Q-Learning to cooperate

Emilio Calvano,
(with Giacomo Calzolari, Sergio Pastorello, Vincenzo Denicoló)

U Bologna and Toulouse School of Economics

September 14, 2018

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Paper, Motivation and Agenda

- Run experiments with AI agents in controlled environments. (computer simulations)
- Inform debate on impact on **algorithms & competition**
- AI may potentially replace human subjects in the lab

1. Iterated prisoner's dilemma
2. Iterated price oligopoly with differentiated goods

Q-Learning

- Perception: ~~state of the board~~ all past prices, demand
- Actions: ~~legal moves~~ own price
- Reward: ~~+1/-1 end of game~~ period profit

Q-learning: why?

- **Natural choice**: designed to "crack" Markov Decision Problems
- **Model free**: versatile
- **Popular:** building block of many deep learning algos e.g. video-games Nature paper: Mnih et al (2015)
- **Not fancy** (tabular solution method)

Only three design dimensions

- Rate of learning $\alpha \in [0, 1]$
- Rate (and type) of experimentation $\beta \geq 0$
- Discounting δ

Baseline Game and setup

Baseline model

- 2 players
- Differentiated goods
- Logit demand, constant mc
- Fully Symmetric

Baseline Implementation

- 1 period memory: state space = last period prices.
- 15 price points $\underline{p} = \frac{9}{10} p_{\text{Nash}}$, $\overline{p} = \frac{11}{10} p_{\text{mon}}$

Departures from baseline (one at a time)

$$\max_{p_i} (p_i - c_i) \frac{e^{\frac{b_i - p_i}{\sigma}}}{\sum_j e^{\frac{b_j - p_j}{\sigma}} + e^{\frac{b_0}{\sigma}}}$$

- **3** players
- **Asymmetric** Demand: $b_1 > b_2$
- **Asymmetric** cost: $c_1 > c_2$
- Demand increases $b_0 \uparrow$
- Differentiation increases $\sigma \uparrow$
- **2** period memory
- **30** price points $\underline{p} = \frac{1}{2} p_{\text{nash}}$, $\overline{p} = \frac{3}{2} p_{\text{mon}}$

Approach

- Look at grid of parameters α, β, δ
- 435 parametrizations in total (baseline).
- Agents play (up to) 1 billion iterations per session
- 1000 sessions for each parametrization
- We report averages across sessions and parameterizations

Q-learning means:

- strategy $\sigma_i^t(p_1^{t-1}, p_2^{t-1})$ evolves over time
- How? actions that 'perform well' are reinforced

We observe both prices and strategies (!) and report on both!

Results

Two Q-learning agents interacting repeatedly typically:

