Neural Population Learning beyond Symmetric Zero-sum Games

Siqi Liu
Google DeepMind and
University College London
London, United Kingdom
liusiqi@google.com

Luke Marris
Google DeepMind and
University College London
London, United Kingdom
marris@google.com

Marc Lanctot
Google DeepMind
Montréal, Canada
lanctot@google.com

Georgios Piliouras
Google DeepMind
London, United Kingdom
gpil@google.com

Joel Z. Leibo
Google DeepMind
London, United Kingdom
jzl@google.com

Nicolas Heess
Google DeepMind
London, United Kingdom
heess@google.com

ABSTRACT

We study computationally efficient methods for finding equilibria in n-player general-sum games, specifically ones that afford complex visuomotor skills. We show how existing methods would struggle in this setting, either computationally or in theory. We then introduce NeuPL-JPSRO, a neural population learning algorithm that benefits from transfer learning of skills and converges to a Coarse Correlated Equilibrium (CCE) of the game. We show empirical convergence in a suite of OpenSpiel games, validated rigorously by exact game solvers. We then deploy NeuPL-JPSRO to complex domains, where our approach enables adaptive coordination in a MuJoCo control domain and skill transfer in capture-the-flag. Our work shows that equilibrium convergent population learning can be implemented at scale and in generality, paving the way towards solving real-world games between heterogeneous players with mixed motives.

KEYWORDS

Game Theory; Deep Learning; Multiagent Reinforcement Learning; Coarse Correlated Equilibrium

ACM Reference Format:
Siqi Liu, Luke Marris, Marc Lanctot, Georgios Piliouras, Joel Z. Leibo, and Nicolas Heess. 2024. Neural Population Learning beyond Symmetric Zero-sum Games. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

1 INTRODUCTION

Purely competitive, symmetric zero-sum games have proven to be popular testbeds for AI research since its early days [5, 6, 28, 31, 34, 35, 38, 39]. Principled algorithms have been developed in this setting, with convergence guarantees to a Nash Equilibrium (NE, [26]) where players can be expected to win (or draw) against *any* opponent. One family of equilibrium convergent methods follows from Fictitious Play (FP, [4]) and Double Oracle (DO, [23]). By learning a set of strategies each best-responding to a mixture over their predecessors, FP and DO converge to an NE even in cyclic games (e.g. *rock-paper-scissors*) where self-play would have made no progress. Policy-Space Response Oracle (PSRO, [14]) extended similar guarantees to extensive-form (EF, [12]) games by constructing a normal-form (NF) metagame whose actions correspond to playing a deep reinforcement learning (RL) policy for an entire episode. Variations of this idea have led to competitive agents in games as complex as StarCraft [39], albeit at significant costs training and evaluating hundreds of independent deep RL agents.

While significant advances have been made in finding Nash Equilibria in symmetric zero-sum games, real-world interactions are often n-player general-sum — between heterogeneous actors with mixed motives. Progress here has been more limited for a few reasons. Computationally, finding exact NE is intractable beyond two-player zero-sum games (i.e. PPAD-complete [9]). More importantly, NE describe an impoverished view of general-sum interactions as it forbids correlated action choices between players. This limitation is subtle but critical: consider a road junction, an NE can only suggest uncorrelated action choices for each driver when much improved outcomes could have been achieved by coordinating drivers with a trusted third-party (e.g. a traffic light). Similar general-sum interactions occur frequently in our society. Fair, mutually beneficial social norms often enable coordination and improve outcomes for all parties.

This observation motivated (Coarse) Correlated Equilibria ((C)CE, [1, 24]), an equilibrium solution concept that allows for coordinated actions between players, mediated by a correlation device that *rational*, *self-interested* players would find beneficial to follow (go on “green”, wait on “red”). (C)CE generalise NE, as they naturally reduce to NE if players are in pure competition and have no way to usefully coordinate. Unlike NE, (C)CE are computationally tractable too, as they can be formulated as a linear program (LP) even in the n-player general-sum setting. Algorithms that offer convergence to (C)CE have also received increased interest in recent years. Following similar iterative best-response arguments as PSRO, [21] proposed Joint PSRO (JPSRO), a population learning algorithm with convergence guarantees to a NF (C)CE in n-player general-sum EF games. Nevertheless, evidence of convergence to (C)CE has been limited to a few research games that can be solved analytically — the costs of representing, training and evaluating a population of independent RL agents *for each player* quickly become intractable, especially in games that demand complex skills.

How can we bring game-theoretic algorithms such as (J)PSRO to real-world games in full generality and at scale? The central

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.

question here, we argue, is that of efficient policy representation. If human players can routinely combine and reuse different skills to develop new strategies in games such as Chess, Go and Poker, could artificial agents reuse fundamental skills such as locomotion, perception and memory across strategies too? Whereas such skills must be learned repeatedly at each iteration in (J)PSRO using independent RL, Neural Population Learning (NeuPL, [16, 19]) transfers and refines such skills across all policies within the population. Conditioned on opponent priors, the same neural network represents diverse best-response policies and implements PSRO in symmetric zero-sum games. This approach offers several advantages. The shared, probabilistic representation of opponent behaviours promotes skill transfer between strategies, allows for online adaptation under different opponent priors and reduces the computational costs of training a population of policies to be comparable to that of self-play. Despite strong empirical results in several test domains, NeuPL remains limited in important ways. First, NeuPL is restricted to symmetric zero-sum games, limiting its generality in practical applications. Second, the convergence guarantees of population learning algorithms with shared strategy representation requires further clarification, as we shall explain in this work. Relatedly, NeuPL fell short of demonstrating empirical convergence to a NE in research games where convergence can be verified empirically.

We address these limitations with NeuPL-JPSRO, a scalable, equilibrium convergent algorithm for n-player general-sum games. We clarify the convergence guarantees of population learning algorithms with shared representation, before motivating fundamental departures from NeuPL that ensure convergence in general games. Empirically, we show that NeuPL-JPSRO converges to a CCE in several OpenSpiel games at a rate of convergence comparable to exact JPSRO where solutions can be evaluated exactly using analytical game solvers. Lastly, we show that unlike JPSRO, NeuPL-JPSRO can be deployed efficiently in complex domains, demonstrating adaptive, coordinated control in MuJoCo control domains and transfer learning of skills in the team strategy game of capture-the-flag that requires spatiotemporal reasoning from partial, visual observations.

Additional experimental details and proofs are available in appendices to the extended version of this paper [20].

2 PRELIMINARIES

We now formally describe CCE [24] and JPSRO [21] before introducing our method NeuPL-JPSRO.

