Open-ended Learning in Symmetric Zero-sum Games

David Balduzzi [1]   **Marta Garnelo** [1]   **Yoram Bachrach** [1]   **Wojciech M. Czarnecki** [1]   **Julien Perolat** [1]
Max Jaderberg [1]   **Thore Graepel** [1]

Abstract

Zero-sum games such as chess and poker are, abstractly, functions that evaluate pairs of agents, for example labeling them 'winner' and 'loser'. If the game is approximately transitive, then self-play generates sequences of agents of increasing strength. However, nontransitive games, such as rock-paper-scissors, can exhibit strategic cycles, and there is no longer a clear objective – we want agents to increase in strength, but against whom is unclear. In this paper, we introduce a geometric framework for formulating agent objectives in zero-sum games, in order to construct adaptive sequences of objectives that yield open-ended learning. The framework allows us to reason about population performance in nontransitive games, and enables the development of a new algorithm (rectified Nash response, PSRO_{rN}) that uses game-theoretic niching to construct diverse populations of effective agents, producing a stronger set of agents than existing algorithms. We apply PSRO_{rN} to two highly nontransitive resource allocation games and find that PSRO_{rN} consistently outperforms the existing alternatives.

1. Introduction

A story goes that a Cambridge tutor in the mid-19th century once proclaimed: "I'm teaching the smartest boy in Britain." His colleague retorted: "I'm teaching the best test-taker." Depending on the version of the story, the first boy was either Lord Kelvin or James Clerk Maxwell. The second boy, who indeed scored highest on the Tripos, is long forgotten.

Modern learning algorithms are outstanding test-takers: once a problem is packaged into a suitable objective, deep (reinforcement) learning algorithms often find a good solution. However, in many multi-agent domains, the question

of what test to take, or what objective to optimize, is not clear. This paper proposes algorithms that adaptively and continually pose new, useful objectives which result in open-ended learning in two-player zero-sum games. This setting has a large scope of applications and is general enough to include function optimization as a special case.

Learning in games is often conservatively formulated as training agents that tie or beat, on average, a fixed set of opponents. However, the dual task, that of *generating useful opponents to train and evaluate against*, is under-studied. It is not enough to beat the agents you know; it is also important to generate better opponents, which exhibit behaviours that you don't know.

There are very successful examples of algorithms that pose and solve a series of increasingly difficult problems for themselves through forms of self-play (Silver et al., 2018; Jaderberg et al., 2018; Bansal et al., 2018; Tesauro, 1995). Unfortunately, it is easy to encounter nontransitive games where self-play cycles through agents without improving overall agent strength – simultaneously improving against one opponent and worsening against another. In this paper, we develop a mathematical framework for analyzing nontransitive games, and present algorithms that systematically uncover and solve the latent problems embedded in a game.

Overview. The paper starts in Section 2 by introducing functional-form games (FFGs) as a new mathematical model of zero-sum games played by parametrized agents such as neural networks. Theorem 1 decomposes any FFG into a sum of transitive and cyclic components. Transitive games, and closely related monotonic games, are the natural setting for self-play, but the cyclic components, present in non-transitive games, require more sophisticated algorithms which motivates the remainder of the paper.

The main problem in tackling nontransitive games, where there is not necessarily a best agent, is understanding what the objective should be. In Section 3, we formulate the global objective in terms of **gamescapes** – convex polytopes that encode the interactions between agents in a game. If the game is transitive or monotonic, then the gamescape degenerates to a one-dimensional landscape. In nontransitive games, the gamescape can be high-dimensional because training against one agent can be fundamentally different

---

[1]DeepMind. Correspondence to: <dbalduzzi@google.com>.

from training against another.

Measuring the performance of individual agents is vexed in nontransitive games. Therefore, in Section 3, we develop tools to analyze populations of agents, including a population-level measure of performance, definition 3. An important property of population-level performance is that it increases transitively as the gamescape polytope expands in a nontransitive game. Thus, we reformulate the problem of learning in games from finding the best agent to growing the gamescape. We consider two approaches to do so, one directly performance related, and the other focusing on a measure of diversity, definition 4. Crucially, the measure quantifies diverse *effective behaviors* – we are not interested in differences in policies that do not lead to differences in outcomes, nor in agents that lose in new and surprising ways.

Section 4 presents two algorithms, one old and one new, for growing the gamescape. The algorithms can be seen as specializations of the policy space response oracle (PSRO) introduced in Lanctot et al. (2017). The first algorithm is Nash response (PSRO$_N$), which is an extension to functional-form games of the double oracle algorithm from McMahan et al. (2003). Given a population, Nash response creates an objective to train against by averaging over the Nash equilibrium. The Nash serves as a proxy for the notion of 'best agent', which is not guaranteed to exist in general zero-sum games. A second, complementary algorithm is the rectified Nash response (PSRO$_{rN}$). The algorithm amplifies strategic diversity in populations of agents by adaptively constructing *game-theoretic niches* that encourage agents to 'play to their strengths and ignore their weaknesses'.

Finally, in Section 5, we investigate the performance of these algorithms in Colonel Blotto (Borel, 1921; Tukey, 1949; Roberson, 2006) and a differentiable analog we refer to as differentiable Lotto. Blotto-style games involve allocating limited resources, and are highly nontransitive. We find that PSRO$_{rN}$ outperforms PSRO$_N$, both of which greatly outperform self-play in these domains. We also compare against an algorithm that responds to the uniform distribution PSRO$_U$, which performs comparably to PSRO$_N$.

Related work. There is a large literature on novelty search, open-ended evolution, and curiosity, which aim to continually expand the frontiers of game knowledge within an agent (Lehman & Stanley, 2008; Taylor et al., 2016; Banzhaf et al., 2016; Brant & Stanley, 2017; Pathak et al., 2017; Wang et al., 2019). A common thread is that of *adaptive objectives* which force agents to keep improving. For example, in novelty search, the target objective constantly changes – and so cannot be reduced to a fixed objective to be optimized once-and-for-all.

We draw heavily on prior work on learning in games, es-

pecially Heinrich et al. (2015); Lanctot et al. (2017) which are discussed below. Our setting resembles multiobjective optimization (Fonseca & Fleming, 1993; Miettinen, 1998). However, unlike multiobjective optimization, we are concerned with *both generating and optimizing* objectives. Generative adversarial networks (Goodfellow et al., 2014) are zero-sum games that do *not* fall under the scope of this paper due to lack of symmetry, see appendix **??**.

Notation. Vectors are columns. The constant vectors of zeros and ones are $\mathbf{0}$ and $\mathbf{1}$. We sometimes use $\mathbf{p}[i]$ to denote the ith entry of vector $\mathbf{p}$. Proofs are in the appendix.

2. Functional-form games (FFGs)

Suppose that, given any pair of agents, we can compute the probability of one beating the other in a game such as Go, Chess, or StarCraft. We formalize the setup as follows.

Definition 1. *Let W be a set of agents parametrized by, say, the weights of a neural net. A **symmetric zero-sum functional-form game** (FFG) is an* antisymmetric *function, $\phi(\mathbf{v}, \mathbf{w}) = -\phi(\mathbf{w}, \mathbf{v})$, that evaluates pairs of agents*

$$\phi : W \times W \to \mathbb{R}.$$

The higher $\phi(\mathbf{v}, \mathbf{w})$, the better for agent $\mathbf{v}$. We refer to $\phi > 0$, $\phi < 0$, and $\phi = 0$ as wins, losses and ties for $\mathbf{v}$.

Note that (i) the strategies in a FFG are *parametrized agents* and (ii) the parametrization of the agents is folded into ϕ, so the game is a composite of the agent's architecture and the environment itself.

Suppose the probability of $\mathbf{v}$ beating $\mathbf{w}$, denoted $P(\mathbf{v} \succ \mathbf{w})$ can be computed or estimated. Win/loss probabilities can be rendered into antisymmetric form via $\phi(\mathbf{v}, \mathbf{w}) := P(\mathbf{v} \succ \mathbf{w}) - \frac{1}{2}$ or $\phi(\mathbf{v}, \mathbf{w}) := \log \frac{P(\mathbf{v} \succ \mathbf{w})}{P(\mathbf{v} \prec \mathbf{w})}$.

Tools for FFGs. Solving FFGs requires different methods to solving normal form games (Shoham & Leyton-Brown, 2008) due to their continuous nature. We therefore develop the following basic tools.

First, the **curry** operator converts a two-player game into a *function from agents to objectives*

$$\left[\phi : W \times W \longrightarrow \mathbb{R} \right] \xrightarrow{\text{curry}} \left[W \longrightarrow \left[W \longrightarrow \mathbb{R} \right] \right]$$
$$\phi(\mathbf{v}, \mathbf{w}) \qquad\qquad \mathbf{w} \mapsto \phi_{\mathbf{w}}(\bullet) := \phi(\bullet, \mathbf{w})$$

Second, an **approximate best-response oracle** that, given agent $\mathbf{v}$ and objective $\phi_{\mathbf{w}}(\bullet)$, returns a new agent $\mathbf{v}' := \text{oracle}(\mathbf{v}, \phi_{\mathbf{w}}(\bullet))$ with $\phi_{\mathbf{w}}(\mathbf{v}') > \phi_{\mathbf{w}}(\mathbf{v}) + \epsilon$, if possible. The oracle could use gradients, reinforcement learning or evolutionary algorithms.

Third, given a population $\mathfrak{P}$ of n agents, the $(n \times n)$ anti-

symmetric **evaluation matrix** is

$$\mathbf{A}_{\mathfrak{P}} := \left\{ \phi(\mathbf{w}_i, \mathbf{w}_j) \; : \; (\mathbf{w}_i, \mathbf{w}_j) \in \mathfrak{P} \times \mathfrak{P} \right\} =: \phi(\mathfrak{P} \otimes \mathfrak{P}).$$

Fourth, we will use the (not necessarily unique) **Nash equilibrium** on the zero-sum matrix game specified by $\mathbf{A}_{\mathfrak{P}}$.

Finally, we use the following **game decomposition.** Suppose W is a compact set equipped with a probability measure. The set of integrable antisymmetric functions on W then forms a vector space. Appendix D shows the following:

Theorem 1 (game decomposition). *Every* FFG *decomposes into a sum of a transitive and cyclic game*

$$\text{FFG} = \textit{transitive game} \oplus \textit{cyclic game}.$$

with respect to a suitably defined inner product.

Transitive and cyclic games are discussed below. Few games are purely transitive or cyclic. Nevertheless, understanding these cases is important since general algorithms should, at the very least, work in both special cases.

2.1. Transitive games

A game is **transitive** if there is a 'rating function' f such that performance on the game is the difference in ratings:

$$\phi(\mathbf{v}, \mathbf{w}) = f(\mathbf{v}) - f(\mathbf{w}).$$

In other words, if ϕ admits a 'subtractive factorization'.

Optimization (training against a fixed opponent). Solving a transitive game reduces to finding

$$\mathbf{v}^* := \underset{\mathbf{v} \in W}{\operatorname{argmax}} \, \phi_{\mathbf{w}}(\mathbf{v}) = \underset{\mathbf{v} \in W}{\operatorname{argmax}} \, f(\mathbf{v}).$$

Crucially, the choice of opponent $\mathbf{w}$ makes no difference to the solution. The simplest learning algorithm is thus to train against a fixed opponent, see algorithm 1.

Algorithm 1 Optimization (against a fixed opponent)

  input: opponent $\mathbf{w}$; agent $\mathbf{v}_1$
  fix objective $\phi_{\mathbf{w}}(\bullet)$
  for $t = 1, \ldots, T$ **do**
    $\mathbf{v}_{t+1} \leftarrow \text{oracle}\left(\mathbf{v}_t, \phi_{\mathbf{w}}(\bullet)\right)$
  end for
  output: $\mathbf{v}_{T+1}$

Monotonic games generalize transitive games. An FFG is monotonic if there is a <u>monotonic function</u> σ such that

$$\phi(\mathbf{v}, \mathbf{w}) = \sigma\big(f(\mathbf{v}) - f(\mathbf{w})\big). \qquad (1)$$

For example, Elo (1978) models the probability of one agent beating another by

$$P(\mathbf{v} \succ \mathbf{w}) = \sigma\big(f(\mathbf{v}) - f(\mathbf{w})\big) \text{ for } \sigma(x) = \frac{1}{1 + e^{-\alpha \cdot x}}$$

for some $\alpha > 0$, where f assigns Elo ratings to agents. The model is widely used in Chess, Go and other games.

