FACEBOOK AI

Mila McGill UNIVERSITY

Trajectory Diversity for Zero-Shot Coordination

Andrei Lupu*, Brandon Cui,
Hengyuan Hu, Jakob Foerster

- We highlight the role of diversity in ZSC
- We leverage diversity in a PBT approach
- We introduce TrajeDi, a general and differentiable objective for training diverse policies

Problem Setting: Zero-Shot Coordination

1. Players agree on training method
2. Task is revealed
3. Players each **train** a policy **independently**
4. Test time: **evaluate cross-play score** between players

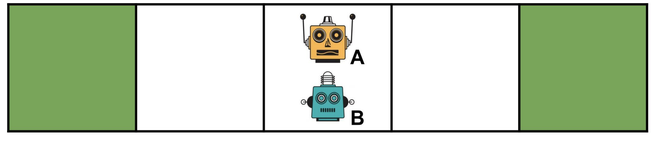


Figure 1: Corridor coordination task admitting three optimal self-play policies, only one of which is suitable for ZSC.

Population Based Approach

- Train *n* policies + common best response (BR)
- Want objective of the type:

$$\mathcal{L}(\mathrm{BR}, \pi_1, ..., \pi_n) = -\Bigg[\sum_{i=1}^{n}(J(\mathrm{BR}, \pi_i) + J(\pi_i, \pi_i)) + J(\mathrm{BR}, \mathrm{BR}) + \alpha Diversity(\pi_1, ..., \pi_n)\Bigg]$$

Jensen Shannon Divergence

- JSD defines distance between policies

$$\mathrm{JSD}(\pi_1, ..., \pi_n) = H(\hat{\pi}) - \sum_{i=1}^{n}\frac{1}{n}H(\pi_i)$$

Trajectory Diversity

- **Action Discounting** allows to tune the sensitivity of the diversity objective

$$\delta_{i,t}(\tau) := \prod_{t'=0}^{T}[\pi_i(a^\tau_{t'}|s^\tau_{t'})]^{\gamma^{|t-t'|}}$$

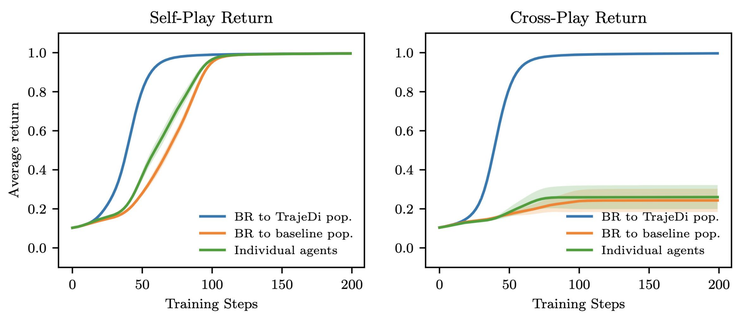
- JSD + Action Discounting = ***TrajeDi Objective***

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

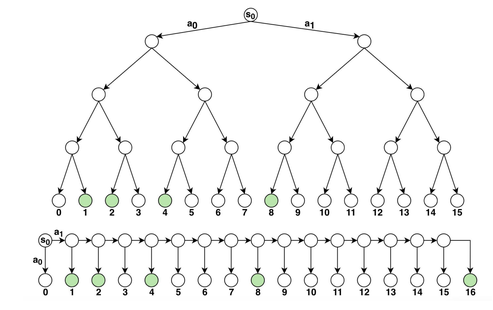
Experiments

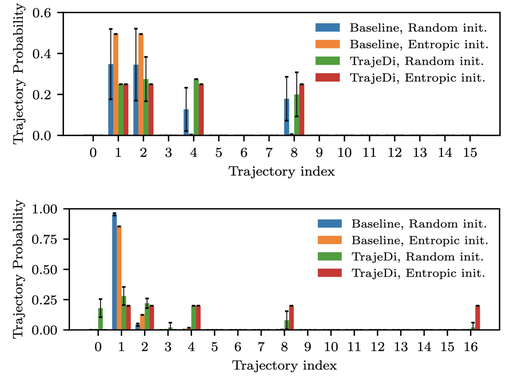
1. Zero-Shot Coordination in a Matrix Game

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



2. Diversity of Solutions in Tree-like MDPs





2. TrajeDi in Hanabi

Method	Self-Play	Cross-Play
Individual Agents (Other-Play)	**24.24±0.02**	23.65±0.06
BR to pool of OP agents	24.17±0.04	23.66±0.07
BR to pool of OP agents + TrajeDi	24.22±0.01	**24.09±0.02**

References

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot,M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S.,Hughes, E., et al. The Hanabi Challenge: A New Frontier for AI Research. *Artificial Intelligence*, 280:103216, 2020.

Hu, H., Lerer, A., Peysakhovich, A., and Foerster, J. "Other-Play" for Zero-Shot Coordination. *arXiv preprint arXiv:2003.02979*, 2020.

Endres, D. M. and Schindelin, J. E. A New Metric for Probability Distributions. IEEE Transactions on Information theory, 49(7):1858–1860, 2003.

Code: https://bit.ly/33NBw5o

Facebook AI Research, Mila, McGill University

Correspondence to: andrei.lupu@mail.mcgill.ca