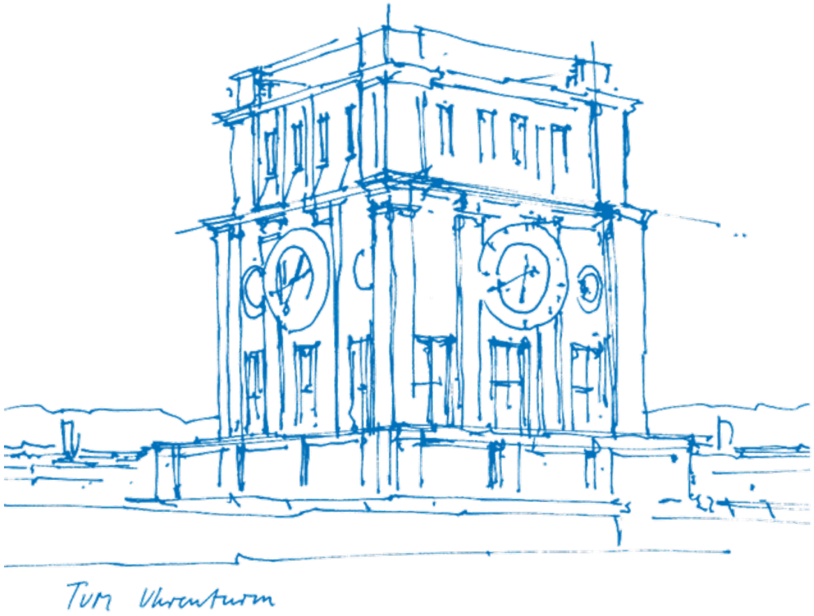
# Deep Reinforcement Learning Finds Bayes–Nash Equilibrium in Competitive Newsvendor Problems

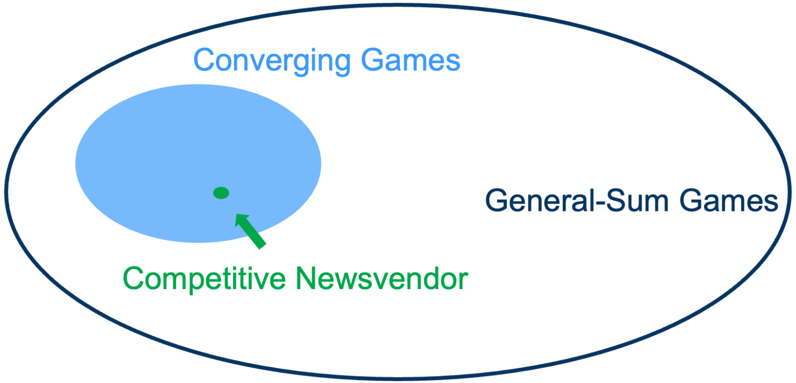
***Kassian Köck*, Fabian R. Pieroth, Martin Bichler**
Technical University of Munich
School of Computation, Information and Technology
Chair of Decision Sciences & Systems
International Conference on Machine Learning 2026



# Learning Equilibria in Games

**Learning in games is notoriously unstable**

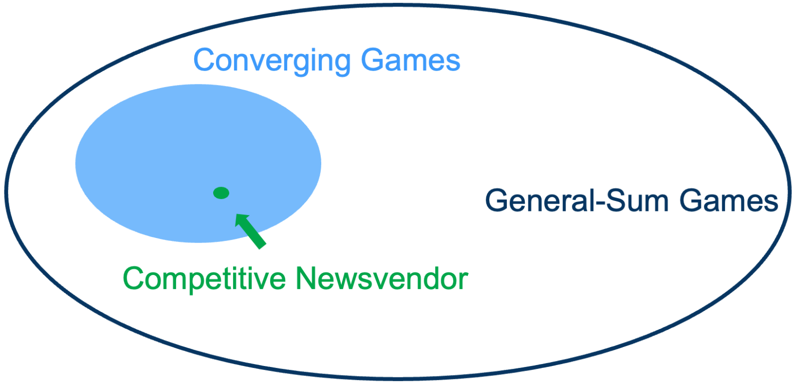
- Independent learners change each other's environment
- Dynamics may converge, cycle, or diverge
- Convergence guarantees are rare