Optimizing against a fixed opponent fares badly in monotonic games. Concretely, if Elo's model holds then <u>training against a much weaker opponent yields no learning signal because the gradient vanishes</u> $\nabla_{\mathbf{v}}\phi(\mathbf{v}_t, \mathbf{w}) \approx 0$ once the sigmoid saturates when $f(\mathbf{v}_t) \gg f(\mathbf{w})$.

Self-play (algorithm 2) <u>generates a sequence of opponents.</u> Training against a sequence of opponents of increasing strength <u>prevents gradients from vanishing due to large skill differentials</u>, so self-play is well-suited to games modeled by eq. (1). Self-play has proven effective in Chess, Go and other games (Silver et al., 2018; Al-Shedivat et al., 2018).

Algorithm 2 Self-play

  input: agent $\mathbf{v}_1$
  for $t = 1, \ldots, T$ **do**
    $\mathbf{v}_{t+1} \leftarrow \text{oracle}\left(\mathbf{v}_t, \phi_{\mathbf{v}_t}(\bullet)\right)$
  end for
  output: $\mathbf{v}_{T+1}$

Self-play *is* an <u>open-ended</u> learning algorithm: it <u>poses and masters a sequence of objectives</u>, rather than optimizing a pre-specified objective. However, <u>self-play assumes transitivity: that local improvements ($\mathbf{v}_{t+1}$ beats $\mathbf{v}_t$) imply global improvements ($\mathbf{v}_{t+1}$ beats $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t$).</u> The assumption fails in nontransitive games, such as the disc game below. Since performance is nontransitive, <u>improving against one agent does not guarantee improvements against others.</u>

2.2. Cyclic games

A game is **cyclic** if

$$\int_W \phi(\mathbf{v}, \mathbf{w}) \cdot d\mathbf{w} = 0 \quad \text{for all} \quad \mathbf{v} \in W. \qquad (2)$$

In other words, <u>wins against some agents are necessarily counterbalanced with losses against others</u>. Strategic cycles often arise when <u>agents play simultaneous move</u> or <u>imperfect information games</u> such as rock-paper-scissors, poker, or StarCraft.

Example 1 (Disc game). *Fix $k > 0$. Agents are $W = \{\mathbf{x} \in \mathbb{R}^2 \; : \; \|\mathbf{x}\|_2^2 \le k\}$ with the uniform distribution. Set*

$$\phi(\mathbf{v}, \mathbf{w}) = \mathbf{v}^{\intercal} \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \mathbf{w} = v_1 w_2 - v_2 w_1.$$

The game is cyclic, see figure 2A.

Example 2 (Rock-paper-scissors embeds in disc game). *Set* $\mathfrak{r}_\epsilon = \frac{\sqrt{3}\epsilon}{2}(\cos 0, \sin 0)$, $\mathfrak{p}_\epsilon = \frac{\sqrt{3}\epsilon}{2}(\cos \frac{2\pi}{3}, \sin \frac{2\pi}{3})$ *and* $\mathfrak{s}_\epsilon = \frac{\sqrt{3}\epsilon}{2}(\cos \frac{4\pi}{3}, \sin \frac{4\pi}{3})$ *to obtain*

$$\mathbf{A}_{\{\mathfrak{r}_\epsilon, \mathfrak{p}_\epsilon, \mathfrak{s}_\epsilon\}} = \begin{bmatrix} 0 & \epsilon^2 & -\epsilon^2 \\ -\epsilon^2 & 0 & \epsilon^2 \\ \epsilon^2 & -\epsilon^2 & 0 \end{bmatrix}.$$

Varying $\epsilon \in [0, 1]$ yields a family of $\mathfrak{r}$-$\mathfrak{p}$-$\mathfrak{s}$ interactions that trend deterministic as ϵ increases, see figure 2B.

Our goal is to extend self-play to general FFGs. The success of optimization and self-play derives from **(i)** repeatedly applying a local operation that **(ii)** improves a transitive measure. If the measure is *not* transitive, then applying a sequence of local improvements can result in no improvement at all. Our goal is thus to find practical substitutes for **(i)** and **(ii)** in general FFGs.

3. Functional and Empirical Gamescapes

Rather than trying to find a single dominant agent which may not exist, we seek to find all the atomic components in "strategy space" of a zero-sum game. That is, we aim to discover the underlying strategic dimensions of the game, and the best ways of executing them. Given such knowledge, when faced with a new opponent, we will not only be able to react to its behavior conservatively (using the Nash mixture to guarantee a tie), but will also be able to optimally exploit the opponent. As opposed to typical game-theoretic solutions, we do not seek a single agent or mixture, but rather a population that embodies a complete understanding of the strategic dimensions of the game.

To formalize these ideas we introduce **gamescapes**, which geometrically represent agents in functional form games. We show some general properties of these objects to build intuitions for the reader. Finally we introduce two critical concepts: **population performance**, which measures the progress in performance of populations, and **effective diversity**, which quantifies the coverage of the gamescape spanned by a population. Equipped with these tools we present algorithms that guarantee iterative improvements in FFGs.

Definition 2. *The **functional gamescape** (FGS) of $\phi : W \times W \to \mathbb{R}$ is the convex set*

$$\mathcal{G}_\phi := \text{hull}\left(\left\{\phi_{\mathbf{w}}(\bullet) : \mathbf{w} \in W\right\}\right) \subset \mathcal{C}(W, \mathbb{R}),$$

where $\mathcal{C}(W, \mathbb{R})$ is the space of real-valued functions on W.

*Given population $\mathfrak{P}$ of n agents with evaluation matrix $\mathbf{A}_\mathfrak{P}$, the corresponding **empirical gamescape** (EGS) is*

$$\mathcal{G}_\mathfrak{P} := \left\{\text{convex mixtures of rows of } \mathbf{A}_\mathfrak{P}\right\}.$$

The FGS represents all the mixtures of objectives implicit in the game. We cannot work with the FGS directly because we cannot compute $\phi_{\mathbf{w}}(\bullet)$ for infinitely many agents. The EGS is a tractable proxy (Wellman, 2006). The two gamescapes represent all the ways agents can-in-principle and are-actually-observed-to interact respectively. The remainder of this section collects basic facts about gamescapes.

Optimization landscapes are a special case of gamescapes. If $\phi(\mathbf{v}, \mathbf{w}) = f(\mathbf{v}) - f(\mathbf{w})$ then the FGS is, modulo constant terms, a single function $\mathcal{G}_\phi = \left\{f(\bullet) - f(\mathbf{w}) : \mathbf{w} \in W\right\}$. The FGS degenerates into a landscape where, for each agent $\mathbf{v}$ there is a unique direction $\nabla\phi_{\mathbf{w}}(\mathbf{v}) = \nabla f(\mathbf{v})$ in weight space which gives the steepest performance increase *against all opponents*. In a monotonic game, the gradient is $\nabla\phi_{\mathbf{w}}(\mathbf{v}) = \sigma' \cdot \nabla f(\mathbf{v})$. There is again a single steepest direction $\nabla f(\mathbf{v})$, with tendency to vanish controled by the ratings differential $\sigma' = \sigma'\big(f(\mathbf{v}) - f(\mathbf{w})\big) \geq 0$.

Redundancy. First, we argue that gamescapes are more fundamental than evaluation matrices. Consider

$$\begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 1 & -1 & -1 \\ -1 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix}.$$

The first matrix encodes rock-paper-scissors interactions; the second is the same, but with two copies of scissors. The matrices are difficult to compare since their dimensions are different. Nevertheless, the gamescapes are equivalent triangles embedded in $\mathbb{R}^3$ and $\mathbb{R}^4$ respectively.

Proposition 2. *An agent in a population is redundant if it is behaviorally identical to a convex mixture of other agents. The EGS is **invariant** to redundant agents.*

Invariance is explained in appendix E.

Dimensionality. The dimension of the gamescape is an indicator of the complexity of both the game and the agents playing it. In practice we find many FFGs have a low dimensional latent structure.

Figure 1 depicts evaluation matrices of four populations of 40 agents. Although the gamescapes could be 40-dim, they turn out to have one- and two-dim representations. The dimension of the EGS is determined by the rank of the evaluation matrix.

Proposition 3. *The EGS of n agents in population $\mathfrak{P}$ can be represented in $\mathbb{R}^r$, where $r = \text{rank}(\mathbf{A}_\mathfrak{P}) \leq n$.*

A low-dim representation of the EGS can be constructed via the Schur decomposition, which is the analog of PCA for antisymmetric matrices (Balduzzi et al., 2018b). The length of the longest strategic cycle in a game gives a lower-bound on the dimension of its gamescape:

Example 3 (latent dimension of long cycles). *Suppose n agents form a long cycle: $\mathfrak{P} = \{\mathbf{v}_1 \xrightarrow{\text{beats}} \mathbf{v}_2 \to \cdots \to \mathbf{v}_n \xrightarrow{\text{beats}} \mathbf{v}_1\}$. Then $\text{rank}(\mathbf{A}_\mathfrak{P})$ is $n - 2$ if n is even and $n - 1$ if n is odd.*

Nash equilibria in a symmetric zero-sum game are (mixtures of) agents that beat or tie all other agents. Loosely speaking, they replace the notion of best agent in games
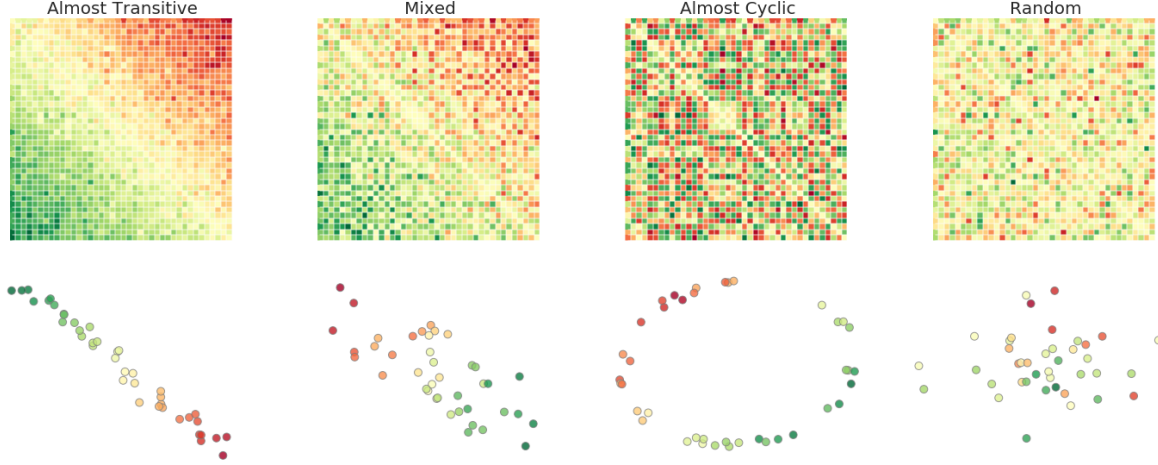
Figure 1. *Low-dim gamescapes of various basic game structures.* **Top row:** Evaluation matrices of populations of 40 agents each; colors vary from red to green as ϕ ranges over $[-1, 1]$. **Bottom row:** 2-dim embedding obtained by using first 2 dimensions of Schur decomposition of the payoff matrix; Color corresponds to average payoff of an agent against entire population; EGS of the transitive game is a line; EGS of the cyclic game is two-dim near-circular polytope given by convex hull of points. For extended version see Figure 6 in the Appendix.

where there is no best agent. Functional Nash equilibria, in the FGS, are computationally intractable so we work with empirical Nash equilibria over the evaluation matrix.

Proposition 4. *Given population $\mathfrak{P}$, the empirical Nash equilibria are*

$$\mathcal{N}_{\mathfrak{P}} = \{\mathbf{p} \text{ distribution} : \mathbf{p}^{\mathsf{T}} \mathbf{A}_{\mathfrak{P}} \succeq \mathbf{0}\}.$$

In other words, Nash equilibria correspond to points in the empirical gamescape that intersect the positive quadrant $\{\mathbf{x} \in \mathcal{G}_{\mathfrak{P}} : \mathbf{x} \succeq \mathbf{0}\}$. The positive quadrant thus provides a set of directions in weight space to aim for when training new agents, see PSRO$_N$ below.

