Centralized Cooperative Exploration Policy for Continuous Control Tasks

Extended Abstract

Chao Li*
Institute of Automation, Chinese Academy of Sciences, China
lichao2021@ia.ac.cn

Chen Gong*
Institute of Automation, Chinese Academy of Sciences, China
gongchen2020@ia.ac.cn

Qiang He
University of Tubingen, Germany
qianghe97@gmail.com

Xinwen Hou†
Institute of Automation, Chinese Academy of Sciences, Beijing, China
xinwen.hou@ia.ac.cn

Yu Liu
Institute of Automation, Chinese Academy of Sciences, Beijing, China
yu.liu@ia.ac.cn

ABSTRACT

Despite recent works making great progress in continuous control tasks, exploration in these tasks has remained insufficiently investigated. This paper proposes CCEP (**C**entralized **C**ooperative **E**xploration **P**olicy), which utilizes estimation biases of value functions to contribute to the exploration capacity. CCEP keeps two value functions initialized with different parameters, and generates diverse policies with multiple exploration styles from a pair of value functions. In addition, a centralized policy framework ensures that CCEP achieves message delivery between multiple policies, furthermore contributing to exploring the environment cooperatively. Extensive experimental results demonstrate that CCEP achieves higher exploration capacity. Empirical analysis shows diverse exploration styles in the learned policies by CCEP, reaping benefits in more exploration regions. Besides, the exploration capabilities of CCEP have been demonstrated to outperform current state-of-the-art methods on multiple continuous control tasks.

KEYWORDS

Cooperative Exploration; Continuous Control Tasks

ACM Reference Format:

Chao Li*, Chen Gong, Qiang He, Xinwen Hou, and Yu Liu. 2023. Centralized Cooperative Exploration Policy for Continuous Control Tasks: 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

In DRL settings, an agent aims to learn an optimal policy to maximize its expected cumulative rewards through trial and error. It is essential that during the training phase, the agent should be encouraged to explore the environments and gather sufficient reward signals for well-training. Therefore, exploration has obsessed with a critical problem: submitting solutions too quickly without sufficient exploration, leading to getting stuck at local minima or even complete failure. Whereas existing exploration methods [1, 3, 6, 7, 11, 16] remain a problematic drawback – lacking diversity to explore. However, in massive situations, diverse styles of exploration are necessary. For instance, in chess games, players should perform different styles of policies to keep competitive.

Our insights originate from a non-trivial phenomenon during the critic update process: the different critic functions may have great differences even if they approximate the same target due to the function approximation error which accumulates to the estimation bias. Although many proposed methods are dedicated to reducing estimation bias [4, 5, 8, 9, 14, 15], this knotty problem is impossible to completely avoid. We raise the question of whether we can "transform an enemy into a friend" – utilizing estimation bias to

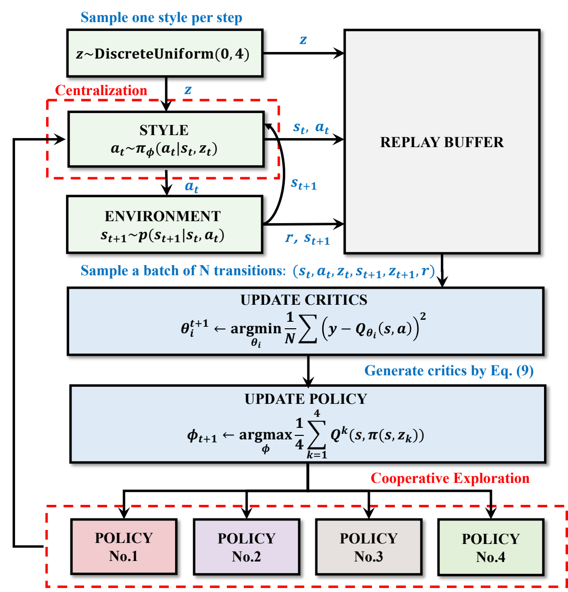

Figure 1: The workflow of CCEP Algorithm. The agent π interacts with the environment with diverse style cooperatively. A centralized policy with four different styles is learned from the multi-styled critics.