2.1 Coarse Correlated Equilibrium (CCE)

Normal-form (NF) games are the simplest game formulation where each player p plays one of its actions $a_p \in \{a_p^0, a_p^1, \dots\} = \mathcal{A}_p$ simultaneously and receives a payoff $G_p : \mathcal{A} \to \mathbb{R}$ as a function of the joint action $a = (a_1, \dots, a_n)$, with $a \in \otimes_p \mathcal{A}_p = \mathcal{A}$ and n the number of players. For player p, we denote $a = (a_p, a_{\neg p})$ with $a_{\neg p} = (\dots, a_{p-1}, a_{p+1}, \dots)$, the actions played by all players except p. Let $\sigma(a) = \sigma(a_p, a_{\neg p})$ denote the probability of players playing the joint action a and σ a probability distribution over the space of joint actions. A pure strategy is an action distribution that is deterministic and a mixed strategy is one that can be stochastic. The value for player p under a mixed joint strategy σ is defined as $\mathbb{E}_{a\sim\sigma}[G_p(a)] = \mathbb{E}_{a\sim\sigma}[G_p(a_p, a_{\neg p})]$. Let $\sigma_{\neg p}(a_{\neg p}) = \sum_{a_p} \sigma(a_p, a_{\neg p})$ be a marginal distribution over all players' action choices other than that of player p. Let $\Delta_{\neg p}$ be the probability simplex of all such distributions. We similarly define marginals $\sigma(a_p) = \sum_{\neg p} \sigma(a_p, a_{\neg p})$.

Let $\lfloor x \rfloor_+ = \max(x, 0)$. The maximum incentive for p to deviate from σ by choosing action a'_p is:

$$\delta_p(\sigma) = \Big\lfloor \max_{a'_p \in \mathcal{A}_p} \Big(\mathop{\mathbb{E}}_{a\sim\sigma} [G_p(a'_p, a_{\neg p}) - G_p(a)] \Big) \Big\rfloor_+. \tag{1}$$

σ is an ϵ-CCE if $\forall p, \delta_p(\sigma) \le \epsilon$. A common metric for quantifying the approximation of σ to a CCE is the CCE gap: $\delta(\sigma) = \sum_p \delta_p(\sigma) \ge 0$. A CCE gap is 0 if and only if σ is a CCE. The value to a player under a CCE is its CCE value. CCE generalises NE in that every NE is a CCE but only CCE that factorise into marginals $\sigma(a) = \prod_p \sigma(a_p)$ are also NEs. Consequently, CCE always exist in finite games [26].

NF CCE can be applied to EF games and we describe a specific construction that allows us to do so. In an EF game, player p follows a policy $\pi_p(\cdot|s)$ that maps a state s to a distribution over actions. Analogous to players taking actions in an NF game, we can construct an NF metagame where playing an action $a_p \in \mathcal{A}_p$ executes a policy $\pi_p \in \Pi_p$ in the EF game for player p. Such a metagame would be large as it contains all enumerated policies as actions. We refer to the joint policy of all players $\pi = (\pi_1, \dots, \pi_n)$ as the joint policy and $\pi_{\neg p} = (\dots, \pi_{p-1}, \pi_{p+1}, \dots)$ as the co-player joint policy for player p. In this metagame, the definition of an NF CCE applies. The $\max_{a'_p \in \mathcal{A}_p}$ operator, from Equation 1, amounts to $\max_{\pi'_p \in \Pi_p}$ or finding a policy π'_p that maximises player p's expected payoff to a co-player mixed-strategy $\sigma_{\neg p}$. The policy that does so is a best-response (BR) to the co-player mixed-strategy $\sigma_{\neg p}$. We study such NF metagames for the rest of this paper.

2.2 Joint Policy-Space Response Oracle (JPSRO)

The action space of an NF metagame can be intractable to enumerate. Nevertheless, we can study *restricted* NF metagames, whose action spaces are subsets of those of the full game. JPSRO [21] implements such an algorithm that converges to an NF CCE of an EF game (Algorithm 1). Starting from an initial *restricted* metagame with a set of starting policies for each player $\Pi^0 = \{\pi_p^0\}_{p=1}^n$, JPSRO computes a BR policy to co-player joint policies $\Pi_{\neg p}^{t-1}$, sampled according to the marginal CCE $\sigma_{\neg p}^{t-1}$ and adds it to player p's set of policies at iteration t. This amounts to adding a metagame action for each player, leading to an expanded metagame at the next iteration. The expected payoff (EP) G^t is re-evaluated at each iteration for all joint policies and σ^t is re-computed using a CCE meta-strategy solver (MSS). σ^t is referred to as a CCE of the *restricted* game, or a CCE mixed-strategy at iteration t. This process terminates when a BR operator cannot find policies $\pi_p^t, \forall p$ that yield an ϵ improvement in expected payoff with $\max_p \delta_p^t = \big\lfloor \mathbb{E}_{\pi\sim\sigma^{t-1}} [G_p(\pi_p^t, \pi_{\neg p}) - G_p(\pi)] \big\rfloor_+ < \epsilon$, resulting in an NF ϵ-CCE of the full EF game.

In short, instead of enumerating policies which is intractable for most NF metagames, JPSRO uses a BR operator to implement the maximisation step from Equation 1 and constructs a sequence of *restricted* metagames whose CCE converge to a CCE of the full game. The BR operator can be implemented using analytical solvers if available (as in [21]) or approximate methods such as RL. We refer to the former case as *exact* JPSRO and the latter *approximate*.

Algorithm 1 JPSRO (CCE) [21]

$\Pi_1^0, \dots, \Pi_n^0 := \{\pi_1^0\}, \dots, \{\pi_n^0\}$
$G^0 \leftarrow \mathrm{EP}(\Pi^0)$
$\sigma^0 \leftarrow \mathrm{MSS}(G^0)$
for $t \in [1, \dots]$ **do**
  for $p \in [1, \dots, n]$ **do**
    $\pi_p^t, \delta_p^t \leftarrow \mathrm{BR}(\Pi_{\neg p}^{t-1}, \sigma_{\neg p}^{t-1})$
    $\Pi_p^t \leftarrow \Pi_p^{t-1} \cup \{\pi_p^t\}$
  end for
  if $\max_p \delta_p^t < \epsilon$ **then** ▷ ϵ-CCE.
    return $(\Pi^{t-1}, \sigma^{t-1})$
  end if
  $G^t \leftarrow \mathrm{EP}(\Pi^t)$ ▷ payoffs.
  $\sigma^t \leftarrow \mathrm{MSS}(G^t)$ ▷ CCE solver.
end for

3 NEUPL-JPSRO

We now describe NeuPL-JPSRO (Algorithm 2), an algorithm that builds on JPSRO but scales up to complex domains using function approximation and deep RL. The key idea behind NeuPL-JPSRO is to parameterise each policy for each player with a strategy embedding vector $v_p^t \in \mathbb{R}^d$, resulting in player-specific strategy embedding vectors $\mathcal{V} = \{\mathcal{V}_1, \dots, \mathcal{V}_n\}$ where $\mathcal{V}_p = \{v_p^0, v_p^1, \dots\}$ represents strategies available to player p. Each strategy embedding vector parameterises a policy $\Pi_\theta(\cdot|s, v)$, using a conditional neural network with parameters θ shared over all players' strategies. Strategy embedding vectors v_p^t are randomly initialised and optimised jointly with the rest of the network parameters θ. For conciseness, $\Pi_\theta^{\mathcal{V}}$ denotes all sets of policies for all players. We omit subscript or superscript on strategy embeddings when referring to all players or all strategies of a player. Strategy embeddings are optimised end-to-end in the same way as the network parameters θ. We use $\pi'(\cdot|s) \leftarrow \pi(\cdot|s)$ to refer to policy π' updating its action distribution to that of π in state s. This can be implemented exactly in the tabular case, and approximately by minimising KL-divergence $D_{\mathrm{KL}}\left[\pi(\cdot|s) \,||\, \pi'(\cdot|s)\right]$ when using function approximation.