The gap between the EGS and FGS. Observing $\mathfrak{r}$-$\mathfrak{p}$ interactions yields different conclusions from observing $\mathfrak{r}$-$\mathfrak{p}$-$\mathfrak{s}$ interactions; it is always possible that an agent that appears to be dominated is actually part of a partially observed cycle. Without further assumptions about the structure of ϕ, it is impossible to draw strong conclusions about the nature of the FGS from the EGS computed from a finite population. The gap is analogous to the exploration problem in reinforcement learning. To discover unobserved dimensions of the FGS one could train against randomized distributions over opponents, which would eventually find them all.

3.1. Population performance

If $\phi(\mathbf{v}, \mathbf{w}) = f(\mathbf{v}) - f(\mathbf{w})$ then improving performance of agent $\mathbf{v}$ reduces to increasing $f(\mathbf{v})$. In a cyclic game, the performance of individual agents is meaningless: beating one agent entails losing against another by eq. (2). We therefore propose a *population* performance measure.

Definition 3. *Given populations $\mathfrak{P}$ and $\mathfrak{Q}$, let $(\mathbf{p}, \mathbf{q})$ be a Nash equilibrium of the zero-sum game on $\mathbf{A}_{\mathfrak{P},\mathfrak{Q}} := \phi(\mathbf{v}, \mathbf{w})_{\mathbf{v}\in\mathfrak{P},\mathbf{w}\in\mathfrak{Q}}$. The **relative population performance** is*

$$v(\mathfrak{P}, \mathfrak{Q}) := \mathbf{p}^{\mathsf{T}} \cdot \mathbf{A}_{\mathfrak{P},\mathfrak{Q}} \cdot \mathbf{q} = \sum_{i,j=1}^{n_1,n_2} A_{ij} \cdot p_i q_j.$$

Proposition 5. *(i) Performance v is independent of the choice of Nash equilibrium. (ii) If ϕ is monotonic then performance compares the best agents in each population*

$$v(\mathfrak{P}, \mathfrak{Q}) = \max_{\mathbf{v}\in\mathfrak{P}} f(\mathbf{v}) - \max_{\mathbf{w}\in\mathfrak{Q}} f(\mathbf{w}).$$

*(iii) If $\text{hull}(\mathfrak{P}) \subset \text{hull}(\mathfrak{Q})$ then $v(\mathfrak{P}, \mathfrak{Q}) \leq 0$ and $v(\mathfrak{P}, \mathfrak{R}) \leq v(\mathfrak{Q}, \mathfrak{R})$ for **any** population $\mathfrak{R}$.*

The first two properties are sanity checks. Property *(iii)* implies growing the polytope spanned by a population improves its performance *against any other population*.

Consider the concentric rock-paper-scissors populations in figure 2B and example 2. The Nash equilibrium is $(0, 0)$, which is a uniform mixture over any of the populations. Thus, the relative performance of any two populations is zero. However, the outer population is better than the inner population at *exploiting* an opponent that only plays, say, rock because the outer version of paper wins more deterministically than the inner version.

Finding a population that contains the Nash equilibrium is necessary but not sufficient to fully solve an FFG. For example, adding the ability to always force a tie to an FFG makes finding the Nash trivial. However, the game can still exhibit rich strategies and counter-strategies that are worth discovering.
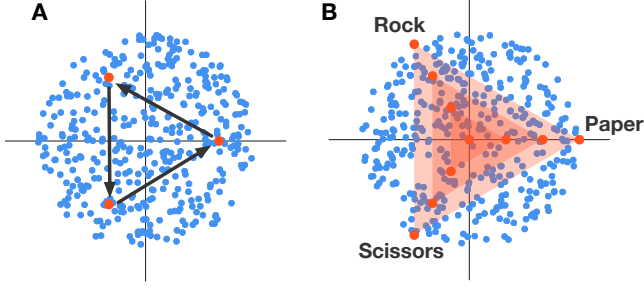
Figure 2. The disc game. **A:** A set of possible agents from the disc game is shown as blue dots. Three agents with non-transitive rock-paper-scissors relations are visualized in red. **B:** Three concentric gamescapes spanned by populations with rock-paper-scissor interactions of increasing strength.

3.2. Effective diversity

Measures of diversity typically quantify differences in weights or behavior of agents but ignore performance. *Effective* diversity measures the variety of effective agents (agents with support under Nash):

Definition 4. *Denote the rectifier by* $\lfloor x \rfloor_+ := x$ *if* $x \geq 0$ *and* $\lfloor x \rfloor_+ := 0$ *otherwise. Given population* $\mathfrak{P}$, *let* $\mathbf{p}$ *be a Nash equilibrium on* $\mathbf{A}_\mathfrak{P}$. *The **effective diversity** of the population is:*

$$d(\mathfrak{P}) := \mathbf{p}^\mathsf{T} \cdot \lfloor \mathbf{A}_\mathfrak{P} \rfloor_+ \cdot \mathbf{p} = \sum_{i,j=1}^{n} \lfloor \phi(\mathbf{w}_i, \mathbf{w}_j) \rfloor_+ \cdot p_i p_j.$$

Diversity quantifies how the best agents (those with support in the maximum entropy Nash) exploit each other. If there is a dominant agent then diversity is zero.

Effective diversity is a matrix norm, see appendix F.2. It measures the $\ell_{1,1}$ volume spanned by Nash supported agents. In figure 2B, there are four populations spanning concentric gamescapes: the Nash at $(0,0)$ and three variants of r-p-s. Going outwards to large gamescapes yields agents that are more diverse and better exploiters.

4. Algorithms

We now turn attention to constructing objectives that when trained against, produce new, effective agents. We present two algorithms that construct a sequence of fruitful local objectives that, when solved, iteratively add up to transitive population-level progress. Importantly, these algorithms output populations, unlike self-play which outputs single agents.

Concretely, we present algorithms that expand the empirical gamescape in useful directions. Following Lanctot et al. (2017), we assume access to a subroutine, or **oracle**, that finds an approximate best response to any mixture $\sum_i p_i \phi_i(\mathbf{w}_i)$ of objectives. The subroutine could be a gradient-based, reinforcement learning or evolutionary algorithm. The subroutine returns a vector in weight-space, in which existing agents can be shifted to create new agents. Any mixture constitutes a valid training objective. However, many mixtures do not grow the gamescape, because the vector could point towards redundant or weak agents.

4.1. Response to the Nash (PSRO$_\mathsf{N}$)

Since the notion of 'the best agent' – one agent that beast all others – does not necessarily exist in nontransitive games, a natural substitute is the mixture over the Nash equilibrium on the most recent population $\mathfrak{P}_t$. The policy space response to the Nash (PSRO$_\mathsf{N}$) iteratively generates new agents that are approximate best responses to the Nash mixture. If the game is transitive then PSRO$_\mathsf{N}$ degenerates to self-play. The algorithm is an extension of the double oracle algorithm (McMahan et al., 2003) to FFGs, see also (Zinkevich et al., 2007; Hansen et al., 2008).

Algorithm 3 Response to Nash (PSRO$_\mathsf{N}$)

  input: population $\mathfrak{P}_1$ of agents
  for $t = 1, \ldots, T$ **do**
    $\mathbf{p}_t \leftarrow$ Nash on $\mathbf{A}_{\mathfrak{P}_t}$
    $\mathbf{v}_{t+1} \leftarrow$ oracle $\left(\mathbf{v}_t, \sum_{\mathbf{w}_i \in \mathfrak{P}_t} \mathbf{p}_t[i] \cdot \phi_{\mathbf{w}_i}(\bullet)\right)$
    $\mathfrak{P}_{t+1} \leftarrow \mathfrak{P}_t \cup \{\mathbf{v}_{t+1}\}$
  end for
  output: $\mathfrak{P}_{T+1}$

The following result shows that PSRO$_\mathsf{N}$ strictly enlarges the empirical gamescape:

Proposition 6. *If* $\mathbf{p}$ *is a Nash equilibrium on* $\mathbf{A}_\mathfrak{P}$ *and* $\sum_i p_i \phi_{\mathbf{w}_i}(\mathbf{v}) > 0$, *then adding* $\mathbf{v}$ *to* $\mathfrak{P}$ *strictly enlarges the empirical gamescape:* $\mathcal{G}_\mathfrak{P} \subsetneq \mathcal{G}_{\mathfrak{P} \cup \{\mathbf{v}\}}$.

A failure mode of PSRO$_\mathsf{N}$ arises when the Nash equilibrium of the game is contained in the empirical gamescape. For example, in the disc game in figure 2 the Nash equilibrium of the entire FFG is the agent at the origin $\mathbf{w} = (0,0)$. If a population's gamescape contains $\mathbf{w} = (0,0)$ – which is true of any r-p-s subpopulation – then PSRO$_\mathsf{N}$ will not expand the gamescape because there is no ϵ-better response to $\mathbf{w} = (0,0)$. The next section presents an algorithm that uses *niching* to meaningfully grow the gamescape, even after finding the Nash equilibrium of the FFG.

Response to the uniform distribution (PSRO$_\mathsf{U}$). A closely related algorithm is fictitious (self-)play (Brown, 1951; Leslie & Collins, 2006; Heinrich et al., 2015). The algorithm finds an approximate best-response to the uniform distribution on agents in the current population: $\sum_{\mathbf{w}_i \in \mathfrak{P}_t} \phi_{\mathbf{w}_i}(\bullet)$. PSRO$_\mathsf{U}$ has guarantees in matrix form games and performs well empirically (see below). However, we do not currently understand its effect on the gamescape.
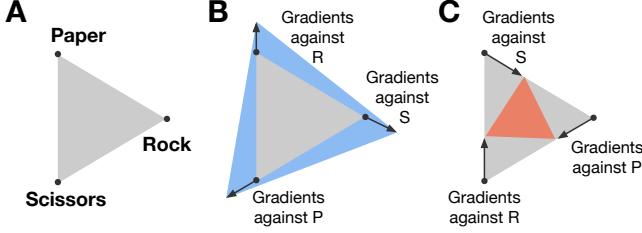
Figure 3. **A:** Rock-paper-scissors. **B:** Gradient updates obtained from PSRO$_\text{rN}$, amplifying strengths, grow gamescape (gray to blue). **C:** Gradients obtained by optimizing agents to reduces their losses shrink gamescape (gray to red).

4.2. Response to the rectified Nash (PSRO$_\text{rN}$)

Response to the *rectified* Nash (PSRO$_\text{rN}$), introduces game-theoretic niches. Each <u>effective agent – that is each agent with support under the Nash equilibrium</u> – is trained against the Nash-weighted mixture of agents that it beats or ties. Intuitively, the idea is to encourage agents to '<u>amplify their strengths and ignore their weaknesses</u>'.

A special case of PSRO$_\text{rN}$ arises when there is a dominant agent in the population, that beats all other agents. The Nash equilibrium is then concentrated on the dominant agent, and PSRO$_\text{rN}$ degenerates to training against the best agent in the population, which can be thought of as a form of self-play (assuming the best agent is the most recent).

Algorithm 4 Response to rectified Nash (PSRO$_\text{rN}$)

  input: population $\mathfrak{P}_1$
  for $t = 1, \ldots, T$ **do**
    $\mathbf{p}_t \leftarrow$ Nash on $\mathbf{A}_{\mathfrak{P}_t}$
    for agent $\mathbf{v}_t$ with positive mass in $\mathbf{p}_t$ **do**
      $\mathbf{v}_{t+1} \leftarrow \text{oracle}\left(\mathbf{v}_t, \sum_{\mathbf{w}_i \in \mathfrak{P}_t} \mathbf{p}_t[i] \cdot \lfloor \phi_{\mathbf{w}_i}(\bullet) \rfloor_+\right)$
    end for
    $\mathfrak{P}_{t+1} \leftarrow \mathfrak{P}_t \cup \{\mathbf{v}_{t+1} : \text{updated above}\}$
  end for
  output: $\mathfrak{P}_{T+1}$

Proposition 7. *The objective constructed by rectified Nash response is effective diversity, definition 4.*

Thus, PSRO$_\text{rN}$ amplifies the positive coordinates, of the Nash-supported agents, in their rows of the evaluation matrix. A pathological mode of PSRO$_\text{rN}$ is when there are many extremely local niches. That is, every agent admits a specific exploit that does not generalize to other agents. PSRO$_\text{rN}$ will <u>grow the gamescape by finding these exploits</u>, generating a large population of highly specialized agents.