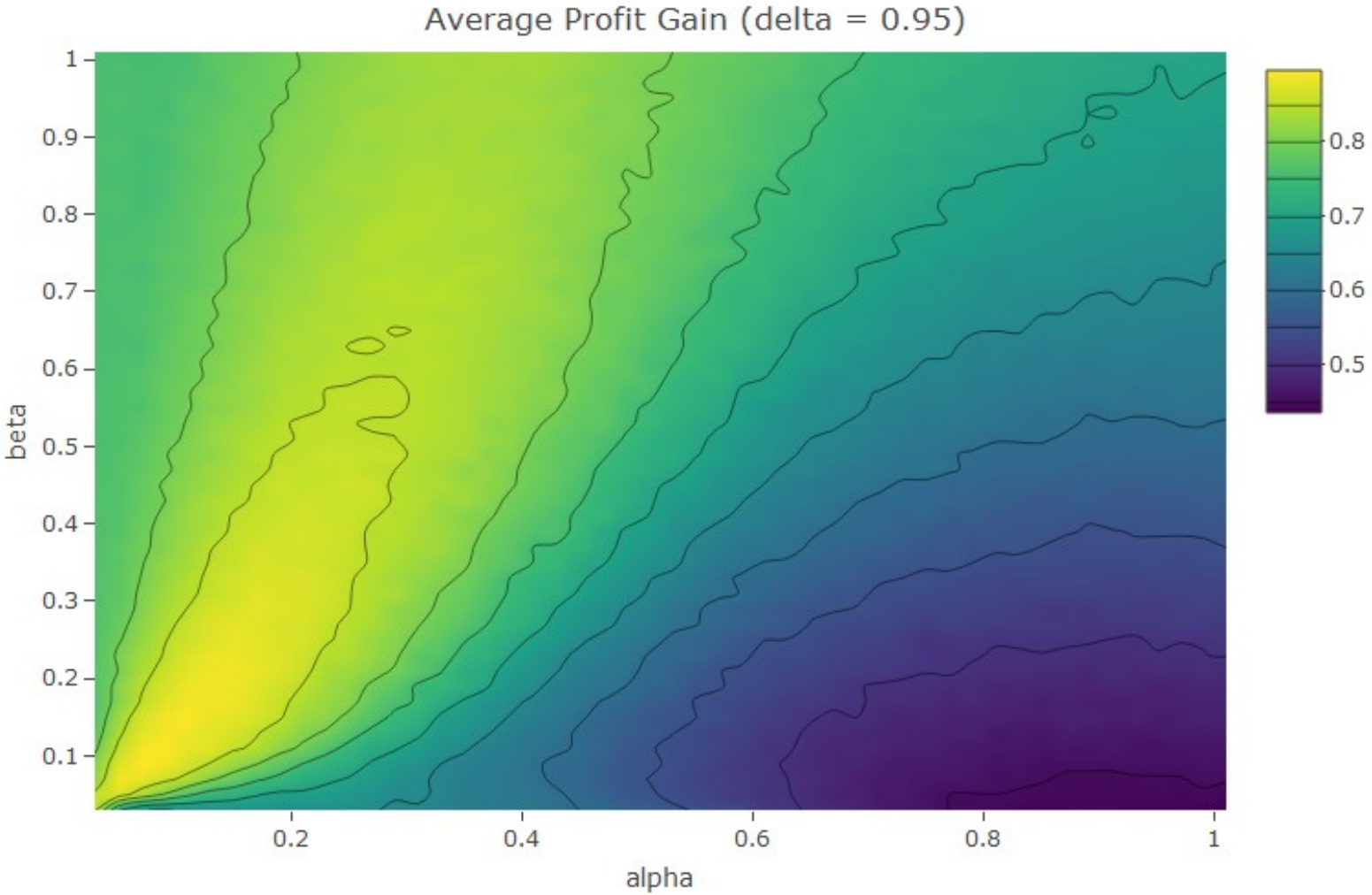
1. Learn to Play (Converge)
2. Learn to Cooperate
3. Learn to Collude

1 - Convergence

- convergence = strategy does not change for 25k iterations.
- 99.9% sessions converge.
- takes 1.6M iterations on average over the grid
- **Not obvious:** no theoretical guarantees due to non stationary environment.
- Somewhat **fast**: few minutes in CPU time
- Somewhat **slow**: they can't "learn by doing."
- They need to be **trained**!

