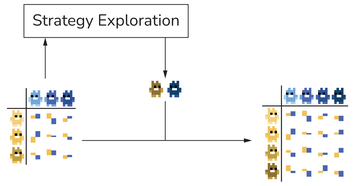
Strategic Knowledge Transfer

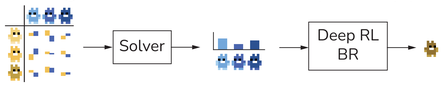
Max Olan Smith, Thomas Anthony, Michael P. Wellman
Journal of Machine Learning Research, 24(233), 2023.

Context

- Games can be solved through empirical game models by proxy.
- Models are built by alternating game-reasoning and strategy exploration.

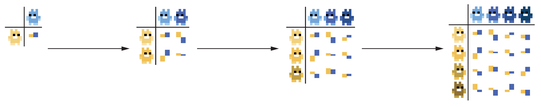


- **Strategy Exploration:** compute best-responses to current solution.
- Complex games make computing best-responses intractable.
- Reinforcement learning can be used to compute approximate responses.



This Work

- **Problem:** reinforcement learned policies require lots of data to train.
- Data can be expensive to collect (e.g., human interactions / simulation).
- Experiential Cost: amount of data used to train policies.
- Game modelling computes a series of policies, each with high cost.



- **Solution:** strategy exploration methods that utilize transfer learning.

Reduce data costs of game learning by transferring value functions from previously learned policies.

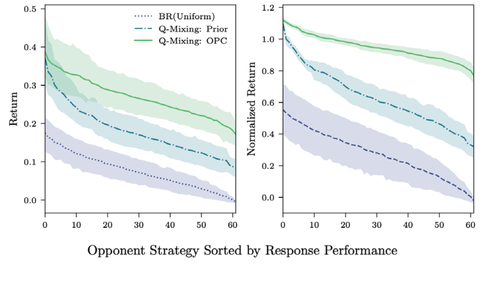
Q-Mixing

- **Idea:** average Q-values following belief over opponents.

$$\text{BR Mixed Strategy} = \sum_{\pi_{-i}} \text{Opponent Likelihood} \cdot \text{BR Pure Strategy}$$

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

Running With Scissors



Prior: use opponent distribution as likelihood.

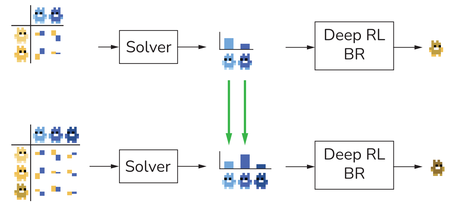
$$\psi_i(\pi_{-i} \mid \quad \sigma_{-i})$$

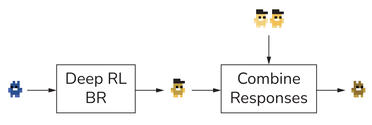
OPC: use evidence during play to inform likelihood.

$$\psi_i(\pi_{-i} \mid o_i, \sigma_{-i})$$

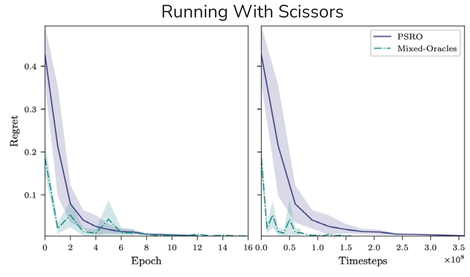
Mixed-Oracles

- **Insight:** each player adds one new policy.
- **Idea:** learn best-response to new policy, and transfer the rest.

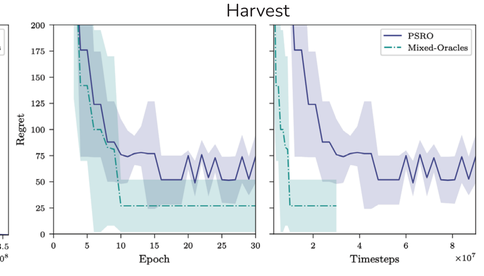




Running With Scissors

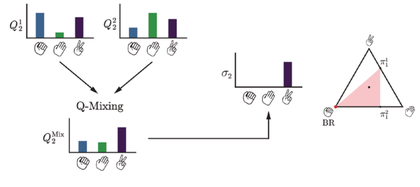


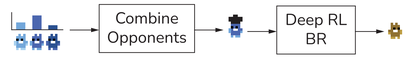
Harvest



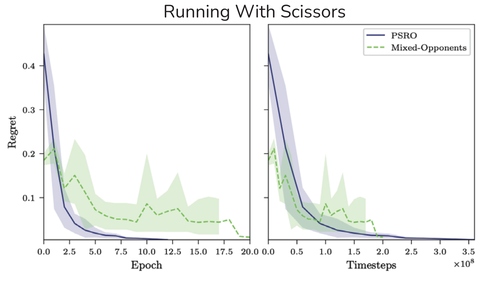
Mixed-Opponents

- **Insight:** aggregate opponent policies using Q-Mixing.
- Discovery strategically important policies faster, train less policies overall.
- Consider each policy's value for all actions, giving rise to unique policy.

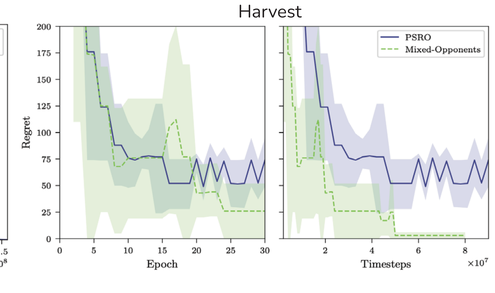




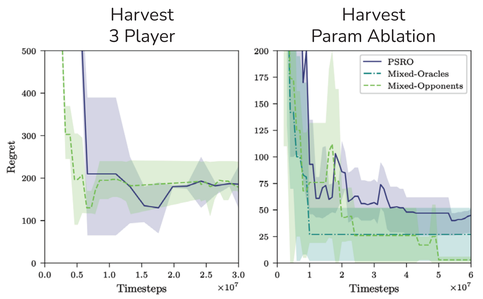
Running With Scissors



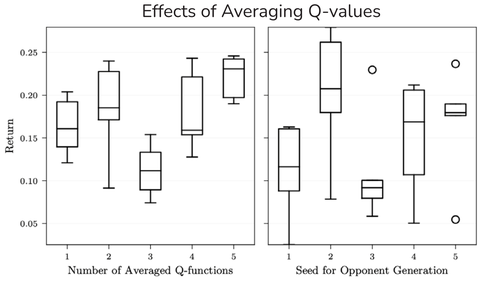
Harvest



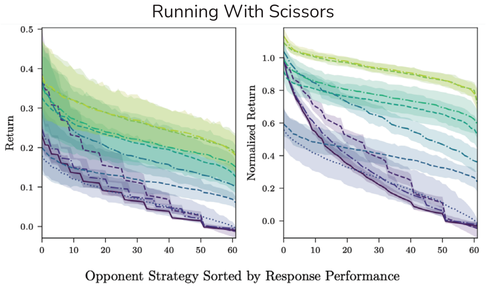
Extras



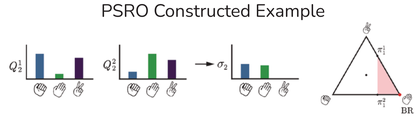
Effects of Averaging Q-values



Running With Scissors



PSRO Constructed Example



Opponent-Policy Identification Game