# Supplementary Material for
# Modelling Behavioural Diversity for Learning in Open-Ended Games

## Contents

# A. A Review of $\alpha$-Rank

$\alpha$-Rank (Omidshafiei et al., 2019) is a new type of solution concept that is built on the *response graph* of a game. In particular, it tries to find the so-called *sink strongly-connected components (SSCC)* nodes on the response graph that have only incoming edges but no outgoing edges. $\alpha$-Rank serves as a promising replacement for NE; the key associated benefits are its uniqueness, and its polynomial-time solvability in multi-player general-sum games.

On the response graph, each joint pure-strategy profile is a node, and a directed edge points from node $\sigma \in \mathbb{S}$ to node $S \in \mathbb{S}$ if 1) $\sigma$ and $S$ differ in only one single player's strategy, and 2) that deviating player, denoted by $i$, benefits from deviating from $S$ to $\sigma$ such that $\boldsymbol{G}^i(\sigma) > \boldsymbol{G}^i(S)$. We are interested in the so-called *sink strongly-connected components (SSCC)* nodes that have only incoming edges but no outgoing edges on the response graph. To find those SSCC nodes, $\alpha$-Rank constructs a random walk along the directed response graph, which can be equivalently described by a Markov chain, with the transition probability matrix $\boldsymbol{C}$ being:

$$\boldsymbol{C}_{S,\sigma} = \begin{cases} \eta \frac{1-\exp\left(-\alpha\left(\boldsymbol{G}^k(\sigma)-\boldsymbol{G}^k(s)\right)\right)}{1-\exp\left(-\alpha m\left(\boldsymbol{G}^k(\sigma)-\boldsymbol{G}^k(s)\right)\right)} & \text{if } \boldsymbol{G}^k(\sigma) \neq \boldsymbol{G}^k(S) \\ \dfrac{\eta}{m} & \text{otherwise} \end{cases}$$

$$\boldsymbol{C}_{S,S} = 1 - \sum_{i\in\mathcal{N}} \boldsymbol{C}_{S,\sigma}$$

In the above Equation, $\eta = (\sum_{i\in\mathcal{N}}(|S^i| - 1))^{-1}$, $m \in \mathbb{N}$ and $\alpha > 0$ are three constants. Large $\alpha$ ensures the Markov chain is irreducible, and thus guarantees the existence and uniqueness of the $\alpha$-Rank solution, which is the resulting unique stationary distribution $\boldsymbol{\pi}$ of the Markov chain, $\boldsymbol{C}^\top \boldsymbol{\pi} = \boldsymbol{\pi}$. The probability mass of each joint strategy in $\boldsymbol{\pi}$ can be interpreted as the longevity of that strategy during an evolution process.

# B. Algorithm for Open-Ended (Meta-)Game Solvers

**Algorithm 1** A General Solver for Open-Ended (Meta-)Games

```
Initialise: the "high-level" policy set $\mathbb{S} = \prod_{i\in\mathcal{N}} \mathbb{S}^i$, the meta-game payoff $\mathsf{M}, \forall S \in \mathbb{S}$, and meta-policy $\boldsymbol{\pi}^i =$ UNIFORM($\mathbb{S}^i$).
for iteration $t \in \{1, 2, ...\}$ do:
    for each player $i \in \mathcal{N}$ do:
        Compute the meta-policy $\boldsymbol{\pi}_t$ by meta-game solver $\mathcal{S}(\mathsf{M}_t)$.
        Find a new policy against others by Oracle: $S^i_t = \mathcal{O}^i(\boldsymbol{\pi}^{-i}_t)$.
        Expand $\mathbb{S}^i_{t+1} \leftarrow \mathbb{S}^i_t \cup \{S^i_t\}$ and update meta-payoff $\mathsf{M}_{t+1}$.
    terminate if: $\mathbb{S}^i_{t+1} = \mathbb{S}^i_t, \forall i \in \mathcal{N}$.
Return: $\boldsymbol{\pi}$ and $\mathbb{S}$.
```

## C. The Effectiveness of Our Method on the RPS-X Game

We show via a worked example why our method will not fail to find the final strategy X in the RPS-X game, whereas current diversifying methods do fail.[1] Assume that we are in the state of the strategy space being $\{R, P, S\}$, we know that PSRO$_{rn}$ from this position will fail as the best response strategy is still within $\{R, P, S\}$ (McAleer et al., 2020). Instead, we show that if we follow our method and select the strategy with the largest expected cardinality as the best response, that strategy $\{X\}$ will be selected to be added to the population. There are four cases to consider, namely those of adding all four strategies to the current population of strategies:

1. Strategy $\{R\}$ is added to the population and we have the full strategy set $\{R, P, S, R\}$, which induces the following meta-game:

$$\mathbf{M} = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix}.$$

   and the following kernel matrix:

$$\boldsymbol{\mathcal{L}} = \mathbf{M}\mathbf{M}^T = \begin{bmatrix} 2 & -1 & -1 & 2 \\ -1 & 3 & -2 & -1 \\ -1 & -2 & 3 & -1 \\ 2 & -1 & -1 & 2 \end{bmatrix}.$$

   which we can use to solve for the expected cardinality:

$$\mathrm{Tr}\left(\mathbf{I} - (\boldsymbol{\mathcal{L}} + \mathbf{I})^{-1}\right) = 1.6667$$

2. Strategy $\{P\}$ is added to the population and we have the full strategy set $\{R, P, S, P\}$. The result is the same as that of adding $\{R\}$ to the population and the expected cardinality is 1.6667.

3. Strategy $\{S\}$ is added to the population and we have the full strategy set $\{R, P, S, S\}$. The result is the same as that of adding $\{R\}$ to the population and the expected cardinality is 1.6667.

4. Strategy $\{X\}$ is added to the population and we have the full strategy set $\{R, P, S, X\}$, which induces the following meta-game:

$$\mathbf{M} = \begin{bmatrix} 0 & -1 & 1 & -2/5 \\ 1 & 0 & -1 & -2/5 \\ -1 & 1 & 0 & -2/5 \\ 2/5 & 2/5 & 2/5 & 0 \end{bmatrix}.$$

   and the following kernel matrix:

$$\boldsymbol{\mathcal{L}} = \mathbf{M}\mathbf{M}^T = \begin{bmatrix} 2.16 & -0.84 & -0.84 & 0 \\ -0.84 & 2.16 & -0.84 & 0 \\ -0.84 & -0.84 & 2.16 & 0 \\ 0 & 0 & 0 & 0.48 \end{bmatrix}.$$

   which we can use to solve for the expected cardinality:

$$\mathrm{Tr}\left(\mathbf{I} - (\boldsymbol{\mathcal{L}} + \mathbf{I})^{-1}\right) = 2.1486$$

   which, due to inducing the largest expected cardinality out of all of the possible options to be added to the population, strategy $\{X\}$ would be added.

[1] Notably, vanilla PSRO will solve this game in this setting as it will just collapse to the Double Oracle method.

