1 Rebuttal Supplementary

1.1 Impact of Various αs

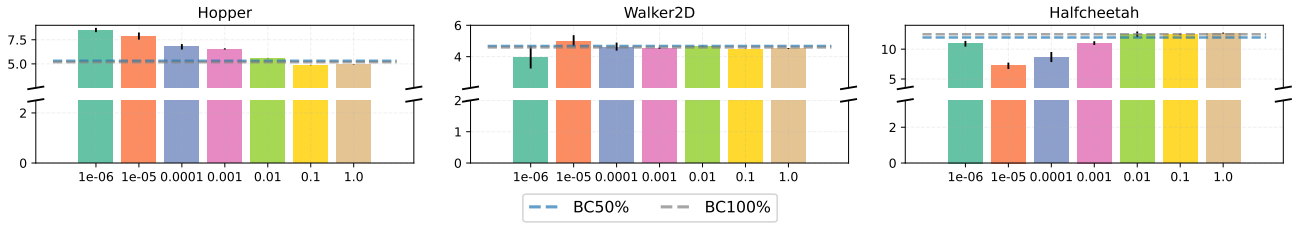

Figure 1: Comparison of different levels of the hyperparameter (α) of ROIDICE in locomotion environments using expert data quality. We average the scores and get $\pm 2\times$ standard error with 5 seeds across 10 episodes.

1.2 Comparison of offline constrained RL algorithms

Data quality	medium	medium-expert	expert
ROIDICE	9.21±0.49	8.29±0.18	8.51±0.25
VOCE 50th	1.807±0.63	1.33±1.28	1.34±0.67
VOCE 80th	1.799±0.62	1.34±1.3	1.47±0.76
CPQ 50th	1.13±0.58	-0.06±0.41	-0.11±0.71
CPQ 80th	-0.10±0.64	-0.15±0.48	-0.70±0.19

Table 1: ROI of ROIDICE compared with offline constrained RL algorithms. We average the each scores and get $\pm 2\times$ standard error with 5 seeds across 10 episodes.

1.3 ROIDICE pseudocode

$$\begin{aligned}
\mathcal{L}_{\nu_\phi} =& \mathbb{E}_{s\sim p_0}\left[t(1-\gamma)\nu(s)\right] \\
&+ \mathbb{E}_{(s,a)\sim d_D}\left[w^*_{\nu,\mu,t}(s,a)(e_\nu(s,a)-\mu c(s,a)) - \alpha f(w^*_{\nu,\mu,t}(s,a),t)\right] \quad (1)\\
\mathcal{L}_{\mu} =& \mathbb{E}_{(s,a)\sim d_D}\left[w^*_{\nu,\mu,t}(s,a)(e_\nu(s,a)-\mu c(s,a)) - \alpha f(w^*_{\nu,\mu,t}(s,a),t)\right] + \mu \quad (2)\\
\mathcal{L}_{t} =& \mathbb{E}_{s\sim p_0}\left[t(1-\gamma)\nu(s)\right] - E_{(s,a)\sim d_D}\left[\alpha f(w^*_{\nu,\mu,t}(s,a),t)\right] \quad (3)
\end{aligned}$$

Algorithm 1 ROIDICE

Input: The offline dataset D, the initial state offline dataset p_0, parameterized Lagrangian multipliers ν_ϕ, μ, t, and policy π_θ.
Output: Optimal policy π^*_θ.
Initialize all parameters.
while convergence **do**
  Update ν_ϕ, μ, t with $\mathcal{L}_{\nu_\phi}$, $\mathcal{L}_\mu$, $-\mathcal{L}_t$
  Update θ with $\mathcal{L}_\theta = -\mathbb{E}_{(s,a)\sim D}\left[\frac{1}{t}\cdot w^*_{\nu,\mu,t}(s,a)\cdot \log \pi_\theta(a|s)\right]$
end while

1.4 ROI of ROIDICE in Safety RL environments

Task	CarGoal	PointPush
ROIDICE	0.160±0.02	0.033±0.01
COptiDICE 50th	0.119±0.01	0.026±0.01
COptiDICE 80th	0.115±0.01	0.022±0.01

Table 2: ROI of ROIDICE compared with offline constrained RL algorithms. We average the each scores and get $\pm 2\times$ standard error with 5 seeds across 10 episodes.