Distributional Reinforcement Learning via Sinkhorn Iterations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distributional reinforcement learning (RL) is a class of state-of-the-art algorithms that estimate the whole distribution of the total return rather than only its expectation. The representation manner of each return distribution and the choice of distribution divergence are pivotal for the empirical success of distributional RL. In this paper, we propose a new class of *Sinkhorn distributional RL (SinkhornDRL)* algorithm that learns a finite set of statistics, i.e., deterministic samples, from each return distribution and then leverages Sinkhorn iterations to evaluate the Sinkhorn distance between the current and target Bellman distributions. Remarkably, Sinkhorn divergence interpolates between the Wasserstein distance and Maximum Mean Discrepancy (MMD). This allows our proposed SinkhornDRL algorithm to find a sweet spot leveraging the geometry of optimal transport based distance and the unbiased gradient estimates of MMD. Finally, experiments on the suit of 55 Atari games reveal the competitive performance of SinkhornDRL algorithm as opposed to existing state-of-the-art algorithms.

1 Introduction

Classical reinforcement learning (RL) algorithms are normally based on the expectation of discounted cumulative rewards that an agent observes while interacting with the environment. Recently, a new class of RL algorithms called *distributional RL* estimates the full distribution of total returns and has exhibited the state-of-the-art performance in a wide range of environments [2, 8, 7, 24, 26, 17].

From the literature of distributional RL, it is easily recognized that algorithms based on either Wasserstein distance or MMD have gained great attention due to their superior performance. As such, their mutual connection from the perspective of mathematical properties intrigues us to explore further in order to design new algorithms. Particularly, Wasserstein distance, long known to be a powerful tool to compare probability distributions with non-overlapping supports, has recently emerged as an appealing contender in various machine learning applications. It is known that Wasserstein distance was long disregarded because of its computational burden in its original form to solve an expensive network flow problem. However, recent works [21, 14] have shown that this cost can be largely mitigated by settling for cheaper approximations through strongly convex regularizers. The benefit of this regularization has opened the path to wider applications of the Wasserstein distance in relevant learning problems, including the design of distributional RL algorithms.

The Sinkhorn divergence [21] introduces the entropic regularization on the Wassertein distance, allowing it tractable for the evaluation especially in high-dimensions. It has been successfully applied in numerous crucial machine learning developments, including the Sinkhorn-GAN [14] and Sinkhorn-based adversarial training [23]. More importantly, it has been shown that Sinkhorn divergence interpolates Wasserstein ditance and MMD, and their equivalence form can be well established in the limit cases [11, 18, 17]. However, a Sinkhorn-based distributional RL algorithm has not yet been

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

formally proposed and its connection with algorithms based on Wasserstein distance and MMD is also less studied. Therefore, a natural question is *can we design a new class of distributional RL algorithms via Sinkhorn divergence, thus bridging the gap between existing two main branches of distributional RL algorithms?* Moreover, the dominant quantile-based algorithms, e.g., QR-DQN [8], aimed at approximating Wasserstein distance, suffers from the non-crossing issue in the quantile estimation [26], while sample-based Sinkhorn algorithm can naturally circumvent this problem.

In this paper, we propose a novel distributional RL algorithm based on *Sinkhorn divergence*. Firstly, we point out the key roles of distribution divergence and representation of value distribution in the design of distributional RL. After a detailed introduction of our proposed SinkhornDRL algorithm, we theoretically analyze its convergence guarantee and moment matching behavior of distributional Bellman operators under Sinkhorn divergence. Thus, a regularized MMD equivalence form of Sinkhorn divergence is derived, interpreting the emprical success of our algorithms in real applications. Finally, we compare the performance of our SinkhornRL algorithm with typical baselines on 55 Atari games, verifying the competitive performance of our proposal. Our approach inspires researchers to find a trade-off that simultaneously leverages the geometry of the Wasserstein distance and the favorable unbiased gradient estimate property of MMD while designing new distributional RL algorithms in the future.

2 Preliminary Knowledge

2.1 Distributional Reinforcement Learning

In the classical RL setting, an agent interacts with an environment via a Markov decision process (MDP), a 5-tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively. P is the environment transition dynamics, R is the reward function and $\gamma \in (0, 1)$ is the discount factor.

From Value function to Value distribution. Given a policy π, the discounted sum of future rewards is a random variable $Z^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$, where $s_0 = s$, $a_0 = a$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, and $a_t \sim \pi(\cdot|s_t)$. In the control setting, expectation-based RL is based on the action-value function $Q^\pi(s,a)$, which is the expectation of $Z^\pi(s,a)$, i.e., $Q^\pi(s,a) = \mathbb{E}\left[Z^\pi(s,a)\right]$. By contrast, distributional RL focuses on the action-value distribution, the full distribution of $Z^\pi(s,a)$, and the incorporation of additional distributional knowledge intuitively interprets its empirical success.

Distributional Bellman operators. For the policy evaluation in expectation-based RL, the action-value function is updated via the Bellman operator $\mathcal{T}^\pi Q(s,a) = \mathbb{E}[R(s,a)] + \gamma \mathbb{E}_{s' \sim p, \pi}\left[Q\left(s', a'\right)\right]$. In distributional RL, the action-value distribution of $Z^\pi(s,a)$ is updated via the distributional Bellman operator $\mathfrak{T}^\pi$

$$\mathfrak{T}^\pi Z(s,a) = R(s,a) + \gamma Z\left(s', a'\right), \tag{1}$$

where $s' \sim P(\cdot|s,a)$ and $a' \sim \pi\left(\cdot|s'\right)$. The equality in Eq. 1 implies that random variables of both sides are equal in distribution. The distributional Bellman operator $\mathfrak{T}^\pi$ is contractive under certain distribution divergence metrics, but the distributional Bellman optimality operator $\mathfrak{T}$ can only converge to a set of optimal non-stationary value distributions in a weak sense [9].

2.2 Divergences between Measures

Optimal Transport (OT) and Wasserstein Distance The optimal transport (OT) metric between two probability measures (μ, ν) supported on two metric spaces is defined as the solution of the linear program:

$$\min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \int c(x,y) \mathrm{d}\Pi(x,y), \tag{2}$$

where c is the cost function and Π is the joint distribution with marginals (μ, ν). Wasserstein distance (a.k.a. earth mover distance) is a special case of optimal transport with the Euclidean norm as the cost function. In particular, given two scalar random variables X and Y, p-Wasserstein metric W_p between the distributions of X and Y can be simplified as

$$W_p(X,Y) = \left(\int_0^1 \left|F_X^{-1}(\omega) - F_Y^{-1}(\omega)\right|^p d\omega\right)^{1/p}, \tag{3}$$

where F^{-1} is the inverse cumulative distribution function of a random variable. The desirable geometric property of Wasserstein distance allows it to recover full support of measures, but it suffers from the curse of dimension [13, 1].

Maximum Mean Discrepancy The squared Maximum Mean Discrepancy (MMD) MMD_k^2 with the kernel k is formulated as

$$\text{MMD}_k^2 = \mathbb{E}\left[k\left(X, X'\right)\right] + \mathbb{E}\left[k\left(Y, Y'\right)\right] - 2\mathbb{E}\left[k(X, Y)\right], \tag{4}$$

where $k(\cdot, \cdot)$ is a continuous kernel on $\mathcal{X}$. X' (resp. Y') is a random variable independent of X (resp. Y). If k is a trivial kernel, MMD degenerates to the energy distance. Mathematically, the "flat" geometry that MMD induces on the space of probability measures does not faithfully lift the ground distance [11], but MMD is cheaper to compute than OT and has a smaller sample complexity, i.e., approximating the distance with samples of measures [13]. We provide the detailed introduction of more distribution divergences in Appendix A.

3 Roles of Distribution Divergence and Representation in distributional RL

3.1 Distributional RL: From Neural Q-Fitted Iteration to Neural Z-Fitted Iteration

Neural Q-Fitted Iteration. It is known that Deep Q Learning [16] can be simplified into *Neural Q-Fitted Iteration* [10] under tricks of experience replay and the target network Q_{θ^*}, where we update parameterized $Q_\theta(s, a)$ in each iteration k:

$$Q_\theta^{k+1} = \underset{Q_\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - Q_\theta^k\left(s_i, a_i\right)\right]^2, \tag{5}$$

where the target $y_i = r(s_i, a_i) + \gamma \max_{a \in \mathcal{A}} Q_{\theta^*}^k\left(s_i', a\right)$ is fixed within every T_{target} steps to update target network Q_{θ^*} by letting $\theta^* = \theta$ and the experience buffer induces independent samples $\left\{(s_i, a_i, r_i, s_i')\right\}_{i \in [n]}$. In an ideal case that neglects the non-convexity and TD approximation errors, we have $Q_\theta^{k+1} = \mathcal{T} Q_\theta^k$, which is exactly equivalent to updating under Bellman optimality operator.

Neural Z-Fitted Iteration. Analogous to neural Q-fitted iteration, we can also simplify value-based distributional RL methods based on a parameterized Z_θ into a *Neural Z-fitted Iteration* as

$$Z_\theta^{k+1} = \underset{Z_\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} d_p(Y_i, Z_\theta^k\left(s_i, a_i\right)), \tag{6}$$

where the target $Y_i = R(s_i, a_i) + \gamma Z_{\theta^*}^k\left(s_i', \pi_Z(s')\right)$ with $\pi_Z(s') = \operatorname{argmax}_{a'} \mathbb{E}\left[Z_{\theta^*}^k(s', a')\right]$ is fixed within every T_{target} steps to update target network Z_{θ^*}, and d_p is a divergence metric between two distributions.