# D. Full Proofs

## D.1. Proposition 4 [G-DPP Diversity Metric]

**Proposition 4** (**G-DPP Diversity Metric**). *The diversity metric, defined as the expected cardinality of a G-DPP, can be computed in $\mathcal{O}(|\mathbb{S}|^3)$ time by the following equation:*

$$\text{Diversity}\left(\mathbb{S}\right) = \mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}}\left[|\mathbf{Y}|\right] = \text{Tr}\left(\mathbf{I} - (\boldsymbol{\mathcal{L}}_\mathbb{S} + \mathbf{I})^{-1}\right).$$

**Proof.** *We can calculate the expected cardinality of a DPP sample by an exponential sum over all subsets of $\mathbb{S}$ as follows (Gillenwater et al., 2018):*

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}}\left[|\mathbf{Y}|\right] &= \sum_{\mathbf{Y}\subseteq\mathbb{S}} |\mathbf{Y}|\, \mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}(\mathbf{Y}) \\ &= \sum_{\mathbf{Y}\subseteq\mathbb{S}} |\mathbf{Y}|\, \frac{\det(\boldsymbol{\mathcal{L}}_\mathbf{Y})}{\det(\boldsymbol{\mathcal{L}}_\mathbb{S} + \mathbf{I})}\end{aligned}$$

*where $\frac{\det(\boldsymbol{\mathcal{L}}_\mathbf{Y})}{\det(\boldsymbol{\mathcal{L}}_\mathbb{S}+\mathbf{I})}$ represents the probability of sampling the subset $\mathbf{Y}$, and therefore we are taking the sum over the probabilities of sampling all different $|\mathbf{Y}|$. Based on the following Lemma from Rising (2013):*

> **Lemma** (Theorem 2.3.9 in Rising (2013))**.** *Let $\mathbf{Y} \sim \text{DPP}(\mathcal{K})$ where $\mathcal{K} = \mathbf{I} - (\boldsymbol{\mathcal{L}} + \mathbf{I})^{-1}$, and let $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of $\mathcal{K}$. Then $|\mathbf{Y}| = \sum_{i=1}^n Z_i$ where $\{Z_i\}_{i=1}^n$ are independent Bernoulli trials with $\mathbb{E}[Z_i] = \lambda_i$.*

*we can write the expectation over the random variable $|\mathbf{Y}|$ by*

$$\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}}\left[|\mathbf{Y}|\right] = \mathbb{E}\left[\sum_{i=1}^n Z_i\right] = \sum_{i=1}^n \mathbb{E}\left[Z_i\right] = \sum_{i=1}^n \lambda_i = \text{Tr}\left(\mathcal{K}\right).$$

*Since $\mathcal{K} = \boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{L}} + \mathbf{I})^{-1} = \mathbf{I} - (\boldsymbol{\mathcal{L}} + \mathbf{I})^{-1}$ (Kulesza et al., 2012), we can relate the eigenvalues of $\boldsymbol{\mathcal{L}}$ and $\mathcal{K}$ as follows,*

$$\lambda_i^{\mathcal{K}} = \frac{\lambda_i^{\boldsymbol{\mathcal{L}}}}{\lambda_i^{\boldsymbol{\mathcal{L}}} + 1}, \quad \forall i$$

*where the superscript references which matrix the eigenvalues belong to. Finally, we have,*

$$\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}}\left[|\mathbf{Y}|\right] = \sum_{i=1}^n \frac{\lambda_i^{\boldsymbol{\mathcal{L}}_\mathbb{S}}}{\lambda_i^{\boldsymbol{\mathcal{L}}_\mathbb{S}} + 1} = \text{Tr}\left(\mathbf{I} - (\boldsymbol{\mathcal{L}}_\mathbb{S} + \mathbf{I})^{\mathbf{-1}}\right)$$

*and we have the expected cardinality of a sample from our DPP based upon the kernel matrix $\boldsymbol{\mathcal{L}}_\mathbb{S}$.* ■

## D.2. Proposition 5 [Maximum Diversity]

**Proposition 5** (**Maximum Diversity**). *The diversity of a population $\mathbb{S}$ is bounded by* $\text{Diversity}\left(\mathbb{S}\right) \leq \text{rank}(\mathbf{M})$*, and if* $\mathbf{M}$ *is normalised (i.e.,* $||\mathbf{M}_{[i,:]}|| = 1, \forall i$*), we have* $\text{Diversity}\left(\mathbb{S}\right) \leq \text{rank}(\mathbf{M})/2$*. In both cases, maximal diversity is reached if and only if* $\mathbf{M}$ *is orthogonal.*

**Proof.** *We start from the the case of an un-normalised meta-game. We can always do the SVD decomposition of* $\mathbf{M}$ *and get*

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^\top$$

*Then, the kernel $\boldsymbol{\mathcal{L}}$ of G-DPP can be written as*

$$\boldsymbol{\mathcal{L}} = \mathbf{M}\mathbf{M}^\top = (\mathbf{U}\Sigma\mathbf{V}^\top)(\mathbf{U}\Sigma\mathbf{V}^\top)^\top = \mathbf{U}\Sigma^2\mathbf{U}^\top$$

*This means that the eigenvalues of $\boldsymbol{\mathcal{L}}$ are $\lambda_i = \sigma_i^2 > 0$ where $\sigma_i$ are the entries of the diagonal of $\Sigma$. Thus, based on Proposition 4, we can write the expected cardinality given kernel $\boldsymbol{\mathcal{L}}$ as*

$$\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}}\left[|\mathbf{Y}|\right] = \sum_i \frac{\lambda_i}{\lambda_i + 1} = \sum_i \frac{\sigma_i^2}{\sigma_i^2 + 1}.$$

*This implies that all eigenvalues are positive and thus, a larger cardinality can only be achieved by either adding more eigenvalues or making the eigenvalues larger. Since we can only get a maximum of $n = \text{rank}(\boldsymbol{M})$ non-zero eigenvalues, making the eigenvalues larger will increase the terms of the summation by $\frac{\lambda_i}{\lambda_i+1} < 1$ which means that*

$$\sup \mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_\mathbb{S}}}\left[|\mathbf{Y}|\right] = \text{rank}(\mathbf{M})$$

*Thus, the maximum achievable diversity is obtained when population $\mathbb{S}$ makes* $\mathbf{M}$ *full rank and $\lambda_i \to \infty, \forall i$.*

*We now prove the case for a normalised (i.e., $||\mathsf{M}_{[i,:]}|| = 1, \forall i$) meta-game. We first show that a DPP is maximised over orthogonal feature vectors. To do this we show that the probability of sampling the whole set of agents is maximised for orthogonal vectors. Starting by noting Hadamard's Inequality, it states that if $\mathsf{N}$ is the matrix having columns $v_i$, then*

$$\Big|\det\big(\mathsf{N}\big)\Big| \leq \prod_{i=1}^{n} \|v_i\|,$$

*where notably equality holds if and only if the vectors $v_i$ are orthogonal. Therefore, by this inequality we also know that the determinant of $\mathsf{M}$ is maximised when the payoffs are orthogonal for each agent, as this allows the equality to hold.*

*Additionally, as we define our $\mathsf{M}$ to be a square matrix of size $M \times M$, we can decompose the determinant of $\boldsymbol{\mathcal{L}}$ as,*

$$\det\big(\boldsymbol{\mathcal{L}}_{\mathcal{Y}}\big) = \det\big(\mathsf{M}\mathsf{M}^{\top}\big) = \det\big(\mathsf{M}\big)\det\big(\mathsf{M}^{\top}\big) = \det\big(\mathsf{M}\big)^2$$

*and therefore we know that $\det(\boldsymbol{\mathcal{L}}_{\mathcal{Y}})$ is maximised whenever $\det(\mathsf{M})$ is maximised. According to the following equation,*

$$\mathbb{P}_{\boldsymbol{\mathcal{L}}}\big(\boldsymbol{Y} = \mathcal{Y}\big) \propto \det\big(\boldsymbol{\mathcal{L}}_{\mathcal{Y}}\big)$$

*we therefore have that the probability of sampling the whole ground set is maximised when the $\det(\boldsymbol{\mathcal{L}}_{\mathcal{Y}})$ is maximised, which coincides with the orthogonality of $\mathsf{M}$.*

*Finally, based on the fact that orthogonal feature vectors maximise the diversity of a population, we show that a meta-game with orthogonal feature vectors will receive an expected cardinality value of $\frac{n}{2}$.*