Rectified Nash responses in the disc game (example 1)**.** The disc game embeds many subpopulations with rock-paper-scissor dynamics. As the polytope they span expands outwards, the interactions go from noisy to deterministic.
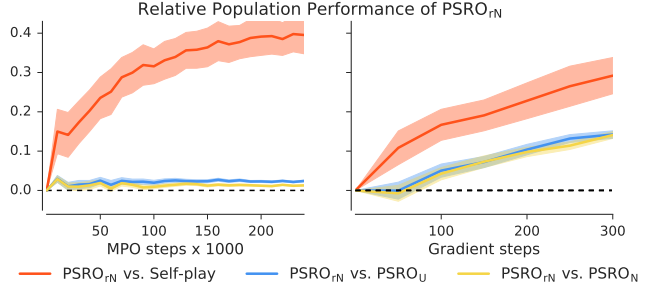

Figure 4. Performance of PSRO$_\text{rN}$ relative to self-play, PSRO$_\text{U}$ and PSRO$_\text{N}$ on Blotto (left) and Differentiable Lotto (right). In all cases, the relative performance of PSRO$_\text{rN}$ is positive, and therefore outperforms the other algorithms.

The disc game is differentiable, so we can use gradient-based learning for the oracle in PSRO$_\text{rN}$. Figure 3B depicts the gradients resulting from training each of rock, paper and scissors against the agent it beats. Since the gradients point *outside* the polytope, training against the rectified Nash mixtures expands the gamescape.

Why ignore weaknesses? A natural modification of PSRO$_\text{rN}$ is to train effective agents against effective agents that they *lose* to. In other words, to force agents to improve their weak points whilst taking their strengths for granted. Figure 3C shows the gradients that would be applied to each of rock, paper and scissors under this algorithm. They point *inwards*, contracting the gamescape. Training rock against paper would make it more like scissors; similarly training paper against scissors would make it more like rock and so on. Perhaps counter-intuitively, building objectives using the weak points of agents does not encourage diverse niches.

5. Experiments

We investigated the performance of the proposed algorithms in two highly nontransitive resource allocation games.

Colonel Blotto is a resource allocation game that is often used as a model for electoral competition. Each of two players has a budget of c coins which they simultaneously distribute over a fixed number of areas. Area a_i is won by the player with more coins on a_i. The player that wins the most areas wins the game. Since Blotto is not differentiable we use maximum a posteriori policy optimization (MPO) (Abdolmaleki et al., 2018) as best response oracle. MPO is an inference-based policy optimization algorithm; many other reinforcement learning algorithms could be used.

Differentiable Lotto is inspired by continuous Lotto (Hart, 2008). The game is defined over a fixed set $\mathcal{C}$ of c 'customers', each being a point in $\mathbb{R}^2$. An agent $(\mathbf{p}, \mathbf{v}) = \{(p_1, \mathbf{v}_1), \ldots, (p_k, \mathbf{v}_k)\}$ distributes one unit of mass over
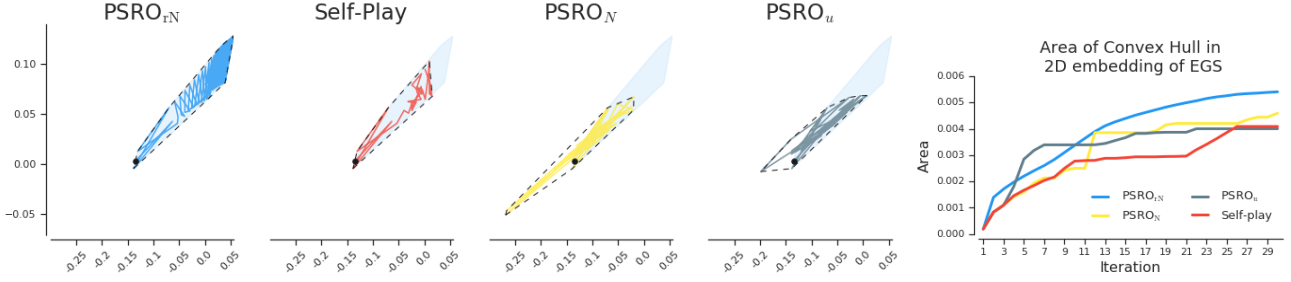
Figure 5. Visualizations of training progress in Differentiable Lotto experiment. **Left:** Comparison of trajectories taken by each algorithm in the 2-dim Schur embedding of the EGS; a black dot represents first agent found by the algorithm and a dashed line represents the convex full. Shaded blue region shows area of the convex hull of PSRO_rN. Notice the PSRO_rN consistent expansion of the convex hull through ladder-like movements. See Figure 7 for an extended version. **Right:** Area of convex hull spanned by populations over time. Note that only PSRO_rN consistently increases the convex hull in all iterations.

k servers, where each server is a point $\mathbf{v}_i \in \mathbb{R}^2$. Roughly, given two agents $(\mathbf{p}, \mathbf{v})$ and $(\mathbf{q}, \mathbf{w})$, customers are softly assigned to the nearest servers, determining the agents' pay-offs. More formally, the payoff is

$$\phi\big((\mathbf{p}, \mathbf{v}), (\mathbf{q}, \mathbf{w})\big) := \sum_{i,j=1}^{c,k} \big(p_j v_{ij} - q_j w_{ij}\big),$$

where the scalars v_{ij} and w_{ij} depend on the distance between customer i and the servers:

$$(v_{i1}, \ldots, w_{ik}) := \mathrm{softmax}(-\|\mathbf{c}_i - \mathbf{v}_1\|^2, \ldots, -\|\mathbf{c}_i - \mathbf{w}_k\|^2).$$

The *width* of a cloud of points is the expected distance from the barycenter. We impose agents to have width equal one. We use gradient ascent as our oracle.

Experimental setups. The experiments examine the performance of self-play, PSRO_U, PSRO_N, and PSRO_rN. We investigate performance under a fixed computational budget. Specifically, we track queries made to the oracle, following the standard model of computational cost in convex optimization (Nemirovski & Yudin, 1983). To compare two algorithms, we report the relative population performance (definition 3), of the populations they output. Computing evaluation matrices is expensive, $\mathcal{O}(n^2)$, for large populations. This cost is not included in our computational model since populations are relatively small. The relative cost of evaluations and queries to the oracle depends on the game.

In Blotto, we investigate performance for $a = 3$ areas and $c = 10$ coins over $k = 1000$ games. An agent outputs a vector in $\mathbb{R}^3$ which is passed to a softmax, $\times 10$ and discretized to obtain three integers summing to 10. Differentiable Lotto experiments are from $k = 500$ games with $c = 9$ customers chosen uniformly at random in the square $[-1, 1]^2$.

Results. Fig 4 shows the relative population performance, definition 3, between PSRO_rN and each of PSRO_N, PSRO_U and self-play: the more positive the number is, the more

PSRO_rN outperforms the alternative method. We find that PSRO_rN outperforms the other approaches across a wide range of allowed compute budgets. PSRO_U and PSRO_N perform comparably, and self-play performs the worst. Self-play, algorithm 2, outputs a single agent, so the above comparison only considers the final agent. If we upgrade self-play to a population algorithm (by tracking all agents produced over), then it still performs the worst in differentiable Lotto, but by a smaller margin. In Blotto, suprisingly, it slightly outperforms PSRO_N and PSRO_U.

Figure 5 shows how gamescapes develop during training. From the left panel, we see that PSRO_rN grows the polytope in a more uniform manner than the other algorithms. The right panel shows the area of the empirical gamescapes generated by the algorithms (the areas of the convex hulls). All algorithms increase the area, but PSRO_rN is the only method that increases the area at every iteration.

6. Conclusion

We have proposed a framework for open-ended learning in two-player symmetric zero-sum games – where strategies are agents, with a differentiable parametrization. We propose the goal of learning should be **(i)** to discover the underlying strategic components that constitute the game and **(ii)** to master each of them. We formalized these ideas using gamescapes, which geometrically represent the latent objectives in games, and provided tools to analyze them. Finally, we proposed and empirically validated a new algorithm, PSRO_rN, for uncovering strategic diversity within functional form games.

The algorithms discussed here are simple and generic, providing the foundations for methods that unify modern gradient and reinforcement-based learning with the adaptive objectives derived from game-theoretic considerations. Future work lies in expanding this understanding and applying it to develop practical algorithms for more complex games.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. Maximum a Posteriori Policy Optimisation. In *ICLR*, 2018.

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., and Abbeel, P. Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments. In *ICLR*, 2018.

Baker, M. Hodge theory in combinatorics. *Bull. AMS*, 55 (1):57–80, 2018.

Balduzzi, D., Racanière, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. In *ICML*, 2018a.

Balduzzi, D., Tuyls, K., Perolat, J., and Graepel, T. Reevaluating Evaluation. In *NeurIPS*, 2018b.

Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., and Mordatch, I. Emergent complexity via multi-agent competition. *ICLR*, 2018.

Banzhaf, W., Baumgaertner, B., Beslon, G., Doursat, R., Foster, J. A., McMullin, B., de Melo, V. V., Miconi, T., Spector, L., Stepney, S., , and White, R. Defining and Simulating Open-Ended Novelty: Requirements, Guidelines, and Challenges. *Theory in Biosciences*, 2016.

Borel, E. La théorie du jeu et les équations intégrales à noyau symétrique. *Comptes rendus de l'Académie des Sciences*, 1921.

Brant, J. C. and Stanley, K. O. Minimal Criterion Coevolution: A New Approach to Open-Ended Search. In *GECCO*, 2017.

Brown, G. Iterative Solutions of Games by Fictitious Play. In Koopmans, T. C. (ed.), *Activity Analysis of Production and Allocation*. Wiley, 1951.

Candogan, O., Menache, I., Ozdaglar, A., and Parrilo, P. A. Flows and Decompositions of Games: Harmonic and Potential Games. *Mathematics of Operations Research*, 36(3):474–503, 2011.

de Jong, E. D. A Monotonic Archive for Pareto-Coevolution. *Evolutionary Computation*, 15(1):61–93, 2001.

Elo, A. E. *The Rating of Chess players, Past and Present*. Ishi Press International, 1978.

Ficici, S. G. and Pollack, J. B. A Game-Theoretic Approach to the Simple Coevolutionary Algorithm. In *Parallel Problem Solving from Nature (PPSN)*, 2000.

Ficici, S. G. and Pollack, J. B. Pareto Optimality in Coevolutionary Learning. In *ECAL*, 2001.

Fonseca, C. and Fleming, P. Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. In *GECCO*, 1993.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *NeurIPS*, 2014.

Hansen, T. D., Miltersen, P. B., and Sørensen, T. B. On Range of Skill. In *AAAI*, 2008.

Hart, S. Discrete Colonel Blotto and General Lotto games. *Int J Game Theory*, 36:441–460, 2008.

Heinrich, J., Lanctot, M., and Silver, D. Fictitious Self-Play in Extensive-Form Games. In *ICML*, 2015.

Hillis, W. D. Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D: Nonlinear Phenomena*, 42(1-3):228–234, 1990.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J. Z., Silver, D., Hassabis, D., Kavukcuoglu, K., and Graepel, T. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv:1807.01281*, 2018.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. Statistical ranking and combinatorial Hodge theory. *Math. Program., Ser. B*, 127:203–244, 2011.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *NeurIPS*, 2017.

Lehman, J. and Stanley, K. O. Exploiting Open-Endedness to Solve Problems Through the Search for Novelty. In *ALIFE*, 2008.

Leslie, D. and Collins, E. J. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.

McMahan, H. B., Gordon, G., and Blum, A. Planning in the presence of cost functions controlled by an adversary. In *ICML*, 2003.

Miettinen, K. *Nonlinear Multiobjective Optimization*. Springer, 1998.

Monroy, G. A., Stanley, K. O., and Miikkulainen, R. Coevolution of neural networks using a layered Pareto archive. In *GECCO*, 2006.

Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Nolfi, S. and Floreano, D. Coevolving predator and prey robots: Do "arms races" arise in artificial evolution? *Artificial Life*, 4(4):311–335, 1998.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven Exploration by Self-supervised Prediction. In *ICML*, 2017.

