Offline RL with Smooth OOD Generalization in Convex Hull and its Neighborhood

Qingmao Yao, Zhichao Lei, Tianyuan Chen, Ziyue Yuan,
Xuefan Chen, Jianxiang Liu, Faguo Wu, Xiao Zhang

ICLR 2025

北京航空航天大學
BEIHANG UNIVERSITY

Outline

Introduction

Methods

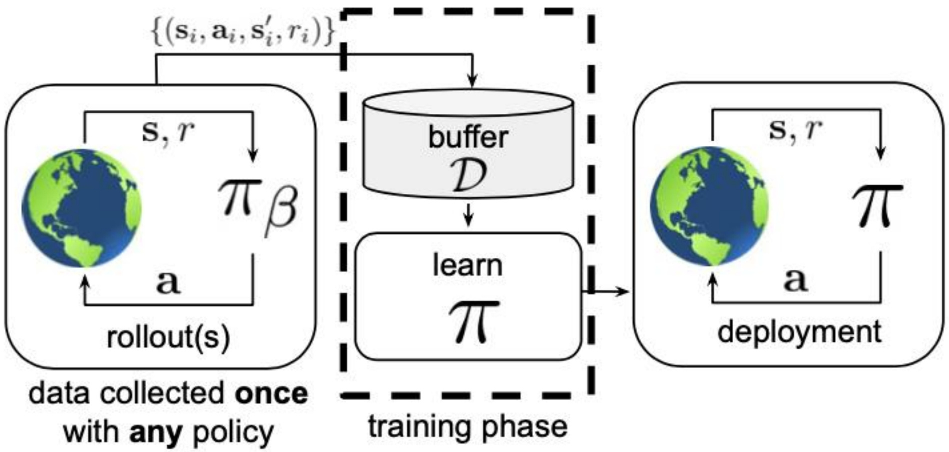
Experiments

Introduction

Background: Offline Reinforcement Learning

- Offline Reinforcement Learning (RL) learns the optimal policy solely from offline datasets $D = \{(s_i, a_i, r_i, s_i', d_i)\}_{i=1}^{N},\ d_i \in \{0, 1\}$

- Key challenges
  - The distribution shift between behavior policy μ (dataset policy) and learned policy π
  - The overestimation issue of out-of-distribution (OOD) actions, leading to suboptimal policy

