Tasks	Proposed	COMBO	BCQ	BEAR	OptiDICE	ATAC	CQL	IQL	TD3+BC
walker2d-medium	80.8 ± 5.1	81.9 ± 2.8	53.1	59.1	21.8 ± 7.1	**89.6**	77.2 ± 4.2	78.3 ± 4.3	81.7 ± 2.3
walker2d-medium-replay	**99.6** ± 2.9	56.0 ± 8.6	15.0	19.2	21.6 ± 2.1	92.5	20.8 ± 1.6	73.9 ± 2.8	34.4 ± 4.2
walker2d-medium-expert	108.2 ± 7.4	103.3 ± 5.6	57.5	40.1	74.8 ± 9.2	**114.2**	103.8 ± 6.9	109.6 ± 7.0	100.5 ± 8.9
walker2d-random	**11.2** ± 3.8	7.0 ± 3.6	4.9	7.3	9.9 ± 4.3	6.8	4.7 ± 1.5	5.8 ± 2.8	3.4 ± 1.7
hopper-medium	94.9 ± 4.3	97.2 ± 2.2	54.5	52.1	94.1 ± 3.7	85.6	74.3 ± 5.8	66.3 ± 6.4	**98.4** ± 1.6
hopper-medium-replay	**113.0** ± 2.1	89.5 ± 1.8	33.1	33.7	36.4 ± 1.1	102.5	32.6 ± 1.9	94.7 ± 1.5	44.4 ± 3.7
hopper-medium-expert	117.8 ± 1.9	111.1 ± 2.9	110.9	96.3	111.5 ± 0.6	**119.2**	111.4 ± 1.2	91.5 ± 2.2	112.4 ± 0.3
hopper-random	**18.7** ± 1.5	17.9 ± 1.4	10.6	11.4	11.2 ± 1.1	17.5	10.7 ± 0.1	10.8 ± 0.6	11.1 ± 0.2
halfcheetah-medium	**58.1** ± 1.4	54.2 ± 1.5	40.7	41.7	38.2 ± 0.1	53.3	37.2 ± 0.3	47.4 ± 1.1	27.8 ± 0.7
halfcheetah-medium-replay	49.3 ± 2.1	**55.1** ± 1.0	38.2	38.6	39.8 ± 0.3	48.0	41.9 ± 1.1	44.2 ± 2.5	48.3 ± 0.7
halfcheetah-medium-expert	**98.5** ± 3.8	90.0 ± 5.6	64.7	53.4	91.1 ± 3.7	94.8	66.7 ± 8.9	86.7 ± 3.6	95.9 ± 3.9
halfcheetah-random	37.6 ± 2.4	**38.8** ± 3.7	2.2	25.1	11.6 ± 1.2	3.9	26.7 ± 1.4	22.4 ± 1.8	26.1 ± 1.8
maze2d-umaze	96.5 ± 27.8	34.2 ± 8.6	12.8	3.4	**111.0** ± 8.3	84.4 ± 24.8	50.5 ± 7.9	41.5 ± 4.7	13.8 ± 22.8
maze2d-medium	137.5 ± 18.9	49.9 ± 13.9	8.3	29.0	145.2 ± 17.5	**152.3** ± 34.6	28.6 ± 9.2	38.5 ± 4.2	59.1 ± 44.7
maze2d-large	**187.8** ± 15.2	128.2 ± 17.3	6.2	4.6	155.7 ± 33.4	142.1 ± 33.8	46.2 ± 16.2	54.2 ± 18.1	87.6 ± 15.4

Table 1: Results for D4RL datasets. Each number is the normalized score of the policy at the last iteration of training, averaged over 5 random seeds. We take results of COMBO, OptiDICE and ATAC from their original papers for Gym locomotion, and run COMBO and ATAC using author-provided implementations for Maze2D. The results of BCQ, BEAR methods from the D4RL original paper. In addition, CQL, IQL and TD3+BC are re-run to ensure a fair evaluation process for all tasks. Our algorithm achieves the best performance in 7 tasks and is comparable to the baselines in the remaining tasks.

Algorithms	Proposed	COMBO	OptiDICE	ATAC	CQL	IQL	TD3+BC
Runtime	3h34min	3h16min	4h17min	6h24min	4h03min	1h58min	1h29min

Table 2: Algorithms' runtime finish the ∼100K halfcheetah-medium-replay task on a GeForce RTX 3090.

