Offline Reinforcement Learning with Wasserstein Regularization via Optimal Transport Maps

Anonymous authors
Paper under double-blind review

Keywords: Offline Reinforcement Learning, Deep Reinforcement Learning, Wasserstein Distance.

Summary

Offline reinforcement learning (RL) aims to learn an optimal policy from a static dataset, making it particularly valuable in scenarios where data collection is costly, such as robotics. A major challenge in offline RL is distributional shift, where the learned policy deviates from the dataset distribution, potentially leading to unreliable out-of-distribution actions. To mitigate this issue, regularization techniques have been employed. While many existing methods utilize density ratio-based measures, such as the f-divergence, for regularization, we propose an approach that utilizes the Wasserstein distance, which is robust to out-of-distribution data and captures the similarity between actions. Our method employs input-convex neural networks (ICNNs) to model optimal transport maps, enabling the computation of the Wasserstein distance in a discriminator-free manner, thereby avoiding adversarial training and ensuring stable learning. Our approach demonstrates comparable or superior performance to widely used existing methods on the D4RL benchmark dataset.

Contribution(s)

1. We introduce a novel regularization method with the Wasserstein distance via optimal transport maps for offline RL, eliminating the need for adversarial training and a discriminator through ICNNs.
   Context: Wu et al. (2019); Asadulaev et al. (2024) performed regularization using the Wasserstein distance in offline reinforcement learning through adversarial learning with a discriminator. Makkuva et al. (2020); Korotin et al. (2021b;a); Mokrov et al. (2021) modeled the Wasserstein distance in a discriminator-free manner using ICNNs in a non-RL domain.
2. We evaluate our proposed method on the D4RL benchmark dataset and find that it achieves performance comparable to or even surpassing that of widely used methods. Additionally, by comparing it with an adversarial training-based approach, we show that our discriminator-free method incorporates Wasserstein distance regularization more effectively for these tasks.
   Context: We compared our method with Kostrikov et al. (2022), which serves as a component of our approach, and a Wu et al. (2019)-based method that performs regularization using the Wasserstein distance via discriminator-based adversarial learning. By keeping the value function learning consistent across these existing methods and the proposed method, we fairly evaluated the effect of our proposed regularization on the policy.

Offline Reinforcement Learning with Wasserstein Regularization via Optimal Transport Maps

Anonymous authors
Paper under double-blind review

Abstract

Offline reinforcement learning (RL) aims to learn an optimal policy from a static dataset, making it particularly valuable in scenarios where data collection is costly, such as robotics. A major challenge in offline RL is distributional shift, where the learned policy deviates from the dataset distribution, potentially leading to unreliable out-of-distribution actions. To mitigate this issue, regularization techniques have been employed. While many existing methods utilize density ratio-based measures, such as the f-divergence, for regularization, we propose an approach that utilizes the Wasserstein distance, which is robust to out-of-distribution data and captures the similarity between actions. Our method employs input-convex neural networks (ICNNs) to model optimal transport maps, enabling the computation of the Wasserstein distance in a discriminator-free manner, thereby avoiding adversarial training and ensuring stable learning. Our approach demonstrates comparable or superior performance to widely used existing methods on the D4RL benchmark dataset.

1 Introduction

In offline reinforcement learning (RL), learning is conducted solely using a pre-collected dataset to maximize return. When the learned policy deviates from the behavior policy of the dataset, issues such as overestimation of values in unseen states and actions arise (Levine et al., 2020). Preventing such divergence remains a central challenge in offline RL. Prior studies introduced regularization methods to mitigate distributional shift, including those based on the f-divergence (Wu et al., 2019; Garg et al., 2023; Sikchi et al., 2024).

Regularization measures based on the density ratio of distributions such as f-divergence can become unstable when the supports of the distributions do not overlap, and these measures do not consider the similarity between variables. Thus, we employ the Wasserstein distance as a regularization term, as it is robust to out-of-distribution data and can incorporate the metric of the variable space. When we apply the Wasserstein distance to RL, we can take into consideration the distances in the continuous action space and can handle out-of-distribution actions.

The Wasserstein distance between probability distributions P and Q is defined as the infimum of the expected value of the distance between corresponding samples over all possible couplings of P and Q. For the 2-Wasserstein distance, if P is absolutely continuous with respect to the Lebesgue measure, there exists a convex function ψ whose gradient $\nabla\psi$ acts as the unique optimal transport map from P to Q. In this setting, the coupling induced by P and the mapping $\nabla\psi$ is the optimal coupling (Brenier, 1991).