3.2 Key Roles of d_p and Z_θ

Within the Neural Z-fitted Iteration framework proposed in Eq. 6, we observe that the choice of representation manner on Z_θ and the metric d_p are pivotal for the distributional RL algorithms. For instance, QR-DQN [8] approximates Wasserstein distance W_p, which leverages quantiles to represent

Algorithm	d_p Distribution Divergence	Representation Z_θ	Convergence Rate of $\mathfrak{T}^\pi$	Sample Complexity of d_p
C51 [2]	Cramér distance	Histogram	$\sqrt{\gamma}$	\
QR-DQN [8]	Wasserstein distance	Quantiles	γ	$\mathcal{O}(n^{-\frac{1}{d}})$
MMDDRL [17]	MMD	Samples	$\gamma^{\alpha/2}$ with kernel k_α	$\mathcal{O}(1/n)$
SinkhornDRL (ours)	Sinkhorn divergence	Samples	γ $(\varepsilon \to 0)$ $\gamma^{\alpha/2}$ $(\varepsilon \to \infty)$	$\mathcal{O}(n^{\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}}})$ $(\varepsilon \to 0)$ $\mathcal{O}(n^{-\frac{1}{2}})$ $(\varepsilon \to \infty)$

Table 1: Comparison between typical distributional RL algorithms under different distribution divergences and represention of Z_θ. $k_\alpha = -\|x - y\|^\alpha$ in MMDDRL, d is the sample dimension and $\kappa = 2\beta d + \|c\|_\infty$, where the cost function c is β-Lipschitz [13]. Sample complexity of MMD can be improved to $\mathcal{O}(1/n)$ using kernel herding technique [5].

the distribution of Z_θ. C51 [2] represents Z_θ via a categorical distribution under the convergence of Cramér distance [3, 19], while MMD distributional RL (MMDDRL) [17] learns samples to represent the distribution of Z_θ based on MMD. We compare characteristics of these distribution divergence, including the convergence rate and sample complexity, in Table 1. Theoretical results regarding Sinkhorn divergence is based on [13] and the detailed convergence proof of other distances is also provided in Appendix A. In summary, we argue that d_p and Z_θ are two crucial factors in distributional RL design, based on which we introduce our Sinkhorn distributional RL.

4 Sinkhorn Distributional RL (SinkhornDRL)

In this section, we firstly introduce Sinkhorn divergence and apply it in distributional RL. Next, we conduct a theoretical analysis about the convergence speed and a new moment matching manner of our algorithm under the Sinkhorn divergence. Finally, a practical Sinkhorn iteration algorithm is introduced to evaluate the Sinkhorn divergence.

4.1 Sinkhorn Divergence and Genetic Algorithm

We design Sinkhorn distributional RL algorithm via Sinkhorn divergence. Sinkhorn divergence [21] is a tractable loss to approximate the optimal transport problem by leveraging an entropic regularization to turn the original Wasserstein distance into a differentiable and more robust quantity. The resulting loss can be computed using Sinkhorn fixed point iterations, which is naturally suitable for modern deep learning frameworks. In particular, the entropic smoothing generates a family of losses interpolating between Wasserstein distance and Maximum Mean Discrepancy (MMD). As such, it allows us to find a sweet trade-off that simultaneously leverages the geometry of Wasserstein distance on the one hand, and the favorable high-dimensional sample complexity and unbiased gradient estimates of MMD. We introduce the entropic regularized Wassertein distance $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$ as

$$\min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \int c(x,y) \mathrm{d}\Pi(x,y) + \varepsilon \mathrm{KL}(\Pi | \mu \otimes \nu), \tag{7}$$

where $\mathrm{KL}(\Pi|\mu \otimes \nu) = \int \log\left(\frac{\Pi(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}\right) \mathrm{d}\Pi(x,y)$ is a strongly convex regularization. The impact of this entropy regularization is similar to ℓ_2 ridge regularization in linear regression. Next, the sinkhorn loss [11, 14] between two measures μ and ν is defined as

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) = 2\mathcal{W}_{c,\varepsilon}(\mu,\nu) - \mathcal{W}_{c,\varepsilon}(\mu,\mu) - \mathcal{W}_{c,\varepsilon}(\nu,\nu). \tag{8}$$

As demonstrated by [11], the Sinkhorn divergence $\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu)$ is convex, smooth and positive definite that metrizes the convergence in law. In statistical physics, $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$ can be re-factored as a projection problem:

$$\mathcal{W}_{c,\varepsilon}(\mu,\nu) := \min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \mathrm{KL}\left(\Pi | \mathcal{K}\right), \tag{9}$$

where $\mathcal{K}$ is the Gibbs distribution with the density function satisfies $d\mathcal{K}(x,y) = e^{-\frac{c(x,y)}{\varepsilon}} d\mu(x) d\nu(y)$. This problem is often referred to as the "static Schrödinger problem" [15, 20] as it was initially considered in statistical physics.

Distributional RL with Sinkhorn Divergence and Particle Representation. The key of applying Sinkhorn divergence in distributional RL is to simply leverage the Sinkhorn loss $\overline{\mathcal{W}}_{c,\varepsilon}$ to measure the distance between the current action-value distribution $Z_\theta(s,a)$ and the target distribution $\mathfrak{T}^\pi Z_\theta(s,a)$, yielding $\overline{\mathcal{W}}_{c,\varepsilon}(Z_\theta(s,a), \mathfrak{T}^\pi Z_\theta(s,a))$ for each s, a pairs. In terms of the representation for $Z_\theta(s,a)$, we employ the unrestricted statistics, i.e., deterministic samples, due to its superiority in MMDDRL [17], instead of using predefined statistic functionals, e.g., quantiles in QR-DQN [8] or histogram partitions in C51 [2]. More concretely, we use neural networks to generate samples that approximate the value distribution. This can be expressed as $Z_\theta(s,a) := \{Z_\theta(s,a)_i\}_{i=1}^N$, where N is the number of generated samples. We refer to the samples $\{Z_\theta(s,a)_i\}_{i=1}^N$ as *particles*. Then we leverage the Dirac mixture $\frac{1}{N}\sum_{i=1}^N \delta_{Z_\theta(s,a)_i}$ to approximate the true density function of $Z^\pi(s,a)$, thus minimizing the Sinkhorn divergence between the approximate distribution and its distributional Bellman target. A detailed and generic distributional RL algorithm with Sinkhorn divergence and particle representation is provided in Algorithm 1.

Algorithm 1 Generic Sinkhorn distributional RL Update

Require: Number of generated samples N, the cost function c and hyperparameter ε.
Input: Sample transition (s, a, r', s')

if Policy evaluation **then**
  $a^* \sim \pi(\cdot|s')$.
else
  $a^* \leftarrow \arg\max_{a' \in \mathcal{A}} \frac{1}{N} \sum_{i=1}^{N} Z_\theta (s', a')_i$
end if
$\mathfrak{T} Z_i \leftarrow r + \gamma Z_{\theta^*} (s', a^*)_i, \forall 1 \leq i \leq N$

Output: $\overline{\mathcal{W}}_{c,\varepsilon} \left(\{Z_\theta(s,a)_i\}_{i=1}^N, \{\mathfrak{T} Z_\theta(s,a)_j\}_{j=1}^N \right)$

Remark. By comparing the state-of-the-art MMDDRL algorithm [17], our Sinkhorn distributional RL simply modifies the distribution divergence. Hence, we can also easily extend our generic Sinkhorn algorithm to DQN-like architecture as well as IQN [7] and FQF [24]. A following question is whether there is any theoretical connection between Sinkhorn distributional RL and algorithms based on MMD and Wasserstein distance. We provide this crucial analysis in Section 4.2

4.2 Theoretical Analysis under Sinkhorn Divergence

Convergence Analysis. Firstly, we denote the supreme form of Sinkhorn divergence as $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}(\mu, \nu)$:

$$\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}(\mu, \nu) = \sup_{(x,a) \in \mathcal{S} \times \mathcal{A}} \overline{\mathcal{W}}_{c,\varepsilon}(\mu(x,a), \nu(x,a)). \tag{10}$$

We will use $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}(\mu, \nu)$ to establish the convergence of $\mathfrak{T}^\pi$ in Theorem 1.

Theorem 1. *If we leverage Sinkhorn loss $\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu)$ in Eq. 8 as the distribution divergence in distributional RL, and* ***choose the unrectified kernel*** $k_\alpha := -\|x - y\|^\alpha$ ***as*** $-c$ ***(***$\alpha > 0$***)****, it holds that*

(1) As $\varepsilon \to 0$, $\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu) \to 2W_\alpha(\mu, \nu)$. When $\varepsilon = 0$, $\mathfrak{T}^\pi$ is a γ-contraction under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$.

(2) As $\varepsilon \to +\infty$, $\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu) \to$ MMD$^2_{k_\alpha}(\mu, \nu)$. When $\varepsilon = +\infty$, $\mathfrak{T}^\pi$ is $\gamma^{\alpha/2}$-contractive under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$.

(3) For any $\varepsilon \in (0, +\infty)$, $\mathfrak{T}^\pi$ is a ***closely*** *non-expansive operator under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$, and the difference term $\Delta(\gamma) \to 0$ as $\gamma \to 1$.*

Proof is provided in Appendix B. Theorem 1 (1) and (2) are follow-up conclusions in terms of the convergence behavior of $\mathfrak{T}^\pi$ based on the interpolation relationship between Sinkhorn divergence with Wasserstein distance and MMD [14]. Our key theoretical contribution is for the general $\varepsilon \in (0, \infty)$, the convergence behavior is determined by the "joint" KL divergence in Eq. 9 between the optimal joint distribution Π^* and the Gibbs distribution associated with the cost function c. We conclude that $\mathfrak{T}^\pi$ is a **close** non-expansive operator and the different term $\Delta(\gamma) \to 0$ as $\gamma \to 1$. Note that γ is normally very close to 1 in practice, and this is beneficial for the convergence of $\mathfrak{T}^\pi$ under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$.

Remark on Theorem 1 (3). If we consider to use Gaussian kernel, we can not guarantee $\mathfrak{T}^\pi$ is closely non-expansive for any $\varepsilon \in (0, \infty)$. This conclusion is consistent with those discussed in MMDDRL [17], where $\mathfrak{T}^\pi$ is generally not a contraction operator under MMD equipped with Gaussian kernels as a counterexample has been pointed out in MMDDRL (when $\varepsilon \to +\infty$). When $\varepsilon \to 0$, the γ-contractive $\mathfrak{T}^\pi$ under Wasserstein distance is also not contradictory to Theorem 1 (3). Moreover, although we can only obtain that $\mathfrak{T}^\pi$ is closely non-expansive, the expectation of Z^π remains a γ-contraction (see Appendix B). In experiments, we thereby use k_α and we can also demonstrate the appealing empirical performance of our SinkhornDRL algorithm in Section 5.

