DGPO: Discovering Multiple Strategies with Diversity-Guided Policy Optimization

Extended Abstract

Wenze Chen
Tsinghua University
Beijing, China
cwz19@mails.tsinghua.edu.cn

Shiyu Huang
4Paradigm Inc.
Beijing, China
huangshiyu@4paradigm.com

Yuan Chiang
Tsinghua University
Beijing, China
yjennice2001@gmail.com

Ting Chen
Tsinghua University
Beijing, China
tingchen@tsinghua.edu.cn

Jun Zhu
Tsinghua University
Beijing, China
dcszj@tsinghua.edu.cn

ABSTRACT

Recent algorithms designed for reinforcement learning tasks focus on finding a single optimal solution. However, in many practical applications, it is important to develop reasonable agents with diverse strategies. In this paper, we propose Diversity-Guided Policy Optimization, an on-policy framework for discovering multiple strategies for the same task. Our algorithm uses diversity objectives to guide a latent code conditioned policy to learn a set of diverse strategies in a single training procedure. Experimental results show that our method efficiently finds diverse strategies in a wide variety of reinforcement learning tasks. We further show that DGPO has similar performance and achieves a higher diversity score or better sample efficiency compared to other baselines.

KEYWORDS

Diversity; Deep reinforcement learning; Probabilistic graphical models

ACM Reference Format:
Wenze Chen, Shiyu Huang, Yuan Chiang, Ting Chen, and Jun Zhu. 2023. DGPO: Discovering Multiple Strategies with Diversity-Guided Policy Optimization: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement learning (RL) has achieved human-level performance in various tasks, e.g., video games [2, 4, 5, 11] and robotics [8, 13]. However, RL algorithms are notorious for highly "overfitting" the given task, i.e., while there is a diverse set of quality solutions for the given problem, RL algorithms can only obtain a single optimal one. Finding a set of qualified diverse strategies is crucial for a robust agent.

In this paper, we proposed Diversity-Guided Policy Optimization (DGPO), an on-policy framework for discovering multiple strategies for the same task. Our contributions are as follows: (1) We formalize two constrained optimization problems to efficiently discover a set of optimal strategies. (2) We carefully designed a novel on-policy algorithm, denoted as DGPO, that can find a diverse set of quality strategies simultaneously. (3) We empirically show that our method achieves competitive performance and has better diversity or sample efficiency than other baselines on various benchmarks.

2 PRELIMINARY

Latent conditioned policy: We consider policy π_θ that is conditioned on latent variable z to model the diverse strategies, where θ is the parameter of the policy π. We denote the latent conditioned policy as $\pi(a|s,z)$ and the latent conditioned critic network as $V^\pi(s,z)$. A latent variable $z \sim p(z)$ will be randomly sampled at the beginning of every episode and policy π will be used to sample a trajectory τ_z, with z being fixed for the entire episode. The latent variable z is sampled from a categorical distribution with the number of categories n_z and $p(z)$ is the uniform distribution.
Discounted state occupancy: The discounted state occupancy measure of policy π is defined as $\rho^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_t^\pi(s)$, where $P_t^\pi(s)$ is the probability that policy π visits state s at time t. The goal of the RL agent is to train a policy π to maximize the accumulated reward $J(\theta) = \mathbb{E}_{z\sim p(z),s\sim\rho^\pi(s),a\sim\pi(\cdot|s,z)}\left[\sum_t \gamma^t r(s_t,a_t)\right]$.

3 METHODOLOGY

We implement DGPO by developing a new variation of PPO [10]. Specifically, DGPO can be separated into two parts: diversity-constrained optimization and extrinsic-reward-constrained optimization. In **diversity-constrained optimization**, the policy maximizes $J(\theta)$ only when its behaviors are sufficiently different from the existing policy's. Otherwise, it updates an intrinsic rewards-based objective to encourage policies to behave differently. We introduce a novel diversity metric to evaluate the diversity score of the given set of policies as below:

$$\begin{aligned}\mathrm{DIV}(\pi_\theta) &= \mathbb{E}_{z\sim p(z)}[\min_{z'\neq z} D_{KL}(\rho^{\pi_\theta}(s|z)||\rho^{\pi_\theta}(s|z'))] \\ &\geq \mathbb{E}_{z\sim p(z),s\sim\rho^\pi(s)}\left[\min_{z'\neq z}\log\frac{p(z|s)}{p(z|s)+p(z'|s)}\right].\end{aligned} \tag{1}$$