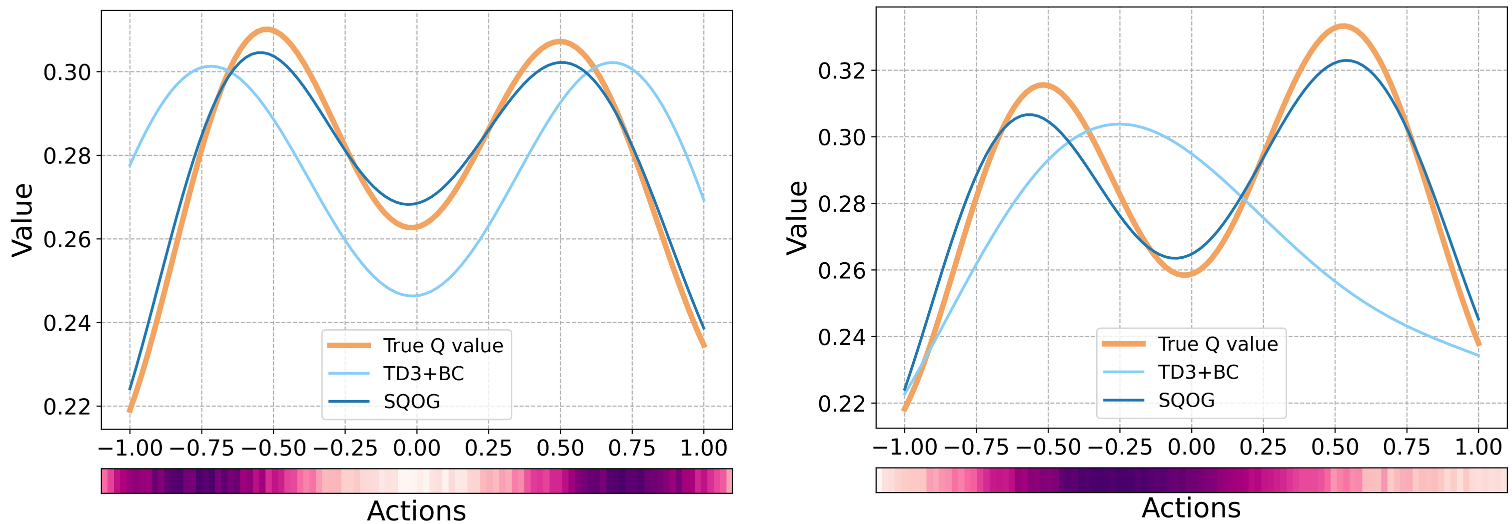


Introduction

Background: Over-constraint issue in Offline RL

- Recent solutions are too conservative, introducing an over-constraint issue.

- Over-constraint issue (on Q-value)



- TD3+BC: the learned policy is overly close to the behavior policy
- SQOG (our method): alleviates the over-constraint issue

- Goal: improve Q-value estimation by enhancing Q-function generalization in dataset OOD regions.

Scott Fujimoto and Shixiang (Shane) Gu. A minimalist approach to offline reinforcement learning. NeurIPS, 2021.

Contributions

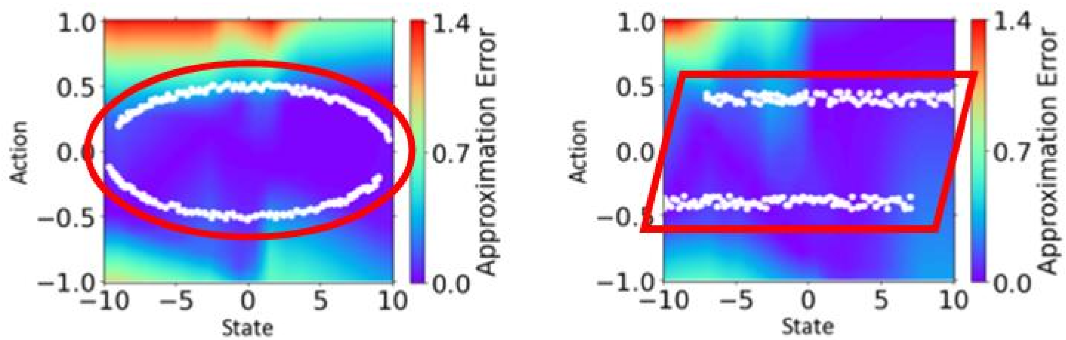
- ✓ Under the safety guarantees of the Convex Hull and its Neighborhood (CHN), we propose the Smooth Bellman Operator (SBO), which enhances Q-function generalization in OOD regions and approximates the true Q-values.
- ✓ Building on SBO, we design an effective algorithm, SQOG, which alleviates the over-constraint issue and obtains SOTA results on D4RL benchmarks.

Methods

Safety guarantees of the generalization within CHN

- **Safety guarantee 1: Q-value difference is controlled within CHN**

  - Previous work (DOGE, 2023) demonstrated that Q-value difference is controlled within the convex hull, we extend this result to the CHN.



- **Safety guarantee 2: Q-function is uniformly continuous within CHN**

These two guarantees ensure safer and more reliable Q-function generalization in OOD regions within CHN!

Jianxiong Li, Xianyuan Zhan, Haoran Xu, Xiangyu Zhu, Jingjing Liu, and Ya-Qin Zhang. When data geometry meets deep function: Generalizing offline reinforcement learning. ICLR, 2023.