*We know that the entries of our kernel $\boldsymbol{\mathcal{L}}$ are simply $L_{ij} = \boldsymbol{w}_i\boldsymbol{w}_j^{\top}$. As long as $||\boldsymbol{w}_i|| = 1, \forall i$ it will be the case that $L_{ii} = 1$. Additionally, as we have stated that the rows of the meta game are orthogonal, $\boldsymbol{w}_i\boldsymbol{w}_j^{\top} = 0$ , $\forall i, j$, so all off-diagonal entries $L_{ij} = 0$. For example, the kernel $\boldsymbol{\mathcal{L}}$ for a meta game of size $3$ would be,*

$$\boldsymbol{\mathcal{L}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

*which is obviously the identity matrix of size M. The identity matrix has M eigenvalues of 1, and therefore we have*

$$\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_{\mathbb{S}}}}\big[|\mathbf{Y}|\big] = \sum_{i=1}^{n} \frac{\lambda_i^{\boldsymbol{\mathcal{L}}}}{\lambda_i^{\boldsymbol{\mathcal{L}}} + 1} = \sum_{i=1}^{n} \frac{1}{2} = \frac{n}{2}.$$

*As the expected cardinality is maximised for orthogonal feature vectors, and we know that orthogonal feature vectors will return an expected cardinality of $\frac{n}{2}$, we know that when our feature vectors are normalised this is the max achievable expected cardinality.* ■

### D.3. Proposition 6 [Diversity vs. Matrix Norm]

**Proposition 6 (Diversity vs. Matrix Norm).** *Maximising the diversity in Eq. (11) also maximises the Frobenius norm of $\|\mathsf{M}\|_F$, but NOT vice versa.*

**Proof.** *Based on Proposition 5, the maximum achievable diversity is:*

$$\sup \mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_{\mathbb{S}}}}\big[|\mathbf{Y}|\big] = \sup \sum_i \frac{\lambda_i}{\lambda_i + 1} = \operatorname{rank}(\mathsf{M}) = n$$

*Since the terms in the above summation are $0 \leq \frac{\lambda_i}{\lambda_i+1} < 1$, the expected cardinality is also maximised for $\lambda_i \to \infty, \forall n$. Therefore, we have that the maximum achievable diversity is attained when the eigenvalues of $\boldsymbol{\mathcal{L}}$ are maximised and that* $\operatorname{rank}(\mathsf{M}) = n$. *In a G-DPP, we have that the kernel matrix is defined as $\boldsymbol{\mathcal{L}} = \mathsf{M}\mathsf{M}^{\top}$, and the Frobenius norm is written as*

$$\begin{aligned} \Big\|\mathsf{M}\Big\|_F &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\Big|\mathsf{M}_{ij}\Big|^2} = \sqrt{\operatorname{Tr}\Big(\mathsf{M}\mathsf{M}^{\top}\Big)} \\ &= \sqrt{\operatorname{Tr}\big(\boldsymbol{\mathcal{L}}\big)} \\ &= \sqrt{\sum_{i=1}^{n}\lambda_i} \end{aligned}$$

*We have shown above that maximising the expected cardinality of $\boldsymbol{\mathcal{L}}$ is achieved by maximising the eigenvalues of $\boldsymbol{\mathcal{L}}$ and it* *directly follows that:*

$$\sup \left\|\mathsf{\mathbf{M}}\right\|_F = \sup \sqrt{\sum_{i=1}^{n} \lambda_i}$$

*is achieved as $\lambda_i \to \infty, \forall i$.*

*Now we allude to why maximising $||\mathsf{\mathbf{M}}||_F$ will not necessarily lead to a large diversity of agents (i.e., the opposite direction of this proposition is not necessarily correct), but our measure based upon expected cardinality will. Firstly, note that given*

$$||\mathsf{\mathbf{M}}||_F = \sqrt{\sum_{i=1}^{n} \lambda_i},$$

*an optimiser may go about increasing this function in two manners. One manner may be to continually add new eigenvalues which will help improve the overall diversity of $\mathsf{\mathbf{M}}$. However, an optimiser may also increase this norm by focusing on one eigenvalue, or a subset of eigenvalues, and increase those to be as large as possible. Our major worry here would be that by focusing on a subset of eigenvalues, new agents that are added to the population may be redundant as they are added to increase the current set of eigenvalues - not to bring in new non-zero eigenvalues. This would lead to a lack of diversity in the population. On the other hand, note the formula for expected cardinality is,*

$$\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_{\mathbb{S}}}}\left[|\mathbf{Y}|\right] = \sum_i \frac{\lambda_i}{\lambda_i + 1}$$

*and note how each eigenvalue can at most contribute a value of 1 to the expected cardinality value. The importance of this is that an optimiser will gain less marginal benefit from increasing an already large eigenvalue than it would from adding a new non-zero eigenvalue to the meta game. Therefore, whilst we can guarantee that our expected cardinality measure will always search for new non-zero eigenvalues (and by that notion, more diverse agents), we can not guarantee the same for the Frobenius norm measure.* ■

### D.4. Proposition 7 [Uniqueness of Diverse Best Response]

**Proposition 7** (**Uniqueness of Diverse Best Response**). *Eq. (11) is a strictly concave function. The resulting best response in Eq. (12) has a unique solution.*

**Proof.** *We study the sign of the second derivative of the diversity measure (Eq.(11)) in a neighbourhood of the positive semidefinte symmetric matrix $\boldsymbol{\mathcal{L}}$. We apply a perturbation to $\boldsymbol{\mathcal{L}}$ such that $\tilde{\boldsymbol{\mathcal{L}}} = \boldsymbol{\mathcal{L}} + \varepsilon\mathbf{A}$ with $\mathbf{A}$ being a symmetric matrix and $\varepsilon \in \mathbb{R}$, as a result, what we need to show is:*

$$\frac{\partial^2}{\partial \varepsilon^2} \operatorname{Tr}\Big(\mathbf{I} - (\tilde{\boldsymbol{\mathcal{L}}} + \mathbf{I})^{-1}\Big)\Big|_{\varepsilon=0} < 0$$

*First, we notice that*

$$\mathbf{I} - (\tilde{\boldsymbol{\mathcal{L}}} + \mathbf{I})^{-1} = \mathbf{I} - \sum_{n=0}^{\infty}(-1)^n(\tilde{\boldsymbol{\mathcal{L}}})^n = -\sum_{n=1}^{\infty}(-1)^n(\tilde{\boldsymbol{\mathcal{L}}})^n = \sum_{n=0}^{\infty}(-1)^n(\boldsymbol{\mathcal{L}} + \varepsilon\mathbf{A})^{n+1}$$

*where we used the matrix expansion $(\mathbf{I} + \tilde{\boldsymbol{\mathcal{L}}})^{-1} = \sum_{n=0}^{\infty}(-1)^n\tilde{\boldsymbol{\mathcal{L}}}^n$. We can always ensure $||\tilde{\boldsymbol{\mathcal{L}}}||_F < 1$ by choosing $\varepsilon$ small enough and redefining $\boldsymbol{\mathcal{L}} := \dfrac{\boldsymbol{\mathcal{L}}}{||\boldsymbol{\mathcal{L}}||_F + \beta}, \beta > 0 \in \mathbb{R}$ if necessary. We note that after this modification $\boldsymbol{\mathcal{L}}$ remains a legitimate matrix for a DPP due to being positive semidefinite, and do not affect the rankings of agents. Then,*

$$\begin{aligned}
\frac{\partial^2}{\partial \varepsilon^2}\Big(\mathbf{I} - (\tilde{\boldsymbol{\mathcal{L}}} + \mathbf{I})^{-1}\Big)\Big|_{\varepsilon=0} &= \frac{\partial^2}{\partial \varepsilon^2}\sum_{n=0}^{\infty}(-1)^n(\boldsymbol{\mathcal{L}} + \varepsilon\mathbf{A})^{n+1}\Big|_{\varepsilon=0} \\
&= \sum_{n=0}^{\infty}(-1)^n\frac{\partial^2}{\partial \varepsilon^2}(\boldsymbol{\mathcal{L}} + \varepsilon\mathbf{A})^{n+1}\Big|_{\varepsilon=0} \\
&= \frac{1}{2}\mathbf{A}^2\sum_{n=0}^{\infty}(-1)^n(n+1)(n)\boldsymbol{\mathcal{L}}^{n-1} + \frac{1}{2}\sum_{n=0}^{\infty}(-1)^n(n+1)(n)\boldsymbol{\mathcal{L}}^{n-1}\mathbf{A}^2 \qquad (1)
\end{aligned}$$

