DeepMind

Multi-Agent Training beyond Zero-Sum with Correlated Equilibrium Meta-Solvers

ICML 2021

Presenter: Luke Marris

Authors

Luke Marris

Paul Muller

Marc Lanctot

Karl Tuyls

Thore Graepel

Current State of the Art

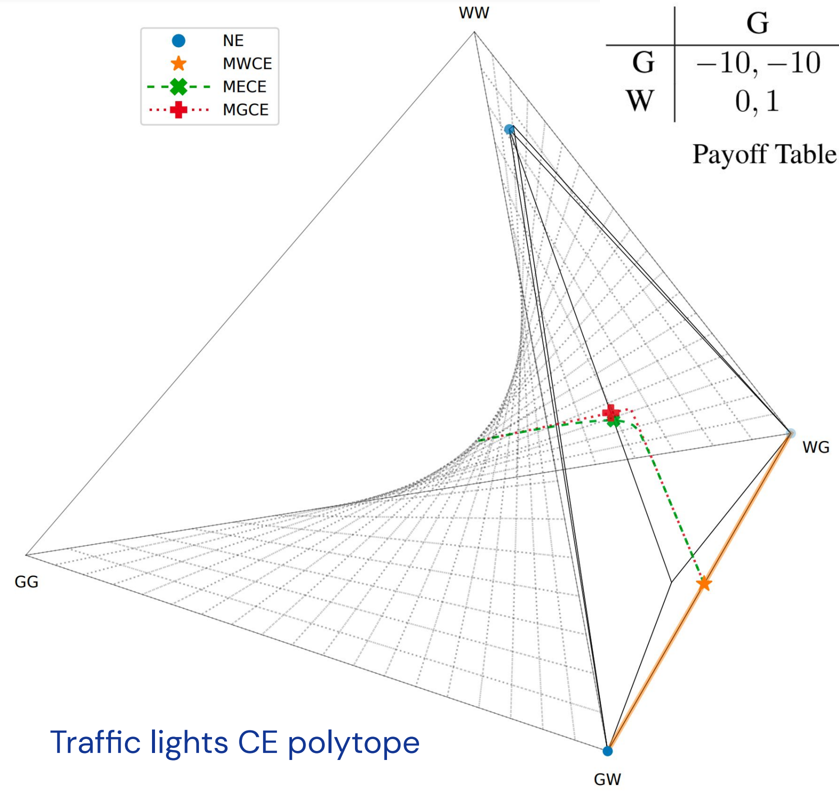
- **Current Progress**: The field has become increasingly competent at solving **two-player**, **zero-sum** games (Go, Chess, StarCraft).
- **Zero-Sum Definition**: Purely competitive class where one player's gain is another player's loss.
- **Properties**: Zero-sum games are easier to solve because there is a principled target objective; the set of Nash equilibrium policies, which are interchangeable and tractable to compute for this class.
- **Real World**: The real world has many games which have more than two players and are not purely competitive ("n-player, general-sum").
- **Previous Work**: Some work on n-player, general-sum (Capture the Flag, Dota) has been impressive but falls short of convincingly solving these games.
- **Blocker**: Progress beyond two-player, zero-sum has been stymied by a a) lack of game theoretic learning algorithms suitable for this setting and b) uncertainty on a suitable solution concept.

Contributions of this work

- **Solution Concept**: We argue that (normal form) correlated equilibria (CEs) and coarse correlated equilibria (CCEs) are suitable target objectives in n-player, general-sum games.
- **Equilibrium Selection**: We suggest a new tractable method of picking between several equilibria ("equilibrium selection problem"): maximum Gini (C)CE (MG(C)CE).
- **Learning Algorithms**: We provide two new algorithms based on Policy Space Response Oracles (PSRO) for training agents in n-player, general-sum games, called JPSRO(CE) and JPSRO(CCE).
- **Convergence**: We mathematically prove that JPSRO(CE) converges to a CE, and JPSRO(CCE) converges to a CCE.
- **Empirical**: We empirically check that this algorithm converges to maximum welfare (C)CE solutions in 3-player Kuhn poker (purely competitive), Trade Comm (purely cooperative), and Sheriff (a mixed cooperative and competitive game). We provide code with this work.
- **Meta-Solver Study**: We evaluate a number of meta-solvers for JPSRO and discuss their strengths and weaknesses.

