Towards Unifying Behavioral and Response Diversity for Open-ended Learning in Zero-sum Games

Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, Zhipeng Hu, Yaodong Yang

1Shanghai Jiao Tong University, 2Netease Fuxi AI Lab, 3Institute for AI, Peking University
liuxiangyu999@sjtu.edu.cn, ying.wen@sjtu.edu.cn, yaodong.yang@pku.edu.cn

Abstract

Measuring and promoting policy diversity is critical for solving games with strong non-transitive dynamics where strategic cycles exist, and there is no consistent winner (e.g., Rock-Paper-Scissors). With that in mind, maintaining a pool of diverse policies via open-ended learning is an attractive solution, which can generate auto-curricula to avoid being exploited. However, in conventional open-ended learning algorithms, there are no widely accepted definitions for diversity, making it hard to construct and evaluate the diverse policies. In this work, we summarize previous concepts of diversity and work towards offering a unified measure of diversity in multi-agent open-ended learning to include all elements in Markov games, based on both Behavioral Diversity (BD) and Response Diversity (RD). At the trajectory distribution level, we re-define BD in the state-action space as the discrepancies of occupancy measures. For the reward dynamics, we propose RD to characterize diversity through the responses of policies when encountering different opponents. We also show that many current diversity measures fall in one of the categories of BD or RD but not both. With this unified diversity measure, we design the corresponding diversity-promoting objective and population effectivity when seeking the best responses in open-ended learning. We validate our methods in both relatively simple games like matrix game, non-transitive mixture model, and the complex Google Research Football environment. The population found by our methods reveals the lowest exploitability, highest population effectivity in matrix game and non-transitive mixture model, as well as the largest goal difference when interacting with opponents of various levels in Google Research Football.

1 Introduction

Many zero-sum games have a strong non-transitive component [2, 4] in the policy space, and thus each player must acquire a diverse set of winning strategies to achieve high unexploitability [36], which has been widely validated by recent studies of constructing AIs with superhuman performance in sophisticated tasks, such as StarCraft [30, 28] and DOTA2 [25, 39]. The non-transitivity in games means there is not a dominating strategy and the set of strategies form a cycle (e.g., the endless cycles among Rock, Paper and Scissors). It is the presence of this special structure in games that requires players to maintain a diverse set of policies. Otherwise, we only need to seek the strongest one. Formally, the necessity of diversity for zero-sum games lies in three ways: (1) policy evaluation: with the presence of non-transitivity, one cannot justify the strength or weakness of a strategy through the outcome of the interaction with a single type of opponent; (2) avoiding being exploited [24]: since in non-transitive games a single strategy can be always beaten by another one, a diverse set of strategies...
allows players to make corresponding responses when encountering different opponents; (3) training adaptable strategies [32]: a diverse set of training opponents helps gradually eliminate the weakness of a strategy, which can adapt to a wide range of opponents with very few interactions at test time.

The open-ended learning framework is a promising direction towards inducing a population of distinct policies in zero-sum games via auto-curricula. Although various open-ended algorithms have been proposed to derive diverse strategies [2, 24, 6, 22], there are no consistent definitions for diversity. One of the most intuitive principles to characterize diversity is to build metrics over the trajectory or state-action distribution [6, 20]. However, this perspective only focuses on the policy behaviors and ignores the reward attributes inherited from the Markov decision process. We argue that this is not reasonable since sometimes a slight difference in the policy can result in a huge difference in the final reward like, in maze. Contrary to this, another line of works builds the diversity measure over empirical payoffs [24, 2, 10], thus revealing the underlying diverse behaviors of a strategy through the responses when encountering distinct opponents.