Popovici, E., Bucci, A., Wiegand, R. P., and Jong, E. D. D. Coevolutionary principles. In Rozenberg, G., Bck, T., and Kok, J. N. (eds.), *Handbook of Natural Computing*, pp. 987–1033. Springer, Berlin, Heidelberg, 2012.

Roberson, B. The Colonel Blotto game. *Economic Theory*, 29(1), 2006.

Rosin, C. D. New methods for competitive coevolution. *Evolutionary Computation*, 5(1):1–29, 1997.

Schmidt, M. D. and Lipson, H. Coevolution of Fitness Predictors. *IEEE Transactions on Evolutionary Computation*, 12(6):736–759, 2008.

Shoham, Y. and Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362: 1140–1144, 2018.

Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beslon, G., Dolson, E., Froese, T., Hickinbotham, S., Ikegami, T., McMullin, B., Packard, N., Rasmussen, S., Virgo, N., Agmon, E., McGregor, E. C. S., Ofria, C., Ropella, G., Spector, L., Stanley, K. O., Stanton, A., Timperley, C., Vostinar, A., and Wiser, M. Open-Ended Evolution: Perspectives from the OEE Workshop in York. *Artificial Life*, 22:408423, 2016.

Tesauro, G. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3):58–68, 1995.

Tukey, J. W. A problem of strategy. *Econometrica*, 17, 1949.

Wang, R., Lehman, J., Clune, J., and Stanley, K. O. Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. In *arXiv:1901.01753*, 2019.

Wellman, M. P. Methods for empirical game-theoretic analysis. In *AAAI*, pp. 1552–1556, 2006.

Zinkevich, M., Bowling, M., and Burch, N. A New Algorithm for Generating Equilibria in Massive Zero-Sum Games. In *AAAI*, 2007.

APPENDIX

A. Behaviorism

The paper adopts a strict form of behaviorism: an agent is what it does. More precisely, an agent *is all that it can do*. Agents are operationally characterized by all the ways they can interact with all other agents. The functional gamescape captures all possible interactions; the empirical gamescape captures all observed interactions.

Here, we restrict to the single facet of pairwise interactions reported by the function ϕ. One could always delve deeper in specific instances. For example, instead of just reporting win/loss probabilities, one could also report summary statistics regarding which units are built or which pieces are taken.

B. Comparison with GANs and co-evolution

Generative adversarial networks are two-player zero-sum games played by neural networks (Goodfellow et al., 2014). The main difference between GANs and FFG's is that GANs are not symmetric. The discriminator and generator play qualitatively different roles – it would be meaningless to plug discriminator weights into the generator or vice versa. In short, the two losses that comprise a GAN have two different semantics.

There is a substantial literature on co-evolution, which has substantial overlap with our setting (Hillis, 1990; Rosin, 1997; Nolfi & Floreano, 1998; Ficici & Pollack, 2000; 2001; de Jong, 2001; Monroy et al., 2006; Schmidt & Lipson, 2008; Popovici et al., 2012). As with GANs, co-evolution often involves *asymmetric* interactions – such as setter-solver, predators and prey, or organisms and parasites. As a result, the semantics are much richer, but also much more difficult to analyze, and therefore beyond the scope of this paper.

C. Convex geometry

A set S contained in some vector space is **convex** if the lines

$$L_{\mathbf{v}, \mathbf{w}} = \left\{ \alpha \cdot \mathbf{v} + (1 - \alpha) \cdot \mathbf{w} : 0 \le \alpha \le 1 \right\}$$

are subsets of S for all $\mathbf{v}, \mathbf{w} \in S$.

Definition 5. *The **convex hull** of a set A is the intersection*

of all convex sets containing A. Alternatively, the convex hull of $A \subset \mathbb{R}^d$ is

$$\text{hull}(A) := \left\{ \sum_{i=1}^{n} \alpha_i \cdot \mathbf{w}_i : \boldsymbol{\alpha} \succeq 0, \boldsymbol{\alpha}^\intercal \mathbf{1} = 1, \mathbf{w}_i \in A \right\}.$$

The convex hull of a set is necessarily convex.

Theorem 8 (Krein-Milman). *An **extreme point** of a convex set S is a point in S which does not lie in any open line segment joining two points of S. If S is convex and compact in a normed space, then S is the closed convex hull of its extreme points.*

Row versus column mixtures. In the main text we define the empirical gamescape $\mathcal{G}_\mathbf{A}$ to be all convex mixtures of *rows* of an antisymmetric matrix $\mathbf{A}$. It follows by antisymmetry of $\mathbf{A}$ that working with convex mixtures of columns obtains the same polytope up to sign.

The same holds for functional gamescape, where we chose to work with convex combinations of functions of the form $\phi_\mathbf{w}(\bullet) = \phi(\bullet, \mathbf{w})$, but could have equivalently (up to sign) worked with convex combinations of functions of the form $\phi(\mathbf{w}, \bullet)$.

D. Proof of theorem 1

We generalize some results from combinatorial Hodge theory (Jiang et al., 2011; Candogan et al., 2011; Balduzzi et al., 2018b) to the setting where flows are antisymmetric *functions* instead of antisymmetric *matrices*.

D.1. Combinatorial Hodge theory

Combinatorial Hodge theory is a combinatorial analog of differential geometry, see Jiang et al. (2011); Candogan et al. (2011); Baker (2018). Consider a fully connected graph with vertex set $[n] = \{1, \ldots, n\}$. Assign a **flow** $\mathbf{A}_{ij} \in \mathbb{R}$ to each edge of the graph. The flow in the opposite direction ji is $\mathbf{A}_{ji} = -\mathbf{A}_{ij}$, so flows are just $(n \times n)$ antisymmetric matrices. The flow on a graph is analogous to a vector field on a manifold.

The **combinatorial gradient** of an n-vector $\mathbf{r}$ is the flow: $\text{grad}(\mathbf{r}) := \mathbf{r}\mathbf{1}^\intercal - \mathbf{1}\mathbf{r}^\intercal$. Flow $\mathbf{A}$ is a **gradient flow** if $\mathbf{A} = \text{grad}(\mathbf{r})$ for some $\mathbf{r}$, or equivalently if $\mathbf{A}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ for all i, j. The **combinatorial divergence** of a flow is the n-vector $\text{div}(\mathbf{A}) := \frac{1}{n}\mathbf{A} \cdot \mathbf{1}$. The divergence measures the contribution to the flow of each vertex, considered as a source. The **curl** of a flow is the three-tensor $\text{curl}(\mathbf{A})_{ijk} = \mathbf{A}_{ij} + \mathbf{A}_{jk} - \mathbf{A}_{ik}$.

D.2. Functional Hodge theory

We now extend the basic tools of combinatorial Hodge theory to the functional setting.

Let W be a compact set with a probability measure. Given function $\phi : W \times W \to \mathbb{R}$, let $\phi(\mathbf{v}, \mathbf{w})$ prescribe the **flow** from $\mathbf{v}$ to $\mathbf{w}$. The opposite flow from $\mathbf{w}$ to $\mathbf{v}$ is $\phi(\mathbf{w}, \mathbf{v}) = -\phi(\mathbf{v}, \mathbf{w})$, so $\phi : W \times W \to \mathbb{R}$ is antisymmetric. The combinatorial setting above arises as the special case where the compact set W is finite.

The **combinatorial gradient**[1] of a function $f : W \to \mathbb{R}$ is the flow $\text{grad}(f) : W \times W \to \mathbb{R}$ given by

$$\text{grad}(f)(\mathbf{v}, \mathbf{w}) := f(\mathbf{v}) - f(\mathbf{w}).$$

The **divergence** of a flow is the function $\text{div}(\phi) : W \to \mathbb{R}$ given by

$$\text{div}(\phi)(\mathbf{v}) := \int_W \phi(\mathbf{v}, \mathbf{w}) \cdot d\mathbf{w}.$$

The **curl** of flow ϕ is $\text{curl}(\phi) : W \times W \times W \to \mathbb{R}$ given by

$$\text{curl}(\phi)(\mathbf{u}, \mathbf{v}, \mathbf{w}) := \phi(\mathbf{u}, \mathbf{v}) + \phi(\mathbf{v}, \mathbf{w}) - \phi(\mathbf{u}, \mathbf{w}).$$

The following proposition lays the groundwork for the Hodge decomposition, proved in the next section.

Lemma 9. *Let $\mathcal{F} = \{f : W \to \mathbb{R} : \int_W f(\mathbf{w})d\mathbf{w} = 0\}$ be the vector space of functions with zero expectation. Then (i) $\text{div} \circ \text{grad}(f) = f$; (ii) $\text{div}(\phi) \in \mathcal{F}$ for all flows ϕ; and (iii) $\text{curl} \circ \text{grad}(f) \equiv 0$ for all $f \in \mathcal{F}$.*

Proof. **(i)** Observe that

$$\begin{aligned} \text{div} \circ \text{grad}(f) &= \int_W \left(f(\mathbf{v}) - f(\mathbf{w}) \right) \cdot d\mathbf{w} \\ &= f(\mathbf{v}) \int_W d\mathbf{w} - \int_W f(\mathbf{w})d\mathbf{w} \\ &= f(\mathbf{v}) - 0 \end{aligned}$$

where $\int_W d\mathbf{w} = 1$ because $d\mathbf{w}$ is a probability measure and $\int_W f(\mathbf{w})d\mathbf{w} = 0$ by assumption.

(ii) Direct computation obtains that

$$\int_W \text{div}(\phi)(\mathbf{v})d\mathbf{v} = \int_{W \times W} \phi(\mathbf{v}, \mathbf{w}) \cdot d\mathbf{v} \cdot d\mathbf{w} = 0$$

by antisymmetry of ϕ.

(iii) Direct computation shows that $\text{curl} \circ \text{grad}(f)$ is

$$\begin{aligned} & \text{grad}(f)(\mathbf{u}, \mathbf{v}) + \text{grad}(f)(\mathbf{v}, \mathbf{w}) - \text{grad}(f)(\mathbf{u}, \mathbf{w}) \\ &= f(\mathbf{u}) - f(\mathbf{v}) + f(\mathbf{v}) - f(\mathbf{w}) - f(\mathbf{u}) + f(\mathbf{w}) \equiv 0 \end{aligned}$$

as required. $\square$

---

[1] Our use of 'gradient' and 'divergence' does not coincide with the usual usage in multivariate calculus. This is forced on us by the terminological discrepancy between combinatorial Hodge theory and multivariate calculus. No confusion should result; we use $\text{grad}(\bullet)$ for combinatorial gradients and ∇ for calculus gradients.

A basic result in combinatorial Hodge theory is the Hodge decomposition, which decomposes the space of flows into gradient-flows and curl-free flows (Jiang et al., 2011; Balduzzi et al., 2018b). The result is analogous to the Helmholtz decomposition in electrodynamics (Balduzzi et al., 2018a). Here, we prove a variant of the Hodge decomposition in the functional setting.

Theorem 1 (Hodge decomposition)**.** *The vector space of games admits an orthogonal decomposition*

{transitive games}⊕{cyclic games} = im(grad)⊕ker(div)

with respect to the inner product on games $\langle \phi, \psi \rangle :=$ $\int_W \phi(\mathbf{v}, \mathbf{w}) \psi(\mathbf{v}', \mathbf{w}') \cdot d\mathbf{v} d\mathbf{w} d\mathbf{v}' d\mathbf{w}'$.

Proof. First we show that im(grad) and ker(div) are orthogonal. Suppose $\phi = \text{grad}(f)$ and $\text{div}(\psi) = 0$. Then

$$\langle \phi, \psi \rangle = \int \Big(f(\mathbf{v}) - f(\mathbf{w}) \Big) \psi(\mathbf{v}', \mathbf{w}') d\mathbf{v} d\mathbf{v}' d\mathbf{w} d\mathbf{w}'$$
$$= \int f(\mathbf{v}) \psi(\mathbf{v}', \mathbf{w}') d\mathbf{v} d\mathbf{v}' d\mathbf{w}'$$
$$- \int f(\mathbf{w}) \psi(\mathbf{v}', \mathbf{w}') d\mathbf{v}' d\mathbf{w} d\mathbf{w}'$$
$$\equiv 0.$$

Second, observe that any flow ϕ can be written as

$$\phi = \underbrace{\text{grad} \circ \text{div}(\phi)}_{\text{grad}(f)} + \underbrace{\big(\phi - \text{grad} \circ \text{div}(\phi) \big)}_{\psi}$$

where $f = \text{div}(\phi)$ and $\psi := \phi - \text{grad} \circ \text{div}(\phi)$ satisfies

$$\text{div}(\psi) = \text{div}(\phi) - \text{div} \circ \text{grad}(\text{div}\, \phi)$$
$$= \text{div}(\phi) - \text{div}\, \phi = 0,$$

because $\text{div} \circ \text{grad}(\text{div}\, \phi) = \text{div}(\phi)$ by lemma 9 **(i)** and **(ii)**. $\square$

E. Proofs of propositions

Notation. The dot product is $\mathbf{v}^\intercal \mathbf{w} = \langle \mathbf{v}, \mathbf{w} \rangle$. Subscripts can indicate the dimensions of vectors and matrices, $\mathbf{v}_{n \times 1}$ or $\mathbf{A}_{m \times n}$, or their entries, $\mathbf{v}_i$ or $\mathbf{A}_{ij}$; no confusion should result. The unit vector with a 1 in coordinate i is $\mathbf{e}_i$.