Why (Coarse) Correlated Equilibrium?

- **Tractable**: Is tractable to compute in n-player, general-sum settings.
- **Convex**: Has a convex polytope of solutions.
- **Coordination**: Allows players coordinate strategies (essential in cooperative games).
- **High Value Joint Policy**: Results in higher value solutions than Nash equilibrium.
- **Principled**: Is a principled, well studied game theoretic solution.



	G	W
G	$-10, -10$	$1, 0$
W	$0, 1$	$0, 0$

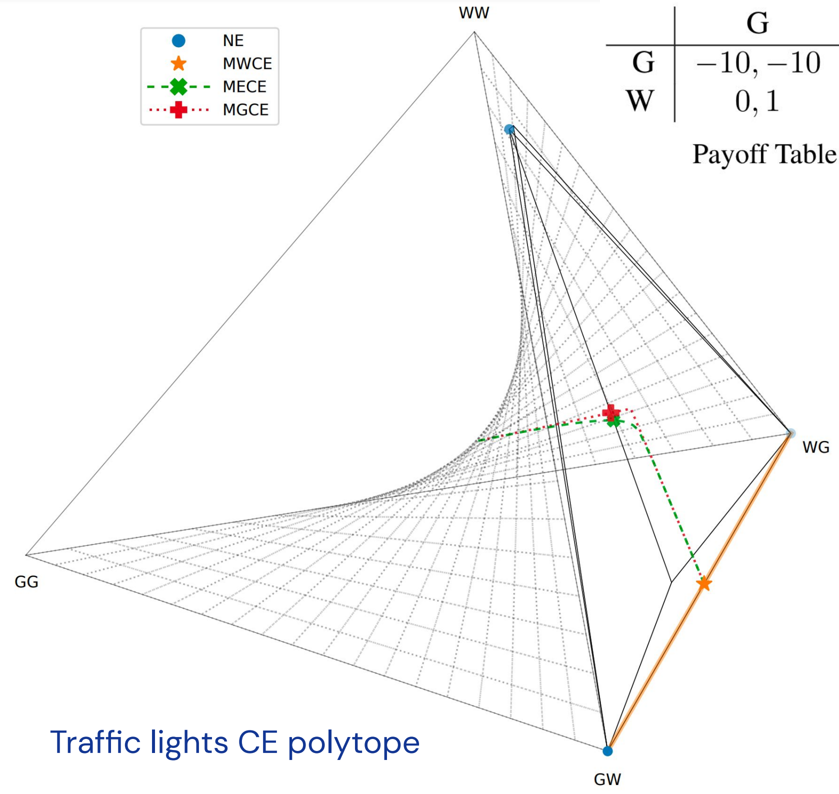
Payoff Table

Traffic lights CE polytope

Maximum Gini (Coarse) Correlated Equilibrium MG(C)CE

Solving equilibrium selection:

- **Objective**: Maximizes the Gini Impurity ($\Sigma\, 1 - \sigma^2$), a quantity closely related to Shannon's entropy.
- **Known Problem Class**: Is a quadratic program so can be computed with many off-the-shelf solvers.
- **Properties**:
  - Scales well when solutions are full-support distributions.
  - Is invariant under affine transforms of the payoff tensor.
  - Can be parameterized by ϵ to produce a family of distributions.



Traffic lights CE polytope

	G	W
G	$-10, -10$	$1, 0$
W	$0, 1$	$0, 0$

Payoff Table

JPSRO - An n-player, general-sum training algorithm

A straightforward extension to PSRO.

- Instead of using factorized distributions (PSRO), JPSRO uses full joint distributions.
- (C)CE meta-solvers (MS) can be used to find a joint distribution.
- Custom best response (BR) operators either converge to a CE or CCE.
- Convergence is achieved when there is no gap (Δ) under the meta-solver distribution.

The output is a joint probability distribution (σ) over set joint policies ($\boldsymbol{\Pi}$).

