Mitigating Constraint Conflict in Offline RL: An Adaptive Weighted Constraint Approach

Extended Abstract

Pengyu Chen
Harbin Institute of Technology
Harbin, China
chenpengyu02@foxmail.com

Shirong Liu
Harbin Institute of Technology
Harbin, China
shirongliu16@gmail.com

Minye Huang
Harbin Institute of Technology
Harbin, China
2212131772@qq.com

Haozhuo Zheng
Harbin Institute of Technology
Harbin, China
24b903065@stu.hit.edu.cn

Haoyu Liu
Harbin Institute of Technology
Harbin, China
liuhaoyu@stu.hit.edu.cn

Wenyu Yuan
Harbin Institute of Technology
Harbin, China
wenyuyuan0904@gmail.com

Yang Liu
Harbin Institute of Technology
Harbin, China
liuyang@hit.edu.cn

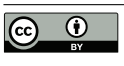
ABSTRACT

Offline Reinforcement Learning allows agents to learn policies from pre-collected datasets by imposing conservative constraints to address the out-of-distribution problem. However, existing methods face a critical challenge of constraint conflict with datasets generated by multiple behavior policies, and these policies suggest conflicting actions that lead the agent towards suboptimal performance. Geometric distance-based and advantage-weighted methods can be employed to address this problem. These methods exhibit several limitations, including sensitivity to low-quality data, high computational cost, and over conservatism. To overcome these limitations, we propose Adaptive Weighted Constraint (AWC), which mitigates constraint conflicts by training a constraint network via adaptive weighted behavior cloning. AWC dynamically assigns importance weights to dataset actions based on their consistency with the current policy, ensuring that the constraint is informed by the behavior and its distance to the policy. Inspired by the robustness of central tendency estimators in statistics, we apply the weighted geometric median of the actions as a stable target for the policy constraint. Experiments on D4RL benchmarks demonstrate AWC outperforms prior methods on a majority of tasks.

KEYWORDS

Offline Reinforcement Learning; Constraint Conflict; Geometric Median

ACM Reference Format:
Pengyu Chen, Shirong Liu, Minye Huang, Haozhuo Zheng, Haoyu Liu, Wenyu Yuan, and Yang Liu. 2026. Mitigating Constraint Conflict in Offline RL: An Adaptive Weighted Constraint Approach: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/CYXQ3092



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CYXQ3092

1 INTRODUCTION

Reinforcement Learning (RL) enables agents to learn optimal decision making through interactions with the environment [29]. However, in many real-world applications, such as robotics and healthcare, continuous interaction with the environment is impractical or unsafe [15, 26, 32, 33]. This limitation has motivated the development of Offline Reinforcement Learning (Offline RL), also known as Batch RL [7, 13]. Its goal is to learn effective policies from pre-collected datasets without interactions with the environment. The key challenge in Offline RL is the Out-of-Distribution (OOD) problem. When the learned policy executes actions that are not well covered by the dataset, the value function can suffer from extrapolation errors [23], leading to catastrophic performance failures [13]. This issue is typically addressed by constraining the policy to remain close to the behavior policy through regularization [6, 7, 11, 30, 31], or by penalizing the value function for OOD actions [1, 10, 12].

While effective for datasets that are collected by a single behavior policy, these methods may be ineffective when the dataset is collected by multiple behavior policies—a common scenario in practice. In such cases, the agent encounters the constraint conflict problem, where there exist conflicting actions from multiple behavior policies, this problem forces the learning policy toward an ineffective average behavior, leading to overly conservative or suboptimal performance [2, 9, 16, 17, 24, 28]. The constraint conflict problem can be addressed through geometric distance-based or advantage-weighted methods. Geometric distance-based methods design constraints by prioritizing proximal actions, based on the actions in the dataset and their geometric distance to the current policy [14, 25, 27]. However, such methods can be highly sensitive

to corrupted or low-quality data: if corrupted or low-quality state-action pairs are chosen to design constraints, they may misguide the learning process and lead to a suboptimal policy. Additionally, they often require exhaustive searches across the entire dataset during the learning process, resulting in high computational costs. Advantage-weighted methods [16, 20–22, 24] guide policies towards high-advantage actions, thereby naturally avoiding constraint conflicts from lower-advantage actions. However, these methods may tend to learn overly conservative policies when dealing with low-quality datasets, significantly limiting their effectiveness [8].

To overcome the limitations of existing methods, we propose a novel approach called Adaptive Weighted Constraint (AWC). Specifically, our method begins by applying an action distance metric to assign weights to the actions in the offline dataset, giving higher weights to actions that are closer to the current policy. Based on these weights, a constraint network is trained to theoretically learn the weighted geometric median as the constraint target. The geometric median is naturally resistant to outliers and effectively identifies central action tendencies, making it suitable for stable policy constraints, especially for low-quality or mixed-quality datasets [3, 4, 18, 19].

In summary, constraint conflict is a critical challenge in offline RL, especially when learning from datasets generated by multiple behavior policies. By combining action-distance weighting with the weighted geometric median as a stable constraint target, our method addresses this issue.

