Budgeting Counterfactual for Offline RL

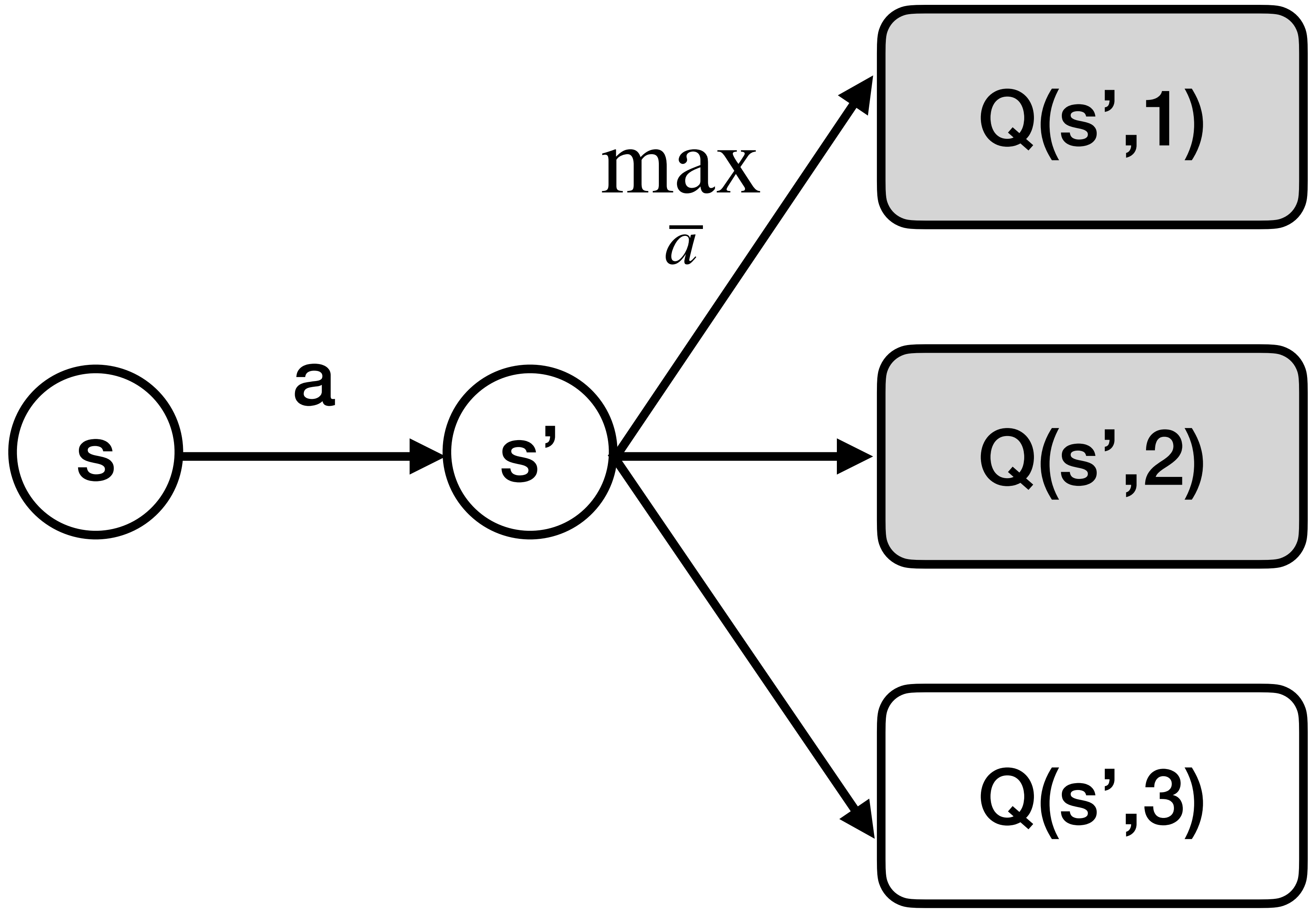
Yao Liu[1], Pratik Chaudhari[1,2], Rasool Fakoor[1]

[1]Amazon, [2]University of Pennsylvania

Why online RL fails in offline

"Extrapolation error"

Bellman backup: $Q(s,a) \leftarrow r(s,a) + \gamma \max_{\bar{a} \in \mathcal{A}} Q(s', \bar{a})$



Unexplored states.

