Computational Behaviors of Learning Agents in Social Dilemmas

TINGWU WANG,

MACHINE LEARNING GROUP,

UNIVERSITY OF TORONTO

Contents

Repeated General-sum Games

1. Repeated general-sum games with social dilemma
    1. Generazation of general-sum matrix games.

$R > P$ Mutual cooperation is preferred to mutual defection.

$R > S$ Mutual cooperation is preferred to being exploited by a defector.

$2R > T + S$

$T > R \quad P > S$

	C	D
C	R, R	S, T
D	T, S	P, P

2. Potential problems
    1. The tasks are arbitary
        1. The dominant strategy might not exisit or hard to discover
        2. Difficulty of understanding opponent's intrinsic intention
    2. Limited computation resource
        1. Behaviors might not be optimal, and involves exploration
        2. Local minimus.
    3. Unstatic distribution of opponent's policy
        1. How to design policy that maintains cooperation?

Contributions of the Project

1. Emiprical analysis of computational behaviors of different agents.
    1. A mixture of trainable agents and agents of pre-defined strategy.

2. Environments of complex social dilemma.
    1. Environments where multiple steps of action are needed before the reward is received.
    2. Uncertanty of the meanings of the action.

3. Design and varify the CMAR agents that could maintain cooperation in certain complex games.

Contents

Environment

1. Iterative Prisoner's Dilemma
    1. Toy Example of iterative prisoner's dilemma.
        1. Each agent observes the history of actions and states.
2. Hunting Maze
    1. Two hunters A and B in the maze.
    2. Three targets (prey) are in the environment.
        1. Prey A for hunter A (coop).
        2. Prey B for hunter B (coop).
        3. A much more valuable Prey C for both hunter (defect).
    3. Hunters could
        1. Cooperate by hunting their own prey.
        2. Defect by stealling the Prey C.
    4. The behavior itself is hard to learn.
        1. How to navigate to different prey?
        2. How to analyse opponent's behavior?
        3. What's the dominant strategy?

Different Types of Agents

1. Agent Types
    1. Naive Agents (trainable)
        1. Optimize the total reward of both agents (most preferred during evolution).
    2. Selfish Agents (trainable)
        1. Truely Rational Agents, only optimizing for itself, always looking for dominant strategy
    3. Adaptive Agents (trainable, hoping to maintain coop)
        1. Dynamically adapting its behavior.
        2. TFT Agents

        An agent using this strategy will first cooperate, then subsequently replicate an opponent's previous action.
    4. Enforcer Agents
    5. Malicious Agents

Optimization of the Policy Using Reinforcement Learning

1. Computational behavior of the agents could be formulated as the results of optimization.
2. Ideally, we should use evolutionary update.
    1. Too time consuming
    2. Variance concerns and local minimus.
3. Policy Gradient Methods

$$J_{avR}(\theta) = \sum d^{\pi_\theta}(s) \sum \pi_\theta(s,a) \mathcal{R}_s^a$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(s,a) \; Q^{\pi_\theta}(s,a) \right]$$

    1. Off-policy learning for the agents (important for our agents)
        1. Behavior policy and target policy could be different.

Approximating Cooperative-Mutual-Assured-Retaliation Agent with Neural Network

1. Inspired by TFT (Tit for Tat) agent
    1. Start with cooperation.
    2. How to classify "cooperating"?

2. Inspired by Mutual Assured Destruction
    1. The threat of using strong weapons against the enemy prevents the enemy's use of those same weapons.

    2. Raise alert level everytime your opponent does something harmful.

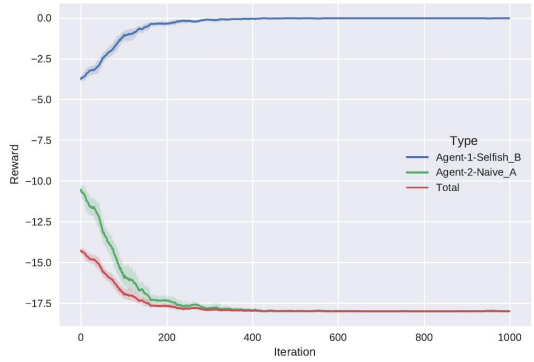
Approximating Cooperative-Mutual-Assured-Retaliation Agent with Neural Network