*The series* $\sum_{n=0}^{\infty}(-1)^n(n+1)(n)\boldsymbol{\mathcal{L}}^{n-1}$ *can be written as*

$$\begin{aligned}
\sum_{n=0}^{\infty}(-1)^n(n+1)(n)\boldsymbol{\mathcal{L}}^{n-1} &= \sum_{n=0}^{\infty}(-1)^{n+1}(n+2)(n+1)\boldsymbol{\mathcal{L}}^{n} \\
&= -\sum_{n=0}^{\infty}(-1)^{n+2}\frac{\partial^2}{\partial\boldsymbol{\mathcal{L}}^2}\boldsymbol{\mathcal{L}}^{n+2} \\
&= -\frac{\partial^2}{\partial\boldsymbol{\mathcal{L}}^2}\sum_{n=0}^{\infty}(-1)^{n+2}\,\boldsymbol{\mathcal{L}}^{n+2} \\
&= -\frac{\partial^2}{\partial\boldsymbol{\mathcal{L}}^2}\left(\boldsymbol{\mathcal{L}}^2-\boldsymbol{\mathcal{L}}^3+\boldsymbol{\mathcal{L}}^4-\ldots\right) \\
&= -\frac{\partial^2}{\partial\boldsymbol{\mathcal{L}}^2}\left((\mathbf{I}+\boldsymbol{\mathcal{L}})^{-1}+\boldsymbol{\mathcal{L}}-\mathbf{I}\right) \\
&= -2(\mathbf{I}+\boldsymbol{\mathcal{L}})^{-3}
\end{aligned}$$

*where we used the matrix expansion* $(\mathbf{I}+\boldsymbol{\mathcal{L}})^{-1}=\sum_{n=0}^{\infty}(-1)^n\boldsymbol{\mathcal{L}}^n$ *which holds under the modifications specified before.*

*Finally, we obtain that*

$$\begin{aligned}
\frac{\partial^2}{\partial\varepsilon^2}\operatorname{Tr}\left(\mathbf{I}-(\tilde{\boldsymbol{\mathcal{L}}}+\mathbf{I})^{-1}\right)\Big|_{\varepsilon=0} &= \operatorname{Tr}\left(\frac{\partial^2}{\partial\varepsilon^2}\left(\mathbf{I}-(\tilde{\boldsymbol{\mathcal{L}}}+\mathbf{I})^{-1}\right)\right)\Big|_{\varepsilon=0} \\
&= -\operatorname{Tr}\left(\mathbf{A}^2(\mathbf{I}+\boldsymbol{\mathcal{L}})^{-3}+(\mathbf{I}+\boldsymbol{\mathcal{L}})^{-3}\mathbf{A}^2\right) \\
&< 0
\end{aligned}$$

*Where the last inequality comes from the fact that both* $(\mathbf{I}+\boldsymbol{\mathcal{L}})^{-3}$ *and* $\mathbf{A}^2$ *are positive definite and the trace of the product of two positive definite matrices is positive. We therefore have shown that the expected cardinality is a strictly concave function, which are known to have a unique optimiser.* ■

### D.5. Theorem 8 [Convergence of Diverse FP]

**Theorem 8 (Convergence of Diverse FP).** *The perturbation sequence induced by diverse FP process is a uniformly bounded martingale difference sequence; therefore, diverse FP shares the same convergence property as GWFP.*

**Proof.** *We begin by proving that our perturbation sequence is a martingale sequence w.r.t. the normalised expected cardinality (this has no effect on rankings of agents),*

$$X_n=\frac{\mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\boldsymbol{\mathcal{L}}_{\mathbb{S}}}}\left[|\mathbf{Y}|\right]}{|\mathcal{Y}|}$$

*Based on Proposition 5, assuming a normalised payoff table, we know* $0\leq X_n\leq 0.5$*. We are to show that* $S_n=\frac{1}{n}\sum_{i=1}^{n}[X_i-\mu]$ *where* $\mu=\mathbb{E}[X_n]$ *is a martingale sequence by proving that it meets the three conditions listed below one by one.*

1. $S_n$ *is a measurable function of* $X_1, X_2, ..., X_n$

   $S_n$ *is a measurable function as it is a partial sum of* $\{X_i\}_{i=1}^{\infty}$.

2. $\mathbb{E}[|S_n|]<\infty, \forall n$

$$\mathbb{E}[|S_n|]\leq\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_i-\mu|]<\infty$$

3. $\mathbb{E}[S_{n+1}|X_1, ...X_n] = S_n$

$$\begin{aligned}
\mathbb{E}[S_{n+1}|X_1, ..., X_n] &= \mathbb{E}[S_n + \frac{1}{n}X_{n+1} - \frac{1}{n}\mu] \\
&= S_n + \mathbb{E}[\frac{1}{n}X_{n+1} - \frac{1}{n}\mu|X_1, ..., X_n] \\
&= S_n - \frac{1}{n}\mu + \mathbb{E}[\frac{1}{n}X_{n+1}|X_1, ..., X_n] \\
&= S_n - \frac{1}{n}\mu + \frac{1}{n}\mu \\
&= S_n
\end{aligned}$$

*and therefore we have that $\{S_k\}_{k=1}^{\infty}$ is a martingale sequence w.r.t. $\{X_k\}_{k=1}^{\infty}$. Now we let $M_k = S_k - S_{k-1}, k = 2, 3, ...$ and show that this is a martingale **difference** sequence under the same three conditions.*

1. *$M_k$ is a measurable function of $X_1, X_2, ..., X_n$*

   *$M_k$ is measurable as both $S_k$ and $S_{k-1}$ are.*

2. $\mathbb{E}[|M_k|] < \infty, \quad \forall k$

$$\mathbb{E}[|M_k|] \leq \mathbb{E}[|S_k|] + \mathbb{E}[|S_{k-1}|] < \infty$$

3. $\mathbb{E}[M_{k+1}|X_1, ..., X_n] = 0$

$$\begin{aligned}
\mathbb{E}[M_{k+1}|X_1, ..., X_n] &= \mathbb{E}[S_{k+1} - S_k|X_1, ..., X_n] \\
&= \mathbb{E}[S_{k+1}|X_1, ..., X_n] - S_k \\
&= S_k - S_k \\
&= 0
\end{aligned}$$

*and we have shown that $\{M_k\}_{k=1}^{\infty}$ is a martingale difference sequence w.r.t. $\{X_k\}_{k=1}^{\infty}$.*
*Next, we show that this martingale sequence is bounded uniformly in $L2$. $M_n$ is said to be bounded in $L2$ if:*

$$\sup_n [M_n^2] < \infty$$

*which can be shown due to the following:*

$$\sup_n \left[M_n^2\right] = \sup_n \left[\frac{1}{n}\sum_{i=1}^{n}\left[X_i - \mu\right]^2\right] < \infty$$

$$\textit{as } \frac{1}{n}\sum_{i=1}^{n}\left[X_i - \mu\right]^2 \leq 0.25, \forall n$$

