

Offline Meta Reinforcement Learning with Weighted Policy Constraints and Proximal Context Collection

Extended Abstract

Haorui Li

State Key Laboratory of Multimodal Artificial Intelligence
Systems, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
lihaorui2021@ia.ac.cn

Jiaqi Liang

State Key Laboratory of Multimodal Artificial Intelligence
Systems, Institute of Automation, CAS
Beijing, China
liangjiaqi2014@ia.ac.cn

Linjing Li

State Key Laboratory of Multimodal Artificial Intelligence
Systems, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
linjing.li@ia.ac.cn

Daniel Zeng

State Key Laboratory of Multimodal Artificial Intelligence
Systems, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
dajun.zeng@ia.ac.cn

ABSTRACT

Offline meta-reinforcement learning (OMRL) encounters two key challenges: effectively learning the meta-policy from offline datasets and correctly inferring unseen tasks. Existing methods often address the first challenge by imposing policy constraints, but are limited by the suboptimal actions in offline datasets. For the second challenge, most focus on meta-training without enhancing task inference during meta-testing. To address these issues, we propose a novel method called weighted policy conStraints and proximal contExt coLlEction sTrategy for OMRL (SELECT). During meta-training, we integrate policy constraints with weighted behavior cloning, allowing for more flexible policy learning while maintaining desirable behaviors. In the meta-testing phase, SELECT introduces a proximal context collection strategy that balances exploration and exploitation. This strategy gathers high-quality context, improving task inference and adaptation to unseen tasks. Experimental results show that SELECT significantly reduces the distributional shift, enhances the meta-policy's generalization, and outperforms state-of-the-art methods across various domains.

KEYWORDS

Reinforcement Learning; Meta Learning; Imitation Learning; Task Adaptation

ACM Reference Format:

Haorui Li, Jiaqi Liang, Linjing Li, and Daniel Zeng. 2025. Offline Meta Reinforcement Learning with Weighted Policy Constraints and Proximal Context Collection: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, Detroit, Michigan, USA, May 19 – 23, 2025, IFAAMAS, 3 pages.



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

A key obstacle in OMRL is the distributional shift between the behavior policy used to collect offline data and the learned policy for new tasks. This mismatch often results in overestimating the values of out-of-distribution (OOD) states, as such errors cannot be corrected through interactions with the environment. Existing methods attempt to address this issue by constraining the learned policy to stay close to the behavior policy. However, when datasets include suboptimal or unsafe actions, these constraints can lead to undesirable policy imitation, ultimately degrading performance. Furthermore, existing studies often overlook context collection during meta-testing, hindering OMRL's generalization to new tasks.

To overcome these limitations, we propose a novel OMRL approach called weighted policy conStraints and proximal contExt coLlEction sTrategy for OMRL (SELECT), designed to optimize the utilization of offline datasets and promote rapid adaptation to new tasks. Our critical insight is that even sub-optimal actions can contain task-relevant information that enhances policy learning and adaptation in OMRL. SELECT builds on two key innovations:

- **Weighted policy constraints:** During meta-training, we combine policy constraint methods with weighted behavior cloning to selectively relax constraints on suboptimal actions while retaining necessary constraints for desirable ones. This balance enhances meta-policy performance without excessive reliance on inferior behaviors.
- **Proximal context collection:** During meta-testing, we introduce a strategy that balances exploration and exploitation by adding Gaussian noise to inferred actions. This approach enables the agent to gather diverse and informative contexts, improving task inference robustness and generalization.

We validate our approach by comparing SELECT with state-of-the-art OMRL methods in environments with varying reward and dynamic functions, such as Point-Robot and MuJoCo simulators [8]. Experimental results show that SELECT consistently outperforms existing methods, achieving superior performance.

2 METHOD

The SELECT method consists of four key components: 1) **Offline data collection**, which involves collecting trajectories using Soft Actor-Critic (SAC) [4] at different training checkpoints to form datasets for training the task inference and conditional policy modules. 2) **Task inference module training**, which updates the context encoder q_ϕ using maximum-minimum mutual information loss, as in CSRO [3], ensuring task embeddings are distinguishable and mitigating the effects of the distribution drift. 3) **Conditional policy module training**, which combines policy constraints with weighted behavior cloning to ensure the policy imitates desirable actions while learning from suboptimal actions to enhance overall performance. 4) **Meta-testing**, which uses a proximal context collection strategy to enable efficient adaptation to new tasks by iteratively exploring tasks, collecting context, updating the task distribution, and improving task inference and action selection.

Our contributions focus on the conditional policy module training and the meta-testing phases, represented by the weighted policy constraints and the proximal context collection strategy, respectively. We provide a concise overview of these two key innovations.

Weighted policy constraints: To effectively leverage offline datasets, we guide the learning policy to prioritize imitation of high-quality behaviors while minimizing the impact of less effective actions. Inspired by [2, 6], we use the following objective function to update the conditional policy π_ω :

$$\mathcal{L}_{actor}(\omega) = \mathbb{E}_{(s,a) \sim \mathcal{D}} [\lambda Q_\theta(s, \pi_\omega(a|s, z)|z) - w(s, a|z)(\pi_\omega(a|s, z) - a)^2] \quad (1)$$

where $z \sim q_\phi(z|c)$ represents the output of the context encoder, c is the collected context, Q_θ denotes the action-value function, and the scalar λ is defined as

$$\lambda = \frac{\alpha}{\frac{1}{N_d} \sum_{(s_i, a_i)} |Q_\theta(s_i, a_i|z)|}, \quad (2)$$

where α is a hyperparameter, and N_d represents the number of transitions (s_i, a_i) in the dataset. The regularization term in the denominator of Equation (2) helps balance the scales of actions and rewards, minimizing the need for hyperparameter tuning.

