Learning from Sparse Offline Datasets via Conservative Density Estimation

Zhepeng Cen [1] Zuxin Liu [1] Zitong Wang [2] Yihang Yao [1] Henry Lam [2] Ding Zhao [1]

Abstract

Offline reinforcement learning (RL) offers a promising direction for learning policies from pre-collected datasets without requiring further interactions with the environment. However, existing methods struggle to handle out-of-distribution (OOD) extrapolation errors, especially in sparse reward or scarce data settings. In this paper, we propose a novel training algorithm called Conservative Density Estimation (CDE), which addresses this challenge by explicitly imposing constraints on the state-action occupancy stationary distribution. CDE overcomes the limitations of existing approaches, such as the stationary distribution correction method, by addressing the support mismatch issue in marginal importance sampling. Our method achieves state-of-the-art performance on the D4RL benchmark. Notably, CDE consistently outperforms baselines in challenging tasks with sparse rewards or insufficient data, demonstrating the advantages of our approach in addressing the extrapolation error problem in offline RL.

1. Introduction

Reinforcement Learning (RL) has witnessed remarkable advancements in recent years (Akkaya et al., 2019; Kiran et al., 2021). Nevertheless, the success of RL relies on continuous online interactions, resulting in high sample complexity and potentially restricting its practical applications in real-world scenarios (Levine et al., 2016; Gu et al., 2022). As a compelling solution, offline RL has been brought to the fore, with the objective of learning effective policies from pre-existing datasets, thereby eliminating the necessity for further environment interactions (Fu et al., 2020; Prudencio et al., 2023).

Despite its benefits, offline RL is not devoid of challenges, most notably the out-of-distribution (OOD) extrapolation errors, which emerge when the agent encounters state-actions that were absent in the dataset. These issues pose significant hurdles when learning policies from datasets with sparse rewards or low coverage of state-action spaces (Levine et al., 2020). To address OOD estimation errors in value-based offline RL, current efforts have primarily revolved around two strategies: pessimism-based methods (Xie et al., 2021a; Shi et al., 2022) and the integration of regularizations (Kostrikov et al., 2021a). However, these approaches hinge on assumptions of the behavior policy. In addition, pessimism-based techniques may be prone to over-pessimism, especially in high-dimensional state-action spaces, while regularization methods often struggle with the tuning of the regularization coefficient (Nachum et al., 2019a). As such, striking the optimal balance of conservativeness, particularly in sparse-reward settings, remains an elusive goal (Nachum & Dai, 2020).

Recent attention has been drawn towards an alternative method that employs importance sampling (IS) for offline data distribution correction (Precup, 2000; Jiang & Li, 2016). Among these, Distribution Correction-Estimation (DICE)-based methods have garnered substantial interest. They use a single marginal ratio to reweight rewards for each state-action pair, thereby achieving a relatively low estimation variance (Nachum et al., 2019b; Zhang et al., 2020; Lee et al., 2021). DICE provides a behavior-agnostic estimation of stationary distributions, presenting a more direct approach to handling the distribution mismatch. However, DICE-based techniques rely on an implicit assumption of the dataset's concentrability(Munos, 2007; Xie et al., 2021b; Li et al., 2022), otherwise the stationary distribution support mismatch between the dataset and policy can cause an arbitrarily large IS ratio, resulting in unstable training and poor performance. This problem can be significantly severe with insufficient data.

To address these challenges, we introduce a novel method, the Conservative Density Estimation (CDE), that integrates the strengths of both pessimism-based and DICE-based approaches. CDE employs the principles of conservative Q-learning (Kumar et al., 2020) in a unique way, incorporating pessimism within the stationary distribution space to achieve a theoretically-backed conservative occupation distribution. On the one hand, CDE does not rely on Bellman

[1] Carnegie Mellon University [2] Columbia University. Correspondence to: Zhepeng Cen <zcen@andrew.cmu.edu>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023. Copyright 2023 by the author(s).

update-style value estimation, favoring a direct behavior-policy-agnostic stationary distribution correction that improves performance in *sparse reward scenarios*. On the other hand, by constraining the density of the stationary distribution induced by OOD state-action pairs, CDE significantly enhances performance in *data-limited settings*. This stands in contrast to the significant performance degradation observed in baseline offline RL methods with diminishing dataset sizes, as CDE maintains high rewards even with only **1% trajectories** in challenging D4RL tasks (Fu et al., 2020).

1. We introduce the first approach to explicitly apply pessimism in the stationary distribution space. Notably, CDE outperforms value-learning-based approaches in sparse reward settings and demonstrates superior performance over DICE-based methods in handling scarce data situations.

2. We present a method that automatically bounds the concentrability coefficient without resorting to the common concentrability assumption (Rashidinejad et al., 2021; Shi et al., 2022; Ma et al., 2022; Zhan et al., 2022), underlining its robustness in managing the OOD extrapolation issue inherent in offline RL.

3. We demonstrate the resilience of CDE in maintaining high rewards even with significantly reduced dataset sizes, such as 1% of trajectories, while prior methods fail. Therefore, our method provides a viable solution for real-world applications where data can be scarce or costly to obtain.

2. Related Work

Offline RL with regularization or constraints. To mitigate OOD issues, Q-value-based methods are often enhanced with regularization or constraint terms (Levine et al., 2020; Prudencio et al., 2022). These techniques restrict the learned policy's deviation from the behavior policy in the dataset, whether through constrained policy spaces with explicit constraints (Fujimoto et al., 2019) or regularizers in the objective (Wu et al., 2019; Peng et al., 2019; Nair et al., 2020; Fujimoto & Gu, 2021). Particularly, the maximum mean discrepancy (MMD) constraint (Kumar et al., 2019) facilities to mitigate the support mismatch in policy space. Alternatively, value regularization is employed to yield lower estimates for unseen states or actions, resulting in conservative policies (Kostrikov et al., 2021a; Kumar et al., 2020). However, those methods may suffer instability from approximation error when learning value with Bellman update iteratively (Fujimoto et al., 2018; Fu et al., 2019; Brandfonbrener et al., 2021), often failing in sparse reward settings even with expert demonstrations. Meanwhile, the reliance on heuristic regularization can lead to overly conservative policies and degrade performance.

Offline RL with marginal importance sampling. The Distribution Correction-Estimation (DICE) method represents a class of approaches that directly address distribution shift using marginal importance sampling, offering reduced estimation variance compared to naive importance weighting (Precup, 2000). These methods reframe the learning objective as maximizing expected reward, using the duality between value-function linear programming and distribution optimization (Nachum et al., 2019b; Nachum & Dai, 2020). DICE calculates the importance ratio using either a forward method that minimizes the residual error of the Bellman equation (Zhang et al., 2020), or a backward method that optimizes the value function via duality (Nachum et al., 2019a). Some variations add a regularization to the objective function, yielding a closed-form solution for the importance ratio (Nachum et al., 2019b; Lee et al., 2021). Despite their ability to provide unbiased policy evaluation, DICE-style methods yield arbitrarily large importance ratios when the dataset lacks sufficient state-action space coverage, a challenge particularly acute in scarce data settings.

3. Method

3.1. Preliminaries

We formulate reinforcement learning problem in the context of a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, r, \gamma, \rho_0 \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ specifies the transition probability $T(s'|s,a)$, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, γ is the discount factor, and $\rho_0 : \mathcal{S} \to [0,1]$ is the initial state distribution. The policy $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ maps from a state to a distribution over actions. Given a policy π, consider the trajectory $\tau = \{s_0, a_0, s_1, a_1, \dots\}$ sampled by π, i.e., $s_0 \sim \rho_0, a_t \sim \pi(\cdot|a_t), s_{t+1} \sim T(\cdot|s_t, a_t)$, the stationary state-action distribution d^π is defined as

$$d^\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s, a_t = a). \qquad (1)$$

The goal of RL is to learn a return-maximization policy: $\pi^* = \arg\max_\pi \mathbb{E}_{\tau \sim \pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. This objective is be equivalent as the expectation of reward (Puterman, 2014): $\pi^* = \arg\max_\pi \mathbb{E}_{s,a \sim d^\pi}[r(s,a)]$.

In offline RL, the agent learns the policy from a pre-collected dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$. For simplicity, we denote the empirical state-action distribution of offline dataset as $d^\mathcal{D}$. The DICE-style methods apply marginal IS to estimate the expection of certain function $f(\cdot,\cdot)$: $\mathbb{E}_{s,a \sim d^\pi}[f(s,a)] = \mathbb{E}_{s,a \sim d^\mathcal{D}}[\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)} f(s,a)]$ with $d^\pi, d^\mathcal{D}$ as target and proposal distributions. The IS estimation can thus be approximated by sampling from offline dataset.

3.2. Conservative Density Estimation

In this section, we present Conservative Density Estimation (CDE), which aims to learns a policy that induces the distribution with conservative density in OOD state-action region. We first consider a f-divergence regularized policy optimization problem (Nachum et al., 2019a;b):

$$\max_{d^\pi \geq 0} \mathbb{E}_{d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^{\mathcal{D}}), \tag{2}$$

$$s.t. \sum_a d^\pi(s,a) = (1-\gamma)\rho_0(s) + \gamma \mathcal{T}_* d^\pi(s), \forall s, \tag{3}$$

where $D_f(d^\pi \| d^{\mathcal{D}}) = \mathbb{E}_{d^{\mathcal{D}}}[f(\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)})]$ is the f-divergence between two distributions, α is the hyperparameter of regularization, $\mathcal{T}_* d^\pi(s) = \sum_{s',a'} T(s|s',a') d^\pi(s',a')$ is the transposed transition operator. Here we adhere to state-wise Bellman flow constraint as (Lee et al., 2021) to incorporate the stochasticity of action distribution on next state $a' \sim \pi(\cdot|s')$ since the state-action-wise constraint can lead to overestimation for 'lucky' samples (Kostrikov et al., 2021b) and instability during training. Particularly, we have following assumption on f function selection:

Assumption 3.1. The f function in f-divergence is strictly convex and continuously differentiable, and $(f')^{-1}(x) \geq 0, \forall x \in \mathbb{R}$.

The previous DICE methods (Nachum et al., 2019b; Nachum & Dai, 2020; Lee et al., 2021) transform constrained optimization problem to unconstrained one in Eq.(3) by Lagrange or Fenchel-Rockafellar duality and evaluate the unconstrained objective by marginal IS with $d^{\mathcal{D}}$ as proposal distribution. However, one **implicit assumption** behind DICE methods is that the support of dataset distribution is large enough and otherwise the density of unseen state-actions in $d^{\mathcal{D}}$ can be zero or arbitrarily small. Therefore, when the support of d^π mismatches $d^{\mathcal{D}}$, there will be a large extrapolation error for OOD state-actions and variance in IS estimation. Meanwhile, the f-divergence regularization is enforced on the support of data distribution and approximated by single or several sample points, failing to serve as an effective supervision to explicitly reduce extrapolation errors for unseen state-actions.