Methods
Smooth Bellman Operator

- Definition of SBO

$$\widetilde{\mathcal{B}}^{\pi} Q(s,a) = (\mathcal{G}_1 \hat{\mathcal{B}}_2^{\pi}) Q(s,a)$$

- Base Bellman operator

$$\hat{\mathcal{B}}_2^{\pi} Q(s,a) = \begin{cases} \boxed{\hat{\mathcal{B}}^{\pi} Q(s,a),} & \hat{\mu}(a|s) > 0 \\ Q(s,a), & \hat{\mu}(a|s) = 0 \; and \; (s,a) \in CHN \end{cases}$$

empirical Bellman operator (CQL, 2020)

- Smooth generalization operator

$$\mathcal{G}_1 Q(s,a) = \begin{cases} Q(s,a), & \hat{\mu}(a|s) > 0 \\ Q(s, \boxed{a_{neighbor}^{in}}), & \boxed{\hat{\mu}(a|s) = 0 \; and \; (s,a) \in CHN} \end{cases}$$

in-sample neighbor action of the OOD action *a*

OOD action within CHN

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. NeurIPS, 2020.

Methods
Theoretical justification for SBO

- Why $\hat{Q}^{\pi}_{\theta}(s, a^{in}_{neighbor})$ is an appropriate OOD target ?

- Goal: let $\hat{Q}^{\pi}_{\theta}(s, a^{ood})$ (the output of value network) approximate the true OOD Q-value $Q^{\pi}(s, a^{ood})$

- Proposition 3: if $\hat{Q}^{\pi}_{\theta}(s, a^{in}) \approx Q^{\pi}(s, a^{in})$, then

$$\|Q^{\pi}(s, a^{ood}) - \hat{Q}^{\pi}_{\theta}(s, a^{in}_{neighbor})\| < \varepsilon$$

- Theorem 1 shows $\hat{Q}^{\pi}_{\theta}(s, a^{in}) \approx Q^{\pi}(s, a^{in})$, when the KL-divergence of learned policy and behavior policy is bound

  - $\|\hat{\mathcal{B}}^{\pi} Q_{\theta} - \mathcal{B}^{\pi} Q_{\theta}\|$ is bound, for all $(s, a) \in \mathcal{D}$
  - $\hat{Q}^{\pi}_{\theta}(s, a^{in})$ closely approximates $Q^{\pi}_{\theta}(s, a^{in})$
  - $Q^{\pi}_{\theta}(s, a^{in})$ is close to true in-sample Q-value $Q^{\pi}(s, a^{in})$ (PRDC, 2023)
  - $\hat{Q}^{\pi}_{\theta}(s, a^{in}) \approx Q^{\pi}(s, a^{in})$

Yuhang Ran, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, and Yang Yu. Policy regularization with dataset constraint for offline reinforcement learning. ICML, 2023.

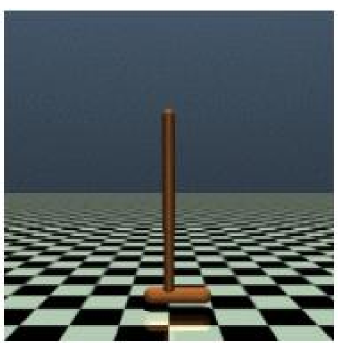
Methods
Effects of the SBO

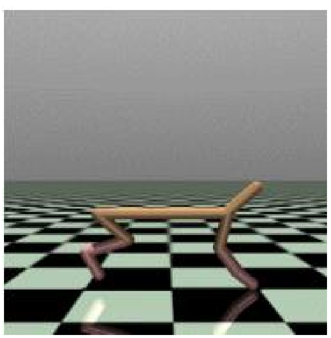
- SBO achieves better Q-value estimation
  - For in-sample evaluation, SBO introduces negligible changes to the empirical Bellman operator. (Theorem 2)
  - For OOD evaluation, SBO helps mitigate underestimation and overestimation. (Theorem 3)
- Based on the SBO, we develop the algorithm SQOG.

