

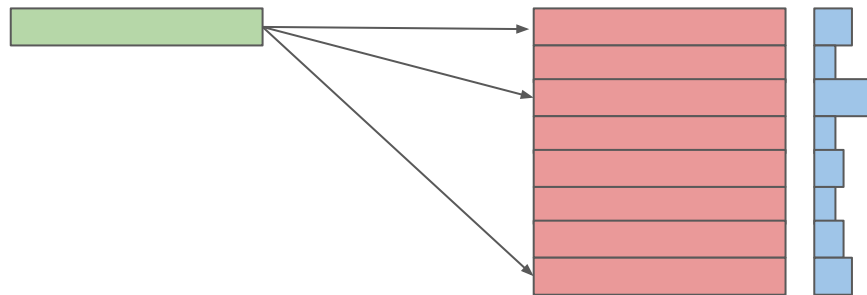
# Toward Optimal and Scalable Double Oracle Algorithms

Stephen McAleer, CMU

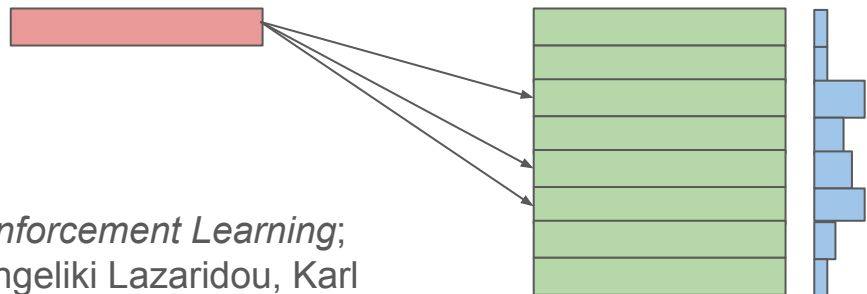
# Policy Space Response Oracles (PSRO)

- Both players learn best response to opponent's meta-Nash
- Meta-Nash converges to a Nash equilibrium

Player 1 Best Responses to Player 2's Meta Nash



Player 2 Best Responses to Player 1's Meta Nash



*A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning;*  
Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl  
Tuyls, Julien Pérolat, David Silver, Thore Graepel. NIPS 2017.

# PSRO Pros and Cons

- **Pros**

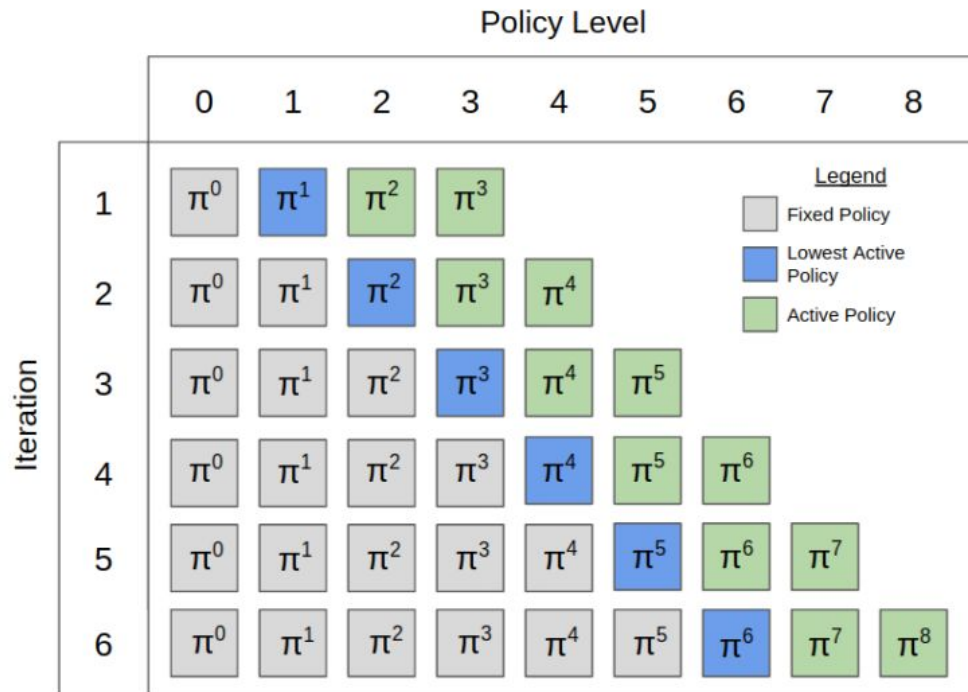
- Can converge faster than NFSP, Deep CFR in certain games
- Easy to use with any existing RL algorithm
- Can handle continuous actions in practice
- Has been used to achieve expert-level performance at Starcraft