To overcome the above issues, we consider a new constraint on the density of $d^\pi(s,a)$ by $\mu(s,a)$: $d^\pi(s,a) \leq \epsilon\mu(s,a), \forall s,a \in \text{supp}(\mu)$, where $\mu(s,a)$ is a distribution on OOD state-action pairs, i.e., $\text{supp}(\mu) \cap \text{supp}(d^{\mathcal{D}}) = \emptyset$. The new optimization problem is formulated as

$$\max_{d^\pi \geq 0} \mathbb{E}_{d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) \tag{4}$$

$$s.t. \sum_a d^\pi(s,a) = (1-\gamma)\rho_0(s) + \mathcal{T}_* d^\pi(s), \forall s \tag{5}$$

$$d^\pi(s,a) \leq \epsilon\mu(s,a), \forall s,a \in \text{supp}(\mu). \tag{6}$$

The corresponding unconstrained problem is $\max_{d^\pi} \min_{\lambda \geq 0, v} \mathcal{L}(d^\pi, v, \lambda)$, where

$$\begin{aligned} \mathcal{L}(d^\pi, v, \lambda) &= \mathbb{E}_{d^\pi}[A(s,a)] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] \\ &\quad - \alpha D_f(d^\pi \| d^{\mathcal{D}}) - \mathbb{E}_\mu\left[\lambda(s,a)\left(d^\pi/\mu(s,a) - \epsilon\right)\right] \end{aligned} \tag{7}$$

and $A(s,a) := r(s,a) + \gamma\mathbb{E}_{s' \sim T(\cdot|s,a)} v(s') - v(s)$ is regarded as advantage function if we interpret $v(s)$ as the V-value of state s. The derivation is attached in Appendix A.1.

In practice, we restrict the state marginal of μ to match the state distribution of dataset $d^{\mathcal{D}}(s)$ as previous OOD querying methods (Kumar et al., 2020; Kostrikov et al., 2021a; Lyu et al., 2022) and shrink the OOD region to unseen actions with existing states. Given a state s in dataset, suppose there exists n actions $a^{(1)}, \ldots, a^{(n)}$ such that $(s, a^{(i)}) \in \mathcal{D}, i = 1, \ldots, n$, we define the set of unseen actions as $\mathcal{A}_{\text{OOD}}(s) := \{a \big| \min_i \|a - a^{(i)}\|_\infty \geq \Delta a\}$. We further adopt a uniform distribution $\pi^\mu(a|s)$ over unseen action space $\mathcal{A}_{\text{OOD}}(s)$ as the policy of μ. Therefore, $\mu(s,a) = d^{\mathcal{D}}(s)\pi^\mu(a|s)$. We want to emphasize that although we constrain the support of μ to unseen actions with existing states, our method is still compatible with other OOD sampling distribution with proper inductive bias.

3.2.1. Policy Evaluation and Improvement

Based on the unconstrained objective in Eq.(7), we first adopt marginal IS to evaluate a policy given its stationary distribution. To avoid the support mismatch issue, we consider a new proposal distribution $\hat{d}^{\mathcal{D}}(s,a) := \zeta d^{\mathcal{D}}(s,a) + (1-\zeta)\mu(s,a)$ in importance sampling, where $\zeta \in (0,1)$ is the mixture coefficient. Therefore, the support of new proposal distribution can cover the target distribution d^π. We further replace the original f-divergence regularizer by $D_f(d^\pi \| \hat{d}^{\mathcal{D}})$ to constrain the density of both OOD and in-support state-actions. Besides, we substitute the importance ratio $w(s,a) = d^\pi(s,a)/\hat{d}^{\mathcal{D}}(s,a)$ for d^π as $\hat{d}^{\mathcal{D}}$ is fixed. The new objective function is

$$\begin{aligned} \mathcal{L}'(w, v, \lambda) &= \zeta\mathbb{E}_{d^{\mathcal{D}}}\left[w(s,a)A(s,a) - \alpha f(w(s,a))\right] \\ &\quad + (1-\zeta)\mathbb{E}_\mu\left[w(s,a)(A(s,a) - \lambda(s,a))\right. \\ &\quad \left. - \alpha f(w(s,a)) + \tilde{\epsilon}\lambda(s,a)\right] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)], \end{aligned} \tag{8}$$

where $\tilde{\epsilon} = \frac{\epsilon}{1-\zeta}$. The derivation is attached in Appendix A.1. With assumption 3.1, the objective in Eq.(4) is convex and thus is equivalent to the minimax problem $\min_{\lambda \geq 0, v} \max_{d^\pi} \mathcal{L}(d^\pi, v, \lambda)$ by Slater's condition. Moreover, the inner maximization has a closed-form solution (Nachum & Dai, 2020; Nachum et al., 2019b;a) and the outer minimization is a convex optimization problem. The proofs are in Appendix A.3A.4.

Proposition 3.2. *With assumption 3.1, the closed-form solution to inner maximization problem* $\max_{w \geq 0} \mathcal{L}'(w, v, \lambda)$ *is*

$$w^*(s,a) = (f')^{-1}(\tilde{A}(s,a)/\alpha), \tag{9}$$

where $\tilde{A}(s,a) := A(s,a) - \mathbf{1}\{(s,a) \in \textit{supp}(\mu)\} \cdot \lambda(s,a)$ *denotes* ***regularized advantage*** *function and* $\mathbf{1}\{\cdot\}$ *is the indicator function.*

Proposition 3.3. *The outer minimization problem* $\min_{\lambda \geq 0, v} \mathcal{L}'(w^*, v, \lambda)$ *is a convex optimization problem.*

Suppose the optimal solution is (λ^*, v^*)*, then* λ^* *has a closed-form solution*

$$\lambda^*(s,a) = \max\{0, A^*(s,a) - \alpha f'(\tilde{\epsilon})\}, \forall s, a \in supp(\mu), \tag{10}$$

where $A^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s,a)} v^*(s') - v^*(s)$*. The optimal regularized advantage is*

$$\tilde{A}^*(s,a) = \begin{cases} A^*(s,a), & (s,a) \in supp(d^{\mathcal{D}}) \\ \min\{\alpha f'(\tilde{\epsilon}), A^*(s,a)\}, & (s,a) \in supp(\mu) \end{cases} \tag{11}$$

Based on the closed-form relation between stationary distribution d^π and value function, we can thus improve the policy by maximizing w.r.t. value function. Since it requires the reward $r(s,a)$ and transition $T(\cdot|s,a)$ to compute regularized advantage function $\tilde{A}(s,a)$, which is available only for $(s,a) \in \mathcal{D}$, we consider function approximation for both V-value v and regularized advantage $\tilde{A}$ by parameters φ and ϕ. The optimization is in two steps: 1) We first optimize v_φ by minimizing the value of states in distribution:

$$\begin{aligned} \min_{\varphi} \mathbb{E}_{d^{\mathcal{D}}} [& w^*(s,a)(r(s,a) + \gamma \mathbb{E}_{s'} v_\varphi(s') - v_\varphi(s)) \\ & - \alpha f(w^*(s,a))] + (1-\gamma)\mathbb{E}_{s_0 \sim \rho_0}[v_\varphi(s_0)]. \end{aligned} \tag{12}$$

2) Then we regress the regularized advantage $\tilde{A}_\phi$ to the optimal $\tilde{A}^*$ in Eq.(11). Specifically, we regress the OOD advantages to $\alpha f'(\tilde{\epsilon})$ if they exceed it and regress the in-distribution advantages to the values from v_φ: $A_v(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} v_\varphi(s') - v_\varphi(s)$. In summary, we optimize the regularized advantage function by following mean squared error (MSE):

$$\begin{aligned} \min_{\phi} & \zeta \mathbb{E}_{d^{\mathcal{D}}}[(\tilde{A}_\phi(s,a) - A_\varphi(s,a))^2] + \\ & (1-\zeta)\mathbb{E}_\mu \left[\left(\max\{\tilde{A}_\phi(s,a) - \alpha f'(\tilde{\epsilon}), 0\} \right)^2 \right], \end{aligned} \tag{13}$$

and obtain the approximated optimal importance ratios for both in-distribution and OOD state-actions:

$$\tilde{w}^*(s,a) = (f')^{-1}(\tilde{A}_\phi(s,a)/\alpha). \tag{14}$$

By definition, the optimal distribution is $d^*(s,a) = \tilde{w}^*(s,a)\hat{d}^{\mathcal{D}}(s,a)$. In practice, we introduce another constraint to enforce $\sum_{s,a} d^*(s,a) = 1$ as (Zhang et al., 2020). See Appendix A.2 for full derivations.

3.2.2. Policy Extraction

Finally, we extract the policy from the learned importance ratios. Note that one policy is uniquely determined given its corresponding stationary distribution. Therefore, we extract the policy by minimizing the KL divergence between the stationary distribution of optimal policy and parameterized policy π_θ, where we approximate $d^{\pi_\theta}(s,a) \approx d^{\mathcal{D}}(s)\pi_\theta(a|s)$:

$$\min_\theta D_{\mathrm{KL}}[d^{\pi_\theta} \| d^*] \approx \mathbb{E}_{\substack{s \sim \mathcal{D} \\ a \sim \pi_\theta}} \left[\log \frac{d^{\mathcal{D}}(s)\pi_\theta(a|s)}{d^*(s,a)} \right] \tag{15}$$

$$= \mathbb{E}_{\substack{s \sim \mathcal{D} \\ a \sim \pi_\theta}}[-\log \tilde{w}^*(s,a)] + \mathbb{E}_{s \sim \mathcal{D}}[D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \hat{\pi}^{\mathcal{D}}(\cdot|s))] \tag{16}$$

where $\hat{\pi}^{\mathcal{D}}(a|s) = \zeta \pi^{\mathcal{D}}(a|s) + (1-\zeta)\pi^\mu(a|s)$ is the mixed behavior policy and $\pi^{\mathcal{D}}(a|s)$ denotes the empirical behavior policy. The mixed policy can be trained via weighted behavioral cloning from dataset and OOD sampling. We will analyze the error induced by state marginal approximation in Theorem 3.7. The final objective in Eq.(16) consists of two components: the maximizing of $\tilde{w}^*$ and minimizing the divergence with mixed behavior policy, indicating the trade-off between performance improvement by maximizing the value and conservative learning to reduce extrapolation error.

Algorithm 1 Conservative Density Estimation

```
Initialize value functions $v_\varphi$, $\tilde{A}_\phi$, mixed behavior policy $\hat{\pi}^{\mathcal{D}}$, policy $\pi_\theta$.
▷ policy evaluation and improvement
for training iteration $i$ do
    Sample batch $\{(s_i, a_i, r_i, s'_i)\}$ from $\mathcal{D}$ and $n$ OOD actions $\{a^{(1)}, \dots, a^{(n)}\}$ for each $s$;
    Update V-value $v_\varphi$ by Eq.(12);
    Update regularized advantage $\tilde{A}_\phi$ by Eq.(13);
    Update $\hat{\pi}^{\mathcal{D}}$ by weighted importance sampling.
end for
▷ policy extraction
for training iteration $j$ do
    Update policy $\pi_\theta$ by Eq.(16).
end for
```

The key steps of complete training procedure are summarized in Algo. 1. See Appendix B.2 for full algorithm and training details. One noteworthy difference from other actor-critic methods is that CDE updates the policy after the value function converges, which improves the learning stability and computation efficiency.

The advantages of CDE over previous DICE methods are two-fold: 1) the proposal distribution (i.e., $\hat{d}^{\mathcal{D}}$) has wider coverage than $d^{\mathcal{D}}$, which mitigates the support mismatch in IS and prevents the arbitrarily large importance ratio; 2) CDE produces a conservative estimation of density in OOD region. Compared to previous conservative methods, CDE determines the degree of conservatism precisely by optimal λ in Proposition 3.3, mitigating overly pessimistic estimation and loss of the generalization ability (Lyu et al., 2022). Furthermore, CDE disentangles two optimization steps, i.e., learning the value function by convex optimization and extracting the policy from the optimal importance ratio, thereby reducing the compounded error amplified by the interleaved optimization (Brandfonbrener et al., 2021).