In this work, based on all previous diversity concepts, we work towards offering a unified view for diversity in an open-ended learning framework by combining both the behavioral attribute and the response attribute of a strategy. The behavioral diversity is formulated through the occupancy measure, which is an equivalent representation of a policy. We hypothesize that the diversity in policy behaviors should be revealed by differences in the state-action distribution, and we use a general divergence family $f$-divergence to indicate the novelty of a new policy. On the other hand, gamescape [2] has been proposed to represent the response capacity of a population of strategies. Based on gamescape, we formulate a new geometric perspective to treat the response diversity by considering the distance to the gamescape.

To summarize, in this paper, we provide the following contributions:

- We formulate the concept of behavioral diversity in the state-action space as the discrepancies of occupancy measures and analyze the optimization methods in both normal-form games and general Markov games.
- We provide a new geometric perspective on the response diversity as a form of Euclidean projection onto the convex hull of the meta-game to enlarge the gamescape directly and propose the optimization lower bound for practical implementation.
- We analyze the limitation of exploitability as the evaluation metric and introduce a new metric with theoretical soundness called population effectivity, which is a fairer way to represent the effectiveness of a population than exploitability [16].

## 2 Preliminaries

### 2.1 Markov Games

The extension of Markov decision processes (MDPs) with more than one agents is commonly modelled as Markov games [18]. A Markov game with $N$ agents is defined by a tuple $< N, S, \{A_i\}_{i=1}^{N}, P, \{r_i\}_{i=1}^{N}, \eta, \gamma >$, where $S$ denotes the state space and $A_i$ is the action space for agent $i$. The function $P$ controls the state transitions by the current state and one action from each agent: $P : S \times A_1 \times \cdots \times A_N \rightarrow \mathcal{P}(S)$, where $\mathcal{P}(S)$ denotes the set of probability distributions over the state space $S$. Given the current state $s_t$ and the joint action $(a_1, \ldots, a_N)$, the transition probability to $s_{t+1}$ is given by $P(s_{t+1}|s_t, a_1, \ldots, a_N)$. The initial state distribution is given by $\eta : S \rightarrow [0, 1]$. Each agent $i$ also has an associated reward function $r_i : S \times A_i \times \cdots \times A_N \rightarrow \mathbb{R}$. Each agent’s goal is to maximize the $\gamma$-discounted expected return $R_i = E[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)]$, where $i$ is a compact representation of all complementary agents of $i$. Specifically, for zero-sum games, the rewards satisfy that $\sum_{i=1}^{N} r_i(s, a) = 0$, and players need to behave competitively to achieve higher rewards.

In multi-agent reinforcement learning (MARL) [37], each agent is equipped with a policy $\pi_i : S \times A_i \rightarrow [0, 1]$ and the joint policy is defined by $\pi(a|s) = \Pi_{i=1}^{N} \pi_i(a_i|s)$. In single-agent reinforcement learning, occupancy measure is a principled way to characterize a policy, which indicates how a policy covers the state-action space. Inspired by the definition from the single-agent setting, we define the joint occupancy measure in MARL induced by the joint policy $\pi(a|s)$ as:
Then the distribution of state-action pairs \( \rho_{\pi}(s, a) = \rho_{\pi}(s)\pi(a|s) \) is called occupancy measure of the joint policy \( \pi \).

### 2.2 Policy Space Response Oracle

Adapted from double oracle [23], policy space response oracle (PSRO) [16, 5] has been serving as a powerful tool to solve the nash equilibrium (NE) in zero-sum games. In PSRO, each player maintains a pool of policies, say \( \mathcal{P}_i = \{\pi_1^i, \ldots, \pi_{M}^i\} \) for player \( i \) and \( \mathcal{P}_{-i} = \{\pi_1^{-i}, \ldots, \pi_{N}^{-i}\} \) for player \(-i\). The so-called meta game \( A_{\pi_i, \mathcal{P}_{-i}} \) has its \((k, j)\) entry as \( \phi_i(\pi_k^i, \pi_j^{-i}) \), where the function \( \phi_i \) encapsulates the reward outcome for player \( i \) like the winning rate or expected return. When player \( i \) adds a new policy \( \pi_{M+1}^i \), it will compute the best response to the mixture of its opponents:

\[
\text{Br}(\mathcal{P}_{-i}) = \max_{\mathcal{P}_{-i}} \sum_{k,j} \sigma^{j}_{-i} E_{\pi_{M+1}^i, \pi_j^{-i}}[r_i(s, a)].
\]

where \( \sigma = (\sigma_i, \sigma_{-i}) \) is a distribution over policies in \( \mathcal{P}_i \) and \( \mathcal{P}_{-i} \), which is usually a NE of \( A_{\mathcal{P}_i, \mathcal{P}_{-i}} \).

The empirical gamescape is introduced by [2] to represent the expressiveness of a population \( \mathcal{P}_i \) in the reward outcome level given the opponent population \( \mathcal{P}_{-i} \):

**Definition 2.** Given population \( \mathcal{P}_i \) and \( \mathcal{P}_{-i} \) with evaluation matrix \( A_{\mathcal{P}_i, \mathcal{P}_{-i}} \), the corresponding empirical gamescape (EGS) for \( \mathcal{P}_i \) is defined as

\[
G_{\mathcal{P}_i, \mathcal{P}_{-i}} := \{\text{convex mixtures of rows of } A_{\mathcal{P}_i, \mathcal{P}_{-i}}\}.
\]

### 2.3 Existing Diversity Measures

As the metric to measure the differences between models, diversity is an important topic in many fields of machine learning, including generative modelling [6], latent variable models [33], and robotics [1]. Specifically, in reinforcement learning (RL), diversity is a useful tool for learning transferable skills [7], boosting explorations [27], or collecting near-optimal policies that are distinct in a meaningful way. Despite the importance of diversity, as shown in Table 1, there has not been a consistent definition of diversity for RL, and various diversity concepts are used. [22] investigated behavioral diversity in multi-agent reinforcement learning through expected action variation, which is modeled as the average total variation distance of two action distributions under certain sampled states. Considering the geometric perspective that the determinant of the kernel matrix represents the volume of a parallelepiped spanned by feature maps, DvD [26] proposed the concept of population diversity using the determinant of the kernel matrix composed by the behavioral embeddings by multiple policies. Thanks to the tools of empirical game theory analysis, diversity can be modeled from the perspective of the empirical game. Effective diversity [2] is formulated as the weighted \( L_{1,1} \) norm of the empirical payoff matrix, which emphasizes what opponents a policy can win against. Also inspired by determinantal point process (DPP) [14, 38], [24] uses the expected cardinality to measure the diversity of a population.
3 A Unified Diversity Measure

Motivated by bisimulation metrics [8] to measure the similarity of two states in MDPs: 
\[ d(s_i, s_j) = \max_{a \in A} (1 - c) \cdot |R_{s_i}^a - R_{s_j}^a| + c \cdot W_1 (\mathbb{P}_{s_i}^a, \mathbb{P}_{s_j}^a; d) \]
that considers both the immediate reward and the following transition dynamics, where \( c \in [0, 1] \) and \( W_1 \) is the 1-Wasserstein distance, we want to build a metric to measure the similarity of two policies in a given Markov game through the task-specific reward attributes and the interaction between policy behaviors and transition dynamics.

We will firstly model the interaction between policy behaviors and transition dynamics through the
where we use the general
Prediction Error for Approximate Optimization.