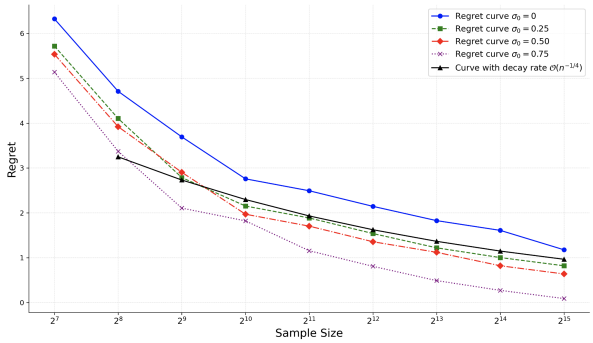


Figure 1: Convegence rate of the near-optimal regret (compete to the optimal policy) on the synthetic dataset with different degrees of exploration σ_0. A smaller σ_0 indicates the training data is less explored. The plot shows that the convergence rate is close to $\mathcal{O}(n^{-1/4})$ in all scenarios, which validates the theoretical regret bound of our practical algorithm in Theorem 5.1 and 5.2.

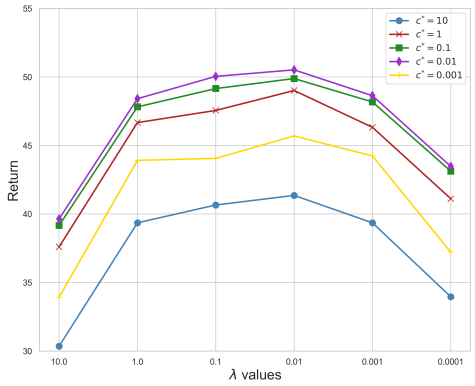
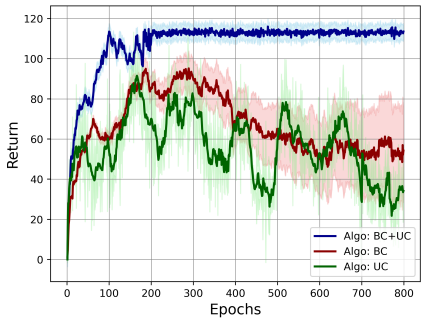
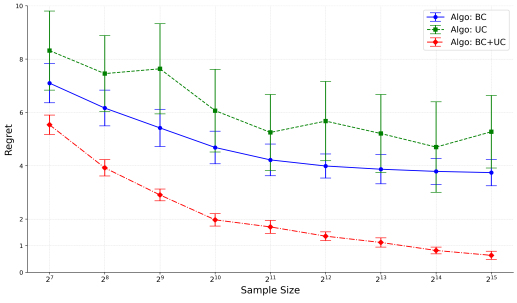


Figure 2: Sensitivity analysis on the effects of hyperparameters λ and c^* for model performance with halfcheetah-medium-replay dataset. Our algorithm has desired robustness over a wide range of hyperparameters.



(a) The plots show the policy performance across optimization epochs of our algorithm with the hopper-medium-replay dataset. The algorithm only with UC (Algo: UC) is unstable, and (Algo: BC) yields suboptimal performance. The stability and performance are greatly improved with both components, where BC provides a consistent estimation guarantee for q^π, UC enhances model extrapolation (discussed in lines 220-230 of the main paper).



(b) This plot shows the regret across sample size on the synthetic dataset. Error bars show 95% confidence intervals. (Algo: BC) or (Algo: UC) has sub-optimal regret and a slower convergence rate than (Algo: BC+UC).

Figure 3: Ablation of the return with/without lower-level components of our algorithm: Bellman consistency (BC) and uncertainty control (UC) in line 153 of the main paper.

Tasks	s-low	s-med	s-high	c-low	c-med	c-high	p-540	p-544	p-552	p-567	p-584	p-596
TD3+BC	2.1 ± 0.7	2.4 ± 0.9	3.0 ± 0.7	44.6 ± 6.2	49.9 ± 5.2	53.8 ± 4.7	17.1 ± 0.8	9.7 ± 0.8	7.4 ± 0.8	27.8 ± 1.5	22.4 ± 1.7	4.8 ± 1.3
Proposed	3.1 ± 0.8	3.6 ± 0.8	4.1 ± 0.6	52.5 ± 6.6	57.8 ± 6.8	59.7 ± 5.9	20.4 ± 0.5	13.2 ± 1.9	7.9 ± 0.9	36.9 ± 1.3	33.1 ± 1.8	6.5 ± 1.1

Table 3: Return for our algorithm and TD3+BC across the original datasets in our paper (Simulated=s, CartPole=c, Patient=p).