A discriminator $q_\phi(z|s_t)$ is trained in a supervised manner to approximate $p(z|s)$, where ϕ is the parameter of the discriminator

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

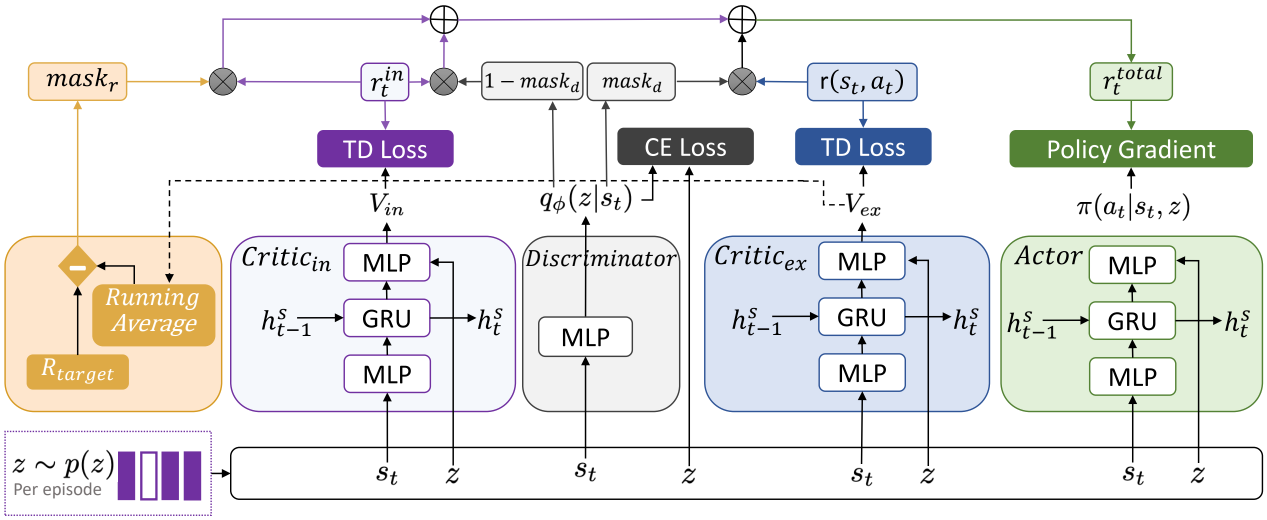


Figure 1: The overall framework of the DGPO algorithm. Top illustrates the way of calculating r_t^{total}, where $mask_r = \mathbb{I}[J(\pi_\theta) \geq R_{target}]$ and $mask_d = \mathbb{I}[\mathrm{DIV}(\pi_\theta) \geq \delta]$. Center shows the network structure and the data flow of the DGPO algorithm. Bottom shows the latent variable sampling process.

network. We introduce the intrinsic reward based on learned discriminator $q_\phi(z|s)$ as below:

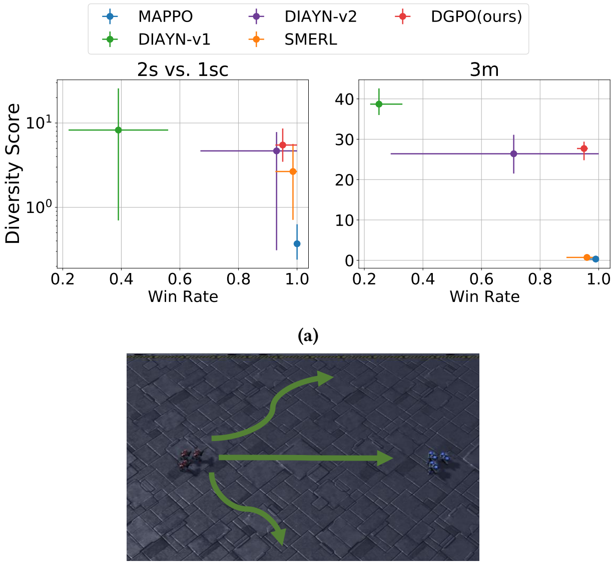
$$r_t^{in} = \min_{z' \neq z} \log \frac{q_\phi(z|s_t)}{q_\phi(z|s_t) + q_\phi(z'|s_t)} \tag{2}$$

