NEURAL INFORMATION PROCESSING SYSTEMS

Q-Distribution guided Q-learning for offline reinforcement learning: Uncertainty penalized Q-value via consistency model

Jing Zhang[1], Linjiajie Fang[1], Kexin Shi[1],Wenjia Wang[2,1], Bing-Yi Jing[3]

[1]Hong Kong University of Science and Technology, Hong Kong, China

[2]Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

[3]Southern University of Science and Technology, Shenzhen, China

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

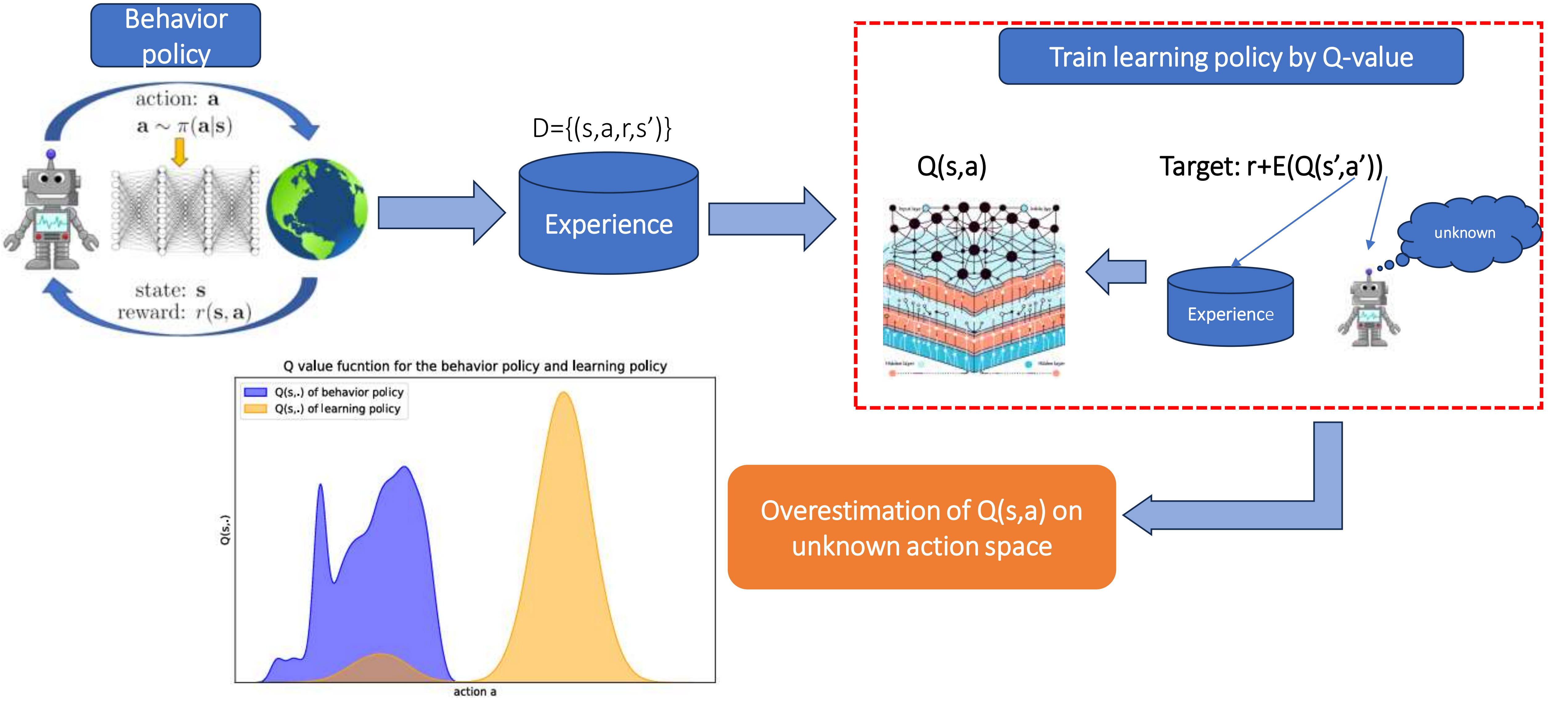
香港科技大學(廣州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY 南方科技大学