\[ \text{Leveraging the powerful tool from occupancy measure matching, we here propose an approximate} \]

\[ \text{measure matching algorithms in imitation learning try to minimize the same objective \[ \text{(Motivated by bisimulation metrics [8])} \]

\[ \text{method to maximize} \]

\[ \text{where} \]

\[ \text{For more general Markov games, computing the exact occupancy measure is intractable. However, notice that we are maximizing a} \]

\[ \text{we are justified in considering diversity from a perspective of the occupancy} \]

\[ \text{otherwise, the joint policy aggregated according the distribution of nash is} \]

\[ \text{Next, we will consider how to promote diversity in the framework of policy space response oracle. Suppose after} \]

\[ \text{which can maximize the discrepancy between old} \]

\[ \text{where we use the general} \]

\[ \text{We firstly investigate the objective under the one-step game by giving the following theorem:} \]

\[ \text{Proposition 1 (Theorem 2 of [29]). If} \]

\[ \text{Usually, the policy} \]

\[ \text{and policies among players are independent, the behavioral diversity can be simplified by:} \]

\[ \text{where} \]

\[ \text{Proof. See Appendix A.1.} \]

\[ \text{For more general Markov games, computing the exact occupancy measure is intractable. However, notice that we are maximizing a} \]

\[ \text{Prediction Error for Approximate Optimization. Inspired by random expert distillation [31], a neural network} \]

\[ \text{a neural network} \]

\[ \text{a neural network} \]

\[ \text{we propose an approximate} \]

\[ \text{method to maximize} \]
Alternative Solutions. There are also many other practical occupancy measure matching algorithms. One popular paradigm is learning a discriminator $D(s, a)$ to classify the state-action pair $(s, a)$ from the distribution $\rho_{\pi_i', \pi_{-i}}$ and the distribution $\rho_{\pi_E}$. Then the trained $D(s, a)$ can be used to construct different intrinsic rewards, which will correspond to different choices of $f$-divergence [13, 11, 9]. One major drawback of this paradigm is that the discriminator depends on the new policy $\pi_i'$ and needs re-training once the policy $\pi_i'$ is updated. Another popular paradigm is to learn an intrinsic reward directly from the target distribution $\rho_{\pi_E}$ like the prediction error. Besides using the prediction error, there are also other choices, including energy-based models (EBM) [19] and expert variance [3]. However, those methods usually require specialized training techniques.

3.2 Response Diversity via Convex Hull Enlargement

Take the two-player game for example. In games with more than two players, one can simply denote players other than player $i$ as player $-i$. To the empirical payoff matrix, another fundamental way to characterize the diversity of a new policy is through the reward outcome from the interaction with many different opponents. Each row in the empirical payoff matrix embeds how the corresponding row policy behaves against different opponent policies. We are therefore justified in using the row vector of the empirical payoff matrix to represent the corresponding row policy.

Formally, assume the row player maintains a pool of policies $\Psi_i = \{\pi_i^1, \ldots, \pi_i^M\}$ and the column player maintains a pool of policies $\Psi_{-i} = \{\pi_{-i}^1, \ldots, \pi_{-i}^N\}$. The induced $(k, j)$ entry in the empirical payoff matrix $A_{\Psi_i \times \Psi_{-i}}$ is given by $\phi_i(\pi_i^k, \pi_{-i}^j)$, where the function $\phi_i$ encapsulates the reward outcome for player $i$ given the joint policy $(\pi_i^k, \pi_{-i}^j)$. Now we can define the diversity measure induced by the reward representations:

$$\text{Div}_{\text{rew}}(\pi_i^{M+1}) = D(a_{M+1} \| A_{\Psi_i \times \Psi_{-i}})$$

$$a_{M+1}^* := (\phi_i(\pi_i^{M+1}, \pi_{-i}^j))_{j=1}^N.$$

$D(a_{M+1} \| A_{\Psi_i \times \Psi_{-i}})$ essentially measures the diversity of the new vector $a_{M+1}$ given the presence of row vectors in $A_{\Psi_i \times \Psi_{-i}}$.

