Self-play Q-Learners Can Provably Collude in the Iterated Prisoner's Dilemma

Quentin Bertrand
Inria

Juan Duque
Mila & Université de Montréal

Emilio Calvano
Luiss University
Toulouse School of Economics

Gauthier Gidel
Mila & Université de Montréal
Canada CIFAR AI Chair

Motivation: Algorithmic Pricing and Collusion

- Algorithmic pricing has supplanted manual pricing
  ↪ $\approx 1/2$ of Amazon's largest third-party sellers
- Academic and institutional concerns over *tacit collusion*
  ↪ OECD, Competition Bureau Canada

Question: *Could pricing algorithms autonomously learn to cooperate, thereby leading to higher prices?*

Contributions:

- For optimistic enough Q-values, *self-play* Q-learning guided agents learn a cooperative policy
- Extend the latter result for $\epsilon >$ 0-greedy Q-learning guided agents
- Empirically demonstrated the convergence to a cooperative policy

Problem Setting:

- Iterated Prisoner's Dilemma
  ↪ $r_{\text{DC}} > r_{\text{CC}} > r_{\text{DD}} > r_{\text{CD}}$ and $2r_{\text{CC}} > r_{\text{CD}} + r_{\text{DC}}$

	Cooperate	Defect
Cooperate	r_{CC} r_{CC}	r_{CD} r_{DC}
Defect	r_{DC} r_{CD}	r_{DD} r_{DD}

- Multi-agent Q-Learning

$$Q^{\star}_{s_t,a^1_t} = \mathbb{E}_{a^2_t \sim \pi_2(\cdot|s_t)} \left(r^1_{a^1_t,a^2_t} + \gamma \max_a Q^{\star}_{(a^1_t,a^2_t),a} \right) .$$

$$Q^{t+1}_{s_t,a^1_t} = Q^{t}_{s_t,a^1_t} + \alpha \left(r^1_{a^1_t,a^2_t} + \gamma \max_{a'} Q^{t}_{(a^1_t,a^2_t),a'} - Q^{t}_{s_t,a^1_t} \right)$$

- Self-play

$$a^1_t, a^2_t \sim \pi(\cdot|s_t) \quad \text{// same } Q\text{-table for } a^1_t \text{ and } a^2_t$$

$$Q^{\star}_{s_t,a^1_t} = \mathbb{E}_{a^2_t \sim \pi(\cdot|s_t)} \left(r_{a^1_t,a^2_t} + \gamma \max_a Q^{\star}_{(a^1_t,a^2_t),a} \right)$$

- Epsilon greedy

$$\pi(a|s) = \begin{cases} 1-\epsilon & \text{if } a = \arg\max_a Q_{s,a} \\ \epsilon & \text{else} \end{cases}$$

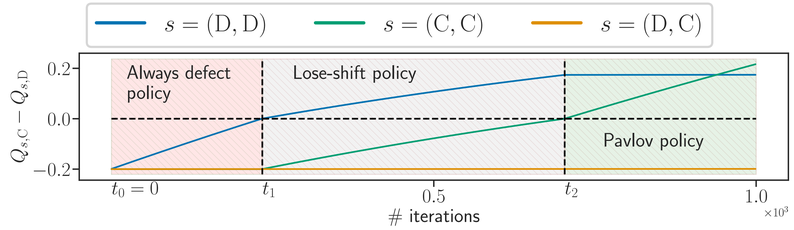
Theoretical Results

Challenges:

- Multiple fixed-point policies
- Show convergence toward a specific policy

Assumptions: Optimistic enough Q-values

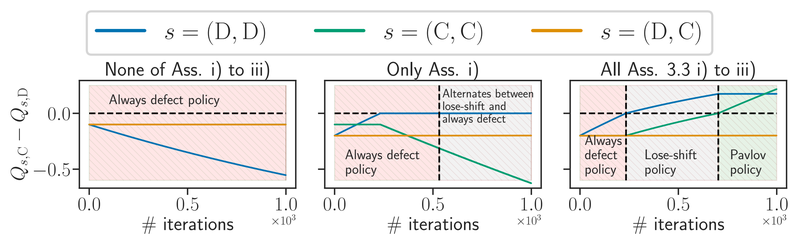
- $\frac{r_{\text{DD}}}{1-\gamma} < Q^{t_0}_{(\text{D,D}),\text{C}}$
- $Q^{t_0}_{(\text{D,D}),\text{C}} < \frac{r_{\text{CC}}}{1-\gamma} - \frac{r_{\text{CC}}-r_{\text{DD}}}{1-\gamma^2} < Q^{t_0}_{(\text{C,C}),\text{C}}$
- $Q^{t_0}_{(\text{C,C}),\text{C}} < \frac{r_{\text{CC}}}{1-\gamma}$



Evolution of the Q-values as a function of the number of iterations. Agents move from the *always defect* policy to the *cooperative Pavlov* policy.

Theorem 1: Fully Greedy Q-learning

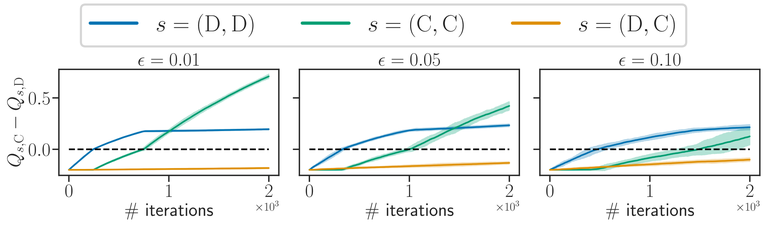
Suppose the initial policy is always defect and the initial state s_0 is `defect` `defect`: $s_0 =$ DD. Under the optimistic Q-values initialization, Q-learning guided agents move away from the *always defect* policy and learn the *cooperative Pavlov policy*.



Theorem 2: ϵ-greedy case with $\epsilon > 0$

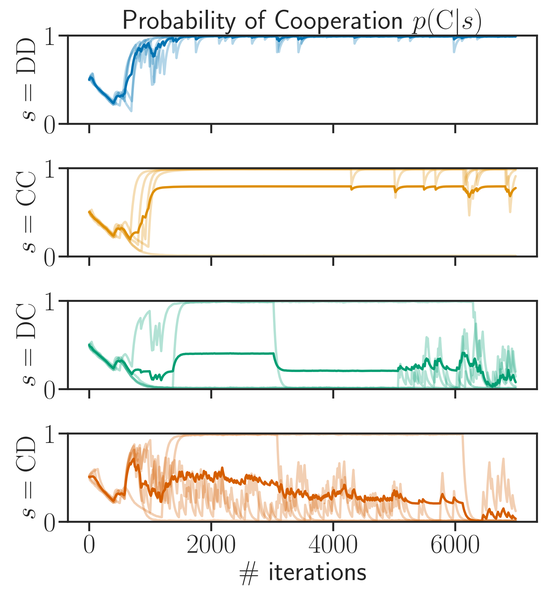
Suppose the initial policy is always defect and the initial state s_0 is `defect` `defect`: $s_0 =$ DD. Under the optimistic Q-values initialization, Q-learning guided agents move away from the *always defect* policy and learn the *cooperative Pavlov policy* with high probability.

Experimental Results



Evolution of the Q-values as a function of the number of iterations. The larger the exploration parameter ϵ, the smaller the number of trajectories yielding cooperation.

Extension to deep Q-learning:



References

E. Calvano, G. Calzolari, V. Denicolo, and S. Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 2020.

J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *Conference on Autonomous Agents and Multiagent System*, 2017.

M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and . Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *NeurIPS*, 2017.