Content

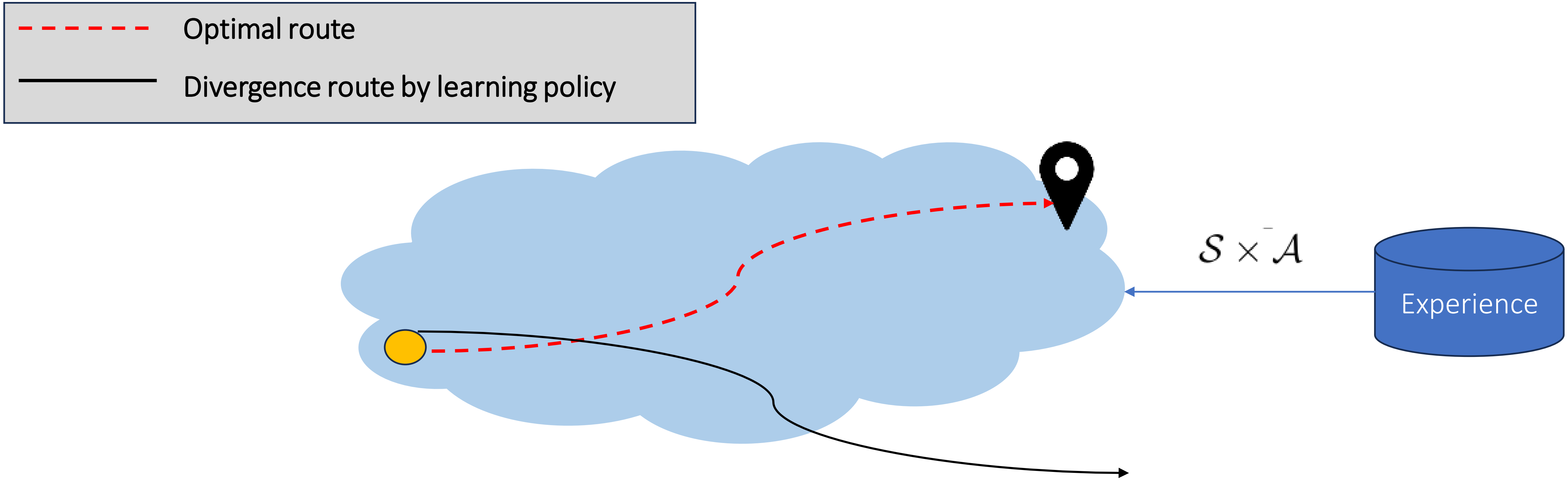
- Problem
- Q-Distribution guided Q-learning(QDQ)
- Theoretical Analysis
- Experiments

Overestimation of Q-value in offline RL



Divergence caused by Q-value overestimation

- Overestimated Q function assume the unknow action correspond to high-reward actions. Then the learning policy prioritizes these risky actions. The accumulate of bootstrap error will lead to a failure.