Variance in $Q(s', \bar{a})$ => Bias in the max

Revisit extrapolation error (EE)

- No EE if we always backup from behavior action:

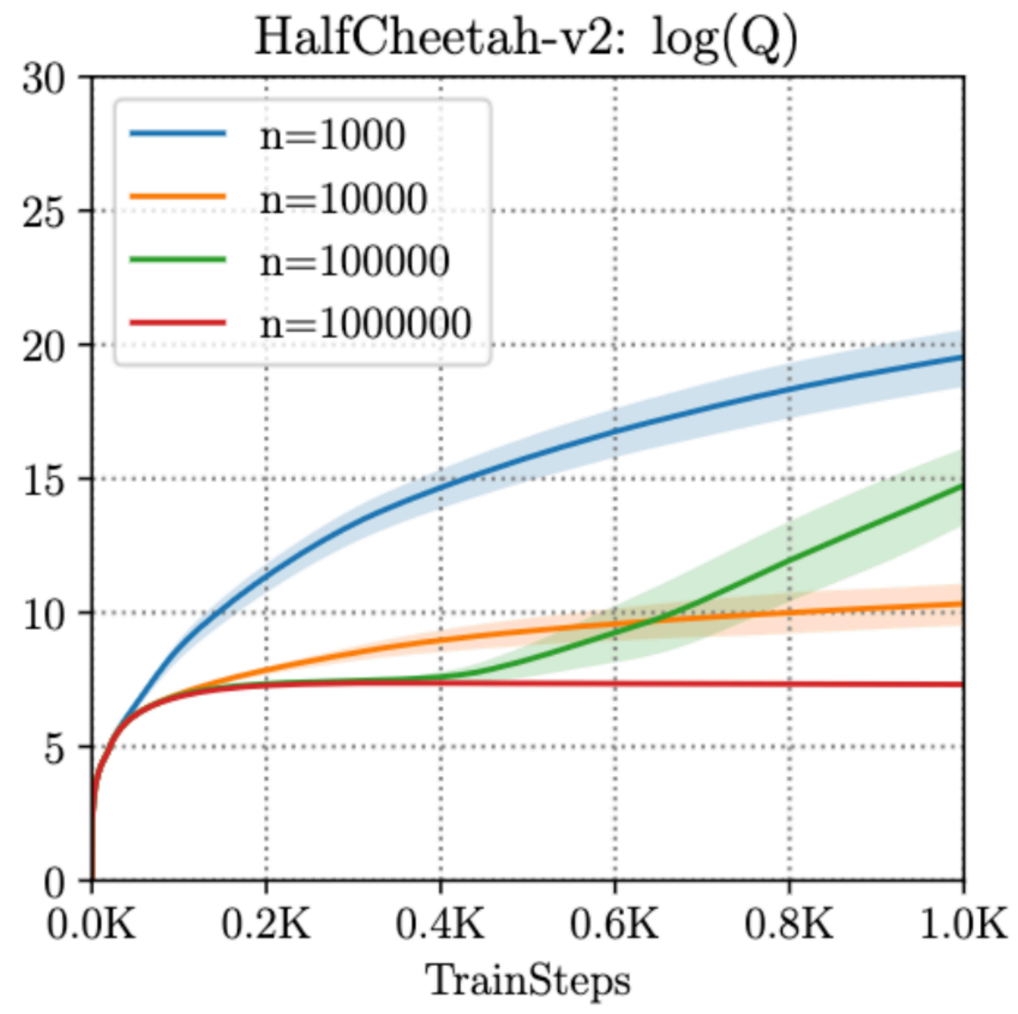
$$Q(s,a) \leftarrow r(s,a) + \gamma \mathbb{E}_{a' \sim \mu} Q(s',a')$$