Regularized Moment Matching under Sinkhorn Divergence. We further examine the potential reason behind the empirical success for SinkhornDRL, although only a non-expansive contraction can be guaranteed for the general case when $\varepsilon \in (0, +\infty)$ as shown in Theorem 1. Inspired by the similar manner in MMDDRL [17], we find that the Sinkhorn divergence with the Gaussian kernel can also promote to match all moments between two distributions. More specifically, the Sinkhorn divergence can be rewritten as a regularized moment matching form in Proposition 1.

Proposition 1. *For $\varepsilon \in (0, +\infty)$, Sinkhorn divergence $\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu)$ associated with Gaussian kernels $k(x, y) = \exp(-(x-y)^2/(2\sigma^2))$ as $-c$, can be equivalent to*

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu) := \sum_{n=0}^{\infty} \frac{1}{\sigma^{2n} n!} \left(\tilde{M}_n(\mu) - \tilde{M}_n(\nu)\right)^2 + \varepsilon \mathbb{E}\left[\log \frac{(\Pi_\varepsilon^*(X, Y))^2}{\Pi_\varepsilon^*(X, X')\Pi_\varepsilon^*(Y, Y')}\right], \tag{11}$$

where Π_ε^* denotes the optimal Π determined by ε by evaluating the Sinkhorn divergence via $\min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu)$. $\tilde{M}_n(\mu) = \mathbb{E}_{x\sim\mu}\left[e^{-x^2/(2\sigma^2)} x^n\right]$, and similarly for $\tilde{M}_n(\nu)$.

We provide the proof of Proposition 1 in Appendix C. Similar to MMDDRL associated with a Gaussian kernel [17], Sinkhorn divergence approximately performs a regularized moment matching scaled by $e^{-x^2/(2\sigma^2)}$. This similar moment matching impact intuitively explains the empirical success of SinkhornDRL as MMDDRL, although the contraction of both MMD with Gaussian kernel [17] and Sinkhorn divergence for general $\epsilon \in (0, +\infty)$ may not be guaranteed.

Equivalence to Regularized MMD distributional RL. Based on Proposition 1, we can immediately establish the connection between Sinkhorn divergence and MMD in Corollary 1, indicating that minimizing Sinkhorn divergence between two distributions is equivalent to minimizing a regularized squared MMD.

Corollary 1. *For $\varepsilon \in (0, +\infty)$ and denote Π_ε^* as the optimal Π by evaluating the Sinkhorn divergence, it holds that*

$$\overline{\mathcal{W}}_{c,\varepsilon} := \mathit{MMD}_{-c}^2(\mu, \nu) + \varepsilon \mathbb{E}\left[\log \frac{(\Pi_\varepsilon^*(X, Y))^2}{\Pi_\varepsilon^*(X, X')\Pi_\varepsilon^*(Y, Y')}\right], \tag{12}$$

where we use $\overline{\mathcal{W}}_{c,\varepsilon}$ to replace $\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu)$ for short.

Proof of Corollary 1 is provided in Appendix C. It is worthy of noting that this equivalence is established for the general case when $\varepsilon \in (0, +\infty)$, and it does not hold in the limit cases when $\varepsilon \to 0$ or $+\infty$. For example, when $\varepsilon \to +\infty$, the second part including ε in Eq. 12 is not expected to dominate. This is owing to the fact that the regularization term would be 0 as $\Pi_\varepsilon^* \to \mu \otimes \nu$ when $\varepsilon \to +\infty$. In summary, even though the Sinkhorn divergence was initially proposed to serve as an entropy regularized Wasserterin distance, it turns out that it is equivalent to a regularized MMD, as revealed in Corollary 1. This connection provides strong evidence for our empirical results, in which SinkhornDRL achieves competitive performance as opposed to MMDDRL.

4.3 Distributional RL via Sinkhorn Iterations

The theoretical analysis in Section 4.2 sheds light on the behavior of distributional RL with Sinkhorn divergence, but another crucial issue we need to address is how to evaluate the Sinkhorn loss effectively. Due to the advantages of Sinkhorn divergence that both enjoys geometry property of optimal transport and the computational effectiveness of MMD, we can utilize Sinkhorn's algorithm, i.e., Sinkhorn Iterations [21, 14], to evaluate the Sinkhorn loss. Notably, Sinkhorn iteration with L steps yields a differentiable and solvable efficiently loss function as the main burden involved in it is the matrix-vector multiplication, which streams well on the GPU with simply adding extra differentiable layers on the typical deep neural network, such as a DQN architecture.

Specifically, given two sample sequences $\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N$ in the distributional RL algorithm, the optimal transport distance is equivalent to the form:

$$\min_{P \in \mathbb{R}_+^{N\times N}} \left\{\langle P, \hat{c}\rangle; P\mathbf{1}_N = \mathbf{1}_N, P^\top \mathbf{1}_N = \mathbf{1}_N\right\}, \tag{13}$$

where the empirical cost function $\hat{c}_{i,j} = c(Z_i, \mathfrak{T}Z_j)$. By adding entropic regularization on optimal transport distance, Sinkhorn divergence can be viewed to restrict the search space of P in the following scaling form:

$$P_{i,j} = a_i \mathcal{K}_{i,j} b_j, \tag{14}$$

where $\mathcal{K}_{i,j} = e^{-\hat{c}_{i,j}/\varepsilon}$ is the Gibbs kernel defined in Eq. 9. This allows us to leverage iterations regarding the vectors a and b. More specifically, we initialize $b_0 = \mathbf{1}_N$, and then the Sinkhorn iterations are expressed as

$$a_{l+1} \leftarrow \frac{\mathbf{1}_N}{\mathcal{K} b_l} \quad \text{and} \quad b_{l+1} \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}^\top a_{l+1}}, \tag{15}$$

Algorithm 2 Sinkhorn Iterations to Approximate $\overline{\mathcal{W}}_{c,\varepsilon}\left(\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N\right)$

Input: Two samples sequences $\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N$, number of Sinkhorn iterations L and hyperparameter ε.

1: $\hat{c}_{i,j} = c(Z_i, \mathfrak{T}Z_j)$ for $\forall i = 1, ..., N, j = 1, ..., N$
2: $\mathcal{K}_{i,j} = \exp(-\hat{c}_{i,j}/\varepsilon)$
3: $b_0 \leftarrow \mathbf{1}_N$
4: **for** $l = 1, 2, ..., L$ **do**
5: $\quad a_l \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}b_{l-1}}, b_l \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}a_l}$
6: **end for**
7: $\widehat{\overline{\mathcal{W}}}_{c,\varepsilon}\left(\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N\right) = \langle (K \odot \hat{c})b, a\rangle$

Return: $\widehat{\overline{\mathcal{W}}}_{c,\varepsilon}\left(\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N\right)$

where $\frac{\cdot}{\cdot}$ indicates an entry-wise division. It has been proven that Sinkhorn iteration asymptotically converges to the true loss in a linear rate [14, 12, 6]. We provide a detailed algorithm description of Sinkhorn iterations in Algorithm 2. With the efficient and differential Sinkhorn iterations, we can easily evaluate the Sinkhorn divergence and thus let our algorithm enjoy its theoretical advantages. In practice, we need to choose L and ε, and we conduct a rigorous sensitivity analysis in Section 5.

5 Experiments

We demonstrate the effectiveness of SinkhornDRL as described in Algorithm 1 on the full 55 Atari 2600 games. Specifically, we leverage the same architecture as QR-DQN [8], and replace the quantiles output with N particles, i.e., samples. In contrast to MMDDRL, SinkhornDRL only changes the distribution divergence from MMD to Sinkhorn divergence, and therefore the potential superiority in the performance can be attributed to the advantages of Sinkhorn divergence. In Section 5.1, we make a rigorous comparison between SinkhornDRL with other typical distributional RL algorithms from the perspectives of learning curves and final ratio improvement of returns. An extensive sensitivity analysis in terms of multiple hyperparameters in SinkhornDRL is provided in Section 5.2.

Baselines. Due to the interpolation characteristic of Sinkhorn divergence between Wassertein distance and MMDDRL, we choose three typical distributional RL algorithms as classic baselines, including QR-DQN [8] that approximates the Wasserstein distance, C51 [2] and MMDDRL [17], as well as DQN [16]. MMDDRL algorithm is implemented with the same architecture as QRDQN, and leverages Gaussian kernels $k_h(x, y) = \exp(-(x-y)^2/h)$ with the kernel mixture trick covering a range of bandwidths h, which is same as the basic setting in the original MMDDQN paper [17]. We deploy all algorithms on 55 Atari 2600 games, and reported results are averaged over 3 seeds with the shade indicating the standard deviation.

Hyperparameter settings. For a fair comparison with QR-DQN, C51 and MMDDRL, we used the same hyperparamters: the number of generated samples $N = 200$, Adam optimizer with $\text{lr} = 0.00005, \epsilon_{\text{Adam}} = 0.01/32$. We used a target network to compute the distributional Bellman target, which fits well in the neural Z-fitted iteration framework. In addition, we choose number of Sinkhorn iterations $L = 10$ and smoothing hyperparameter $\varepsilon = 10.0$ in Section 5.1 as they are not sensitive within a proper interval as demonstrated in Section 5.2. We choose the unrectified kernel as the cost function, i.e.,$-c = k_\alpha$, and select $\alpha = 2$ in k_α in our SinkhornDRL algorithm.