While NeuPL-JPSRO closely follows JPSRO, a key issue arises with this integrated approach to policy representation — changes to one policy now affect behaviours of others. This could be beneficial in the form of skill transfer, as shown in [19] (and the rest of this work), but may also affect the convergence guarantees of the algorithm. We discuss the convergence properties of population learning algorithms with shared representation in Section 3.1 before explaining how NeuPL-JPSRO in theory guarantees convergence to a CCE with a continual learning approach. Section 3.2 examines how function approximation can be used in each of the key operations of NeuPL-JPSRO to scale up to large games. In particular, best-response learning can benefit from skill transfer and expected payoff (EP) evaluation can build on the learned strategy embeddings. Our approach here generalises NeuPL which is specialised to strategy representation in symmetric zero-sum games. The scaling approaches described here are also in contrast to exact JPSRO that does not generalise to large games, or approximate JPSRO, which quickly becomes computationally prohibitive at scale.

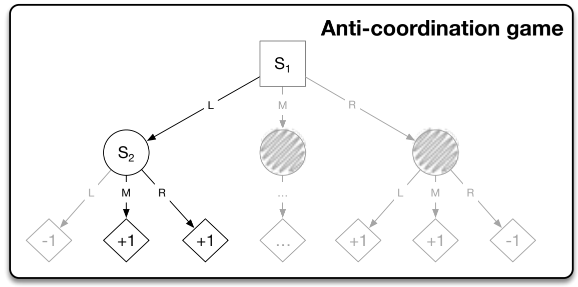

Figure 1: A turn-based two-player zero-sum game where player 1 publicly chooses a direction that player 2 is rewarded for avoiding. Terminal nodes show the payoffs of player 2.

3.1 Convergence to Equilibria

A key element of the convergence arguments of NeuPL [16, 19] relies on the stationarity of co-player policies at each best-responding iteration τ. Given stationary co-player policies, it follows that a max-entropy RL algorithm would converge uniquely and the resulting policy would stand as a stationary co-player policy for all iterations $t > \tau$ that follow. We show that the use of a max-entropy RL algorithm does *not* uniquely identify a best-response policy with a counter-example (Figure 1), unless additional assumptions are made on the dynamics of the game and the co-players. Specifically, we argue that max-entropy RL only results in stationary action distributions in *states it visits when best-responding*, which is a subset of all states where it may need to act as a co-player. This means that during best-response learning at iteration t, co-player policies developed at earlier iterations $\tau < t$ may be non-stationary in some states, and their best-responses may in turn change. In other words, best-response learning for iterations $t > \tau$ may be optimising towards "moving targets", and may never converge.

To make this concrete, consider the toy example shown in Figure 1. In this game, player 1 publicly declares a direction and player 2 is rewarded for avoiding it. Suppose player 1 plays a deterministic policy that always chooses "L" in state S_1 to which player 2 best-responds, a sample-based maximum-entropy BR operator would assign equal probability to "M" and "R" in state S_2, converging to the maximum-entropy BR which is stationary in all states reachable under the deterministic player 1 policy. Nevertheless, player 2 may be forced to act in additional (dashed) states in subsequent iterations with a different player 1 strategy – indeed, with function approximation and shared representation, player 2 would behave unpredictably in these states yet any behaviour in these states would still constitute a valid best-response to player 1's original strategy. This is problematic for convergence, as the best-response to player 2's policy may change arbitrarily and there is no guarantee that the population would expand over all finite policies in the limit. This is true unless the best-responding policy can reach all states under the co-player joint policies (Definition 3.1). Under the full-support condition, NeuPL policies will remain stationary in all states. Indeed, the two domains considered in [19] likely satisfy this full-support condition: *rock-paper-scissors* is an NF game, and in *running-with-scissors* co-player policies are stochastic, under partial observability and entropy-maximising, resulting in full-support reach probabilities for all states. This need not be the case in general games. In fact, the problem is particularly salient in games where best-response strategies are often deterministic (e.g. *goofspiel*, also known as the game of pure-strategies [30]). We evaluate NeuPL-JPSRO in such domains in our experiments in Section 4 to show its robustness in converging to a CCE even in the absence of the full-support condition.

Definition 3.1 (Full-Support)**.** All states can be reached with positive probability at each iteration under the co-player joint-policy.

How does NeuPL-JPSRO ensure convergence to an equilibria while maintaining the computational efficiency of shared strategy representation? The key idea behind NeuPL-JPSRO is to take a continual learning approach that removes the need for the full-support condition entirely.

Algorithm 2 NeuPL-JPSRO (Ours)

$\mathcal{V}_1, \dots, \mathcal{V}_n = \{\nu_p^0\}, \dots, \{\nu_n^0\}$ ▷ With $\Pi_\theta(\cdot|s,\nu)$.
$G^0 \leftarrow \mathrm{EP}(\Pi_\theta^{\mathcal{V}})$
$\sigma^0 \leftarrow \mathrm{MSS}(G^0)$
for $t \in [1, \dots]$ **do** ▷ NeuPL-JPSRO iterations.
  $\hat{\theta} \leftarrow \theta, \hat{\mathcal{V}} \leftarrow \mathcal{V}$ ▷ Reference policy parameters.
  for $p \in [1, \dots, n]$ **do**
    $\pi_p^t, \delta_p^t \leftarrow \mathrm{BR}(\Pi_{\hat{\theta}}^{\hat{\mathcal{V}}}, \sigma_{\neg p}^{t-1})$ ▷ See Section 3.2.1.
    $\forall s, \Pi_\theta(\cdot|s,\nu_p^t) \leftarrow \pi_p^t(\cdot|s)$ ▷ Distill.
    $\forall s, \Pi_\theta(\cdot|s,\nu_p) \leftarrow \Pi_{\hat{\theta}}(\cdot|s,\hat{\nu}_p)$ ▷ Regularise.
    $\mathcal{V}_p \leftarrow \mathcal{V}_p \cup \{\nu_p^t\}$
  end for
  if $\max_p \delta_p^t < \epsilon$ **then**
    return $(\Pi_{\hat{\theta}}^{\hat{\mathcal{V}}}, \sigma^{t-1})$
  end if
  $G^t \leftarrow \mathrm{EP}(\Pi_\theta^{\mathcal{V}})$ ▷ See Section 3.2.3.
  $\sigma^t \leftarrow \mathrm{MSS}(G^t)$
end for