Here, we consider the case where P is fixed, and Q is optimized within an objective function that includes $W_2^2(P, Q)$. When Q is directly modeled using a generator and the Wasserstein distance is computed with a discriminator, as in WGAN (Arjovsky et al., 2017), additional instability occurs due to adversarial training. Moreover, since controlling the Lipschitz constant of the discriminator is inherently difficult, accurately computing the Wasserstein distance is challenging. To address this issue, we propose optimizing a convex function ψ in place of Q, based on Brenier's theorem (Brenier, 1991). Since ψ is learned by minimizing the L^2 distance instead of using adversarial training, the learning process is relatively stable. Furthermore, as long as ψ remains convex, it is guaranteed to approximate the Wasserstein distance between P and some distribution Q, ensuring that the exact Wasserstein distance is consistently computed, even during training.

We apply this approach to policy regularization in offline RL by applying Wasserstein distance regularization to the visitation distribution. Specifically, we model the visitation distribution $d^\pi(s, a)$ of the learned policy as the distribution transported from the dataset distribution $d^{\mathcal{D}}(s, a)$ through the gradient of a convex function. This transport map corresponds to the optimal transport map in the 2-Wasserstein distance $W_2^2(d^{\mathcal{D}}, d^\pi)$. By learning a parameterized convex function that maximizes the objective regularized by the Wasserstein distance, we can obtain d^π without adversarial training. The policy π is then learned from samples drawn from $d^\pi(s, a)$.

We employ input-convex neural networks (ICNNs) (Amos et al., 2017) as the parameterized convex function. By integrating this policy learning method with existing in-sample value function learning methods, we propose a simple Wasserstein regularized algorithm that only requires additional ICNN training. We refer to our approach as *Q-learning regularized by Direct Optimal Transport modeling* (Q-DOT) and evaluate its performance through experiments.

We conduct experiments using the D4RL benchmark dataset (Fu et al., 2020) and compare our method with widely used existing approaches. The results demonstrate that our proposed method achieves performance comparable or superior to existing methods. Furthermore, we compare our method with adversarial training-based Wasserstein distance regularization methods that use a discriminator, confirming that our discriminator-free approach is more stable and effective.

Our study makes the following key contributions:

- We introduce a novel regularization method with the Wasserstein distance via optimal transport maps for offline RL, eliminating the need for adversarial training and a discriminator through ICNNs.
- We evaluate our proposed method on the D4RL benchmark dataset and find that it achieves performance comparable to or even surpassing that of widely used methods. Additionally, by comparing it with an adversarial training-based approach, we show that our discriminator-free method incorporates Wasserstein distance regularization more effectively for these tasks.

2 Preliminaries

2.1 Reinforcement Learning

Reinforcement learning (RL) is a framework for sequential decision-making, where an agent interacts with an environment modeled as a Markov decision process (MDP). An MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma, d_0)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(s'|s, a)$ is the transition probability distribution, $r(s, a)$ is the reward function, $\gamma \in (0, 1)$ is the discount factor and d_0 is the probability distribution of initial states. The agent follows a policy $\pi(a|s)$, which defines a probability distribution over actions given a state, aiming to maximize the expected cumulative reward: $\mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$, where T is a task horizon.

2.2 Offline RL with Regularization

Offline RL focuses on learning an optimal policy purely from a fixed dataset $\mathcal{D} = \{(s, a, r, s')\}$ collected by an unknown behavior policy $\pi_{\mathcal{D}}$. Since the learned policy π may select actions outside the support of $\mathcal{D}$, distributional shift issues arise, causing erroneous value estimates and degraded performance.

To mitigate distributional shift, regularization techniques are employed to constrain the learned policy. Regularization is sometimes applied to the divergence between the learned policy π and the dataset policy $\pi_{\mathcal{D}}$ (Garg et al., 2023; Xu et al., 2023). In this study, following Nachum & Dai (2020); Sikchi et al. (2024), we consider regularization based on the visitation distributions $d^{\mathcal{D}}$ and d^{π}. In this case, the optimization problem with regularization is formulated as follows:

$$\begin{aligned} \max_{d \geqslant 0} \quad & \mathbb{E}_{d(s,a)}[r(s,a)] - \alpha D(d(s,a) \| d^{\mathcal{D}}(s,a)) \\ \text{s.t.} \quad & \sum_{a} d(s,a) = (1-\gamma) d_0(s) + \gamma \sum_{s',a'} d(s',a') p(s|s',a'), \end{aligned} \tag{1}$$

where D is a divergence that measures the deviation between distributions, and α is a hyperparameter that adjusts the strength of regularization. From Lagrange duality, this constrained optimization problem is equivalent to the following min-max problem (Nachum & Dai, 2020; Sikchi et al., 2024):