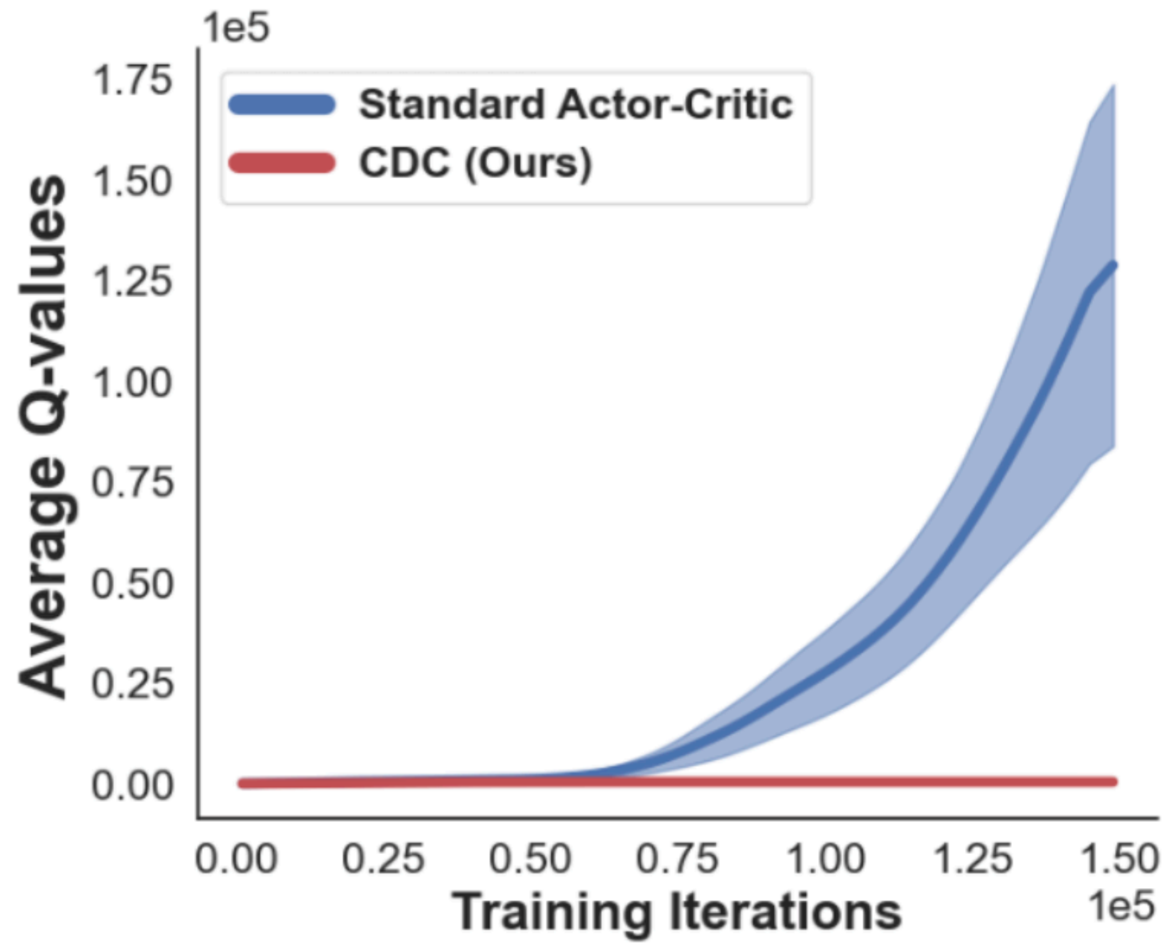
- The EE comes from counterfactual $\arg\max_{\bar{a}} Q(s', \bar{a})$, and it amplify itself by Bellman backup:

$$Q(s,a) \leftarrow TQ(s,a) := r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s',a')$$

$$Q_k \leftarrow TQ_{k-1} \leftarrow T \circ TQ_{k-2} \leftarrow \ldots \leftarrow (T)^k Q_0$$

Empirically, EE was reported to increase nearly exponentially





Budgeting Counterfactual

- **Observation 1**: Bellman backup with $\pi \neq \mu$ or $\max$ backup (counterfactual decisions) results in exponentially larger divergence $(1+\delta)^H$.

- **Observation 2**: Controlling local divergence δ does not stop the exponential increase.

Key idea: only apply $Q(s,a) \leftarrow r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s',a')$ with <u>a limited number</u> of steps <u>in one trajectory</u>. For the rest of decision steps use μ.

Budgeting Counterfactual

How to decide when to take the greedy action and when to take μ?

- Dynamic programming

$$TQ(s,b,a) := \mathbb{E}_{s,a,s',a'}\left[r(s,a) + \gamma \begin{cases} \max\{\max_{\bar{a}\in\mathcal{A}} Q(s',b-1,\bar{a}), Q(s',b,a')\} & b>0 \\ Q(s',b,a') & b=0 \end{cases}\right]$$

where $(s,a,s',a') \in \mathcal{D}$.