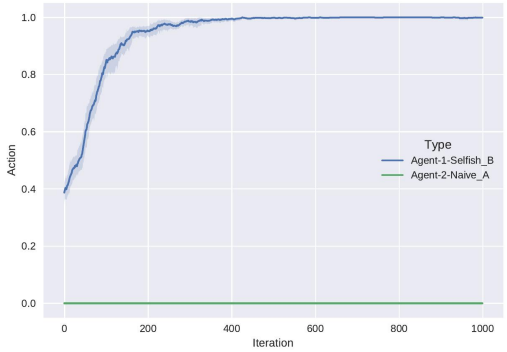
1. CMAD (Cooperative-Mutual-Assured-Retaliation) agents
    1. Assumption: All agents' state distribution (or initial state distribution) is the same or symmetry.
    2. Maintaining two target policies (off-policy learning):
        1. Cooperation Policy for optimizing total reward
        2. Retaliation Policy for optimizing the negative of total reward of opponent
    3. Behavior policy is the Cooperative-Mutual-Assured-Retaliation policy.
        1. Start with cooperating.
        2. Inference Model from the distribution of behavior policy of your opponent.
            1. If this policy is close to the distribution of your target coop policy. Use cooperative policy.
            2. Otherwise: Retaliation policy.

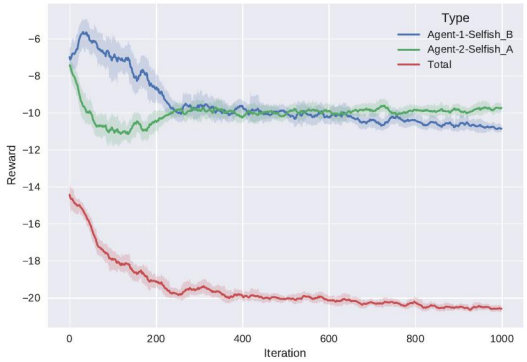
Contents