Proof of proposition 2. Before proving the proposition, we first tighten up the terminology.

Two agents $\mathbf{v}$ and $\mathbf{v}'$ are behaviorally identical if $\phi(\mathbf{v}, \mathbf{w}) = \phi(\mathbf{v}', \mathbf{w})$ for all $\mathbf{w} \in W$. Given population $\mathfrak{P}$, two agents are empirically indistinguishable if $\phi(\mathbf{v}, \mathbf{w}) = \phi(\mathbf{v}', \mathbf{w})$ for all $\mathbf{w} \in \mathfrak{P}$.

Definition 6. *Population $\mathfrak{Q}$ is redundant relative to population $\mathfrak{P}$ if every agent in $\mathfrak{Q}$ is empirically indistinguishable from a convex mixture of agents in $\mathfrak{P}$.*

More formally, for all $\mathbf{v}$ in $\mathfrak{Q}$ there is an agent $\mathbf{v}' \in \mathcal{G}_{\mathfrak{P}}$ such that $\phi(\mathbf{v}, \mathbf{w}) = \phi(\mathbf{v}', \mathbf{w})$ for all $\mathbf{w} \in \mathfrak{P}$.

Next, we need a notion of equivalence that applies to objects of different dimensions.

Definition 7. *Two polytopes $\mathcal{P} \subset \mathbb{R}^m$ and $\mathcal{Q} \subset \mathbb{R}^{m+n}$ are **equivalent** if there is an orthogonal projection $\boldsymbol{\pi}_{(m+n) \times m}$ with left inverse $\boldsymbol{\xi}_{m \times (m+n)}$ (i.e. satisfying $\boldsymbol{\xi} \cdot \boldsymbol{\pi} = \mathbf{I}_{m \times m}$) such that*

$$\mathcal{P} = \mathcal{Q} \cdot \boldsymbol{\pi} \quad and \quad \mathcal{Q} = \mathcal{P} \cdot \boldsymbol{\xi}.$$

Proposition 2. *Suppose $\mathfrak{Q} \supset \mathfrak{P}$ is redundant relative to $\mathfrak{P}$. Then the gamescapes $\mathcal{G}_{\mathfrak{P}}$ and $\mathcal{G}_{\mathfrak{Q}}$ are equivalent.*

Proof. Suppose that $\mathfrak{P}$ has m elements and $\mathfrak{Q}$ has $m + n$ elements. We assume, without loss of generality, that the elements of $\mathfrak{Q} \supset \mathfrak{P}$ are ordered such that the first m coincide with $\mathfrak{P}$. The $(m + n) \times (m + n)$ evaluation matrix $\mathbf{A}_{\mathfrak{Q}}$ thus contains the $(m \times m)$ evaluation matrix $\mathbf{A}_{\mathfrak{P}}$ as a 'top left' sub-block.

By assumption (since the additional n agents in $\mathfrak{Q}$ are convex combinations of agents in $\mathfrak{P}$), evaluation matrix $\mathbf{A}_{\mathfrak{Q}}$ has the block form

$$\mathbf{A}_{\mathfrak{Q}} = \begin{bmatrix} \mathbf{A}_{\mathfrak{P}} & -\mathbf{A}_{\mathfrak{P}}^\intercal \mathbf{M}^\intercal \\ \mathbf{M}\mathbf{A}_{\mathfrak{P}} & -\mathbf{M}\mathbf{A}_{\mathfrak{P}}^\intercal \mathbf{M}^\intercal \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{m \times m} \\ \mathbf{M}_{n \times m} \end{bmatrix} \cdot \mathbf{A}_{\mathfrak{P}} \cdot \begin{bmatrix} \mathbf{I}_{m \times m} \\ \mathbf{M}_{n \times m} \end{bmatrix}^\intercal$$

where $\mathbf{M}$ specifies the convex combinations. It follows that $\mathcal{G}_{\mathfrak{Q}}$ is generated by convex mixtures of the first m rows of $\mathbf{A}_{\mathfrak{Q}}$:

$$\begin{bmatrix} \mathbf{A}_{\mathfrak{P}} & -\mathbf{A}_{\mathfrak{P}}^\intercal \mathbf{M}^\intercal \end{bmatrix}.$$

Now, let $\mathbf{A} := \mathbf{A}_{\mathfrak{P}}$

$$\mathbf{B} := \begin{bmatrix} \mathbf{A}_{\mathfrak{P}} & -\mathbf{A}_{\mathfrak{P}}^\intercal \mathbf{M}^\intercal \end{bmatrix} = \mathbf{A}_{\mathfrak{P}} \cdot \begin{bmatrix} \mathbf{I}_{m \times m} \\ \mathbf{M} \end{bmatrix}^\intercal,$$

and let

$$\boldsymbol{\xi} := \begin{bmatrix} \mathbf{I}_{m \times m} & \mathbf{M}^\intercal \end{bmatrix} \quad and \quad \boldsymbol{\pi} := \begin{bmatrix} \mathbf{I}_{m \times m} \\ \mathbf{0} \end{bmatrix}.$$

Then $\mathbf{A}$ and $\mathbf{B}$ are equivalent under definition 7. It follows that the gamescape of $\mathbf{A}$ is the orthgonal projection under $\boldsymbol{\pi}$ of the gamescape of $\mathbf{B}$, and that $\mathbf{B}$ can be recovered from $\mathbf{A}$ by applying $\boldsymbol{\xi}$, the right inverse of $\boldsymbol{\pi}$. $\square$

Proof of proposition 3.

Proposition 3. *The EGS of population $\mathfrak{P}$ can be represented in $\mathbb{R}^r$, where $r = \text{rank}(\mathbf{A}_{\mathfrak{P}})$.*

Proof. The rank of an antisymmetric matrix $\mathbf{A}$ is even, so let $r = 2k$. The Schur decomposition, see (Balduzzi et al., 2018b), factorizes an antisymmetric matrix as

$$\mathbf{A}_{n \times n} = \mathbf{W}_{n \times 2k} \cdot \mathbf{J}_{2k \times 2k} \cdot \mathbf{W}_{2k \times n}^\mathsf{T}$$

where $\mathbf{J}_{2k \times 2k} = \sum_{i=1}^{k} \lambda_i \cdot (\mathbf{e}_{2i} \mathbf{e}_{2i-1}^\mathsf{T} - \mathbf{e}_{2i-1} \mathbf{e}_{2i}^\mathsf{T})$ with $\lambda_i > 0$ and the rows of $\mathbf{W}_{n \times 2k}$ are orthogonal.

Let $\mathbf{G} := \mathbf{W}_{n \times 2k} \cdot \mathbf{J}_{2k \times 2k}$. We claim the empirical gamescape, given by convex mixtures of rows of $\mathbf{A}$, is isomorphic to the polytope given by convex mixtures of rows of $\mathbf{G}$, which lives in $\mathbb{R}^{2k} = \mathbb{R}^r$. To see this, observe that $\mathbf{A} = \mathbf{G} \cdot \mathbf{W}^\mathsf{T}$ and the columns of $\mathbf{W}^\mathsf{T}$ are orthogonal. $\square$

Proof of proposition 4.

Proposition 4. *Geometrically, the empirical Nash equilibria are*

$$\mathcal{N}_{\mathfrak{P}} = \{\mathbf{p} \text{ distribution} : \mathbf{p}^\mathsf{T} \mathbf{A}_{\mathfrak{P}} \succeq \mathbf{0}\}.$$

Proof. Recall that the Nash equilibria, of the column-player, in a two-player zero-sum game specified by matrix $\mathbf{A}$, are the solutions to the linear program:

$$\max_{v \in \mathbb{R}} v$$
$$\text{s.t.} \quad \mathbf{p}^\mathsf{T} \mathbf{A} \succeq v \cdot \mathbf{1}$$
$$\mathbf{p} \succeq \mathbf{0} \text{ and } \mathbf{p}^\mathsf{T} \mathbf{1} = 1$$

where the resulting v^* is the value of the game. Since $\mathbf{A}$ is antisymmetric, the value of the game is zero and the result follows. $\square$

Proof of proposition 5.

Proposition 5. *(i) Performance v is independent of the choice of Nash equilibrium. (ii) If ϕ is monotonic then performance compares the best agents in each population*

$$v(\mathfrak{P}, \mathfrak{Q}) = \max_{\mathbf{v} \in \mathfrak{P}} f(\mathbf{v}) - \max_{\mathbf{w} \in \mathfrak{Q}} f(\mathbf{w}).$$

*(iii) If $\text{hull}(\mathfrak{P}) \subset \text{hull}(\mathfrak{Q})$ then $v(\mathfrak{P}, \mathfrak{Q}) \leq 0$ and $v(\mathfrak{P}, \mathfrak{R}) \leq v(\mathfrak{Q}, \mathfrak{R})$ for **any** population $\mathfrak{R}$.*

Proof. **(i)** The value $\mathbf{p}^\mathsf{T} \mathbf{A} \mathbf{q}$ of a zero-sum two-player matrix game is independent of the choice of Nash equilibrium $(\mathbf{p}, \mathbf{q})$.

(ii) If $\phi = \sigma(\text{grad}(f))$ then

$$\mathbf{A}_{\mathfrak{P}} = \left(\sigma\big(f(\mathbf{w}_i) - f(\mathbf{w}_j)\big) \right)_{i,j=1}^n$$

and the Nash equilibrium for each player is to concentrate its mass on the set

$$\underset{i \in [n]}{\arg\max} \, f(\mathbf{w}_i).$$

The result follows immediately.

(iii) If $\text{hull}(\mathfrak{P}) \subset \text{hull}(\mathfrak{Q})$ then the Nash $\mathbf{p}$ on $\mathfrak{P}$ of the row player can be reconstructed by the column player (since every mixture within $\mathfrak{P}$ is also a mixture of agents in $\mathfrak{Q}$. Thus, $v(\mathfrak{P}, \mathfrak{Q})$ is at most zero.

Similarly, if $\text{hull}(\mathfrak{P}) \subset \text{hull}(\mathfrak{Q})$ then every mixture available to $\mathfrak{P}$ is also available to $\mathfrak{Q}$, so $\mathfrak{P}$'s performance against any other population $\mathfrak{R}$ will be the same or worse than the performance of $\mathfrak{Q}$. $\square$

Proof of proposition 6.