# Learning Equilibria in Games

**Learning in games is notoriously unstable**

- Independent learners change each other's environment
- Dynamics may converge, cycle, or diverge
- Convergence guarantees are rare




**Do DRL agents reliably converge to equilibrium strategies in Competitive Newsvendor games?**

# The Competitive Newsvendor Problem

**Model**

- Two retailers buy $y_i$ items at cost $c_i$

# The Competitive Newsvendor Problem

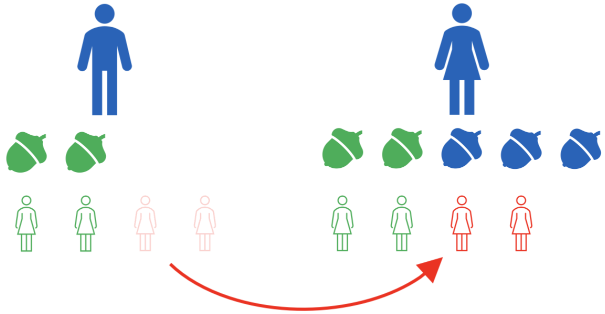
**Model**

- Two retailers buy $y_i$ items at cost $c_i$
- Independent uniform demand $x_i$

# The Competitive Newsvendor Problem

**Model**

- Two retailers buy $y_i$ items at cost $c_i$
- Independent uniform demand $x_i$
- Stockouts generate spillover demand $d_i$



# The Competitive Newsvendor Problem

**Model**

- Two retailers buy $y_i$ items at cost $c_i$
- Independent uniform demand $x_i$
- Stockouts generate spillover demand $d_i$
- Sell price $s$
- Profit for retailer $i$ is

$$\pi_i(\mathbf{y}, c_i, \mathbf{x}) := -c_i y_i + s \cdot \min(y_i, x_i + d_i).$$

- Retailers aim to maximize expected profit

$$J_i(\mathbf{y}, c_i) = \mathbb{E}_{\mathbf{x}}[\pi_i(\mathbf{y}, c_i, \mathbf{x})],$$

- Cost $c_i$ is either common knowledge (complete information) or private (Bayesian)

# Strict Monotonicity

**The payoff gradient of the Competitive Newsvendor game is strictly monotone.**

Vector of marginal profits $\mathbf{l}(\mathbf{y},\mathbf{c}) := \frac{\partial \mathbf{J}(\mathbf{y},\mathbf{c})}{\partial \mathbf{y}}$

For all inventory profiles $\mathbf{y} \neq \mathbf{y}'$,

$$\langle \mathbf{l}(\mathbf{y},\mathbf{c}) - \mathbf{l}(\mathbf{y}',\mathbf{c}), \mathbf{y} - \mathbf{y}' \rangle < 0.$$

# Strict Monotonicity

**The payoff gradient of the Competitive Newsvendor game is strictly monotone.**

Vector of marginal profits $\mathbf{l}(\mathbf{y},\mathbf{c}) := \frac{\partial \mathbf{J}(\mathbf{y},\mathbf{c})}{\partial \mathbf{y}}$

For all inventory profiles $\mathbf{y} \neq \mathbf{y}'$,

$$\langle \mathbf{l}(\mathbf{y},\mathbf{c}) - \mathbf{l}(\mathbf{y}',\mathbf{c}), \mathbf{y} - \mathbf{y}' \rangle < 0.$$

**Consequences**

- Unique Nash equilibrium
- Unique Bayes–Nash equilibrium
- Stable variational-inequality dynamics

# Strict Monotonicity

**The payoff gradient of the Competitive Newsvendor game is strictly monotone.**

Vector of marginal profits $\mathbf{l}(\mathbf{y},\mathbf{c}) := \frac{\partial \mathbf{J}(\mathbf{y},\mathbf{c})}{\partial \mathbf{y}}$