3.3. Theoretical Analysis

CDE adopts a proposal distribution with broader support in marginal IS and explicitly constrains the stationary distribution density of the OOD region, resulting in a theoretical bound for the importance ratio, also known as concentrability coefficent (Munos, 2007; Rashidinejad et al., 2021).

Proposition 3.4 (Upper bound of concentrability ratio on OOD state-actions)**.** *With assumption 3.1, the theoretical optimal importance ratio is upper bounded by* $w^*(s,a) \le \tilde{\epsilon}, \forall (s,a) \in \mathit{supp}(\mu)$.

The proof of Proposition 3.4 is in Appendix A.5. It should be noted that an unbounded importance ratio can cause unstable training for importance-sampling-based methods (Shi et al., 2022). We further bound the function approximation $\tilde{w}^*$ in Eq.(14) with following continuity assumption:

Assumption 3.5 (Lipschitz continuity of $A_\phi(s,a)$)**.** There is a constant $L > 0$ such that $\forall a, a' \in \mathcal{A}_{\text{OOD}}(s), \forall s \in \mathcal{D}$,

$$|A_\phi(s,a) - A_\phi(s,a')| \le L\|a - a'\|_\infty .$$

Theorem 3.6 (Upper bound of function approximated concentrability ratio)**.** *Suppose that 1) the action space is* d*-dim, i.e.,* $\mathcal{A} \subset \mathbb{R}^d$*, 2) the diameter of* $\mathcal{A}$ *is* M*, i.e.,* $\|a_1 - a_2\|_\infty \le M, \forall a_1, a_2 \in \mathcal{A}$*, and 3) there are at least* N *OOD action samples from* μ *given any state* $s \in \mathcal{D}$*. When the continuity assumption 3.5 holds,* $\forall (s,a) \in \mathit{supp}(\mu)$*, with probability at least* $1 - \delta, \delta > 0$*, we have*

$$(f')\left(\tilde{w}^*(s,a)\right) \le f'(\epsilon) + \frac{\xi}{\alpha} + \frac{L}{\alpha}\left(\Delta a^d + \frac{M^d}{N}\log\frac{1}{\delta}\right)^{1/d}, \tag{17}$$

where ξ *is the maximum residual error of OOD regression in Eq.(13),* Δa *is the radius of in-distribution region as previously defined.*

The proof of Theorem 3.6 is in Appendix A.6. Notably, the upper bound shrinks with high probability as the number of OOD samples increases and it requires more samples for the same bound when the dimension of $\mathcal{A}$ increases.

Proposition 3.4 and Theorem 3.6 show that CDE inherently bounds the OOD concentrability coefficient. This coefficient is frequently assumed to be bounded in the variance/performance analysis in both off-policy evaluation and offline RL domains (Rashidinejad et al., 2021; Ma et al., 2022; Zhan et al., 2022), as an unbounded concentrability coefficient can introduce instability during training. As such, the CDE framework shows promise as a potential tool for reducing variance or establishing performance lower bounds in future research.

Meanwhile, CDE evaluates the policy within the stationary distribution space, enabling the computation of performance differences between policies based on the discrepancies in their respective stationary distributions. Consequently, we can establish the following bound on the performance gap between the learned and optimal policies.

Theorem 3.7 (The upper bound of performance gap)**.** *Suppose the maximum reward is* $R_{max} = \max_{s,a}\|r(s,a)\|$*, let* $V^\pi(\rho_0) := \mathbb{E}_{s_0\sim\rho_0}[V^\pi(s_0)]$ *denote the performance given a policy* π*. For policy* π *optimized by Eq.(16) and* N *transition data from* $d^{\mathcal{D}}$*, under mild assumptions, we have*

$$V^*(\rho_0) - V^\pi(\rho_0) \le \frac{4R_{max}}{1-\gamma} D_{\text{TV}}(d^{\mathcal{D}}(s)\|d^*(s)) + e_N$$

and e_N *converges in probability to zero at the rate* $N^{-\frac{1}{4+h}}, \forall h > 0$*, i.e.,* $N^{\frac{1}{4+h}} e_N \xrightarrow{N\to\infty} 0$ *in probability. Here,* $d^{\mathcal{D}}(s), d^*(s)$ *denote the state marginal of* $d^{\mathcal{D}}, d^*$*, and* $V^*(\rho_0)$ *denotes the performance of optimal policy.*

The full assumptions and proof are in Appendix A.7. The performance gap bound comprises two elements: 1) the discrepancy between the state distribution of the data and the optimal policy, and 2) the number of training samples. The first element stems from the state-marginal approximation in Eq.(15) during policy extraction. Importantly, this bound explicitly highlights two **crucial factors** influencing the final performance of the learned policy: the performance of behavior policy $\pi^{\mathcal{D}}$ and the size of the offline dataset. It provides a quantitative illustration of how the offline RL problem difficulty increases as the performance of behavior policy degrades and the dataset size decreases.

4. Experiment

In this section, we aim to study if CDE can truly combine the advantages of both pessimism-based methods and the DICE-based approaches. We are particularly interested in two main questions:

(1) Does CDE incorporate the strengths of the stationary-distribution correction training framework when handling ***sparse reward settings***?

(2) Can CDE's explicit density constraint effectively manage out-of-distribution (OOD) extrapolation issues in situations with ***insufficient datasets***?

Tasks. To answer these questions, we adopt 3 Maze2D datasets, 8 Adroit datasets, and 9 MuJoCo (random, medium, medium-expert) datasets from the D4RL benchmark (Fu et al., 2020). We use the normalized score as the evaluation metric. Note that the Maze2D and Adroit tasks are challenging due to their **sparse rewards**. To assess performance under **scarce data conditions**, we employ a random sampling strategy on the full datasets (details are provided in

Table 1: Normalized scores of CDE against other baselines. The scores are taken average over the evaluations of final 20% training steps with 3 seeds, where each evaluation is tested on 20 trajectories. In MuJoCo tasks, ”-r”, ”-m”, ”-m-e” represent ”-random”, ”-medium”, ”-medium-expert”. The versions of tasks are ”-v1” for Maze2D and Adroit and ”-v2” for MuJoCo. We **bold** the mean values that $\geq 0.95 *$ highest value.

Task	BC	BCQ	CQL	IQL	TD3+BC	AlgaeDICE	OptiDICE	CDE
maze2d-umaze	3.8	32.8	5.7	50.0	41.5	-15.7	111.0	**122.6±3.9**
maze2d-medium	30.3	20.7	5.0	31.0	76.3	10.0	145.2	**153.2±12.0**
maze2d-large	5.0	47.8	12.5	58.0	77.8	-0.1	155.7	**206.9±14.4**
halfcheetah-r	2.3	2.2	**18.6**	13.5	11.0	-0.3	2.5	2.4±1.4
walker2d-r	1.7	6.9	2.5	8.0	**8.5**	0.5	1.2	2.3±3.7
hopper-r	4.8	10.6	2.5	4.4	1.6	0.9	22.3	**26.9±5.8**
halfcheetah-m	42.6	45.7	**49.1**	46.8	**48.0**	-2.2	46.0	43.7±2.6
walker2d-m	75.3	73.9	**83.3**	77.4	**84.7**	1.2	67.5	70.1±5.9
hopper-m	52.9	53.3	**64.6**	50.3	57.7	0.3	47.1	50.1±4.1
halfcheetah-m-e	55.2	76.0	75.6	80.3	**84.9**	-0.8	72.5	76.6±8.9
walker2d-m-e	**107.5**	**109.8**	**109.5**	105.3	**110.3**	1.1	99.0	**107.5±16.1**
hopper-m-e	52.5	83.0	102.0	91.5	98.0	0.4	95.5	**107.3±4.0**
pen-human	63.9	68.9	37.5	71.5	2.0	-3.3	11.9	**72.1±15.8**
hammer-human	1.2	0.5	**4.4**	1.4	1.4	0.3	0.3	1.9±0.7
door-human	2.0	0.0	**9.9**	4.3	-0.3	0.0	0.1	7.7±3.3
relocate-human	0.1	-0.1	0.2	0.1	-0.3	-0.1	-0.1	**0.3±0.1**
pen-expert	85.1	**114.9**	107.0	**111.7**	79.1	-3.5	83.3	105.0±12.3
hammer-expert	**125.6**	107.2	86.7	116.3	3.1	0.3	**127.1**	**126.3±3.4**
door-expert	34.9	99.0	101.5	103.8	-0.3	0.0	**105.7**	**105.9±0.3**
relocate-expert	**101.3**	41.6	95.0	**102.7**	-1.5	-0.1	99.8	**102.6±1.9**
Total Score	848	994.7	973.1	1128.3	783.5	-11.1	1293.6	**1491.4**

Section 4.2). For all tasks, we use the latest versions of the datasets.

Baselines. We compare our CDE method with a collection of state-of-the-art offline RL baselines spanning different categories. These include: 1) behavior cloning (BC); 2) BCQ(Fujimoto et al., 2019) as a direct policy constraint method; 3) CQL(Kumar et al., 2020) as a value regularization method; 4) IQL(Kostrikov et al., 2021b) as an asymmetric Q-learning method; 5) TD3+BC(Fujimoto & Gu, 2021) as an implicit policy regularization method; 6) AlgaeDICE(Nachum et al., 2019b) as a policy-gradient-based DICE method; and 7) OptiDICE(Lee et al., 2021) as an in-sample (without OOD querying) DICE method. More details regarding baselines are available in Appendix B.

We use tanh-squashed Gaussian policy for CDE's policy π following SAC(Haarnoja et al., 2018) and tanh-squashed Gaussian mixture model for mixed empirical behavior policy $\hat{\pi}^{\mathcal{D}}$ to improve the expressivity for multi-modality of offline data from composite policies (e.g., medium-expert tasks in MuJoCo) or non-Markovian policies (e.g., Maze2D tasks). We choose soft-chi function $f_{\text{soft}-\chi^2}$ (Lee et al., 2021) in f-divergence:

$$f_{\text{soft}-\chi^2} = \begin{cases} x \log x - x + 1, & 0 < x < 1, \\ (x-1)^2/2, & x \geq 1. \end{cases} \tag{18}$$

and thus the $(f')^{-1}$ is equal to ELU function (Clevert et al., 2015), which satisfies Assumption 3.1 and also avoids the gradient vanishing problem for small values when computing importance ratios.

For consistent evaluation and fair comparison, we use the same set of hyper-parameters for experiments in the same task domain. We evaluate all methods every 1000 training steps and compute a mean value over 20 evaluated trajectories. The final scores are the mean of evaluation values in last 20% training steps. Full experimental details are included in Appendix B.