Theoretical justification

Theorem 1: There is a unique fixed point of T, that is

$$Q^{\star}(s, b, a) := \max_{\pi} E\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a, b_0 = b; \pi\right] \text{ s.t. } b_t \geq 0,\ \forall t \geq 0$$

where $b_t = b_{t-1} - 1\{\pi(\cdot \mid s_{t-1}, b_{t-1}) \neq \mu(\cdot \mid s_{t-1})\}$.

- The fixed point iteration on T converge to the optimal value function under a constrain on the number of counterfactual decisions.

Algorithm: BCOL

Continuous action space: $\pi_\phi(\,\cdot\mid s) \approx \arg\max$

$$\hat{T}Q_\theta(s,b,a) := r(s,a) + \gamma \begin{cases} \max\{\mathbb{E}_{\bar{a}\sim\pi_\phi} Q_\theta(s',b-1,\bar{a}), Q_\theta(s',b,a')\} & b > 0 \\ Q_\theta(s',b,a') & b = 0 \end{cases}$$

Actor loss: $-\sum_{b=0}^{B} \mathbb{E}_{s\sim D, a\sim\pi_\phi(\cdot|s,b)} Q_\theta(s,b,a)$

Critic loss: $\sum_{b=0}^{B} \mathbb{E}_{(s,a,s',a')\sim D}\left[\left(Q_\theta(s,b,a) - \hat{T}Q_{\bar{\theta}}(s,b,a)\right)^2\right]$

Inference

How to select the action at test-time based on π_ϕ and Q_θ

- $Q_\theta(s, b, a)$: the optimal value starting from (s, a), using at most b counterfactual decisions in the future.

- $\pi_\phi(\cdot \mid s, b)$: the optimal counterfactual decision given s and at most b counterfactual decisions in the future.

Initialize $b = B$, select action and update b by:

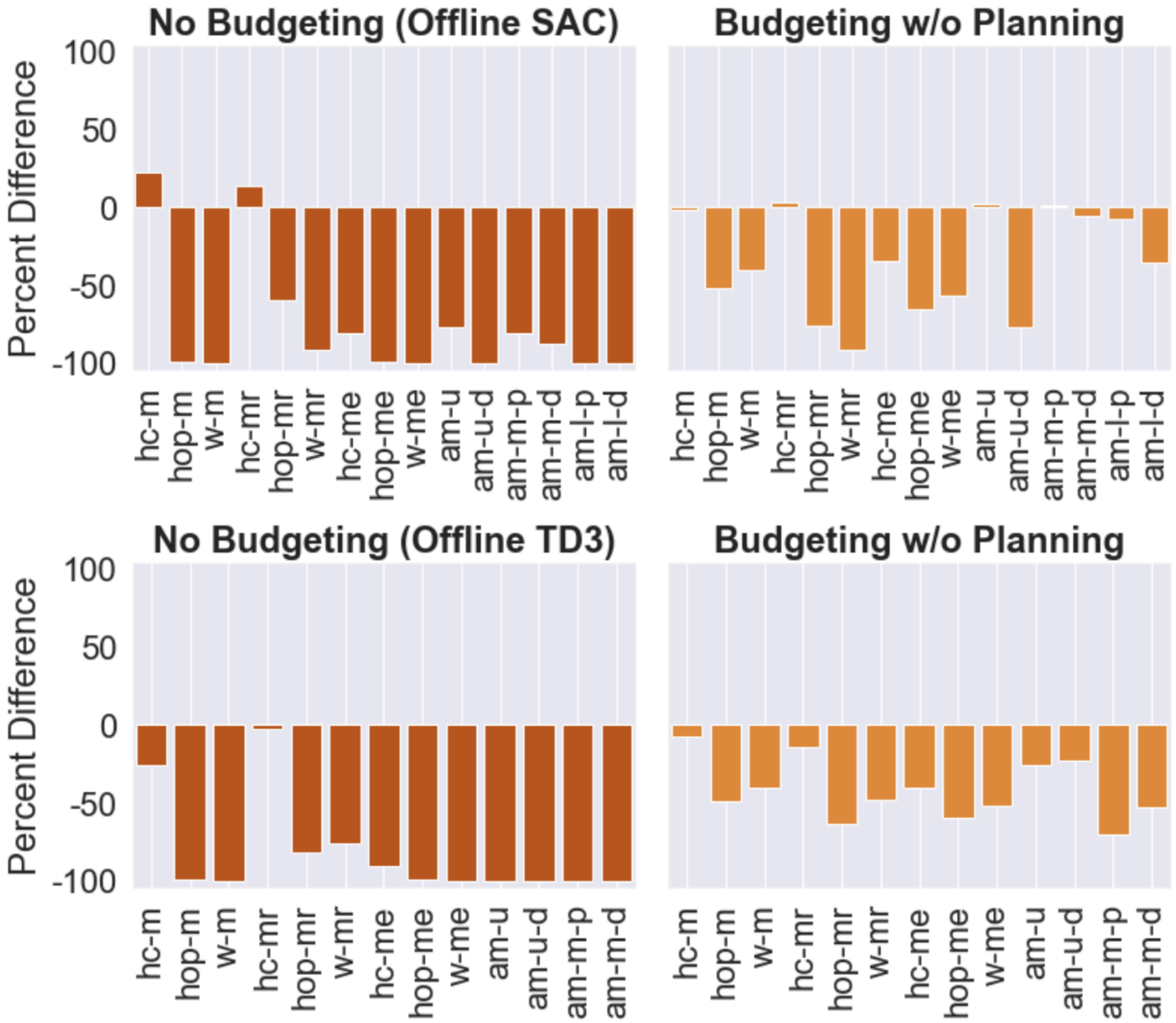
$a \sim \mu(\cdot \mid s), b \leftarrow b$ if $\mathbb{E}_{\overline{a} \sim \pi(s,b)} Q(s, b-1, \overline{a}) \leq \mathbb{E}_{\overline{a} \sim \mu(s)} Q(s, b, \overline{a})$ or $b = 0$

