Towards Robust Ad Hoc Teamwork Agents By Creating Diverse Training Teammates

Arrasy Rahman, **Elliot Fosong**, **Ignacio Carlucho** and **Stefano V. Albrecht**
School of Informatics, University of Edinburgh, UK
{arrasy.rahman, e.fosong, ignacio.carlucho, s.albrecht}@ed.ac.uk

Abstract

Ad hoc teamwork (AHT) is the problem of creating an agent that must collaborate with previously unseen teammates without prior coordination. Many existing AHT methods can be categorised as type-based methods, which require a set of predefined teammates for training. Designing teammate types for training is a challenging issue that determines the generalisation performance of agents when dealing with teammate types unseen during training. In this work, we propose a method to discover diverse teammate types based on maximising best response diversity metrics. We show that our proposed approach yields teammate types that require a wider range of best responses from the learner during collaboration, which potentially improves the robustness of a learner's performance in AHT compared to alternative methods.

1 Introduction

Ad hoc teamwork (AHT) is the challenging problem of creating a single agent that must collaborate with previously unseen teammates without prior coordination. The controlled agent, which we refer to as the *learner*, shares an environment with collaborative teammates that have common objective with the learner. Note that the collaborative teammates may have other individual objectives which encourage specific focus on different skills to achieve the common objective [Mirsky *et al.*, 2022]. Effective collaboration in AHT then requires a learner which adapts its policy according to teammates' skills.

Many AHT approaches can be categorised as type-based methods [Albrecht *et al.*, 2016; Barrett *et al.*, 2017]. A type formally encapsulates all important information on a teammate's decision making process, which are unknown to the learner. Assuming access to teammates with predefined types during training, type-based AHT methods learn to infer teammates' types during interaction. Inferred teammate types are then used to select the learner's optimal action for collaboration.

Due to the diverse teammate types that can be encountered during interaction, it is often impossible to predefine all possible teammate types in many real-world AHT applications. Existing type-based methods can still collaborate with teammates with previously unseen types assuming that unknown types are a probabilistic mixture of types seen in training [Barrett *et al.*, 2017], or that all types can be mapped into a continuous type space by trained neural networks [Rahman *et al.*, 2021; Papoudakis *et al.*, 2021]. Nevertheless, type-based AHT methods can perform poorly if the teammates' behaviour are highly different from those seen in training.

In this work, we propose a method that generates a diverse set of collaborative teammates, that can be used for training learners that can collaborate with a wider range of teammate types. Our method optimises a diversity metric which is based on the cross-play performance of best-response policies for each generated teammate. We demonstrate that our approach discovers teammate policies which potentially improves learner's generalisation performance compared to prior methods employing alternative definitions of diversity [Lupu *et al.*, 2021].

2 Related Works

Type-based AHT: Type-based AHT methods [Barrett *et al.*, 2017; Albrecht *et al.*, 2016; Rahman *et al.*, 2021] assume access to predefined teammate policies for learning. In these methods, training data is gathered by letting the learner interact with the predefined teammates. The learner is subsequently trained to infer the types of teammates based on the limited experience that are gathered by interacting with them. Inferred teammate types are finally used to decide the learner's optimal action for collaborating with its teammates. For instance, a learner could employ an expert policy to choose its optimal action when collaborating with a certain type of teammate.

Teammate Policy Generation: Diverse teammate policy generation has been explored in previous work on AHT and other closely related problems such as Zero-Shot Coordination (ZSC) [Hu *et al.*, 2020]. Several works on this area formulate diversity in terms of information theoretic measures defined over the learner' generated trajectories [Parker-Holder *et al.*, 2020; Lupu *et al.*, 2021; Xing *et al.*, 2021; Lucas and Allen, 2022]. Despite its prevalence, previous works [Lupu *et al.*, 2021; Liu *et al.*, 2021] highlighted that training with teammates generated by tra-

jectory diversity-based methods does not always lead to improved learner's robustness. This is because many teammate behaviours producing distinct trajectories entail the same learner's best response policy during collaboration. While Liu *et al.* [2021] also proposed an approach based on the best response policies' performance, their approach is limited to zero-sum games.

3 Background & Setting

Interaction between the learner, its teammates, and the environment is modelled as a Stochastic Game (SG), which is a tuple defined as $< N, S, \{\mathcal{A}^i\}_{i=1}^{|N|}, P, \{\mathcal{R}^i\}_{i=1}^{|N|}, \mu >$. SGs extends the concept of MDP to problems where multiple agents interact with each other. N, $\mathcal{S}$, and $\mathcal{A}_i$ denote the set of agents, state space, and action space of agent i respectively. $P : S \times \mathcal{A}^1 \times ... \times \mathcal{A}^{|N|} \rightarrow \Delta S$ and $\mathcal{R}^i : S \times \mathcal{A}^1 \times ... \times \mathcal{A}^{|N|} \rightarrow \mathbb{R}$ denotes the transition function and agent i's reward function, respectively.

While our approach can be extended to more complicated settings, this work only considers two-player games ($N = \{1, 2\}$). We also assume that the SG's reward function only models the shared objectives of the collaborative teammates being generated, i.e. $\mathcal{R}^1 = \mathcal{R}^2 = \mathcal{R}$. An approach to discover collaborative teammates that utilise different behaviours in maximising the shared objective is finally presented in Section 4.

