Toward Adaptive Conservatism in Offline RL

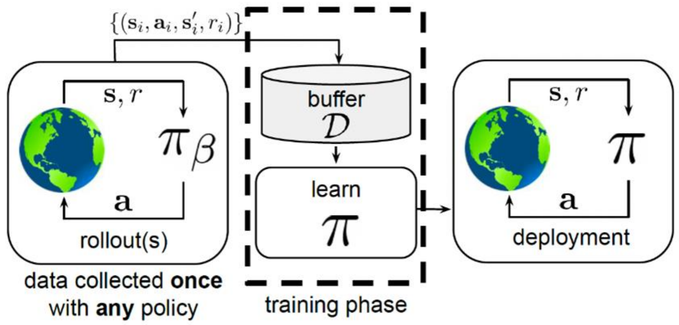
Team 6: Jongmin Lee, Beomhan Baek, Seungsu Han, Dongwon Kim, Mincheol Cho

TL;DR

- Offline RL often suffers from distributional shift and the overestimation of OOD actions, and CQL has been proposed to mitigate this issue by promoting pessimism.
- We propose **Probabilistic Conservative Q-learning (PCQ)**, a novel method that **adaptively penalizes** OOD actions based on their likelihood under the behavior policy.
- We provide theoretical guarantees that PCQ yields pessimistic Q-value estimates under linear MDP settings with mild assumptions.
- Empirical evaluations on MuJoCo and Antmaze benchmarks show that PCQ achieves good performance, compared to previous CQL and SCQ.

Introduction

Offline RL.



- **Distributional shift**: trained under behavior policy π_β, evaluated on learned policy π
- **Overestimated** Q-values on **out-of-distribution** (OOD) actions

Policy Evaluation in Actor-Critic.

$$\theta^{k+1} \leftarrow \text{argmin}_\theta \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \widehat{\mathcal{B}}^{\pi_{\phi_k}}[Q_{\theta_k}(s',a')]\right)^2\right]$$

Conservative Q-Learning (CQL).

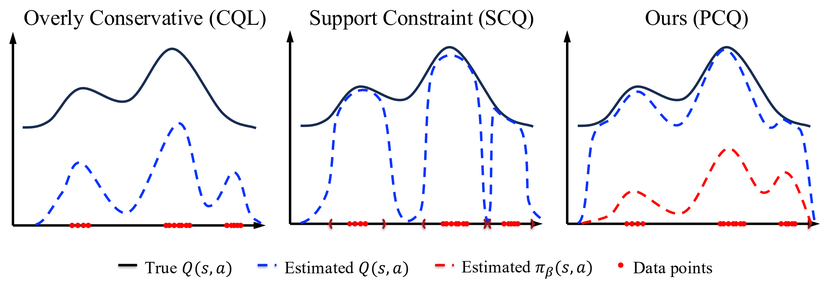
$$\theta^{k+1} \leftarrow \text{argmin}_\theta\ \alpha(\mathbb{E}_{s,a\sim\mathcal{D},a\sim\pi_{\phi_k}(\cdot|s)}[Q_\theta(s,a)] - \mathbb{E}_{s,a\sim\mathcal{D}}[Q_\theta(s,a)])$$
$$+\frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \widehat{\mathcal{B}}^{\pi_{\phi_k}}[Q_{\theta_k}(s',a')]\right)^2\right]$$

Strategically Conservative Q-Learning (SCQ).

$$\theta^{k+1} \leftarrow \text{argmin}_\theta\ \alpha\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_{ood}(\cdot|s)}[Q_\theta(s,a)]$$
$$+\frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \widehat{\mathcal{B}}^{\pi_{\phi_k}}[Q_{\theta_k}(s',a')]\right)^2\right]$$

where $\text{Proj}_\mathcal{D}(a|s) = \text{argmin}_{(s,a')\in\mathcal{D}}\|a - a'\|$,
$A_{ood}(s) = \{a : \|a - \text{Proj}_\mathcal{D}(a|s)\|_2 \geq \delta\}$,
$\pi_{ood} \propto \pi_{\phi_k} 1_{\{a\in A_{ood}(s)\}}$

Method & Analysis



Probabilistic Conservative Q-learning (PCQ).

$$\theta^{k+1} \leftarrow \text{argmin}_\theta\ \alpha_k\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_{\phi_k}(\cdot|s)}[w(s,a)Q_\theta(s,a)]$$
$$+\frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \widehat{\mathcal{B}}^{\pi_{\phi_k}}[Q_{\theta_k}(s',a')]\right)^2\right]$$

where $w(s,a) = f(-\log\hat{\pi}_\beta(a \mid s))$ is a **continuous, likelihood-weighted** penalty for OOD actions, f is a **nondecreasing** function and $\hat{\pi}_\beta$ denotes the **distribution of offline behavior policy**, estimated by Conditional Variational Autoencoder (CVAE) .