- **Cons**

- Sequential algorithm, requires training a new best response every iteration
- Convergence guarantees on normal form of the game, exponential in # of infostates
- Exploitability can increase from one iteration to the next
- Strategies added every iteration are not optimal

# Pipeline PSRO

- Fixed and active policies
- Each active policy plays against meta-Nash of policies below it
- Once lowest active policy plateaus, it becomes fixed and a new policy is added
- Inherits same convergence guarantees as PSRO



*Pipeline PSRO: A Scalable Approach for Finding Approximate Nash Equilibria in Large Games*; Stephen McAleer\*, John Lanier\*, Pierre Baldi, Roy Fox. NeurIPS 2020.

# Pipeline PSRO Results: Barrage Stratego

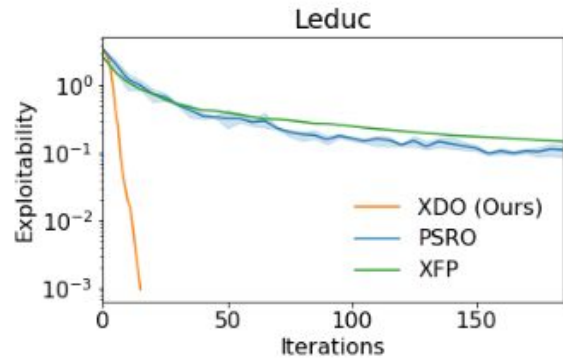
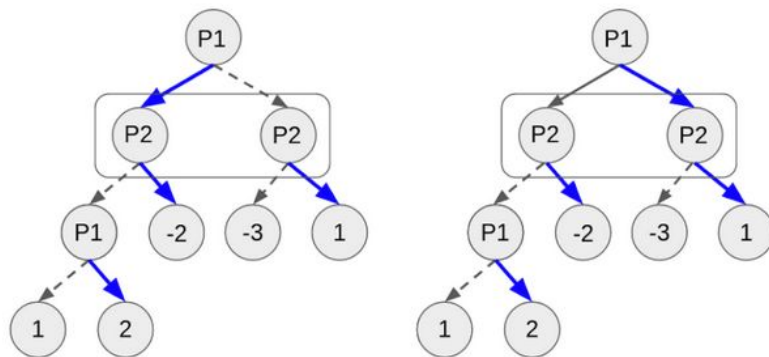
|   |   |  |  |   |   |   |    |    |   |
|---|---|--|--|---|---|---|----|----|---|
|   |   |  |  |   |   |   |    |    |   |
|   |   |  |  |   |   |   | F  |    |   |
|   |   |  |  | 9 |   | 2 | 10 |    |   |
| 2 | 3 |  |  |   | B |   |    | S  |   |
|   |   |  |  |   |   |   |    |    |   |
|   |   |  |  |   |   |   |    |    |   |
|   | S |  |  |   | 2 |   | 3  |    |   |
|   |   |  |  | 2 |   |   |    | 10 |   |
|   |   |  |  |   |   |   | 9  |    |   |
|   |   |  |  |   |   |   |    | B  | F |

Table 1: P2SRO Results vs. Existing Bots

| Name                    | P2SRO Win Rate vs. Bot |
|-------------------------|------------------------|
| Asmodeus                | 81%                    |
| Celsius                 | 70%                    |
| Vixen                   | 69%                    |
| Celsius1.1              | 65%                    |
| <b>All Bots Average</b> | <b>71%</b>             |