4.1. Main results on D4RL benchmark

Table 1 presents the normalized scores of our method and baselines. In dense-reward MuJoCo tasks, CDE obtains competitive performances compared to the best performance of prior methods in random and medium tasks, while achieving better performance than most baselines in medium-expert tasks, which validates the ability of CDE that extracts the optimal policy from mixed-level data. In the **sparse-reward** Maze2D and Adroit domains, where the agents are more vulnerable to the value function approximation error due to the sparsity of rewards, CDE consistently outperforms all other baselines in almost all tasks. Specifically, CDE achieves state-of-the-art performance in Maze2D domain by a large margin. The substantial improvement over standard-RL-based methods indicates that CDE can miti-

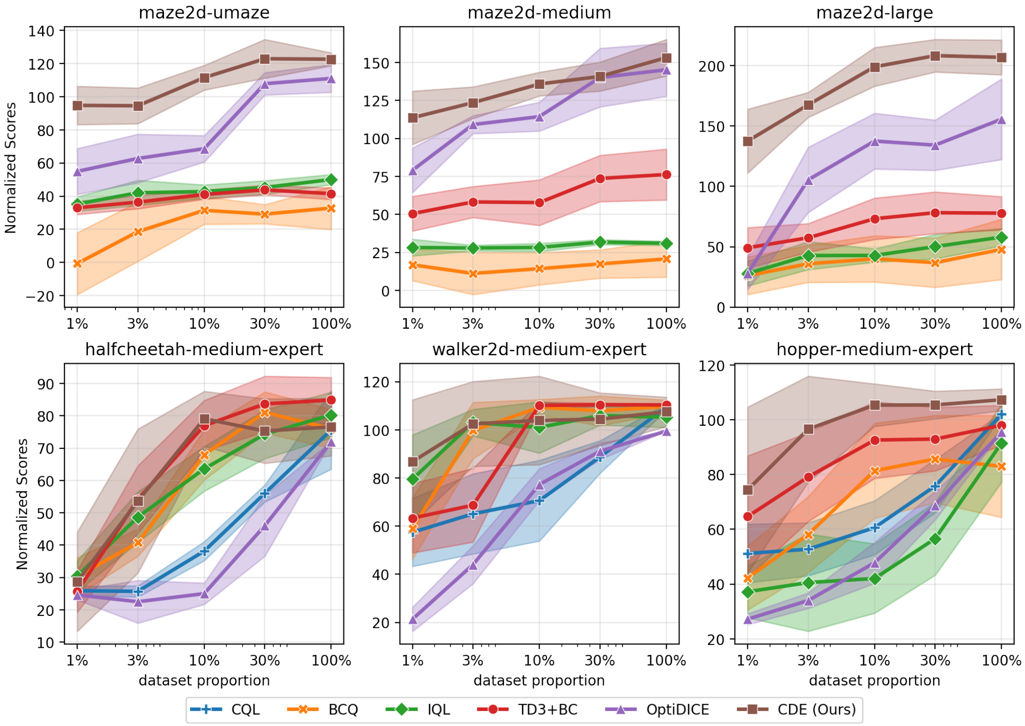


Figure 1: The results on sub-datasets with different dataset sizes.

gate compounded value estimation error by leveraging a closed-form optimal value solution to replace bootstrapping value update.

Another notable observation is that CDE exceeds both AlgaeDICE and OptiDICE in most tasks. AlgaeDICE falls short because it updates the policy via high-variance policy gradients, as opposed to extracting from optimal importance ratios. While OptiDICE exhibits comparable performance in tasks like Maze2D, it tends to overestimate unseen state-action pairs as it applies f-divergence regularization exclusively to in-distribution regions, thus leading to out-of-support issues. By introducing pessimism on the importance ratios of OOD regions, CDE mitigates potential overestimation, resulting in superior performance, especially in more challenging tasks like Maze2D ”-large” and Adroit ”-human”.

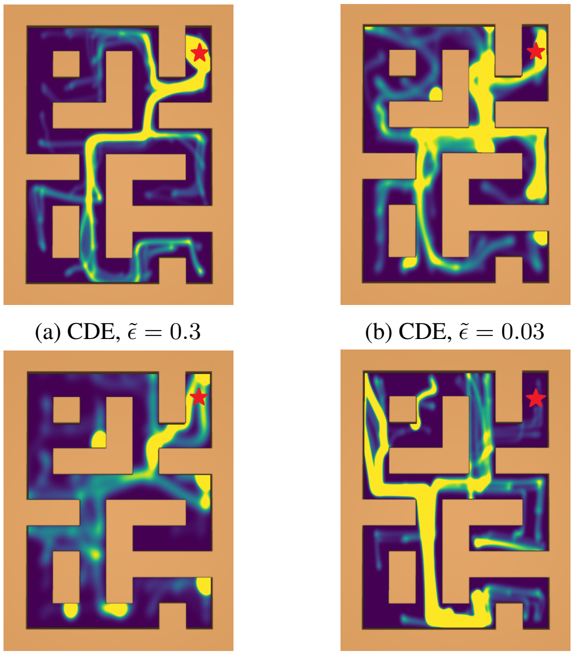
4.2. Comparative experiments on scarce data setting

To investigate further into situations where offline data is scarce and the state-action occupation of the empirical distribution is sparse, we examine the performance of different methods across datasets of varying sizes. In these circumstances, agents are likely to confront a more severe distribution shift problem, causing simple imitation learning to fail or experience a significant performance drop.

Given the extreme difficulty of the Adroit ”-human” tasks due to their high-dimensional space, narrow data distribution, and data scarcity (25 trajectories), we select Maze2D and MuJoCo tasks as our testing platforms. We randomly sample 1%, 3%, 10%, and 30% of trajectories from standard datasets to create our sub-datasets. Our chosen baselines include BCQ, CQL, IQL, TD3+BC, and OptiDICE. Due to the inferior performance of CQL, we exclude it from the Maze2D tasks experiments. The final results represent the average across 5 seeds, excluding the minimum and maximum values. The rest of the evaluation process aligns with the procedures used in the full dataset experiments.

The comparison results are shown in fig. 1, we defer the comparison of MuJoCo medium tasks to Appendix B. In the Maze2D domain, CDE consistently achieves the highest scores across all dataset sizes. In the MuJoCo domain, CDE significantly outperforms the baselines on the hopper tasks and exhibits comparable performance to other methods when the dataset size is relatively larger. For insufficient data settings, all methods demonstrate poorer performance in halfcheetah, while CDE maintains high scores and experiences less reduction in walker2d. Notably, OptiDICE displays less robustness against scarce data, undergoing a sharp performance drop despite achieving comparable performance in the full dataset setting. Moreover, considering the original Adroit ”-human” tasks already contain scarce data, OptiDICE also falls short, as shown in Table 1. This is because the narrow distribution of scarce data exacerbates the support mismatch problem in OptiDICE, leading to a significant bias in the importance-sampling-based off-policy evaluation. Conversely, CDE employs a mixed data policy, successfully mitigating large distribution shifts between the stationary distribution support of the dataset and the learned policy.

These results underscore the effectiveness of integrating conservatism into density estimation as implemented in CDE. However, it prompts the question: given that conservative value function estimation is also prevalent in standard-RL-based offline RL methods (e.g., CQL), why do they underperform in insufficient data settings? To answer this question, we delve deeper into the relationship between the performance of CDE and the degree of conservativeness.


(a) CDE, $\tilde{\epsilon} = 0.3$

(b) CDE, $\tilde{\epsilon} = 0.03$

(c) CDE, $\tilde{\epsilon} = 0.003$

(d) CQL

Figure 2: The heatmaps of agents with different levels of conservatism in maze2d-large environment. Yellow denotes the high occupation probability. The average normalized scores of four policies are 221, 150, 46 and 12. The starting point of each trajectory may vary but the destination is the same, i.e., the red star in the figures. Smaller $\tilde{\epsilon}$ indicates more conservative policy. The yellow accumulation points except the destination indicate that the agent is stuck at those regions.

4.3. Ablation study on the level of conservatism

We select the maze2d-large environment as our experimental platform as the agent's position directly represents the stationary distribution of the learned policy. We contrast CDE and CQL, two representative categories of pessimism augmentation in offline learning. According to Theorem 3.6, the theoretical OOD importance ratio is upper bounded by $\alpha f'(\tilde{\epsilon})$. Hence, we modify the degree of conservatism by altering $\tilde{\epsilon}$; a smaller $\tilde{\epsilon}$ imposes stricter constraints on density, resulting in more conservative policies.

The heatmap in fig.2, which represents the stationary distribution of position based on 100 trajectories, reveals that the probability mass at the midway and starting points increases as the level of conservatism escalates. In particular, in fig.2c, the agent is trapped at yellow points due to the overly strict constraint. Although stronger regularization can reduce OOD extrapolation error, excessive conservatism can harm generalization and lead to significant performance degradation. Compared to CDE, CQL policy is less likely to be trapped in single points but still struggles to reach the destination. There are two main reasons: 1) the sparse reward setting makes it challenging to estimate the Q value accurately by Bellman bootstrapping; 2) CQL applies value regularization by a distribution that evolves along the policy update, which is less stable and harder to adjust the level of conservation. In contrast, the CDE employs the closed-form relation between value and density to explicitly constrain the stationary distribution space, allowing for more precise control over the level of conservatism.

5. Conclusion

In this work, we propose CDE, a new offline RL approach, derived from the perspective of stationary state-action occupation. CDE applies the pessimism mechanism on stationary distribution and enjoys the benefits from both fields. CDE evaluates the policy in a behavior-agnostic manner, estimating the state-action density with proper conservatism, which makes it perform advantageously in sparse reward and scarce data settings. We further provide the theoretical analysis for the importance-sampling ratios and performance of CDE. Extensive experimental results demonstrated remarkable improvements over previous baselines in challenging tasks, highlighting its practical potential for real-world applications.

The major limitations include that CDE requires the strict alignment of initial state distribution in offline data and online environments, which restricts its performance in inconsistent settings. One potentially negative impact of CDE is that there will be a risk of algorithmic bias if the offline data used for training is not representative, leading to unfair outcomes. Therefore, it's crucial that these technologies are developed ethically and thoughtfully to mitigate potential negative consequences.

References

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Brandfonbrener, D., Whitney, W., Ranganath, R., and Bruna, J. Offline rl without off-policy evaluation. *Advances in neural information processing systems*, 34:4933–4946, 2021.

Cheng, G. and Huang, J. Z. Bootstrap consistency for general semiparametric m-estimation. 2010.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Fu, J., Kumar, A., Soh, M., and Levine, S. Diagnosing bottlenecks in deep q-learning algorithms. In *International Conference on Machine Learning*, pp. 2021–2030. PMLR, 2019.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., Yang, Y., and Knoll, A. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Hong, Z.-W., Agrawal, P., des Combes, R. T., and Laroche, R. Harnessing mixed offline reinforcement learning datasets via trajectory weighting. In *The Eleventh International Conference on Learning Representations*.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23 (6):4909–4926, 2021.

Kostrikov, I., Fergus, R., Tompson, J., and Nachum, O. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pp. 5774–5783. PMLR, 2021a.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021b.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Luenberger, D. G. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Lyu, J., Ma, X., Li, X., and Lu, Z. Mildly conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2206.04745*, 2022.

Ma, S. and Kosorok, M. R. Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217, 2005.

Ma, Y. J., Yan, J., Jayaraman, D., and Bastani, O. How far i'll go: Offline goal-conditioned reinforcement learning via f-advantage regression. *arXiv preprint arXiv:2206.03023*, 2022.

Munos, R. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Nachum, O. and Dai, B. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019a.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Nair, A., Gupta, A., Dalal, M., and Levine, S. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. gendice regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Prudencio, R. F., Maximo, M. R., and Colombini, E. L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *arXiv preprint arXiv:2203.01387*, 2022.

