Enhancing Offline Safe Reinforcement Learning with Trajectory-Constrained Diffusion Planning

Extended Abstract

Hengrui Zhang
Beijing Jiaotong University, Beijing
Key Laboratory of Traffic Data
Mining and Embodied Intelligence
Beijing, China
18112037@bjtu.edu.cn

Youfang Lin
Beijing Jiaotong University, Beijing
Key Laboratory of Traffic Data
Mining and Embodied Intelligence
Beijing, China
yflin@bjtu.edu.cn

Shuo Shen
Interactive Entertainment Group,
Tencent
Shanghai, China
svenshen@tencent.com

Hanfeng Lin
Beijing Jiaotong University, Beijing
Key Laboratory of Traffic Data
Mining and Embodied Intelligence
Beijing, China
hanfenglin.new@gmail.com

Peng Cheng
Beijing Jiaotong University, Beijing
Key Laboratory of Traffic Data
Mining and Embodied Intelligence
Beijing, China
pcheng6@126.com

Sheng Han
Beijing Jiaotong University, Beijing
Key Laboratory of Traffic Data
Mining and Embodied Intelligence
Beijing, China
shhan@bjtu.edu.cn

Kai Lv
Beijing Jiaotong University, Beijing
Key Laboratory of Traffic Data
Mining and Embodied Intelligence
Beijing, China
lvkai@bjtu.edu.cn

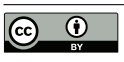
ABSTRACT

Recent approaches have utilized the RL via Supervised Learning (RvS) framework to model offline safe RL. However, these methods overlook the fundamental differences between reward maximization and constraint satisfaction, treating them identically with guidance sampling, and requiring different hyperparameters for different constraint conditions. To address these limitations, we propose a novel framework, the Trajectory-Constrained Diffusion Planner (TCDP), which reframes offline safe RL as a product of trajectory conditional probabilities and energy functions. Additionally, we introduce Cost-returns-To-Go relabeling with Data Augmentation (CTGDA) and the Quantile Normalization (QN) technique, enabling the adaptation to various constraints without retraining or extensive hyperparameter adjustments.

KEYWORDS

Offline Safe Reinforcement Learning; Diffusion Planning

ACM Reference Format:
Hengrui Zhang, Youfang Lin, Shuo Shen, Hanfeng Lin, Peng Cheng, Sheng Han, and Kai Lv. 2025. Enhancing Offline Safe Reinforcement Learning with Trajectory-Constrained Diffusion Planning: 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.



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

Safe Reinforcement Learning (RL) [3, 5, 11, 22] aims to maximize the agent's rewards while adhering to specified constraints. However, RL is inherently a trial-and-error process that necessitates continuous interaction with the environment to optimize policies [19]. Ensuring safety during these interactions is often challenging, thus highlighting the need for offline safe RL [12, 13, 20].

Recent research has explored RL via Supervised Learning (RvS) [2] to model offline safe RL. Among the approaches leveraging RvS modeling, some utilize the return-conditioned sequential modeling capability of Transformers [14, 21], incorporating Cost Returns as tokens to adapt policies to different constraints. Others employ Diffusion techniques to learn trajectory distributions [10, 17], ultimately using classifier guidance to generate trajectories that meet varying constraints.

However, these methods face the following challenges: (1) They handle reward maximization and constraint satisfaction similarly. This implies a lack of distinction between reward maximization and constraint satisfaction, hindering fine-grained adjustments. (2) Different constraint conditions necessitate varying hyperparameter adjustments, requiring knowledge of the maximum achievable return under different constraints or adjusting the degree of reward and constraint during guidance sampling.