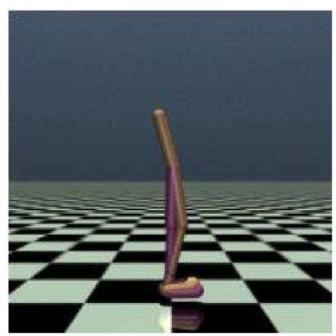
D4RL benchmarks

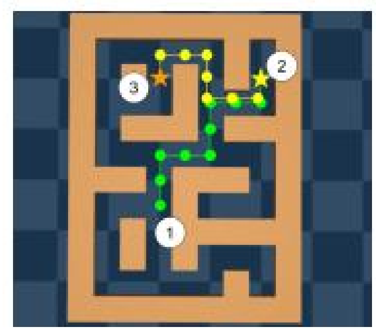
- D4RL is the **most widely** used benchmarks in offline RL.
- **Gym-Mujoco** are the **lomocotion** tasks (Hopper, Halfcheetah, Walker2d)
- **Maze2D** is a **navigation** task requiring a 2D agent to reach a fixed goal location.
- **Adroit** involves **controlling Hand robot** tasked with hammering a nail, opening a door, twirling a pen, or picking up and moving a ball.
- Performance Evaluation: **Normalized score** (100 → expert, 0 → random).

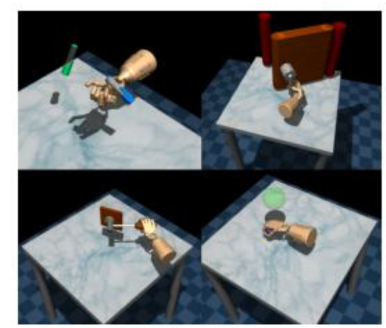
$$\text{normalized score} = 100 * \frac{\text{score} - \text{random score}}{\text{expert score} - \text{random score}}$$


Hopper


Halfcheetach


Walker2d


Maze2d


Adroit

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021.

Experiments
Results on D4RL Benchmarks

- SQOG obtains SOTA results on benchmark datasets

Dataset	BC	TD3+BC	CQL	IQL	DOGE	MCQ	SQOG
halfcheetah-r	2.2±0.0	11.0±1.1	17.5±1.5	13.1±1.3	17.8±1.2	23.6±0.8	**25.6±0.4**
hopper-r	3.7±0.6	8.5±0.6	7.9±0.4	7.9±0.2	21.1±12.6	**31.0±1.7**	15.6±3.3
walker2d-r	1.3±0.1	1.6±1.7	5.1±1.3	5.4±1.2	0.9±2.4	10.3±6.8	**17.7±3.5**
halfcheetah-m	43.2±0.6	48.3±0.3	47.0±0.5	47.4±0.2	45.3±0.6	58.3±1.3	**59.2±2.4**
hopper-m	54.1±3.8	59.3±4.2	53.0±28.5	66.2±5.7	98.6±2.1	73.6±10.3	**100.6±0.7**
walker2d-m	70.9±11.0	83.7±2.1	73.3±17.7	78.3±8.7	86.8±0.8	**88.4±1.3**	82.9±0.8
halfcheetah-m-r	37.6±2.1	44.6±0.5	45.5±0.7	44.2±1.2	42.8±0.6	**51.5±0.2**	46.4±1.2
hopper-m-r	16.6±4.8	60.9±18.8	88.7±12.9	94.7±8.6	76.2±17.7	99.5±1.7	**100.9±5.1**
walker2d-m-r	20.3±9.8	81.8±5.5	81.8±2.7	73.8±7.1	87.3±2.3	83.3±1.9	**88.3±3.5**
halfcheetah-m-e	44.0±1.6	90.7±4.3	75.6±25.7	86.7±5.3	78.7±8.4	85.4±3.4	**92.6±0.4**
hopper-m-e	53.9±4.7	98.0±9.4	105.6±12.9	91.5±14.3	102.7±5.2	106.1±2.3	**109.2±2.8**
walker2d-m-e	90.1±13.2	110.1±0.5	107.9±1.6	109.6±1.0	**110.4±1.5**	110.3±0.1	109.0±0.3
Mujoco Average	36.5	58.2	61.8	59.9	64.1	68.4	**70.7**
Maze2d Average	-2.0	35.0	19.6	37.2	-	102.2	**124.7**
Adroit Total	93.9	0.0	93.6	110.7	-	123.3	**149.6**
Runtime (h)	0.3	0.4	10.8	0.4	0.9	8.0	0.4