Prudencio, R. F., Maximo, M. R., and Colombini, E. L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Seno, T. and Imai, M. d3rlpy: An offline deep reinforcement learning library. *The Journal of Machine Learning Research*, 23(1):14205–14224, 2022.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pp. 19967–20025. PMLR, 2022.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Xu, T., Li, Z., and Yu, Y. Error bounds of imitating policies and environments. *Advances in Neural Information Processing Systems*, 33:15737–15749, 2020.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

A. Supplementary Derivations and Proofs

A.1. Derivation of Eq.(7)(8)

For Eq.(7), the Lagrangian for equation 4-6 is

$$\max_{d^\pi} \min_{\lambda \geq 0, v} \mathcal{L}(d^\pi, v, \lambda) := \mathbb{E}_{\substack{(s,a)\sim d^\pi \\ s'\sim T(\cdot|s,a)}} [r(s,a)] - \sum_{s,a \in \text{supp}(\mu)} \lambda(s,a)[d^\pi(s,a) - \epsilon\mu(s,a)] + \tag{19}$$

$$\sum_s v(s)[(1-\gamma)\rho_0(s) + \gamma \mathcal{T}_* d^\pi(s) - \sum_a d^\pi(s,a)] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) \tag{20}$$

$$= \mathbb{E}_{d^\pi}[r(s,a)] - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) \tag{21}$$

$$+ (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] + \sum_{\bar{s},\bar{a}} v(s) T(s|\bar{s},\bar{a}) d^\pi(\bar{s},\bar{a}) - \mathbb{E}_{d^\pi}[v(s)] \tag{22}$$

$$= \mathbb{E}_{d^\pi}[r(s,a)] - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) \tag{23}$$

$$+ (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] + \sum_{s,a} v(s') T(s'|s,a) d^\pi(s,a) - \mathbb{E}_{d^\pi}[v(s)] \tag{24}$$

$$= \mathbb{E}_{d^\pi}[r(s,a)] - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) \tag{25}$$

$$+ (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] + \mathbb{E}_{s,a\sim d^\pi, s'\sim T(\cdot|s,a)}[v(s')] - \mathbb{E}_{d^\pi}[v(s)] \tag{26}$$

$$= \mathbb{E}_{d^\pi}\left[r(s,a) + \mathbb{E}_{s'\sim T(\cdot|s,a)}[v(s')] - v(s)\right] - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] \tag{27}$$

$$- \alpha D_f(d^\pi \| d^{\mathcal{D}}) + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] \tag{28}$$

$$= \mathbb{E}_{d^\pi}[A(s,a)] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] \tag{29}$$

For Eq.(8), we have

$$\mathcal{L}'(w, v, \lambda) = \mathbb{E}_{d^\pi}[A(s,a)] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] - \alpha D_f(d^\pi \| \hat{d}^{\mathcal{D}}) - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] \tag{30}$$

$$= \mathbb{E}_{\hat{d}^{\mathcal{D}}}\left[\frac{d^\pi}{\hat{d}^{\mathcal{D}}} A(s,a) - \alpha f\left(\frac{d^\pi}{\hat{d}^{\mathcal{D}}}\right)\right] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] \tag{31}$$

$$= \mathbb{E}_{\hat{d}^{\mathcal{D}}}\left[w(s,a)A(s,a) - \alpha f(w(s,a))\right] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] - \mathbb{E}_\mu \left[\lambda(s,a)\left(\frac{d^\pi}{\mu} - \epsilon\right)\right] \tag{32}$$

$$= \zeta \mathbb{E}_{d^{\mathcal{D}}}\left[w(s,a)A(s,a) - \alpha f(w(s,a))\right] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] \tag{33}$$

$$+ (1-\zeta)\mathbb{E}_\mu\left[w(s,a)(A(s,a) - \lambda(s,a)) - \alpha f(w(s,a)) + \tilde{\epsilon}\lambda(s,a)\right], \tag{34}$$

The above derivations lead to the Eq.(7)(8) in the main content.

A.2. Derivation of normalization for stationary distribution

In practice, the optimal distribution d^* may not satisfy $\sum_{s,a} d^*(s,a) = 1$ due to function approximation error. Therefore, we explicitly enforce the $\sum_{s,a} d^*(s,a) = 1$ (Zhang et al., 2020) to make d^* a valid distribution, which is equivalent to $\mathbb{E}_{\hat{d}^{\mathcal{D}}} w^*(s,a) = 1$.

With new normalization constraint, the corresponding unconstrained problem becomes

$$\min_{\lambda \geq 0, v, \eta} \max_w \mathcal{L}(w; v, \lambda, \eta) := \zeta \mathbb{E}_{d^{\mathcal{D}}}\left[w(s,a)A(s,a) - \alpha f(w(s,a))\right] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] \tag{35}$$

$$+ (1-\zeta)\mathbb{E}_\mu\left[w(s,a)(A(s,a) - \lambda(s,a)) - \alpha f(w(s,a)) + \tilde{\epsilon}\lambda(s,a)\right] + \eta(1 - \mathbb{E}_{\hat{d}^{\mathcal{D}}} w^*(s,a)) \tag{36}$$

$$= \zeta \mathbb{E}_{d^{\mathcal{D}}}\left[w(s,a)(A(s,a) - \eta) - \alpha f(w(s,a))\right] + (1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] \tag{37}$$

$$+ (1-\zeta)\mathbb{E}_\mu\left[w(s,a)(A(s,a) - \lambda(s,a) - \eta) - \alpha f(w(s,a)) + \tilde{\epsilon}\lambda(s,a)\right] + \eta, \tag{38}$$

where η is the dual variable of the normalization constraint. Therefore, we only need to replace $\tilde{A}$ by $\tilde{A} - \eta$ for optimization w.r.t. v_φ and π_θ. Meanwhile, we will also update η by gradient descent. See more details in full algorithm in Appendix B.2.4.

A.3. Proof for Proposition 3.2

Let $\frac{\partial \mathcal{L}'(w,v,\lambda)}{\partial w} = 0$ and we have

$$\zeta \mathbb{E}_{d^{\mathcal{D}}}[A(s,a) - \alpha f'(w(s,a))] + (1-\zeta)\mathbb{E}_{d^{\mathcal{D}}}[A(s,a) - \lambda(s,a) - \alpha f'(w(s,a))] = 0 \tag{39}$$

Separate state-action space $\mathcal{S} \times \mathcal{A}$ into the support of $d^{\mathcal{D}}$ and μ, then we can get the solution as Eq.(9). Meanwhile, $w^* \geq 0$ always holds by assumption 3.1. Therefore, the solution is valid and is exactly the optimal solution.

The closed-form solution to optimal importance ratio can also be derived by Fenchel-Rockafellar dual form of f-divergence (Nachum et al., 2019b; Nachum & Dai, 2020), which leads to the same results.

A.4. Proof for Proposition 3.3

Notice that the convexity of dual function, which corresponds to $g(v,\lambda) := \max_w \mathcal{L}'(w,v,\lambda)$ in our setting, is proved by previous literature (Proposition 1, section 8.3 in (Luenberger, 1997)).

Then we prove the closed form of optimal λ^*. Consider the partial differential of $\mathcal{L}'(w^*,v,\lambda)$ w.r.t λ:

$$\begin{aligned}
\frac{\partial \mathcal{L}'(w^*,v,\lambda)}{\partial \lambda} =& \frac{\partial}{\partial \lambda}(1-\zeta)\mathbb{E}_\mu\left[(f')^{-1}\left(\frac{A-\lambda}{\alpha}\right)(A-\lambda) - \alpha f\left((f')^{-1}\left(\frac{A-\lambda}{\alpha}\right)\right) + \lambda\tilde{\epsilon}\right] && (40)\\
=&(1-\zeta)\mathbb{E}_\mu\left[((f')^{-1})'\left(\frac{A-\lambda}{\alpha}\right)\left(-\frac{1}{\alpha}\right)(A-\lambda) - (f')^{-1}\left(\frac{A-\lambda}{\alpha}\right)\right. && (41)\\
&\left. -\alpha f'\left((f')^{-1}\left(\frac{A-\lambda}{\alpha}\right)\right)((f')^{-1})'\left(\frac{A-\lambda}{\alpha}\right)\left(-\frac{1}{\alpha}\right) + \tilde{\epsilon}\right] && (42)\\
=&(1-\zeta)\mathbb{E}_\mu\left[-((f')^{-1})'\left(\frac{A-\lambda}{\alpha}\right)\frac{A-\lambda}{\alpha} - (f')^{-1}\left(\frac{A-\lambda}{\alpha}\right)\right. && (43)\\
&\left. +((f')^{-1})'\left(\frac{A-\lambda}{\alpha}\right)\frac{A-\lambda}{\alpha} + \tilde{\epsilon}\right] && (44)\\
=&(1-\zeta)\mathbb{E}_\mu\left[-(f')^{-1}\left(\frac{A-\lambda}{\alpha}\right) + \tilde{\epsilon}\right] && (45)
\end{aligned}$$

We omit (s,a) for A and λ functions for brevity. By assumption 3.1, f is convex and f' is monotonic increasing. Therefore, when $A(s,a) \leq \alpha f'(\tilde{\epsilon})$, the gradient of λ is always non-negative for $\lambda \geq 0$; otherwise, the gradient equals to zero when $\lambda = A(s,a) - \alpha f'(\tilde{\epsilon})$.

Therefore, the optimal solution of λ is

$$\lambda^*(s,a) = \max\{0, A(s,a) - \alpha f'(\tilde{\epsilon})\}. \tag{46}$$

Plug-in the λ^* to Proposition 3.2 and then we can get the optimal regularized advantage function $\tilde{A}^*$.

A.5. Proof for Proposition 3.4

Combine equation 9 and equation 10,

$$\begin{aligned}
w^*(s,a) :=& (f')^{-1}\left(\frac{A(s,a) - \lambda^*(s,a)}{\alpha}\right) && (47)\\
=& (f')^{-1}\left(\frac{A(s,a) - \max\{0, A(s,a) - \alpha f'(\tilde{\epsilon})\}}{\alpha}\right) && (48)\\
=& (f')^{-1}\left(\min\left\{\frac{A(s,a)}{\alpha}, f'(\tilde{\epsilon})\right\}\right) && (49)
\end{aligned}$$

By Assumption 3.1, f' is strictly increasing and so is $(f')^{-1}$. As a result,

$$w^*(s,a) = (f')^{-1}\left(\min\left\{\frac{A(s,a)}{\alpha}, f'(\tilde{\epsilon})\right\}\right) = \min\{(f')^{-1}(A(s,a)/\alpha), \tilde{\epsilon}\} \tag{50}$$

A.6. Proof for Theorem 3.6

We first give the following lemma.