Inspired by the intuition of the convex hull that indicates the representational capacity of a pool of policies, the diverse new policy should seek to enlarge the convex hull of reward vectors as large as possible. To characterize the contribution of a vector to the enlargement of the convex hull directly, we define the novel diversity measure as a form of Euclidean projection:

$$\text{Div}_{\text{rew}}(\pi_i^{M+1}) = \min_{\beta \geq 1} \| A_{\Psi_i \times \Psi_{-i}}^T \beta - a_{M+1} \|^2_2.$$

Unfortunately, there is no closed-form solution to this optimization problem. To facilitate the optimization, we propose a practical and differential lower bound:

**Theorem 2.** For a given empirical payoff matrix $A$ and the reward vector $a_{M+1}$, the lower bound of $\text{Div}_{\text{occ}}$ is given by:

$$\text{Div}_{\text{rew}}(\pi_i^{M+1}) \geq \sigma_{\min}(A)(1 - 1^T (A^T)^+ a_{M+1})^2 + ||(I - A^T (A^T)^+) a_{M+1}||_2^2,$$

where $(A^T)^+$ is the Moore–Penrose pseudoinverse of $A^T$, and $\sigma_{\min}(A)$ is the minimum singular value of $A$.

**Proof.** See Appendix A.2.

Let $F(\pi_i^{M+1})$ be the right hand of the inequality. Then $F(\pi_i^{M+1})$ serves as a lower bound of $\text{Div}_{\text{rew}}(\pi_i^{M+1})$.

4 A Unified Diverse Objective for Best Response

Equipped with the unified diversity measure, we are ready to propose the diversity-aware response during each iteration of PSRO:

$$\arg \max_{\pi_i'} \mathbb{E}_{s, a \sim \rho_{\pi_i', \pi_{-i}}} [r(s, a)] + \lambda_1 \text{Div}_{\text{occ}}(\pi_i') + \lambda_2 \text{Div}_{\text{rew}}(\pi_i')$$

If both $\lambda_1$ and $\lambda_2$ are 0, then objective is a normal best response.
Assume the policy weighted intrinsic reward. Another issue we need to address is to sample
Algorithm 1
Therefore, the smaller exploitability means the joint policy is closer to the Nash equilibrium.

Joint Optimization.
It shows how much each player gains by deviating to their best responses:

\[ \text{Expl}(\pi) = \sum_{i=1}^{N} \left( \max_{\pi_i} \phi_i(\pi_i, \pi_{-i}) - \phi_i(\pi_i, \pi_{-i}) \right). \] (9)

Therefore, the smaller exploitability means the joint policy is closer to the Nash equilibrium.

### 4.1 On the Optimization of Diverse Regularizers

**Discussions on Optimizing Behavioral Diversity.** As discussed in Section 3.1, diversity in the occupancy measure level is fully compatible with the reinforcement learning task since the agent can get an intrinsic reward \( r^{\text{int}}_i(s, a) \) to indicate the novelty of a state-action pair \((s, a)\). Therefore, to optimize the first two items in the objective, we only need to add the original reward by the \( \lambda_1 \)-weighted intrinsic reward. Another issue we need to address is to sample \((s, a)\) from the distribution \( \rho_{\pi_E} \), which has been mentioned in Section 3. Since \( \pi_E \) is not a true policy but only a hypothetical policy aggregated according to the mixture \((\sigma_i, \sigma_{-i})\), sampling from \( \rho_{\pi_E} \) is equivalent to sampling from \( \rho_{\pi^*_i, \pi^*_j} \) with probability \( \sigma_i \sigma_j \).

**Discussions on Optimizing Response Diversity.** Optimizing \( \text{Div}_{\text{rew}} \) is not so easy since it involves an inner minimization problem. Fortunately, we have derived a closed-form low-bound \( F \), which can serve as a surrogate for the outer maximization.