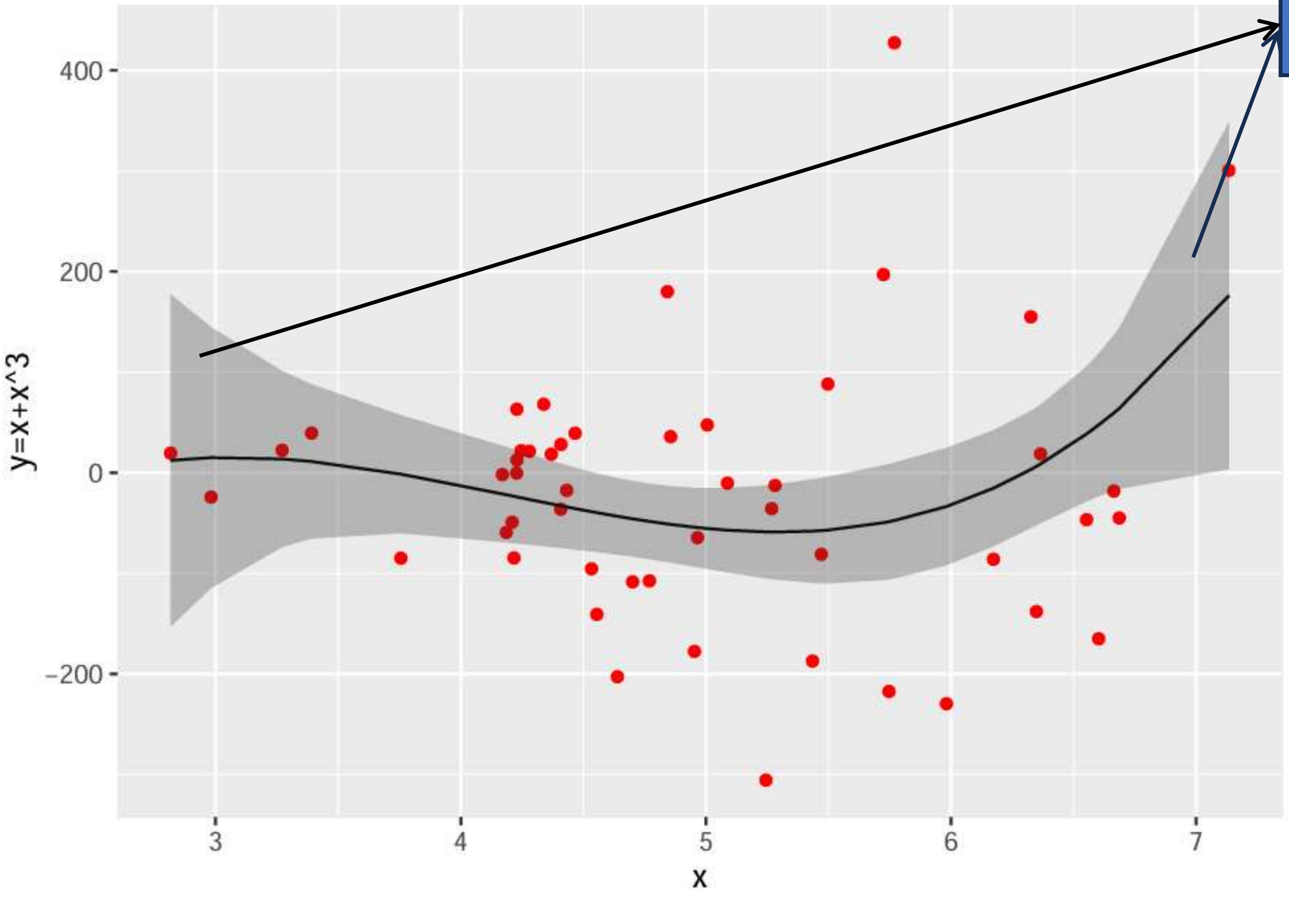


Content

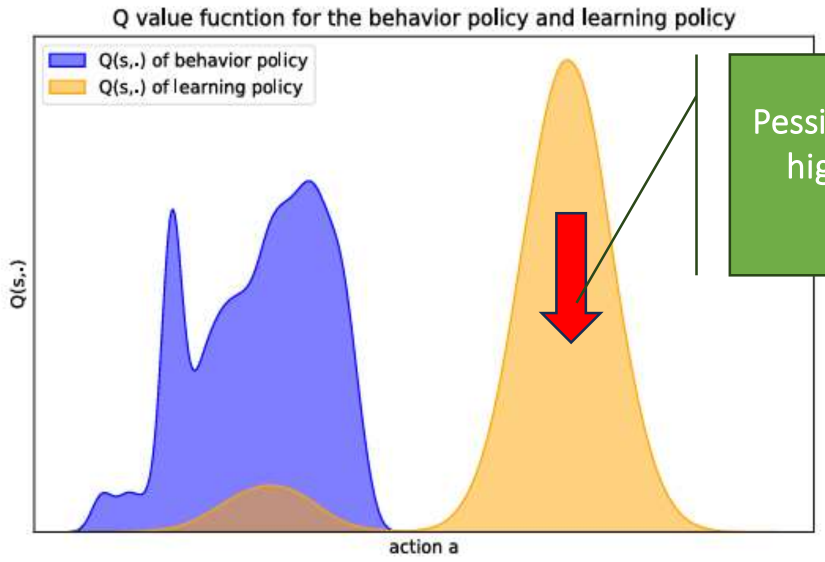
- Problem
- Q-Distribution guided Q-learning(QDQ)
- Theoretical Analysis
- Experiments