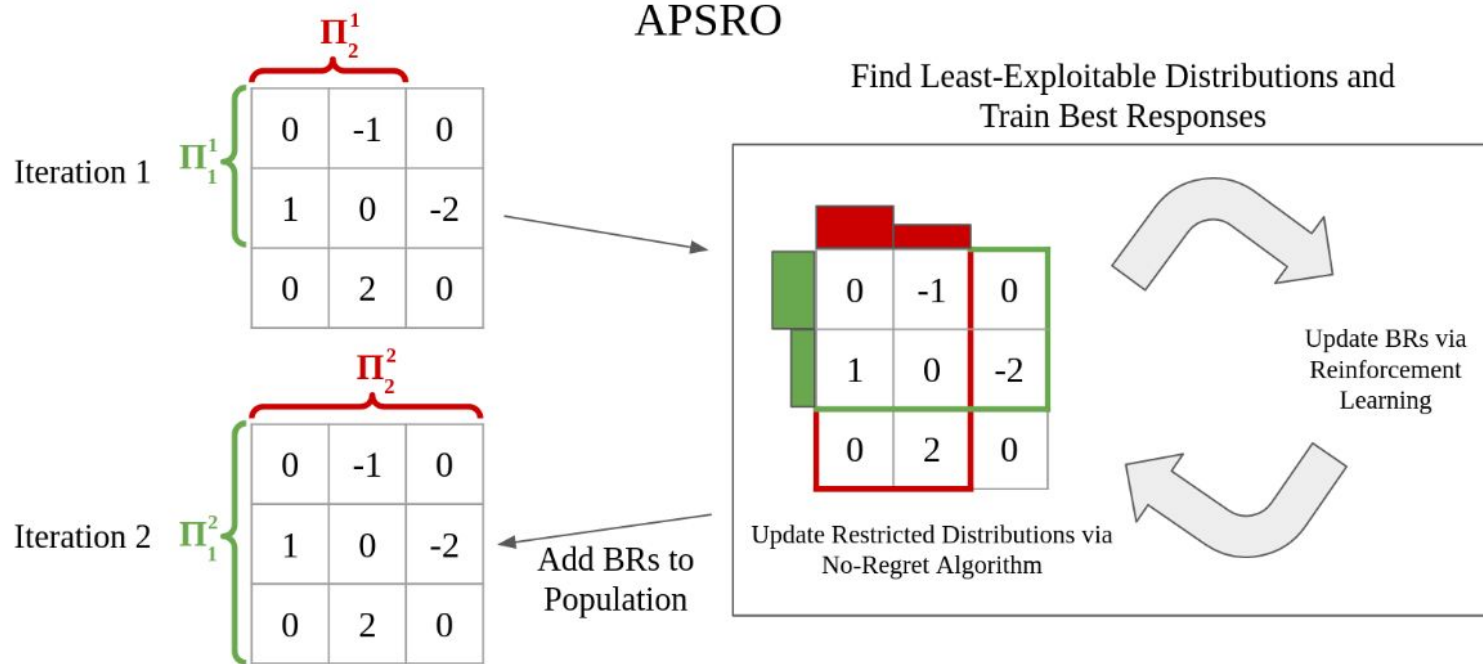
# (N)XDO

- Same as PSRO, but meta-Nash is computed in extensive form of the game
- Restricted game is created by restricting the actions to be choosing a best response from the population
- This restricted game is solved to get meta-Nash
- Linear convergence instead of exponential



*XDO: A Double Oracle Algorithm for Extensive-Form Games*; Stephen McAleer, John Lanier, Kevin Wang, Pierre Baldi, Roy Fox. NeurIPS 2021.

# Anytime PSRO (APSRO)



*Anytime PSRO for Two-Player Zero-Sum Games*; Stephen McAleer, Kevin Wang, Marc Lanctot, John Lanier, Pierre Baldi, Roy Fox. AAI RLG Workshop 2022.