Inspired by [6], we adopt the advantage function as the criterion for determining whether to imitate a given state-action pair:

$$w(s, a|z) = \mathbf{1}[A(s, a|z) > 0] = \mathbf{1}[Q_\theta(s, a|z) - V_\delta(s|z) > 0]. \quad (3)$$

In Equation (3), $A(s, a|z)$, $Q_\theta(s, a|z)$, and $V_\delta(s|z)$ represent the estimated advantage function, action-value function, and value function, respectively, conditioned on the task representation z over the mini-batch. To update the action-value and value functions, we follow the method in [2], using dual target networks for Q_θ updates, with V_δ relying on the outputs of the target Q networks.

Proximal context collection strategy: To enhance diversity and reduce bias in task inference during meta-testing, we use the learned meta-policy as the exploration strategy for new tasks. Before executing the action $\pi_\omega(a|s, z)$, we add random noise $\epsilon_{test} \sim \mathcal{N}(0, \sigma_{test})$, which is drawn from a Gaussian distribution with a standard deviation of σ_{test} . The final action is then defined as

$$\hat{a} = \pi_\omega(a|s, z) + \text{clip}(\epsilon_{test}, \epsilon_{test}^{\min}, \epsilon_{test}^{\max}), \quad (4)$$

where $\text{clip}(x, x_{\min}, x_{\max})$ constrains x within the range $[x_{\min}, x_{\max}]$.

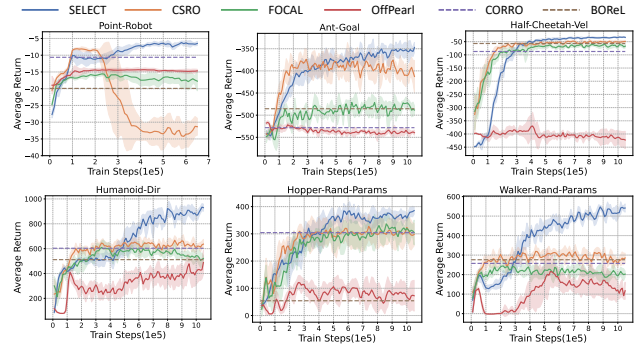


Figure 1: The average return during the online test phase for unseen test tasks compared to other OMRL methods. The solid line represents the mean value across five seeds, and the shaded region shows the standard deviation.

3 EXPERIMENTS

We evaluate the proposed method on the Point-Robot and MuJoCo physics simulators [8], which are often used to access the OMRL performance. We compare SELECT with OffPearl [7], FOCAL [5], BOREL [1], CORRO [9] and CSRO [3].

After training on numerous pre-collected datasets, we conducted online testing with limited data to ensure applicability. Figure 1 presents the average return and standard deviation of SELECT compared to other OMRL baselines across five seeds, where higher average returns indicate better performance. The results demonstrate that SELECT consistently outperforms the baselines across all six environments, with notable improvements in Point-Robot, Humanoid-Dir, and Walker-Rand-Params. These findings validate the effectiveness of our approach, highlighting its robust performance and adaptability across diverse tasks.

4 CONCLUSION

In this paper, we propose SELECT, a novel context-based method that enhances offline datasets and improves context collection during meta-testing. During the meta-training, we combine policy constraints with weighted behavior cloning to focus on desirable behaviors while allowing exploration of undesirable actions. Moreover, in the meta-testing phase, we introduce a proximal context collection strategy to improve the quality of sampled contexts, which enables robust task inference and effective task-specific action execution. Experimental results show that SELECT significantly reduces the distributional shift, enhances the performance and generalization capability for the meta-policy across multiple challenging domains.

ACKNOWLEDGMENTS

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27030100 and the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems MAIS2024310, and in part by the National Natural Science Foundation of China under Grants 72293575 and Grant 72293573.

REFERENCES

- [1] Ron Dorfman, Idan Shenfeld, and Aviv Tamar. 2021. Offline Meta Reinforcement Learning – Identifiability Challenges and Effective Data Collection Strategies. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 34. 4607–4618.
- [2] Scott Fujimoto and Shixiang (Shane) Gu. 2021. A Minimalist Approach to Offline Reinforcement Learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 34. 20132–20145.
- [3] Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, Qi Guo, Ling Li, and Yunji Chen. 2023. Context Shift Reduction for Offline Meta-Reinforcement Learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 36. 80024 – 80043.
- [4] 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 International Conference on Machine Learning*, Vol. 80. 1856–1865.
- [5] Lanqing Li, Rui Yang, and Dijun Luo. 2021. FOCAL: Efficient Fully-Offline Meta-Reinforcement Learning via Distance Metric Learning and Behavior Regularization. In *Proceeding of the International Conference on Learning Representations*.
- [6] Zhiyong Peng, Changlin Han, Yadong Liu, and Zongtan Zhou. 2023. Weighted Policy Constraints for Offline Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9435–9443.
- [7] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. 2019. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *Proceedings of the International Conference on Machine Learning*, Vol. 97. 5331–5340.
- [8] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In *Proceedings of the International Conference on Intelligent Robots and Systems*. 5026–5033.
- [9] Haoqi Yuan and Zongqing Lu. 2022. Robust Task Representations for Offline Meta-Reinforcement Learning via Contrastive Learning. In *Proceedings of the International Conference on Machine Learning*, Vol. 162. 25747–25759.