$$\begin{aligned} & \min_{V} \max_{d \geqslant 0} \mathbb{E}_{(s,a)\sim d}\left[r(s,a) - \alpha D\left(d(s,a) \| d^{\mathcal{D}}(s,a)\right)\right] \\ & \qquad + \sum_{s} V(s) \left((1-\gamma) d_0(s) + \gamma \sum_{s',a'} d(s',a') p(s|s',a') - \sum_{a \in A} d(s,a) \right), \qquad (2) \\ & = \min_{V} \max_{d \geqslant 0} (1-\gamma) \mathbb{E}_{d_0(s)}[V(s)] + \mathbb{E}_{(s,a)\sim d}\left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V(s') - V(s) \right] \\ & \qquad - \alpha D\left(d(s,a) \| d^{\mathcal{D}}(s,a)\right), \qquad (3) \\ & = \min_{V} \max_{d \geqslant 0} \, (1-\gamma) \mathbb{E}_{d_0(s)}[V(s)] + \mathbb{E}_{(s,a)\sim d}[Q(s,a) - V(s)] - \alpha D(d(s,a) \| d^{\mathcal{D}}(s,a)). \qquad (4) \end{aligned}$$

In the next section, we introduce a discriminator-free regularization method using this equation with the Wasserstein distance.

2.3 Wasserstein Distance

The Wasserstein distance, particularly the 2-Wasserstein distance, is widely used to measure the discrepancy between two probability distributions. Given two distributions P and Q on $\mathbb{R}^D$ with finite second order moments, the 2-Wasserstein distance is defined as follows:

$$W_2^2(P,Q) \coloneqq \min_{\xi \in \Pi(P,Q)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|x-y\|_2^2 \, d\xi(x,y), \tag{5}$$

where $\Pi(P,Q)$ denotes the set of all joint distributions whose marginals are P and Q. The Wasserstein distance captures the geometric discrepancy between probability distributions. Unlike density-ratio-based measures such as the KL divergence, which can diverge when the supports of the distributions do not overlap, the Wasserstein distance is less prone to divergence and serves as a robust measure for out-of-distribution data.

Brenier (1991) showed that if P is absolutely continuous with respect to the Lebesgue measure, there exists a convex function $\psi : \mathbb{R}^D \to \mathbb{R} \cup \{\infty\}$ whose gradient $\nabla\psi : \mathbb{R}^D \to \mathbb{R}^D$ serves as the unique optimal transport map from P to Q. Consequently, the unique optimal transport plan is $\xi^* = [\mathrm{id}_{\mathbb{R}^D}, T^*]\sharp P$, with $T^* = \nabla\psi$. Here, for any measurable mapping $T : \mathbb{R}^D \to \mathbb{R}^D$, $T\sharp P$ denotes the push-forward of P under T, and $\mathrm{id}_{\mathbb{R}^D}$ is the identity map on $\mathbb{R}^D$. Then, $Q = \nabla\psi\sharp P$, and the 2-Wasserstein distance can be expressed as:

$$W_2^2(P,Q) = \int_{\mathbb{R}^D} \|x - \nabla\psi(x)\|_2^2 \, dP(x). \tag{6}$$

The convex function $\psi(x)$ can be modeled using ICNNs (Amos et al., 2017), and its gradient can be used as a mapping to compute the Wasserstein distance (Makkuva et al., 2020; Korotin et al., 2021a;b; Mokrov et al., 2021).

3 Offline RL with Wasserstein Regularization via Optimal Transport Maps

In this section, we propose a method for regularization using the Wasserstein distance in offline RL without relying on a discriminator. This algorithm involves learning a value function $Q_\theta(s,a)$ and $V_\phi(s)$, a generator $d_\omega(s,a)$ corresponding to the visitation distribution, and a policy $\pi_\rho(a \mid s)$. The parameters $\theta, \phi, \omega, \rho$ represent the respective neural network parameters.

3.1 Learning the Visitation Distribution and Policy

We begin by describing the key component of this approach: learning the visitation distribution $d_\omega(s,a)$. From Eq. (4), the objective function for d_ω when the regularization measure is the squared 2-Wasserstein distance is given by:

$$J(\omega) = \mathbb{E}_{s,a\sim d_\omega}[Q_\theta(s,a) - V_\phi(s)] - \alpha W_2^2(d_\omega(s,a)\|d^{\mathcal{D}}(s,a)). \tag{7}$$

That is, the learned policy's visitation distribution is optimized to maximize the advantage while being regularized to prevent excessive deviation from the dataset distribution $d^{\mathcal{D}}$.