Lemma A.1. *Suppose that 1) the action space is d-dim, i.e., $\mathcal{A} \subset \mathbb{R}^d$, 2) the diameter of $\mathcal{A}$ is M, i.e., $\|a_1 - a_2\|_\infty \leq M, \forall a_1, a_2 \in \mathcal{A}$, and 3) there are N action samples from μ given any state $s \in \mathcal{D}$, denoted by $(s, a_1), \ldots, (s, a_N)$, and μ is a uniform distribution over OOD action space. Let $\delta > 0$, $(s,a) \in \mathcal{D}$, $\tilde{a} \in \mathcal{A}_{OOD}(s)$. We have*

$$\mathbb{P}\left(\min_{i=1,\ldots,N} \|\tilde{a} - a_i\|_\infty > \delta\right) \leq \left(1 - \frac{\delta^d - \Delta a^d}{M^d}\right)^N \tag{51}$$

Proof. Let $B_\infty(x,y) = \{x' \in \mathbb{R}^d : \|x - x'\|_\infty \leq y\}$ denote the d-dim Euclidean Ball under $\|\cdot\|_\infty$. The volume of $B_\infty(x,y)$ is then given by $\text{Vol}(B_\infty(x,y)) = 2^d y^d$. We have

$$\mathbb{P}(\|\tilde{a} - a_1\|_\infty > \delta) = 1 - \mathbb{P}(\|\tilde{a} - a_1\|_\infty \leq \delta) = 1 - \mathbb{P}(a_1 \in B_\infty(\tilde{a}, \delta)) \tag{52}$$

Recall that $(s, a_1), \ldots, (s, a_N)$ are $i.i.d.$ samples from uniform distribution on $\mathcal{A} \backslash B_\infty(a, \Delta a)$. Thus, we can establish the following equality

$$\begin{aligned}
\mathbb{P}(a_1 \in B_\infty(\tilde{a}, \delta)) &= \int_{\mathbb{R}^d} \mathbf{1}\{x \in B_\infty(\tilde{a}, \delta)\}\mu(x)dx && (53)\\
&= \int_{\mathbb{R}^d} \mathbf{1}\{x \in B_\infty(\tilde{a}, \delta)\}\frac{\mathbf{1}\{x \in \mathcal{A} \backslash B_\infty(a, \Delta a)\}}{\text{Vol}(\mathcal{A} \backslash B_\infty(a, \Delta a))}dx && (54)\\
&= \frac{1}{\text{Vol}(\mathcal{A} \backslash B_\infty(a, \Delta a))}\int_{\mathbb{R}^d} \mathbf{1}\{x \in B_\infty(\tilde{a}, \delta) \cap \mathcal{A} \backslash B_\infty(a, \Delta a)\}dx && (55)\\
&= \frac{\text{Vol}(B_\infty(\tilde{a}, \delta) \cap \mathcal{A} \backslash B_\infty(a, \Delta a))}{\text{Vol}(\mathcal{A} \backslash B_\infty(a, \Delta a))} && (56)
\end{aligned}$$

Since the action space $\mathcal{A}$ is bounded with radius M,

$$\text{Vol}(\mathcal{A} \backslash B_\infty(a, \Delta a)) \leq \text{Vol}(B_\infty(a, M)) \tag{57}$$

In addition, notice that

$$\text{Vol}(B_\infty(\tilde{a}, \delta) \cap \mathcal{A} \backslash B_\infty(a, \Delta a)) \geq \text{Vol}(B_\infty(\tilde{a}, \delta)) - \text{Vol}(B_\infty(a, \Delta a)) \tag{58}$$

Combining the above inequalities and plugging in the formula for d-dim ball under $\|\cdot\|_\infty$, we have

$$\begin{aligned}
&\mathbb{P}(\|\tilde{a} - a_1\|_\infty > \delta) && (59)\\
=&1 - \mathbb{P}(a_1 \in B_\infty(\tilde{a}, \delta)) && (60)\\
\leq&1 - \frac{\text{Vol}(B_\infty(\tilde{a}, \delta)) - \text{Vol}(B_\infty(a, \Delta a))}{\text{Vol}(B_\infty(a, M))} && (61)\\
=&1 - \frac{\delta^d - \Delta a^d}{M^d} && (62)
\end{aligned}$$

By independence between the OOD samples

$$\begin{aligned}
&\mathbb{P}\left(\min_{i=1,\ldots,N} \|\tilde{a} - a_i\|_\infty > \delta\right) && (63)\\
=&\mathbb{P}\left(\bigcap_{i=1}^N \{\|\tilde{a} - a_i\|_\infty > \delta\}\right) = \mathbb{P}(\|\tilde{a} - a_1\|_\infty > \delta)^N \leq \left(1 - \frac{\delta^d - \Delta a^d}{M^d}\right)^N && (64)
\end{aligned}$$

This finishes the proof. □

As a remark, if we consider $\|\cdot\|_p$ instead of $\|\cdot\|_\infty$, the result would still be the same. Now we give the proof of Theorem 3.6.

Proof. Let $(s, a) \in \mathcal{D}$ and suppose that $(s, a_1), \ldots, (s, a_N)$ are the $i.i.d.$ samples from μ. Let $a' \in \{a_1, \ldots, a_N\}$ be the OOD sample that is closest to a under $\|\cdot\|_\infty$ (i.e., $a' = \arg\min_{x \in \{a_1, \ldots, a_N\}} \|x - a\|_\infty$). Since the maximum regression residual error is ξ, we have

$$\tilde{A}_\phi(s, a') \leq \alpha f'(\tilde{\epsilon}) + \xi. \tag{65}$$

Then, by assumption 3.5, we have

$$\tilde{A}_\phi(s, a) \leq \tilde{A}_\phi(s, a') + |\tilde{A}_\phi(s, a) - \tilde{A}_\phi(s, a')| \leq \alpha f'(\tilde{\epsilon}) + L \cdot \|a - a'\|_\infty + \xi \tag{66}$$

Let $\tilde{\delta} > 0$ and $\delta' = \frac{\alpha}{L}(f'(\tilde{\epsilon} + \tilde{\delta}) - f'(\tilde{\epsilon}) - \frac{\xi}{\alpha})$. Suppose $\delta' > 0$, by Lemma A.1, and using the fact that $1 + x \leq e^x, \ \forall x \in \mathbb{R}$, we have

$$\mathbb{P}(\|a' - a\|_\infty \leq \delta') \geq 1 - \left(1 - \frac{\delta'^d - \Delta a^d}{M^d}\right)^N \geq 1 - e^{-N \frac{\delta'^d - \Delta a^d}{M^d}} \tag{67}$$

Combine equation 66 and equation 67, we have, with probability at least $1 - e^{-N \frac{\delta'^d - \Delta a^d}{M^d}}$,

$$\tilde{A}_\phi(s, a) \leq \alpha f'(\tilde{\epsilon}) + L\delta' + \xi = \alpha f'(\tilde{\epsilon} + \tilde{\delta}) \tag{68}$$

Recall that $\tilde{w}^*(s, a) := (f')^{-1}(\tilde{A}_\phi(s, a)/\alpha)$. By equation 68, we have

$$\tilde{w}^*(s, a) := (f')^{-1}(\tilde{A}_\phi(s, a)/\alpha) \leq (f')^{-1}(f'(\tilde{\epsilon} + \tilde{\delta})) = \tilde{\epsilon} + \tilde{\delta} \tag{69}$$

with probability at least $1 - e^{-N \frac{\delta'^d - \Delta a^d}{M^d}}$, where $\delta' = \frac{\alpha}{L}(f'(\tilde{\epsilon} + \tilde{\delta}) - f'(\tilde{\epsilon}) - \frac{\xi}{\alpha})$. The inequality step in equation 69 follows from the fact that f' is increasing.

Let $\delta \in (0, 1)$. Consider $\tilde{\delta} = (f')^{-1}(f'(\tilde{\epsilon}) + \frac{\xi}{\alpha} + \frac{L}{\alpha}(\Delta a^d + \frac{M^d}{N} \log \frac{1}{\delta})^{\frac{1}{d}}) - \tilde{\epsilon}$. First, we verify $\delta' > 0$ with this choice of $\tilde{\delta}$.

$$\delta' := \frac{\alpha}{L}\left(f'(\tilde{\epsilon} + \tilde{\delta}) - f'(\tilde{\epsilon}) - \frac{\xi}{\alpha}\right) \tag{70}$$

$$= \frac{\alpha}{L}\left(f'(\tilde{\epsilon}) + \frac{\xi}{\alpha} + \frac{L}{\alpha}\left(\Delta a^d + \frac{M^d}{N} \log \frac{1}{\delta}\right)^{\frac{1}{d}} - f'(\tilde{\epsilon}) - \frac{\xi}{\alpha}\right) \tag{71}$$

$$= \left(\Delta a^d + \frac{M^d}{N} \log \frac{1}{\delta}\right)^{\frac{1}{d}} \tag{72}$$

$$> 0 \tag{73}$$

Substitute $\tilde{\delta}$ back into equation 69. We get

$$w^*(s, a) \leq \tilde{\epsilon} + (f')^{-1}\left(f'(\tilde{\epsilon}) + \frac{\xi}{\alpha} + \frac{L}{\alpha}\left(\Delta a^d + \frac{M^d}{N} \log \frac{1}{\delta}\right)^{\frac{1}{d}}\right) - \tilde{\epsilon} \tag{74}$$

$$= (f')^{-1}\left(f'(\tilde{\epsilon}) + \frac{\xi}{\alpha} + \frac{L}{\alpha}\left(\Delta a^d + \frac{M^d}{N} \log \frac{1}{\delta}\right)^{\frac{1}{d}}\right) \tag{75}$$

with probability of at least

$$
\begin{aligned}
&1-e^{-N\frac{\delta'^d-\Delta a^d}{M^d}} \tag{76}\\
=&1-\exp\left(-N\frac{(\frac{\alpha}{L}(f'(\tilde{\epsilon}+\tilde{\delta})-f'(\tilde{\epsilon})-\frac{\xi}{\alpha}))^d-\Delta a^d}{M^d}\right) \tag{77}\\
=&1-\exp\left(-N\frac{(\frac{\alpha}{L}(f'(\tilde{\epsilon})+\frac{\xi}{\alpha}+\frac{L}{\alpha}(\Delta a^d+\frac{M^d}{N}\log\frac{1}{\delta})^{\frac{1}{d}}-f'(\tilde{\epsilon})-\frac{\xi}{\alpha}))^d-\Delta a^d}{M^d}\right) \tag{78}\\
=&1-\exp\left(-N\frac{\Delta a^d+\frac{M^d}{N}\log\frac{1}{\delta}-\Delta a^d}{M^d}\right) \tag{79}\\
=&1-\delta \tag{80}
\end{aligned}
$$

This finishes the proof of Theorem 3.6. □

A.7. Proof of Theorem 3.7

In this section, we consider the performance of our policy as the sample size N grows.

Let $d^{\mathcal{D}}$ denote the data distribution from which $\mathcal{D}$ is obtained. Thus $\mathcal{D}$ can be viewed as N $i.i.d.$ samples from $d^{\mathcal{D}}$. In this section, we use the notation with subscript $\mathcal{D}_N$ to denote $\mathcal{D}$ to address the number of data and avoid ambiguity.

Recall that π_θ minimizes the following objective

$$\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}(d^{\mathcal{D}}(s_i)\pi_\theta(\cdot|s_i)\|d^*(s_i,\cdot)) \tag{81}$$

which is the empirical version of the following expectation

$$\mathbb{E}_{s\sim d^{\mathcal{D}}}[D_{\mathrm{KL}}(d^{\mathcal{D}}(s_i)\pi_\theta(\cdot|s_i)\|d^*(s_i,\cdot))] \tag{82}$$

We make following assumptions:

Assumption A.2. Denote the space of parameter θ in the policy extraction step by Θ. Let $g_\theta(s) := D_{\mathrm{KL}}(\pi_\theta(\cdot|s)\|\pi^*(\cdot|s))$, where $d^*(s)$ denotes the state marginal of d^*. Then, the function class $\mathcal{F} = \{g_\theta(\cdot) : \mathcal{S} \to \mathbb{R} | \theta \in \Theta\}$ is $d^{\mathcal{D}}$-Donsker. And $\mathrm{Var}_{s\sim d^{\mathcal{D}}(s)}(g_\theta(s)) < \infty$ for all $\theta \in \Theta$.

Assumption A.2 guarantees the consistency of θ trained with dataset $\mathcal{D}_N$, which is a common assumption when considering training with finite samples (Van der Vaart, 2000; Geer, 2000; Ma & Kosorok, 2005; Cheng & Huang, 2010). A sufficient condition for Assumption A.2 is Θ being bounded, together with a Lipschitz-type condition on $\mathcal{F}$ (Van der Vaart, 2000).