Instead of concurrently optimising at all iterations as in NeuPL, learning in NeuPL-JPSRO proceeds iteratively — at iteration t, the best-response policies for each player $\pi_p^t, \forall p \in [n]$ are optimised against stationary co-player policies (we describe how π_p^t are computed efficiently next). To ensure co-player stationarity, we introduce a set of reference policies $\Pi_{\hat{\theta}}(\cdot|s,\hat{\nu}_p^\tau), \forall \tau < t, \forall p \in [n]$, whose parameters $\hat{\theta}$ and $\hat{\mathcal{V}}$ are held fixed for the duration of one iteration. Within one iteration, the best-response policies $\pi_p^t, \forall p \in [n]$ being optimised are distilled into the neural population $\forall s, \Pi_\theta(\cdot|s,\nu_p^t) \leftarrow \pi_p^t(\cdot|s)$. At the same time, the behaviours of all existing strategies are held stationary via regularisation $\forall s, \Pi_\theta(\cdot|s,\nu_p) \leftarrow \Pi_{\hat{\theta}}(\cdot|s,\hat{\nu}_p)$ in all states that player p may reach. We recall that both distillation and regularisation are implemented as a minimisation problem of the KL-divergence between two policies. Our approach here draws inspiration from [32] where a conditional model continuously compresses existing skills while incorporating new ones with a shared skill latent space. We can now formally state the convergence guarantees of NeuPL-JPSRO with a proof that trivially extends from the convergence arguments of JPSRO [21].

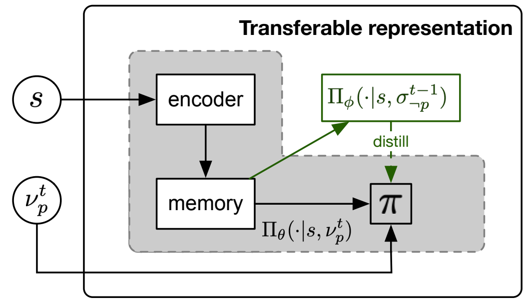

Figure 2: Efficient best-response solving by reusing transferable representation from the policy population $\Pi_\theta^{\mathcal{V}}$. At iteration t, the policy head Π_ϕ (green) reuses the encoder and memory representation from Π_θ (gray) to learn a best-response to co-player mixed-strategy $\sigma_{\neg p}^{t-1}$. The best-response policy is concurrently distilled into the neural population of policies $\Pi_\theta^{\mathcal{V}}(\cdot|s,\nu_p^t)$ under the strategy embedding vector ν_p^t.

Theorem 3.2 (CCE Convergence)**.** *When using a CCE meta-strategy solver in NeuPL-JPSRO, and when distill and regularise operators are exact, the sequence of mixed-strategy converges to a normal-form CCE under the meta-strategy distribution.*

Proof. In this case, NeuPL-JPSRO is equivalent to JPSRO which is known to converge, as proved in [21]. □

Finally, we generalise the convergence arguments of JPSRO [21] which assumes that the BR operator returns deterministic policies, to also consider the case of stochastic policies as is often the case in policy-gradient RL methods. In fact, we show that by encouraging specific types of stochastic policies such as the ones that are entropy-maximising, we can guarantee that the population learning process terminates. We defer formal arguments to Appendix A for completeness and use max-entropy RL algorithms for best-response solving in this work in practice.

3.2 Scaling to large games

3.2.1 BR learning to CCE co-player mixed-strategies. At iteration t, NeuPL-JPSRO solves for a BR policy against co-player CCE mixed-strategy $\sigma_{\neg p}^{t-1}$ for every player p. For games that cannot be solved analytically, approximate methods such as independent deep RL can be used. Nevertheless, doing so at every iteration would be impractical (as we show in Section 4.3). Learning from scratch when facing skilled co-players becomes challenging at later iterations, especially when basic skills (e.g. locomotion, perception and recurrent memory) are required for any meaningful exploration to occur. Instead, we implement the BR operator using a policy network that benefits from transferable skills shared by the population of policies $\Pi_\theta^{\mathcal{V}}$. Figure 2 shows how an auxiliary policy head $\Pi_\phi(\cdot|s,\sigma_{\neg p}^{t-1})$ reuses the strategy-agnostic encoder and memory networks, and learns to best-respond to co-player mixed-strategies $\sigma_{\neg p}^{t-1}$. Care needs to be taken when representing $\sigma_{\neg p}^{t-1}$ in neural networks: the number of co-player joint strategies grows exponentially and a useful representation would reflect the player's probabilistic prior over co-players' joint strategies. We represent a co-player mixed-strategy $\sigma_{\neg p}$ with

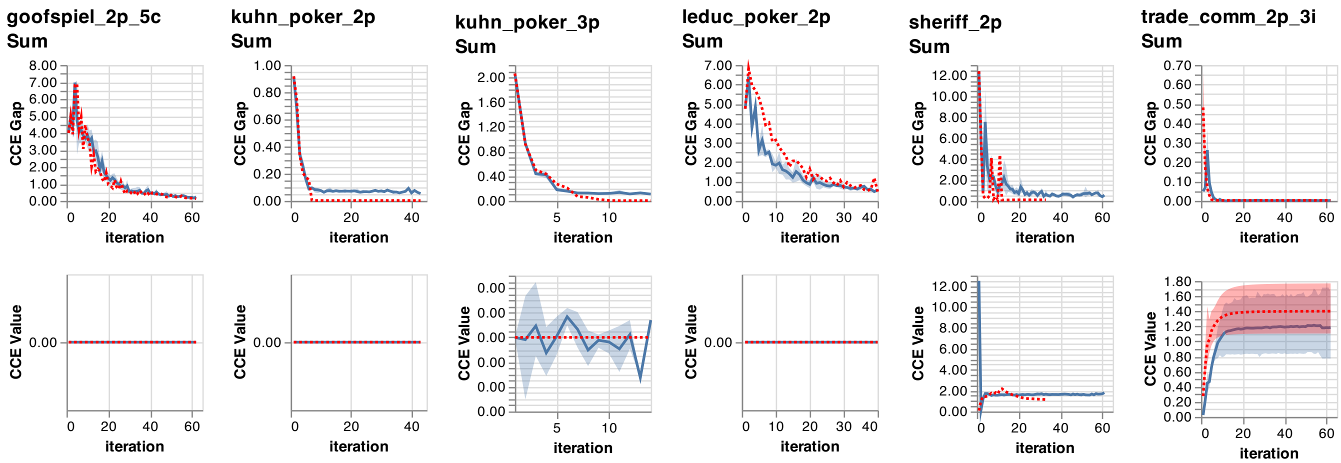

Figure 3: Exact CCE gaps and CCE values in 6 OpenSpiel games for NeuPL-JPSRO (Blue) compared JPSRO (Red) using exact best-response and expected payoff solvers averaged over 5 seeds.

a weighted representation $g(\mathcal{V}, \sigma_{\neg p})$ as follows

$$g(\mathcal{V}, \sigma_{\neg p}) = \sum_{a_{\neg p}} \sigma_{\neg p}(a_{\neg p}) f(\dots, v_{p-1}^{a_{p-1}}, v_{p+1}^{a_{p+1}}, \dots) \tag{2}$$

considering only TOP-K co-player joint strategies $a_{\neg p}$ under $\sigma_{\neg p}$ for the summation operation. We describe further details in Appendix B.1 where we show that $k = 96$ almost always leads to a lossless representation of co-player joint strategies in our experiments, as well as how player symmetry can be leveraged in the encoding function f for further representation efficiency.