$a \sim \pi_\phi(\cdot \mid s, b), b \leftarrow b - 1$ otherwise

Results

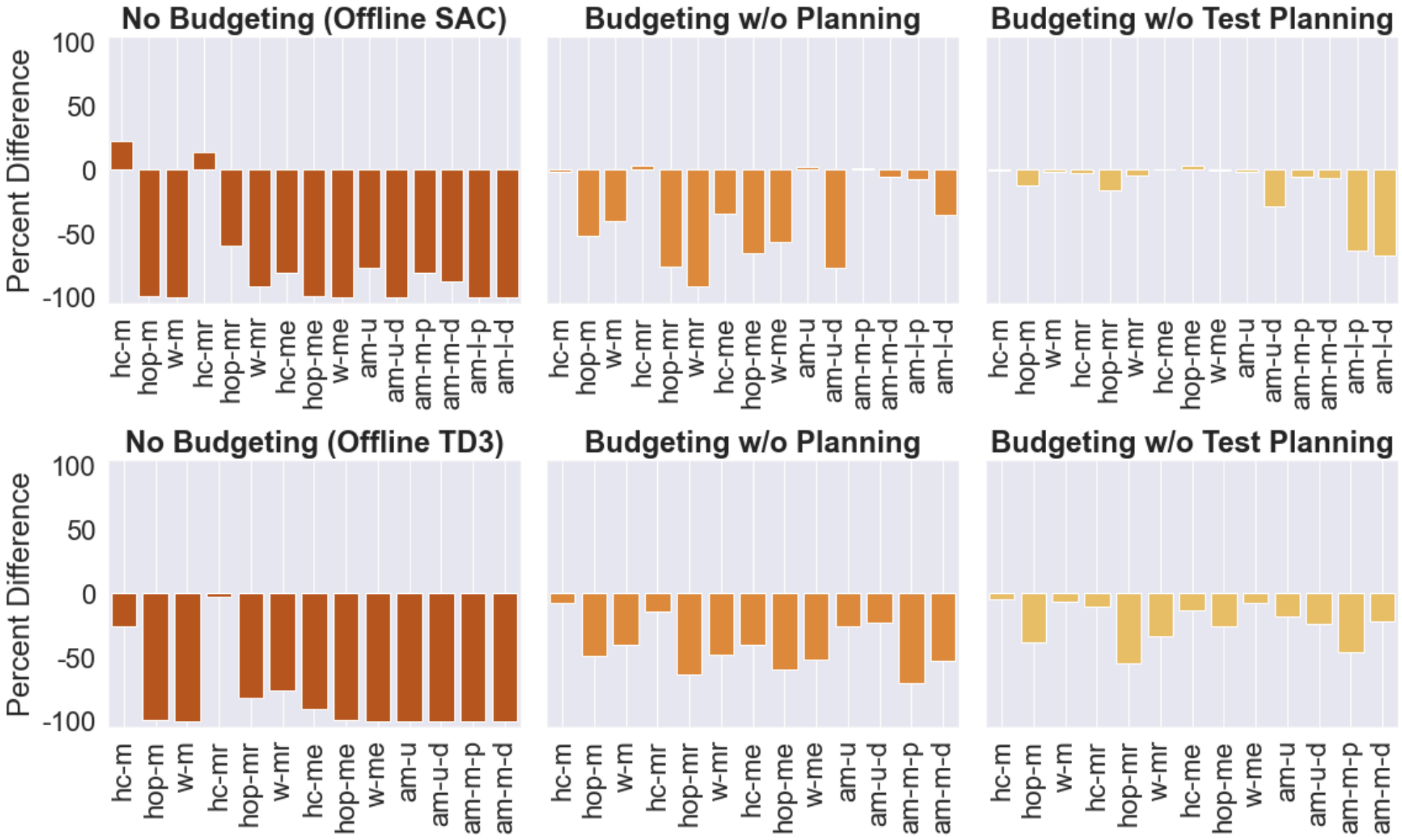
Task Name	BC	IQL	Onestep	TD3+BC	BCOL (TD3)	CQL	CDC	BCOL (SAC)
halfcheetah-m	42.6	47.4	55.6	<u>48.4</u>	45.0	46.1	<u>**62.5**</u>	50.1
hopper-m	52.9	66.3	83.3	59.4	<u>**85.8**</u>	64.6	<u>84.9</u>	83.2
walker2d-m	75.3	78.3	**85.6**	<u>84.5</u>	76.7	74.5	70.7	<u>84.1</u>
halfcheetah-mr	36.6	44.2	42.5	<u>44.4</u>	40.9	45.4	<u>**52.3**</u>	46.2
hopper-mr	18.1	94.7	71.0	50.1	<u>83.4</u>	92.3	87.4	<u>**99.8**</u>
walker2d-mr	26.0	73.9	71.6	<u>80.2</u>	49.7	83.7	<u>**87.8**</u>	86.0
halfcheetah-me	55.2	86.7	**93.5**	<u>91.5</u>	88.7	<u>87.3</u>	66.3	86.9
hopper-me	52.5	91.5	102.1	100.5	<u>106.8</u>	<u>**109.2**</u>	83.2	99.0
walker2d-me	107.5	109.6	**110.9**	<u>110.1</u>	108.5	109.9	103.9	<u>**110.9**</u>
antmaze-u	54.6	87.5	64.3	<u>**96.3**</u>	93.3	<u>94.0</u>	93.6	90.3
antmaze-u-d	45.6	62.2	60.7	<u>71.7</u>	68.0	47.3	57.3	<u>**90.0**</u>