2(a) - Cooperation over the grid

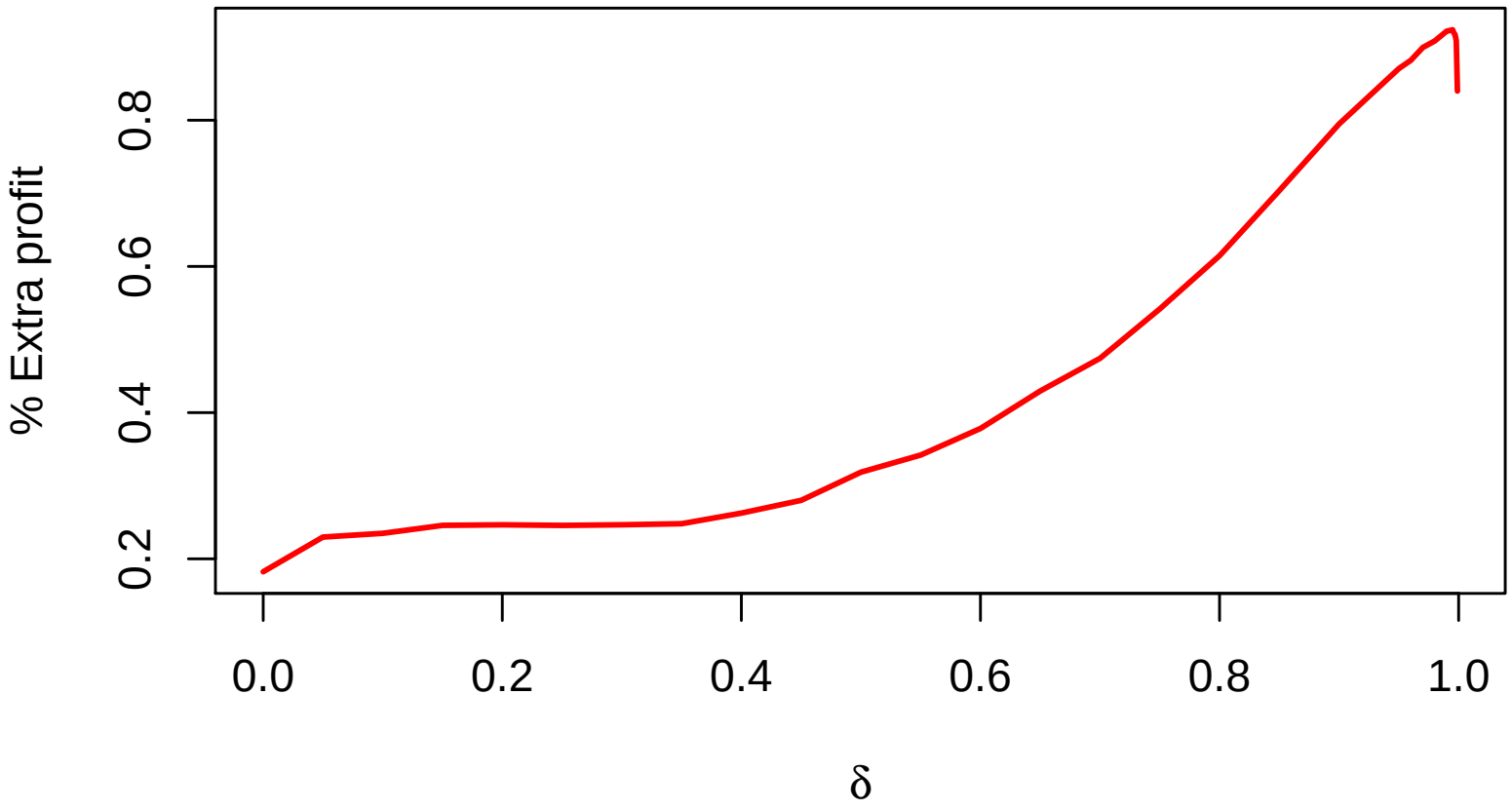
- Let $\Delta = \pi^{\text{collusion}} - \pi^{\text{nash}}$ be the 'extra profit'



