Exploration-Exploitation in Multi-Agent Competition

Convergence with Bounded Rationality

Stefanos Leonardos, Georgios Piliouras and Kelly Spendlove

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

UNIVERSITY OF OXFORD

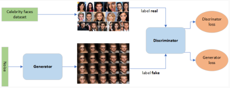
35th Conference on Neural Information Processing Systems, December 6-14, 2021

Introduction: Multi-Agent Competition

Motivation

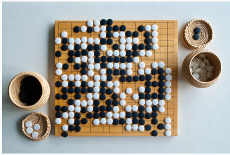
Many recent ML and AI advances involve competitive interactions between 2-agents

- generative adversarial networks (GANs)
- actor-critic systems
- competitive game-playing: chess, Go



Modelled as strictly-competitive, 2-agent, zero-sum games or variants thereof

- multiple equilibria but unique value
- equilibrium strategies are exchangeable
- optimization-driven algorithms perform well



What happens beyond these settings?

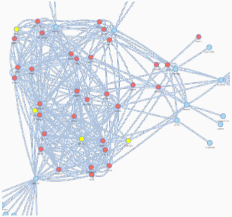
Motivation: Open Questions

In multi-agent competition, many properties of the 2-agent settings collapse

- multiple but payoff-diverse equilibria
- exploration-exploitation for equilibrium selection

Multi-agent vs 2-agent competition

- not only significantly harder
- but also qualitatively different



Research goals: in networks of strictly competitive games

- convergence of exploration-exploitation dynamics
- equilibrium selection with payoff-diverse equilibria

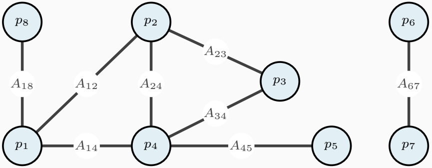
Game-Theoretic Model

Weighted Zero-sum Polymatrix Games

A weighted zero-sum polymatrix game (WZPG), $\Gamma = \Big((V,E), (S_k, w_k)_{k\in V}, (\mathbf{A}_{kl})_{[k,l]\in E}\Big)$

$$u_k(\mathbf{x}) := \mathbf{x}_k^\top \sum_{[k,l]\in E} \mathbf{A}_{kl}\mathbf{x}_l = \mathbf{x}_k^\top r_k(\mathbf{x}_{-k})$$

$$\sum_{k\in V} w_k u_k(\mathbf{x}) = 0, \quad \text{for all } \mathbf{x} \in \Delta.$$



Nash Equilibrium (NE): a strategy profile, $\mathbf{p} = (p_k)_{k\in V} \in \Delta$, with one strategy for each agent $k \in V$, $\mathbf{p}_k = (p_{ki})_{i\in S_k} \in \Delta_k$ such that

$$u_k(\mathbf{p}) \geq u_k(x_k, \mathbf{p}_{-k}), \quad \text{for all } x_k \in \Delta_k, k \in V.$$

Properties: WZPGs capture complexities of multi-agent competition

- multiple NE with non-unique payoff values and non-exchangeable NE strategies

Joint Learning Model

Q-Learning Dynamics (QLD)

Q-value updates and Boltzmann selection probabilities for all agents $k \in V$

$$\frac{\dot{x}_{ki}}{x_{ki}} = \underbrace{r_{ki}(\mathbf{x}_{-k}) - \mathbf{x}_k^\top r_k\left(\mathbf{x}_{-k}\right)}_{\text{exploitation}} - T_k \underbrace{\left[\ln\left(x_{ki}\right) - \mathbf{x}_k^\top \ln\left(\mathbf{x}_k\right)\right]}_{\text{exploration}}, \tag{1}$$

Exploration rates T_k:

- $T_k = 0$: select action with highest Q-value (exploitation)
- $T_k \to \infty$: uniformly randomize over actions (exploration)

Interpretation of T_k's:

- physics: temperature of the system
- behavioral: agents bounded rationality or discounting of past payoffs
- algorithmic: regularization to avoid boundary or local optima

Solution Concept: QRE

Quantal Response Equilibria

Quantal Response Equilibria (QRE), $\mathbf{p} = (p_k)_{k \in V}$, of Γ

- standard solution concept in games with bounded rationality
- logit (softmax) form that depends on exploration rates

$$p_{ki} = \frac{\exp\left(r_{ki}/T_k\right)}{\sum_{j \in S_k} \exp\left(r_{kj}/T_k\right)}, \quad \text{for all } i \in S_k, k \in V. \tag{2}$$

- may be very different from NE, but not when T_k are close to 0.

Theorem (Interior Fixed Points of QLD)

The interior fixed points, $\mathbf{p} = (p_k)_{k \in V}$, of the Q-learning dynamics in an arbitrary game Γ with positive exploration rates, $T_k > 0$, always exist and coincide with the QRE of Γ.

A strategy profile $\mathbf{p} = (p_k)_{k \in V}$ is an interior fixed point of QLD if the RHS in (1) is 0.