Proposition 6. *If $\mathbf{p}$ is a Nash equilibrium on $\mathbf{A}_{\mathfrak{P}}$ and $\sum_i p_i \phi_{\mathbf{w}_i}(\mathbf{v}) > 0$, then adding $\mathbf{v}$ to $\mathfrak{P}$ strictly enlarges the gamescape: $\mathcal{G}_{\mathfrak{P}} \subsetneq \mathcal{G}_{\mathfrak{P} \cup \{\mathbf{v}\}}$.*

Proof. We are required to show that $\mathbf{v}$ is not a convex combination of agents in $\mathfrak{P}$. This follows by contradiction: if $\mathbf{v}$ were a convex combination of agents in $\mathfrak{P}$ then it would either tie or lose to the Nash distribution – whereas $\mathbf{v}$ beats the Nash. $\square$

Proof of proposition 7

Proposition 7. *The objective constructed by rectified Nash response is effective diversity, definition 4.*

Proof. Rewrite

$$d(\mathfrak{P}) = \sum_{i,j=1}^n \lfloor \phi(\mathbf{w}_i, \mathbf{w}_j) \rfloor_+ \cdot p_i p_j$$
$$= \sum_i p_i \sum_j p_j \lfloor \phi_{\mathbf{w}_j}(\mathbf{w}_i) \rfloor_+ =: \sum_i p_i \cdot h_i(\mathbf{w}_i)$$

where $h_i(\mathbf{w}_i) := \sum_j p_j \lfloor \phi_{\mathbf{w}_j}(\mathbf{w}_i) \rfloor_+$. Recall that the rectified Nash response objectives are

$$\mathbb{E}_{\mathbf{w} \sim \mathbf{p}_t} \left[\lfloor \phi_{\mathbf{w}}(\mathbf{v}_t) \rfloor_+ \right].$$

Finally, notice that $h_j(\mathbf{v}_t) = \mathbb{E}_{\mathbf{w} \sim \mathbf{p}_t} \left[\lfloor \phi_{\mathbf{w}}(\mathbf{v}_t) \rfloor_+ \right]$ as required. $\square$

F. Rectified Nash response and a reduction to rock-paper-scissors

In this section, we present some basic analysis of the geometry of the gamescape and its relation to Nash equilibria.

F.1. Nash reweighting of evaluation matrices

Let $\mathbf{A}$ be an antisymmetric matrix with Nash equilibrium $\mathbf{p}$ (not necessarily unique).

Lemma 10. *The Nash reweighted matrix* $\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}$ *defined by setting its* ij^{th} *entry to*

$$(\mathbf{p} \odot \mathbf{A} \odot \mathbf{p})_{ij} := A_{ij} \cdot p_i \cdot p_j$$

is (i) antisymmetric and (ii) satisfies $(\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}) \cdot \mathbf{1} = \mathbf{0}$ *and* $\mathbf{1}^\mathsf{T} \cdot (\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}) = \mathbf{0}^\mathsf{T}$. *That is, all its rows and columns sum to zero.*

Proof. **(i)** Antisymmetry holds since $A_{ij} p_i p_j + A_{ji} p_j p_i = A_{ij} p_i p_j - A_{ij} p_i p_j = 0$ by antisymmetry of $\mathbf{A}$.

(ii) We show that all the entries of the vector $(\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}) \cdot \mathbf{1}$ are zero, by showing that they are all nonnegative and that they sum to zero.

Direct computation obtains that the ith entry of $(\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}) \cdot \mathbf{1}$ is $p_i \sum_j A_{ij} p_j = p_i \cdot (\mathbf{A}\mathbf{p})[i]$. Recall that $\mathbf{A}\mathbf{p} \succeq \mathbf{0}$ since $\mathbf{p}$ is a Nash equilibrium and the value of the game is zero (since $\mathbf{A}$ is antisymmetric). Thus, all the entries of $(\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}) \cdot \mathbf{1}$ are nonnegative. Finally, $\sum_{ij} A_{ij} p_i p_j = \mathbf{p}^\mathsf{T} \mathbf{A} \mathbf{p} = 0$ because $\mathbf{A}$ is antisymmetric. $\square$

F.2. Diversity as a matrix norm

We show that effective diversity, definition 4, arises as a particular matrix norm. Specifically, the diversity of a population can be rewritten as a matrix norm $\ell_{p,q}$ for $p = q = 1$. An interesting direction, beyond the scope of this paper, is to investigate the properties and performance of alternate diversity measures based on more common matrix norms such as the Frobenius and spectral norms.

Recall that

$$\|\mathbf{A}\|_{p,q} := \left(\sum_i \left(\sum_j |a_{ij}|^p \right)^{q/p} \right)^{1/q}$$

and so, in particular,

$$\|\mathbf{A}\|_{1,1} := \sum_{ij} |a_{ij}|.$$

Lemma 11. *The diversity of population* $\mathfrak{P}$ *is*

$$d(\mathfrak{P}) = \frac{1}{2} \|\mathbf{p} \odot \mathbf{A}_{\mathfrak{P}} \odot \mathbf{p}\|_{1,1}$$

Proof. Notice that

$$\|\mathbf{p} \odot \mathbf{A}_{\mathfrak{P}} \odot \mathbf{p}\|_{1,1} = \sum_{ij} |\phi(\mathbf{w}_i, \mathbf{w}_j)| \cdot p_i p_j$$

$$= \sum_{ij} 2\lfloor \phi(\mathbf{w}_i, \mathbf{w}_j) \rfloor_+ \cdot p_i p_j$$

where the second equality follows by antisymmetry of ϕ. $\square$

F.3. Reduction to the rock-paper-scissors

Let $\mathbf{A}$ be an antisymmetric matrix with Nash equilibrium $\mathbf{p}$ (not necessarily unique). Suppose that $\mathbf{p}$ has nonzero mass in at least three[2] entries. Pick a row of $\mathbf{A}$ with positive mass under the Nash – which we can assume to be the first row without loss of generality.

Rock-paper-scissors reduction. Construct weight vectors:

$$\mathbf{p}_{\mathfrak{r}}[i] := \begin{cases} \mathbf{p}[1] & \text{if } i = 1 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{p}_{\mathfrak{p}}[i] := \begin{cases} \mathbf{p}[i] & \text{if } \phi(\mathbf{w}_1, \mathbf{w}_i) > 0 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{p}_{\mathfrak{s}}[i] := \begin{cases} \mathbf{p}[i] & \text{if } \phi(\mathbf{w}_1, \mathbf{w}_i) \le 0 \text{ and } i \ne 1 \\ 0 & \text{else} \end{cases}$$

and corresponding mixture agents:

$$\mathfrak{r} := \sum_{i \in [n]} \mathbf{p}_{\mathfrak{r}}[i] \cdot \mathbf{w}_i$$

$$\mathfrak{p} := \sum_{i \in [n]} \mathbf{p}_{\mathfrak{p}}[i] \cdot \mathbf{w}_i$$

$$\mathfrak{s} := \sum_{i \in [n]} \mathbf{p}_{\mathfrak{s}}[i] \cdot \mathbf{w}_i$$

Mixtures of agents are evaluated in the obvious way, for example

$$\phi(\mathfrak{r}, \mathfrak{s}) = \sum_{i,j \in [n]} \mathbf{p}_{\mathfrak{r}}[i] \cdot \mathbf{p}_{\mathfrak{s}}[j] \cdot \phi(\mathbf{w}_i, \mathbf{w}_j).$$

Proposition 12 (the reduction creates a rock-paper-scissors meta-game)**.** *The evaluation matrix for the population* $\mathfrak{T} = \{\mathfrak{r}, \mathfrak{p}, \mathfrak{s}\}$ *has the form*

$$\mathbf{A}_{\mathfrak{T}} = \begin{pmatrix} 0 & \alpha & -\alpha \\ -\alpha & 0 & \alpha \\ \alpha & -\alpha & 0 \end{pmatrix}$$

for some $\alpha = \phi(\mathfrak{r}, \mathfrak{s}) > 0$.

Proof. In general, a (3×3) antisymmetric matrix must have the form

$$\begin{pmatrix} 0 & \alpha & -\beta \\ -\alpha & 0 & \gamma \\ \beta & -\gamma & 0 \end{pmatrix}$$

for some $\alpha, \beta, \gamma \in \mathbb{R}$. The content of the proposition is **(i)** that $\alpha > 0$ and **(ii)** that $\alpha = \beta = \gamma$.

---

[2] If the Nash equilibrium has support on two agents, say $\mathbf{v}$ and $\mathbf{w}$, then necessarily $\phi(\mathbf{v}, \mathbf{w}) = 0$.

(i) It is clear that $\alpha > 0$ since $\alpha := \phi(\mathfrak{r}, \mathfrak{s}) > 0$ by construction.

(ii) The matrix $\mathbf{A}_{\mathfrak{T}}$ can alternatively be constructed as follows. Take the Nash reweighted matrix $\mathbf{p} \odot \mathbf{A} \odot \mathbf{p}$ and group the rows and columns into rock, paper and scissor blocks, to obtain a matrix with 3×3 blocks. Sum the entries within each of the blocks to obtain the 3×3 matrix $\mathbf{A}_{\mathfrak{T}}$. Lemma 10 then implies that $\mathbf{A}_{\mathfrak{T}} \cdot \mathbf{1} = \mathbf{0}$, which in turn implies $\alpha = \beta = \gamma$. $\qquad \square$

G. Examples of functional-form games

Go, Chess, Starcraft, 1v1 DOTA, poker and other two-player zero-sum games are functional-form games when played by parametrized agents. The function ϕ is difficult to compute, especially when agents are stochastic. Nevertheless, it can be estimated and approximate best-responses to agents or mixtures of agents can be estimated by reinforcement learning algorithms.

In this section, we present a variety of explicit, mathematical constructions of FFG's.

G.1. Symplectic games

Let $W \subset \mathbb{R}^{2d}$ and define $\phi : W \times W \to \mathbb{R}$ as

$$\phi(\mathbf{v}, \mathbf{w}) = \mathbf{v}^\mathsf{T} \cdot \mathbf{\Omega}_{2d} \cdot \mathbf{w} \qquad (3)$$

where $\mathbf{\Omega}_{2d}$ is the antisymmetric matrix

$$\mathbf{\Omega}_{2d} = \begin{pmatrix} 0 & -1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} = \sum_{i=1}^{d} \mathbf{e}_{2i}\mathbf{e}_{2i-1}^\mathsf{T} - \mathbf{e}_{2i-1}\mathbf{e}_{2i}^\mathsf{T}$$

Transitive special case. If $d = 2$ and $\mathbf{v}$ has the form $\mathbf{v} = (\alpha, v)$ then

$$\phi(\mathbf{v}, \mathbf{w}) = \mathbf{v}^\mathsf{T} \cdot \mathbf{\Omega}_2 \cdot \mathbf{w} = \alpha(v - w)$$

which gives a transitive game with $f(\mathbf{v}) = \alpha \cdot v$.

Cyclic special case. A high dimensional cyclic game is obtained by setting

$$W = \{\mathbf{v} \in \mathbb{R}^{2d} \: : \: \|\mathbf{v}\|_2^2 < k\}$$

for some $k > 0$ and setting ϕ per eq. (3).

Mixed transitive and cyclic. The following symplectic game incorporates both transitive and cyclic aspects. Suppose $d > 2$ and

$$W = \{\mathbf{v} \in \mathbb{R}^{2d} \: : \: v_1 = 1, |v_2| \le 100, \text{ and } \|\mathbf{v}_{3:2d}\|_2^2 \le 1\}.$$

In other words, $\mathbf{v} \in W$ has the form $\mathbf{v} = (1, v_1, v_2, \ldots, v_{2d})$ where $v_1 = 1$, $v_2 \in [-100, 100]$ and $(v_3, \ldots, v_{2d})$ lies in the unit ball in $\mathbb{R}^{2d-2}$.

Then $\phi : W \times W \to \mathbb{R}$ given by

$$\phi(\mathbf{v}, \mathbf{w}) = \mathbf{v}^\mathsf{T} \cdot \mathbf{\Omega}_{2d} \cdot \mathbf{w}$$

as above specifies a game that resembles a cigar – with both transitive (given by the first two coordinates) and cyclic (given by the last $2d - 2$ coordinates) structure.

Long cycles. We provide a little more detail on example 3. Suppose n agents form a long cycle: $\mathfrak{P} = \{\mathbf{v}_1 \xrightarrow{\text{beats}} \mathbf{v}_2 \to \cdots \to \mathbf{v}_n \xrightarrow{\text{beats}} \mathbf{v}_1\}$. Then

$$\mathbf{A}_{\mathfrak{P}} = \sum_{i=1}^{n} \mathbf{e}_{\lfloor i \rfloor_n} \mathbf{e}_{\lfloor i+1 \rfloor_n}^\mathsf{T} - \mathbf{e}_{\lfloor i-2 \rfloor_n} \mathbf{e}_{\lfloor i-1 \rfloor_n}^\mathsf{T},$$

where $\lfloor i \rfloor_n$ is i modulo n. Direct computation shows $\text{rank}(\mathbf{A}_{\mathfrak{P}})$ is $n - 2$ if n is even and $n - 1$ if n is odd.