$\%\Delta(\alpha, \beta)$

2(b) - Cooperation & discounting

- % Extra profit Δ as a function of δ for $\alpha = 0.15$, $\beta = 0.3$



3 - Learn to collude: Impulse response of prices

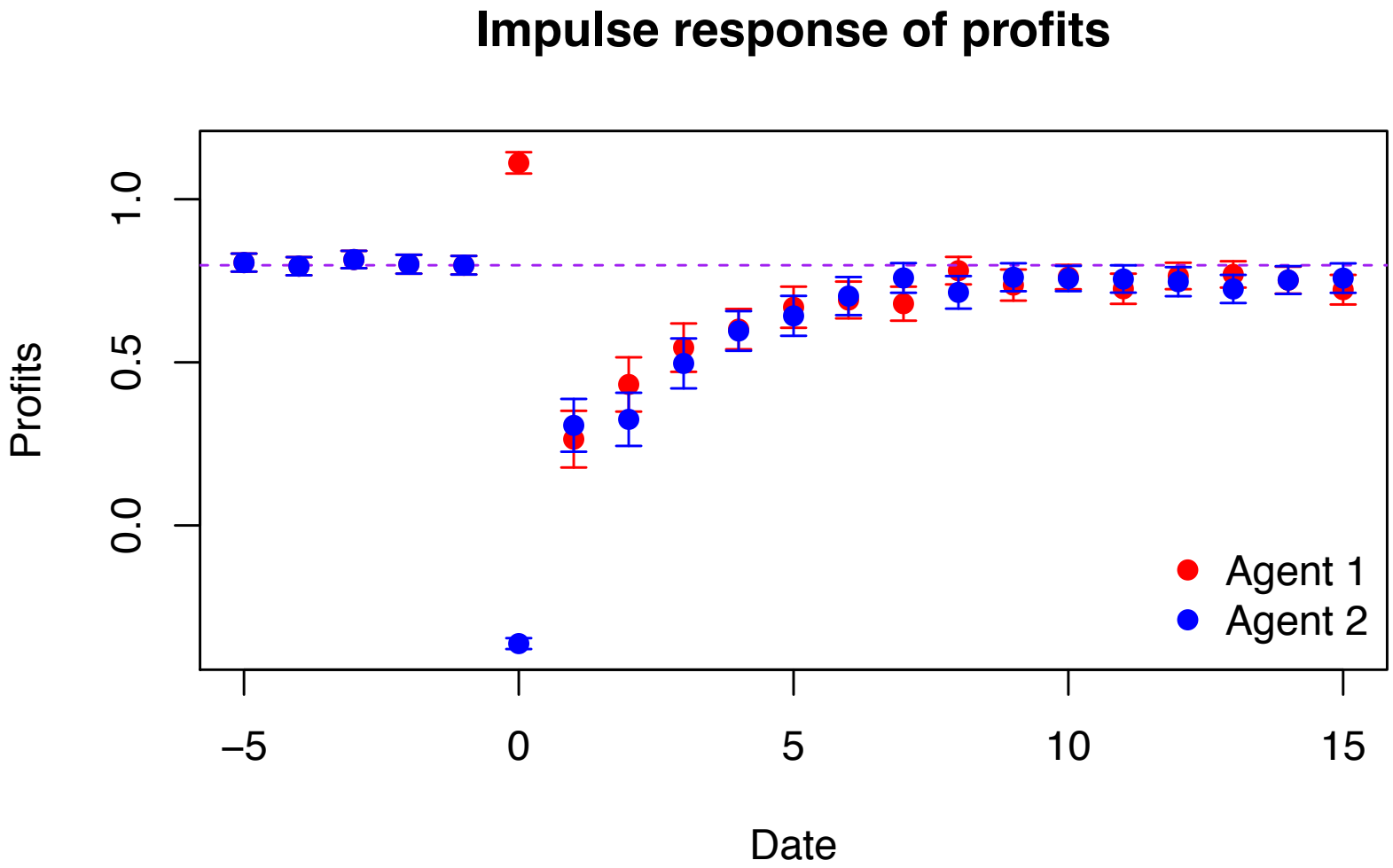
- Let agents play according to learnt strategies
- Agent 1 (blue line) deviates charging p_{nash} at $t = 0$
- Showing average of 200 impulse responses to such shock

parameters: $\delta = 0.95, \alpha = 0.05, \beta = 0.3$

Same exercise - just zooming in

Same exercise, looking at profits

- Normalized $1 = \pi^{\text{collusive}}$



Robustness

- % extra profit Δ over the baseline grid

	max	min	avg	median
Baseline	99.1	12.7	57.7	59.1
3 players	80.4	32.1	67.4	69.3
30 prices	86.5	26.9	70.1	73.91
Asymmetric demand: $b_1 = 1.5b_2$	59.7	6.8	36	36.8
Asymmetric cost: $c_1 = 1.5c_2$	85.2	8.3	47.8	47.7
Differentiation $\downarrow$: $\sigma' = \sigma/5$	97.9	12.6	57.7	58.6
2 Period Memory (in progress...)	?	?	?	?