Assumption A.3. Suppose the policy extracted from Eq.(16) is π, define the state marginal of $d^{\mathcal{D}}, d^\pi, d^*$ as $d^{\mathcal{D}}(s), d^\pi(s), d^*(s)$, then

$$D_{\mathrm{TV}}(d^\pi(s)\|d^{\mathcal{D}}(s)) \le D_{\mathrm{TV}}(d^*(s)\|d^{\mathcal{D}}(s)) \tag{83}$$

The Assumption A.3 holds in general because empirically, the performance of learned policy π is in between $\pi^{\mathcal{D}}$ and π^*, indicating that the stationary state distribution of learned policy d^π is closer to dataset distribution than the optimal state distribution.

Then we introduce the following lemma based on Lemma 6 in (Xu et al., 2020):

Lemma A.4. *Suppose the maximum reward is $R_{max} = \max_{s,a}\|r(s,a)\|$, $V^\pi(\rho_0) := \mathbb{E}_{s_0\sim\rho_0}[V^\pi(s_0)]$ denote the performance given a policy π, then with assumption A.3,*

$$|V^\pi(\rho_0) - V^*(\rho_0)| \le \frac{4R_{max}}{1-\gamma} D_{\mathrm{TV}}(d^*(s)\|d^{\mathcal{D}}(s)) + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{d^{\mathcal{D}}(s)}[D_{\mathrm{TV}}(\pi(\cdot|s)\|\pi^*(\cdot|s))], \tag{84}$$

where $d^\pi(s), d^{\mathcal{D}}(s)$ denote the state marginal of $d^\pi, d^{\mathcal{D}}$ and $d^{\mathcal{D}}\pi(s,a) := d^{\mathcal{D}}(s)\pi(a|s)$.

Proof.

$$
\begin{aligned}
&|V^{\pi}(\rho_0) - V^{*}(\rho_0)| && (85)\\
=&\frac{1}{1-\gamma}\left|\mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)] - \mathbb{E}_{(s,a)\sim d^{*}}[r(s,a)]\right| && (86)\\
\le&\frac{R_{\text{max}}}{1-\gamma}\sum_{s,a}|d^{\pi}(s,a) - d^{*}(s,a)| && (87)\\
=&\frac{2R_{\text{max}}}{1-\gamma}D_{\text{TV}}(d^{\pi}\|d^{*}) && (88)\\
\le&\frac{2R_{\text{max}}}{1-\gamma}\left(D_{\text{TV}}(d^{\pi}\|d^{\mathcal{D}}\pi) + D_{\text{TV}}(d^{\mathcal{D}}\pi\|d^{\mathcal{D}}\pi^{*}) + D_{\text{TV}}(d^{\mathcal{D}}\pi^{*}\|d^{*})\right) && (89)\\
=&\frac{2R_{\text{max}}}{1-\gamma}\left(D_{\text{TV}}(d^{\pi}(s)\|d^{\mathcal{D}}(s)) + D_{\text{TV}}(d^{\mathcal{D}}(s)\|d^{*}(s))\right) + \frac{2R_{\text{max}}}{1-\gamma}\mathbb{E}_{s\sim d^{\mathcal{D}}(\cdot)}[D_{\text{TV}}(\pi(\cdot|s)\|\pi^{*}(\cdot|s))] && (90)\\
\le&\frac{4R_{\text{max}}}{1-\gamma}D_{\text{TV}}(d^{*}(s)\|d^{\mathcal{D}}(s)) + \frac{2R_{\text{max}}}{1-\gamma}\mathbb{E}_{d^{\mathcal{D}}(s)}[D_{\text{TV}}(\pi(\cdot|s)\|\pi^{*}(\cdot|s))] && (91)
\end{aligned}
$$

The Eq.(89) follows the triangle inequality of TV distance. □

Now we give the complete statement and proof of Theorem 3.7.

Theorem A.5. *Suppose the maximum reward is $R_{max} = \max_{s,a}\|r(s,a)\|$, let $V^{\pi}(\rho_0) := \mathbb{E}_{s_0\sim\rho_0}[V^{\pi}(s_0)]$ denote the performance given a policy π. For policy π_θ optimized by Eq.(16) and N transition data from $d^{\mathcal{D}}$, if π_θ is a universal approximator, under Assumption A.2 and A.3, we have*

$$V^{*}(\rho_0) - V^{\pi_\theta}(\rho_0) \le \frac{4R_{max}}{1-\gamma}D_{\text{TV}}(d^{\mathcal{D}}(s)\|d^{*}(s)) + e_N$$

and e_N converges in probability to zero at the rate $N^{-\frac{1}{4+h}}, \forall h > 0$, i.e., $N^{\frac{1}{4+h}}e_N \xrightarrow{N\to\infty} 0$ in probability.

Proof. By Lemma A.4, it remains to establish the vanishing rate of $e_N := \frac{2R_{\text{max}}}{1-\gamma}\mathbb{E}_{d^{\mathcal{D}}}[D_{\text{TV}}(\pi\|\pi^{*})]$. By Pinsker's inequality and Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}_{s\sim d^{\mathcal{D}}}[D_{\text{TV}}(\pi(\cdot|s)\|\pi^{*}(\cdot|s))] \le&\mathbb{E}_{s\sim d^{\mathcal{D}}}[\sqrt{2D_{\text{KL}}(\pi(\cdot|s)\|\pi^{*}(\cdot|s))}] && (92)\\
\le&\sqrt{2\mathbb{E}_{s\sim d^{\mathcal{D}}}[D_{\text{KL}}(\pi(\cdot|s)\|\pi^{*}(\cdot|s))]} && (93)
\end{aligned}
$$

Recall that the π_θ minimizes an empirical expectation

$$
\begin{aligned}
&\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}D_{\text{KL}}(d^{\mathcal{D}}(s_i)\pi_\theta(\cdot|s_i)\|d^{*}(s_i,\cdot)) && (94)\\
=&\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{a\sim\pi_\theta}\left[\log\frac{d^{\mathcal{D}}(s_i)\pi_\theta(a|s_i)}{d^{*}(s_i)\pi^{*}(a|s_i)}\right] && (95)\\
=&\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{a\sim\pi_\theta}\left[\log\frac{\pi_\theta(a|s_i)}{\pi^{*}(a|s_i)} + \log\frac{d^{\mathcal{D}}(s)}{d^{*}(s_i)}\right] && (96)\\
=&\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{a\sim\pi_\theta}\left[\log\frac{\pi_\theta(a|s_i)}{\pi^{*}(a|s_i)}\right] && (97)\\
=&\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}D_{\text{KL}}(\pi_\theta(\cdot|s_i)\|\pi^{*}(\cdot|s_i)). && (98)
\end{aligned}
$$

i.e., the objective is equivalent to minimizing the KL divergence over policy distribution. Since π_θ is a universal approximator, then it exactly minimizes the objective, i.e.,

$$\frac{1}{N}\sum_{i=1}^{N}D_{\text{KL}}(\pi_\theta(\cdot|s_i)\|\pi^{*}(\cdot|s_i)) = 0 \tag{99}$$

Use notation $g_\theta(s) = D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi^*(\cdot|s))$ as defined in Assumption A.2. By Assumption A.2, $\sqrt{N}(\mathbb{E}_{s \sim d^{\mathcal{D}}}[g_\theta(s)] - \frac{1}{N}\sum_{i=1}^N g_\theta(s_i))$ converges in distribution to a normal distribution with mean 0 and variance $\mathrm{Var}_{s \sim d^{\mathcal{D}}}(g_\theta(s)) < \infty$. (see e.g., (Van der Vaart, 2000))

As a result, for any $h > 0$,

$$N^{\frac{1}{2+h}} \left(\mathbb{E}_{s \sim d^{\mathcal{D}}}[g_\theta(s)] - \frac{1}{N}\sum_{i=1}^N g_\theta(s_i) \right) \xrightarrow{N \to \infty} 0, \quad \text{in probability} \tag{100}$$

Therefore,

$$\begin{aligned} \mathbb{E}_{s \sim d^{\mathcal{D}}}[D_{\mathrm{TV}}(\pi_\theta(\cdot|s) \| \pi^*(\cdot|s))] \leq & \sqrt{2\mathbb{E}_{s \sim d^{\mathcal{D}}}[D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \pi^*(\cdot|s))]} & (101) \\ = & \sqrt{2\mathbb{E}_{s \sim d^{\mathcal{D}}}[g_\theta(s)]} & (102) \\ = & \sqrt{2 \left(\mathbb{E}_{s \sim d^{\mathcal{D}}}[g_\theta(s)] - \frac{1}{N}\sum_{i=1}^N g_\theta(s) \right)} & (103) \end{aligned}$$

Combine with equation 100, for any $h > 0$, we have

$$\begin{aligned} N^{\frac{1}{4+h}} e_N = & N^{\frac{1}{4+h}} \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{s \sim d^{\mathcal{D}}}[D_{\mathrm{TV}}(\pi(\cdot|s) \| \pi^*(\cdot|s))] & (104) \\ \leq & \frac{2R_{\max}}{1-\gamma} \sqrt{2N^{\frac{1}{2+h/2}} \left(\mathbb{E}_{s \sim d^{\mathcal{D}}}[g_\theta(s)] - \frac{1}{N}\sum_{i=1}^N g_\theta(s) \right)} \xrightarrow{N \to \infty} 0, \text{ in probability} & (105) \end{aligned}$$

This finishes the proof. □

B. More Experiment Details

B.1. Tasks and baselines

As stated in the main content, we adopt ”-v1” tasks for Maze2D and Adroit domains while using ”-v2” tasks for MuJoCo domain. Note many methods only provides the evaluation on ”-v0” tasks in their original papers. Meanwhile, the results from ”-v0” and ”-v1” for Maze2D and Adroit tasks are comparable since there is only a minor fix to timeout flag issue, while there are some major bug fixes from ”-v0,-v1” to ”-v2” in MuJoCo tasks.[1]. Therefore, we directly adopt reported ”-v0” scores for Maze2D and Adroit tasks and rerun the experiments for MuJoCo tasks if ”-v2” scores are absent. For the scarce dataset settings, we rerun all adopted baselines.

We rerun the baselines for these tasks using their official codes or the d3rlpy library (Seno & Imai, 2022), whose hyperparameters are kept the same as the original papers for consistency and fair evaluation. We find that some results presented in table. 1 are slightly different from the official paper, e.g., the scores of IQL and TD3+BC on MuJoCo medium tasks. Specifically, the scores of BCQ are significantly better than the reported ones in the previous benchmark(Fu et al., 2020; Prudencio et al., 2023) due to the implementation improvement in the d3rlpy library.

B.2. Full algorithm and details of CDE

In this section, we present the full algorithm and implementation details. Without otherwise statements, the parameterization of policies or critics (i.e., value functions) defaults to be neural networks (NN).

