Carnegie Mellon University

Learning from Sparse Offline Datasets via Conservative Density Estimation

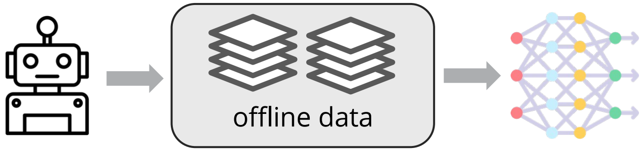
Zhepeng Cen[1], Zuxin Liu[1], Zitong Wang[1], Yihang Yao[1], Henry Lam[1], Ding Zhao[1]

[1]CMU, [2]Columbia University

Learning from Sparse Offline Datasets via Conservative Density Estimation

- TL;DR: We propose a new offline reinforcement learning (RL) method to improve the performances in sparse reward and scarce data settings.

- Offline RL:
  - Learn a policy from fixed dataset
  - Without further interaction with Environment

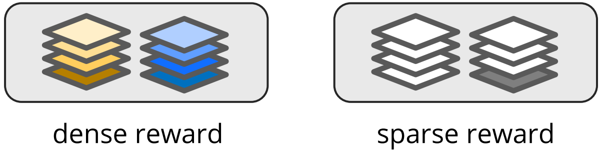


Background

- Challenges of offline RL in terms of **data**
  - **Sparse reward setting**

(e.g., reward > 0 only when reaching the goal)

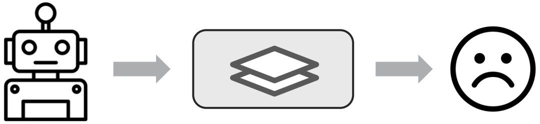
➔ It makes it hard to tell whether a policy is good or not, especially with Bellman-style value learning



  - **Scarce data setting**

➔ The coverage of offline data on state-action space is not enough

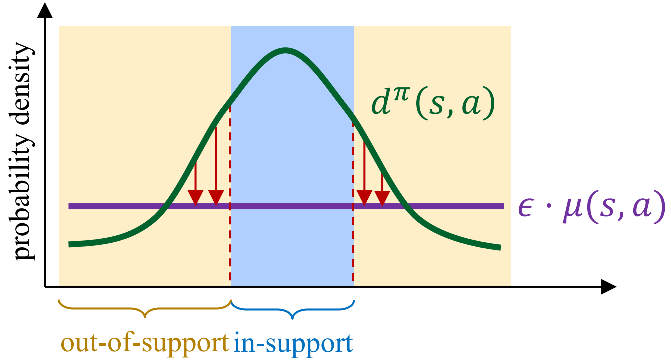
➔ The out-of-distribution (OOD) issue is more severe



Carnegie Mellon University

Method - Conservative Density Estimation

- We formulate the optimization problem in terms of stationary distribution $d^\pi(s,a)$
  - Based on Distribution Correction Estimation (D?CE)
  - Additional constraint: be conservative on OOD region
    - *Mitigate the support mismatch issue*



$$\max_{d^\pi \geq 0} \mathbb{E}_{d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^\mathcal{D})$$ → maximize regularized reward

$$s.t. \sum_a d^\pi(s,a) = (1-\gamma)\rho_0(s) + \mathcal{T}_* d^\pi(s), \forall s$$ → d^π should be valid

$$d^\pi(s,a) \leq \epsilon\mu(s,a), \forall s, a \notin \text{supp}(d^\mathcal{D})$$ → be conservative on unseen region

Carnegie Mellon University

Method

- How to solve the above constrained optimization problem?
  - Let $\hat{d}^D(s,a) = \zeta d^D(s,a) + (1-\zeta)\mu(s,a)$, $w(s,a) = \frac{d^\pi(s,a)}{\hat{d}^D(s,a)}$
  - ➔ $\min_{\lambda \geq 0, v} \max_{w} \mathcal{L}(w; v, \lambda)$

- Nice properties for solving this min-max problem.
  - Inner max problem has a closed-form solution:

    $w^*(s,a) = (f')^{-1}\big(\tilde{A}(s,a)/\alpha\big), \tilde{A}$ can be represented by v, λ
  - Outer min problem: $\min_{\lambda \geq 0, v} \mathcal{L}(w^*; v, \lambda)$ is a convex optimization
  - Mitigate the value estimation error compared to Bellman update

Method

- The training pipeline of our method CDE:
  - Policy evaluation: solve the optimal (w^*, v^*, λ^*) from the minimax problem
  - Policy extraction: $\min_\theta D_{KL}[d^{\pi_\theta} | w^* \hat{d}^D] \approx \min_\theta E\left[-\log w^*(s,a) + D_{KL}[\pi_\theta | \pi_D]\right]$