Overestimation on OOD region=high uncertainty of estimation

- Pessimistic Q-value function by the Q-value uncertainty



High uncertainty when using out of distribution point

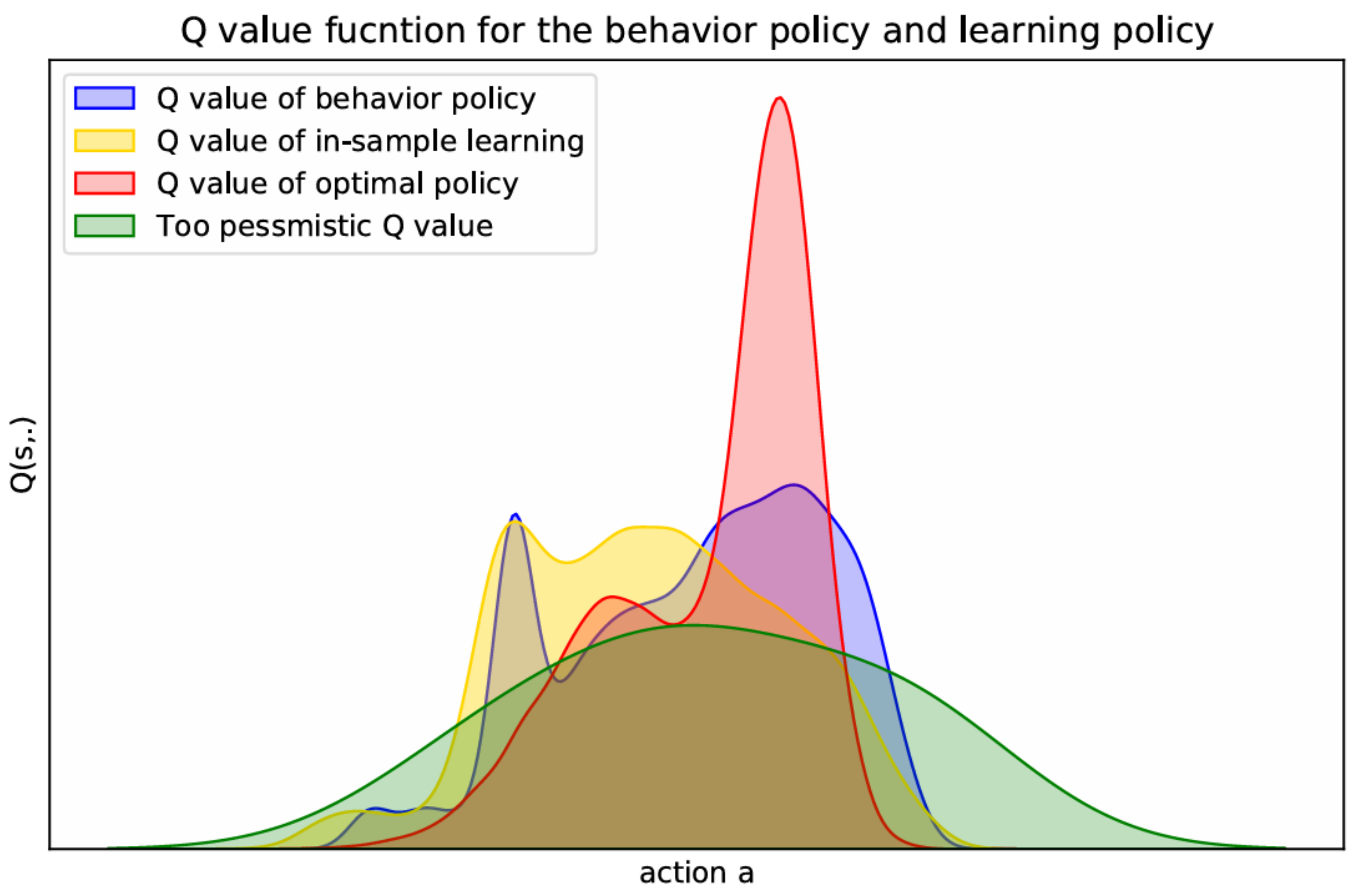


Pessimistic Q-value with high uncertainty on OOD region

QDQ: estimate Q-value uncertainty directly

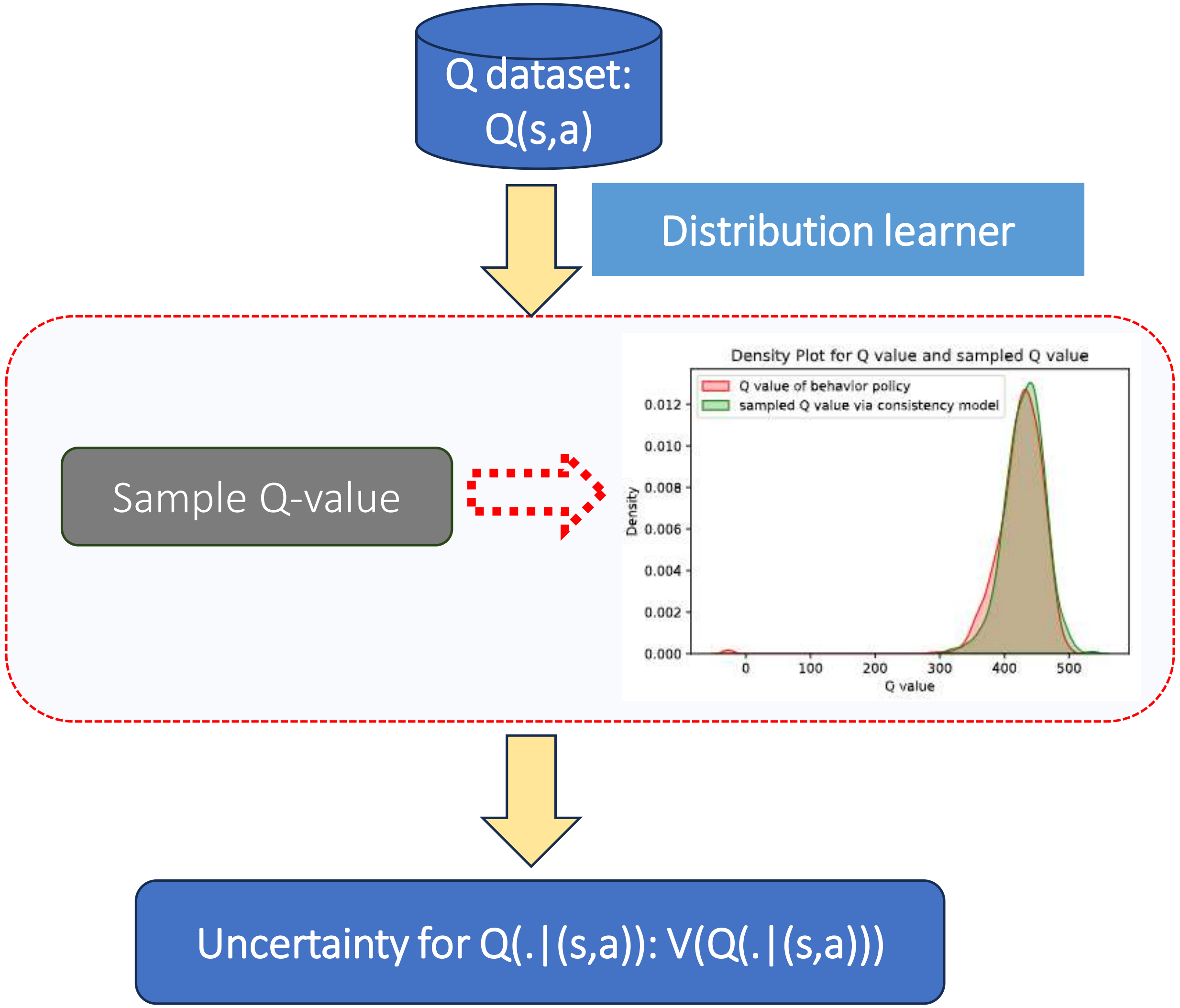
Challenge of estimate Q-value uncertainty

- Being too conservative in Q-value estimation
- Fail to approach a tight lower confidence bound
- Mimic the Q-value of the behavior policy



QDQ: estimate Q-value uncertainty directly

- Behavior policy and learning policy share the same uncertainty set over actions.
- The core concept revolves around learning the distribution of the Q-value of the behavior policy and quantifying uncertainty by bootstrap samples.