3.2.2 BR learning to any co-player mixed-strategy. The weighted co-player representation (Equation 2) allows us to replicate results from [16] too, where a conditional policy $\pi(\cdot|s, \sigma_{\neg p})$ responds Bayes-optimally to any co-player mixed-strategies $\sigma_{\neg p} \in \Delta_{\neg p}$. We demonstrate this potential for online adaption in Section 4.2 where a player collaborates with different co-players under uncertain prior beliefs. We make a simple modification to Algorithm 2 where we additionally sample arbitrary distributions over co-player strategies $\sigma_{\neg p} \sim \Pr(\Delta_{\neg p})$ and optimise $\Pi_\phi(\cdot|s, \sigma_{\neg p})$ to best-respond accordingly. At convergence, $\Pi_\phi(\cdot|s, \sigma_{\neg p})$ behaves Bayes-optimally under any prior $\sigma_{\neg p} \in \Delta_{\neg p}$ over co-player joint-strategies.

3.2.3 Expected payoff evaluation. Payoff tensor G^t needs to be evaluated at each iteration to update the metagame CCE mixed-strategy $\sigma^t \leftarrow \mathrm{MSS}(G^t)$. This can be costly: the number of joint strategies to evaluate grows exponentially in the number of players and estimating payoffs under each joint strategy may require many simulations in the absence of an analytical solver. We leverage learned strategy embedding vectors $\mathcal{V}$ and continuously optimises a payoff estimator network $G(a) \leftarrow \psi_w(v_1^{a_1}, \dots, v_n^{a_n}) \in \mathbb{R}^n$ that predicts payoffs to each player under a joint strategy $a = (a_1, \dots, a_n)$. This network is used in all our experiments in lieu of the EP operator. Similar to [19], the payoff estimator network is trained to minimise the same loss function as the action-value function of the underlying RL agent. As ψ_w is only conditioned on joint strategy embedding vectors and unaware of the state-action pairs, it is therefore regressing towards the expected returns for each player under a specific joint-strategy, with the expectation over the state and action distribution when players play out the specific joint-strategy a. We describe how this payoff estimator network is implemented and optimised in Appendix B.3. This payoff estimation network removes the need for evaluating payoff tensors through simulation, making payoff estimation at every NeuPL-JPSRO iteration practical and efficient.

4 RESULTS

We now present our empirical results that aim to answer two questions. First, we verify rigorously that NeuPL-JPSRO converges to a CCE in non-monotone strategy games for which exact CCE gaps and CCE values can be computed using analytical solvers. Second, we demonstrate its efficiency and generality by applying it to complex domains that involve realistic physics, partial observability and team-play. Additional technical details on compute, network architecture and game settings are provided in Appendix C.

4.1 Convergence in n-player general-sum games

To demonstrate empirical convergence to a CCE in n-player general-sum games, we investigated a diverse set of 6 strategy games from the OpenSpiel task suite [13] as described in Appendix C.1. Figure 3 summarises our results in each game reporting the sum of CCE gaps and CCE values across players when playing their equilibrium strategies of the *restricted* metagame at each iteration. For comparison, we show the same metrics for exact JPSRO in red, where entropy-maximising BRs are computed analytically at each iteration. 5 independent trials have been run for each game for each algorithms. For all games except `trade_comm`, the initial policy acts uniformly in all states[1]. The evaluation of CCE gaps and CCE values are exact in that the only input to the evaluation procedure from NeuPL-JPSRO are the sets of trained policies $\Pi_\theta^{\mathcal{V}}$. The value

[1]In `trade_comm`, learning a maximum-entropy BR to a uniform policy makes no progress: for a policy that ignores the message received, the BR would be agnostic to what message to send.

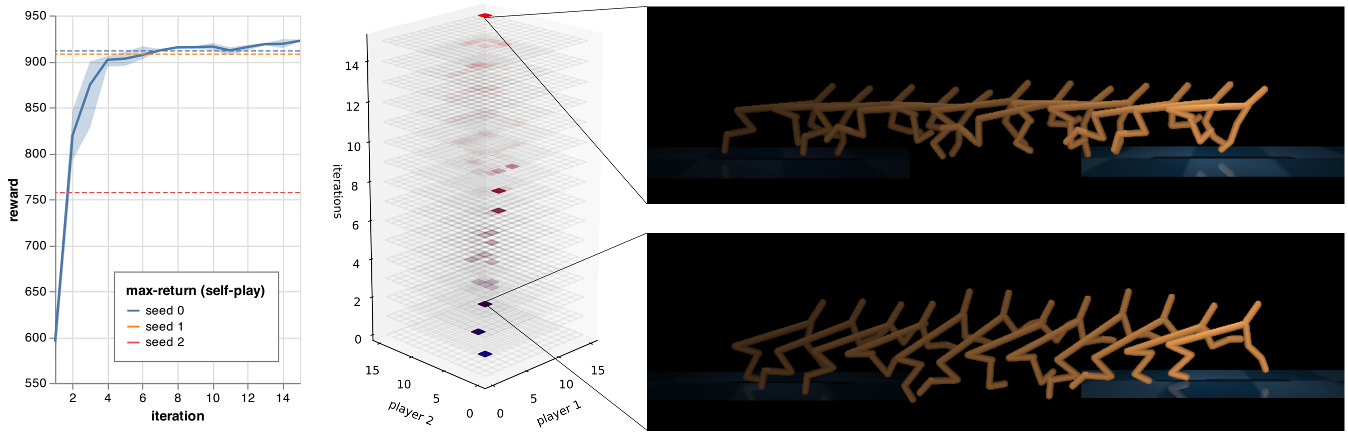

Figure 4: Emergence of cooperation in MuJoCo multiagent `cheetah_run`**. (Left) Expected returns achieved at each iteration (solid) compared to the maximum return obtained by independent trials where players optimise through self-play (dashed). The average return at iteration 16 is comparable to that of SoTA single-agent RL [33] (Middle) The sequence of CCE best-responded to at each iteration for player 1 (rear leg) and player 2 (front leg). (Right) Visualization of the learned behaviours at iteration 3 where the rear leg player raises the front leg player and at iteration 16 where both players cooperate competently.**

of each policy, the optimal deviation actions as well as the CCE distributions are computed using analytical solvers. For completeness, detailed results from each trial and for each player are shown in Appendix C.1. For instance, inspecting per-player CCE values shows that NeuPL-JPSRO has recovered the last-mover advantage in poker. The sustained stability of the CCE values and CCE gaps at each NeuPL-JPSRO iteration suggest empirical co-player stationarity, necessary under our convergence arguments.

