Q-Learning in Pricing Games

Pranjal Rawat

May 15, 2023

Georgetown University

Introduction

What helps or hinders algorithmic collusion when Q-learning pricing bots interact in the market?

We will consider 7 factors:

- Algorithms: exploratory strategies, discount factor, update rule
- Feedback: past prices, reputation
- Environment: number of competitors, seasonal demand

Setup

Stage Game

Two firms compete in a differentiated products pricing game:

$$D(p, p') = \frac{e^{(a_1 - p)/\mu}}{e^{(a_1 - p)/\mu} + e^{(a_2 - p')/\mu} + e^{a_0/\mu}}$$

$$\pi(p, p') = (p - c)D(p, p')$$

Parameters:

- a_i: Quality of outside, own and opponents' good
- μ: Product own-price sensitivity
- c: constant marginal cost

Rewards

$$a_0 = 2, a_1 = 20, a_2 = 20, \mu = 0.1, c = 0$$

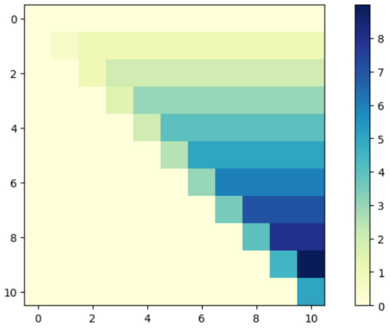


Figure 1: Stage Game Rewards for Firm 1

Dynamics

- Action: $p_t \in \{0, 1, 2, 3, 4...10\}$
- State: s_t can be opponents' past price p'_{t-1}
- Policy: $\sigma(s_t)$ is a strategy of choosing actions given states
- Discount factor: γ

For policy σ, every state s has a **expected return**, assuming the opponents plays by σ'.

$$V_{\sigma,\sigma'}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \pi(p_t, p'_t) \mid s_0 = s, \sigma, \sigma'\right]$$

$$Q_{\sigma,\sigma'}(s, p) = \pi(p, p') + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t \pi(p_t, p'_t) \mid \sigma, \sigma'\right]$$

Q-Learning

"Experience based equilibrium" is $(\bar{\sigma}, \bar{\sigma'})$ arrived at by an iterative process. One such process is Multi-agent Q-learning:

- Guess $Q_0(s, p)$
- at t, do:
  - observe s_t
  - take action p_t from **exploratory strategy**
  - collect reward π_t
  - observe transition s_{t+1}
  - update Q at point (s_t, p_t):

$$Q_{t+1}(s_t, p_t) = (1-\alpha)Q_t(s_t, p_t) + \alpha(\pi_t + \gamma max_{p'} Q_t(s_{t+1}, p'))$$

  - Learning rate α: weight given to current experiences over past

As bots' Q tables stabilize, $\bar{\sigma}(s) = \text{argmax}_p Q(s, p)$ is going to be optimal against $\bar{\sigma'}(s)$ and both will be best responses to each other.

Exploration

Given s_t, choose p_t:

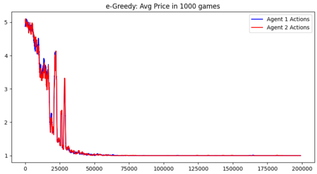
- Random: $P(p_t|s_t) = 1/|A|$
- Greedy: $p_t = \text{argmax}_p Q_t(s_t, p)$
- ϵ-Greedy: Random with $1 - \epsilon$ and greedy with ϵ
- Boltzmann Exploration:

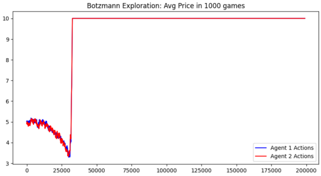
$$P(p_t|s_t) = \frac{e^{Q_t(s_t,p_t)/\beta}}{\sum_{p'} e^{Q_t(s_t,p')/\beta}}$$

Results

Simulation 1: Exploration

No state or discounting $\gamma = 0$, $\alpha = 0.3$. Boltzmann Exploration leads to collusion while ϵ-Greedy does not.

