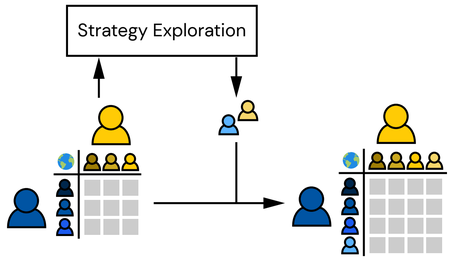
Iterative Empirical Game Solving via Single Policy Best Response

Max Olan Smith, Thomas Anthony, Michael P. Wellman

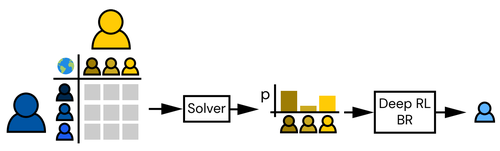
On each epoch, train against a single opponent policy rather than a distribution; reducing variance and focusing training on salient strategic knowledge.

Motivation

Policy-Space Response Oracles (PSRO) [1] is a general framework for learning policies in multiagent systems. It works by building an empirical game, a simulated model of a game, through iterations of empirical game analysis and deep reinforcement learning (Deep RL).



This work investigates reducing the cumulative training time incurred by repeated applications of Deep RL.



Strategy Exploration Problem: how to choose which policies to add to the empirical game?

PSRO's answer
1. Solve empirical game
2. Best-Response via RL in full game

Three issues:
1. Opponents secretly sample policies at beginning of episode. Results in high variance in state-outcomes for the learner.
2. Failing to exploit previous training with opponent policies that have already been encountered .
3. The best response to empirical game's solution may not be the most useful new strategy.

Core Idea:
- Replace mixed-strategy opponent with a single opponent policy.
- Reduce variance induced by unobserved mixture of opponents.
- Select opponent policy that represents more salient strategic knowledge.

Contact: mxsmith@umich.edu

[1] Lanctot, et al.. A unified game-theoretic approach to multiagent reinforcement learning. NeurIPS'17.

Work at the University of Michigan was supported in part by MURI grant W911NF-13-1-0421 from the US Army Research Office.

Mixed-Oracles

- During each epoch of PSRO only a single policy is added to each player's strategy set.

- RL must relearn against old opponent policies.

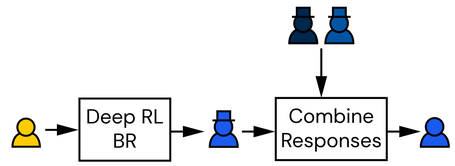
Epoch 2:

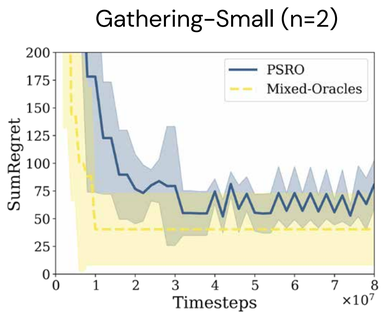
Epoch 3:

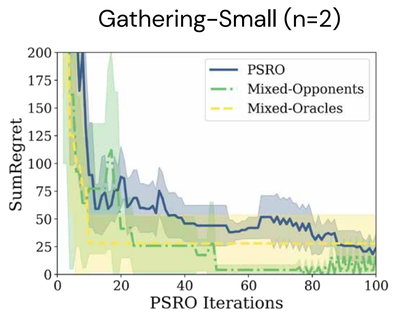
- **Idea:**
  - Train best-response to new opponent policy.
  - Transfer respones from older opponent policies.

- Use Q-Mixing to combine responses:

$$Q_{\pi_i}(o_i, a_i \mid \sigma_{-i}) = \sum_{\pi_{-i}} \psi_i(\pi_{-i} \mid o_i, \sigma_{-i}) \cdot Q_{\pi_i}(o_i, a_i \mid \pi_{-i})$$







Mixed-Opponents

- Responding to the empirical game's solution may not represent the most salient training objective.

- For example, this may generate a policy very similar to an existing policy.

- Ideally, objective enables more efficient search of the game's strategy space.

- **Idea:**
  - Combine opponent policies by averaging their Q-values.
  - Considers the value of all actions.
  - Constructs unique greedy opponent policy.