- SQOG consistently attains the highest scores on most datasets (8/12) and achieves the highest average scores (bold) across the Mujoco, Maze2d, and Adroit tasks, with low computational cost.

Experiments
Sanity Check: alleviation of the over-constraint issue

SQOG alleviates the over-constraint issue

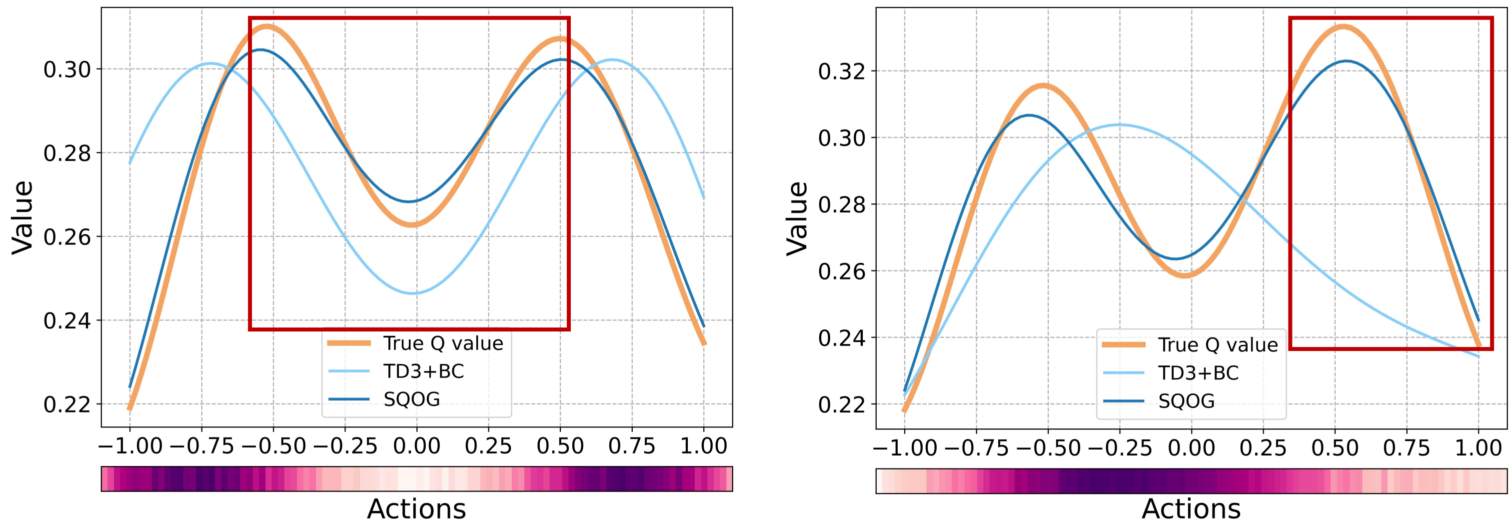


Fig. Q values estimation of actions in two key states.

- The highest true value exists in [-0.50, 0.50] (left), which corresponds to OOD regions within the convex hull.
- The highest true value exists in [0.30, 1.00] (right), corresponding to OOD regions in the neighborhood of the convex hull.