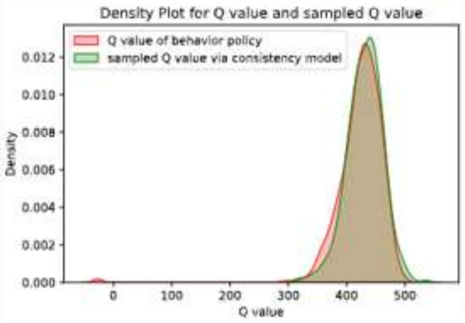
QDQ: estimate Q-value uncertainty directly

QDQ: Trajectory-level truncated Q-value

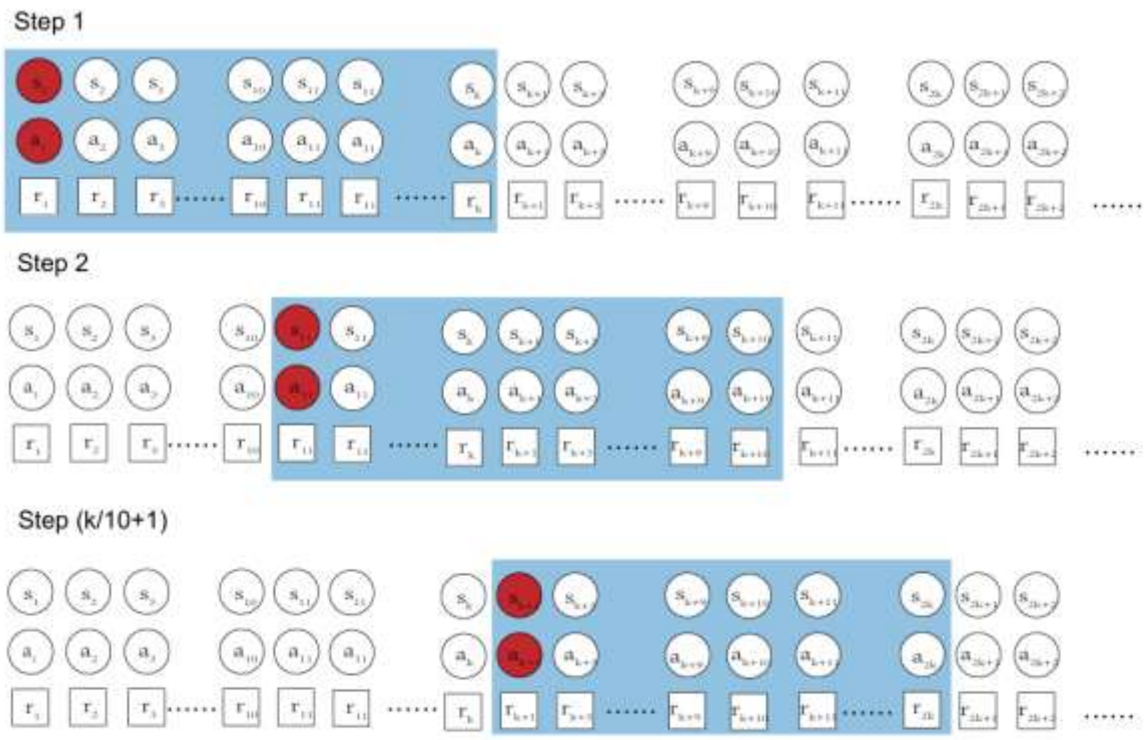
Q dataset: Q(s,a)

Distribution learner

Sample Q-value



Uncertainty for Q(.|(s,a)): V(Q(.|(s,a)))



$$Q^{\pi_\beta}_{\mathcal{T}}(s_i,a_i)=\sum_{m=i}^{\mathcal{T}}\gamma^{m-1}r(s_m,a_m)\times t(s_m,a_m), t(s_m,a_m)=\begin{cases}0, & terminal,\\ 1, & otherwise.\end{cases}$$

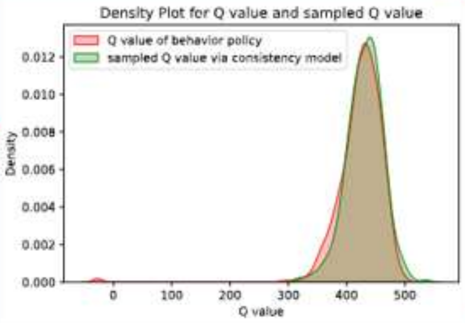
QDQ: estimate Q-value uncertainty directly

QDQ: Learn the distribution of Q-value

Q dataset: Q(s,a)

Distribution learner

Sample Q-value



Uncertainty for Q(.|(s,a)): V(Q(.|(s,a)))