*where the last line holds due to $0 \leq X_i \leq 0.5$. Therefore, if $M_n \in L2$, then the martingale sequence $M$ is bounded in $L2$ if and only if:*

$$\sum_{k \geq 1} \mathbb{E}\left[(M_k - M_{k-1})^2\right] < +\infty$$

*which we can show in the following manner:*

$$\begin{aligned}
\sum_{k \geq 1} \mathbb{E}\left[(M_k - M_{k-1})^2\right] &= \sum_{k \geq 1}\left[\mathbb{E}[M_k^2 - 2M_kM_{k-1} + M_{k-1}^2]\right] \\
&= \sum_{k \geq 1}\left[\mathbb{E}[M_k^2] + \mathbb{E}[M_{k-1}^2] - \mathbb{E}[2M_kM_{k-1}]\right] \\
&= \sum_{k \geq 1}\left[\mathbb{E}[M_k^2] + \mathbb{E}[M_{k-1}^2]\right] \\
&< \infty
\end{aligned}$$

*as $M_k \in L2, \forall k$. where we note that $\mathbb{E}[2M_k M_{k-1}] = 0$ due to the martingale difference sequence properties of $M_k$. Finally, as we satisfy both of the conditions of GWFP, namely that our expected cardinality function is strictly concave as shown in Proposition 7 and the perturbation meets the condition of Eq. (5) we have that as $\tau \to 0$ our diverse FP process will almost surely result in a GWFP process, which are known to converge in two-player zero-sum games and potential games (Leslie & Collins, 2006).* ■

### D.6. Proposition 9 [Gamescape Enlargement]

**Proposition 9** (**Gamescape Enlargement**). *Adding a new best-response policy $S_{\boldsymbol{\theta}}$ via Eq. (13) strictly enlarges the gamescape. Formally, we have*

$$\text{Gamescape}\left(\mathbb{S}\right) \subsetneq \text{Gamescape}\left(\mathbb{S} \cup \left\{S_{\boldsymbol{\theta}}\right\}\right)$$

**Proof.** *According to Proposition 5 and Proposition 7, maximising the expected cardinality of the DPP implies there exists a unique solution that makes the meta-game payoff table $\mathbf{M}$ full rank and increases the eigenvalues. If $\mathbf{M}$ is full rank, this means that the new row in the meta game $\mathbf{M}$ corresponding to the new policy $S_{\boldsymbol{\theta}}$ must be linearly independent to the other rows and thus it must not be a convex combination of the other rows, as a result, the gamescape is strictly enlarged. Such a result is also expected because the diversity in terms of expected cardinality promotes orthogonality, which is a stronger property than linear independency.* ■

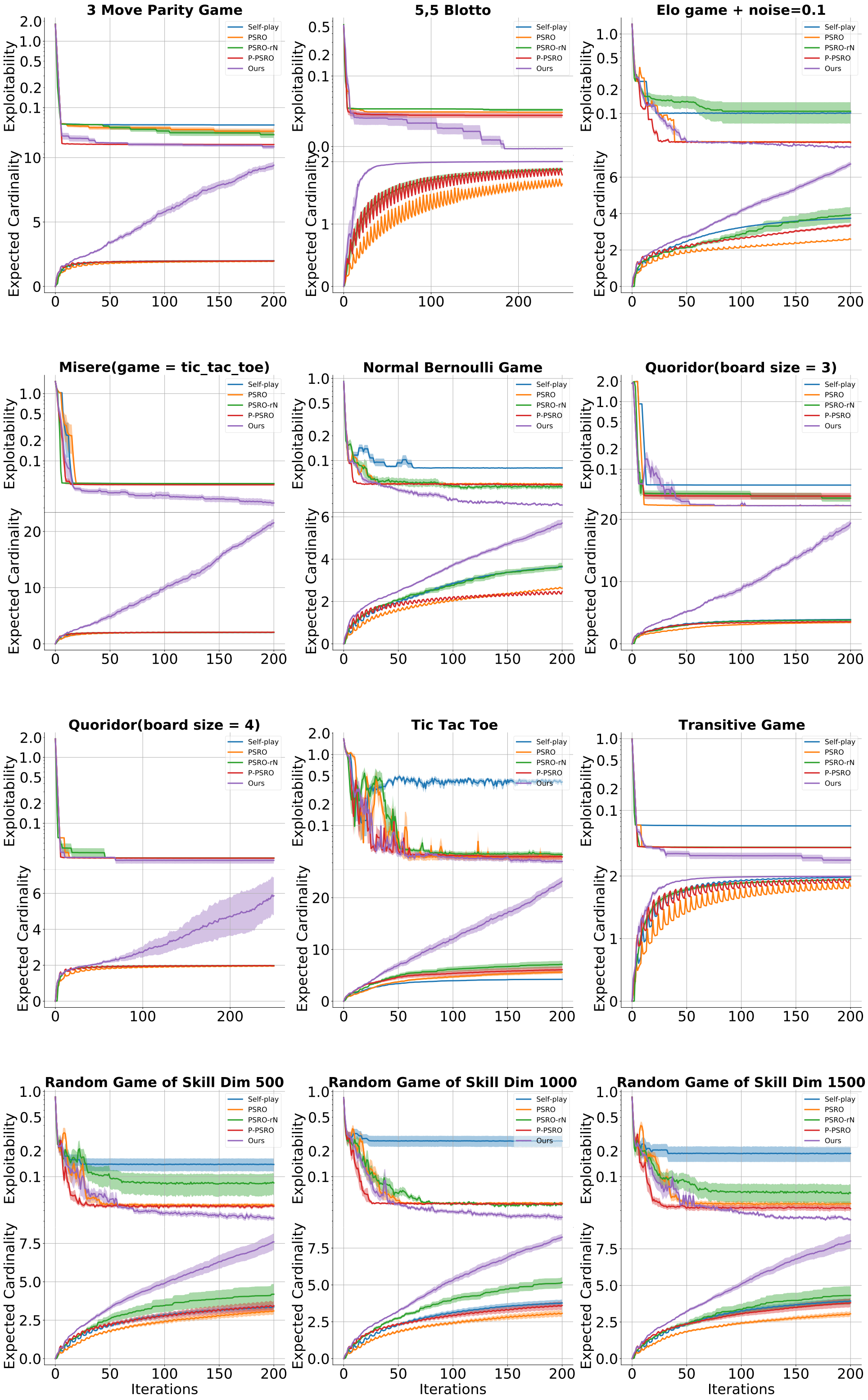
### D.7. Theorem 10 [Convergence of Diverse $\alpha$-PSRO]