Experiments
Sanity Check: alleviation of the over-constraint issue

SQOG alleviates the over-constraint issue

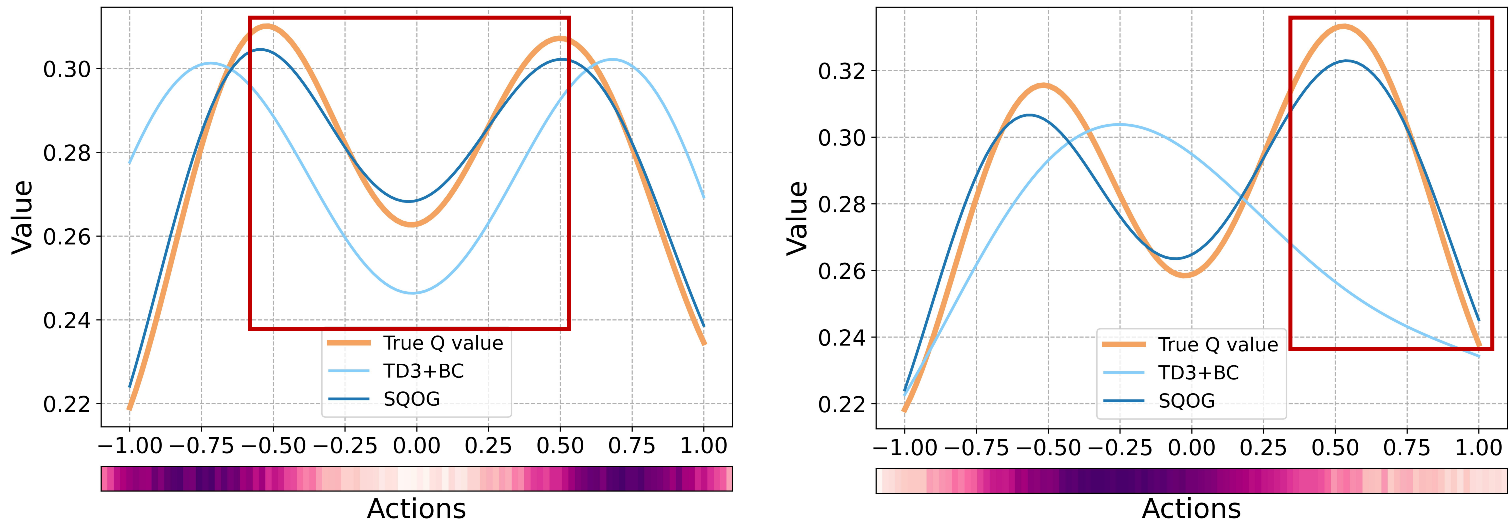


Fig. Q values estimation of actions in two key states.

- TD3+BC encounters the over-constraint issue in these OOD regions, failing to leverage implicit OOD information within CHN.
- SQOG accurately estimates Q-values through smooth OOD generalization within the CHN (convex hull and its neighborhood).

Experiments
Generalizability of SBO

- SBO is a versatile plug-in for policy constraint methods.

Dataset	BRAC	BRAC+SBO
halfcheetah-medium	49.8±1.2	**54.3±1.2**
hopper-medium	3.6±3.1	**90.9±2.9**
walker2d-medium	7.8±8.1	**85.6±4.3**
halfcheetah-medium-replay	41.8±6.2	**47.8±2.0**
hopper-medium-replay	28.8±20.3	**61.1±11.9**
walker2d-medium-replay	8.5±3.0	**67.6±11.0**
Mujoco Average	23.4	**67.9**
Improvement	-	**190.2%**
pen-human	19.2±16.3	**69.7±8.7**
pen-cloned	28.4±23.4	**69.0±14.8**
Adroit Average	23.8	**69.4**
Improvement	-	**191.6%**

- A significant performance improvement when SBO is added to BRAC.
- SBO serves as a valuable complement to policy constraint methods.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *CoRR*, abs/1911.11361, 2019.

Summary and Takeaways

- We present a method that broadly alleviates the over-constraint issue in policy constraint methods, achieving SOTA performance with low computational cost.
  - Better Q-value estimation leads to better policy performance.
  - Neighboring in-sample Q-values serve as appropriate targets for over-constrained OOD Q-values.

Thank you！