CDSA: Conservative Denoising Score-based Algorithm for Offline Reinforcement Learning

Extended Abstract

Zeyuan Liu*
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
gritmaybe@gmail.com

Kai Yang*
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
yk22@mails.tsinghua.edu.cn

Jiafei Lyu
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
lvjf20@mails.tsinghua.edu.cn

Xiu Li†
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
li.xiu@sz.tsinghua.edu.cn

ABSTRACT

Distribution shift is a major obstacle in offline reinforcement learning (RL). While existing conservative offline RL algorithms perform well in learning in-distribution policies, they often fail to generalize to unseen actions. To address this issue, we propose leveraging knowledge derived from the gradient fields of the dataset's density to refine and adjust the original actions. Building on this, we introduce the Conservative Denoising Score-based Algorithm (CDSA), which utilizes score-based diffusion models to estimate the gradients of the dataset density and generates action correction subcomponents to refine the actions. This approach enables more accurate and efficient decision-making during the testing phase in Markov Decision Process (MDP) environments. By decoupling conservatism constraints from the policy, our method is broadly applicable to various offline RL algorithms. Experiments demonstrate that our approach significantly enhances baseline performance on D4RL datasets and exhibits plug-and-play compatibility with different pre-trained offline RL policies.

KEYWORDS

Reinforcement Learning, Offline RL, Score-based Diffusion

ACM Reference Format:
Zeyuan Liu*, Kai Yang*, Jiafei Lyu, and Xiu Li†. 2025. CDSA: Conservative Denoising Score-based Algorithm for Offline Reinforcement Learning: 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.

*: Equal contribution. †: Corresponding authors.

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

Offline RL algorithms [5, 13–15, 18, 25] are prone to producing inaccurate predictions and catastrophic action commands when queried outside of the distribution of the training data, which leads to catastrophic outcomes. To strike a suitable trade-off between learning an improved policy and minimizing the divergence from the behavior policy, aiming to avoid errors due to distribution shift, previous work has provided various perspectives, including constraining the system in the training dataset distribution [5, 11, 13], developing a distributional critic to leverage risk-averse measures [16, 22], and reconstructing the density function of the training dataset [5, 6, 11, 17, 19]. However, most previous conservative offline RL algorithms failed to fully disentangle the knowledge related to conservatism from the algorithm's training process. This knowledge is typically incorporated into functions such as the final policy or critics, rendering it inseparable from other subcomponents. we explore the possibility of learning conservatism-related knowledge exclusively from the training dataset to obtain a plug-and-play decision adjuster. One intuitive approach is to leverage the density distribution of each dataset to guide the agent towards states located in areas of high density as much as possible. This can be achieved by adjusting the actions within the dataset to steer transitions towards states with higher density.

Inspired by the recent success of diffusion models [7–9, 20, 21, 23] and their applications in reinforcement learning (RL), we introduce the Conservative Denoising Score-based Algorithm (CDSA), a method designed to refine actions during testing without altering the training process of the original offline RL algorithm (as illustrated in Figure 1). CDSA modifies generated actions in a way that minimally disrupts the original decision-making, avoiding the need for subjective human input or rigid constraints that could limit the algorithm's adaptability. To mitigate the negative impacts of network uncertainty, we employ a strategy of performing only a single inference step with the inverse dynamics model at each timestep, effectively reducing potential errors. Experimental results show that CDSA consistently improves the performance of various offline RL baseline algorithms and can be seamlessly integrated

without requiring fine-tuning or additional conservatism-related training.

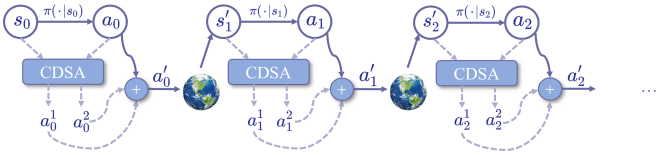

Figure 1: CDSA generates action subcomponents, utilizing conservatism-related knowledge acquired from the training dataset, to be added to the action generated from a pretrained policy π.