Consistency model

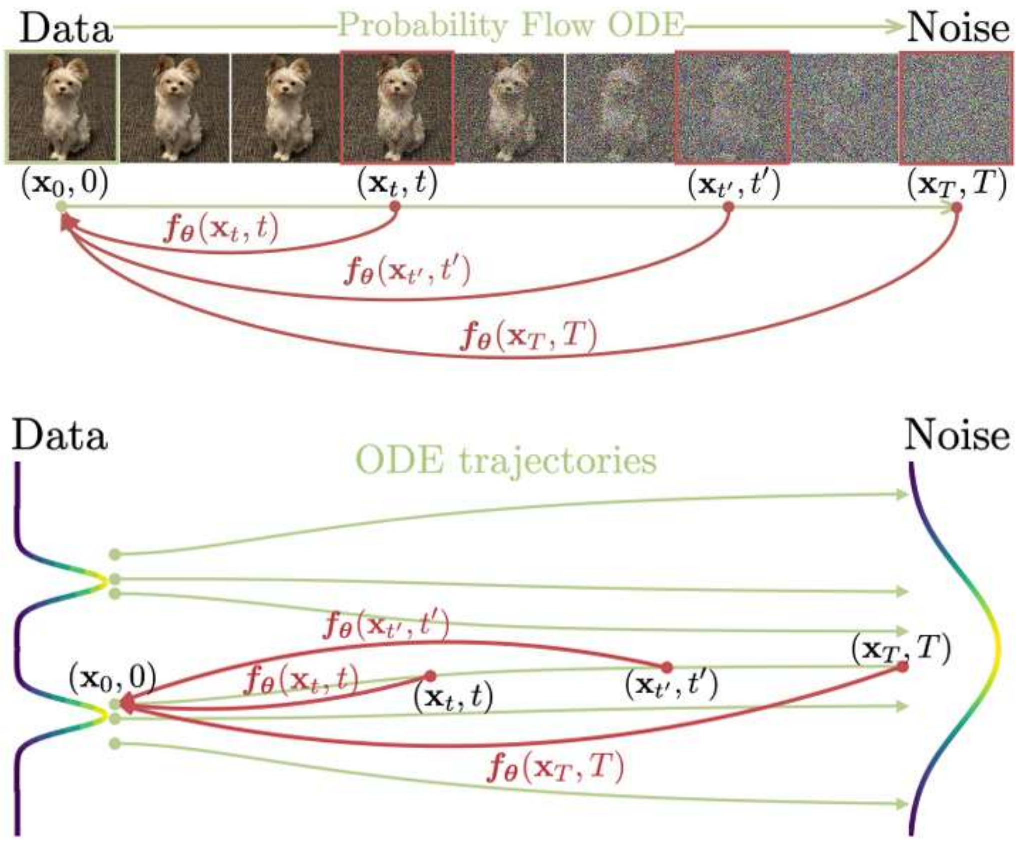


Figure 2: Consistency models are trained to map points on any trajectory of the PF ODE to the trajectory's origin.

Source: The right figure is from Figure 2 of the paper consistency model(https://arxiv.org/abs/2303.01469)

QDQ: estimate Q-value uncertainty directly

QDQ: RL paradigm

Recover Q-value function: uncertainty-aware learning objective.

$$Q_L(s',a') = \frac{1}{\mathcal{H}_Q(a'|s')} Q(s',a') 1_{(a' \in \mathcal{U}(Q))} + \beta Q(s',a') 1_{(a' \notin \mathcal{U}(Q))}.$$

$$\mathcal{L}_{adv}(Q) = \min_{\theta} \{\alpha \mathcal{L}(Q)_H + (1-\alpha)\mathcal{L}(Q)_L\}.$$

$$\mathcal{H}_Q(a'|s') = \sqrt{V(X_\epsilon|(\boldsymbol{s'},\boldsymbol{a'}))}$$

Improve the learning policy.

$$\mathcal{L}_\phi(\pi) = \max_{\phi} \left[\mathbb{E}_{s\sim\mathbb{P}_\mathcal{D}(s), a\sim\pi_\phi(\cdot|s)}[Q_\theta(s,a)] + \gamma \mathbb{E}_{a\sim\mathcal{D}}[\log \pi_\phi(a)]\right].$$

Content

- Problem
- Q-Distribution guided Q-learning(QDQ)
- Theoretical Analysis
- Experiments
- Conclusion and Future Work

QDQ: Theoretical Analysis

- Convergence of learned Q-value distribution

Theorem 4.1 (Informal). *Under some mildly condition, the truncated Q-value $Q_{\mathcal{T}}^{\pi_\beta}$ converge in-distribution to the true true Q-value Q^{π_β}.*

$$F_{Q_{\mathcal{T}}^{\pi_\beta}}(x) \to F_{Q^{\pi_\beta}}(x), \mathcal{T} \to +\infty. \tag{10}$$

- Consistency model is suitable for estimating uncertainty

Theorem 4.2 (Informal). *Following the assumptions as in [20], $f_\theta(x, T|(s, a))$ is L-Lipschitz. We also assume the truncated Q-value is bounded by $\mathcal{H}$. The action a broadly influences $V(X_\epsilon|(s, a))$ by:* $|\frac{\partial var(X_\epsilon)}{\partial a}| = O(L^2 T\sqrt{\log n})\mathbf{1}$.

