vMFER: von Mises-Fisher Experience Resampling Based on Uncertainty of Gradient Directions for Policy Improvement of Actor-Critic Algorithms

Extended Abstract

Yiwen Zhu
Zhejiang University
Hangzhou, China
evanzhu@zju.edu.cn

Jinyi Liu
Tianjin University
Tianjin, China
jyliu@tju.edu.cn

Wenya Wei
Zhejiang University
Hangzhou, China
wwy_vivian@qq.com

Qianyi Fu
Zhejiang University
Hangzhou, China
aeroqianyi@gmail.com

Yujing Hu*
NetEase Fuxi AI Lab
Hangzhou, China
huyujing@corp.netease.com

Zhou Fang*
Zhejiang University
Hangzhou, China
zfang@zju.edu.cn

Bo An
Nanyang Technological University
Singapore
boan@ntu.edu.sg

Jianye Hao
Tianjin University
Tianjin, China
jianye.hao@tju.edu.cn

Tangjie Lv
NetEase Fuxi AI Lab
Hangzhou, China
hzlvtangjie@corp.netease.com

Changjie Fan
NetEase Fuxi AI Lab
Hangzhou, China
fanchangjie@corp.netease.com

ABSTRACT

Reinforcement Learning (RL) is a widely employed technique in decision-making problems, encompassing two fundamental operations – policy evaluation and policy improvement. Actor-critic algorithms dominate the field of RL, but there is a challenge in improving their learning efficiency. To address this, ensemble critics are often employed to enhance policy evaluation efficiency. However, when using multiple critics, the actor in the policy improvement process can obtain different gradients. Previous studies have combined these gradients without considering their disagreements. Therefore, optimizing the policy improvement process is crucial to enhance the learning efficiency of actor-critic algorithms. This study focuses on investigating the impact of gradient disagreements caused by ensemble critics on policy improvement. We introduce the concept of uncertainty of gradient directions as a means to measure the disagreement among gradients utilized in the policy improvement process. Through measuring the disagreement among gradients, we find that transitions with lower uncertainty of gradient directions are more reliable in the policy improvement process. Building on this analysis, we propose a method called von Mises-Fisher Experience Resampling (vMFER), which optimizes the policy improvement process by resampling transitions and assigning higher confidence to transitions with lower uncertainty of gradient directions. Our experiments on Mujoco robotic control tasks and robotic arm tasks with sparse rewards demonstrate that vMFER significantly outperforms the benchmark and is particularly well-suited for ensemble structures in RL.

KEYWORDS

Resample; Uncertainty; Von Mises-Fisher Distribution

ACM Reference Format:
Yiwen Zhu, Jinyi Liu, Wenya Wei, Qianyi Fu, Yujing Hu[1], Zhou Fang[1], Bo An, Jianye Hao, Tangjie Lv, and Changjie Fan. 2024. vMFER: von Mises-Fisher Experience Resampling Based on Uncertainty of Gradient Directions for Policy Improvement of Actor-Critic Algorithms: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Recent advancements in reinforcement learning (RL) have shown remarkable progress, notably in complex task handling [1–3]. Yet, the challenge of enhancing learning efficiency in RL remains.

At its core, RL involves policy evaluation and improvement [4]. Methods like Double Q-learning [5], SAC [6, 7], TD3 [8], and REDQ [9] have optimized policy evaluation using ensemble critics. However, these critics can lead to disagreements during policy improvement. Typically, methods aggregate multiple gradients into one,

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

*The corresponding authors.

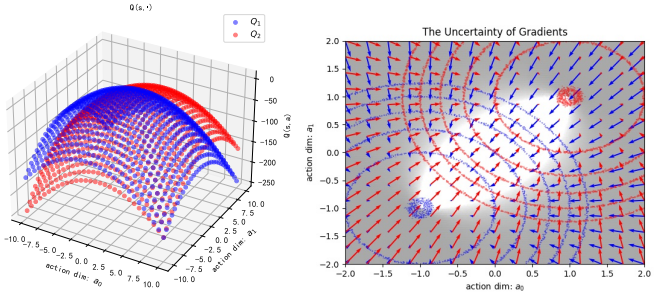

Figure 1: A simple example to demonstrate the the uncertainty of gradients in policy improvement. *Left*: The ensemble Q-values; *Right*: The uncertainty of gradients caused by the ensemble structure in policy improvement process.

overlooking these disagreements. Alternatives like delayed policy update in REDQ and TD3, while ensuring more reliable gradients, ignore discrepancies among transitions.

We propose an approach to optimize policy improvement by selectively avoiding delayed updates for transitions with reliable gradients under current ensemble critics. We introduce a metric to measure the gradients' reliability and identify appropriate transitions for policy improvement. As ensemble critics become more accurate, the concentration of policy gradient directions increases. We define the uncertainty of gradient directions to assess the reliability of transitions under the current ensemble structure. From a directional statistics view [10], we model these directions as a distribution, using the von Mises-Fisher distribution [11] for computational efficiency.