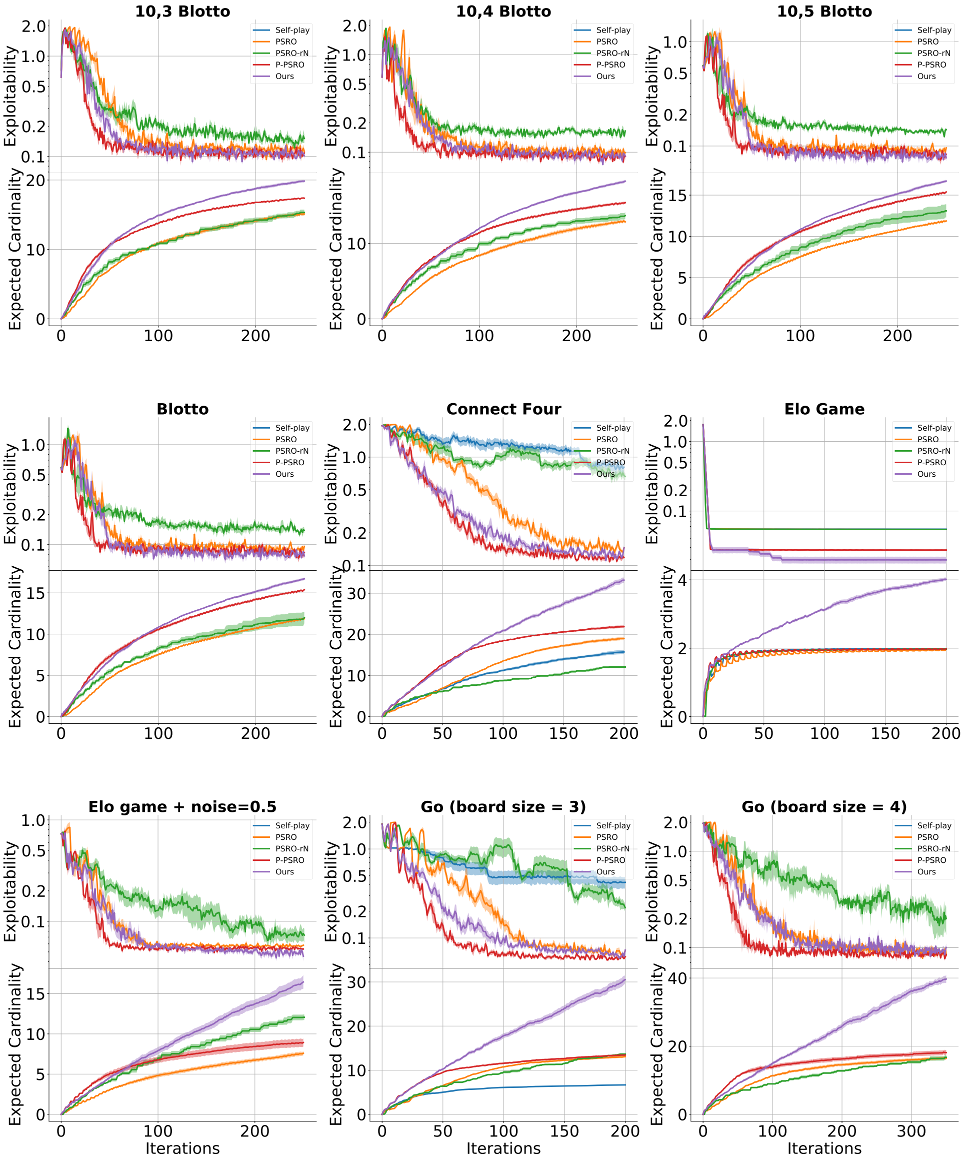
**Theorem 10** (**Convergence of Diverse $\alpha$-PSRO**). *Diverse $\alpha$-PSRO with the Oracle of Eq. (14) converges to the sub-cycle of the unique SSCC in the two-player symmetric games.*

**Proof.** *Our proof follows the same argument as (Muller et al., 2019) Proposition 3, which can be adapted in the content of our diverse PBR oracle in Eq. (14). The uniqueness of the SSCC follows from the fact that in the single-population case, the response graph is fully-connected. Suppose at termination of $\alpha$-DPP, the $\alpha$-population contains no strategy within the SSCC, and let $s$ be a strategy in the $\alpha$-DPP population. We claim that $s$ attains a higher value for the objective defining the $\alpha$-DPP oracle than any strategy in the $\alpha$-DPP population, which contradicts the fact that $\alpha$-DPP has terminated. By virtue of being in the SSCC we have that $\boldsymbol{M}^1(s, s') > \boldsymbol{M}^2(s', s)$ for all $s'$ outside the SSCC, and in particular for all $s' \in \mathbb{S}_t^i$, thus the PBR objective for $s$ is 1.*

*If a member of the underlying game's SSCC appears in the $\alpha$-DPP population, this member will induce its own meta-SSCC in the meta game's response graph which will have positive probability under the $\alpha$-Rank distribution for the meta-game, and the DPP oracle for this meta-SSCC will always return a member of the underlying game's SSCC. This is because the only strategies that receive a non-zero quality score are those that are members of the underlying game's SSCC, and we ignore strategies with a quality of zero. If the only strategy that returns a non-zero quality score, which must be a member of the SSCC, is already in the population, the corresponding meta-SSCC already contains a cycle of the underlying SSCC. Note that if the meta-SSCC does not contain a cycle, it must be a singleton. Either this singleton is equal to the full SSCC of the underlying game (in which we have $\alpha$-fully converged), or it is not, in which case the DPP oracle must return a new strategy from the underlying SSCC, contradicting our assumption that is has terminated.* ■

# E. Additional Results on Real-World Meta-Games

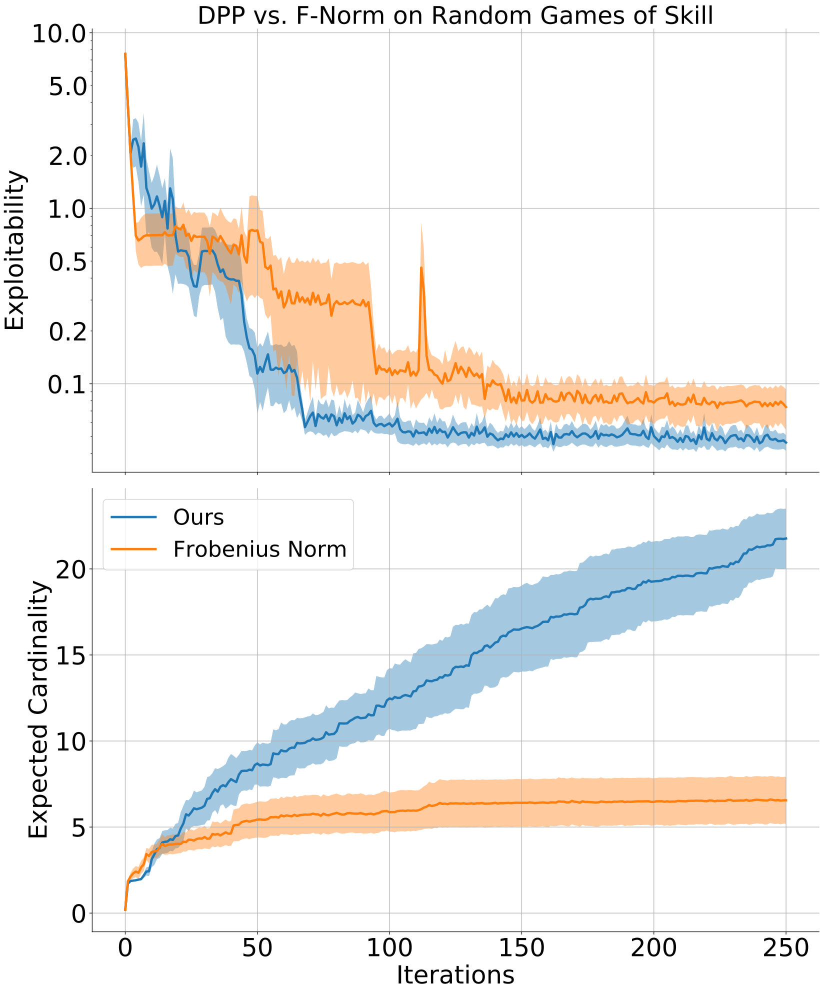



10,3 Blotto
10,4 Blotto
10,5 Blotto
Blotto
Connect Four
Elo Game
Elo game + noise=0.5
Go (board size = 3)
Go (board size = 4)
Self-play
PSRO
PSRO-rN
P-PSRO
Ours
Exploitability
Expected Cardinality
Iterations
0.1
0.2
0.5
1.0
2.0
0
5
10
15
20
30
40
50
100
150
200
300

# F. Empirical Result of Frobenius Norm vs. Expected Cardinality

We illustrate here a counter-example of when replacing our expected cardinality term in Eq. (13) with a Frobenius norm measure on the meta-game $\mathbf{M}$ over the population $\mathbb{S}^1 \cup \{S_\theta\}$ will fail to improve both the expected cardinality of $\mathbf{M}$ and so be unable to break the exploitability plateau of PSRO.