For all inventory profiles $\mathbf{y} \neq \mathbf{y}'$,

$$\langle \mathbf{l}(\mathbf{y},\mathbf{c}) - \mathbf{l}(\mathbf{y}',\mathbf{c}), \mathbf{y} - \mathbf{y}' \rangle < 0.$$

**Consequences**

- Unique Nash equilibrium
- Unique Bayes–Nash equilibrium
- Stable variational-inequality dynamics

**Monotone VI methods**

- Extragradient (Korpelevich)
- Mirror-Prox
- Proximal Point Method
- Optimistic Gradient Methods

$\Longrightarrow$ **(Bayes–)Nash equilibria can be computed via monotone variational-inequality methods**

# Independent Learners

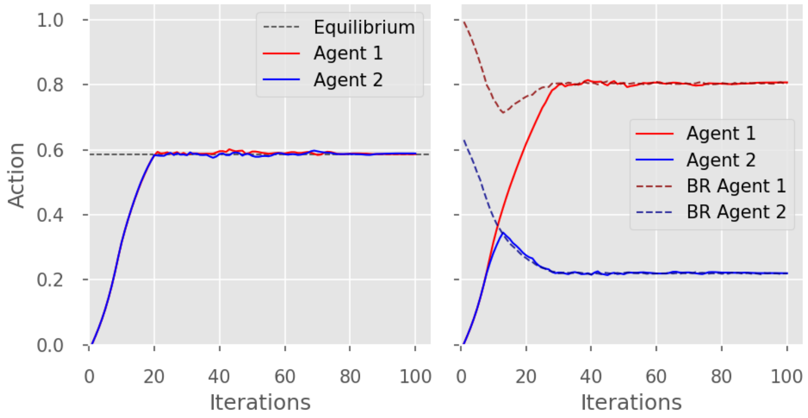
**Does convergence to equilibrium strategies also hold for standard reinforcement learners?**

# Independent Learners

**Does convergence to equilibrium strategies also hold for standard reinforcement learners?**

**Experiment**

- Independent agents in self-play
- Verification via best-response checks
- Quick convergence to equilibrium strategies in all settings



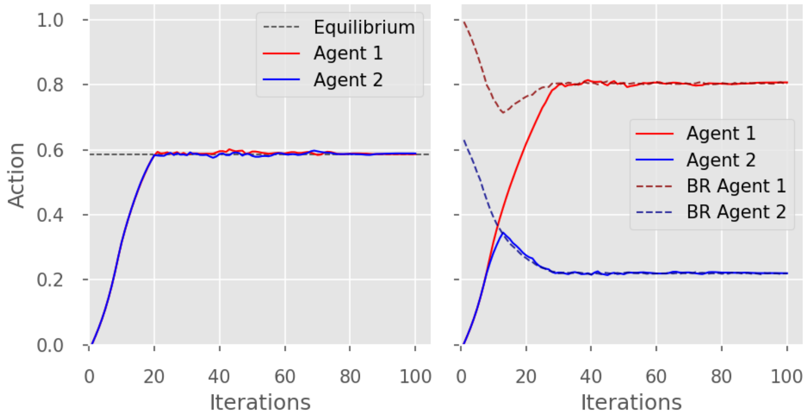

PPO agents converging in both symmetric and asymmetric (here: complete-information) settings

# Independent Learners

**Does convergence to equilibrium strategies also hold for standard reinforcement learners?**

**Experiment**

- Independent agents in self-play
- Verification via best-response checks
- Quick convergence to equilibrium strategies in all settings




PPO agents converging in both symmetric and asymmetric (here: complete-information) settings

**$\Longrightarrow$ Standard reinforcement learners, despite not being monotone VI methods, still find equilibrium strategies efficiently**

# Takeaways

**Do DRL agents reliably converge to equilibrium strategies in Competitive Newsvendor games?**

Theory $\swarrow$

$\searrow$ Observation

**The payoff gradient is strictly monotone**

$\Downarrow$

**Monotone VI methods converge to Nash equilibria**

**Standard reinforcement learners empirically converge to Nash equilibria**