4 Method

The goal of our approach is to produce a set of teammate policies for training, $K = \{\pi^1, \pi^2, ..., \pi^{|K|}\}$, which is optimised to be as diverse as possible. Unlike many prior approaches, we do not define policy diversity in terms of information theoretic measurements, typically defined over distribution of trajectories from policies in K. To illustrate the failure mode of trajectory diversity-based methods, consider a two-player grid world where agents are rewarded if together they occupy the upper left or lower right corner grid. Assuming $|K| = 2$ and agents' actions that only allow movement across the four cardinal directions, optimising trajectory diversity may produce teammates that use different routes to arrive at the same corner. The learner will then never learn to visit the other corner and will perform poorly against a new teammate who moves towards the other corner. We now provide a diversity metric that provides better teammate populations for training robust learners, which in this example is one where each teammate goes towards a different corner.

Cross-Play Matrix: We prevent the emergence of teammate populations requiring the same best response policies for collaboration by optimising a best response (BR) diversity metric. Given a state s, BR diversity metrics at s are defined over a $|K| \times |K|$ cross-play matrix, $C^K(s)$, where the element in row i and column j, $C^K_{i,j}(s)$, is defined as:

$$\mathbb{E}_{a_t^1 \sim \pi^{-i}, a_t^2 \sim \pi^i} \left[\sum_{t=0}^{\infty} \mathcal{R}(s_t, < a_t^1, a_t^2 >) \middle| s_0 = s \right], \quad (1)$$

while π^{-i} denotes the BR policy to collaborate with $\pi^i \in K$.

Optimised Objective: There are two prerequisites for optimising K in terms of BR diversity. First, π^i and π^{-i} must be jointly trained for all $i \in K$ to ensure that π^{-i} is a BR policy to π^i. Second, a BR diversity metric has to be defined to optimise K. Assuming that the associated policies in K and their BR policies are parameterised by θ, we maximise the following objective to train policies in K:

$$O(\theta) = \text{Tr}(C^{K_\theta}(s)) + \text{Div}(C^{K_\theta}(s)), \quad (2)$$

with K_θ being the set of policies induced by θ.

Maximising the trace of the cross-play matrix in the right hand side of Equation 2 ensures that π^{-i} is best response to π^i for all $i \in K_\theta$. On the other hand, the second term encourages an increase in the BR diversity metric which we define as:

$$\text{Div}(C^{K_\theta}(s)) = \text{Det}(\kappa(C^{K_\theta}(s))), \quad (3)$$

with $\kappa(C^{K_\theta}(s))$ being a $|K| \times |K|$ matrix containing the return similarity between BR policies for $\pi^i \in K$. Assuming $C^{K_\theta}_{i,\cdot}(s)$ denotes the i^{th} row of $C^{K_\theta}(s)$, the elements of κ are defined as:

$$\kappa_{i,j}(C^{K_\theta}(s)) = \exp\left(- \frac{|C^{K_\theta}_{i,\cdot}(s) - C^{K_\theta}_{j,\cdot}(s)|^2}{\sigma^2} \right) \quad (4)$$

This determinant-based metric penalises K when two different policies $\pi^i, \pi^j \in K$ have BR policies that yield similar performances when interacting with other policies in K.

Optimisation Technique: Since (2) is differentiable with respect to the elements of C^{K_θ}, updating θ based on Equation (2) can be done via the chain rule. In our case, we use a technique similar to DDPG [Lillicrap *et al.*, 2015]. We specifically design a critic that estimates Equation (1) and backpropagate gradients from Equation (2) through the critic.

5 Preliminary Experiments

We performed preliminary experiments in a 5×5 grid world where two agents are initially positioned at $(2, 2)$. The agents' actions allow them to move alongside the four cardinal directions. These agents are subsequently given a reward of 1 if they manage to arrive at $(0, 0)$ or $(4, 4)$ together. Assuming $|K| = 2$, our approach finds K where π^1 consistently goes to $(0, 0)$ while π^2 always moves towards $(4, 4)$. By contrast, methods that optimise other trajectory diversity metrics, such as Jensen-Shannon Divergence, result can result in both π^1 and π^2 moving towards only one of $(0, 0)$ or $(4, 4)$. Optimising trajectory diversity may therefore cause the learner to not reach destination grids not visited by generated teammates, which leads learners to fail when collaborating with new teammates who move towards the unreached destination grid.

6 Conclusion and Future Work

In this work, we proposed a promising approach to improve the robustness of type-based AHT learners. Our approach generates diverse sets of teammates which require highly different best responses during collaboration. We observe that training against the different teammates forces the learner to

learn a more robust policy that effectively deals with the different teammate behaviours.

Following promising initial results, additional evaluation against prior work in more complex environments is required. Furthermore, evaluation of the learners resulting from training against the generated teammates are also required. Finally, further exploration on other BR diversity metrics may also provide interesting insights for future work.

References

Stefano V. Albrecht, Jacob W. Crandall, and Subramanian Ramamoorthy. Belief and truth in hypothesised behaviours. *Artificial Intelligence*, 235:63–94, 2016.

Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132–171, 2017.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, ZHIPENG HU, and Yaodong Yang. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 941–952. Curran Associates, Inc., 2021.

Keane Lucas and Ross E Allen. Any-play: An intrinsic augmentation for zero-shot coordination. *arXiv preprint arXiv:2201.12436*, 2022.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork: Definitions, methods, and open problems. *arXiv preprint arXiv:2202.10450*, 2022.

Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. Agent modelling under partial observability for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof M Choromanski, and Stephen J Roberts. Effective diversity in population based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:18050–18062, 2020.

Arrasy Rahman, Niklas Höpner, Filippos Christianos, and Stefano V. Albrecht. Towards open ad hoc teamwork using graph-based policy learning. In *International Conference on Machine Learning*, volume 139. PMLR, 2021.

Dong Xing, Qianhui Liu, Qian Zheng, and Gang Pan. Learning with generated teammates to achieve type-free ad-hoc teamwork. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 472–478. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.