Specifically, we use a 150 dimensional Random Game of Skill (Czarnecki et al., 2020) where 25 of the strategies in the underlying game are particularly bad. In this setting, the Frobenius norm struggles as the poor-performing strategies all induce a large norm, and are therefore selected to be added to $\mathbb{S}^1$. On the other hand, at worst our expected cardinality measure will select one of these poor-performing strategies, but the remaining poor strategies are no longer diverse and are not selected. Therefore, whilst we suspect the Frobenius norm measure can work as a diversity measure, it is not as general as our metric and can fail under some circumstances, i.e. poor-performing strategies in the underlying game.



## G. Hyper-parameter Settings for All Experiments

The code can be found in: `https://github.com/oslumbers/diverse-psro`

*Table 1.* Hyper-parameter settings for our diverse-PSRO method *vs.* other baselines on four experiments.

| Settings | Value | Description |
|---|---|---|
| | **Real-World Meta-Games** | |
| Oracle method | Diverse best response | subroutine of getting oracles |
| Learning rate | 0.5 | Learning rate for agents |
| Improvement Threshold | 0.03 | Convergence criteria |
| Metasolver | Fictitious play | Metasolver method |
| Metasolver iterations | 1000 | # iterations for metasolver |
| # of threads in pipeline | 2 | # learners in pipeline PSRO |
| # of seeds | 10 | # trials |
| DPP weighting | 0.15 | Weight of the DPP best response |
| | **Non-Transitive Mixture Model** | |
| Oracle | Gradient Ascent | subroutine of getting oracles |
| Optimizer | Adam | Gradient ascent optimizer |
| Learning rate | 0.1 | Learning rate for optimizer |
| Betas | (0.9, 0.99) | Betas parameter for optimizer |
| $\boldsymbol{\pi}^i$ | $\boldsymbol{\pi}_k^i = \exp\left(-(x_i - \mu_k)^\top \Sigma (x_i - \mu_k)/2\right)$ | strategy from 2D coordinates |
| Game Payoff | $\boldsymbol{\pi}^{1,\top} \begin{bmatrix} 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & 0 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 0 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 0 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 0 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & 0 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{bmatrix} \boldsymbol{\pi}^2 + \frac{1}{2}\sum_{k=1}^{7}(\boldsymbol{\pi}_k^1 - \boldsymbol{\pi}_k^2)$ | Payoff for $\boldsymbol{\pi}^1$ against $\boldsymbol{\pi}^2$ |
| $\Sigma$ | $1/2\boldsymbol{I}$ | Covariance matrix for Gaussians |
| $\mu_1$ | $(2.872, -0.025)$ | position of the first Gaussian |
| $\mu_2$ | $(1.8105, 2.2298)$ | position of the second Gaussian |
| $\mu_3$ | $(1.8105, -2.2298)$ | position of the third Gaussian |
| $\mu_4$ | $(-0.61450, 2.8058)$ | position of the fourth Gaussian |
| $\mu_5$ | $(-0.61450, -2.8058)$ | position of the fifth Gaussian |
| $\mu_6$ | $(-2.5768, 1.2690)$ | position of the sixth Gaussian |
| $\mu_7$ | $(-2.5768, -1.2690)$ | position of the seventh Gaussian |
| Strategy initialisation variance | $0.1^2$ | Variance of Gaussian distribution at 0 |
| Metasolver | Fictitious play | Metasolver method |
| Metasolver iterations | 1000 | # iterations for metasolver |
| Iterations | 50 | # training iterations |
| DPP weight at iteration $t$ | $\dfrac{0.7}{1 + \exp\left(-0.25(t - 25)\right)}$ | Weight of the DPP best response |
| # of threads in pipeline | 4 | # learners in pipeline PSRO |
| # of seeds | 10 | # of trials |
| | **Colonel Blotto** | |
| Oracle method | Zero-order | subroutine of getting oracles |
| Oracle iterations | 50 | new strategies per iteration |
| $\mu$ | 0.1 | Learning rate |
| Metasolver | Linear Programming | Metasolver method |
| # of seeds | 10 | # of trials |
| DPP weight | 0.15 | Weight of the DPP best response |
| # of areas | 3 | # of areas to distribute coins |
| # of coins | 10 | # of coins to distribute |
| | **$\alpha$-PSRO** | |
| Oracle method | PBR / DPP-PBR | subroutine of getting oracles |
| Metasolver | $\alpha$-Rank | Metasolver Method |
| Iterations | 50 | # of training iterations |
| # of seeds | 20 | # of trials |
| $\alpha$ | Infinite | The $\alpha$ in $\alpha$-Rank |

# H. Implementation of Oracles

## H.1. Pseudocodes

**Algorithm 2** Diverse Best Response Oracle

```
Inputs:
Player Populations $\mathbb{S}_t = \prod_{i \in \mathcal{N}} \mathbb{S}_t^i$, with $S_t^i \in \mathbb{S}_t^i$ parametrised by $\theta_{S_t^i}$
Metapolicies $\boldsymbol{\pi}_t = \prod_{i \in \mathcal{N}} \boldsymbol{\pi}_t^i$
Learning rate $\mu$
Diversity probability $\lambda$

function oracle($\mathbb{S}_t, \boldsymbol{\pi}_t, \mu, \lambda$)
    Compute $\mathbf{BR}_{qual} = \mathbf{BR}^i(\mathbb{S}_t^{-i}, \boldsymbol{\pi}_t^{-i})$
    for each pure strategy $P_j$ do:
        Update meta-payoff $\mathbf{M}_j = \mathbf{M}(\mathbb{S}_t^i \cup \{P_j\})$
    Compute $\mathbf{BR}_{div} = \arg\max_{P_j} \left( \mathrm{Tr}\left( \mathbf{I} - (\mathbf{M}_j \mathbf{M}_j^\top + \mathbf{I})^{-1} \right) \right)$.
    Choose $\mathbf{BR} = \mathbf{BR}_{div}$ with probability $\lambda$ else $\mathbf{BR} = \mathbf{BR}_{qual}$
    Update $\theta_{S_t^i} = \mu \theta_{S_t^i} + (1 - \mu) \theta_{\mathbf{BR}}$
    Return: $S_t^i$
```

**Algorithm 3** Diverse Gradient Ascent Oracle

```
Inputs:
Player Populations $\mathbb{S}_t = \prod_{i \in \mathcal{N}} \mathbb{S}_t^i$, with $S_t^i \in \mathbb{S}_t^i$ parametrised by $\theta_{S_t^i}$
Metapolicies $\boldsymbol{\pi}_t = \prod_{i \in \mathcal{N}} \boldsymbol{\pi}_t^i$
Number of training updates $N_{train}$
Diversity weight $\lambda$

function oracle($\mathbb{S}_t, \boldsymbol{\pi}_t, N_{train}, \lambda$)
    Randomly initialise a new $S^{train}$
    for $j = 1, \ldots, N_{train}$:
        Compute payoff $p_j$ of $S^{train}$
        Compute meta-payoff $\mathbf{M}_j = \mathbf{M}(\mathbb{S}_t^i \cup \{S^{train}\})$
        Compute diversity $d_j = \mathrm{Tr}\left( \mathbf{I} - (\mathbf{M}_j \mathbf{M}_j^\top + \mathbf{I})^{-1} \right)$
        Compute loss $l_j = -(1 - \lambda) p_j - \lambda d_j$
        Update $\theta_{S^{train}}$ to minimise $l_j$ using a gradient based optimization method
    Return: $S^{train}$
```