Deformed symplectic games. Let $\mathbf{f}_i : \mathbb{R}^n \to \mathbb{R}^{2d}$ be differentiable multivariate functions and suppose $g_i : \mathbb{R} \to \mathbb{R}$ are odd for all i, i.e. satisfy $g_i(-x) = -g_i(x)$. A large class of differentiable FFG's are of the form

$$\phi(\mathbf{v}, \mathbf{w}) := \sum_i g_i\Big(\mathbf{f}_i(\mathbf{v})^\mathsf{T} \cdot \mathbf{\Omega}_{2d} \cdot \mathbf{f}_i(\mathbf{w})\Big). \qquad (4)$$

where $\phi : W \times W \to \mathbb{R}$ for some $W \subset \mathbb{R}^n$. The class is further enlarged by considering linear combinations of functions of the form in eq. (4).

A strange example. Let

$$W = \{\mathbf{v} \in \mathbb{R}^{2d} \: : \: \|\mathbf{v}\|_1 = 1\}$$

equipped with the uniform probability measure and suppose $\mathbf{f} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is given by

$$\mathbf{f}(\mathbf{v}) = \left(v_1^{99}, \ldots, v_{2d}^{99}\right)$$

It follows that $\mathbf{f}(W) \subset W$, and moreover that the image of W under $\mathbf{f}$ is concentrated on extremely sparse vectors, with most entries near zero. Finally consider the FFG given by

$$\phi(\mathbf{v}, \mathbf{w}) := \mathbf{f}_i(\mathbf{v})^\mathsf{T} \cdot \mathbf{\Omega}_{2d} \cdot \mathbf{f}_i(\mathbf{w}).$$

G.2. Variants of Colonel Blotto

The Colonel Blotto game was introduced in Borel (1921). The game concerns two players that have two distribute limited resources over several battlefields.

Continuous Lotto   Fix $a > 0$. Continuous Lotto is a two-player symmetric zero-sum game introduced in Hart (2008). An 'agent' is a real-valued non-negative random variable X satisfying $\mathbb{E}[X] = a$. Given agents X and Y, define

$$\phi(X, Y) := P(X > Y) - P(X < Y)$$

where

$$P(X > Y) = \int F_X(y) f_Y(y) dy$$

for $F_X(y) = \int_{x=-\infty}^{y} f_X(x) dx$.

Picking different parametrized families of random variables, for example mixtures of k Diracs or mixtures of k Gaussians, yields a variety of functional form games.

Differentiable Lotto.   Many variants of the distance game, described in the main text, can be constructed, to obtain more examples of differentiable zero-sum symmetric functional-form games.

H. Details of 2D embedding procedure

In Figure 1 and 6 we used payoff tables generated in the following way. First, we sampled $N \times N$ matrix E of independent samples from Normal distribution with $\sigma = 0.02$, and then transformed it to an antisymmetric matrix by putting $P_{\text{random}} = (E - E^{\text{T}})/2$ which became the Random payoff. Next we created a purely transitive payoff matrix and added P_{random} to it, creating Almost Transitive payoff P_{trans} and analogously for Almost Cyclic P_{cyclic}. The Mixed payoff is a convex combination of the above, namely $P_{\text{mixed}} = 0.65 P_{\text{trans}} + 0.35 P_{\text{cyc}}$. Finally the almost monotonic payoff has been obtained by using sign function applied point-wise to each element of P_{tran} before the noise was added, so $P_{\text{monotonic}} = \text{sign}(P_{\text{trans}} - E) + E$.

In order to obtain embeddings we used following methods for each of payoffs P:

- **Schur**, Schur decomposition of P has been computed $P = QUQ^{-1}$, and then first two dimencions of U were used as an embedding.

- **PCA**, Principal Component Analysis was applied to N points, where attributes of ith datapoint were taken to be an ith row of the P and 2 most informative dimensions were used.

- **SVD**, Singular-value Decomposition $P = U\Sigma V^*$ was used analogously to the PCA, by keeping dimensions with two highest singular values.

- **tSNE**, t-Distributed Stochastic Neighbor Embedding, applied analogously to PCA was used with perplexity 9 (selected empirically, with various choices leading

to significantly different embeddings, none of which seemed to capture gamescape structure in a more meaningful way).
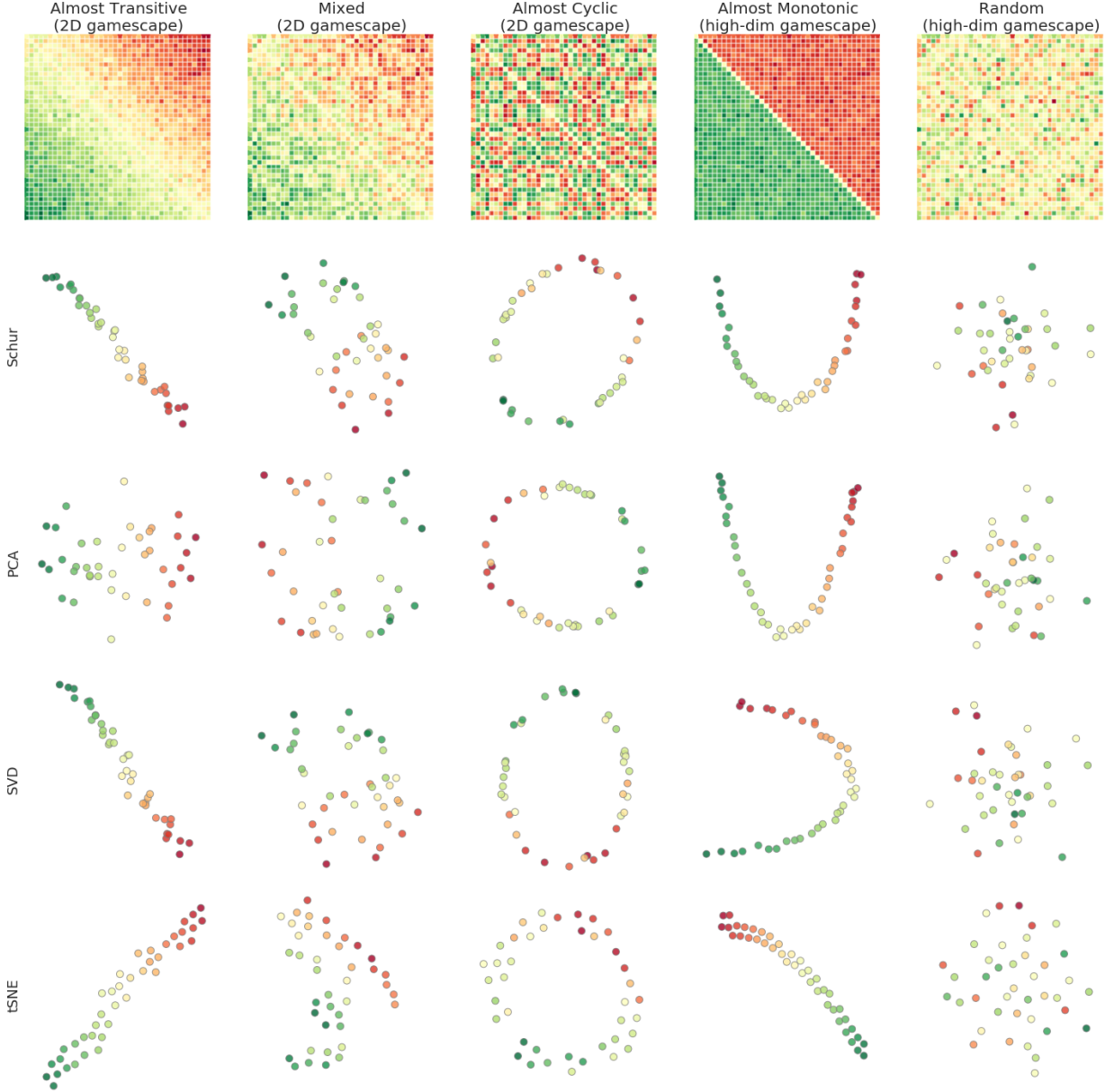
Figure 6. Low-dim gamescapes of various basic game structures. Extended version of Figure 1. **Top row:** Evaluation matrices of populations of 40 agents each; colors vary from red to green as ϕ ranges over $[-1, 1]$. **Bottom row:** 2D embedding obtained by using first 2 dimensions of various embedding techniques; Color corresponds to average payoff of an agent against entire population; EGS of the transitive game is a line; EGS of the cyclic game is two-dim near-circular polytope given by convex hull of points.
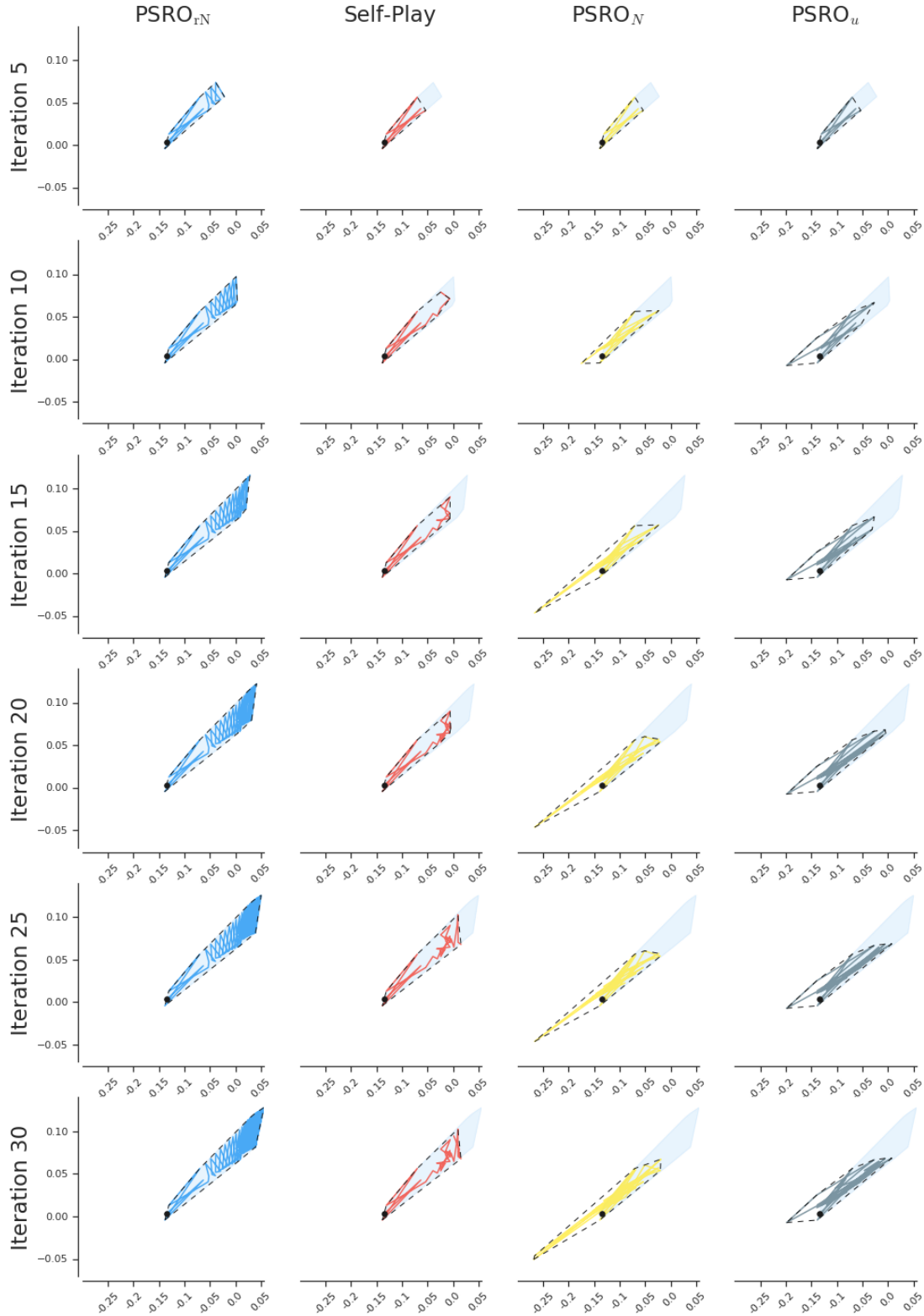
Figure 7. Extended version of Figure 5. Trajectories in 2D Schur embedding space while training on Differentiable Blotto game. Black dot represents initial agent, and dashed lines correspond to a convex hull of the population. Shaded blue region represents area of the convex hull of the PSRO$_{rN}$ population at a given iteration.