We make the following remarks. First, we observe empirical convergence towards a CCE in all games with both methods. In some games, NeuPL-JPSRO discovered and represented up to 64 policies for each player, far exceeding the size of the population reported in prior works [19]. This demonstrates the potential of NeuPL-JPSRO to converge in games with long strategy cycles such as `goofspiel`. Second, neither NeuPL-JPSRO nor JPSRO converged to CCE with specific properties. In particular, the values of the CCE do not converge to the values of the maximum-welfare CCE in non-zero-sum games (`sheriff` and `trade_comm`)[2]. Equilibrium selection [3] remains an open question: every CCE describes a rational, stable state of the system in that no one has an incentive to deviate from their equilibrium strategy but only certain equilibria are socially valuable or fair. NeuPL-JPSRO converges to one such equilibrium, but to which remains unclear.

Additionally, we note that in some games, NeuPL-JPSRO, an approximate method, has observed faster convergence to a CCE in some games than exact JPSRO in early iterations (e.g. Leduc Poker in the first 20 iterations). This maybe counter-intuitive at first, but we shall explain why this could be the case in practice through an example. Consider the game of *rock-paper-scissors* and suppose we start off the population with a deterministic always-rock policy. An exact best-response to this initial policy would be always-paper and an approximate one (perhaps due to entropy maximisation) may randomise but with a strong preference for paper (e.g. 80%-paper, 10% rock and 10% scissors). In each scenario, the CCE between the first two strategies amounts to always playing the latest strategy — always-paper is played in the former, and a mostly-paper strategy is played in the latter. It is therefore unsurprising that the exact case has a higher CCE gap as always-paper is more exploitable (to always-scissors) than its approximate, randomising counterpart. Indeed, in population learning algorithms such as PSRO, the rate of convergence is not necessarily faster when the best-response solver is exact. The empirical rate of convergence to an equilibrium depends on the dynamics of the game and the choice of initial policy as well. Nevertheless, we show NeuPL-JPSRO to converge empirically, at a rate comparable to exact JPSRO.

[2]Our results for the JPSRO baseline differ from what have been reported in [21]. This is because the default analytical BR solver in OpenSpiel chooses the first action deterministically among indifferent ones. We used a maximum-entropy solver instead, which removes this implicit coordination bias.

4.2 Online adaptation in multiagent MuJoCo domains

Humans often adapt to each other in an online fashion, whether it's in competition, cooperation, or a mixture of both. Through iterative BR solving, population learning can generate diverse and strategically relevant behaviours. Prior work leveraged this property in a competitive setting [16], where a single conditional policy $\pi(\cdot|s, \sigma_{\neg p})$ is trained to best-respond to arbitrary metagame mixed-strategies $\sigma_{\neg p} \in \Delta_{\neg p}$. By interpreting $\sigma_{\neg p}$ as a *prior* over co-player joint strategies, $\pi(\cdot|s, \sigma_{\neg p})$ trades off exploration and exploitation Bayes-optimally, maximising its expected returns. We study if NeuPL-JPSRO can lead to similar adaptive behaviours but in a cooperative setting. We construct a cooperative common-payoff physically-simulated game using `cheetah_run` [37] where player 1 and player 2 control the rear and front legs respectively and receive the same forward velocity reward. Both players observe all joint positions of the cheetah embodiment but do not observe each

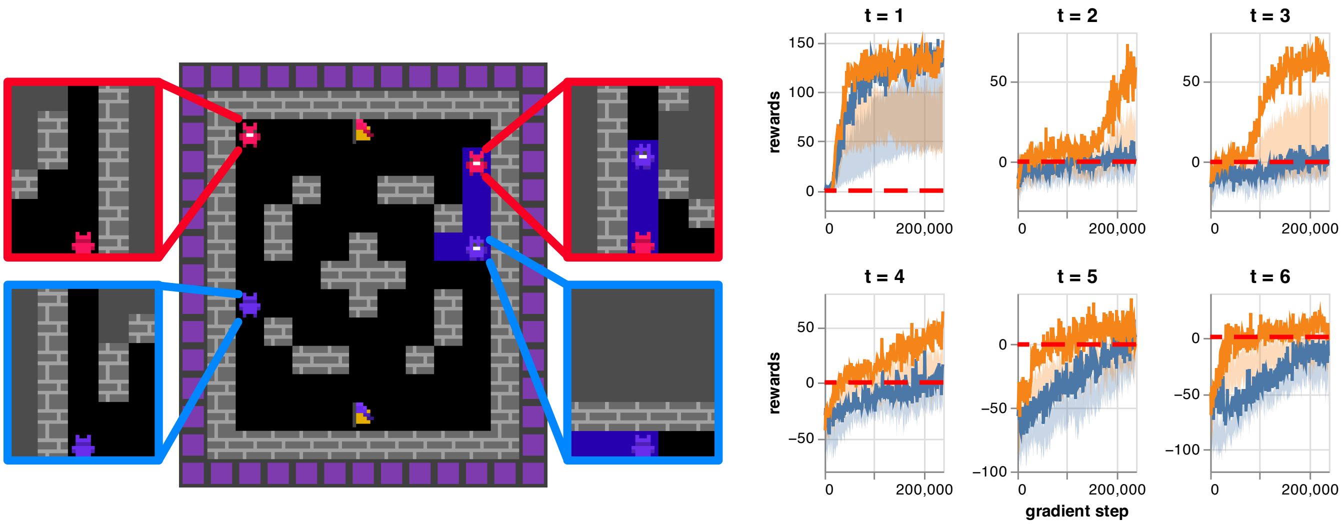

Figure 5: (Left) A 4-player `capture-the-flag` environment showing first-person views for each player. (Right) Convergence to a CCE shown by the diminishing incentive to deviate to an independent BR across iterations. (Blue) Expected returns of independent RL exploiter policies optimised against the marginal CCE mixed-strategy $\sigma^t_{\neg p}$ from NeuPL-JPSRO at iteration t. (Orange) Same as Blue, but initialised with pre-trained encoder and memory network parameters (as in Figure 2). (Red) CCE values at each NeuPL-JPSRO iteration. Solid lines show the optimistic *best-of-six* exploiter returns.

other's joint actuation. To do well, players must build a model of the co-player at play and react accordingly.