**Algorithm 4** Diverse Zero-order Oracle

```
Inputs:
Player Populations $\mathbb{S}_t = \prod_{i \in \mathcal{N}} \mathbb{S}_t^i$, with $S_t^i \in \mathbb{S}_t^i$ parametrised by $\theta_{S_t^i}$
Metapolicies $\boldsymbol{\pi}_t = \prod_{i \in \mathcal{N}} \boldsymbol{\pi}_t^i$
Learning rate $\mu$
Noise parameter $\sigma$
Diversity weight $\lambda$

function $\texttt{oracle}(\mathbb{S}_t, \boldsymbol{\pi}_t, \mu, \sigma, \lambda)$
    Sample a single $S_t^i \in \mathbb{S}_t^i$ with probability $\boldsymbol{\pi}_t^i$
    for perturbation $j \in 1, 2, \ldots$ do:
        $S_j = \texttt{RandomPerturbation}(S_t^i, \mu, \sigma)$
        Update meta-payoff $\mathbf{M}_j = \mathbf{M}(\mathbb{S}_t^i \cup \{S_j\})$
        Compute payoff $p_j$ of $S_j$
        Compute diversity $d_j = \mathrm{Tr}\left(\mathbf{I} - (\mathbf{M}_j \mathbf{M}_j^\top + \mathbf{I})^{-1}\right)$ .
    Return: $\arg\max_{S_j} (1 - \lambda) p_j + \lambda d_j$
function $\texttt{RandomPerturbation}(S_t^i, \mu, \sigma)$
    Generate noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
    Add noise $\theta_{S_t^i} = \mu \theta_{S_t^i} + (1 - \mu)\varepsilon$
    Return $S_t^i$
```

**Algorithm 5** Diverse Reinforcement Learning Oracle

```
Inputs:
Player Populations $\mathbb{S}_t = \prod_{i \in \mathcal{N}} \mathbb{S}_t^i$, with $S_t^i \in \mathbb{S}_t^i$ parametrised by $\theta_{S_t^i}$
Metapolicies $\boldsymbol{\pi}_t = \prod_{i \in \mathcal{N}} \boldsymbol{\pi}_t^i$,
Learning rate $\mu$,
Noise parameter $\sigma$,
Diversity weight $\lambda$,
function $\texttt{oracle}(\mathbb{S}_t, \boldsymbol{\pi}_t, \mu, \sigma, \lambda)$
    Use an RL algorithm to find $\mathbf{BR}_{RL} = \mathbf{BR}_{RL}(\mathbb{S}_t^{-i}, \boldsymbol{\pi}_t^{-i})$
    for perturbation $j \in 1, 2, \ldots$ do:
        $\mathbf{BR}_j = \texttt{RandomPerturbation}(\mathbf{BR}_{RL}, \mu, \sigma)$
        Update meta-payoff $\mathbf{M}_j = \mathbf{M}(\mathbb{S}_t^i \cup \{\mathbf{BR}_j\})$
        Compute payoff $p_j$ of $S_j$
        Compute diversity $d_j = \mathrm{Tr}\left(\mathbf{I} - (\mathbf{M}_j \mathbf{M}_j^\top + \mathbf{I})^{-1}\right)$ .
    Return: $\arg\max_{S_j} (1 - \lambda) p_j + \lambda d_j$
function $\texttt{RandomPerturbation}(\mathbf{BR}_{RL}, \mu, \sigma)$
    Generate noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
    Add noise $\theta_{\mathbf{BR}} = \mu \theta_{\mathbf{BR}_{RL}} + (1 - \mu)\varepsilon$
    Return BR
```

### H.2. Time Complexity: PSRO vs. Diverse PSRO on Normal-Form Games

**Time Complexity of PSRO on Solving Normal-Form Games**

**Input:** Candidate population $\Pi \in \mathbb{R}^{P \times D}, P \ll D$, game $\boldsymbol{G} \in \mathbb{R}^{D \times D}$

1. Compute meta-game $\boldsymbol{M} = \Pi \boldsymbol{G} \Pi^\top \rightarrow \mathcal{O}(PDP + DDP) = \mathcal{O}(P^2D + D^2P)$
2. Compute metanash $\pi$ of meta-game $\boldsymbol{M} \in \mathbb{R}^{P \times P} \rightarrow \mathcal{O}(P^3)$
3. Aggregate the population of strategies $q = \pi^\top \Pi \rightarrow \mathcal{O}(PD)$
4. **For** each pure strategy $p_i$ of a total of D **do**
   - Compute payoff $p_i \boldsymbol{G} q \rightarrow \mathcal{O}(D^2)$

**Then PSRO takes** $\mathcal{O}(P^2D + D^2P + P^3 + PD + D^3) = \mathcal{O}(D^3)$

**Time Complexity of Diverse PSRO on Solving Normal-Form Games**

**Input:** Candidate population $\Pi \in \mathbb{R}^{P \times D}, P \ll D$, game $\boldsymbol{G} \in \mathbb{R}^{D \times D}$

1. Compute meta-game $\boldsymbol{M} = \Pi \boldsymbol{G} \Pi^\top \rightarrow \mathcal{O}(PDP + DDP) = \mathcal{O}(P^2D + D^2P)$
2. Compute metanash $\pi$ of meta-game $\boldsymbol{M} \in \mathbb{R}^{P \times P} \rightarrow \mathcal{O}(P^3)$
3. Aggregate the population of strategies $q = \pi^\top \Pi \rightarrow \mathcal{O}(PD)$
4. **For** each pure strategy $p_i$ of a total of D **do**
   - Compute payoff $p_i \boldsymbol{G} q \rightarrow \mathcal{O}(D^2)$
   - Update training strategy resulting in $\mathbf{BR}_{div} \rightarrow \mathcal{O}(D)$
   - Compute meta-game $\boldsymbol{M} = \Pi \boldsymbol{G} \Pi^\top \rightarrow \mathcal{O}(PDP + DDP) = \mathcal{O}(P^2D + D^2P)$
   - Compute metanash $\pi$ of meta-game $\boldsymbol{M} \in \mathbb{R}^{P \times P} \rightarrow \mathcal{O}(P^3)$
   - Compute $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathsf{M}}\boldsymbol{\mathsf{M}}^\top \rightarrow \mathcal{O}(P^3)$
   - Compute expected cardinality of $\boldsymbol{\mathcal{L}}$ in Proposition 4, the complexity is $\mathcal{O}(P^3)$

**Then Diverse PSRO takes** $\mathcal{O}(P^2D + D^2P + P^3 + PD + D^3 + D^2 + PD^3 + P^2D^2 + 3P^3) = \mathcal{O}(PD^3)$

# References

Czarnecki, W. M., Gidel, G., Tracey, B., Tuyls, K., Omidshafiei, S., Balduzzi, D., and Jaderberg, M. Real world games look like spinning tops. *arXiv*, pp. arXiv–2004, 2020.

Gillenwater, J. A., Kulesza, A., Vassilvitskii, S., and Mariet, Z. E. Maximizing induced cardinality under a determinantal point process. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 6911–6920. Curran Associates, Inc., 2018.

Kulesza, A., Taskar, B., et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

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

McAleer, S., Lanier, J., Fox, R., and Baldi, P. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games. *arXiv preprint arXiv:2006.08555*, 2020.

Muller, P., Omidshafiei, S., Rowland, M., Tuyls, K., Perolat, J., Liu, S., Hennes, D., Marris, L., Lanctot, M., Hughes, E., et al. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2019.

Omidshafiei, S., Papadimitriou, C., Piliouras, G., Tuyls, K., Rowland, M., Lespiau, J.-B., Czarnecki, W. M., Lanctot, M., Perolat, J., and Munos, R. $\alpha$-rank: Multi-agent evaluation by evolution. *Scientific reports*, 9(1):1–29, 2019.

Rising, J. K. Advances in the theory of determinantal point processes. 2013.