Algorithm 2 JPSRO

$\Pi_1^0, \ldots, \Pi_n^0 \leftarrow \{\pi_1^0\}, \ldots, \{\pi_n^0\}$

$G^0 \leftarrow \mathrm{ER}(\Pi^0)$

$\sigma^0 \leftarrow \mathrm{MS}(G^0)$

for $t \leftarrow \{1, \ldots\}$ **do**

  for $p \leftarrow \{1, \ldots, n\}$ **do**

    $\{{}^1\pi_p^t, \ldots\}, \{{}^1\Delta_p^t, \ldots\} \leftarrow \mathrm{BR}_p(\Pi^{0:t-1}, \sigma^{t-1})$

    $\Pi_p^{0:t} \leftarrow \Pi_p^{0:t-1} \cup \{{}^1\pi_p^t, \ldots\}$

  $G^{0:t} \leftarrow \mathrm{ER}(\Pi^{0:t})$

  $\sigma^t \leftarrow \mathrm{MS}(G^{0:t})$

  if $\sum_{p,c} {}^c\Delta_p^t = 0$ **then**

    break

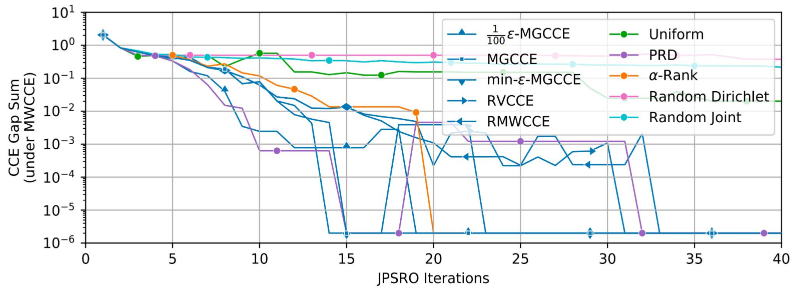
return $\Pi^{0:t}, \sigma^t$

JPSRO(CCE) Empirical Results

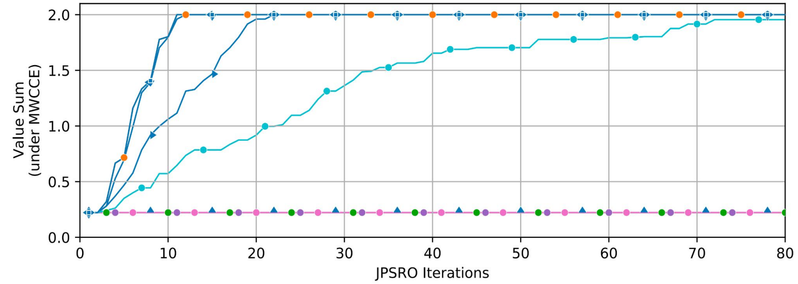
We verified our algorithm on a variety of games

1. Converges to within numerical precision to (coarse) correlated equilibria.
2. Tends to find high value equilibria (usually the maximum welfare equilibria).
3. Verified that classic meta-solvers either do not perform as well or make no progress at all