The diversity objective $J_{Div}(\theta)$, which has a similar definition to $J(\theta)$ but considers intrinsic rewards, can be defined as $J_{Div}(\theta) = \mathbb{E}_{z \sim p(z), s \sim \rho^\pi(s), a \sim \pi(\cdot|s,z)}[\sum_t \gamma^t r_t^{in}]$. To implement diversity-constrained optimization, we optimize the extrinsic-rewards objective $J(\theta)$ when the data's diversity metric $Div(\pi_\theta) \geq \delta$, where δ is a hyper-parameter, and optimize $J_{Div}(\theta)$ when $Div(\pi_\theta) < \delta$. In **extrinsic-reward-constrained optimization**, the policy will optimize $J_{Div}(\theta)$ when $J(\theta)$ is greater than a threshold R_{target}, where R_{target} is a hyper-parameter. Fig. 1 summarizes the DGPO algorithm.

4 EXPERIMENTS

We evaluate our algorithm on multiple RL benchmarks, i.e., Multi-agent Particle-world Environment (MPE) [7], StarCraft II Micromanagement Challenge (SMAC) [9], and Atari [1]. We compare our algorithm with 4 baseline algorithms, including MAPPO [12], DIAYN [3], SMERL [6], and RSPO [14]. Due to the space limit, we only show part of the empirical results here.

We conduct experiments on two StarCraft II maps, i.e., *2s_vs._1sc* and *3m* from the SMAC benchmark. In each map, we set $n_z = 3$ and measure the average winning rates over 5 seeds for all the algorithms. Fig. 2(a) shows that while other algorithms are either unstable in terms of diversity or performance, DGPO can find a set of quality diverse strategies. In fact, DGPO has higher diversity scores and competitive performance in comparison with baseline algorithms. Fig. 2(b) visualizes three strategies obtained by DGPO on *3m* map. In this map, we control three agents (in red) to combat build-in agents (in blue). We visualize the moving trajectories of our agents in green arrows. Empirical results show that in addition to moving forward (the middle arrow) to attack the enemies directly, DGPO agents produce kiting strategies, i.e., our agents keep switching between attacking and moving upward or downward. Through this, they can attack enemies and avoid damage from them.



(a)

(b)

Figure 2: Experimental results in two SMAC scenarios.

5 CONCLUSIONS

In this paper, we proposed Diversity-Guided Policy Optimization (DGPO), an on-policy algorithm that can efficiently discover diverse quality strategies. DGPO formulates the training process as two constrained optimization problems and solves them as a probabilistic inference task. Empirical results indicate that DGPO has competitive performance and sample efficiency with state-of-the-art on-policy RL algorithms and achieves the highest diversity score in comparison with baseline algorithms in various domains.

REFERENCES

[1] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.

[2] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[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] Shiyu Huang, Wenze Chen, Longfei Zhang, Ziyang Li, Fengming Zhu, Deheng Ye, Ting Chen, and Jun Zhu. 2021. TiKick: Towards Playing Multi-agent Football Full Games from Single-agent Demonstrations. *arXiv preprint arXiv:2110.04507* (2021).

[5] Shiyu Huang, Hang Su, Jun Zhu, and Ting Chen. 2019. Combo-action: Training agent for fps game with auxiliary tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 954–961.

[6] Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. 2020. One solution is not all you need: Few-shot extrapolation via structured maxent rl. *Advances in Neural Information Processing Systems* 33 (2020), 8198–8210.

[7] Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[8] Antonin Raffin, Ashley Hill, René Traoré, Timothée Lesort, Natalia Díaz-Rodríguez, and David Filliat. 2018. S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning. *arXiv preprint arXiv:1809.09369* (2018).

[9] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).

[10] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[11] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[12] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2021. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. *arXiv preprint arXiv:2103.01955* (2021).

[13] Chao Yu, Xinyi Yang, Jiaxuan Gao, Huazhong Yang, Yu Wang, and Yi Wu. 2021. Learning Efficient Multi-Agent Cooperative Visual Exploration. *arXiv preprint arXiv:2110.05734* (2021).

[14] Zihan Zhou, Wei Fu, Bingliang Zhang, and Yi Wu. 2022. Continuously Discovering Novel Strategies via Reward-Switching Policy Optimization. *arXiv preprint arXiv:2204.02246* (2022).