Assume the policy \( \pi'_i \) is parameterized by \( \theta \) as \( \pi'_i(\theta) \). Then the gradient of \( F \) with respect to \( \theta \) is given by:

\[
\frac{\partial F}{\partial \theta} = \frac{\partial F}{\partial \theta} \frac{\partial F}{\partial a_{M+1}} = \left( \frac{\partial \phi_i(\pi'_i(\theta), \pi^1_{-i})}{\partial \theta}, \ldots, \frac{\partial \phi_i(\pi'_i(\theta), \pi^N_{-i})}{\partial \theta} \right) \frac{\partial F}{\partial a_{M+1}}.
\]

\( \frac{\partial F}{\partial a_{M+1}} \) controls weights of the policy gradient backpropagated from different opponents policies \( \pi_{-i} \).

For practical implement, we sample an opponent \( j \) according to the absolute values of \( \frac{\partial F}{\partial a_{M+1}} \) and then train \( \pi'_i \) against the opponent \( \pi^j_{-i} \) using gradient descent or ascent, which depends on the sign of the \( j \)th entry of \( \frac{\partial F}{\partial a_{M+1}} \).

**Joint Optimization.** One issue worth our notice is that the update direction of \( \text{Div}_{\text{rew}} \) will heavily rely on the initialization of \( \pi'_i(\theta) \). A bad initialization of \( \theta \) will make the response diversity tell \( \pi'_i \) to update toward worse rewards. Therefore, we propose to first optimize the first two items in the objective jointly and then optimize \( \pi'_i \) using \( \text{Div}_{\text{rew}} \). The final unified diverse response with gradient-based optimization is described in Algorithm 1.

In addition to the gradient-based optimization, we also provide other kinds of optimization oracles suitable for different games. Pseudocodes can be found in Appendix G.

### 4.2 Evaluation Metrics

**Exploitability.** Exploitability [16] measures the distance of a joint policy from the Nash equilibrium. It shows how much each player gains by deviating to their best responses:

\[ \text{Expl}(\pi) = \sum_{i=1}^{N} \left( \max_{\pi_i} \phi_i(\pi_i, \pi_{-i}) - \phi_i(\pi_i, \pi_{-i}) \right). \] (9)
**Population Effectivity.** Note the limitation of exploitability is that it only measures how exploitable a single joint policy is. Therefore, to evaluate the effectiveness of a population, we first need to get an aggregated policy from a population, and we usually use the Nash aggregated policy output by PSRO. Since the Nash is computed over the meta game, which varies with the opponents, the aggregation may be sub-optimal and cannot be used to represent a population. Intuitively, the aggregation weights, and further, the evaluation of a population should not be determined by the population that a specific opponent holds. To address this issue, we propose a generalized opponent-free concept of exploitability called Population Effectivity (PE) by looking for the optimal aggregation in the worst cases:

$$\text{PE}(\{\pi_i^k\}_{k=1}^N) = \min_{\pi_i} \max_{\alpha_i \geq 0} \sum_{k=1}^N \alpha_k \phi_i(\pi_i^k, \pi_{-i}) .$$  \hfill (10)

PE is again a NE over a two-player zero-sum game, where the player owning the population optimizes towards an optimal aggregation denoted by $\alpha$, while the opponent can search over the entire policy space. Similar ideas of revealing the relationship between the minimax objective and diversity can be also found in [40, 34]. Next, we offer a simple example to further illustrate the limitations of exploitability and superiority of PE.

**Example 1.** Consider the matrix game Rock-Scissor-Paper, where $\phi_1(\pi_1, \pi_2) = \pi_1^T A \pi_2$ and $\phi_2(\pi_2, \pi_1) = -\pi_2^T B \pi_1$, $\pi_1 \in \mathbb{R}^3$, $\pi_2 \in \mathbb{R}^3$, $A = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$, $B = -A^T$. Suppose player 1 holds the population $\mathcal{P}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, i.e. $\{\text{Rock, Scissor, Paper}\}$ and $\mathcal{P}_2 = \{1, 0\}$, i.e. $\{\text{Rock}\}$. Then the meta-game $A_{\mathcal{P}_1 \times \mathcal{P}_2} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$. The nash aggregated joint policy $(\pi_1, \pi_2) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. Now we can compute $\text{Expl}((\pi_1, \pi_2))$ as:

$$\text{Expl}((\pi_1, \pi_2)) = \max_{\pi_1} \phi_1(\pi_1', \pi_2) - \phi_1(\pi_1, \pi_2) + \max_{\pi_2} \phi_2(\pi_2', \pi_1) - \phi_2(\pi_2, \pi_1) \hfill (11)$$

$$= \max_{\pi_1} \phi_1(\pi_1', \pi_2) + \max_{\pi_2} \phi_2(\pi_2', \pi_1) = 2. \hfill (12)$$

Now notice that the contribution of player 1 to the exploitability is $\max_{\pi_1'} \phi_1(\pi_1', \pi_2)$, which equals 1. However, it is not reasonable that player 1 and 2 have the same contribution to the exploitability since player 1 has a perfect diverse policy set. Instead, if we use PE as the metric:

$$\text{PE}(\mathcal{P}_1) = 0 ,$$

which justifies that player 1 has already found a perfect population.

In the following theorem, we show that PE is a generalized notion of exploitability under certain conditions and has some desirable properties:

**Theorem 3.** Population effectivity has the following properties:

**P1. Equivalence:** If $N = 1$ and the underlying game $\phi_i(\cdot, \cdot)$ is a symmetric two-player zero-sum game, PE is equivalent to exploitability.

**P2. Monotonicity:** If there are two populations $\mathcal{P}_i$, $\Omega_i$ and $\mathcal{P}_i \subseteq \Omega_i$, then $\text{PE}(\mathcal{P}_i) \leq \text{PE}(\Omega_i)$, while the relationship for exploitability of the Nash aggregated policies of $\mathcal{P}_i$ and $\Omega_i$ may or may not hold.

**P3. Tractability:** If the underlying game $\phi_i(\cdot, \cdot)$ is a matrix game, then computing PE is still solving a matrix game.

**Proof.** See Appendix A.3. \hfill $\square$
Figure 1: (a): Exploitability vs. training iterations (the number of times the optimization oracle is called) on the AlphaStar game. (b): Negative Population Effectivity vs. training iterations on the AlphaStar game. Ablation studies of P-PSRO only with BD or RD are also reported.

Figure 2: Exploration trajectories during training process on Non-Transitive Mixture Games.

5 Experiments

To verify that our diversity-regularized best response algorithm can induce a diverse and less exploitable population, we compare our methods with state-of-the-art game solvers including Self-play [12], PSRO [16], PSRO-rN [2], Pipeline-PSRO (P-PSRO) [21], DPP-PSRO [24]. In this section, we want to demonstrate the effectiveness of our method to tackle the non-transitivity of zero-sum games, which can be shown via higher PE, lower exploitability, and diverse behaviors. Beyond the simple games, we also have the results on the complex Google Research Football game, and our methods can still work. In all the following experiments, we choose the appropriate diversity weights $\lambda_1$ and $\lambda_2$ by extensive hyper-parameter tuning. We also conduct ablation study by choosing different $\lambda_1$ and $\lambda_2$ in Appendix F. The environment details are in Appendix D, and the hyper-parameter settings for each experiment are in Appendix E.3.

Real-World Games. [4] studies the properties of some complex real-world games, including AlphaStar and AlphaGo. We test our method on the empirical games generated through the process of solving these real-world games. In Figure 1a, we report the exploitabilities of different algorithms during solving the AlphaStar game, which contains the meta-payoffs for 888 RL policies. We report values of exploitability and PE during the growth of the population in Figure 1a and Figure 1b. The result shows that with our unified diversity regularizer, our methods achieve the smallest exploitability and largest population effectivity, while most baselines fail to recover the diverse strategies and are easily exploited.