To address these issues, we transform offline safe RL into a new target distribution: the product of trajectory conditional probabilities and energy functions. This transformation allows us to handle

Table 1: Difference between methods

Method	Objective function	Backbone	Guidance
TCDP	$q(\tau\|c \le b)\exp(\beta R(\tau))$	Diffusion	classifier guidance for reward maximization classifier-free guidance for safety constraint satisfaction
TREBI	$q(\tau)\exp(\beta(R(\tau)-nC(\tau)))$	Diffusion	classifier guidance
CDD	$q(\tau\|c=b, r=r_{b,\max})$	Diffusion	classifier-free guidance
CDT	$q(\tau\|c=b, r=r_{b,\max})$	Decision Transformer	/

constraint satisfaction and reward maximization differently. We then employ diffusion techniques to learn and generate this trajectory distribution, naming our method the Trajectory-Constrained Diffusion Planner (TCDP).

2 METHODS

Similar to the probabilistic inference model and notation outlined in [8], we utilize $O = 1$ to denote that a trajectory is optimal, and $O_c = 1$ to indicate that a trajectory satisfies the constraints, specifically $C(\tau) \le b$. According to [8], we have $p(O = 1|\tau) \propto \exp(\beta R(\tau))$. Since whether a trajectory meets constraints does not affect its optimality, we obtain $p(O = 1|\tau, O_c = 1) \propto \exp(\beta R(\tau))$. For our objective $p(\tau|O, O_c)$, we derive:

$$\begin{aligned} p(\tau|O, O_c) &= \frac{p(\tau, O|O_c)p(O_c)}{p(O|O_c)p(O_c)} = \frac{p(O|\tau, O_c)p(\tau|O_c)}{p(O|O_c)} \\ &\propto p(O|\tau,)p(\tau|O_c) \propto \exp(\beta R(\tau))p(\tau|O_c). \end{aligned} \tag{1}$$

Given that the posterior of interest is intractable, an auxiliary trajectory distribution $q(\tau)$ is introduced to serve as an approximation. We then utilize the Kullback-Leibler divergence $D_{KL}(q(\tau)||p(\tau|O, O_c))$ to derive the expression for $q(\tau)$.

In pursuit of the optimal trajectory distribution $q(\tau)$, we employ the Diffusion Planning technique as introduced by [1, 6]. Specifically, we denote $(s_t, s_{t+1}, \cdots, s_{t+H-1})_k$ as the trajectory τ_k, where H represents the length of the trajectory, k denotes the timestep in the diffusion forward process, and t represents the planning timestep. We define $q_k(\tau_k)$ and $p_k(\tau_k)$ as the marginal distribution of the forward diffusion process at time k like [15, 23]. Then, we propose the following assumption that $q_{k0}(\tau_k|\tau_0)$ and $p_{k0}(\tau_k|\tau_0)$ share the same noise transition distribution. To learn and generate the final distribution $q_k(\tau_k)$, we use denoising score matching [18] to train a conditional score network $z_\theta(\tau_k, k)$ to approximate $\nabla\tau_k \log q_k(\tau_k)$ as follows:

$$\min_\theta \int q_k(\tau_k)\left[\|z_\theta(\tau_k, k) + \sigma_k \nabla_{x_k} \log q_k(\tau_k)\|^2\right] \mathrm{d}\tau_k, \tag{2}$$

In practical implementations, we aim for our learned policies to adapt to different constraint thresholds b without the need for retraining. This requires replacing O_c with a general condition c. We introduce Cost-returns-To-Go relabeling with Data Augmentation (CTGDA) and the Quantile Normalization (QN) technique, enabling the adaptation to various constraints without retraining or extensive hyperparameter adjustments. CTGDA enhances the data to generate trajectories that meet different constraints without being affected by noisy data. QN normalizes the trajectory rewards under various constraints using quantile normalization to maximize rewards within the current constraint without being influenced by data from other constraints. We use DiT [16] as the backbone and employ classifier-free guidance to sample trajectories that meet the constraints. The noisy trajectory τ_k and additional conditional information (noise timesteps k and cost returns c) are fed into the DiT Block with adaLN-Zero.

In comparison to previous algorithms, our approach uniquely integrates both classifier guidance and classifier-free guidance concepts, whereas other methods have adopted a singular guidance mechanism. This dual guidance facilitates more nuanced planning by considering the distinct characteristics of reward maximization and safety constraint satisfaction. Table 1 summarizes the differences between our method and others.