5.1 Performance of SinkhornDRL

Figure 1 illustrates that SinkhornDRL can achieve the competitive performance across 55 Atari games compared with various baseline algorithms with different metrics d_p and representation manners on Z_θ. On a large number of games, e.g., Tennis, Seaquest and Atlantis, SinkhornDRL can significantly outperform other baselines, especially on Tennis where other algorithms even fail to converge. The improvement of SinkhornDRL over MMDDRL empirically verifies the regularization advantage of the Sinkhorn as analyzed in Corollary 1. On some games, e.g., Breakout, Pong and SpaceInvaders,

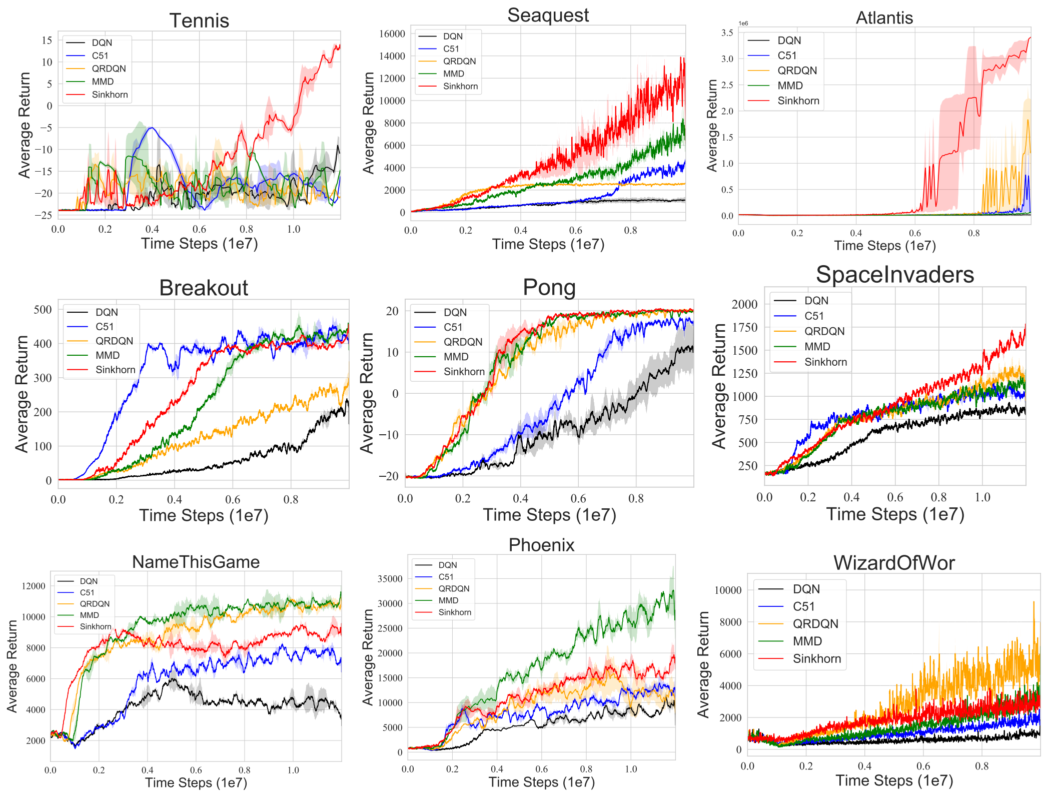


Figure 1: Learning curves of SinkhornDRL algorithm compared with DQN, C51, QR-DQN and MMD, on nine typical Atari games over 3 seeds.

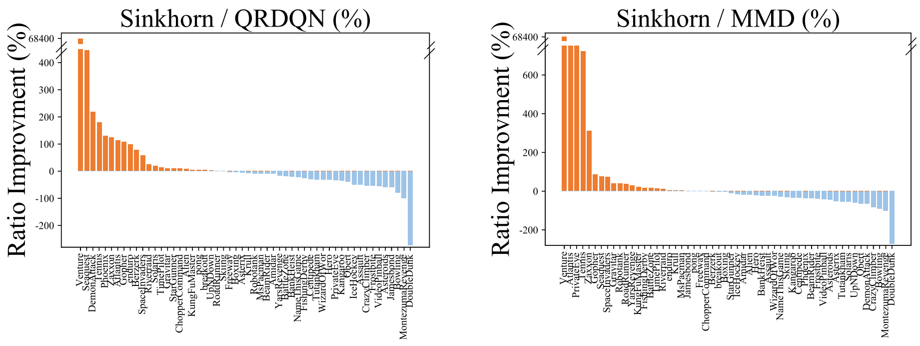
SinkhornDRL is on par with MMDDRL and other baselines, while on the last row in Figure 1, SinkhornDRL is slightly inferior to the state-of-the-art algorithm. We provide learning curves of all typical distributional RL algorithms on all 55 Atari games in Appendix E, where SinkhornDRL still achieves the competitive performance in general.

To further demonstrate theoretical properties of SinkhornDRL in Theorem 1, we conduct a ratio improvement comparison across 55 Atari games between SinkhornDRL with QRDQN and MMDDRL, respectively. Figure 2 showcases that by comparing with QRDQN (left), SinkhornDRL achieves better performance across more than half of considered games. More importantly, the superiority of SinkhornDRL is significant across a large amount of games, including Venture, Seaquest, Tennis and Phoenix. This empirical outperformance verifies the effectiveness and potential of smoothing Wassertein distance in distributional RL, e.g., Sinkhorn divergence. In contrast with MMDDRL, the superiority of SinkhornDRL is reduced with the performance improvement only on a small proportion of games, while a remarkable boost of performance for SinkhornDRL on a large amount of games can be easily observed. We also report mean and median of best human-normalized scores in Table 2 of Appendix D, where SinkhornDRL achieves almost state-of-the-art performance as MMDDRL on average.

Therefore, we conclude that SinkhornDRL is competitive with the state-of-the-art distributional RL algorithms, e.g., MMDDRL, and can be extremely superior over existing algorithms on a large proportion of games. This empirical success can be owing to theoretical advantage of Sinkhorn divergence that simultaneously makes full use of the data geometry from Wasserstein distance and the unbiased gradient estimate property from MMD, which coincides with results in Theorem 1.

5.2 Sensitivity Analysis and Computational Cost

Figure 3 (a) suggests the performance of our algorithm is robust to ε in a certain range, e.g., $[1, 500]$, facilitating its deployment in practice. If we increase ε, SinkhornDRL's performance tends to MMD,

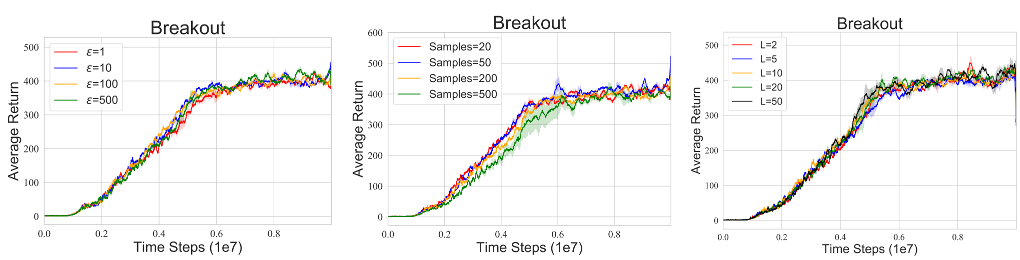


(a) Sinkhorn vs QRDQN (b) Sinkhorn vs MMDDRL

Figure 2: Ratio improvement of return for Sinkhorn distributional RL algorithm over QRDQN (left) and MMDDRL (right) over 3 seeds. For example, the ratio improvement is calculated by (Sinkhorn - QRDQN) / QRDQN in the left.

while if we gradually decline ε, SinkhornDRL's performance tends to QR-DQN. It is also noted that Sinkhorn iterations in Algorithm 2 will suffer from the numerical instability issue under an overly small or large ε. More results with the discussion are provided in Appendix F. It is also illustrated that our algorithm is insensitive to the number of iterations L and samples N as well, but an overly large N can slightly worsen the performance of SinkhornDRL, and at the same time increases the computational burden. Therefore, a proper number of samples, e.g., 200, is sufficient to attain an appealing performance with the computational effectiveness.

For the computation cost, SinkhronDRL indeed increases around 50$\%$ computation cost compared with QR-DQN and C51, but only slightly increases the cost (by around 20$\%$) in contrast to MMDDRL. Detailed comparison is given in Appendix F.



(a) Hyper-parameter ε (b) Number of Samples (c) Sinkhorn Iterations L

Figure 3: Sensitivity analysis of SinkhornDRL on Breakout regarding ε, number of samples, and number of iteration L. Learning curves are reported over 3 seeds.

6 Discussions and Conclusion

The main limitation of our proposal is that the superiority over existing state-of-the-art algorithms may not be sufficiently significant. To extend our algorithm for better performance, implicit generative models, including parameterizing the cost function in Sinkhorn loss, can be further incorporated. We leave it as the future work. Moreover, other divergences, e.g., thoses that can also smooth Wassertein distance, can also be applied into the design of distributional RL algorithms in the future.

In this paper, a novel family of distributional RL algorithms based on Sinkhorn Divergence is proposed that accomplishes a competitive performance compared with the-state-of-the-art distributional RL algorithms on 55 Atari games. Theoretical analysis about the convergence and moment matching behavior is provided along with a rigorous empirical verification. Albeit being associated with MMD algorithm, distributional RL with Sinkhorn divergence is complementary to previous algorithms, leading to an important contribution among the research community.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *International Conference on Machine Learning (ICML)*, 2017.

[3] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[4] Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G Bellemare. Dopamine: A research framework for deep reinforcement learning. *CoRR abs/1812.06110*, 2018.

[5] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. *UAI, 109–116. AUAI Press*, 2012.

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[7] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. *International Conference on Machine Learning (ICML)*, 2018.

[8] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[9] Odin Elie and Charpentier Arthur. *Dynamic Programming in Distributional Reinforcement Learning*. PhD thesis, Université du Québec à Montréal, 2020.

[10] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

[11] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[12] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[13] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[14] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

[15] Christian Léonard. A survey of the schr\" odinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[17] Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning with maximum mean discrepancy. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[18] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[19] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

[20] Ludger Rüschendorf and Wolfgang Thomsen. Closedness of sum spaces andthe generalized schrödinger problem. *Theory of Probability & Its Applications*, 42(3):483–494, 1998.

