Dual-Policy-Guided Offline Reinforcement Learning with Optimal Stopping

Extended Abstract

Weibo Jiang
Tsinghua Shenzhen International
Graduate School, Tsinghua University
Shenzhen, China
jwb21@mails.tsinghua.edu.cn

Shaohui Li
Tsinghua Shenzhen International
Graduate School, Tsinghua University
Shenzhen, China
lishaohui@sz.tsinghua.edu.cn

Zhi Li
Tsinghua Shenzhen International
Graduate School, Tsinghua University
Shenzhen, China
zhilizl@sz.tsinghua.edu.cn

Yuxin Ke
Tsinghua Shenzhen International
Graduate School, Tsinghua University
Shenzhen, China
keyx22@mails.tsinghua.edu.cn

Zhizhuo Jiang
Tsinghua Shenzhen International
Graduate School, Tsinghua University
Shenzhen, China
jiangzhizhuo@sz.tsinghua.edu.cn

Yaowen Li
Tsinghua Shenzhen International
Graduate School, Tsinghua University
Shenzhen, China
liyw23@sz.tsinghua.edu.cn

Yu Liu*
Department of Electronics, Tsinghua
University
Beijing, China
liuyu77360132@126.com

ABSTRACT

Policy-guided offline reinforcement learning (POR) decomposes the offline reinforcement learning (offline RL) problem into goal estimation and goal-conditioned execution subproblems, leading to improved performance. However, we reveal that the preciseness of the estimated goal massively affects the performance and robustness of the trained goal-conditioned policy. To overcome this problem, we propose an offline RL model with dual guide-policies to improve the preciseness of the goal and reduce the variance. The proposed dual-policy-guided offline RL (Dual POR) adopts an integrating function, which balances the goals predicted by two guide-policies to obtain a refined goal. Moreover, we employ the optimal stopping strategy to schedule the training process, which dramatically shortens the training process and improves the generalization. The proposed Dual POR achieves state-of-the-art performance on the D4RL datasets with reduced variances. The improvements in high-complexity tasks are even significant, which indicates the potential of the proposed Dual POR in real-world applications.

KEYWORDS

Offline Reinforcement Learning; Goal-conditioning; Optimal Stopping; Imitation Learning

ACM Reference Format:
Weibo Jiang, Shaohui Li, Zhi Li, Yuxin Ke, Zhizhuo Jiang, Yaowen Li, and Yu Liu. 2024. Dual-Policy-Guided Offline Reinforcement Learning with Optimal Stopping: 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

Goal-conditioning methods reduce the complicated optimization problem into goal-conditioned subproblems [4, 7, 10]. The recent policy-guided method [9] incorporates the goal-conditioning insight into Offline RL by simultaneously learning a guide-policy to offer a goal (target state) for the goal-conditioned execute-policy. Compared to the previous works, the policy-guided method achieves remarkable improvement with a simple imitation learning strategy over the classical locomotion problem, *i.e.*, AntMaze. However, previous offline RL approaches employ a single policy to decide actions. Therefore, the performance massively depends on this policy network, resulting in diminished robustness and inadequate interpretability. The limited dataset (*i.e.*, trajectories from expert policy) used for offline RL makes the problem even more severe.

This paper proposes a dual-policy-guided offline RL (Dual POR) model to improve the goal estimation accuracy and reduce the variance. We introduce an integrating function to generate the integrated goal regarding the two goals from the guide-policies. Furthermore, we employ an optimal stopping approach to refine the execute-policy and avoid over-fitting on the limited offline dataset. Experimental results show that the proposed Dual POR model archives the state-of-the-art performance on the popular offline RL dataset (*i.e.*, the D4RL dataset [3]) with improved performance and reduced variance. In summary, the contributions of this paper are listed below: *i)* We propose a dual-policy-guided offline RL

*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).

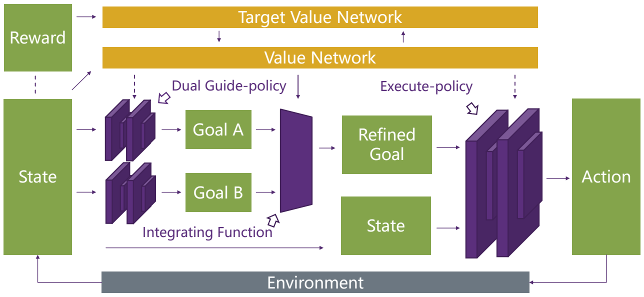

Figure 1: Framework of the proposed dual-policy-guided offline RL (Dual POR) model.

(Dual POR) model, which offers refined goals for goal-conditioned execute-policy. *ii)* We introduce an integrating function and the optimal stopping algorithm into the Dual POR model for enhanced generalization. *iii)* The proposed method achieves state-of-the-art performance on the offline RL dataset with improved robustness.

2 METHODS

Figure 1 depicts the proposed dual-policy-guided offline RL. We achieve the offline RL with dual goal-estimation guide-policies and a goal-conditioned execute-policy. Specifically, two goals whose names are *Goal A* and *Goal B* are derived from the dual guide-policies. Both goals are then fed to the integrating function to generate the *refined goal*. The *refined goal* and the original state are taken into the execute-policy to generate the action. The (target) value network estimates the state value based on the reward. It works as the weight of the training target of the dual guide-policies and the execute-policy, and as flexible weights of the integrating function to balance the two estimated goals.

