Generating Diverse Cooperative Agents by Learning Incompatible Policies

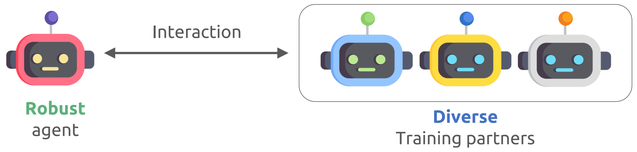
Rujikorn Charakorn[1], Poramate Manoonpong[1,2], Nat Dilokthanakul[3]

[1] VISTEC - [2] SDU - [3] KMITL

A4BM ICML | 2022

Introduction

Training a robust cooperative agent that can work with unseen agents is useful. However, training such a robust agent requires diverse training partner agents. In spite of its importance, obtaining diverse partners is still an open problem. Many prior works propose to generate diverse agents by changing the **state-action distribution** [1] or **joint trajectory distribution** [2,3] of the agents. However, changes in such distribution **might not** lead to high-level behavioral difference [3].



Learning incompatible policies (LIPO)

In this work, we propose an alternative way to diversify behaviors using information from the **task's objective**. Specifically, LIPO trains incompatible policies to generate diverse agents. We show theoretically that ***incompatible policies are not similar*** to each other.

LIPO objective

Considering $\pi_A = (\pi_A^1, \pi_A^2), \pi_B = (\pi_B^1, \pi_B^2)$, we want to find π_A that is not similar π_B. Through our theoretical results, we know that incompatible policies are not similar. Therefore, we propose that such a policy π_A can be learned by maximizing the self-play return while minimizing the cross-play return

$$\max_{\pi_A} \underbrace{J_{SP}(\pi_A)}_{\text{Self-play return}} - \underbrace{\lambda_{XP} J_{XP}(\pi_A, \pi_B)}_{\text{Cross-play return}}$$

Essentially, this objective gives a **competent** π_A that is **incompatible** with π_B.

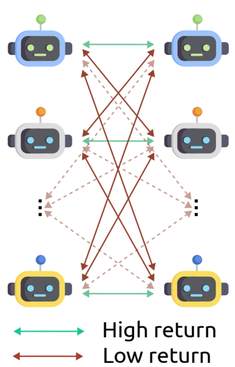
LIPO for population-based training

We extend the objective from considering only two joint policies to a population of N joint policies, $\mathcal{P} = \{\pi_i | 1 \leq i \leq N\}$. For a joint policy π_A in a population $\mathcal{P}$, its objective now has an aggregated cross-play return term

$$\max_{\pi_A} J_{SP}(\pi_A) - \underbrace{\lambda_{XP} \tilde{J}_{XP}(\pi_A, \mathcal{P})}_{\text{Aggregated Cross-play return}}$$

$$\text{where } \tilde{J}_{XP}(\pi_A, \mathcal{P}) = \max_{\pi_B \in \mathcal{P}_{-A}} J_{XP}(\pi_A, \pi_B)$$

$$\mathcal{P}_{-A} = \mathcal{P} \backslash \{\pi_A\}$$



Utilizing a mutual information (MI) objective

It is possible that there exist different behaviors that are fully **compatible**. We propose to capture such behavioral variations by using a **mutual information** objective. Specifically, we condition the policy on a latent variable z such that π_A has the following form

$$\pi_A(a|\tau) = \mathbb{E}_{z^1 \sim p(z^1), z^2 \sim p(z^2)} \pi_A^1(a^1|\tau^1, z^1)\, \pi_A^2(a^2|\tau^2, z^2)$$

Then we can maximize the mutual information between observation-action pair and the latent variable by maximizing the lower bound of the MI objective [2,4,5]

$$L_{MI}(\pi_A, \phi_A) = -\frac{1}{2} \sum_{i=1}^{2} \mathbb{E}_{z^i, (o^i, a^i)} \log q_{\phi_A}(z^i | o^i, \pi_A^i(\cdot\,|o^i, z^i))$$

Therefore, the overall training objective of a policy $\mathcal{P}_{-A}$ within a population $\mathcal{P}$ is

$$\max_{\pi_A, \phi_A} J_{SP}(\pi_A) - \lambda_{XP} \tilde{J}_{XP}(\pi_A, \mathcal{P}) - \lambda_{MI} \tilde{J}_{MI}(\pi_A, \phi_A)$$