To model d_ω, we apply Brenier's theorem. Specifically, we parameterize a convex function ψ_ω using an ICNN and define d_ω as the push-forward from $d^{\mathcal{D}}$ through the gradient: $d_\omega = \nabla\psi_\omega \sharp d^{\mathcal{D}}$. This means that samples from d_ω are obtained as the gradient of the convex function $\nabla\psi_\omega(x)$, where samples x are drawn from the offline distribution $d^{\mathcal{D}}$. Consequently, the Wasserstein distance can be evaluated as:

$$W_2^2(d_\omega, d^{\mathcal{D}}) = \mathbb{E}_{x\sim d^{\mathcal{D}}}\left[\|x - \nabla\psi_\omega(x)\|_2^2\right], \tag{8}$$

where x represents a state-action pair vector. Accordingly, the objective function for d_ω is formulated as:

$$J_\psi(\omega) = \mathbb{E}_{(s,a)\sim d_\omega}\Big[Q_{\hat{\theta}}(s,a) - V_\phi(s)\Big] - \alpha\mathbb{E}_{x\sim d^{\mathcal{D}}}\Big[\|x - \nabla\psi_\omega(x)\|_2^2\Big]. \tag{9}$$

where $\hat{\theta}$ represents the parameters of the target network. There are design choices regarding how to treat x as a combination of state and action and how to model ψ. Since our primary focus when learning a policy from d is on changes in action rather than changes in state, we opt to keep the state unchanged. Instead, we condition on the state and only modify the action. Specifically, for $(s,a) \sim d^{\mathcal{D}}$, we derive $a' = \nabla\psi_s(a)$ and treat (s,a') as a sample from d, where $\nabla\psi_s(a)$ represents the gradient of a convex function with respect to a, conditioned on the state s. In this setup, the Wasserstein distance is computed based on the visitation distribution conditioned on the state. Since this conditioned distribution is equivalent to the policy, it serves as a form of policy regularization, akin to the method proposed by Wu et al. (2019).

In this manner, Wasserstein distance regularization is incorporated into learning without requiring a discriminator. This modeling approach offers additional benefits. Makkuva et al. (2020) reported that while mappings using conventional neural networks, such as those in Arjovsky et al. (2017), are constrained to be continuous, gradient-based modeling allows for the learning of discontinuous mappings. Our method also benefits from this property. However, in general settings, gradient-based methods face challenges, such as the inability to generate more samples than those available in the offline dataset and the tendency to collapse into an identity mapping when return maximization is absent, preventing the generation of new data. Nevertheless, in offline RL, it is reasonable to use only reliable data transformed from the offline dataset to avoid out-of-distribution issues. Moreover, the hyperparameter α allows for adjustment between behavior cloning, which corresponds to the identity mapping, and RL. In other words, this method enables discriminator-free modeling without significant issues, making it well-suited for offline RL.

For policy learning from the learned d_ω, we utilize Advantage Weighted Regression (AWR) (Nair et al., 2021), a method commonly used in offline RL. While existing approaches such as Nair et al.

(2021); Kostrikov et al. (2022); Garg et al. (2023) maximize the log-likelihood of offline dataset state-action pairs weighted by the advantage, our method instead maximizes the log-likelihood of state-action pairs sampled from the learned d_ω, which are obtained as transformed versions of offline data through $\nabla\psi_\omega$. Thus, the loss function is formulated as follows:

$$L_\pi(\rho) = -\mathbb{E}_{(s,a)\sim d_\omega}\left[\exp\left(\beta(Q_\theta(s,a) - V_\phi(s))\right)\log \pi_\rho(a|s)\right]. \tag{10}$$

3.2 Learning the Value Function

There are two common approaches to incorporating regularization into policy learning: *value penalty* and *policy regularization* (Wu et al., 2019). Value penalty methods introduce a penalty term into the value function, whereas policy regularization directly applies a penalty to the policy itself. Although value penalty-based learning can be implemented by optimizing the value function according to Eq. (4), this results in an adversarial learning setup where d_ω is maximized while V is minimized, leading to instability.

To address this issue, we adopt policy regularization instead. Specifically, the policy is trained using the method derived from the regularized objective, as mentioned above, while the value function is learned without regularization. For the learning of the value function, we employed Implicit Q-Learning (IQL) (Kostrikov et al., 2022) for stability. The IQL enables Q-learning-based value function learning through the expectile regression, allowing us to avoid out-of-distribution samples while maintaining an in-sample learning approach. The loss functions for Q and V are formulated as follows:

$$L_V(\phi) = \mathbb{E}_{(s,a)\sim D}\left[L_\tau^2(Q_{\hat{\theta}}(s,a) - V_\phi(s))\right], \tag{11}$$

$$L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim D}\left[(r(s,a) + \gamma V_\phi(s') - Q_\theta(s,a))^2\right]. \tag{12}$$

By integrating this value function learning approach with the previously described learning of d_ω and π_ρ, we achieve a policy regularization-based method that incorporates the Wasserstein distance regularization without relying on a discriminator or adversarial training.

We name this method *Q-learning regularized by Direct Optimal Transport modeling* (Q-DOT) and evaluate it through experiments. The corresponding pseudocode is presented in Algorithm 1.