[21] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[22] Gábor J Székely. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18, 2003.

[23] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pages 6808–6817. PMLR, 2019.

[24] Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32:6193–6202, 2019.

[25] Shangtong Zhang. Modularized implementation of deep rl algorithms in pytorch. `https://github.com/ShangtongZhang/DeepRL`, 2018.

[26] Fan Zhou, Jianing Wang, and Xingdong Feng. Non-crossing quantile regression for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[27] Florian Ziel. The energy distance for ensemble and scenario reduction. *arXiv preprint arXiv:2005.14670*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We provide the discussion about the limitation of our proposal in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Please refer to Appendix B and C.
   (b) Did you include complete proofs of all theoretical results? [Yes] Please refer to Appendix B and C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Our implementation is adapted from Pytorch distributional RL modules [25].
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide the comparison of computational cost in Figure 12 of Appendix F.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Definition of distances and Contraction

Definition of distances. Given two random variables X and Y, p-Wasserstein metric W_p between the distributions of X and Y is defined as

$$W_p(X,Y) = \left(\int_0^1 \left| F_X^{-1}(\omega) - F_Y^{-1}(\omega) \right|^p d\omega \right)^{1/p} = \|F_X^{-1} - F_Y^{-1}\|_p, \tag{16}$$

which F^{-1} is the inverse cumulative distribution function of a random variable with the cumulative distribution function as F. Further, ℓ_p distance [9] is defined as

$$\ell_p(X,Y) := \left(\int_{-\infty}^{\infty} |F_X(\omega) - F_Y(\omega)|^p \, \mathrm{d}\omega \right)^{1/p} = \|F_X - F_Y\|_p \tag{17}$$

The ℓ_p distance and Wassertein metric are identical at $p = 1$, but are otherwise distinct. Note that when $p = 2$, ℓ_p distance is also called Cramér distance [3] $d_C(X,Y)$. Also, the Cramér distance has a different representation given by

$$d_C(X,Y) = \mathbb{E}|X - Y| - \frac{1}{2}\mathbb{E}\left|X - X'\right| - \frac{1}{2}\mathbb{E}\left|Y - Y'\right|, \tag{18}$$

where X' and Y' are the i.i.d. copies of X and Y. Energy distance [22, 27] is a natural extension of Cramér distance to the multivariate case, which is defined as

$$d_E(\mathbf{X},\mathbf{Y}) = \mathbb{E}\|\mathbf{X} - \mathbf{Y}\| - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \mathbf{X}'\| - \frac{1}{2}\mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|, \tag{19}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are multivariate. Moreover, the energy distance is a special case of the maximum mean discrepancy (MMD), which is formulated as

$$\text{MMD}(\mathbf{X},\mathbf{Y};k) = \left(\mathbb{E}\left[k\left(\mathbf{X},\mathbf{X}'\right)\right] + \mathbb{E}\left[k\left(\mathbf{Y},\mathbf{Y}'\right)\right] - 2\mathbb{E}[k(\mathbf{X},\mathbf{Y})]\right)^{1/2} \tag{20}$$

where $k(\cdot,\cdot)$ is a continuous kernel on $\mathcal{X}$. In particular, if k is a trivial kernel, MMD degenerates to energy distance. Additionally, we further define the supreme MMD, which is a functional $\mathcal{P}(\mathcal{X})^{\mathcal{S}\times\mathcal{A}} \times \mathcal{P}(\mathcal{X})^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ defined as

$$\text{MMD}_\infty(\mu,\nu) = \sup_{(x,a)\in\mathcal{S}\times\mathcal{A}} \text{MMD}_\infty(\mu(x,a),\nu(x,a)) \tag{21}$$

We further present the convergence rate under different distribution divergences.

- $\mathcal{T}^\pi$ is γ-contractive under the supreme form of Wassertein distance W_p.
- $\mathcal{T}^\pi$ is $\gamma^{1/p}$-contractive under the supreme form of ℓ_p distance.
- $\mathcal{T}^\pi$ is $\gamma^{\alpha/2}$-contractive under MMD_∞ with the kernel $k_\alpha(x,y) = -\|x-y\|^\alpha, \forall\alpha > 0$.

Proof of Contraction.

- Contraction under supreme form of Wasserstein diatance is provided in Lemma 3 [2].
- Contraction under supreme form of ℓ_p distance can refer to Theorem 3.4 [9].
- Contraction under MMD_∞ is provided in Lemma 6 [17].

B Proof of Theorem 1

Proof. **1.** As $\varepsilon \to 0$ and $c = -k_\alpha$, it is obvious to observe that Sinkhorn loss degenerates to the wasserstein distance. We also have the conclusion that the distributional Bellman operator $\mathfrak{T}^\pi$ is γ-contractive under the supreme form of Wasserstein diatance, the proof of which is provided in Lemma 3 [2]. Since the above conclusion is made directly based on the limiting case when $\varepsilon = 0$, for an unspecified ε, we need a more rigorous proof. We show that their distance difference is **at most an infinitesimal** δ.

Firstly, as $\mathcal{W}_{c,\varepsilon} \to W_\alpha$ and the regularization term is non-negative, using the language of (ε, δ) definition, we have: for $\forall\delta$, there exists a small positive constant a, such that $\mathcal{W}_{c,\varepsilon} - W_\alpha < \delta$ when $\epsilon \leq a$. Based on that, we have the contraction conclusion:

$$\begin{aligned}\overline{\mathcal{W}}^{\infty}_{-\kappa_\alpha,\varepsilon}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) &= \overline{\mathcal{W}}^{\infty}_{-\kappa_\alpha,\varepsilon}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) - W^{\infty}_{\alpha}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) + W^{\infty}_{\alpha}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) \\ &\leq \delta + W^{\infty}_{\alpha}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2),\end{aligned} \tag{22}$$

where the second term $W^{\infty}_{\alpha}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2)$ is contractive, and thus for the unspecified ε, the only difference from the limting $\varepsilon = 0$ is an infinitesimal δ, which will vanish as $\varepsilon \to 0$ or $(a \to 0)$.

2. As $\varepsilon \to \infty$, our complete proof is inspired by [18, 14]. Recap the definition of squared MMD is

$$\mathbb{E}\left[k\left(\mathbf{X}, \mathbf{X}'\right)\right] + \mathbb{E}\left[k\left(\mathbf{Y}, \mathbf{Y}'\right)\right] - 2\mathbb{E}[k(\mathbf{X}, \mathbf{Y})]$$

When the kernel function k degenerates to a unrectified $k_\alpha(x, y) := -\|x - y\|^\alpha$ for $\alpha \in (0, 2)$, the squared MMD would degenerate to

$$\mathbb{E}\|\mathbf{X} - \mathbf{X}'\|^\alpha + \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|^\alpha - 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\alpha$$

On the other hand, we have the Sinkhorn loss as

$$\overline{\mathcal{W}}_{c,\infty}(\mu, \nu) = 2\mathcal{W}_{c,\infty}(\mu, \nu) - \mathcal{W}_{c,\infty}(\nu, \nu) - \mathcal{W}_{c,\infty}(\mu, \nu)$$

Denoting Π_ε be the unique minimizer for $\overline{\mathcal{W}}_{c,\varepsilon}$, it holds that $\Pi_\varepsilon \to \mu \otimes \nu$ as $\varepsilon \to \infty$. That being said, $\mathcal{W}_{c,\infty}(\mu, \nu) \to \int c(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) + 0 = \int c(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$. If $c = -k_\alpha = \|x - y\|^\alpha$, we eventually have $\mathcal{W}_{-k_\alpha,\infty}(\mu, \nu) \to \int \|x - y\|^\alpha \mathrm{d}\mu(x)\mathrm{d}\nu(y) = \mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\alpha$. Finally, we can have

$$\overline{\mathcal{W}}_{-k_\alpha,\infty} \to 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\alpha - \mathbb{E}\|\mathbf{X} - \mathbf{X}'\|^\alpha - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|^\alpha$$

which is exactly the form of squared MMD. Now the key is prove that $\Pi_\varepsilon \to \mu \otimes \nu$ as $\varepsilon \to \infty$.

Firstly, it is apparent that $\mathcal{W}_{c,\varepsilon}(\mu, \nu) \leq \int c(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$ as $\mu \otimes \nu \in \Pi(\mu, \nu)$. Let $\{\varepsilon_k\}$ be a positive sequence that diverges to ∞, and Π_k be the corresponding sequence of unique minimizers for $\mathcal{W}_{c,\varepsilon}$. According to the optimality condition, it must be the case that $\int c(x, y)\mathrm{d}\Pi_k + \varepsilon_k \mathrm{KL}(\Pi_k, \mu \otimes \nu) \leq \int c(x, y)\mathrm{d}\mu \otimes \nu + 0$ (when $\Pi(\mu, \nu) = \mu \otimes \nu$). Thus,

$$\mathrm{KL}\left(\Pi_k, \mu \otimes \nu\right) \leqslant \frac{1}{\varepsilon_k}\left(\int c \,\mathrm{d}\mu \otimes \nu - \int c \,\mathrm{d}\Pi_k\right) \to 0.$$

Besides, by the compactness of $\Pi(\mu, \nu)$, we can extract a converging subsequence $\Pi_{n_k} \to \Pi_\infty$. Since KL is weakly lower-semicontinuous, it holds that

$$\mathrm{KL}\left(\Pi_\infty, \mu \otimes \nu\right) \leqslant \liminf_{k \to \infty} \mathrm{KL}\left(\Pi_{n_k}, \mu \otimes \nu\right) = 0$$

Hence $\Pi_\infty = \mu \otimes \nu$. That being said that the optimal coupling is simply the product of the marginals, indicating that $\Pi_\varepsilon \to \mu \otimes \nu$ as $\varepsilon \to \infty$. As a special case, when $\alpha = 1$, $\overline{\mathcal{W}}_{-k_1,\infty}(u, v)$ is equivalent to the energy distance

$$d_E(\mathbf{X}, \mathbf{Y}) := 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\| - \mathbb{E}\|\mathbf{X} - \mathbf{X}'\| - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|. \tag{23}$$

In summary, if the cost function is the rectified kernel k_α, it is the case that $\overline{\mathcal{W}}_{-k_\alpha,\varepsilon}$ converges to the squared MMD as $\varepsilon \to \infty$. According to [17], $\mathfrak{T}^\pi$ is $\gamma^{\alpha/2}$-contractive in the supreme form of MMD with the rectified kernel k_α.