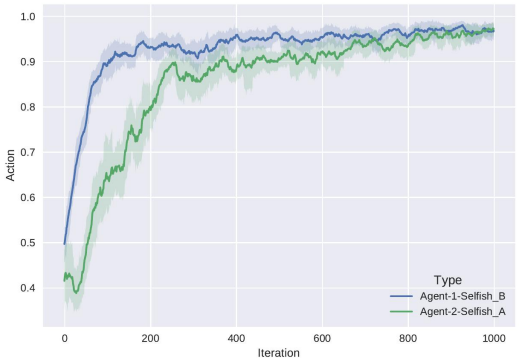
Results of Prisoner's Dilemma

1. Verify the selfish agents' defection



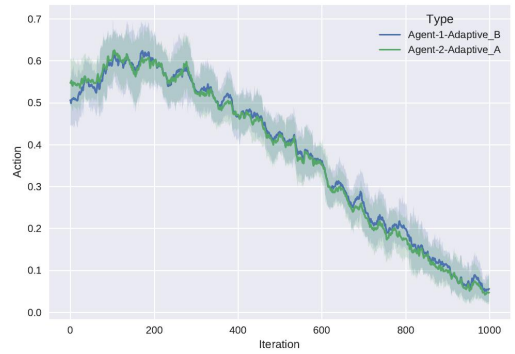


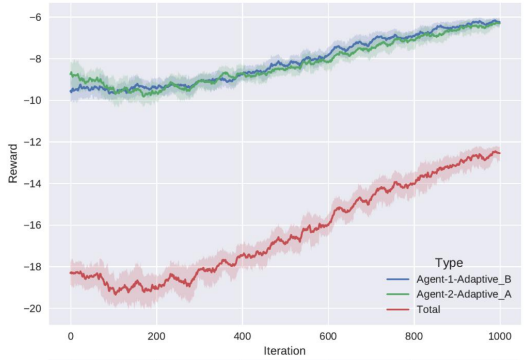


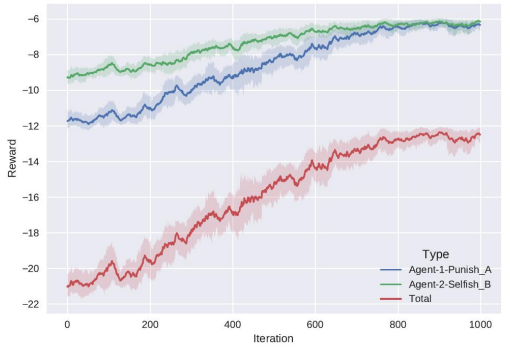


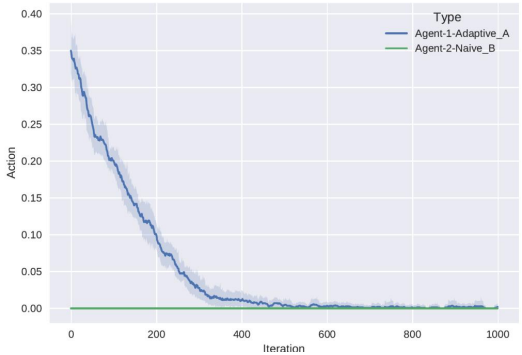
Results of Prisoner's Dilemma

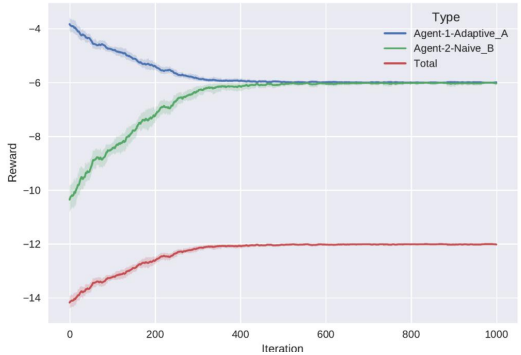
1. Verify the CMAR agents' behavior

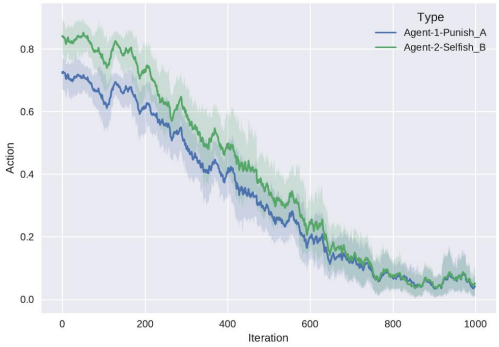






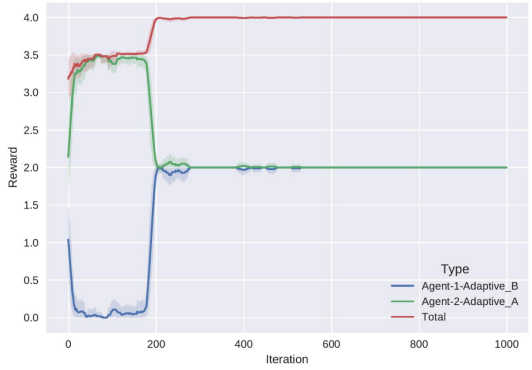


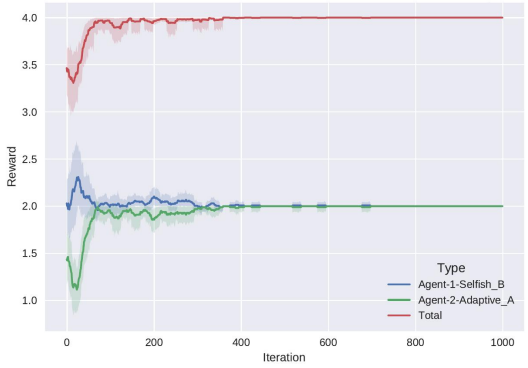


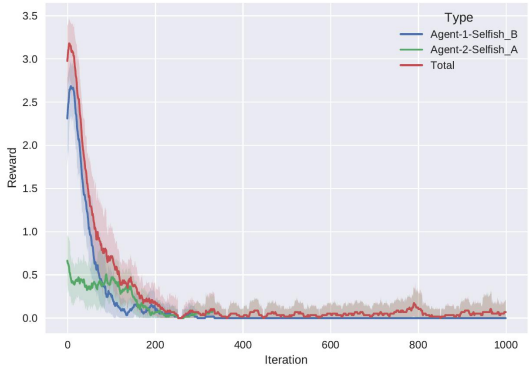


Hunting Maze

1. Similarly, for complex cases we have
    1. Selfish agents will still defect
    2. CMAR agents are cooperating
    3. CMAR agents will force the selfish agents into cooperation







Contents

Reference

To be listed in the course report.