*These two authors contributed equally.
†The corresponding author.

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.

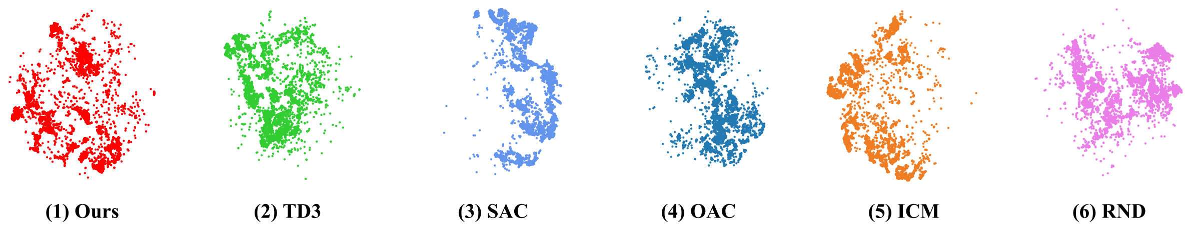

Figure 2: Measuring the exploration region. The points represent region explored by each method in 10 episodes. All the states get dimension reduction by the same t-SNE transformation for better visualization.

Table 1: Max Average Return over 10 trials of 1 million time steps. The mean values have been listed. The maximum value for each task is bolded.

Env	Ours	OAC	SAC	TD3	ICM	RND
HCheetah	**11945**	9921	11129	9758	10085	10629
Hopper	**3636**	3364	3357	3479	3504	3419
Walker2d	**4706**	4458	4349	4229	4255	4197
Ant	**5630**	4519	5084	5142	5166	4990
Pusher	-21	-25	**-20**	-25	-23	-23
Humanoid	5325	**5747**	5523	5356	5374	5490

enhance the RL algorithm performance. Inspired by this, we use the estimation bias gap between these two value functions, termed as *controversy*, to encourage more exploration. In particular, our intuition can be ascribed that controversy in the value estimation will lead to sub-optimal policies. These policies have a bias toward message acquisition known as the *style*.

This paper highlights that controversy can be utilized to encourage policies to yield multiple styles which encourages explorations. Our paper contributes three aspects. (1) We first describe that the estimation bias in double value functions can lead to various exploration *styles*. (2) This paper proposes the CCEP algorithm (Please refer to [10] for the full version of this paper), encouraging diverse exploration for environments by cooperation from multi-styled policies. (3) Finally, in CCEP, we design a novel framework, termed as the centralized value function framework, which is updated by experience collected from all the policies and accomplishes the message delivery mechanism between different policies. Extensive experiments are conducted on the MuJoCo platform to evaluate the effectiveness of our method. The results reveal that the proposed CCEP approach attains substantial improvements in both average return and sample efficiency on the baseline across selected environments. Besides, CCEP also allows agents to explore more states during the same training time steps as the baseline.

2 OUR METHOD

CCEP start by maintaining double randomly initialized value functions Q_{θ_1} and Q_{θ_2} with parameters θ_1 and θ_2 respectively and update the value function with TD3 [4]. But the two randomly initialized value functions potentially have different value estimations for a given state-action pair due to the accumulated function approximation error. This difference leads to the result that the two critics may give two different suggestions for the best action choice. These different criterias for a given state-action pair may lead to a different style of action choice. It is reasonable the estimation is radical if we choose the maximum value of the two to estimate and the estimation is conservative if we choose the minimum value of the two. Additionally, rather of constantly providing conservative or radical estimates for the current batch of state-action pairs, we would like to take into account random conservative or radical estimates. Thus, we involve four critics during the update of policy networks. There exists controversy among these critics, and the controversy can influence the performance of the policy learned.
With four critics, we train a centralized cooperative policy to encourage multi-styled explorations through diverse value estimations. The target is to train multiple policies, with each policy targeting an individual value function. We express the policy function as $\pi(s, z)$, with state s and latent variable z as input. The latent variable z, which is a one-hot label in our method, identifies different policies. For a given latent variable z, the policy targets z-th value functions. With different latent variable z, the policy shows diverse styles due to the multi-styled targets. We make an experiment showing that there exists different exploration preferences for these policies. In the exploration procedure, we randomly sample latent variable z and make decisions by policy $\pi(s, z)$. This approach enables diverse styles to be applied at each time step. The workflow of CCEP Algorithm is shown in Figure 1. Broadly speaking, our exploration policy has the following characteristics: Multi-styled, Centralized, and Cooperative.