Experimental results

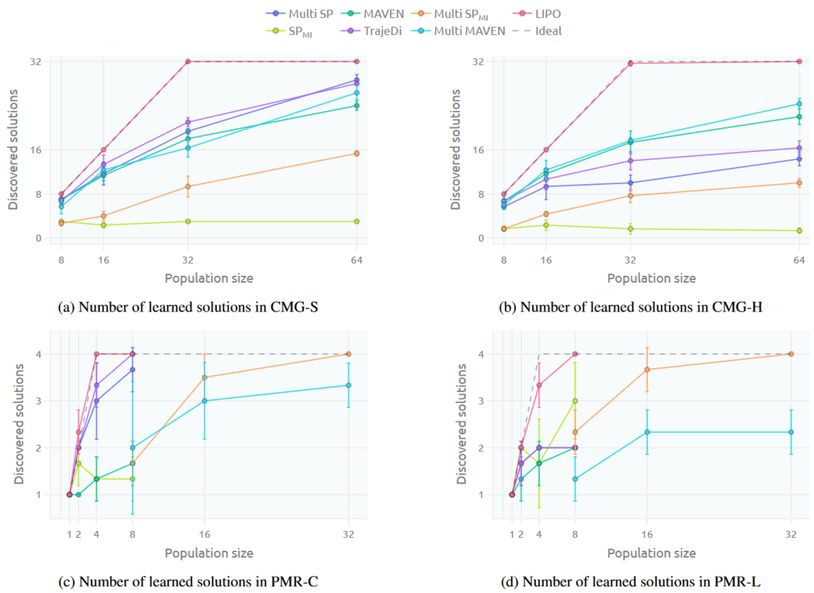


Figure 1: Number of discovered solutions in different population sizes.

We experiment with two cooperative environments: One-Step Cooperative Matrix Game (CMG) and Point Mass Rendezvous (PMR) with two different settings in each environment. In all scenarios, LIPO can find more solutions than the baselines. LIPO is also better at discovering sub-optimal solutions that exist in CMG-S and PMR-L.

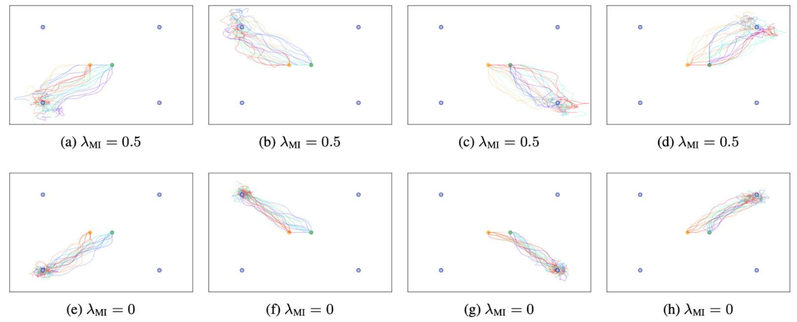


Figure 2: Trajectories of agents trained by LIPO with (a,b,c,d) and without (e,f,g,h) the MI objective in PMR-C. Each row shows four joint policies produced with a single run of LIPO training. Different colors of the trajectories correspond to different values of the latent variable.

We investigate further into how the MI objective affects the behaviors of produced policies. We can see the effect of the MI objective in the distributions of the trajectories, which exhibit larger variations given a small MI regularization. With or without the MI regularization, LIPO discovers all the landmarks with N=4.

Discussion

An agent trained with LIPO are incentivized to act **adversarially** toward agents that behave differently from itself. This behavior might not be desirable for certain downstream tasks. For example, agents produced by LIPO might not be suitable for interacting with humans as they would refuse to conform with the user. However, we believe that training an adaptive agent with these agents, which is the main motivation of this work, would have an opposite effect, in that the adaptive agent would try to comply with what its current partner is doing. We are investigating this potential benefit, and we will include the result in our future work.

References

[1] Lucas, Keane, and Ross E. Allen. "Any-Play: An Intrinsic Augmentation for Zero-Shot Coordination." *AAMAS* 2022.

[2] Mahajan, Anuj, et al. "Maven: Multi-agent variational exploration." *NeuRIPS* 2019.

[3] Lupu, Andrei, et al. "Trajectory diversity for zero-shot coordination." *ICML* 2021.

[4] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *ICLR* (2013).

[5] Osa, Takayuki, Voot Tangkaratt, and Masashi Sugiyama. "Discovering diverse solutions in deep reinforcement learning by maximizing state–action-based mutual information." *Neural Networks* (2022).