2 METHODOLOGY

In this section, we provide the implementation details of our proposed method, CDSA (Conservative Denoising Score-based Algorithm).

Learning the Gradient Field from Dataset. The core idea of our CDSA is to align agent behavior with high-density regions of the dataset's state-action distribution, $p_{\text{data}}(s, a)$. Inspired by score matching, we approximate the gradient fields $\nabla_{(s,a)} \log p_{\text{data}}(s, a)$ to identify directions that increase trajectory likelihood. Two independent score-based diffusion models are trained to separately estimate gradients for actions and states, avoiding dependency challenges between s and a.

Action and State Gradients. For action correction, a score-based diffusion model $g_\theta(s, a)$ is trained to approximate $\nabla_a \log p_{\text{data}}(s, a)$. Using denoising score matching, we perturb state-action pairs with Gaussian noise and optimize g_θ to predict the score of perturbed data. Similarly, a score-based diffusion model $h_\varphi(s, a)$ learns $\nabla_s \log p_{\text{data}}(s, a)$ by perturbing states while keeping actions fixed. Both networks minimize losses that enforce alignment with the true score, enabling reliable gradient estimation.

Inverse Dynamics Model. To translate state gradients into actionable corrections, we introduce an inverse dynamics model $I_\phi(s, \tilde{s})$. This network predicts the action required to transition from state s to a target state $\tilde{s} = s + h_\varphi(s, a)$, which is guided by the state gradient. Trained via imitation learning on the dataset, I_ϕ ensures that state-based adjustments remain feasible within the environment dynamics.

Integration for Action Correction. During policy execution, the original action a_o from a baseline RL algorithm is adjusted using the learned gradients: $a_1 = g_\theta(s, a_o)$ directly modifies the action, while $a_2 = I_\phi(s, s + h_\varphi(s, a_o))$ incorporates state-driven guidance. The final action a combines these terms linearly with hyperparameters K_1, K_2: $a = a_o + K1 * a_1, +K2 * a_2$. This iterative correction process enhances conservatism without requiring additional critics or actors.

By unifying gradient estimation with inverse dynamics, CDSA provides a lightweight yet effective mechanism for offline RL, ensuring actions remain within dataset-supported regions.

3 EXPERIMENTS

We evaluate CDSA on D4RL benchmarks [3], including MuJoCo (Hopper, HalfCheetah, Walker2d) with datasets of varying quality (random, medium, expert) and AntMaze (umaze, medium, large) for navigation tasks. Baselines include IQL [10], POR [24], and established methods such as One-step [1], 10%BC [2], TD3+BC [4], CQL [12], and CODAC [16]. We train IQL and POR for 1M steps and integrate CDSA with their pretrained policies, fine-tuning its gradient models for 10K steps.

The results are summarized in Figure 2. CDSA consistently improves baseline performance. On MuJoCo, it achieves significant gains by stabilizing actions to avoid unsafe states. For AntMaze, CDSA boosts success rates by up to 40.5%, particularly excelling in simpler mazes where distribution alignment is critical. The method's single-step inverse dynamics inference minimizes error propagation, and its compatibility with diverse baselines (IQL/POR) highlights versatility without requiring additional training. Results validate CDSA's ability to enhance conservatism while maintaining algorithmic flexibility.

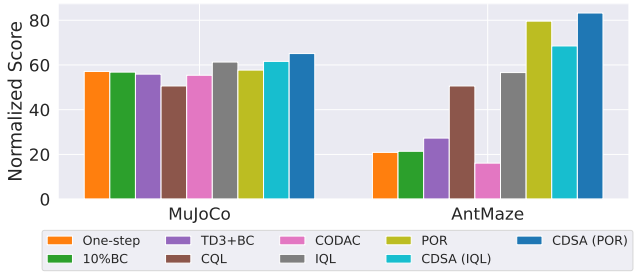

Figure 2: Average normalized scores of algorithms for MuJoCo and AntMaze. The scores are taken over the final 20 evaluations for MuJoCo and 100 evaluations for AntMaze.