QDQ: Theoretical Analysis

- Convergence of QDQ algorithm

Theorem 4.3 (Informal). *The Q-value function of QDQ can converge to a fixed point of the Bellman equation:* $Q(s,a) = \mathscr{F}Q(s,a)$, *where the Bellman operator* $\mathscr{F}Q(s,a)$ *is defined as:*

$$\mathscr{F}Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P_{\mathcal{D}}(s')}\{\max_{a'}[\alpha Q(s',a') + (1-\alpha)Q_L(s',a')]\}. \tag{11}$$

Theorem 4.4 (Informal). *Under mild conditions, with probability* $1-\eta$ *we have*

$$\left\|Q^{\Delta} - Q^*\right\|_{\infty} \leq \epsilon, \tag{12}$$

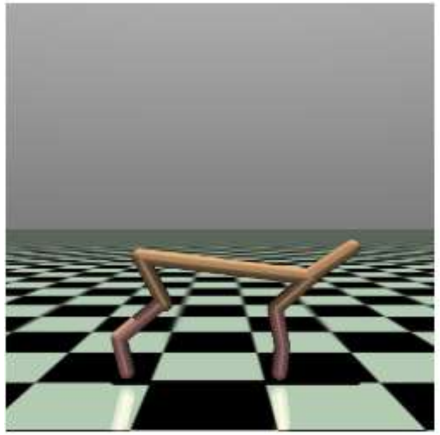
where Q^{Δ} *is learned by the uncertainty-aware loss in Eq.* $\boxed{7}$, ϵ *is error rate related to the difference between the classical Bellman operator* $\mathcal{B}Q$ *and the QDQ bellman operator* $\mathscr{F}Q$.

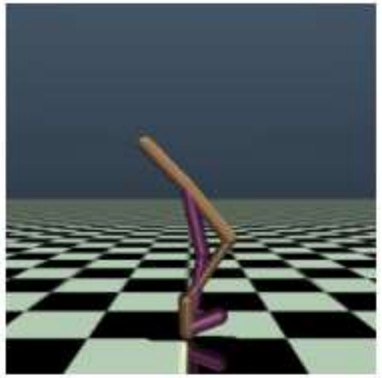
Content

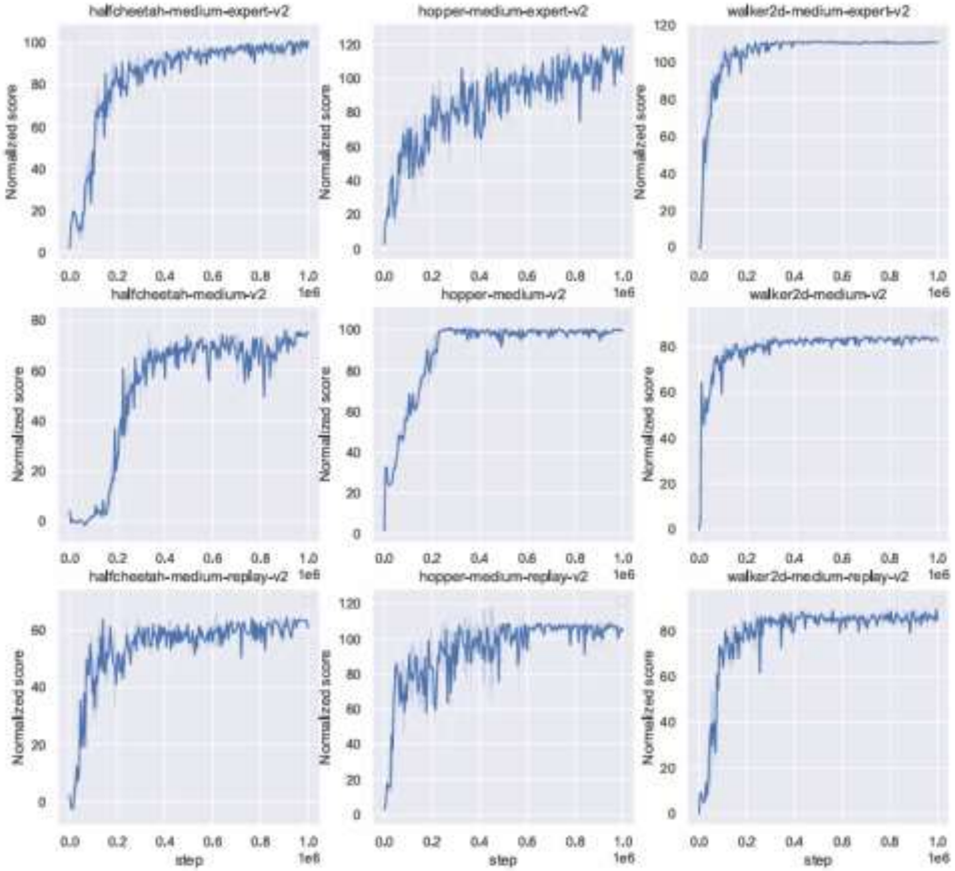
- Problem
- Q-Distribution guided Q-learning(QDQ)
- Theoretical Analysis
- Experiments

Experiments

- Gym-MuJoCo tasks.







