

Pick Your Battles: Interaction Graphs as Population-Level Objectives for Strategic Diversity

Extended Abstract

Marta Garnelo
DeepMind

Wojciech Marian Czarnecki
DeepMind

Siqi Liu
DeepMind

Dhruva Tirumala
DeepMind

Junhyuk Oh
DeepMind

Gauthier Gidel
DeepMind

Hado van Hasselt
DeepMind

David Balduzzi
DeepMind

ABSTRACT

Strategic diversity is often essential in games: in multi-player games, for example, evaluating a player against a diverse set of strategies will yield a more accurate estimate of its performance. Furthermore, in games with non-transitivities diversity allows a player to cover several winning strategies. However, despite the significance of strategic diversity, training agents that exhibit diverse behaviour remains a challenge. In this paper we study how to construct diverse populations of agents by carefully structuring how individuals within a population interact. Our approach is based on *interaction graphs*, which control the flow of information between agents during training and can encourage agents to specialise on different strategies, leading to improved overall performance. We provide evidence for the importance of diversity in multi-agent training and analyse the effect of applying different interaction graphs on the training trajectories, diversity and performance of populations in a range of games.

KEYWORDS

Populations; Reinforcement Learning; Non-transitive games

ACM Reference Format:

Marta Garnelo, Wojciech Marian Czarnecki, Siqi Liu, Dhruva Tirumala, Junhyuk Oh, Gauthier Gidel, Hado van Hasselt, and David Balduzzi. 2021. Pick Your Battles: Interaction Graphs as Population-Level Objectives for Strategic Diversity: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, Online, May 3–7, 2021, IFAAMAS, 3 pages.

1 INTRODUCTION

Most interesting real-world games and tasks involve separate and potentially competing objectives (are **multi-agent** (MA) in nature) and do not admit a single winning strategy that beats all (i.e. have some **non-transitive** element). Training agents that master these types of games poses several challenges. For example, in MA games performance is only defined relative to other players, not absolutely. Additionally, as a result of the non-transitive nature of these games, performance against one opponent can often be uninformative or even misleading about performance against other opponents.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online. © 2021 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

In this paper we show that a number of the challenges associated with non-transitive MA environments can be addressed by maximising strategic diversity of agents. At a higher level, diversity of an agent population is beneficial in three different situations: 1. As a *learner population* diversity leads to better test time performance. 2. *Trainer populations* (the opponents of a learner population during training) constitute better training adversaries if they cover a diverse set of strategies. 3. Finally, diverse *evaluator populations* estimate the performance of the learner population more accurately.

Given the importance of strategic diversity we thus pose the question of how to methodically train such diverse populations. In this paper we opt for a population-level approach and define *structured population-level objectives* via interaction graphs that specify the objective of each agent in a population in terms of (mixtures of) other agents. A key observation is that training the same agent against different opponents results in different training behaviour as well as different final performance of said agent.

We use this insight to train strategically diverse agents by systematically selecting the opponents that an agent encounters during training. To this end we introduce *interaction graphs* as a framework to describe the training interactions between agents in a population. Depending on the properties of the graph the resulting population will exhibit varying levels of strategic diversity and performance. We study this effect for a number of different graphs on a modified version of Rock-Paper-Scissors, Blotto and Starcraft. Finally, it is important to note, that the setup analysed here is qualitatively different from a standard approach, where a best response is found with respect a distribution of previously trained (fixed) agents [3, 5]. Instead, interaction graphs describe a matchmaking schedule for co-training players. Our main contributions are:

- (1) We provide evidence for the importance of diversity in non-transitive MA games.
- (2) We introduce the interaction graph framework to describe the control of training interactions in populations.
- (3) We analyse the effect of different population graphs on the resulting populations for three different games.

2 EXPERIMENTS

Environments. We consider three environments with different types of non-transitive structures: (1) GMM-RPS, a variant of continuous Rock-Paper-Scissors (RPS) that combines the cyclic nature

of RPS with a transitive strength element for each pure strategy. (2) Colonel Blotto is a two-player, zero-sum resource distribution game. It is well-studied in game-theory where it's usually of particular interest because of its highly non-transitive strategy space. (3) The StarCraft II environment [6] is a real-time strategy 2-player game with highly non-transitive game dynamics. In addition to being significantly more complex than the previously described games it also has a temporal element that the other two lack.

2.1 Graphs

We compare nine different graph structures to start characterising the effect of restricting the training interactions within populations. We distinguish between fixed and adaptive interaction graphs; **Fixed interaction graphs** are defined at the beginning of training and remain fixed throughout. We compare fixed graphs with different properties: fully connected and self-play [4] graphs that maximise and minimise information flow respectively. Cyclical graphs with and without hierarchies as well as a non-cyclic hierarchical graph [3]. **Adaptive interaction graphs** start out fully connected and the edges are then continuously updated during training as a function of their relative performance against the other agents in the population (e.g. train against those that are better/worse than you) [1].

