperative navigation
  - predator-prey



Conclusion

- MADDPG was more effective than traditional RL.
- Applicable to any multi-agent algorithm.

Future work

- solve the problem;
  Input for Q grows with number of agents

Thank you for listening!