3 EXPERIMENTS

Our algorithm was evaluated using the DSRL benchmark [13], which encompasses three task sets: Safety-Gymnasium [7], Bullet-Safety-Gym [4] and MetaDrive [9].

In our results, TCDP outperforms other baselines by maximizing rewards while satisfying constraints in the majority of tasks, with a particularly notable improvement in the MetaDrive task. BC-Safe consistently ensures constraint satisfaction across most tasks by training exclusively on safe trajectories. However, due to the presence of suboptimal data in the dataset, BC-Safe fails to achieve competitive rewards compared to our algorithm. FISOR, being a hard constraint algorithm, offers superior safety compared to other soft constraint algorithms. Nevertheless, its overly conservative policy results in lower rewards. BCQ-Lag, CPQ, and COptiDICE struggle to balance rewards and constraints effectively, leading to subpar results. CDT and TREBI, like our algorithm, belong to the RVS approach and support training under multiple constraint conditions. However, CDT requires tuning the maximum return values for each constraint, yet still violates constraints in simpler tasks (BallRun, BallCircle). TREBI's reliance on training a classifier for noisy data leads to inaccurate classifier guidance, resulting in poor performance in many tasks.

Our algorithm does not require setting different parameters for different constraints. We only need to determine a single parameter β, which remains unchanged across different constraints. In contrast, CDD and CDT require setting $r_{b,\max}$, necessitating extensive hyperparameter tuning that often fails to yield optimal results. TREBI, besides sharing the parameter β with our algorithm, also introduces n, which demands substantial hyperparameter adjustments, resulting in suboptimal performance under most constraints.

REFERENCES

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. 2022. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657* (2022).

[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] Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16, 1 (2015), 1437–1480.

[4] Sven Gronauer. 2022. BULLET-SAFETY-GYM: AFRAMEWORK FOR CONSTRAINED REINFORCEMENT LEARNING. (2022).

[5] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. 2022. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *arXiv preprint arXiv:2205.10330* (2022).

[6] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning*.

[7] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. 2023. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems* 36 (2023).

[8] Sergey Levine. 2018. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909* (2018).

[9] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence* 45, 3 (2022), 3461–3475.

[10] Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. 2023. Safe offline reinforcement learning with real-time budget constraints. In *International Conference on Machine Learning*. PMLR, 21127–21152.

[11] Puze Liu, Haitham Bou-Ammar, Jan Peters, and Davide Tateo. 2024. Safe Reinforcement Learning on the Constraint Manifold: Theory and Applications. *arXiv preprint arXiv:2404.09080* (2024).

[12] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. 2022. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*. PMLR, 13644–13668.

[13] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. 2023. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303* (2023).

[14] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. 2023. Constrained decision transformer for offline safe reinforcement learning. In *International Conference on Machine Learning*. PMLR, 21611–21630.

[15] Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. 2023. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In *International Conference on Machine Learning*. PMLR, 22825–22855.

[16] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

[17] Ralf Römer, Lukas Brunke, Martin Schuck, and Angela P Schoellig. [n.d.]. Safe Offline Reinforcement Learning using Trajectory-Level Diffusion Models. In *ICRA 2024 Workshop {\textemdash} Back to the Future: Robot Learning Going Probabilistic*.

[18] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

[19] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[20] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. 2022. Constraints penalized q-learning for safe offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8753–8760.

[21] Qin Zhang, Linrui Zhang, Haoran Xu, Li Shen, Bowen Wang, Yongzhe Chang, Xueqian Wang, Bo Yuan, and Dacheng Tao. 2023. Saformer: A conditional sequence modeling approach to offline safe reinforcement learning. *arXiv preprint arXiv:2301.12203* (2023).

[22] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. 2023. State-wise safe reinforcement learning: A survey. *arXiv preprint arXiv:2302.03122* (2023).

[23] Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. 2024. Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model. *arXiv preprint arXiv:2401.10700* (2024).