A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning

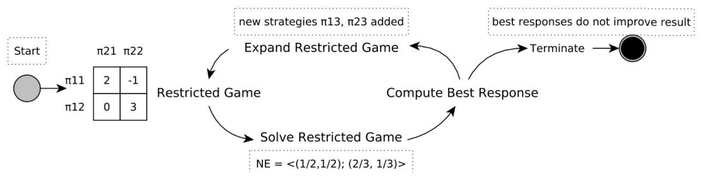
Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, Thore Graepel

DeepMind

Introduction & Motivation

- Renewed interest in multiagent reinforcement learning (MARL)
  - Naive solution: Independent RL (InRL):
  - Every agent treats other agents as part of their local environment
    - **Good**: simple
    - **Bad**: no convergence guarantees
    - **Ugly**: overfitting, lack of generalization
- New metric to quantify the **joint policy correlation (JPC)**
- We present a new algorithm:
  - Based on **empirical game-theoretic analysis** (Wellman '06, Walsh et al. '02)
  - Uses SoTA **deep reinforcement learning**: The Reactor (Gruslys et. al. '17)
  - Generalizes fictitious play, iterated best resp., InRL, double oracle
  - Reduces JPC & finds trade-off between best resp./Nash

Our algorithm is largely inspired by double oracle (McMahan, Gordon, and Blum '03)



Joint Policy Correlation (JPC) in Independent RL

- Start **D** independent training instances → (π_1^d, π_2^d), for $d \in \{1, ..., D\}$
- Produce **JPC Matrix** : $\mathrm{a}_{i,j} = \mathbb{E}_{\rho \sim \pi_1^i, \pi_2^j}[\textsc{Return}(\rho)]$
- Define:
  - **Avg. prop. loss** (from miscoordination): $\mathrm{R}_{-} = \frac{\bar{D} - \bar{O}}{\bar{D}}$

InRL - Laser Tag (small2)

InRL - Laser Tag (small4)

Policy-Space Response Oracles (PSRO)

Player *i* produces approx. best resp. **oracles**: $\Pi_i = (\pi_{i,0}, \pi_{i,1}, \cdots, \pi_{i,K})$

and maintains a **meta-strategy** $\sigma_i = (\sigma_{i,0}, \sigma_{i,1}, \cdots, \sigma_{i,K}) \in \Delta(\Pi)$

All sets of oracles together induce an **empirical game** (Π^K, U^{Π^k})

Deep Cognitive Hierarchies (DCH)

- Parallel PSRO
- Choose K *levels*
- Start N K processes (dist. oracles)
- Async. save + refresh $\pi_{i,k}, \sigma_{i,k}$
- Similar to Camerer et al.
- Space $O(N^2K^2) < O(K^N)$

Meta-Solvers

Main idea: "solve" the empirical game using learning from expert advice (see Cesa-Bianchi & Lugosi '06)

Limited feedback? No problem: use bandit versions!

- Hedge / EXP3 (Auer et. al, 1995)
- Regret Matching (Hart & Mas-Colell '00, '01)
- Projected Replicator Dynamics New!

$$\mathrm{U}^{\Pi} = (\mathbf{A}, \mathbf{B}) \quad \frac{dx_k}{dt} = x_k[(\mathbf{Ay})_k - \mathbf{x}^T\mathbf{Ay}], \quad \frac{dy_k}{dt} = y_k[(\mathbf{x}^T\mathbf{B})_k - \mathbf{x}^T\mathbf{By}],$$
$$\sigma = (\mathbf{x}, \mathbf{y}) \quad \mathrm{x} \leftarrow P(\mathbf{x} + \delta\tfrac{d\mathbf{x}}{dt}) \quad \mathrm{y} \leftarrow P(\mathbf{y} + \delta\tfrac{d\mathbf{y}}{dt})$$

Force **exploratory** meta-strategies

$$\sigma_{i,k} \geq \frac{\gamma}{K+1}$$

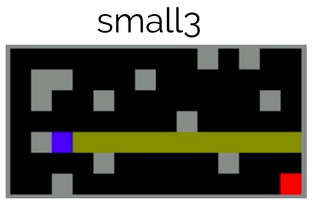
Algorithm	Meta-Strategy
Fictitious Play	$\sigma_i = \left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}, 0\right)$
Iterated Best Response	$\sigma_i = (0, 0, \cdots, 1, 0)$
Independent RL	$\sigma_i = (0, 0, \cdots, 0, 1)$
Double Oracle	$\sigma_i \in \mathrm{NashEq}(U^{\Pi})$

Results

First-Person Gridworld Games

small2

small3



Game	\bar{D}	\bar{O}	R- (InRL)	R- (DCH)	Delta
LT small2	30.4	20.1	34.2 %	5.5 %	- 28.7 %
LT small3	23.1	9.1	62.5 %	8.2 %	- 54.3 %
LT small4	20.2	5.7	71.7 %	15.0 %	- 56.7 %
Gathering	147.3	146.8	-		
Pathfind	108.7	106.3	-		

Laser Tag:

- Moves: Turn left, right, fwd, back, strafe left, strafe right, timeout beam
- +1 for a tag
- Agents have local views
- Agents re-enter at spawn points

JPC reduction using DCH @ 10 levels ($\sigma_{i,10}$)

- No meta-strategy ($\pi_{i,10}$)? **14.7%, 27%, 11.8%**
- Five/three levels ($\sigma_{i,5}, \sigma_{i,3}$) small4: **15.6, 24.6**

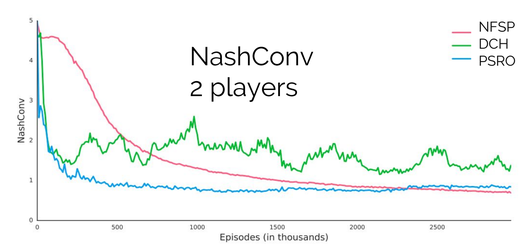
Leduc Poker

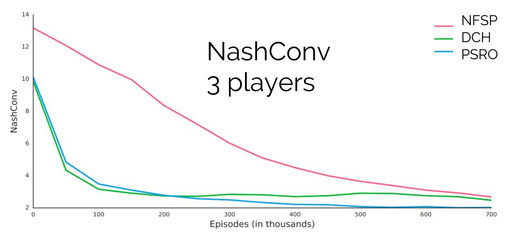
- 1 chip ante: 2 rounds
- 6-card deck, 2 suits, 1 priv card
- Limit raises: 2 per rnd (2,4)
- Public card after rnd 2

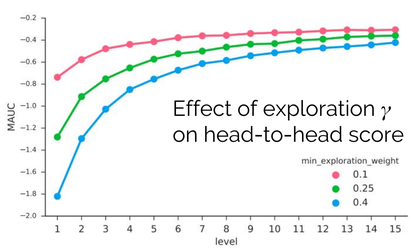
Compare to:

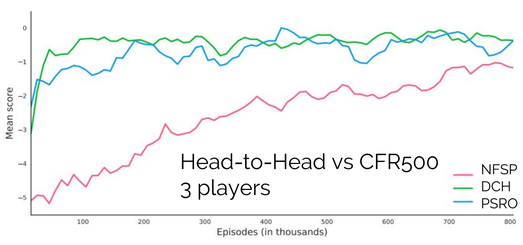
- NFSP (Heinrich & Silver '16)
- CFR (Zinkevich et. al. '07)

$$\textsc{NashConv}(\sigma) = \sum_i^n \max_{\sigma_i'} u_i(\sigma_i', \sigma_{-i}) - u_i(\sigma)$$









Effect of levels/hierarchy on NashConv