2 METHODOLOGY

Our method begins by assigning adaptive weights to actions in the offline dataset $\mathcal{D}$ based on their proximity to the current policy. For each given state s, we consider all associated actions $a \in \mathcal{D}_s$, where $\mathcal{D}_s = \{a \mid (s,a) \in \mathcal{D}\}$ denotes the set of dataset actions paired with state s. The weighting scheme prioritizes actions that are geometrically closer to the current policy's output, with weights increasing as the Euclidean distance decreases. Specifically, we compute weights as follows:

$$w(s,a) = \frac{1}{\|\pi_{\text{actor}}(s) - a\|_2^2 + \epsilon_w}, \tag{1}$$

where $\pi_{\text{actor}}(s)$ represents the current actor policy being optimized, and ϵ_w is a hyperparameter, which is set to a small positive value to prevent division by zero and to bound the maximum weight. Then we normalize these weights across all actions associated with the state s:

$$w'(s,a) = \frac{w(s,a)}{\sum_{a' \in \mathcal{D}_s} w(s,a')}. \tag{2}$$

To construct effective policy constraints while avoiding interference from irrelevant or conflicting actions, we apply a filtering mechanism. We select actions whose normalized weights exceed a threshold ϵ_a, forming a candidate set $\mathcal{A}_c(s)$:

$$\mathcal{A}_c(s) = \{a \mid a \in \mathcal{D}_s, w'(s,a) > \epsilon_a\}, \tag{3}$$

which removes actions from $\mathcal{D}_s$ that lie far from the current policy in the action space, thereby reducing the influence of potentially misleading behaviors.

To address outliers and conflicting actions, we utilize the geometric median as the constraint target, inspired by its statistical robustness properties. For a given state s, the constraint target $a_{\text{gm}}(s)$ is defined as the weighted geometric median of the filtered action candidate set $\mathcal{A}_c(s)$. As it lacks a closed-form solution, the target is implicitly defined by the following optimization problem:

$$a_{\text{gm}}(s) = \arg\min_{y \in \mathbb{R}^{|\mathcal{A}|}} \sum_{a \in \mathcal{A}_c(s)} w'(s,a) \cdot \|y - a\|_2 \tag{4}$$

It has been proven that the geometric median provides strong theoretical advantages as a robust estimator of centrality [19]. This may be especially beneficial in offline RL with mixed-quality datasets, where suboptimal or outlier actions could mislead the policy.

The constraint network parameters θ are optimized to learn $a_{\text{gm}}(s)$ by minimizing the expected weighted distance to high-importance actions within each batch, which is defined by the following loss function based on Equation (8):

$$\mathcal{L}_{\text{con}}(\theta) = \mathbb{E}_{s \sim \mathcal{B}, a \sim \mathcal{A}_c(s)} \left[w'(s,a) \cdot \|\pi_{\text{con}}(s;\theta) - a\|_2 \right], \tag{5}$$

The actor network is by minimizing the following loss function:

$$\begin{aligned} \mathcal{L}_{\text{actor}}(\phi) = \mathbb{E}_{s \sim \mathcal{D}} \big[& -\lambda Q(s, \pi_{\text{actor}}(s;\phi)) \\ & + \|\pi_{\text{actor}}(s;\phi) - \pi_{\text{con}}(s;\theta)\|_2^2 \big], \end{aligned} \tag{6}$$

where the first term maximizes the expected return through the Q-function, the second term regularizes the actor policy towards the constraint network output, and the hyperparameter λ balances these two objectives.

3 EXPERIMENTS

To evaluate the effectiveness of the proposed approach, we conduct experiments on the D4RL benchmarks [5]. Table 1 reports the average normalized scores and standard deviations across 10 random seeds. AWC demonstrates strong performance, achieving the best results in 5 out of 9 tasks.

Table 1: Average normalized scores. Scores with the highest mean are highlighted.

Task name(-v2)	TD3+BC	A2PO	wPC	PRDC	DOGE	ours
halfcheetah-m	48.3	47.1	53.3	**63.5**	45.3	57.1 ± 0.7
halfcheetah-m-r	44.6	44.8	48.3	**55.0**	42.8	49.4 ± 0.5
halfcheetah-m-e	90.7	95.6	93.7	94.5	78.7	**101.1** ± 1.0
walker2d-m	83.7	84.9	86.0	85.2	86.8	**91.5** ± 5.7
walker2d-m-r	81.8	82.8	89.9	92.0	87.3	**96.8**±1.4
walker2d-m-e	110.1	112.1	110.1	111.2	110.4	**112.8** ± 0.4
hopper-m	59.3	80.3	86.5	**100.3**	98.6	89.9 ±7.2
hopper-m-r	60.9	101.6	97.0	100.1	76.2	**102.1** ± 0.3
hopper-m-e	98.4	**113.4**	95.7	109.2	102.7	111.9 ± 2.0

4 CONCLUSION

This paper introduces the Adaptive Weighted Constraint, a novel approach for effectively resolving constraint conflicts in offline RL. AWC achieves effective policy learning from mixed-quality datasets. First, we develop an adaptive weighting scheme that dynamically prioritizes actions according to their relevance to the current policy. Second, we propose a constraint method based on the weighted geometric median, which demonstrates strong robustness against outlier actions and conflicting constraints. Evaluations on the D4RL benchmarks show that our method achieves competitive performance.