Dual guide-policies. The role of guide-policies is to estimate the goal based on the current state. The corresponding mapping is $g_\omega : s \longrightarrow s'$ as we take the next state as the goal. Based on AWR [8], superior results can be achieved by computing the Bellman residual for the value functions of the current state and the next state and utilizing the resulting values as advantage weight for the guide-policy. Then, the optimization objective for the guide-policy can be formulated as Eq. (1).

$$J(\omega_i) = \max_{\omega_i} \mathbb{E}_{(s,s')}\left[e^{\left(r+\gamma V_{\phi'}(s')-V_\phi(s)\right)} \log p\left(s'|g_{\omega_i}(s)\right)\right], \quad (1)$$

where $g_{\omega_i}(s)$ with $i \in \{A, B\}$ are two guide-policies and ω_i are their leanable parameters. Here, we denote the two estimated goals from the guide-policies as s'_A and s'_B. Two guide-policies are trained separately with different initialization. They offer two independent goal estimates from two diverse fitted models, which provide more sufficient information than a single guide-policy.

Integrating function. We assign weights to balance these two goals when integrating them to obtain the high-precision refined goal. The goal with a higher value is assigned a weight of η ($\eta \in [0, 1]$), while the weight $(1 - \eta)$ is allocated to the other goal. The integrating function is formulated as Eq. (2).

$$s'_{\text{ref}} = \eta \cdot \underset{s'_i \in \{s'_A, s'_B\}}{\arg\max}\ V(s'_i) + (1-\eta) \cdot \underset{s'_i \in \{s'_A, s'_B\}}{\arg\min}\ V(s'_i) \quad (2)$$

Optimal stopping. The model further adopts the optimal stopping strategy to prevent overfitting and reduce training time. General offline reinforcement learning task requires a series of evaluations during the model training. A well-trained policy should stop training to avoid overfitting, preventing performance degradation. The secretary problem of the optimal stopping theory concludes that the well-trained policy should measure T/e rounds and terminate once the rest one evaluation score exceeds the maximum score in measurement period. Here T is the total evaluation round, and e is the natural logarithm.

3 EXPERIMENTS

The proposed Dual POR is compared with latest arts including CQL [6], RvS [2], IQL [5] and POR [9]. The benchmark dataset is D4RL [3] which follows the settings of prior arts above. The score is formluated as the D4RL normalized score ranging from 0 to 100, with 100 as expert performance and 0 as ramdom performance.

Table 1: Comparison of Dual POR and other baseline methods on AntMaze datasets.

D4RL Dataset	CQL	RvS	IQL	POR	Dual POR
antmaze-u	74.0	65.4	87.5	90.6±7.1	**94.8±3.5**
antmaze-u-d	**84.0**	60.9	66.2	71.3±12.1	83.6±**3.7**
antmaze-m-p	61.2	58.1	71.2	84.6±5.6	**94.0±2.5**
antmaze-m-d	53.7	67.3	70.0	79.2±**3.1**	**90.4**±4.6
antmaze-l-p	15.8	32.4	39.6	58.0±12.4	**76.8±3.7**
antmaze-l-d	14.9	36.9	47.5	73.4±**8.5**	**73.6±10.0**
AntMaze mean	36.4	53.5	63.6	76.2±8.1	**85.5±4.7**

Table 1 shows the performance on the average score and the variance of each D4RL AntMaze dataset. The proposed Dual POR achieves state-of-the-art performance on the D4RL datasets with reduced variances. We also conduct the comparison experiment on Gym-MuJoCo [1], including Halfcheetah, Hopper, and Walker2D datasets. Dual POR achieves an average score of 84.4, exceeding the prior art of an average score of 81.8. Moreover, the optimal stopping strategy reduces the training epoch by 33.6% and 27.7% for AntMaze and Gym-MuJoCo tasks, respectively.

4 CONCLUSIONS

This paper presents a dual-policy-guided offline reinforcement learning model utilizing dual guide-policies. We incorporate the value function into policy generation, which enriches the information available for policy generation within this evaluation paradigm. We apply the optimal stopping strategy to model evaluation, helping us capture optimal policy quickly. This work shows the improvement of employing more accurate goals in goal-conditioning offline RL methods. More theoretical explorations are expected to gain further improvement.

ACKNOWLEDGMENTS

This work was partially supported by grants from the National Natural Science Foundation of China (No. 62293544, 62022092, and 62301299) and the China Postdoctoral Science Foundation (No. 2022M720077).

REFERENCES

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

[2] Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. 2022. RvS: What is Essential for Offline RL via Supervised Learning?. In *International Conference on Learning Representations*. https://openreview.net/forum?id=S874XAIpkR-

[3] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. arXiv:2004.07219 [cs.LG]

[4] Leslie Pack Kaelbling. 1993. Learning to achieve goals. In *IJCAI*, Vol. 2. Citeseer, 1094–8.

[5] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*.

[6] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.

[7] Minghuan Liu, Menghui Zhu, and Weinan Zhang. 2022. Goal-Conditioned Reinforcement Learning: Problems and Solutions. In *European Conference on Artificial Intelligence*. 5502–5511.

[8] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177* (2019).

[9] Haoran Xu, Li Jiang, Jianxiong Li, and Xianyuan Zhan. 2022. A Policy-Guided Imitation Approach for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

[10] Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. 2022. Rethinking Goal-Conditioned Supervised Learning and Its Connection to Offline RL. In *International Conference on Learning Representations*.