4 Experiments

4.1 Experimental Setup

In this section, we evaluate the effectiveness of the proposed method using the D4RL benchmark (Fu et al., 2020). For comparison, we consider widely used offline RL methods (Fujimoto & Gu, 2021; Kostrikov et al., 2022; Kumar et al., 2020; Chen et al., 2021), and refer to the scores reported in Kostrikov et al. (2022). In addition, we implement and experiment with an adversarial learning-based method that incorporates regularization using the Wasserstein distance, which we refer to as Adversarial Wasserstein (AdvW). In AdvW, the policy is updated based on Wu et al. (2019), and its objective is defined as follows:

Algorithm 1 Q-DOT

```
Input: Offline dataset D = {(s, a, r, s')}
Initialize: Q_θ, V_φ, π_ρ and ICNN ψ_ω
for each update step do
    Sample mini-batch {(s, a, r, s')} from D
    Update V_φ by minimizing Eq. (11)
    Update Q_θ by minimizing Eq. (12)
    Update ψ_ω by maximizing Eq. (9)
    Update π_ρ by minimizing Eq. (10)
end for
```

$$\max_\pi \mathbb{E}_{s\sim\mathcal{D},a\sim\pi}\left[Q(s,a) - \alpha W(\pi(\cdot|s), \pi_{\mathcal{D}}(\cdot|s))\right]. \tag{13}$$

Table 1: The average normalized return on D4RL tasks. For our method, Q-DOT, the mean and standard error over six random seeds are reported.

Dataset	BC	10%BC	DT	TD3+BC	CQL	IQL	AdvW	Q-DOT (Ours)
halfcheetah-medium-v2	42.6	42.5	42.6	**48.3**	44.0	**47.4**	**48.6**	**47.9**$_{\pm 0.1}$
hopper-medium-v2	52.9	56.9	67.6	59.3	58.5	66.3	61.2	**76.7**$_{\pm 3.2}$
walker2d-medium-v2	75.3	75.0	74.0	**83.7**	72.5	78.3	80.7	**83.0**$_{\pm 0.8}$
halfcheetah-medium-replay-v2	36.6	40.6	36.6	**44.6**	**45.5**	**44.2**	44.2	**43.7**$_{\pm 0.6}$
hopper-medium-replay-v2	18.1	75.9	82.7	60.9	95.0	94.7	48.1	**97.4**$_{\pm 1.2}$
walker2d-medium-replay-v2	26.0	62.5	66.6	**81.8**	77.2	73.9	68.9	70.7$_{\pm 4.0}$
halfcheetah-medium-expert-v2	55.2	**92.9**	86.8	90.7	**91.6**	86.7	22.6	89.6$_{\pm 1.7}$
hopper-medium-expert-v2	52.5	**110.9**	107.6	98.0	105.4	91.5	17.6	93.1$_{\pm 13.0}$
walker2d-medium-expert-v2	107.5	**109.0**	108.1	**110.1**	**108.8**	**109.6**	92.5	**110.3**$_{\pm 0.1}$
locomotion total	466.7	666.2	672.6	677.4	698.5	692.4	480.1	**712.4**
antmaze-umaze-v0	54.6	62.8	59.2	78.6	74.0	**87.5**	83.2	**87.8**$_{\pm 1.1}$
antmaze-umaze-diverse-v0	45.6	50.2	53.0	71.4	**84.0**	62.2	51.0	70.2$_{\pm 3.8}$
antmaze-medium-play-v0	0.0	5.4	0.0	10.6	61.2	**71.2**	46.0	68.2$_{\pm 1.5}$
antmaze-medium-diverse-v0	0.0	9.8	0.0	3.0	53.7	**70.0**	42.5	66.2$_{\pm 5.5}$
antmaze-large-play-v0	0.0	0.0	0.0	0.2	15.8	39.6	12.5	**49.0**$_{\pm 4.2}$
antmaze-large-diverse-v0	0.0	6.0	0.0	0.0	14.9	**47.5**	8.2	40.7$_{\pm 4.9}$
antmaze total	100.2	134.2	112.2	163.8	303.6	378.0	243.3	**382.0**
kitchen-complete-v0	**65.0**	-	-	-	43.8	62.5	4.2	**64.2**$_{\pm 3.4}$
kitchen-partial-v0	38.0	-	-	-	49.8	46.3	24.6	**71.3**$_{\pm 1.3}$
kitchen-mixed-v0	**51.5**	-	-	-	**51.0**	**51.0**	21.7	42.9$_{\pm 4.3}$
kitchen total	154.5	-	-	-	144.6	159.8	50.4	**178.3**

where $\pi_{\mathcal{D}}$ represents the behavior policy used for dataset collection. The Wasserstein regularization is computed using the dual form with a discriminator g, following Arjovsky et al. (2017): $W(p,q) = \sup_{g:\|g\|_L \leq 1} \left(\mathbb{E}_{x\sim p}[g(x)] - \mathbb{E}_{x\sim q}[g(x)]\right)$. The training of both the discriminator and the policy follows the official implementation of Wu et al. (2019). The discriminator is trained with a gradient penalty to enforce the Lipschitz constraint. For value function training, we observed that on-policy training, as in Wu et al. (2019), failed to achieve decent learning in the D4RL benchmark. Therefore, we adopted in-sample learning of IQL, similar to our proposed method, resulting in a fair comparison. The hyperparameter α in AdvW was tuned using the values $(0.3, 1, 3, 10, 30)$ specified in Wu et al. (2019). However, we found that the learned policy outputs sometimes deviated significantly from the dataset actions. To address this, we further tested larger values $(10^2, 15^2, 20^2)$.