- Theoretical analysis:
  - The importance ratio $w^*(s,a)$ on unseen region is bounded
  - The performance gap to optimal policy is bounded
    - *in terms of (1) the quality of dataset, and (2) the size of dataset*

Carnegie Mellon University

Experiment – sparse reward setting

- Performances on D4RL sparse-reward tasks

Task	BC	BCQ	CQL	IQL	TD3+BC	Algae-DICE	OptiDICE	CDE
maze2d-umaze	3.8	32.8	5.7	50.0	41.5	-15.7	111.0±8.3	**134.1**±10.4
maze2d-medium	30.3	20.7	5.0	31.0	76.3	10.0	145.2±17.5	**146.1**±13.1
maze2d-large	5.0	47.8	12.5	58.0	77.8	-0.1	155.7±33.4	**210.0**±13.5
pen-human	63.9	68.9	37.5	71.5	2.0	-3.3	42.1±15.3	**72.1**±15.8
hammer-human	1.2	0.5	**4.4**	1.4	1.4	0.3	0.3±0.0	1.9±0.7
door-human	2.0	0.0	**9.9**	4.3	-0.3	0.0	0.1±0.1	7.7±3.3
relocate-human	0.1	-0.1	0.2	0.1	-0.3	-0.1	-0.1±0.1	**0.3**±0.1
pen-expert	85.1	**114.9**	107.0	111.7	79.1	-3.5	80.9±31.4	105.0±12.3
hammer-expert	125.6	107.2	86.7	116.3	3.1	0.3	**127.0**±3.0	**126.3**±3.4
door-expert	34.9	99.0	101.5	103.8	-0.3	0.0	103.4±2.8	**105.9**±0.3
relocate-expert	101.3	41.6	95.0	**102.7**	-1.5	-0.1	99.7±4.2	**102.6**±1.9

Carnegie Mellon University

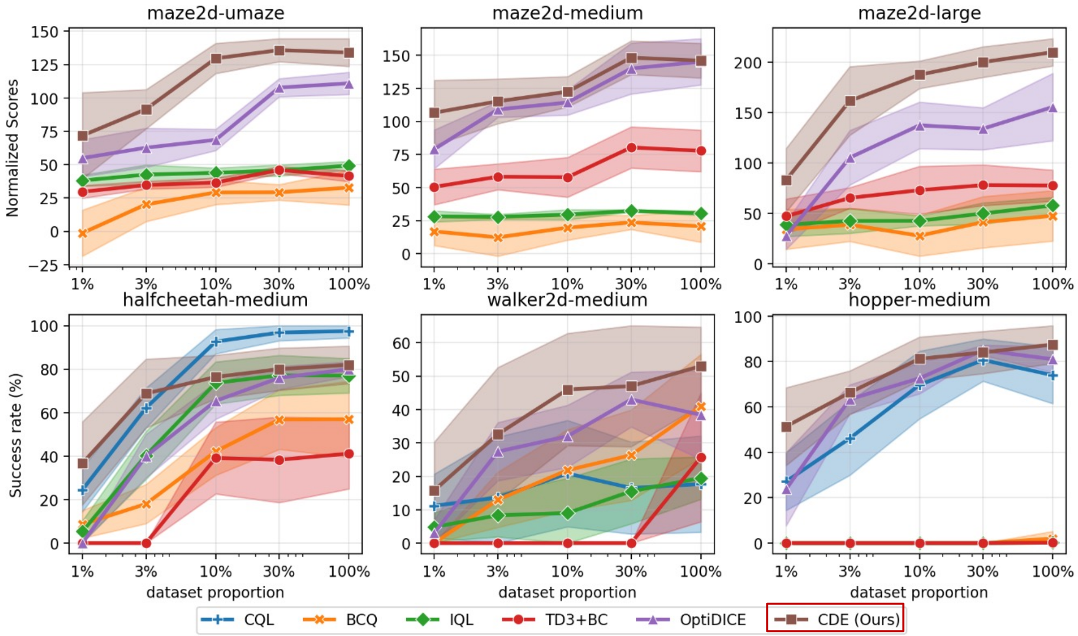
Experiment – scarce data setting

- Test the performances with a small proportion of original dataset

- Sparse-MuJoCo tasks adopt binary sparse rewards:

$$r_{\text{sparse}}(s_t, a_t) = \mathbf{1}\left(\sum_{\tau=0}^{t} r_\tau \geq R\right)$$

(s_t, a_t) is in the trajectory $\{s_0, a_0, r_0, s_1, \ldots\}$



Summary

- We propose an offline RL method CDE by applying pessimism from the perspective of stationary distribution.
- Our method shows better performances in sparse reward or scarce data settings.