B.2.1. Value functions separation

In CDE, we learn both the V-value function and the advantage function. The former can incorporate the stochasticity of action distribution to reduce the instability, and the latter is to generalize the optimal importance ratios to OOD regions since the reward and transition probability functions for unseen transition (s, a, r, s') are absent in offline datasets. Meanwhile, we take two steps to train V-value and advantage functions instead of optimizing the objective function in Eq.(8). Note that the objective can be separated into in-distribution and OOD parts:

$$\begin{aligned}
&\zeta\mathbb{E}_{d^{\mathcal{D}}}\left[w(s,a)A(s,a)-\alpha f(w(s,a))\right]+(1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)] && (106)\\
&\quad+(1-\zeta)\mathbb{E}_{\mu}\left[w(s,a)(A(s,a)-\lambda(s,a))-\alpha f(w(s,a))+\tilde{\epsilon}\lambda(s,a)\right], && (107)\\
=&\zeta(\mathbb{E}_{d^{\mathcal{D}}}\left[w(s,a)A(s,a)-\alpha f(w(s,a))\right]+((1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)])) && (108)\\
&\quad+(1-\zeta)(\mathbb{E}_{\mu}\left[w(s,a)(A(s,a)-\lambda(s,a))-\alpha f(w(s,a))+\tilde{\epsilon}\lambda(s,a)\right]+(1-\gamma)\mathbb{E}_{\rho_0}[v(s_0)]) && (109)
\end{aligned}$$

where the in-distribution part of the objective (i.e., Eq.(108)) corresponds to the learning objective of the V-value function in Eq.(12), which is also the dual form of following constrained optimization:

$$\max_{d^\pi\geq 0}\mathbb{E}_{d^\pi}[r(s,a)]-\alpha D_f(d^\pi\|d^{\mathcal{D}}) \quad (110)$$

$$s.t.\sum_a d^\pi(s,a)=(1-\gamma)\rho_0+\mathcal{T}_*d^\pi(s),\forall s,a\in\text{supp}(d^{\mathcal{D}}). \quad (111)$$

The main difference of it from the previous one in Eq.(3) is that it constrains the state-action in the support of offline datasets. Therefore, the objective for V-value function learning in Eq.(12) is still a convex optimization problem.

B.2.2. Mixed behavior policy training

The mixed behavior policy $\hat{\pi}^{\mathcal{D}}$ is the mixture of the behavior policy of the dataset and a uniform policy over OOD action space. Since we only require access to the KL divergence between π_θ and $\hat{\pi}^{\mathcal{D}}$, we can either approximate $\hat{\pi}^{\mathcal{D}}$ or $\hat{\pi}$ and augment it with a uniform distribution on OOD regions. The direct learning on $\hat{\pi}^{\mathcal{D}}$ can be implemented by weighted BC from state pairs $\{(s,a_{\text{in}}),(s,a_{\text{out}}^{(1)}),(s,a_{\text{out}}^{(2)}),\dots,(s,a_{\text{out}}^{(n)})\}$, where $(s,a_{\text{in}})\in\mathcal{D}$ and $(s,a_{\text{out}}^{(1)}),\dots,(s,a_{\text{out}}^{(n)})$ are from OOD sampling.

In practice, we also apply trajectory reweighting for behavior policy training. We reweight each trajectory according to its return. Specifically, given a trajectory τ, denote the normalized score of return as $G(\tau)$, we use the Boltzmann distribution

[1] Seed more details in the official repository: Maze2D, Adroit, MuJoCo.

over trajectories as Eq.(10) in (Hong et al.), i.e.,

$$w(\tau_i) = \frac{G(\tau_i)/\tau}{\sum_{\tau_j \in \mathcal{D}} G(\tau_j)/\tau} \tag{112}$$

where $w(\cdot)$ means the new weight of trajectory and temperature τ is a constant. Therefore, the behavior policy will be trained via weighted BC with computed weights. Notably, the trajectory reweighting will not influence other training parts.

Table 2: The shared hyperparameters.

Hyperparameters	**values**
hidden layers of policy π_θ	[256,256]
hidden layers of $\hat{\pi}^{\mathcal{D}}$	[256,256]
number of mixtures of $\hat{\pi}^{\mathcal{D}}$	3
hidden layers of V-value v_φ	[256,256]
hidden layers of advantage A_ϕ	[256,256]
activation function of networks	ReLU
NN optimizer	Adam
NN learning rate	3e-4
discount factor γ	0.99
batch size	512
mixture coefficient ζ	0.9
max OOD IS ratio $\tilde{\epsilon}$	0.3
number of OOD action sampling N_{OOD}	5

B.2.3. Hyperparameters

Before training NN, we standardize the observation and reward and scale the reward by multiplying 0.1 (Lee et al., 2021). Note that we will extract the policy after the optimization over value functions converges. In practice, we set the warm-up training step, and the policy π_θ will not start until warm-up training ends. We set the f-divergence coefficient $\alpha = 0.01$ except maze2d-umaze and maze2d-medium ($\alpha = 0.001$) tasks for fair comparisons, which is *significantly different from previous DICE paper (Lee et al., 2021) that finetunes and assigns different hyperparameters for every task*. The other shared hyperparameters are summaries in table 2.

B.2.4. Full algorithm

In this section, we present the full algorithm in Algorithm. 2. In practice, we continue to train value functions and behavior policy after warm-up steps since we find that it improves the performances in most tasks.

B.3. More experiment results

The experiment results on MuJoCo medium tasks are shown in fig. 3. We can find that most methods perform similarly in medium tasks, and our method also obtains comparable performances. Actually, there is no significant performance decrease when we shrink the size of the dataset. This is because the trajectories in these datasets are highly homogeneous, and thus the policy can be learned well from only a small proportion of offline data.

The training curves of full dataset experiments are shown in fig. 4. The training steps start from a non-zero number because of the warm-up step, i.e., we learn the policy after the value function almost converges. The warm-up step is 20,000 for the maze2d environment and 40,000 for other environments.

We can observe that our method converges extremely fast and is very stable during training. This is because CDE employs convex optimization to solve the value function and extracts the optimal policy in a manner of supervised learning. On the contrary, the previous methods (e.g., Q-learning-based methods (Kumar et al., 2020; Kostrikov et al., 2021b)) are prone to over-fitting in training due to the interleaved optimization of value and policy (Brandfonbrener et al., 2021), which may lead to large compounded errors and performance decrease especially with long training steps.

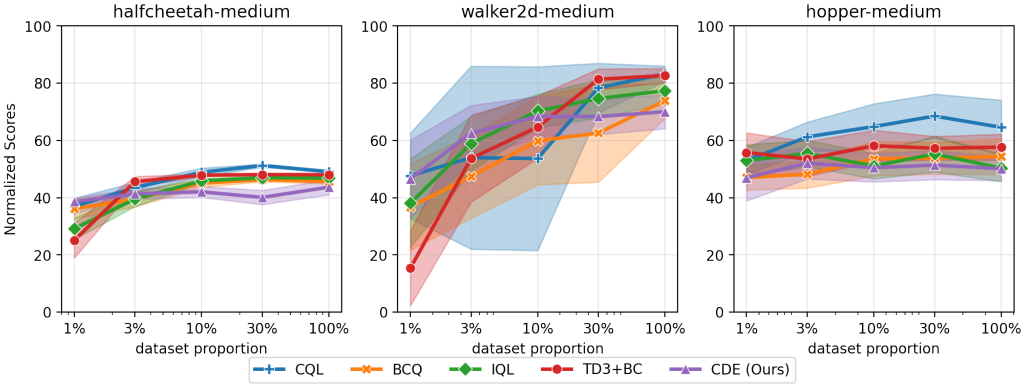


Figure 3: The results on sub-datasets with different dataset sizes for mujoco medium tasks.

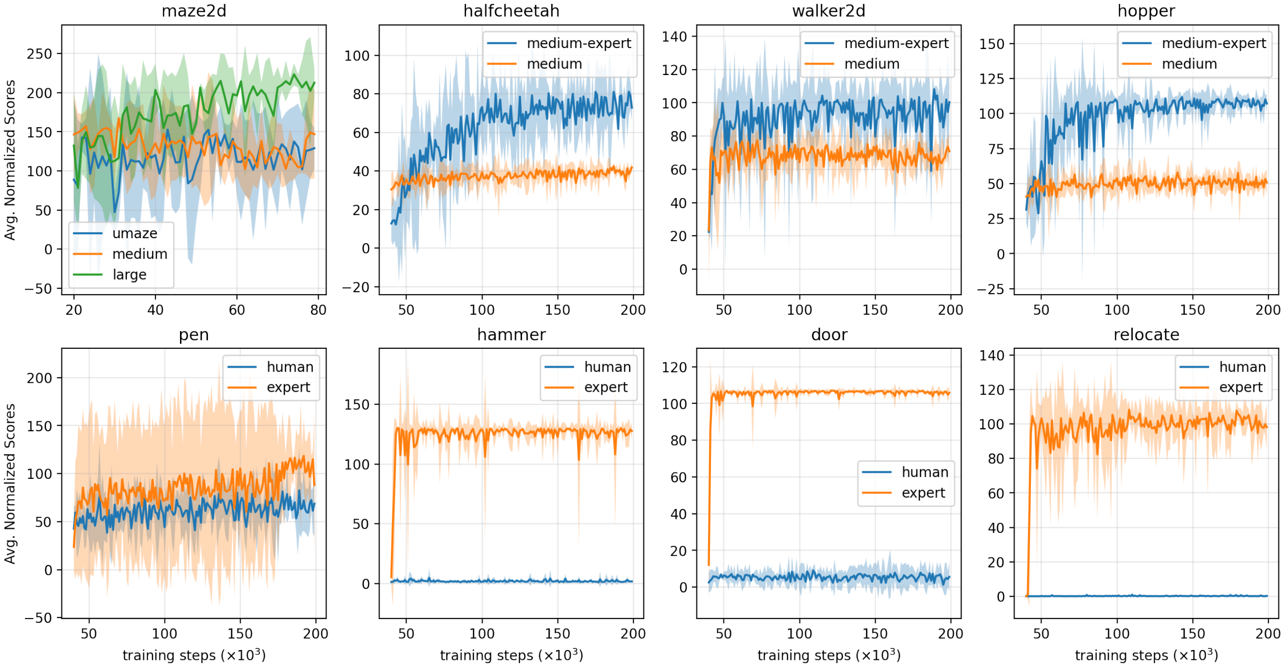


Figure 4: The training curves of CDE. The shadow region indicates the standard deviation of mean values across different seeds instead of the standard deviation of 20 evaluation trajectories, which is much larger than the one in the figure.

Algorithm 2 Full Algorithm of Conservative Density Estimation

```
Initialize value functions $v_\varphi, \tilde{A}_\phi$, mixed behavior policy $\hat{\pi}^{\mathcal{D}}$, policy $\pi_\theta$.
for training iteration $i$ do
    ▷ policy evaluation and improvement
    Sample batch $\{(s_i, a_i, r_i, s'_i)\}$ from $\mathcal{D}$ and $n$ OOD actions $\{a^{(1)}, \dots, a^{(n)}\}$ for each $s$;
    Compute regularized advantage function $\tilde{A}(s, a)$ via $v_\phi$ for in-distribution and $\tilde{A}_\phi$ for OOD state-actions.
    $\tilde{A}(s, a) \leftarrow \tilde{A}(s, a) - \eta, \forall s, a$ to normalize optimal distribution.
    Update V-value $v_\varphi$ by Eq.(12);
    Update regularized advantage $\tilde{A}_\phi$ by Eq.(13);
    Update distribution normalizer $\eta$ by gradient descent: $\eta \leftarrow \eta - \alpha_\eta(1 - \mathbb{E}[w^*(s, a)])$.
    Update $\hat{\pi}^{\mathcal{D}}$ by weighted importance sampling.
    ▷ policy extraction
    if $i \geq$ warm-up steps then
        Update policy $\pi_\theta$ by Eq.(16) with entropy regularization.
    end if
end for
```