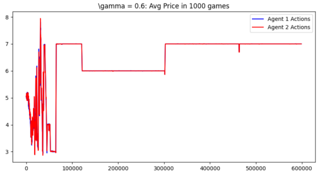


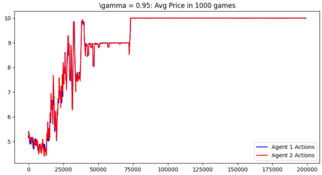


Exploring more in lower profit regimes aids collusion (Waltman et al 2008, Dolgopolov 2022).

Simulation 2: Patience

No state, $\alpha = 0.3$ and ϵ-Greedy exploration.

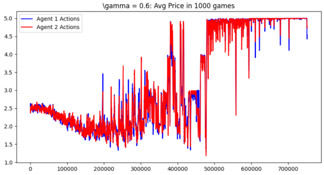


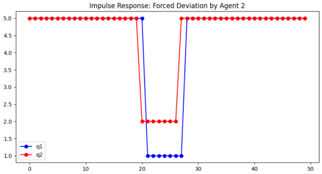


Being patient provides impetus for collusion.

Simulation 3: Memory

State p'_{t-1}, $\gamma = 0.95$, $\alpha = 0.3$ and ϵ-Greedy exploration.

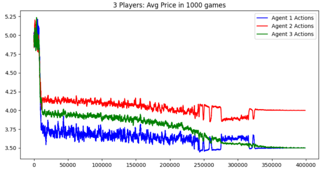


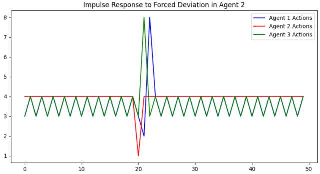


Threat of retaliation via memory supports collusion. (Axelrod 1980, Calvano et al 2020).

Simulation 4: Competitors

State (p'_{t-1}, p''_{t-1}), $\gamma = 0.95$, $\alpha = 0.1$, ϵ-Greedy exploration, $\mu = 2$.

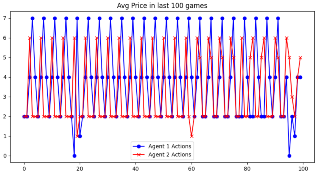


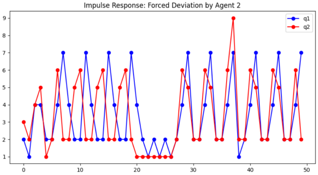


Large numbers increase competitive pressure, but it may not be enough.

Simulation 5: Cycles

Every even period, demand for product reduces by small amount. State p'_{t-1}, discounting $\gamma = 0.95$, $\alpha = 0.3$ and ϵ-Greedy exploration.

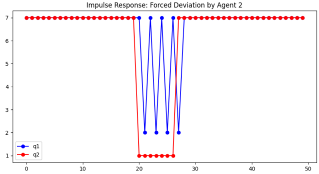


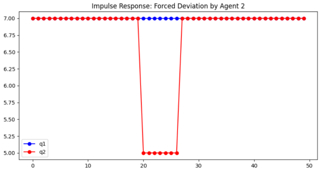


Cyclical variation is not a barrier to collusion, and Edgeworth cycles are common (Klein 2021).

Simulation 6: Reputation

State R'_{t-1}, which is set to 0 if opponent prices 2 units below, else 1.
$\gamma = 0.95$, $\alpha = 0.3$, ϵ-Greedy exploration.



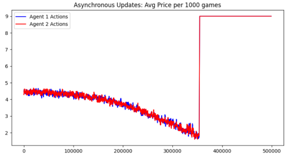


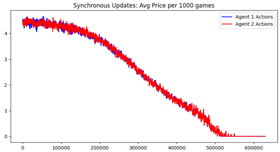
Reputation simplifies tracking others' actions (Nowak 2006).

Simulation 7: Knowledge

No state, no discounting, $\alpha = 0.3$, Boltzmann exploration. Firms know reward table, so can update entire Q-table at once:

$$Q_{t+1} = (1-\alpha)Q_t + \alpha\pi(., p'_t)$$





Knowing demand/rewards perfectly prevents collusion (Asker et al 2022, Banchio and Skrzypacz 2022).

Conclusions

There are many factors that can help or hinder collusion. The nature of the learning algorithm, the environment and the feedback loop are all important.

Future research should explore:

- intelligent and faster learning - model-based, expert recommendation
- dynamics - delayed reward/feedback, adjustment costs
- alternative mechanisms - networks, group formation