3 EXPERIMENTS

To evaluate our method, we test our algorithm on the suit of MuJoCo [12] continuous control tasks. We show the max average return over 10 trials of 1 million time steps in Table 1. Further, We compare the exploration capabilities of CCEP with that of baselines [1, 2, 4, 7, 11] (Figure 2). For a fair comparison, these methods are trained in Ant-v3 with the same seed at half of the training process. In order to get reliable results, the states explored are gathered in 10 episodes with different seeds. We apply the same t-SNE [13] transformation to the states explored by all of the algorithms for better visualization. While there are great differences between the states explored by different algorithms, the result shows that our algorithm (red) explores a wider range of states involving that other algorithms has explored.

REFERENCES

[1] Yuri Burda, Harrison Edwards, Amos Storkey, et al. 2019. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[2] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. 2019. Better Exploration with Optimistic Actor Critic. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 1785–1796. https://proceedings.neurips.cc/paper/2019/hash/a34bacf839b923770b2c360eefa26748-Abstract.html

[3] Justin Fu, John D. Co-Reyes, and Sergey Levine. 2017. EX2: Exploration with Exemplar Models for Deep Reinforcement Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 2574–2584.

[4] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018 (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, PMLR, Stockholmsmässan, Stockholm, Sweden, 1582–1591. http://proceedings.mlr.press/v80/fujimoto18a.html

[5] Chen Gong, Qiang He, Yunpeng Bai, Xinwen Hou, et al. 2021. Wide-Sense Stationary Policy Optimization with Bellman Residual on Video Games. In *2021 IEEE International Conference on Multimedia and Expo, ICME*. IEEE, 1–6.

[6] Chen Gong, Zhou Yang, Yunpeng Bai, Jieke Shi, Arunesh Sinha, Bowen Xu, David Lo, Xinwen Hou, and Guoliang Fan. 2022. Curiosity-Driven and Victim-Aware Adversarial Policies. In *Proceedings of the 38th Annual Computer Security Applications Conference* (Austin, TX, USA) *(ACSAC '22)*. 186–200. https://doi.org/10.1145/3564625.3564636

[7] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018 (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, PMLR, Stockholmsmässan, Stockholm, Sweden, 1856–1865.

[8] Hado Hasselt. 2010. Double Q-learning. In *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Eds.), Vol. 23. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2010/file/091d584fced301b442654dd8c23b3fc9-Paper.pdf

[9] Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 2094–2100.

[10] Chao Li, Chen Gong, Qiang He, Xinwen Hou, and Yu Liu. 2023. Centralized Cooperative Exploration Policy for Continuous Control Tasks. https://doi.org/10.48550/ARXIV.2301.02375

[11] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 2778–2787.

[12] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 5026–5033. https://doi.org/10.1109/IROS.2012.6386109

[13] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[14] Wei Wei, Yujia Zhang, Jiye Liang, Lin Li, and Yyuze Li. 2022. Controlling Underestimation Bias in Reinforcement Learning via Quasi-median Operation. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 8 (Jun. 2022), 8621–8628. https://doi.org/10.1609/aaai.v36i8.20840

[15] Dongming Wu, Xingping Dong, Jianbing Shen, and Steven C. H. Hoi. 2020. Reducing Estimation Bias via Triplet-Average Deep Deterministic Policy Gradient. *IEEE Transactions on Neural Networks and Learning Systems* 31, 11 (2020), 4933–4945. https://doi.org/10.1109/TNNLS.2019.2959129

[16] Yuzhong Zhao, Yuanqiang Cai, Weijia Wu, and Weiqiang Wang. 2022. Explore Faster Localization Learning For Scene Text Detection. *arXiv preprint arXiv:2207.01342* (2022).