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] Xi Chen, Ali Ghadirzadeh, Tianhe Yu, Jianhao Wang, Alex Yuan Gao, Wenzhe Li, Liang Bin, Chelsea Finn, and Chongjie Zhang. 2022. Lapo: Latent-variable advantage-weighted policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 36902–36913.

[3] P Thomas Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. 2008. Robust statistics on Riemannian manifolds via the geometric median. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[4] P Thomas Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. 2009. The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage* 45, 1 (2009), S143–S152.

[5] 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]

[6] Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems* 34 (2021), 20132–20145.

[7] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*. PMLR, 2052–2062.

[8] Ranting Hu. 2025. CAWR: Corruption-Averse Advantage-Weighted Regression for Robust Policy Optimization. arXiv:2506.15654 [cs.LG] https://arxiv.org/abs/2506.15654

[9] Li Jiang, Sijie Cheng, Jielin Qiu, Haoran Xu, Wai Kin Chan, and Zhao Ding. 2024. Offline Reinforcement Learning with Imbalanced Datasets. arXiv:2307.02752 [cs.LG] https://arxiv.org/abs/2307.02752

[10] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline Reinforcement Learning with Implicit Q-Learning. arXiv:2110.06169 [cs.LG] https://arxiv.org/abs/2110.06169

[11] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems* 32 (2019).

[12] 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.

[13] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643 [cs.LG] https://arxiv.org/abs/2005.01643

[14] Jianxiong Li, Xianyuan Zhan, Haoran Xu, Xiangyu Zhu, Jingjing Liu, and Ya-Qin Zhang. 2023. When Data Geometry Meets Deep Function: Generalizing Offline Reinforcement Learning. arXiv:2205.11027 [cs.LG] https://arxiv.org/abs/2205.11027

[15] Zhongyu Li, Xuxin Cheng, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. 2021. Reinforcement learning for robust parameterized locomotion control of bipedal robots. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2811–2817.

[16] Tenglong Liu, Yang Li, Yixing Lan, Hao Gao, Wei Pan, and Xin Xu. 2024. Adaptive Advantage-Guided Policy Regularization for Offline Reinforcement Learning. arXiv:2405.19909 [cs.LG] https://arxiv.org/abs/2405.19909

[17] Yihuan Mao, Chengjie Wu, Xi Chen, Hao Hu, Ji Jiang, Tianze Zhou, Tangjie Lv, Changjie Fan, Zhipeng Hu, Yi Wu, et al. 2024. Stylized offline reinforcement learning: Extracting diverse high-quality behaviors from heterogeneous datasets. In *The Twelfth International Conference on Learning Representations*.

[18] Stanislav Minsker. 2015. Geometric median and robust estimation in Banach spaces. *Bernoulli* 21, 4 (Nov. 2015). https://doi.org/10.3150/14-bej645

[19] Stanislav Minsker and Nate Strawn. 2024. The geometric median and applications to robust mean estimation. *SIAM Journal on Mathematics of Data Science* 6, 2 (2024), 504–533.

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

[21] 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).

[22] 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*. 9435–9443.

[23] Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. 2023. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems* 35, 8 (2023), 10237–10257.

[24] Yunpeng Qing, Shunyu liu, Jingyuan Cong, Kaixuan Chen, Yihe Zhou, and Mingli Song. 2024. A2PO: Towards Effective Offline Reinforcement Learning from an Advantage-aware Perspective. arXiv:2403.07262 [cs.LG] https://arxiv.org/abs/2403.07262

[25] Yuhang Ran, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, and Yang Yu. 2023. Policy Regularization with Dataset Constraint for Offline Reinforcement Learning. arXiv:2306.06569 [cs.LG] https://arxiv.org/abs/2306.06569

[26] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. 2017. Deep reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:1704.02532* (2017).

[27] Yi Shen and Hanyan Huang. 2024. Hypercube Policy Regularization Framework for Offline Reinforcement Learning. arXiv:2411.04534 [cs.LG] https://arxiv.org/abs/2411.04534

[28] Anikait Singh, Aviral Kumar, Quan Vuong, Yevgen Chebotar, and Sergey Levine. 2022. Offline RL With Realistic Datasets: Heteroskedasticity and Support Constraints. arXiv:2211.01052 [cs.LG] https://arxiv.org/abs/2211.01052

[29] R.S. Sutton and A.G. Barto. 1998. *Reinforcement learning: An introduction*. MIT press Cambridge.

[30] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. 2022. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193* (2022).

[31] Jialong Wu, Haixu Wu, Zihan Qiu, Jianmin Wang, and Mingsheng Long. 2022. Supported Policy Optimization for Offline Reinforcement Learning. arXiv:2202.06239 [cs.LG] https://arxiv.org/abs/2202.06239

[32] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–36.

[33] Wenxuan Zhou, Sujay Bajracharya, and David Held. 2021. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*. PMLR, 1719–1735.