Trajectory Diversity for Zero-Shot Coordination

Extended Abstract

Andrei Lupu
McGill University, Mila[1]
andrei.lupu@mail.mcgill.ca

Hengyuan Hu
Facebook AI Research
hengyuan@fb.com

Jakob Foerster
Facebook AI Research
jnf@fb.com

ABSTRACT

We study the problem of zero-shot coordination (ZSC), where agents must independently produce strategies for a collaborative game that are compatible with novel partners not seen during training. In particular, our first contribution is to consider the need for diversity in generating such agents. Because self-play agents control their own trajectory distribution during training, their policy only performs well on this exact distribution. As a result, they achieve low scores in ZSC, since playing with another agent is likely to put them in situations they have not encountered during training. To address this issue, we train a common best response (BR) to a population of agents, which we regulate to be as diverse as possible. For that purpose, we introduce *Trajectory Diversity* (TrajeDi) - a differentiable objective for generating diverse reinforcement learning (RL) policies. We present TrajeDi as a generalization of the Jensen-Shannon divergence (JSD) between policies and motivate it experimentally in a simple matrix game, where it allows to find the unique ZSC-optimal solution.

ACM Reference Format:
Andrei Lupu, Hengyuan Hu, and Jakob Foerster. 2021. Trajectory Diversity for Zero-Shot Coordination: 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

In this paper, we use policy diversity within population based training (PBT) to improve cross-play (XP) scores in the ZSC framework.

ZSC [8] is the problem of independently training two or more agents in a cooperative game such that their strategies are compatible and achieve high return when paired together at test time. Since it is impossible to agree on an arbitrary strategy with all humans ahead of time, solving ZSC is required for human-AI cooperation, such as in rescue robots or self-driving cars.

The challenge of the ZSC framework arises from the fact that many collaborative settings admit multiple joint strategies that are optimal yet incompatible. Then, if we naively train two independent agents in self-play (SP) [10], there is no guarantee that they will converge to compatible policies.

To address this problem, we can rely on the game-theoretic relationships between the optimal SP policies. Notably, if we have access to the entire solution space, we can train an agent to be the common BR to the largest possible subset of that space. The resulting agent would then be robust to the maximum number of potential partners, making it a prime candidate for ZSC.

This approach allows to train good policies for ZSC, but requires access to a diverse pool of optimal policies to serve as the training set for the BR. To that end, we introduce *TrajeDi*, a differentiable objective allowing to drive diversity within a pool of policies in the context of PBT. Specifically, TrajeDi works as a generalization of the JSD between the different policies and, unlike other methods, is especially designed for use in partially observable multi-agent settings.

2 RELATED WORKS

There is a growing corpus of works featuring a measure of diversity in RL. Many leverage it as a means of exploration [2, 5, 7] or to learn less redundant options, in the case of hierarchical RL [4, 6, 9]. Diversity has also been used in a reward-free context as a method of pre-training [3].

Regardless of the application, past works tend to formulate diversity either per state ($\pi(a|s)$) [2, 7], as a function of state distributions ($\mathbb{P}(s|\pi)$) [3, 4, 9] or of state-action distributions ($\mathbb{P}(s, a|\pi)$) [5]. Unfortunately, these formulations are inadequate for multi-agent RL [1, 11], where non-Markovian dynamics require actions to be conditioned on the entire observation history. Since multi-agent settings are ultimately our main concern, we formulate TrajeDi accordingly, making it the first trajectory-based ($\mathbb{P}(\tau|\pi)$) diversity objective, to the best of our knowledge.

Finally, our work addresses directly the challenges of the ZSC framework introduced by Hu et al., which we summarize in section 3. For this setting, Hu et al. propose a suitable training method called "*Other-Play*" that leverages domain knowledge of the game symmetries to find unambiguous solutions that perform well. Since symmetries are not always present or known, we instead rely on the structure of the policy space to find such solutions. That being said, the two methods are compatible, and they address different aspects of ZSC. As such, we expect to use them jointly when applying TrajeDi to more complex settings in the future.

3 SETTING AND BACKGROUND

We assume a collaborative Dec-POMDP $\mathcal{M} = (k, \mathcal{S}, \mathcal{A}, P, r, o, \gamma_{\mathcal{M}}, T)$, where a joint-policy π over the k agents selects joint-actions $a = (a^1, ..., a^k) \in \mathcal{A}$ based on observations $o(s) = (o^1(s), ..., o^k(s))$ of environment states $s \in \mathcal{S}$, with probability $\pi(a|s)$. The environment dynamics are governed by unknown transition probabilities $P(s_{t+1}|s_t, a_t)$, upon which all agents receive a reward $r(s_t, a_t)$. Finally, $\gamma_{\mathcal{M}} \in [0, 1]$ and T are the reward discount factor and horizon, respectively, and so we let $R = \sum_{t=0}^{T} \gamma_{\mathcal{M}}^t r(s_t, a_t)$ be the discounted return.

