ENOTO: Improving Offline-to-Online Reinforcement Learning with Q-Ensembles

Extended Abstract

Kai Zhao
College of Intelligence and Computing, Tianjin University
Tianjin, China
kaizhao@tju.edu.cn

Jianye Hao*
College of Intelligence and Computing, Tianjin University
Tianjin, China
jianye.hao@tju.edu.cn

Yi Ma
College of Intelligence and Computing, Tianjin University
Tianjin, China
mayi@tju.edu.cn

Jinyi Liu
College of Intelligence and Computing, Tianjin University
Tianjin, China
jyliu@tju.edu.cn

Yan Zheng
College of Intelligence and Computing, Tianjin University
Tianjin, China
yanzheng@tju.edu.cn

Zhaopeng Meng
College of Intelligence and Computing, Tianjin University
Tianjin, China
mengzp@tju.edu.cn

ABSTRACT

Offline reinforcement learning (RL) is a learning paradigm where an agent learns from a fixed dataset of experience. However, learning solely from a static dataset can limit the performance due to the lack of exploration. To overcome it, offline-to-online RL combines offline pre-training with online fine-tuning, which enables the agent to further refine its policy by interacting with the environment in real-time. Despite its benefits, existing offline-to-online RL methods suffer from performance degradation and slow improvement during the online phase. To tackle these challenges, we propose a novel framework called **EN**semble-based **O**ffline-**T**o-**O**nline (ENOTO) RL. By increasing the number of Q-networks, we seamlessly bridge offline pre-training and online fine-tuning without degrading performance. Moreover, to expedite online performance enhancement, we appropriately loosen the pessimism of Q-value estimation and incorporate ensemble-based exploration mechanisms into our framework. Experimental results demonstrate that ENOTO can substantially improve the training stability, learning efficiency, and final performance of existing offline RL methods during online fine-tuning on a range of locomotion tasks, significantly outperforming existing offline-to-online RL methods.

KEYWORDS

Offline Reinforcement Learning; Online Fine-tuning; Offline-to-online RL; Ensemble Methods

ACM Reference Format:
Kai Zhao, Jianye Hao*, Yi Ma, Jinyi Liu, Yan Zheng, and Zhaopeng Meng. 2024. ENOTO: Improving Offline-to-Online Reinforcement Learning with Q-Ensembles: 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

Offline-to-online RL has the potential to address the limitations of offline RL, such as the sub-optimality of learned policy. Furthermore, starting with an offline RL policy can achieve strong performance with fewer online environment samples, compared to collecting large amounts of training data by rolling out policies from scratch. Prior offline-to-online RL researches have shown that inefficient policy improvement due to pessimistic learning and unstable learning caused by the distributional shift are the main problems [5, 7, 8]. Existing offline-to-online RL methods have attempted to address these challenges through implicit policy constraints [8], filtering offline data used for online fine-tuning [5–7], adjusting policy constraint weights carefully [12], or training more online policies [11]. Nevertheless, these methods still face performance degradation in some tasks and settings, and their performance improvement in the online phase is limited.

Taking inspiration from leveraging Q-ensembles in offline RL [1], we find that Q-ensembles help to alleviate unstable training and performance degradation, and can serve as a more flexible pessimistic term by encompassing various target computation and exploration methods during the online fine-tuning phase. Based on this discovery, we propose an **EN**semble-based **O**ffline-**T**o-**O**nline (ENOTO) RL framework that bridges offline pre-training and online fine-tuning. Our main contributions are summarized as follows:

- We demonstrate the effectiveness of Q-ensembles in bridging the gap between offline pre-training and online fine-tuning, providing a solution for mitigating the common problem of unstable training and performance drop.
- We propose a unified framework ENOTO for offline-to-online RL, which enables a wide range of offline RL algorithms to transition from pessimistic offline pre-training to optimistic online fine-tuning, leading to stable and efficient performance improvement.
- We empirically validate the effectiveness of ENOTO on various benchmark tasks, including locomotion and navigation

* Corresponding author.

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).

tasks, and verify that ENOTO achieves state-of-the-art performance in comparison to all baseline methods.

2 ENSEMBLE-BASED OFFLINE-TO-ONLINE REINFORCEMENT LEARNING

Based on our observations, we find that Q-ensembles can maintain certain conservative capabilities to mitigate unstable training and performance drop, functioning as a more versatile constraint method for exploring more diverse actions during online fine-tuning compared to offline RL algorithms. Consequently, we propose our **EN**nsemble-based **O**ffline-**T**o-**O**nline (ENOTO) RL Framework. Specifically, we first integrate offline RL algorithm with Q-ensembles in the offline phase. Then in the online phase, we remove the original pessimistic term. Furthermore, we compute the target as the expectation of all the targets, where the expectation is taken over all N-choose-2 pairs of Q-functions. Additionally, we introduce SUNRISE to encourage exploration. Although each individual design decision in ENOTO may seem relatively simple, their specific combination outperforms baselines in terms of training stability, learning efficiency and final performance. Algorithm 1 summarizes the offline and online procedures of ENOTO.