antmaze-m-p	0.0	**71.2**	0.3	1.7	<u>12.3</u>	62.4	59.5	<u>70.0</u>
antmaze-m-d	0.0	70.0	0.0	0.3	<u>14.0</u>	<u>**74.3**</u>	64.6	72.3
antmaze-l-p	0.0	**39.6**	0.0	0.0	0.0	34.2	33.0	<u>35.6</u>
antmaze-l-d	0.0	**47.5**	0.0	0.3	0.0	<u>40.7</u>	25.3	37.6
mujoco total	466.7	692.4	716.0	669.2	<u>685.6</u>	713.0	699.0	<u>**746.0**</u>
antmaze total	100.2	378.0	125.3	171.3	<u>187.7</u>	352.9	333.5	<u>**396.0**</u>
Total	566.9	1070.4	841.3	840.2	<u>873.3</u>	1065.9	1032.5	<u>**1142.0**</u>

Dynamic programming matters



- With budget but randomly select where to spend the budget

Dynamic programming matters



- Train Q_θ and π_ϕ with budget
- Randomly select between π_ϕ and μ during test

Summary

- Offline RL suffers from extrapolation errors on counterfactual actions
- New algorithm: behavior cloning + few key counterfactual actions.
- Use dynamic programming to find where to do counterfactual decisions.
- No additional regularization, simple yet effective compared with SOTA offline RL methods.