Linear MDP.

An MDP is **linear** there exist $\psi: \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d, g: \mathcal{S}\to\mathbb{R}^d$, and $w \in \mathbb{R}^d$ such that such that $r(s,a) = \langle\psi(s,a), w\rangle$, $P(s' \mid s,a) = \langle\psi(s,a), g(s')\rangle$.

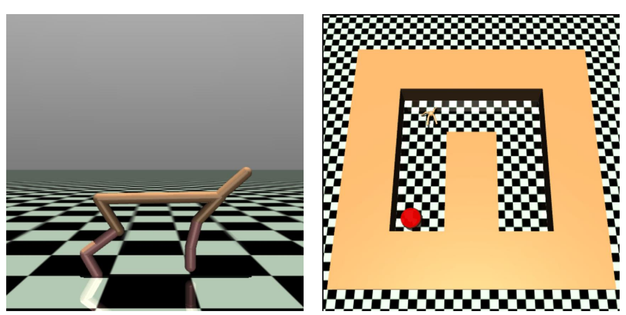
Theorem [Pessimistic Q-value estimates].

Let $\hat{Q}^k$ and Q^k be the solutions to our objective function with and without sample error, respectively. Suppose that the MDP is linear with respect to specific features. Then, with high probability, there exists a parameter α_k in the objective function such that

$$\hat{Q}^k \leq Q^k + \epsilon$$

for every k.

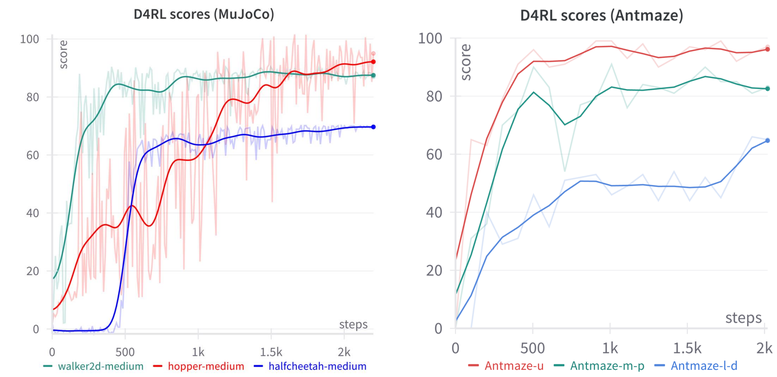
Experimental Setups



- (Left) **Gym-MuJoCo** Environment (Right) **Antmaze** Environment
- **MuJoCo**: Halfcheetah, hopper, walker2d; medium
- **Antmaze**: u(umaze), m(medium), l(large), d (diverse), p(play)

Results

Score Curves of PCQ.



Performance of PCQ.

MuJoCo Tasks

Task Name	TD3+BC	CQL	SCQ	PCQ (ours)
halfcheetah-m	54.7 ± 0.9	45.3 ± 0.7	$\mathbf{68.3 \pm 1.6}$	67.37 ± 3.1
hopper-m	60.9 ± 7.6	64.0 ± 0.8	89.5 ± 5.2	$\mathbf{90.61 \pm 7.1}$
walker2d-m	77.0 ± 2.9	79.5 ± 1.2	86.9 ± 0.6	$\mathbf{87.32 \pm 0.4}$

AntMaze Tasks

Task Name	TD3+BC	CQL	SCQ	PCQ (ours)
antmaze-u	66.3 ± 6.2	74.0	$\mathbf{97.8 \pm 1.1}$	96.0 ± 2.0
antmaze-u-d	53.8 ± 8.5	84.0	$\mathbf{89.5 \pm 9.3}$	58.0 ± 9.7
antmaze-m-p	26.5 ± 18.4	61.2	81.1 ± 16.2	$\mathbf{82.14 \pm 8.9}$
antmaze-m-d	25.9 ± 15.3	53.7	$\mathbf{79.4 \pm 10.1}$	28.4 ± 10.4
antmaze-l-p	0.0 ± 0.0	15.8	$\mathbf{67.2 \pm 10.5}$	62.2 ± 12.1
antmaze-l-d	0.0 ± 0.0	14.9	60.4 ± 17.6	$\mathbf{62.2 \pm 9.1}$

Discussions & Conclusion

Discussions.

- MuJoCo: PCQ **consistently** matches or outperforms SCQ
- Antmaze: PCQ's performances **vary** across tasks
- Impact of incorrect prediction of $\hat{\pi}_\beta$ by CVAE is more significant to PCQ; underlying errors by using ELBO

Limitations & Future Works.

- Lack of computing resource
- Insufficient hyperparameter tuning
- Further improvement by choosing different f
- Ablation studies on pretrained CVAE
- More grounded theoretical analysis