Cooperative Multi-Agent Reinforcement Learning in Convention Reliant Environments

Doctoral Consortium

Jarrod Shipton
University of the Witwatersrand
Johannesburg, South Africa
Jarrod.Shipton@wits.ac.za

ABSTRACT

Multi-Agent Reinforcement Learning (MARL) has seen a move towards creating algorithms which can be trained to work cooperatively with partners. Typically MARL is done in self-play (SP). Recent works show agents trained with SP often achieve near optimal results when paired with one another, however, they form arbitrary play conventions which can perform poorly when mismatched. This led to research into algorithms which have been developed to form strategies which avoid the need for convention matching and allow for zero-shot coordination (ZSC) with any novel partner. ZSC solves the problem of convention matching, and is useful in short interactions, however in pro-longed or repeated interaction this comes at the cost of optimality. Avoiding conventions leaves the challenge of being unable to exploit *known, existing* conventions and achieve higher levels of optimality. In this work we use population training with a belief of the partner type to exploit conventions which could exist, leading to high rewards over pro-longed interactions. We demonstrate that our method is able to better adapt in convention reliant environments over repeated interactions than current state-of-the-art competing ZSC methods.

KEYWORDS

Multi-agent Reinforcement Learning, Ad hoc Coordination, Zero-shot coordination, Conventions, K-level reasoning

ACM Reference Format:
Jarrod Shipton. 2024. Cooperative Multi-Agent Reinforcement Learning in Convention Reliant Environments: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

There has been a substantial increase in interest in the field of Reinforcement Learning (RL), particularly that of using it to solve problems involving cooperation between many different agents, examples include self driving cars, robot assistants and robots in warehouses. Multi-Agent Reinforcement Learning (MARL) has been used with varying levels of success in these cooperative environments enabling two or more agents to be trained to work collaboratively toward a common goal. It has been established that training agents in self-play (SP) can achieve emergent behaviours in which agents adopt different conventions to solve a problem, however a mismatch in convention could lead to sub-optimal or even disastrous results. For instance, in driving, adherence to a unified convention, such as driving on the left or the right, is crucial to prevent collisions. This work introduces a strategy to address convention mismatches by creating a population of agents with diverse conventions and learns to identify which convention should be adopted for a given group of agents.

2 BACKGROUND

MARL research is typically done using self-play (SP), where agents learn through interaction with clones of themselves. Agents trained with SP tend to perform well in MARL environments by learning arbitrary conventions. An example of the success of SP can be seen with the research done on Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning (SAD) [5] where the agents are trained to play the game Hanabi, a particularly complex, partially observable environment which has been set as benchmark environment [1]. The work done in SAD led to the emergent behaviours in which agents use arbitrary conventions to solve the environment. These conventions are an important phenomenon which leads to solutions of problems which *wouldn't otherwise exist*, however, these only work when paired with agents with matching conventions. It was later shown in Other Play (OP) [7], utilizing cross-play (XP) to pair agents from different SP sessions, that conventions are often mismatched. The authors of OP proposed exploiting symmetries in the environment to reduce the number of conventions that could be formed, making it less likely for a mismatch. Noting the weakness of needing to know symmetries of the environment for OP to work, further research was done in Synchronous K-Level Reasoning with a Best Response (SyKLRBR) [4], and Off Belief Learning (OBL) [6] in which the authors employ alternative strategies to achieve zero-shot coordination (ZSC). This is done by using a training strategy similar to Cognitive Hierarchies (CH) [2] and K-level Reasoning (KLR) [3] which trains k-level agents to have a general best response (BR) to previous levels of cognitive reasoning. The initial policy acts randomly to ensure that actions infer no extra meaning outside of what they reveal about the environment so that conventions can't be formed. This means they should, *in principle*, achieve zero-shot coordination (ZSC). Another approach to solving the problem is to develop a diverse set of training partners and learning a BR to these agents, this is explored in Entropy-regularized Deep Recurrent Q-Network (EDRQN) [12], Trajectory Diversity for Zero-Shot Coordination (TrajeDI) [10], and Any-Play [9], all of which are capable of achieving ZSC.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

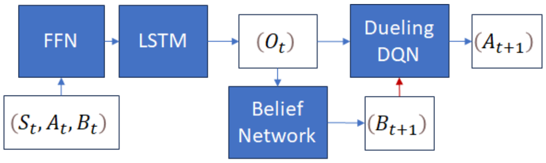

Figure 1: The network structure of our method. Note that the red arrow from B_{t+1} to the dueling DQN represents a break in the back propagation since the belief network is trained as an auxiliary task.

3 METHOD AND PRELIMINARY RESULTS

In XP settings, ZSC techniques tend to outperform SP methods *on average*, despite SP's potential for higher rewards within matching conventions. The potential SP has for higher rewards suggests that we should consider exploiting known conventions in repeated interactions. This would require being able to identify a set of diverse conventions over repeated interactions. The identification of the correct conventions would help achieve higher rewards in later interactions than a ZSC general best response despite the potential to incur early performance losses due to mismatches.