Non-Transitive Mixture Games. This zero-sum two-player game consists of $2l+1$ equally-distanced Gaussian humps on the 2D plane. Each player chooses a point in the 2D plane, which will be translated into a $(2l + 1)$-dimensional vector $\pi_i$ with each coordinate being the density in the corresponding Gaussian distribution. The payoff of the game is given by:

$$\phi_i(\pi_i, \pi_{-i}) = \pi_i^T S \pi_{-i} + 1^T (\pi_i - \pi_{-i}).$$

According to the delicately designed $S$, this game involves both the transitive component and non-transitive component, which means the optimal strategy should stay close to the center of the Gaussian and explore all the Gaussian distributions equally.

We firstly visualize the exploration trajectories during different algorithms solving the game in Figure 2. It shows that the best response algorithm regularized by both BD and RD achieves the most diverse trajectories. Although our algorithm finds the most diverse trajectories, such superiority is not revealed by the metric of exploitability in the last row of Table 2. On the other hand, we also report the PE values for the final population generated by different algorithms in Table 2. It can found that our unified diversity regularizer can always help PSRO dominate other baselines in terms of...
As depicted in Figure 3, policies trained by PSRO with both BD and RD achieve the largest goal differences when playing against the built-in bots. Moreover, they have an average of 60% win-rate over other baseline methods (see the table in the Appendix C). We do not report the results of DPP-PSRO since it needs evolutionary updates and cannot scale to such a complex setting. We also abandon the pipeline trick for ease of implement since it does not affect the relative performance among algorithms. Additionally, the discussion of robustness of policies trained with different methods, the network architectures, the hyperparameters, and other detailed experimental setups can also be found in Appendix C.

6 Conclusions

This paper investigated a new perspective on unifying diversity measures for open-ended learning in zero-sum games, which shapes an auto-curriculum to induce diverse yet effective behaviors. To this end, we decomposed the similarity measure of MDPs into behavioral and response diversity and showed that most of the existing diversity measures for RL can be concluded into one of the categories of them. We also provided the corresponding diversity-promoting objective and optimization methods, which consist of occupancy measure mismatching and convex hull enlargement. Finally, we proposed population effectiveness to overcome the limitation of exploitability in measuring diverse policies for open-ended algorithms. Experimental results demonstrated our method is robust to both highly non-transitive games and complex games like the Google Research Football environment.
We also include a discussion on the limitation of our method here. Regarding the limitations from the perspective of the proposed algorithm, one limitation is that the diversity weights $\lambda_1$ and $\lambda_2$ in our paper are manually tuned. Our methods target the non-transitivity in zero-sum games. Therefore, the weights for a game should be related to how strong the non-transitive component is. In future work, we will work towards quantifying the non-transitivity component of a given game automatically and determining the diversity weights corresponding. Similar ideas have been tested in single-agent RL cases on learning the discount rate [35]. In addition, our methods also inherit the limitations from the framework of PSRO. The advantage of the PSRO based methods also depends on the amount of non-transitivity of the environment [4]. To be specific, if there aren’t too many cyclic policies involved in the environment (such as Rock->Paper->Scissor), the newest model generated by the naive self-play training paradigm could probably be the strongest one (a Nash Policy), which implies that a naive self-play would suffice to solve the problem. The PSRO methods will then lose their advantage under such circumstances since they need extra computation resources to maintain meta-payoffs. Regarding the limitations from the perspective of real-world application, the game dynamics could be complex and there could be a lot of randomness involved in the real-world games. Thus, the approximated best response trained by reinforcement learning algorithms towards a certain policy could be inaccurate. In the Google Research Football experiment, we empirically save the checkpoint when the model win-rate is stable (i.e. the change of win-rate is less than 0.05 during two checks with the check frequency is 1000 model steps) or the training model steps reach an upper bound of 50000. These criteria could be potentially improved or further studied for all the PSRO based methods.

Acknowledgments

This work is supported by Shanghai Sailing Program (21YF1421900). The authors thank Minghuan Liu for many helpful discussions and suggestions.

References