For the unspecified ε, we can get the similar result to the case of $\varepsilon \to 0$. For $\forall\delta$, there exists a large positive constant M, such that $\mathrm{MMD}^2_{k_\alpha} - \mathcal{W}_{c,\varepsilon} < \delta$ when $\epsilon \geq M$. Based on that, we have the contraction conclusion:

$$\begin{aligned}\overline{\mathcal{W}}^{\infty}_{-\kappa_\alpha,\varepsilon}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) &= \overline{\mathcal{W}}^{\infty}_{-\kappa_\alpha,\varepsilon}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) - \mathrm{MMD}^2_\infty(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) + \mathrm{MMD}^2_\infty(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) \\ &\leq \mathrm{MMD}^2_\infty(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) - \delta,\end{aligned} \tag{24}$$

where the first term $\mathrm{MMD}^2_\infty(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2)$ is $\gamma^{\frac{\alpha}{2}}\}$-contractive, and thus for the unspecified ε, the only difference from the limiting $\varepsilon = \infty$ is an infinitesimal δ, which will vanish as $\varepsilon \to +\infty$ or $(M \to +\infty)$.

3. For $\varepsilon \in (0, +\infty)$, a key observation for the analysis is that the Sinkhorn divergence would degenerate to a two-dimensional KL divergence, and therefore embraces a similar convergence behavior to KL divergence. Concretely, according to the equivalent form of $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$ in Eq. 9, it can be expressed as the KL divergence between an optimal joint distribution and a Gibbs distribution associated with the cost function:

$$\mathcal{W}_{c,\varepsilon}(\mu,\nu) := \mathrm{KL}\left(\Pi^*(\mu,\nu)|\mathcal{K}(\mu,\nu)\right), \tag{25}$$

where Π^* is the optimal joint distribution. Thus, the total Sinkhorn divergence is expressed as

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) := 2\mathrm{KL}\left(\Pi^*(\mu,\nu)|\mathcal{K}(\mu,\nu)\right) - \mathrm{KL}\left(\Pi^*(\mu,\mu)|\mathcal{K}(\mu,\mu)\right) - \mathrm{KL}\left(\Pi^*(\nu,\nu)|\mathcal{K}(\nu,\nu)\right). \tag{26}$$

Due to the form of $\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu)$, the convergence behavior is determined by $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$, which is similar to the behavior of KL divergence. Thus, we will focus on the convergence analysis of $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$. We firstly elaborate a Lemma regarding to the convergence under KL divergence.

Lemma 1. *Denote the supreme of D_{KL} as D_{KL}^{∞}, we have: (1) $\mathfrak{T}^\pi$ is a non-expansive operator under D_{KL}^{∞}, i.e., $D_{KL}^{\infty}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) \leq D_{KL}^{\infty}(Z_1, Z_2)$, (2) the expectation of Z^π is still γ-contractive under D_{KL}^{∞}, i.e., $\|\mathbb{E}\mathfrak{T}^\pi Z_1 - \mathbb{E}\mathfrak{T}^\pi Z_2\|_\infty \leq \gamma \|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_\infty$.*

Proof. (1) We recap three crucial properties of a divergence metric. The first is *scale sensitive* **(S)** (of order β, $\beta > 0$), i.e., $d_p(cX, cY) \leq |c|^\beta d_p(X,Y)$. The second property is *shift invariant* **(I)**, i.e., $d_p(A+X, A+Y) \leq d_p(X,Y)$. The last one is *unbiased gradient* **(U)**. We use p and q to denote the density function of two random variables X and Y, and thus $D_{\mathrm{KL}}(X,Y)$ is defined as $D_{\mathrm{KL}}(X,Y) = \int_{-\infty}^{\infty} p(x)\frac{p(x)}{q(x)}\,\mathrm{d}x$. Firstly, we show that $D_{\mathrm{KL}}(X,Y)$ is NOT scale sensitive:

$$\begin{aligned} D_{\mathrm{KL}}(aX, aY) &= \int_{-\infty}^{\infty} \frac{1}{a}p(\frac{x}{a}) \log \frac{\frac{1}{a}p(\frac{x}{a})}{\frac{1}{a}q(\frac{x}{a})}\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} p(y)\log\frac{p(y)}{q(y)}\,\mathrm{d}y \\ &= D_{\mathrm{KL}}(X,Y), \text{ with } \beta = 0 \end{aligned} \tag{27}$$

We further show that $D_{\mathrm{KL}}(X,Y)$ is shift invariant:

$$\begin{aligned} D_{\mathrm{KL}}(A+X, A+Y) &= \int_{-\infty}^{\infty} p(x-A)\log\frac{p(x-A)}{q(x-A)}\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} p(y)\log\frac{p(y)}{q(y)}\,\mathrm{d}y \\ &= D_{\mathrm{KL}}(X,Y) \end{aligned} \tag{28}$$

Moreover, it is well-known that KL divergence has unbiased sample gradients [3]. The supreme D_{KL} is a functional $\mathcal{P}(\mathcal{X})^{\mathcal{S}\times\mathcal{A}} \times \mathcal{P}(\mathcal{X})^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ defined as

$$D_{\mathrm{KL}}^{\infty}(\mu,\nu) = \sup_{(x,a)\in\mathcal{S}\times\mathcal{A}} D_{\mathrm{KL}}(\mu(x,a), \nu(x,a)) \tag{29}$$

Therefore, we prove $\mathfrak{T}^\pi$ is at best a non-expansive operator under the supreme form of D_{KL}:

$$\begin{aligned} &D_{\mathrm{KL}}^{\infty}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) \\ &= \sup_{s,a} D_{\mathrm{KL}}(\mathfrak{T}^\pi Z_1(s,a), \mathfrak{T}^\pi Z_2(s,a)) \\ &= \sup_{s,a} D_{\mathrm{KL}}(R(s,a) + \gamma Z_1(S',A'), R(s,a) + \gamma Z_2(S',A')) \\ &= D_{\mathrm{KL}}(Z_1(S',A'), Z_2(S',A')) \\ &\leq \sup_{s',a'} D_{\mathrm{KL}}(Z_1(s',a'), Z_2(s',a')) \\ &= D_{\mathrm{KL}}^{\infty}(Z_1, Z_2) \end{aligned} \tag{30}$$

There we have $D_{\mathrm{KL}}^{\infty}(\mathfrak{T}^{\pi}Z_1, \mathfrak{T}^{\pi}Z_2) \leq D_{\mathrm{KL}}^{\infty}(Z_1, Z_2)$, implying that $\mathfrak{T}^{\pi}$ is a non-expansive operator under D_{KL}^{∞}.

(2) This statement is an immediate conclusion based on the Lemma 4 in [2]. We give the proof for the completeness. This conclusion holds because the $\mathfrak{T}^{\pi}$ degenerates to $\mathcal{T}^{\pi}$ regardless of the metric d_p. Specifically, due to the linearity of expectation, we obtain that

$$\|\mathbb{E}\mathfrak{T}^{\pi}Z_1 - \mathbb{E}\mathfrak{T}^{\pi}Z_2\|_{\infty} = \|\mathcal{T}^{\pi}\mathbb{E}Z_1 - \mathcal{T}^{\pi}\mathbb{E}Z_2\|_{\infty} \leq \gamma\|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_{\infty}. \tag{31}$$

This implies that the expectation of Z under D_{KL} exponentially converges to the expectation of Z^*, i.e., γ-contraction.

□

We show that $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$ is NOT scale sensitive. Firstly, we denote Π^2 as the optimal joint distribution for (U,V) and thus we write the explicit form of Sinkhorn divergence $W_{c,\varepsilon}(U,V)$ between two random variables U and V:

***

$$\begin{aligned} W_{c,\varepsilon}(U,V) &= \mathrm{KL}(\Pi^2||\mathcal{K}) && (32)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Pi^2(x,y)\log\frac{\Pi^2(x,y)}{\frac{1}{Z_2}e^{-\frac{c(x,y)}{\epsilon}}\mu(x)\nu(y)}dxdy, && (33)\end{aligned}$$

***

where the normalization factor Z_2 for the Gibbs kernel $\mathcal{K}$ is $Z_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-\frac{c(x,y)}{\epsilon}}\mu(x)\nu(y)dxdy$ and $\mu(x),\nu(y)$ are the marginal density funtion of U and V with respect to x and y. We also denote Π^1 as the optimal joint distribution for (aU, aV). A key proof element is about the Gibbs kernel $\mathcal{K}$. By definition, the pdf of $\mathcal{K}(U,V) \propto e^{\frac{-c(x,y)}{\varepsilon}}\mu(x)\nu(y)$. After a scaling transformation, the pdf of aU and aV with respect to x and y would be $\frac{1}{a}\mu(\frac{x}{a})$ and $\frac{1}{a}\nu(\frac{y}{a})$. Thus $\mathcal{K}(2U,2V) \propto e^{\frac{-c(x,y)}{\varepsilon}}\frac{1}{a}\mu(\frac{x}{a})\frac{1}{a}\nu(\frac{y}{a})$. The new normalization factor Z_1 is $Z_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{a^2}e^{-\frac{c(x',y')}{\epsilon}}\mu(x'/a)\nu(y'/a)dx'dy' = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-\frac{c(ax,ay)}{\epsilon}}\mu(x)\nu(y)dxdy$, the cost function of which is different from Z_2. For $\Pi^2(U,V)$, the scaled pdf of $\Pi^2(aU,aV)$ would be $\frac{1}{a^2}\Pi^2(\frac{x}{a},\frac{y}{a})$. Then we have the following results:

***

$$\begin{aligned} \mathcal{W}_{c,\varepsilon}(aU,aV) &= \mathrm{KL}(\Pi^1||\mathcal{K}) && (34)\\ &\leq \mathrm{KL}(\Pi^2||\mathcal{K}) && (35)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{a^2}\Pi^2(\frac{x'}{a},\frac{y'}{a})\log\frac{\frac{1}{a^2}\Pi^2(\frac{x'}{a},\frac{y'}{a})}{\frac{1}{a^2}\frac{1}{Z_1}e^{-\frac{c(x',y')}{\epsilon}}\mu(\frac{x'}{a})\nu(\frac{y'}{a})}dx'dy', && (36)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Pi^2(x,y)\log\frac{\Pi^2(x,y)}{\frac{1}{Z_1}e^{-\frac{c(ax,ay)}{\epsilon}}\mu(x)\nu(y)\frac{Z_2}{Z_2}}dxdy, && (37)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Pi^2(x,y)(\log\frac{\Pi^2(x,y)}{\frac{1}{Z_1}e^{-\frac{c(ax,ay)}{\epsilon}}\mu(x)\nu(y)} + \log\frac{Z_1}{Z_2})dxdy, && (38)\\ &\overset{c=-k_\alpha, a\leq 1}{\leq} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Pi^2(x,y)\log\frac{\Pi^2(x,y)}{\frac{1}{Z_1}e^{-\frac{c(x,y)}{\epsilon}}\mu(x)\nu(y)}dxdy + \log\frac{Z_1}{Z_2}\cdot 1, && (39)\\ &= \mathcal{W}_{c,\varepsilon}(U,V) + \Delta_{\mu,\nu}^{c}(a), && (40)\end{aligned}$$

***

where the second positive term $\Delta_{\mu,\nu}^{c}(a) = \log\frac{Z_1}{Z_2}$ satisfies $\Delta_{\mu,\nu}^{c}(a) \to 0$ as $a \to 1$ (in practice γ is very close to 1). The second inequality holds for the general ε because for the unrectified kernel $k_\alpha = -\|x-y\|^\alpha$ **with** $a \leq 1$, for any ε and x, y we have

$$(ax-ay)^\alpha \leq |a|^\alpha(x-y)^\alpha \leq (x-y)^\alpha e^{-\frac{c(ax,ay)}{\varepsilon}} \geq e^{-\frac{c(x,y)}{\varepsilon}} \Rightarrow e^{-\frac{c(ax,ay)}{\varepsilon}}\mu(x)\nu(y) \geq e^{-\frac{c(x,y)}{\varepsilon}}\mu(x)\nu(y)$$

However, under this condition, $Z_1 \geq Z_2$ and thus $\Delta^c_{\mu,\nu}(a) \geq 0$, but $\Delta^c_{\mu,\nu}(a) \to 0$ as $a \to 1$ (in practice γ is very close to 1). We think there is indeed a gap between a (close) non-expansion property of Sinkhorn divergence and the empirical success of SinkhornDRL algorithm. The inequality is established based on the unrectified kernel, but it is tricky to find the contrative property for Sinkhorn divergence with the Gaussian kernel for any ε and x, y. Thus, it is fair that some counterexamples may exist for the non-contractive $\mathfrak{T}^\pi$ under Sinkhorn divergence, which is also consistent with the counterexample MMD with Gaussian kernel (when $\varepsilon \to \infty$).

Now we show that $\mathcal{W}_{c,\varepsilon}$ is shift invariant:

$$\begin{aligned}\mathcal{W}_{c,\varepsilon}(A+X, A+Y) &= \int_{-\infty}^{\infty} \Pi^*(x-A, y-A) \log \frac{\Pi^*(x-A, y-A)}{\frac{1}{\mathcal{Z}} e^{-\frac{c(x-A,y-A)}{\varepsilon}}} \, \mathrm{d}x \, \mathrm{d}y \\ &= \mathcal{W}_{c,\varepsilon}(X, Y).\end{aligned} \tag{41}$$

According to the equation of $\overline{\mathcal{W}}_{c,\varepsilon}$, it holds the same properties as $\mathcal{W}_{c,\varepsilon}$, i.e., shift invariant and scale sensitive. Thus, we derive the convergence of distributional Bellman operator $\mathfrak{T}^\pi$ under the supreme form of $\overline{\mathcal{W}}_{c,\varepsilon}$, i.e., $\overline{\mathcal{W}}^\infty_{c,\varepsilon}$:

$$\begin{aligned}&\overline{\mathcal{W}}^\infty_{c,\varepsilon}(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) \\ &= \sup_{s,a} \overline{\mathcal{W}}_{c,\varepsilon}(\mathfrak{T}^\pi Z_1(s,a), \mathfrak{T}^\pi Z_2(s,a)) \\ &= \overline{\mathcal{W}}_{c,\varepsilon}(R(s,a) + \gamma Z_1(s',a'), R(s,a) + \gamma Z_2(s',a')) \\ &\leq \overline{\mathcal{W}}_{c,\varepsilon}(Z_1(s',a'), Z_2(s',a')) + \Delta^{-k_\alpha}_{s',a',s,a}(\gamma) \\ &\leq \sup_{s',a'} \overline{\mathcal{W}}_{-k_\alpha,\varepsilon}(Z_1(s',a'), Z_2(s',a')) + \sup_{s,a,s',a'} \Delta^{-k_\alpha}_{s',a',s,a}(\gamma) \\ &= \overline{\mathcal{W}}^\infty_{-k_\alpha,\varepsilon}(Z_1, Z_2) + \Delta(\gamma)\end{aligned} \tag{42}$$

where the first inequality comes from the scale sensitivity proof, and we denote $\sup_{s,a,s',a'} \Delta^{-k_\alpha}_{s',a',s,a}(\gamma) = \Delta(\gamma)$ for short. Since $\Delta(\gamma) \to 0$ as $\gamma \to 1$, we can conclude that $\mathfrak{T}^\pi$ is **closely** a non-expansive operator regardless of the cost function form c when $\varepsilon \in (0, \infty)$. The γ-contraction of the expectation of Z^π can be similarly proved as the KL divergence in Lemma 1. □

C Proof of Proposition 1 and Corollary 1

Proof. As we leverage Π^* to denote the optimal Π by evaluating the Sinkhorn divergence via $\min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu; k)$, the Sinkhorn divergence can be composed in the following form:

$$\begin{aligned}&\overline{\mathcal{W}}_{c,\varepsilon}(\mu, \nu; k) \\ &= 2\mathrm{KL}\left(\Pi^*(\mu,\nu) | \mathcal{K}_{-k}(\mu,\nu)\right) - \mathrm{KL}\left(\Pi^*(\mu,\mu) | \mathcal{K}_{-k}(\mu,\mu)\right) - \mathrm{KL}\left(\Pi^*(\nu,\nu) | \mathcal{K}_{-k}(\nu,\nu)\right) \\ &= 2(\mathbb{E}_{X,Y}\left[\log \Pi^*(\mu,\nu)\right]) + \frac{1}{\varepsilon}\mathbb{E}_{X,X'}\left[c(X,Y)\right]) - (\mathbb{E}_{X,X'}\left[\log \Pi^*(\mu,\nu)\right]) + \frac{1}{\varepsilon}\mathbb{E}_{X,Y}\left[c(X,Y)\right]) \\ &- (\mathbb{E}_{Y,Y'}\left[\log \Pi^*(\nu,\nu)\right]) + \frac{1}{\varepsilon}\mathbb{E}_{Y,Y'}\left[c(Y,Y')\right]) \\ &= \mathbb{E}_{X,X',Y,Y'}\left[\log \frac{(\Pi^*(X,Y))^2}{\Pi^*(X,X')\Pi^*(Y,Y')}\right] + \frac{1}{\varepsilon}(\mathbb{E}_{X,X'}\left[k(X,X')\right] + \mathbb{E}_{Y,Y'}\left[k(Y,Y')\right] - 2\mathbb{E}_{X,X'}\left[k(X,Y)\right]) \\ &= \mathbb{E}_{X,X',Y,Y'}\left[\log \frac{(\Pi^*(X,Y))^2}{\Pi^*(X,X')\Pi^*(Y,Y')}\right] + \frac{1}{\varepsilon}\mathrm{MMD}^2_{-c}(\mu,\nu)\end{aligned} \tag{43}$$

where the cost function c in the Gibbs distribution $\mathcal{K}$ is minus Gaussian kernel, i.e., $c(x,y) = -k(x,y) = e^{-(x-y)/(2\sigma^2)}$. Till now, we have shown the result in Corollary 1.

Next, we use Taylor expansion to prove the moment matching of MMD. Firstly, we have the following equation:

$$\begin{aligned}
\mathrm{MMD}^2_{-c}(\mu,\nu) &= \mathbb{E}_{X,X'}\left[k(X,X')\right] + \mathbb{E}_{Y,Y'}\left[k(Y,Y')\right] - 2\mathbb{E}_{X,X'}\left[k(X,Y)\right] \\
&= \mathbb{E}_{X,X'}\left[\phi(X)^\top\phi(X')\right] + \mathbb{E}_{Y,Y'}\left[\phi(Y)^\top\phi(Y')\right] - 2\mathbb{E}_{X,X'}\left[\phi(X)^\top\phi(Y)\right] \\
&= \mathbb{E}\|\phi(X)-\phi(Y)\|^2
\end{aligned} \tag{44}$$

We expand the Gaussian kernel via Taylor expansion, i.e.,

$$\begin{aligned}
k(x,y) &= e^{-(x-y)^2/(2\sigma^2)} \\
&= e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} e^{\frac{xy}{\sigma^2}} \\
&= e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \sum_{n=0}^{\infty} \frac{1}{\sqrt{n!}}(\frac{x}{\sigma})^n \frac{1}{\sqrt{n!}}(\frac{y}{\sigma})^n \\
&= \sum_{n=0}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{n!}}(\frac{x}{\sigma})^n e^{-\frac{y^2}{2\sigma^2}} \frac{1}{\sqrt{n!}}(\frac{y}{\sigma})^n \\
&= \phi(x)^\top\phi(y)
\end{aligned} \tag{45}$$