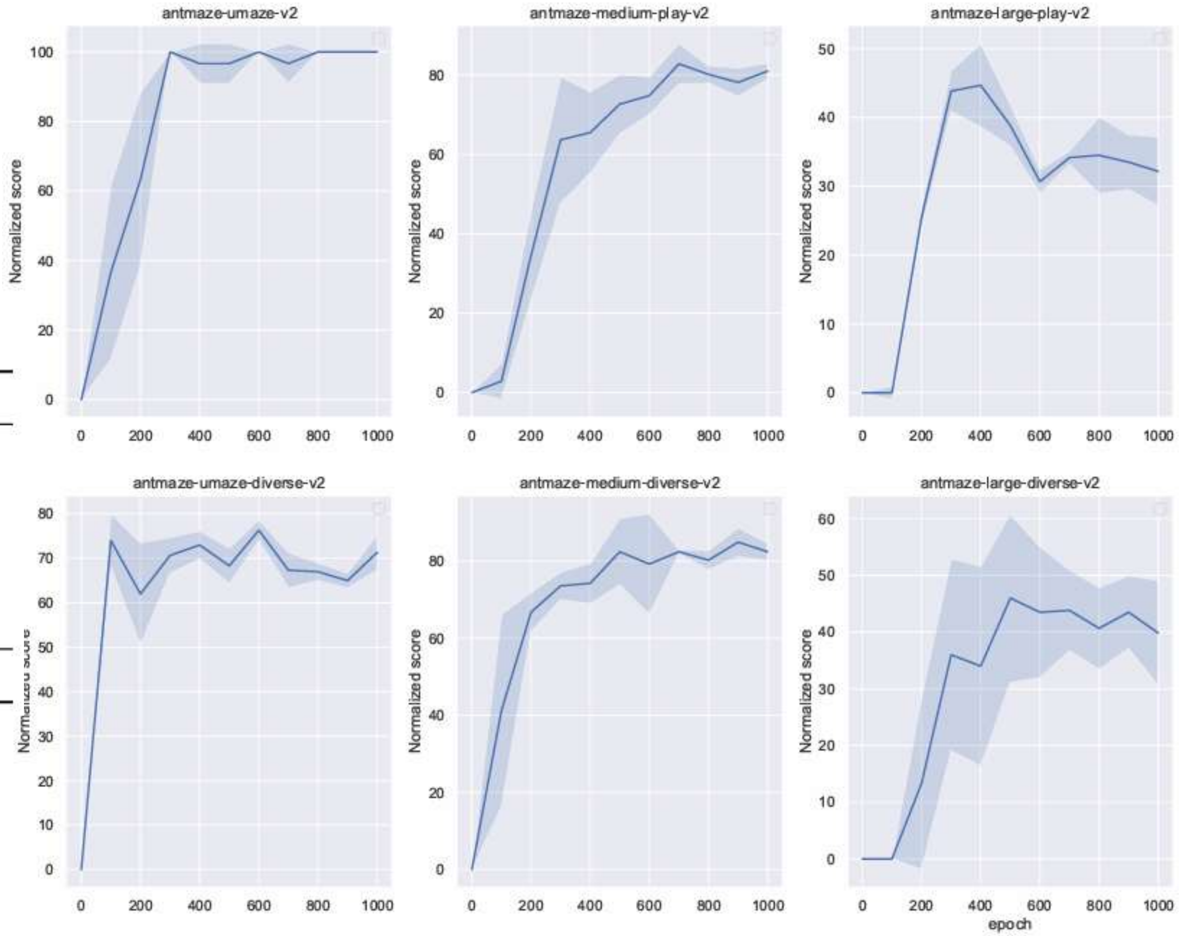
Dataset	BC	AWAC	DT	TD3+BC	CQL	IQL	UWAC	MCQ	EDAC	PBRL	QDQ(Ours)
ha-med	42.6	43.5	42.6	48.3	44.0	47.4	42.2	64.3	65.9	57.9	**74.1**±1.7
ho-med	52.9	57.0	67.6	59.3	58.5	66.2	50.9	78.4	**101.6**	75.3	**99.0**±0.3
wa-med	75.3	72.4	74.0	83.7	72.5	78.3	75.4	91.0	**92.5**	89.6	86.9±0.08
ha-med-r	36.6	40.5	36.6	44.6	45.5	44.2	35.9	56.8	61.3	45.1	**63.7**±2.9
ho-med-r	18.1	37.2	82.7	60.9	95.0	94.7	25.3	101.6	101.0	100.6	**102.4**±0.28
wa-med-r	26.0	27.0	66.6	81.8	77.2	73.8	23.6	91.3	87.1	77.7	**93.2**±1.1
ha-med-e	55.2	42.8	86.8	90.7	91.6	86.7	42.7	87.5	**106.3**	92.3	99.3±1.7
ho-med-e	52.5	55.8	107.6	98.0	105.4	91.5	44.9	112.3	110.7	110.8	**113.5**±3.5
wa-med-e	107.5	74.5	108.1	110.1	108.8	109.6	96.5	114.2	114.7	110.1	**115.9**±0.2
Total	466.7	450.7	672.6	684.6	677.4	698.5	437.4	797.4	841.1	759.4	**848.0**±11.8

Experiments

- AntMaze tasks.

AntMaze

Dataset	BC	TD3+BC	DT	Onestep RL	AWAC	CQL	IQL	QDQ(Ours)
umaze	54.6	78.6	59.2	64.3	56.7	74.0	87.5	**98.6**±2.8
umaze-diverse	45.6	71.4	53.0	60.7	49.3	**84.0**	62.2	67.8±2.5
medium-play	0.0	10.6	0.0	0.3	0.0	61.2	71.2	**81.5**±3.6
medium-diverse	0.0	3.0	0.0	0.0	0.7	53.7	70.0	**85.4**±4.2
large-play	0.0	0.2	0.0	0.0	0.0	15.8	**39.6**	35.6±5.4
large-diverse	0.0	0.0	0.0	0.0	1.0	14.9	**47.5**	31.2±4.5
Total	100.2	163.8	112.2	125.3	142.4	229.8	378	**400.1**±23.0



Thank you for your attention!