Our goal is to verify that by letting players control parts of the same body, the population of policies $\Pi^{\mathcal{V}}_\theta$ would develop diverse behaviour patterns, while the conditional BR policy $\Pi_\phi(\cdot|s, \sigma_{\neg p})$ learns to adapt to different partners in an online fashion. Figure 4 (Left) shows our findings. With NeuPL-JPSRO, the two players discovered a sequence of 16 strategies using the same conditional network, continuously improving the expected returns across iterations. The final averaged return outperforms self-play baselines with less variances across seeds. Figure 4 (Middle) visualised the sequence of CCE joint distributions discovered over 16 iterations. In earlier iterations, the player controlling the rear leg learned to raise its partner so as to minimise the disruption caused by an unskilled co-player (shown in Figure 4 (Bottom-Right)). This strategy became dominant in earlier iterations as it maximises returns for the pair. In subsequent iterations, the front leg player discovered policies that can cooperate with the rear leg player, leading to coordinated joint policies as shown in Figure 4 (Top-Right). We show this emergence of partnership visually over the first 4 iterations[3]. More interestingly, by conditioning the final BR policy $\Pi_\phi(\cdot|s, \bar{\sigma}_{\neg p})$ with a uniform co-player prior $\bar{\sigma}_{\neg p}$, we observed that the same policy is capable of probing its front leg partner through interaction and adapt accordingly online[4]. Following initial feedback, the rear leg player either took control of the full body if its partner appears uncooperative or worked effectively with a front leg partner that proved competent. Our findings are consistent with [16], who considered agents adapting to multiple opponents in a competitive setting. We show here that similar adaptive behaviors can emerge in a cooperative setting with asymmetric roles using NeuPL-JPSRO.

4.3 Strategic team-play in capture-the-flag

NeuPL-JPSRO efficiently scales up to games requiring transferable skills while converging to an approximate CCE. In this section, we investigate `capture-the-flag`, a 4-player game where players compete in teams to capture the opponents' flag and bring it back to the home base under partial observability.

A global view of the environment as well as players' first-person observation of the game are shown in Figure 5 (Left). Compared to the standard environment described in [15], we modified the visibility mechanism such that brick walls would restrict players' view (shown in grey). This puts further emphasis on players' abilities to infer others' strategies under partial visibility. We used the default rewards, which are sparse: players on the same team receive a reward of +10 (-10) upon capturing (conceding) a flag and zero otherwise. Players on the same team do not have explicit communication channels: they can observe floor paint left by co-players, current status of flag capture and the behaviours of other players to form implicit conventions. The opponent's flag is captured when returned to one's home if and only if their own flag remains at home too. Players may paint the floor and they cannot move if the tile they stand on has been painted over by an opponent. A player is removed if they are tagged twice within a short time and they recover more quickly if standing on tiles painted in their own colour. Despite the complexity of the game, NeuPL-JPSRO successfully discovered a sequence of 8 strategies, using a single conditional network. It is challenging to offer a succinct view of the learning dynamics considering the size of the payoff tensor, but Appendix C.3 shows an overall trend of progress: joint strategies

[3] cheetah-run: https://youtu.be/-bBR6Vtu0sI
[4] cheetah-adapt: https://youtu.be/zMwhWgafAK4.

developed at later iterations tend to be less exploitable, dominating earlier combinations of opponent strategies.

We visualise these strategies[5]. While this environment is a simplified variant of the classic 3D game studied in [11], we observe similar behaviour patterns to emerge here as well. Throughout the first few iterations, players incrementally learned to implement coordinated strategies such as "Home Base Defence", "Opponent Base Camping" and finally "Teammate Following". Other skills such as timing the tagging cool-down mechanism have emerged, too.

It is difficult to verify convergence to a CCE in a game of this complexity. Nevertheless, we offer empirical evidence of convergence to a CCE by independently training best-responding RL policies against marginal CCE mixed-strategies of the restricted metagames at each iteration $\{\sigma^t\}_{t=0}^T$, using standard RL algorithms. This is similar to the exact CCE gap evaluation in Section 4.1 but with an approximate BR solver. Figure 5 (Right) shows our findings. Compared to the CCE value to player p when playing according to the joint CCE distribution σ^t (Red dashed), a standard RL algorithm managed to find policies with better returns in the first 5 iterations (Orange and Blue). These iterations have not converged to a CCE of the game as profitable deviations can be made by unilaterally switching to these independently optimised BR policies. At iteration 6, independent RL training can no longer outperform. This shows that we are in close proximity of a CCE and little room for improvement remains compared to what player p's equilibrium strategy already implements. We note that we report the *maximum* expected returns observed over 6 independently trained BR policies. This is because our goal is to certify if *any* profitable deviation actions can be found. We report the average case in Appendix C.3 where independent RL struggled to outperform the CCE equilibrium strategies as soon as iteration 5. Our exploiter results suggest that RL policies optimised via self-play would *at best* match the CCE solution of NeuPL-JPSRO. We note that the CCE value to each player at every iteration must be zero, as the game is zero-sum.

An orthogonal benefit of NeuPL-JPSRO lies in its potential in transfer learning across players and strategies. This is particularly attractive in domains such as ours that afford transferable skills. To do well, players must learn to represent visual observation history so as to infer other players' strategies under partial observability. In close combat, players must understand the cool-down mechanism of the tagging action and proactively retreat to a safe distance in a stand off. Figure 5 (Right) confirms our hypothesis, showing that a randomly initialised RL policy struggled to best-respond to strong opponents (Blue) when a policy partly initialised with pre-trained encoder and memory network parameters succeeded (Orange). This highlights the importance of transfer learning in complex games — strong opponents tend to create difficult exploration problems for randomly initialised RL policies. In this instance, approximate JPSRO would have led to the incorrect conclusion that no further improvement can be made after 3 iterations, forfeiting strategically interesting joint-strategies of this game. NeuPL-JPSRO naturally promotes a progressive learning curriculum and benefit from transfer learning of strategy-agnostic skills.

[5] capture-the-flag: https://youtu.be/z5EeMfcOo7A. Each player's first-person view is annotated in white with the iteration of the strategy at play. The cumulative return to each team is shown at the top.

5 RELATED WORK

A rich body of literature have focused on providing convergence guarantees to equilibria in games of different degrees of generalities. Two-player zero-sum games, in particular, attracted attention given the tractability and interchangeability of NE in this setting. Families of NE-convergent methods have been developed, with recent successes in scaling to complex EF games of real-world interests [4, 14, 23, 29]. Progress beyond two-player zero-sum games trails behind in comparison. Principled no-regret learning methods [7, 8] have been proposed, converging to the generalised solution concept of (C)CE [1, 24]. Nevertheless, there has been limited success in deploying these methods to in many practical applications due to their computational challenges in scaling to large games.

Of particular relevance to our work are population learning methods that leverage latest advances in function approximation and deep RL. Several works proposed to learn a population of deep RL agents and observed emergent complex behaviours through population-based interactions [2, 11, 17, 18, 39]. While these works demonstrated what could be achieved by multiagent RL at scale, their convergence characteristics remain to be understood [10]. In contrast, [14, 19, 22, 36] proposed methods that leverage deep RL while retaining the game-theoretic convergence guarantees. Fewer works in this category ventured beyond two-player zero-sum games. [25] built on the scalable meta-strategy solver of α-rank [27]. [21] extended PSRO to the n-player general-sum case with theoretical convergence guarantees to CCE but does not scale up to domains with complex action and observation spaces.