[1] Work done while at Facebook AI Research.

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.

The probability of a trajectory $\tau = (s_0, a_0, ..., s_{T-1}, a_{T-1}, s_T) \in \mathcal{T}$ under a policy π is given by $\mathbb{P}(\tau|\pi) = \varepsilon(\tau)\pi(\tau)$, where $\varepsilon(\tau) := \mathbb{P}(s_0)\prod_{t=0}^{T-1} P(s_{t+1}|a_t, s_t)$ summarizes the environment dynamics and $\pi(\tau) := \prod_{t=0}^{T-1} \pi(a_t|s_t)$ wraps the policy actions. The behaviour of a policy π is then characterized by the distribution it produces over trajectories and it is optimal if it maximizes $J(\pi) = \mathbb{E}_{\tau\sim\pi}\{R(\tau)\}$.

Zero-Shot Coordination. In the ZSC framework [8], two players must each learn a joint policy for a collaborative game, by a method of their choosing. However, they must perform their training individually, agreeing only on the training procedure before the task is specified. After training, they are paired together and evaluated without any possibility of adaptation to their partner. Consequently, the players cannot rely on arbitrary conventions, unlike in SP, and must instead employ a training procedure that will produce compatible policies when run separately.

4 METHOD

We first assume a population $(\pi_1, ..., \pi_n)$ of n policies, which we train in SP, but also regularize to be as diverse as possible. Additionally, we have a $(n+1)$-th policy denoted BR, which we train to be a common BR to every π_i. As a result, we must also optimize for the XP performance between each policy in the population and the BR. The global population loss is then

$$\mathcal{L}(\text{BR}, \pi_1, ..., \pi_n) = -\left[\sum_{i=1}^{n} \left(J(\text{BR}, \pi_i) + J(\pi_i, \pi_i)\right) + J(\text{BR}, \text{BR}) + \alpha \text{JSD}_\gamma(\pi_1, ..., \pi_n)\right], \quad (1)$$

where $J(\pi, \mu)$ is the expected return when pairing policy π with policy μ, α is a tunable weight and JSD_γ is the TrajeDi objective.

In particular, for a policy π_i unrolled on trajectory τ, let the local action kernel be defined as $\delta_{i,t}(\tau) := \prod_{t'=0}^{T} \left[\pi_i(a^\tau_{t'}|s^\tau_{t'})\right]^{\gamma^{|t-t'|}}$, where γ is a discounting factor in $[0, 1]$. Furthermore, let $\hat{\delta}_t(\tau) := \sum_{i=1}^{n} \frac{1}{n}\delta_{i,t}(\tau)$ be the average local action kernel. Then,

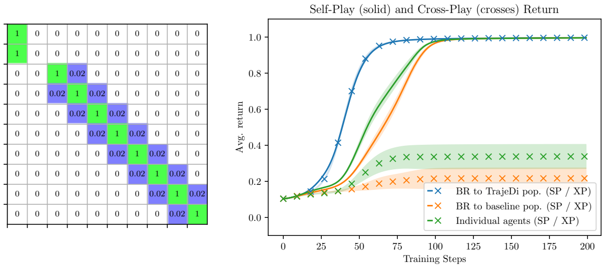
$$\text{JSD}_\gamma(\pi_1, ..., \pi_n) := -\frac{1}{n}\sum_{i=1}^{n}\sum_{\tau}\mathbb{P}(\tau|\pi_i)\sum_{t=0}^{T}\frac{1}{T}\log\frac{\hat{\delta}_t(\tau)}{\delta_{i,t}(\tau)} \quad (2)$$

is the TrajeDi objective, where tuning γ allows to measure diversity at a more local ($\gamma = 0$) or global ($\gamma = 1$) level. In fact, for $\gamma = 1$, eq. 2 is exactly equal to the JSD over the trajectory distribution induced by each policy, thus making TrajeDi a strict generalization.

The ability to tune TrajeDi for a local scope is crucial in settings admitting a high number of near-identical optimal policies. Otherwise, those policies could saturate the diversity objective by producing trajectories that differ only in trivial ways (e.g. a single action in a long sequence) that are irrelevant to the global behaviour.