4 CONCLUSION

Our work introduces the CDSA algorithm, which learns gradient fields from data and utilizes them to acquire action subcomponents. Action adjusted by these action subcomponents guide state-action pairs towards high-density regions within the dataset distribution, mitigating exposure to unfamiliar states. Since CDSA focuses solely on learning gradient fields from data, independent of RL baseline algorithms, it seamlessly integrates with various algorithms such as CQL, IQL, and POR. Our experiments in offline settings demonstrate that our method effectively navigates away from hazardous areas and makes decisions within familiar scenarios within the dataset distribution. Combining baseline algorithms with CDSA leads to improved performance on D4RL datasets across various qualities.

ACKNOWLEDGMENTS

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404. The authors also thank the anonymous reviewers for valuable comments.

REFERENCES

[1] David Brandfonbrener, Will Whitney, Rajesh Ranganath, and Joan Bruna. 2021. Offline rl without off-policy evaluation. *Advances in Neural Information Processing Systems* 34 (2021), 4933–4946.

[2] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[3] 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).

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

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

[6] Carles Gelada and Marc G Bellemare. 2019. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3647–3655.

[7] Chunming He, Chengyu Fang, Yulun Zhang, Tian Ye, Kai Li, Longxiang Tang, Zhenhua Guo, Xiu Li, and Sina Farsiu. 2023. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. *arXiv preprint arXiv:2311.11638* (2023).

[8] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. 2023. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22046–22055.

[9] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. 2024. Diffusion Models in Low-Level Vision: A Survey. *arXiv preprint arXiv:2406.11138* (2024).

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

[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] Jiafei Lyu, Aicheng Gong, Le Wan, Zongqing Lu, and Xiu Li. 2023. State Advantage Weighting for Offline RL. In *International Conference on Learning Representation tiny paper*. https://openreview.net/forum?id=PjypHLTo29v

[14] Jiafei Lyu, Xiu Li, and Zongqing Lu. 2022. Double Check Your State Before Trusting It: Confidence-Aware Bidirectional Offline Model-Based Imagination. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[15] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. 2022. Mildly Conservative Q-learning for Offline Reinforcement Learning. In *Thirty-sixth Conference on Neural Information Processing Systems*.

[16] Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. 2021. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 19235–19247.

[17] Rowan McAllister, Gregory Kahn, Jeff Clune, and Sergey Levine. 2019. Robustness to out-of-distribution inputs via task-aware generative uncertainty. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2083–2089.

[18] Zhongjian Qiao, Jiafei Lyu, Kechen Jiao, Qi Liu, and Xiu Li. 2024. SUMO: Search-Based Uncertainty Estimation for Model-Based Offline Reinforcement Learning. *CoRR* abs/2408.12970 (2024).

[19] Charles Richter and Nicholas Roy. 2017. Safe visual navigation via deep learning and novelty detection. (2017).

[20] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019).

[21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[22] Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. 2021. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371* (2021).

[23] Arash Vahdat, Karsten Kreis, and Jan Kautz. 2021. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems* 34 (2021), 11287–11302.

[24] Haoran Xu, Li Jiang, Jianxiong Li, and Xianyuan Zhan. 2022. A policy-guided imitation approach for offline reinforcement learning. *arXiv preprint arXiv:2210.08323* (2022).

[25] Junjie Zhang, Jiafei Lyu, Xiaoteng Ma, Jiangpeng Yan, Jun Yang, Le Wan, and Xiu Li. 2023. Uncertainty-driven Trajectory Truncation for Model-based Offline Reinforcement Learning. *CoRR* abs/2304.04660 (2023).