6 LIMITATIONS

A common limitation of equilibrium convergent population learning algorithms is the lack of guarantee on which equilibrium the population would converge to. This applies to our work too (Section 4.1). While we can expect convergence to an equilibrium, we cannot predict if the solution would be *desirable* (e.g. that it is social-welfare maximising). In competitive games, equilibrium selection is less critical as NEs are interchangeable. This is in contrast to general-sum games where players might find themselves significantly less well off in some equilibria than others. Computationally, while NeuPL-JPSRO addressed many of the bottlenecks of prior methods, it remains challenging to scale up to *many* players each with *many* strategies due to the size of the payoff tensor. A future direction would be to consider sample-based equilibria solutions, without needing to tabulate the entire payoff tensor upfront.

7 CONCLUSIONS

We proposed NeuPL-JPSRO as an efficient and scalable algorithm to solving n-player general-sum EF games that provably converges to a NF CCE. We verified its convergence empirically, across diverse test domains ranging from research strategy games to games that require deep RL. We showed that NeuPL-JPSRO can adapt to diverse co-players in non-zero-sum settings and demonstrated the importance of transfer learning in solving games with transferable skills. Our method is computationally accessible, paving the way for deploying game-theoretic solutions to real-world general games.

REFERENCES

[1] Robert J Aumann. 1974. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics* 1, 1 (1974), 67–96.

[2] Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. 2018. Emergent Complexity via Multi-Agent Competition. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sy0GnUxCb

[3] Siddharth Barman and Katrina Ligett. 2015. Finding any nontrivial coarse correlated equilibrium is hard. *ACM SIGecom Exchanges* 14, 1 (2015), 76–79.

[4] George W. Brown. 1951. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation* (1951). arXiv:13(1):374–376 [cs.LG]

[5] Noam Brown and Tuomas Sandholm. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359, 6374 (2018), 418–424.

[6] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence* 134, 1-2 (2002), 57–83.

[7] Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge university press.

[8] Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. 2021. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems* 34 (2021), 27604–27616.

[9] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. 2009. The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39, 1 (2009), 195–259.

[10] Marta Garnelo, Wojciech Marian Czarnecki, Siqi Liu, Dhruva Tirumala, Junhyuk Oh, Gauthier Gidel, Hado van Hasselt, and David Balduzzi. 2021. Pick Your Battles: Interaction Graphs as Population-Level Objectives for Strategic Diversity. In *AAMAS*.

[11] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 364, 6443 (2019), 859–865.

[12] H. W. Kuhn and AW Tucker. 1957. Extensive games and the problem and information. *Contributions to the Theory of Games, II, Annals of Mathematical Studies* 28 (1957), 193–216.

[13] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *CoRR* abs/1908.09453 (2019). arXiv:1908.09453 [cs.LG] http://arxiv.org/abs/1908.09453

[14] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3323fe11e9595c09af38fe67567a9394-Paper.pdf

[15] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *International Conference on Machine Learning*. PMLR, 6187–6199.

[16] Siqi Liu, Marc Lanctot, Luke Marris, and Nicolas Heess. 2022. Simplex Neural Population Learning: Any-Mixture Bayes-Optimality in Symmetric Zero-sum Games. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 13793–13806. https://proceedings.mlr.press/v162/liu22h.html

[17] Siqi Liu, Guy Lever, Nicholas Heess, Josh Merel, Saran Tunyasuvunakool, and Thore Graepel. 2019. Emergent Coordination Through Competition. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BkG8sjR5Km

[18] Siqi Liu, Guy Lever, Zhe Wang, Josh Merel, SM Ali Eslami, Daniel Hennes, Wojciech M Czarnecki, Yuval Tassa, Shayegan Omidshafiei, Abbas Abdolmaleki, et al. 2022. From motor control to team play in simulated humanoid football. *Science Robotics* 7, 69 (2022), eabo0235.

[19] Siqi Liu, Luke Marris, Daniel Hennes, Josh Merel, Nicolas Heess, and Thore Graepel. 2022. NeuPL: Neural Population Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=MIX3fJkl_1

[20] Siqi Liu, Luke Marris, Marc Lanctot, Georgios Piliouras, Joel Z. Leibo, and Nicolas Heess. 2024. Neural Population Learning beyond Symmetric Zero-sum Games. arXiv:2401.05133 [cs.AI]

[21] Luke Marris, Paul Muller, Marc Lanctot, Karl Tuyls, and Thore Graepel. 2021. Multi-Agent Training beyond Zero-Sum with Correlated Equilibrium Meta-Solvers. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 7480–7491. http://proceedings.mlr.press/v139/marris21a.html

[22] Stephen Mcaleer, JB Lanier, Roy Fox, and Pierre Baldi. 2020. Pipeline PSRO: A Scalable Approach for Finding Approximate Nash Equilibria in Large Games. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 20238–20248. https://proceedings.neurips.cc/paper/2020/file/e9bcd1b063077573285ae1a41025f5dc-Paper.pdf

[23] H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. 2003. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning* (Washington, DC, USA, 2003-08-21) *(ICML'03)*. AAAI Press, 536–543.

[24] Hervé Moulin and J-P Vial. 1978. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory* 7, 3 (1978), 201–221.

[25] Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. 2020. A Generalized Training Approach for Multiagent Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkl5kxrKDr

[26] John Nash. 1951. Non-Cooperative Games. *Annals of Mathematics* 54, 2 (1951), 286–295. http://www.jstor.org/stable/1969529

[27] Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. 2019. α-rank: Multi-agent evaluation by evolution. *Scientific reports* 9, 1 (2019), 1–29.

[28] Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. 2022. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science* 378, 6623 (2022), 990–996.

[29] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. 2021. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*. PMLR, 8525–8535.

[30] Sheldon M Ross. 1971. Goofspiel—the game of pure strategy. *Journal of Applied Probability* 8, 3 (1971), 621–625.

[31] Arthur L Samuel. 1967. Some studies in machine learning using the game of checkers. II—Recent progress. *IBM Journal of research and development* 11, 6 (1967), 601–617.

[32] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*. PMLR, 4528–4537.

[33] Bobak Shahriari, Abbas Abdolmaleki, Arunkumar Byravan, Abe Friesen, Siqi Liu, Jost Tobias Springenberg, Nicolas Heess, Matt Hoffman, and Martin Riedmiller. 2022. Revisiting Gaussian mixture critics in off-policy reinforcement learning: a sample-based approach. *arXiv preprint arXiv:2204.10256* (2022).

[34] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[35] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[36] Max Smith, Thomas Anthony, and Michael Wellman. 2020. Iterative Empirical Game Solving via Single Policy Best Response. In *International Conference on Learning Representations*.

[37] Yuval Tassa, Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy P. Lillicrap, and Nicolas Manfred Otto Heess. 2020. dm_control: Software and Tasks for Continuous Control. *Softw. Impacts* 6 (2020), 100022. https://api.semanticscholar.org/CorpusID:219980295

[38] Gerald Tesauro et al. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[39] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.