2.2 Evaluation metrics

Qualitative methods: because the strategy space of GMM-RPS is \mathbb{R}^2 , we can plot how the strategies of the different agents in a population change during training depending on the interaction graph.

Quantitative methods: we measure the effective diversity and relative population performance (RPP) [1] of the populations trained with different interaction graphs on GMM-RPS and Blotto (Given the complexity of Starcraft we evaluate the populations on different but related metrics of performance and diversity, see [2] for details). In order to measure the RPP we need an evaluator baseline. We compare learned populations with high and low measured effective diversity as well as a 'ground truth' population containing all the strategies in the Nash if they are known. Finally, we also measure the convergence of the agents' policies to evaluate whether they get stuck in cycles and the coverage to evaluate how much of the strategy space a population covers between all its agents.

3 DISCUSSION

We characterise the differences across all nine interaction graphs by looking at the evaluation metrics mentioned above. In the following we summarise the main findings.

Diverse evaluator populations estimate performance more accurately. We show that the performance of 90 populations evaluated by an evaluator population with high effective diversity matches the ground truth (absolute) performance better than that of a low diversity evaluator population. This effect seems to be stronger as the game increases in complexity.

Diverse learner populations perform better. We show that for the simple games high effective diversity is correlated with strong performance.

Graphs influence the training behaviour of populations. We show that the training trajectories that result from training on

different interaction graphs vary drastically for GMM-RPS. Given the simplicity of the game most populations display one of two possible behaviours: the agents either all synchronise within a population or they cover different areas of the action space independently. In general, interaction graphs that allow for cyclic training interactions cover all modes, while those that don't contain cycles end up cycling.

Graphs influence the effective diversity of populations. We can further quantify this behaviour by measuring the average effective diversity obtained by populations trained on the different interaction graphs. We show that interaction graphs that allow for cyclic interactions between agents have a larger spread across action space. For GMM-RPS and Blotto this spread translates to higher effective diversity.

Graphs with cycles encourage specialisation and increase effective diversity in simple non-transitive games. We show that the directed nature of cycles allows individual agents to focus on a subset of the population (that does not necessarily focus on them) and thus to specialise. Populations trained with undirected graphs, on the other hand, tend to collapse to the same strategy as the symmetry in the connections means agents have the same objective. As a result populations trained with cyclic interaction graphs have higher effective diversity.

A fixed graph structure is powerful when it matches the underlying game structure, otherwise adapting graphs might be a better choice. The hierarchical cycle, for example, works well on the RPS-like games as it matches the underlying structure. It does not, however, perform as well on Blotto which has a richer strategy structure. The adaptive graphs, on the other hand, find a good approximation in either case.

Focusing on those that are better than you makes you less exploitable and focusing on those that are worse than you makes you a better exploiter. Focusing on opponents that beat you involves learning best response against a more diverse set of strategies which encourages agents to be more robust. Playing against those you are beating already, on the other hand, allows agents to specialise. As a result populations might become more exploitable as they might get stuck on weak enemies.

Individual convergence is not as important as population-level convergence for diversity and coverage. Individual agents may cycle as long as all important strategies are covered.

When moving to significantly more complex environments some fundamental insights hold, but some do not. We have chosen simple games as a starting point for our analysis. While most insights hold across these games, they may not translate to significantly more complex games such as StarCraft. In fact, some intuitions, e.g. the usefulness of directed graphs, the fact that the wrong fixed graph can hinder learning or that focusing on agents you beat allows you to specialise seem to agree with the results obtained on StarCraft. However, it is also clear that one should be careful to translate graphs or particular methods directly from very simple environments to more complex ones. Stark difference in game dynamics might lead to unexpected failure modes (e.g. the collapse of rectified Nash onto a single Nash agent that it can't recover from) or unforeseen successes (e.g. the ability of the fully connected graph to explore the strategy space).

For an extended version of this paper we refer the reader to [2].

REFERENCES

- [1] David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech M Czarnecki, Julien Pérolat, Max Jaderberg, and Thore Graepel. 2019. Open-ended learning in symmetric zero-sum games. *International Conference on Machine Learning* (2019).
- [2] Marta Garnelo, Wojciech Marian Czarnecki, Siqi Liu, Dhruva Tirumala, Junhyuk Oh, Gauthier Gidel, William Hawkins, Hado van Hasselt, and David Balduzzi. 2021. Pick Your Battles: Interaction Graphs as Population-Level Objectives for Strategic Diversity. *arXiv preprint* (2021).
- [3] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*. 4190–4203.
- [4] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *Nature* 550, 7676 (2017), 354–359.
- [5] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.
- [6] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).