Main Result

Convergence of Q-Learning to QRE in Multi-Agent Competition

Main Theorem (Informal)

Let Γ be a WZPG, with positive exploration rates, $T_k > 0$, for all $k \in V$. There exists a unique QRE, $\mathbf{p}$*, such that any trajectory,* $\mathbf{x}(t)$*, of the Q-learning dynamics starting from an arbitrary interior point, converges to* $\mathbf{p}$ *exponentially fast.*

Takeaways

- despite the diversity of NE, we have uniqueness of QRE
- as $T_k \to 0$, QRE approaches a NE of Γ: way out of tight spot of equilibrium selection

Remarks

- tight assumptions: if $T_k = 0$ for some $k \in V$, then QLD may converge to the boundary even for interior starting points.
- prior work: QLD provably converges in multi-agent coordination.

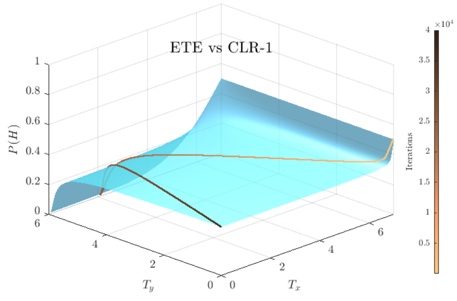
Experiments

Visualization of the QRE Manifold

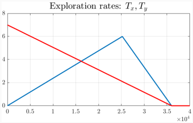
Asymmetric Matching Pennies (AMPs): 2-agent, weighted zero-sum game with

$$\mathbf{A} = \begin{pmatrix} 2 & -2 \\ 0 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 4 & 0 \\ -4 & -4 \end{pmatrix},$$

so that $\mathbf{A} + 0.5 \cdot \mathbf{B}^\top = 0$ and a unique interior NE at $(\mathbf{p}, \mathbf{q}) = ((1/3, 2/3), (2/3, 1/3))$.



- ETE: explore-then-exploit
- CLR-1: cyclical learning rate (1-cycle)

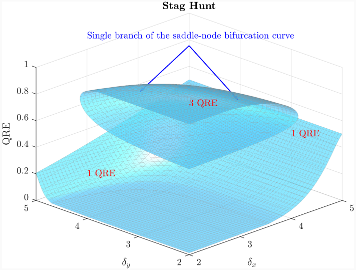


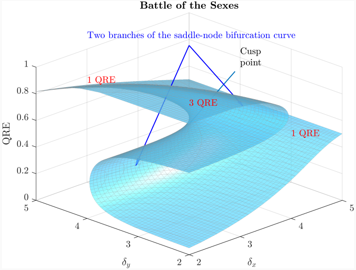
- QRE manifold and exploration path

Excursion: QLD in Multi-Agent *Coordination*

Multi-agent learning in coordination settings (prior work)[1]

- QLD provably converges in multi-agent weighted potential games
- multiple QRE, but bifurcation phenomena explain equilibrium selection
- equilibrium selection after exploration depends on a game's geometry

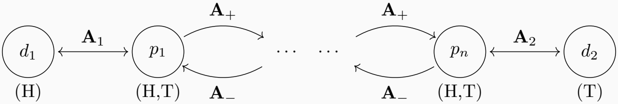




[1] *S. Leonardos, G. Piliouras, Exploration-Exploitation in Multi-Agent Learning: Catastrophe Theory Meets Game Theory*, AAAI-21, Best paper award.

Convergence to QRE

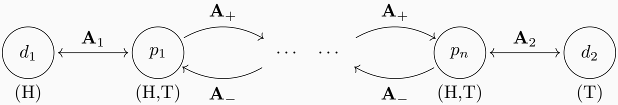
Match-Mismatch Game (MMG): line-network WZPG with

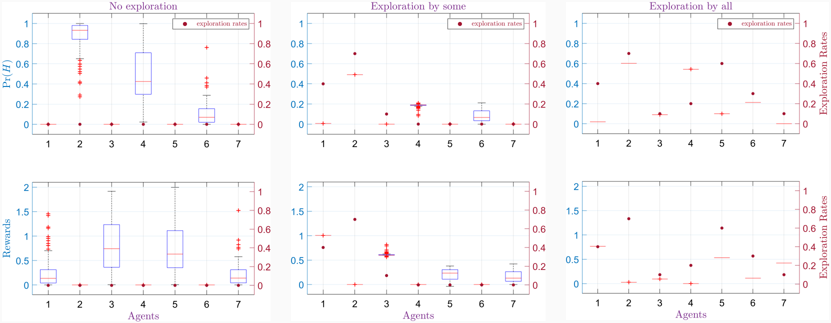


- $\mathbf{A}_+ = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, $\mathbf{A}_- = -\mathbf{A}_+$ and $\mathbf{A}_1 = \mathbf{A}_2 = (1, -1)$
- first and last are dummy agents with fixed actions
- goal: mismatch the previous and match the next agent
- infinite many NE: $(T, H/T, T, H/T, \dots)$

Convergence to QRE

Match-Mismatch Game (MMG): line-network WZPG with





Convergence result is tight

Conclusions

Takeaways: Q-Learning and Quantal Response Equilibria

Multi-agent competition

- despite the diversity of NE, we have uniqueness of QRE
- QLD converges to QRE and solves the equilibrium selection problem

Multi-agent coordination (prior work)

- QLD converges to QRE in multi-agent weighted potential games
- even with multiple QRE, bifurcation phenomena explain equilibrium selection

Next steps

- *Can we go beyond that: mixed games with both cooperation and competition?*
- *Effects of exploration on individual/social welfare* after equilibrium selection?

Thank you