Hence, we propose a method which allows for the creation and identification of a diverse set of training partners with varying conventions. We use a training strategy similar to SyKLRBR. We use synchronous training of M different synchronous KLR training runs, depicted in Figure 2. We direct readers to SyKLRBR [4] for details on the implementation, however we will state the differences in our work. We use M initial policies π_0^m with different static distributions of actions, including one uniform random distribution as is used in SyKLRBR. The initial policies do not learn over time, consistent with the random initial policy in SyKLRBR. We ensure these initial policies action distributions have a Jensen–Shannon divergence which is above some value θ to ensure diverse behaviours. This allows for up to M different conventions to be learned. We use the same training architecture used in SyKLRBR, that is a Recurrent Replay Distributed DQN (R2D2) [8]. This is an architecture which uses an LSTM and the duelling network architecture, as seen in [11]. We make modifications to the network structure for our method, depicted in Figure 1, this is to account for the belief B_t of which of the $M \times K$ partners we are paired with at time t. We take the observed state S_t, the previous action taken A_t and the partner belief B_t encoded by the LSTM to give the output O_t. This output is then used to get an updated belief of the partner we are playing with B_{t+1}. This belief is trained as an auxiliary task with actual knowledge of which of the $M \times K$ partners we are paired with. The combination of O_t and B_{t+1} is then given to a dueling DQN to select an action A_{t+1}. Once the $M \times K$ partners are trained we then train a final policy π_{K+1} with this set of partners and itself.

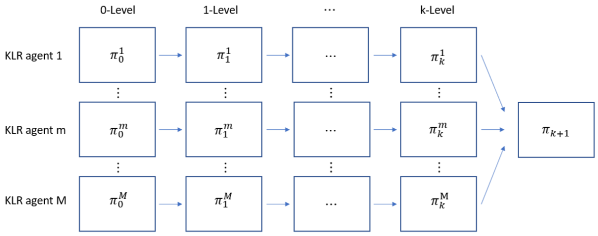

Figure 2: The structure of the M K-level policies with (K+1)-level final policy.

Method	Self-Play	Cross-play	Rules Bots
Our Method	40 ± 2.67	37 ± 4.12	37 ± 4.85
SyKLRBR	25 ± 0	25 ± 0	25 ± 0
EDRQN	25 ± 0	25 ± 0	25 ± 0
TrajeDI	25 ± 0	25 ± 0	25 ± 0

Table 1: Results of repeated signalling game.

To evaluate we take a modified and repeated version of the toy environment first presented in OBL [6]. This is a team game in which two partners need to signal to their teammate which animal is behind a wall by pressing one of a number of buttons and guess the animal without any prior communication about which animal is mapped to with which button, allowing for conventions to be formed. Teams are rewarded with a score of 10 for each correct answer, punished with -10 for each incorrect answer and receive half the score if they cheat and reveal the animal. In a single iteration of this game, a mismatch in convention is disastrous and we suggest ZSC techniques for one time interactions, however in reality we more often encounter repeated interactions. Hence this modified version is repeated 5 times for each pairing. A perfect score of 50 is possible if the conventions are perfectly matched, and 25 when the ZSC strategy of cheating is used. We scaled the number of animals and buttons up to 3, allowing for 6 permutations of mappings between animal and button. The length of the interaction needed to determine the partner convention would increase based on the number of possible conventions as well as how subtle differences in behaviour may be. For example, with n buttons and animals, you need a minimum of n repeated interactions to determine the partner convention, thus our method is suggested for scenarios with more repeated interactions than there are conventions. We made 6 pre-programmed bots which use these mappings for conventions or cheating and then paired trained agent with these bots randomly over 100 games each. We tested 6 training runs of each of the following methods: our method, SyKLRBR, EDRQN, and TrajeDi on this environment. We found that SyKLRBR, EDRQN, and TrajeDi find general best responses and all settle on the ZSC strategy of cheating, while our method tried to coordinate with the strategy it was playing with achieving higher scores on average. The results in Table 1 show that our method outperforms state-of-the-art ZSC methods in this *repeated* game scenario in three play settings: SP, XP and with rules bots.

4 CONCLUSION

We show that identifying and matching diverse conventions from a population of policies is beneficial and often outperforms general best response ZSC strategies over *repeated interactions.* We plan to further this by extending it to other domains such as Hanabi.

REFERENCES

[1] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[2] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 119, 3 (2004), 861–898.

[3] Miguel A Costa-Gomes and Vincent P Crawford. 2006. Cognition and behavior in two-person guessing games: An experimental study. *American economic review* 96, 5 (2006), 1737–1768.

[4] Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob Foerster. 2021. K-level Reasoning for Zero-Shot Coordination in Hanabi. *Advances in Neural Information Processing Systems* 34 (2021).

[5] Hengyuan Hu and Jakob N Foerster. 2019. Simplified action decoder for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1912.02288* (2019).

[6] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. 2021. Off-belief learning. In *International Conference on Machine Learning*. PMLR, 4369–4379.

[7] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "Other-Play" for Zero-Shot Coordination. In *International Conference on Machine Learning*. PMLR, 4399–4410.

[8] Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. 2018. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*.

[9] Keane Lucas and Ross E Allen. 2022. Any-Play: An Intrinsic Augmentation for Zero-Shot Coordination. *arXiv preprint arXiv:2201.12436* (2022).

[10] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. 2021. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*. PMLR, 7204–7213.

[11] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1995–2003.

[12] Dong Xing, Qianhui Liu, Qian Zheng, Gang Pan, and ZH Zhou. 2021. Learning with Generated Teammates to Achieve Type-Free Ad-Hoc Teamwork.. In *IJCAI*. 472–478.