# Self-Play PSRO

|    |    |    |    |    |    |
|----|----|----|----|----|----|
| 0  | -1 | 0  | 0  | 0  | 1  |
| 1  | 0  | -1 | 0  | 0  | 0  |
| 0  | 1  | 0  | -1 | 0  | 0  |
| 0  | 0  | 1  | 0  | -1 | 0  |
| 0  | 0  | 0  | 1  | 0  | -1 |
| -1 | 0  | 0  | 0  | 1  | 0  |

Inner Iteration 1

|    |    |    |    |    |    |
|----|----|----|----|----|----|
| 0  | -1 | 0  | 0  | 0  | 1  |
| 1  | 0  | -1 | 0  | 0  | 0  |
| 0  | 1  | 0  | -1 | 0  | 0  |
| 0  | 0  | 1  | 0  | -1 | 0  |
| 0  | 0  | 0  | 1  | 0  | -1 |
| -1 | 0  | 0  | 0  | 1  | 0  |

Inner Iteration 10

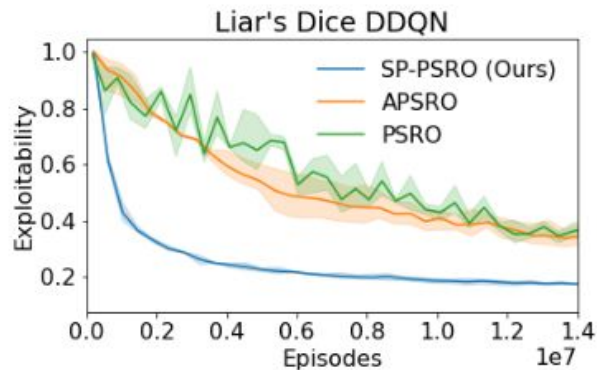
|    |    |    |    |    |    |
|----|----|----|----|----|----|
| 0  | -1 | 0  | 0  | 0  | 1  |
| 1  | 0  | -1 | 0  | 0  | 0  |
| 0  | 1  | 0  | -1 | 0  | 0  |
| 0  | 0  | 1  | 0  | -1 | 0  |
| 0  | 0  | 0  | 1  | 0  | -1 |
| -1 | 0  | 0  | 0  | 1  | 0  |

Time-Averaged New Strategy

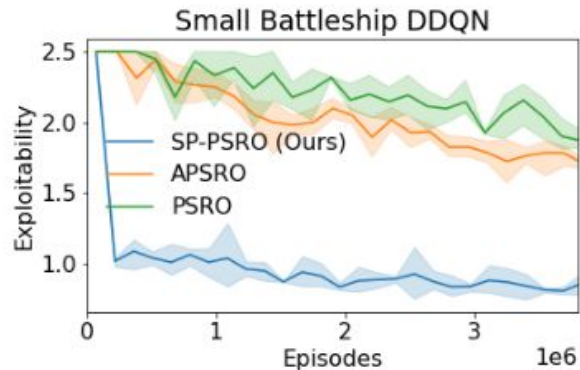
 Fixed Strategies

 New Strategy

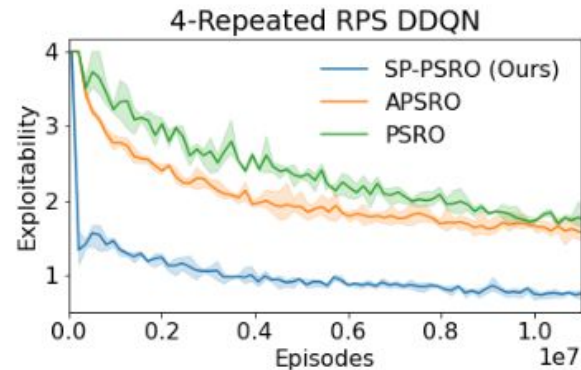
 Opponent Best Response



(a) DRL Liars Dice



(b) DRL Battleship



(c) DRL Repeated RPS

Figure 5: Extensive-form games with DDQN best responses