The expectile parameter τ used for value function estimation in both our proposed method and AdvW is a hyperparameter, and we adopted the same values as in Kostrikov et al. (2022). The implementation of ICNN utilizes Cuturi et al. (2022). The network consists of a two-layer fully connected architecture with 256 hidden units per layer, identical to the other actor and critic networks. Additional implementation details and other hyperparameters are provided in the Supplementary Materials.

4.2 Results on the D4RL Benchmark

Table 1 presents the results on the D4RL benchmark, which consists of the locomotion, antmaze, and kitchen domains. In terms of total return, our proposed method consistently achieved the best or comparable performance across all domains compared to the baselines. When analyzing individual datasets, our method attained the highest return on several datasets. Notably, it outperformed the best existing methods by significant margins in hopper-medium-v2 and kitchen-partial-v0, achieving improvements of +10.4 and +21.5, respectively. For the Antmaze domain, our method performed similarly to IQL. Since Antmaze requires trajectory stitching (Zhuang et al., 2024), regularization alone does not substantially enhance stitching ability. This suggests that further performance improvements would require incorporating additional techniques beyond regularization.

On the other hand, AdvW exhibited lower performance across all domains in terms of total score. In particular, it performed poorly on datasets that contain successful trajectories, such as halfcheetah-

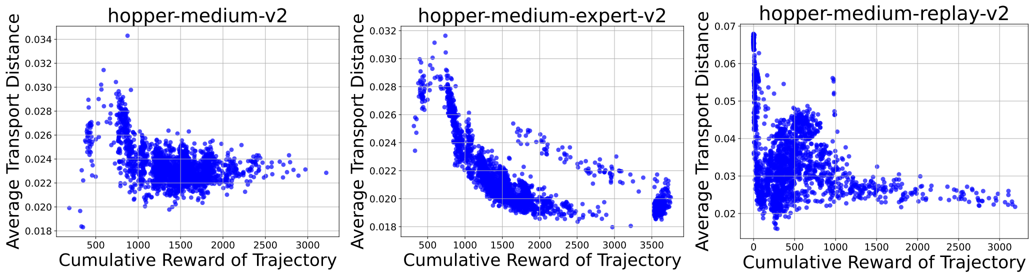

Figure 1: The relationship between trajectory quality and transport-induced distance. The x-axis represents the cumulative reward of each trajectory, while the y-axis shows the average L2 norm of state-action pair differences before and after transport.

medium-expert-v2, hopper-medium-expert-v2, and kitchen-complete-v0, where strong regularization is crucial. This was the case even when using large values for α (e.g., $10^2, 15^2, 20^2$). These AdvW results are similar to the BRAC results reported in Zhang et al. (2021) using a discriminator with f-divergence, suggesting that behavior cloning generally becomes more challenging when adversarial learning is involved.

These results suggest that adversarial learning-based regularization via Wasserstein distance is inherently unstable and challenging. In contrast, the discriminator-free training approach of our proposed method demonstrates effectiveness in achieving high scores consistently.

4.3 Trajectory Quality and Transport Distance

We analyze the relationship between trajectory quality and state-action pair transformations induced by a trained ICNN mapping. Specifically, we visualize the relationship between the cumulative rewards of trajectories in the offline dataset and the average transport-induced distance over state-action pairs. The horizontal axis represents the cumulative reward of each trajectory, while the vertical axis indicates the average L2 norm of the difference between state-action pairs before and after transport.

The results from three tasks show that lower-reward trajectories exhibit greater transport distances for their state-action pairs. This suggests that the optimization objective in Figure 1 effectively regularizes state-action transformations by primarily modifying low-quality trajectories while preserving high-quality ones. Additional results for other tasks are provided in the Supplementary Material.

5 Related Work

Offline RL aims to learn policies solely from pre-collected data. A central challenge in this setting is addressing the distributional shift between the state-action distribution of the learned policy and that of the offline dataset. When the distributional shift is large, the value of actions not observed in the offline dataset may be overestimated (Levine et al., 2020). A particularly simple approach to mitigating this discrepancy is presented in (Fujimoto & Gu, 2021). Fujimoto & Gu (2021) propose a policy regularization method based on TD3, an off-policy technique commonly used in online RL, by incorporating a behavior cloning term into the policy learning process. The behavior cloning term is defined as the squared error between the action output by the learned policy and the action in the dataset. This corresponds to the 2-Wasserstein distance between the dataset policy and the learned policy in cases where the learned policy is deterministic. The promising performance of this simple method suggests the effectiveness of using a notion of action similarity, such as the Wasserstein distance, as a regularization term.