Our novel von Mises-Fisher Experience Resampling (vMFER) algorithm uses gradient direction uncertainty to resample transitions, enhancing policy improvement efficiency by favoring transitions with lower uncertainty.

Our contributions include:

(1) Introducing an uncertainty metric for assessing transition reliability in policy improvement, analyzing gradient direction discrepancies caused by ensemble critics.
(2) Proposing the vMFER algorithm, improving learning efficiency and optimality in actor-critic algorithms by resampling based on gradient uncertainty, applicable to most ensemble-structured actor-critic algorithms.
(3) Demonstrating vMFER's effectiveness in complex control tasks [12, 13], highlighting its broad applicability.

2 MOTIVATION & APPROACH

Figure 1 elucidates the challenges faced during the policy improvement phase. To further underscore the importance of accounting for gradient uncertainty in this process, we use a straightforward shooting task as an illustrative example, presented in Figure 2. All three methods employ identical ensemble critics, ensuring no variation in the policy evaluation phase. However, during policy improvement process, the '**Uniform**' approach randomly samples transitions, the '**Uncertainty**' approach selects transitions with lower uncertainty

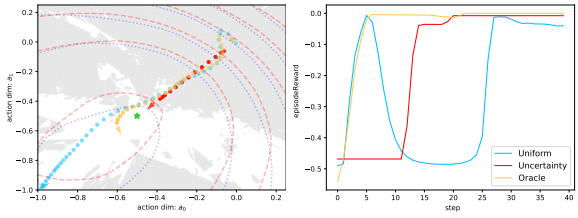

Figure 2: In a one-step MDP shooting scenario, we assess the impact of gradient uncertainty on performance. *Left*: Various methods' policy trajectories in action space, optimal action denoted by a green star. *Right*: Episode returns comparison.

of gradient directions, and the '**Oracle**' approach uses transitions that directly advance the action towards the optimum.

To further extend our methodology into practical RL applications and enhance the policy improvement process, we approach from a directional statistics perspective [14–16]. We employ the von Mises-Fisher distribution [17–19] to model the gradient directions corresponding to each transition during policy improvement. This method quantifies the uncertainty of gradient directions for each transition under the current ensemble critics. The formulations are presented in Eq. (1) and Eq.(2), which correspond to resampling probabilities based on uncertainty and rank, respectively.

$$P(j|x(s_j),\mathcal{D}) = \frac{\exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x(s_j))}{\sum_{i}^{M}\exp(\mathbf{R}(s_i)\mu^{\mathrm{T}}(s_i)x(s_i))} \tag{1}$$

$$P(j|x(s_j),\mathcal{D}) = \frac{rank(\exp(\mathbf{R}(s_j)\mu^{\mathrm{T}}(s_j)x(s_j)))^{-1}}{\sum_{i}^{M} rank(\exp(\mathbf{R}(s_i)\mu^{\mathrm{T}}(s_i)x(s_i)))^{-1}}. \tag{2}$$

3 EXPERIMENTS & CONCLUSION

We conducted experiments across six MuJoCo tasks (Hopper, Ant, Swimmer, HalfCheetah, Humanoid, Walker), evaluating the average performance improvement our method brings compared to baseline algorithms, as shown in Table 1.

The results consistently indicate performance enhancements from our method, demonstrating compatibility with most algorithms using the Actor-Critic framework. Overall, vMFER proves to be an effective scheme for optimizing the policy improvement process, offering significant performance gains with a marginal additional computational cost.

	SAC	TD3	SAC+PER
baseline	100%	100%	100%
vMFER (rank)	106.84%	111.62%	102.09%
vMFER (uncertainty)	**113.78%**	**117.75%**	**107.17%**

Table 1: Average performance improvement of vMFER over baseline, calculated by aggregating performance gains across all tasks.

REFERENCES

[1] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, pages 2226–2240, 2023.

[2] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426, 2023.

[3] Saminda Wishwajith Abeyruwan, Laura Graesser, David B D'Ambrosio, Avi Singh, Anish Shankar, Alex Bewley, Deepali Jain, Krzysztof Marcin Choromanski, and Pannag R Sanketi. i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops. In *Conference on Robot Learning*, pages 212–224, 2023.

[4] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[5] Hado Hasselt. Double q-learning. *Advances in Neural Information Processing Systems*, 23, 2010.

[6] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

[7] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[8] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596, 2018.

[9] Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

[10] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional Statistics*, volume 2. Wiley Online Library, 2000.

[11] Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

[12] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[13] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

[14] Christopher Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974.

[15] John T Kent. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

[16] Frank Bowman. *Introduction to Bessel functions*. Courier Corporation, 2012.

[17] Geoffrey S Watson. Distributions on the circle and sphere. *Journal of Applied Probability*, 19(A):265–280, 1982.

[18] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

[19] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of I_s (x). *Computational Statistics*, 27:177–190, 2012.