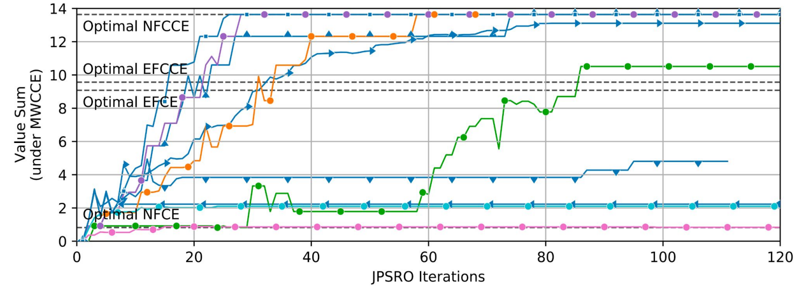
Kuhn Poker: 3-Player Pure Competition



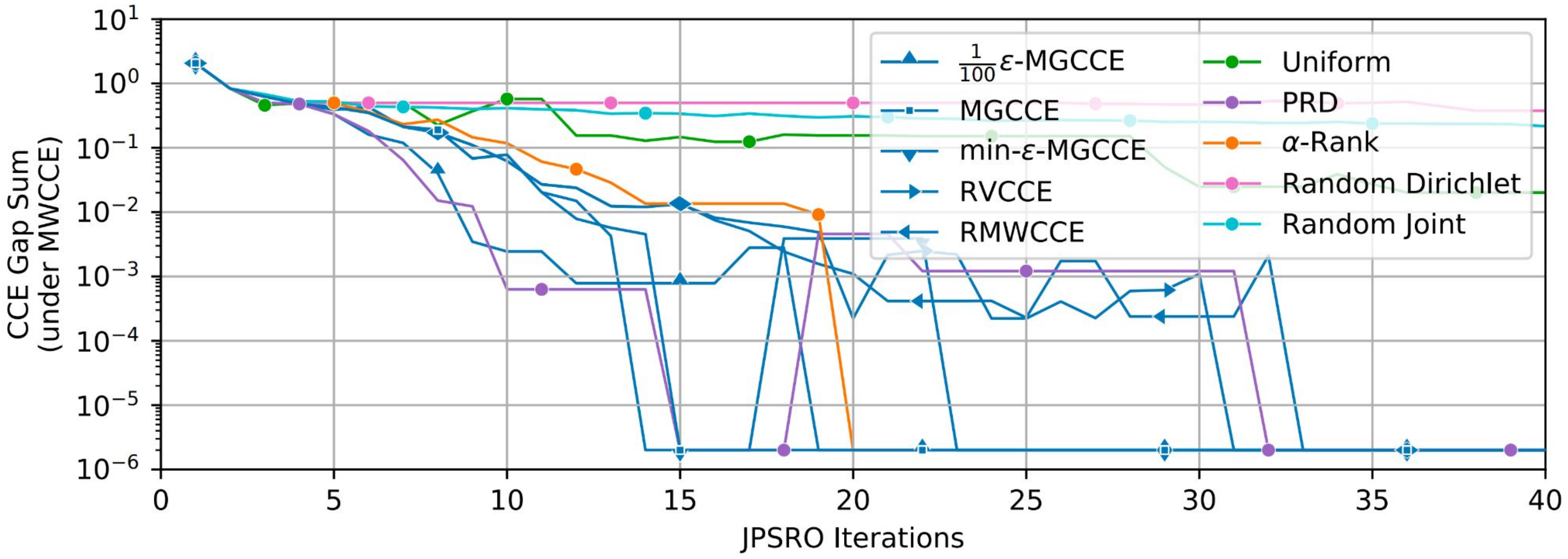
Trade Comm: Pure cooperation



Sheriff: Mixed cooperation and competition

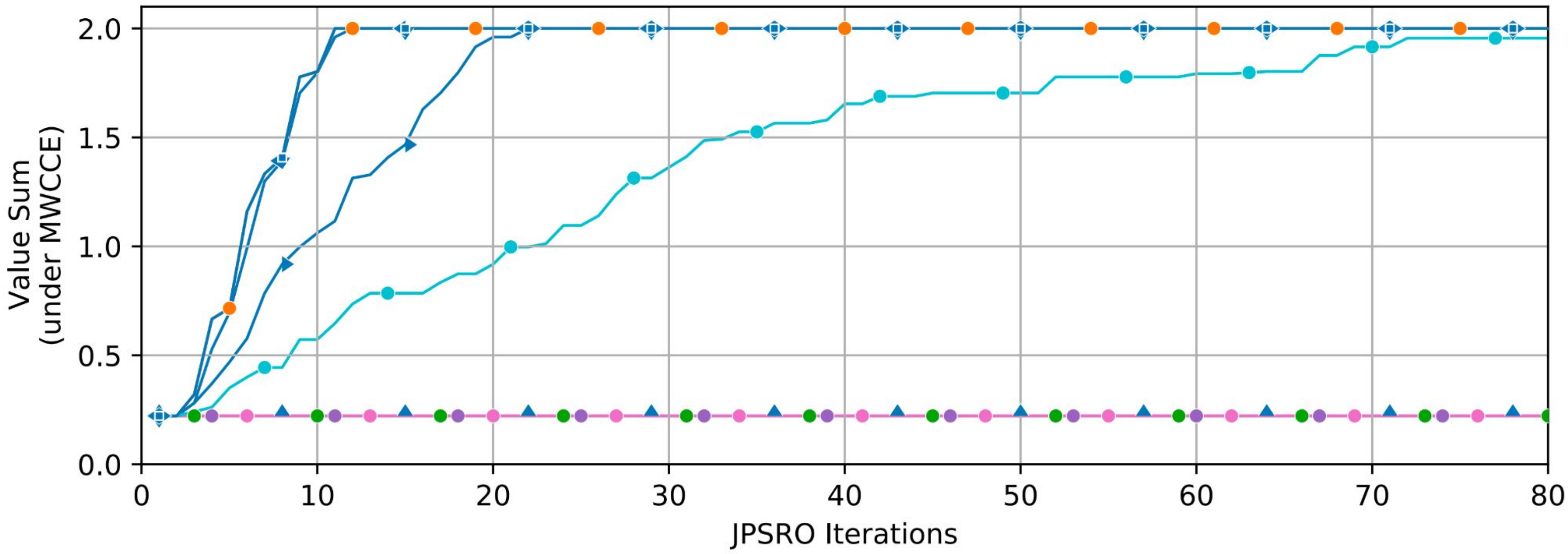


JPSRO(CCE) Empirical Results
3-Player Kuhn Poker



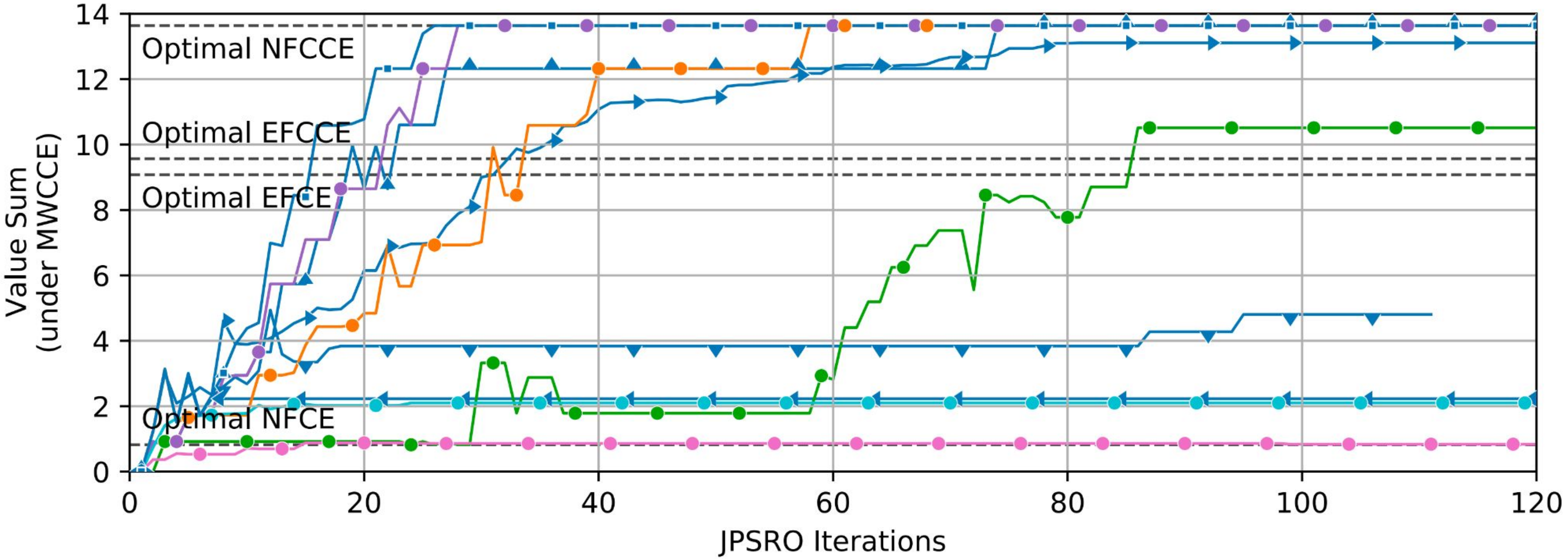
CCE meta-solvers converge to within numerical precision of a CCE.

JPSRO(CCE) Empirical Results
3-Item Trade Comm



CCE meta-solvers rapidly converge to maximum welfare solution. Other meta-solvers flatline.

JPSRO(CCE) Empirical Results
Sheriff



CCE meta-solvers rapidly converge to maximum welfare solution. Other meta-solvers struggle.

Limitations and Future Work

- **RL and Function Approximation**: We believe it is easy to modify JPSRO to use RL for the best response operator, enabling more complex games to be tackled.
- **Scaling**: Although this work proves theoretically a way to converge to normal form (C)CEs for any n-player, general-sum game, there are still significant challenges in scaling to large number of players, mainly due to large payoff tensors.
- **Centralized**: JPSRO is (in part) a centralized training algorithm. Further work to enable fully decentralized training would be beneficial.

Thank you for listening!
See you at the poster!