Wu et al. (2019) experimented with value penalty and policy regularization using various divergence measures. Regularization based on f-divergence and the Wasserstein distance was also explored, where optimization was performed using a dual-form discriminator-based approach. However, as demonstrated with AdvW, even when in-sample maximization was incorporated into value function learning, adversarial learning with a discriminator did not perform well on the D4RL dataset, highlighting the necessity of discriminator-free learning. Asadulaev et al. (2024) proposed a method that extends BRAC with Wasserstein distance regularization. Similar to BRAC, their approach employs adversarial learning using a discriminator. However, their main proposed method is based on Tarasov et al. (2023), which incorporates a large network and multiple techniques, making a fair comparison challenging. Their approach, which formulates offline RL as an Optimal Transport problem, could incorporate our ICNN-based modeling, which may be considered a future direction.

Several studies, including Kostrikov et al. (2022); Xu et al. (2023); Garg et al. (2023); Sikchi et al. (2024), have proposed in-sample maximization approaches. These methods avoid overestimation caused by out-of-distribution actions by training exclusively with dataset actions, without sampling actions from the learned policy. Kostrikov et al. (2022) treat the value function as a distribution with inherent action-related randomness and estimates an expectile with $\tau \approx 1$ using expectile regression to approximate the optimal value function in an in-sample manner, similar to Q-learning. The policy is learned through Advantage Weighted Regression (Nair et al., 2021), where behavior cloning is weighted by an advantage function derived from the learned value function, ensuring that regularization is applied only during policy learning. Garg et al. (2023); Xu et al. (2023) propose algorithms that incorporate regularization terms based on reverse KL divergence and other f-divergence measures into the RL objective. Sikchi et al. (2024) employ a regularization term based on the visitation distribution of each policy, following Nachum & Dai (2020), where f-divergence is used as a measure. Since these in-sample maximization approaches decouple value function learning from policy learning and propose novel methods for value function training, they can be combined with our policy learning method.

Input Convex Neural Networks (ICNNs) (Amos et al., 2017) are neural networks designed such that their outputs form a convex function with respect to the inputs. Based on Brenier's theorem (Brenier, 1991), the gradient of an ICNN can be utilized as a push-forward map, enabling the modeling of the Wasserstein distance even in high-dimensional data settings (Makkuva et al., 2020; Korotin et al., 2021a;b; Mokrov et al., 2021). The use of an ICNN-based generator for minimizing the Wasserstein distance involves transforming existing data. If there is no objective such as return maximization when minimizing the Wasserstein distance using the ICNN-based generator, the transport map simply becomes an identity mapping, rendering it incapable of generating new data and thus meaningless. However, when an objective is introduced, the strength of the regularization term can be adjusted via a hyperparameter α, allowing for a gradual increase in the deviation from the identity mapping. These characteristics make Wasserstein regularization using ICNNs especially well-suited to offline RL. To the best of our knowledge, our proposed approach is the first to introduce discriminator-free Wasserstein distance regularization with ICNNs in RL. This method has the potential for further development beyond offline RL, extending to other RL settings.

6 Conclusion

In this study, we proposed a novel offline RL method that leverages Wasserstein distance as a regularization technique without requiring adversarial learning with a discriminator. By utilizing the gradient of input-convex neural networks (ICNNs) to model the optimal transport mapping, our approach effectively regularizes the learned policy while maintaining stability and efficiency. Through experiments using the D4RL benchmark dataset, we demonstrated that our method performs comparably to or better than established baseline approaches, including adversarial Wasserstein distance regularization methods that rely on a discriminator. These results highlight the effectiveness of our discriminator-free approach in mitigating distributional divergence while ensuring robust policy learning in offline RL settings. Our findings suggest that Wasserstein distance regularization via ICNN-based optimal transport mapping offers a promising direction for future research in RL.

References

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/amos17b.html.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Arip Asadulaev, Rostislav Korst, Alexander Korotin, Vage Egiazarian, Andrey Filchenkov, and Evgeny Burnaev. Rethinking optimal transport in offline reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hKloKv7pR2.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020. URL https://arxiv.org/abs/2004.07219.

Scott Fujimoto and Shixiang Gu. A minimalist approach to offline reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Divyansh Garg, Joey Hejna, Matthieu Geist, and Stefano Ermon. Extreme q-learning: Maxent RL without entropy. In *International Conference on Learning Representations*, 2023.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, San Diega, CA, USA, 2015.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=bEoxzW_EXsa.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=3tFAs5E-Pe.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. URL https://arxiv.org/abs/2005.01643.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/makkuva20a.html.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34, 2021.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020. URL https://arxiv.org/abs/2001.01866.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2021. URL https://arxiv.org/abs/2006.09359.

Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. Dual RL: Unification and new methods for reinforcement and imitation learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xt9Bu66rqv.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vqGWslLeEw.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019. URL https://arxiv.org/abs/1911.11361.

Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Victor Wai Kin Chan, and Xianyuan Zhan. Offline RL with no OOD actions: In-sample learning via implicit value regularization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ueYYgo2pSSU.

Chi Zhang, Sanmukh Kuppannagari, and Prasanna Viktor. Brac+: Improved behavior regularized actor critic for offline reinforcement learning. In Vineeth N. Balasubramanian and Ivor Tsang (eds.), *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 204–219. PMLR, 17–19 Nov 2021. URL https://proceedings.mlr.press/v157/zhang21a.html.

Zifeng Zhuang, Dengyun Peng, Jinxin Liu, Ziqi Zhang, and Donglin Wang. Reinformer: max-return sequence modeling for offline rl. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 62707–62722. PMLR, 2024.

Supplementary Materials

The following content was not necessarily subject to peer review.

7 Experimental Details

In AdvW and Q-DOT, the actor, critic, discriminator (for AdvW), and ICNN (for Q-DOT) are all two-layer MLPs with ReLU activations and 256 hidden units. The learning rate for all updates was set to 3×10^{-4} using the Adam optimizer (Kingma & Ba, 2015). The expectile parameter τ was set to the same value as in IQL: 0.7 for Mujoco locomotion tasks and Kitchen tasks, and 0.9 for Antmaze tasks. For AdvW, the parameter α was selected from the values explored in Wu et al. (2019) as well as larger values, choosing the optimal one from $(0.3, 1, 3, 10, 30, 10^2, 15^2, 20^2)$. The selected values for Mujoco locomotion, Antmaze, and Kitchen were 3, 1, and 30, respectively. In Q-DOT, α was selected from $(1, 5, 10, 20, 10^2, 20^2)$, which includes large values, because W_2^2 was often computed as the squared difference of values below 1, resulting in extremely small values. Meanwhile, β was swept over the range $(0.5, 3, 10, 20)$, which is close to the values reported in Kostrikov et al. (2022). The selected (α, β) pairs for Mujoco locomotion, Antmaze, and Kitchen were $(20, 3)$, $(20, 20)$, and $(20^2, 0.5)$, respectively. Other implementation details follow Kostrikov et al. (2022). The code is provided in the Supplementary Materials.

8 Trajectory Quality and Transport Distance

The results of other locomotion task are shown in Figure 2. In the Walker2d environment, similar to the Hopper environment, the transport distance was larger for lower-quality trajectories. In contrast, this tendency was not as clearly observed in the HalfCheetah environment. A smaller transport distance indicates that the transport that increases the advantage is not being identified by the value function. Thus, learning a value function capable of effectively transforming low-quality trajectories remains a challenge for future research in such tasks.

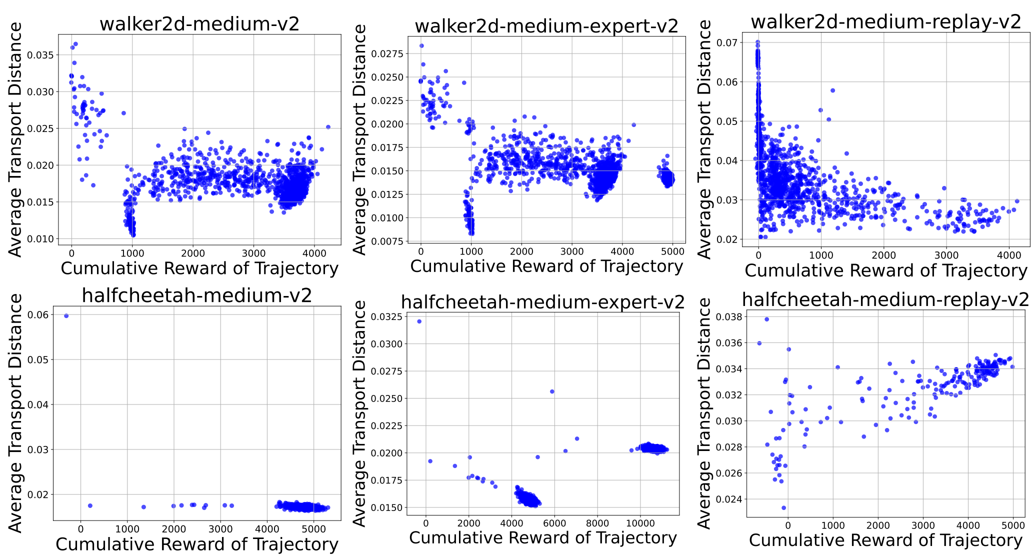

Figure 2: The relationship between trajectory quality and transport-induced distance.