5 EXPERIMENTS

We provide empirical evidence that training a common BR to a TrajeDi-enhanced population results in a robust agent for ZSC[2].



1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0	1	0.02	0	0	0	0	0	0
0	0	0.02	1	0.02	0	0	0	0	0
0	0	0	0.02	1	0.02	0	0	0	0
0	0	0	0	0.02	1	0.02	0	0	0
0	0	0	0	0	0.02	1	0.02	0	0
0	0	0	0	0	0	0.02	1	0.02	0
0	0	0	0	0	0	0	0.02	1	0.02
0	0	0	0	0	0	0	0	0.02	1

(a) Matrix Game

(b) Returns

Figure 1: a) Collaborative matrix game used to evaluate XP performance. b) SP (solid line) and XP (crosses) returns on the matrix game. Results for BRs to TrajeDi populations, BRs to baseline populations, and individual agents. TrajeDi reliably drives the BR to converge to the unique solution that generalizes well in ZSC.

Specifically, we consider the single step collaborative matrix game visualized in Figure 1a. In this game, player 1 must select a row while player 2 chooses a column independently. Once done, the actions are revealed, and both agents get the reward associated with the intersection of their choices.

We train 50 TrajeDi-regularized populations of two agents each complete with a common BR, as given by eq. 1. We evaluate the BRs and compare against 50 common BRs to unregularized populations, as well as to 50 individual agents trained independently.

We plot the performance in both SP and XP in Figure 1b. In SP, all schemes successfully achieve optimal return, but the BRs to TrajeDi populations do so faster. This is likely due to the TrajeDi objective providing a richer reward signal, but also due to the advantage diversity provides with exploration.

In XP, the TrajeDi curve is identical, demonstrating that BRs trained on diverse populations have indeed learned the most general solution (first or second row as the row player and first column as the column player). Meanwhile, individual agents and BRs to baseline populations perform only slightly better than chance, as off-diagonal payoffs tend to make the policies diverge from the TrajeDi solution. Moreover, there is no significant difference between the XP performance of the two baselines, supporting the hypothesis that PBT with a common BR can easily collapse to SP without a diversity bonus.

6 CONCLUSION AND FUTURE WORK

In this work, we first highlighted the role of diversity in ZSC and leveraged it within a PBT approach to produce agents that are robust to a larger number of potential XP partners. To this end, we introduced TrajeDi, a differentiable diversity objective that generalizes the JSD over trajectory distributions. Finally, we showed that TrajeDi allows to reliably converge to the ZSC-optimal solution in a collaborative matrix game, thus greatly improving XP scores over our baselines.

In the near future, we wish to further detail the theoretical properties of TrajeDi and show its compatibility with policy gradient methods, including in the context of batch RL. Additionally, we aim to evaluate the scalability of our method in a large scale ZSC setting, such as the collaborative card game Hanabi, and compare it to a broader set of baselines.

[2] The code is available online and can be run in-browser: https://bit.ly/33NBw5o

REFERENCES

[1] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2010. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1* (2010), 183–221.

[2] Andrew Cohen, Xingye Qiao, Lei Yu, Elliot Way, and Xiangrong Tong. 2019. Diverse exploration via conjugate policies for policy gradient methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3404–3411.

[3] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070* (2018).

[4] Carlos Florensa, Yan Duan, and Pieter Abbeel. 2017. Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012* (2017).

[5] Tanmay Gangwani, Qiang Liu, and Jian Peng. 2018. Learning self-imitating diverse policies. *arXiv preprint arXiv:1805.10309* (2018).

[6] Tuomas Haarnoja, Kristian Hartikainen, Pieter Abbeel, and Sergey Levine. 2018. Latent space policies for hierarchical reinforcement learning. *arXiv preprint arXiv:1804.02808* (2018).

[7] Zhang-Wei Hong, Tzu-Yun Shann, Shih-Yang Su, Yi-Hsiang Chang, Tsu-Jui Fu, and Chun-Yi Lee. 2018. Diversity-driven exploration strategy for deep reinforcement learning. In *Advances in Neural Information Processing Systems*. 10489–10500.

[8] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. " Other-Play" for Zero-Shot Coordination. *arXiv preprint arXiv:2003.02979* (2020).

[9] Yuhang Song, Jianyi Wang, Thomas Lukasiewicz, Zhenghua Xu, and Mai Xu. 2019. Diversity-driven extensible hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4992–4999.

[10] Gerald Tesauro. 1994. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation* 6, 2 (1994), 215–219.

[11] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2019. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635* (2019).