Algorithm 1 ENOTO: **EN**semble-based **O**ffline-**T**o-**O**nline RL Framework

```
Input: Offline dataset $D_{offline}$, offline RL algorithm OfflineRL
Output: Offline to online learning algorithm
// Offline Phase
Turning offline RL algorithm OfflineRL into OfflineRL-N with integration of Q-ensembles.
Training OfflineRL-N using $D_{offline}$
// Online Phase
Removing original pessimistic term in OfflineRL (if possible) and thus turn OfflineRL-N to OnlineRL-N
Setting the Q-target computation method to WeightedMinPair and obtain OnlineRL-N + WeightedMinPair
Introducing SUNRISE to encourage exploration and obtain OnlineRL-N + WeightedMinPair + SUNRISE
return OfflineRL-N → OnlineRL-N + WeightedMinPair + SUNRISE
```

3 EXPERIMENTS

We evaluate our ENOTO framework on MuJoCo [10] locomotion tasks, i.e., HalfCheetah, Walker2d, and Hopper from the D4RL benchmark suite [2]. The goal of each task is to move forward as fast as possible, while keeping the control cost minimal. To demonstrate the applicability of ENOTO on various suboptimal datasets, we use three dataset types: medium, medium-replay, and medium-expert. The medium datasets contain rollouts from medium-level policies. The medium-replay datasets encompass all samples collected during the training of a medium-level agent from scratch. In the case of the medium-expert datasets, half of the data comprises rollouts from medium-level policies, while the other half consists of rollouts from expert-level policies. In this study, we exclude the random and the expert datasets, as in typical real-world scenarios, we rarely use a random policy or have an expert policy for system control. We utilize the v2 version of each dataset. We pre-train the agent for 1M training steps in the offline phase and perform online fine-tuning for 250K environmental steps.

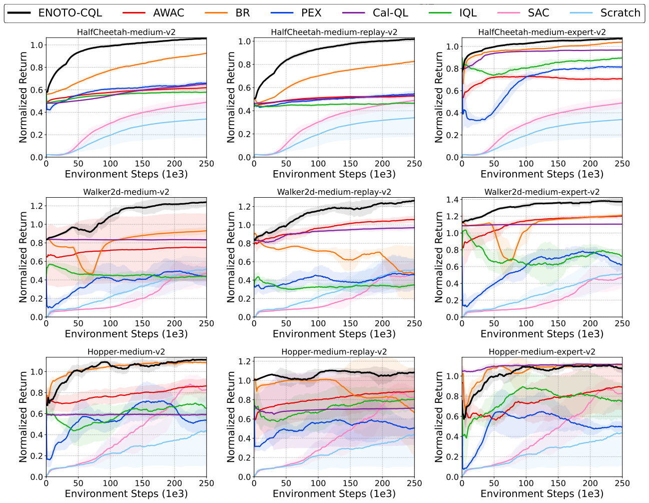

Figure 1: Online learning curves of different methods across five seeds on MuJoCo locomotion tasks. The solid lines and shaded regions represent mean and standard deviation, respectively.

We consider the following methods as baselines, including AWAC [8], BR [5], PEX [11], Cal-QL [9], IQL [4], SAC [3] and Scratch (i.e. SAC-N + WeightedMinPair + SUNRISE). Fig 1 shows the performance of the ENOTO-CQL method (ENOTO instantiated on CQL) and baseline methods during the online fine-tuning phase. We can see that our ENOTO method surpasses the baseline approaches in terms of training stability, learning efficiency, and final performance across most tasks.

4 CONCLUSIONS AND FUTURE WORK

In this work, we have demonstrated that Q-ensembles can be efficiently leveraged to alleviate unstable training and performance drop, and serve as a more flexible constraint method for online fine-tuning. Based on this observation, we propose Ensemble-based Offline-to-Online (ENOTO) RL Framework, which enables pessimistic offline RL algorithms to perform optimistic online fine-tuning and improve their performance efficiently while maintaining stable training process. We conducted experiments on a wide range of tasks to demonstrate its effectiveness. Future work will concentrate on reducing the computational overhead to make ENOTO more scalable and practical for large-scale problems and real-world applications, enabling the development of more efficient and reliable offline-to-online RL systems.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant Nos. 92370132), the National Key R&D Program of China (Grant No. 2022ZD0116402) and the Xiaomi Young Talents Program of Xiaomi Foundation.

REFERENCES

[1] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. 2021. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems* 34 (2021), 7436–7447.

[2] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219* (2020).

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

[4] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169* (2021).

[5] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. 2022. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*. PMLR, 1702–1712.

[6] Yihuan Mao, Chao Wang, Bin Wang, and Chongjie Zhang. 2022. MOORe: Model-based Offline-to-Online Reinforcement Learning. *arXiv preprint arXiv:2201.10070* (2022).

[7] Max Sobol Mark, Ali Ghadirzadeh, Xi Chen, and Chelsea Finn. 2022. Fine-tuning Offline Policies with Optimistic Action Selection. In *Deep Reinforcement Learning Workshop NeurIPS 2022*.

[8] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359* (2020).

[9] Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. 2023. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *arXiv preprint arXiv:2303.05479* (2023).

[10] 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*. IEEE, 5026–5033.

[11] Haichao Zhang, We Xu, and Haonan Yu. 2023. Policy Expansion for Bridging Offline-to-Online Reinforcement Learning. *arXiv preprint arXiv:2302.00935* (2023).

[12] Yi Zhao, Rinu Boney, Alexander Ilin, Juho Kannala, and Joni Pajarinen. 2022. Adaptive behavior cloning regularization for stable offline-to-online reinforcement learning. *arXiv preprint arXiv:2210.13846* (2022).