Therefore, we have

$$\begin{aligned}
\mathrm{MMD}^2_{-c}(\mu,\nu) &= \sum_{n=0}^{\infty} \frac{1}{\sigma^{2n} n!} \left(\mathbb{E}_{x\sim\mu}\left[e^{-x^2/\left(2\sigma^2\right)}x^n\right] - \mathbb{E}_{x\sim\nu}\left[e^{-y^2/\left(2\sigma^2\right)}y^n\right]\right)^2 \\
&= \sum_{n=0}^{\infty} \frac{1}{\sigma^{2n} n!} \left(\tilde{M}_n(\mu) - \tilde{M}_n(\nu)\right)^2
\end{aligned} \tag{46}$$

$\tilde{M}_n(\mu) = \mathbb{E}_{x\sim\mu}\left[e^{-x^2/\left(2\sigma^2\right)}x^n\right]$, and similarly for $\tilde{M}_n(\nu)$. The conclusion is the same as the moment matching in [17]. Finally, due to the equivalence of $\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu)$ after multiplying ε, we have

$$\begin{aligned}
\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu;k) &:= \mathrm{MMD}^2_{-c}(\mu,\nu) + \varepsilon\mathbb{E}\left[\frac{(\Pi^*(X,Y))^2}{\Pi^*(X,X')\Pi^*(Y,Y')}\right] \\
&= \sum_{n=0}^{\infty} \frac{1}{\sigma^{2n} n!} \left(\tilde{M}_n(\mu) - \tilde{M}_n(\nu)\right)^2 + \varepsilon\mathbb{E}\left[\frac{(\Pi^*(X,Y))^2}{\Pi^*(X,X')\Pi^*(Y,Y')}\right],
\end{aligned} \tag{47}$$

This result is also equivalent to Theorem 1, where Π^* would degenerate to $\mu\otimes\nu$ as $\varepsilon\to+\infty$. In that case, the first regularization term would vanish, and thus the Sinkhorn divergence degrades to a MMD loss, i.e., $\mathrm{MMD}^2_{-c}(\mu,\nu)$.

□

D Human-normalized Scores

Our implemnetation is based on [25] and all the experimental settings, including parameters are identical to the distributional RL baselines implemented by [25]. The main results about mean and

	Mean	**Median**	**>Human**	**>DQN**
DQN	173 %	49 %	17	0
C51	309 %	77 %	26	42
QR-DQN-1	430 %	104 %	31	47
MMDQN	**600** %	94 %	27	43
SinkhornDRL	570 %	89 %	27	42

Table 2: Mean and median of best human-normalized scores across 55 Atari 2600 games. The results for all considered algorithms are aaveraged over 3 seeds.

median human-normalized scores of all considered distributional RL algorithms are reported in Table 2. Note that our implementaion is based on Pytorch, and thus the results in Table 2 are not exactly same as results implemented based on Dopamine framework [4]. However, Table 2 also suggests that our SinkhornDRL algorithm can achieve almost state-of-the-art performance in terms of mean human-normalized scores. We argue that although it seems that SinkhronDRL is on par with MMD across all games, our algorithm significant outperforms MMDDRL on a large amount of Atari games, as suggested in Figure 2. The detailed comparison based on learning curves is also exhibited in Appendix E.

E More experimental Results

We provide learning curves of DQN, QRDQN, C51, MMD and SinkhornDRL algorithms on all 55 Atari games in Figures 4 5 6 7 8 9. It illustrates that SinkhornDRL dramatically surpasses the other distributional RL algorithms on a large amount of environments, e.g., Venture, Atlantis, Tennis and SpaceInvader, and presents competitive performance or is only slightly inferior as opposed to the state-of-the-art baselines on other games. Note that the average improvement of SinkhornDRL on Venture game is significant owing to one to two times convergence of SinkhornDRL algorithm over 3 seeds, while the other baselines do not converge over the considered seeds. Although this improvement may also suffer from the instability issue, its occasional success for our SinkhornDRL algorithm also presents huge potential on some complicated environments. We leave the further exploration on the advantage and potential of SinkhornDRL algorithm as the future work.

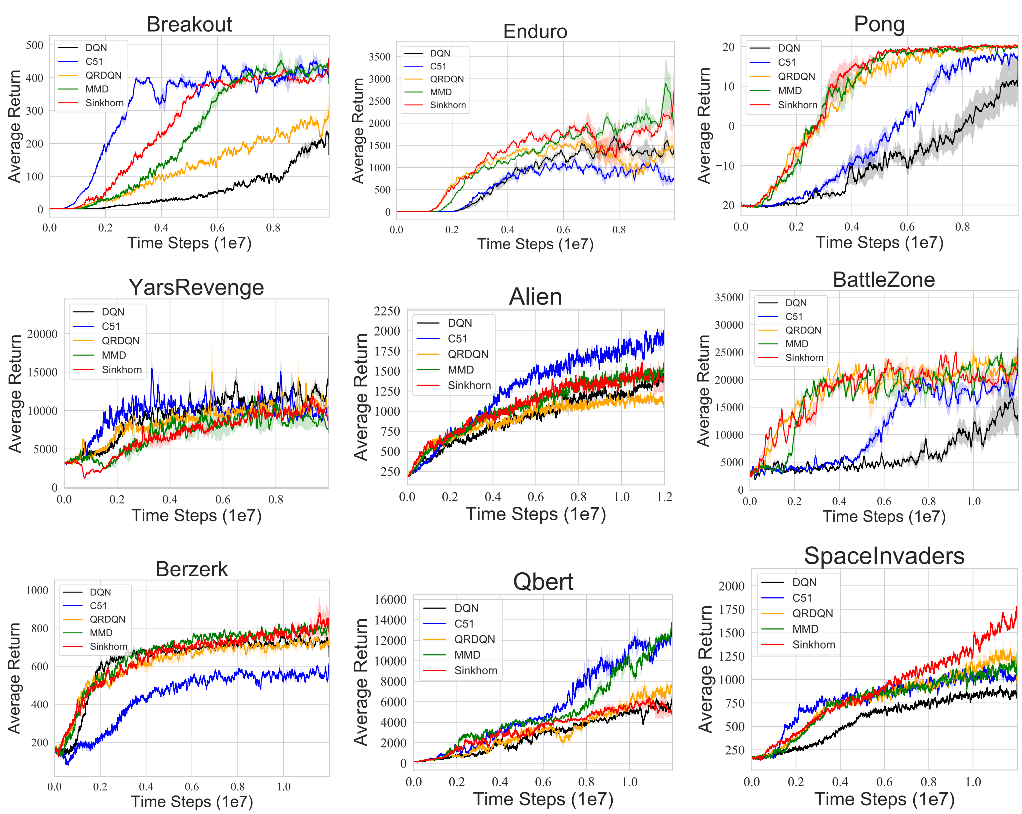


Figure 4: Performance of SinkhornDRL compared with DQN, C51, QRDQN and MMD on Breakout, Enduro, Pong, YarRevenge, Alien, BattleZone, Berzerk, Qbert and SpaceInvader.

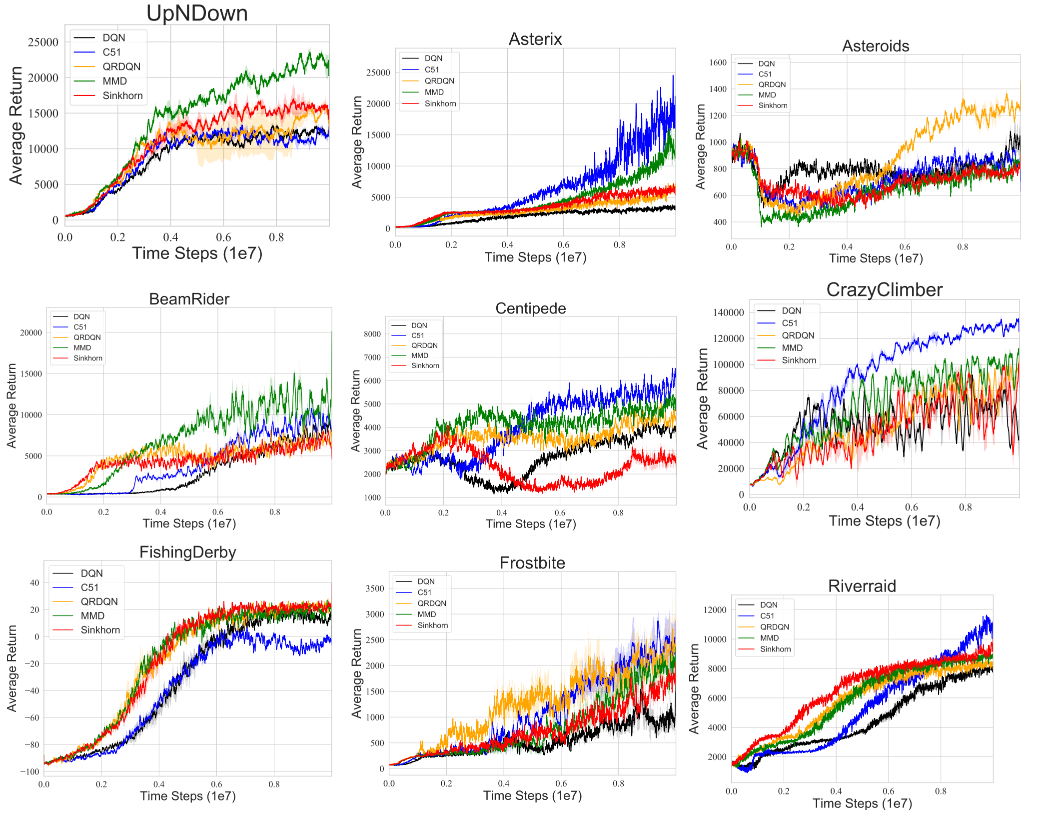


Figure 5: Performance of SinkhornDRL compared with DQN, C51, QRDQN and MMD on UpNDown, Asterix, Asteriods, BeamRider, Centipede, FishingDerby, Frostbite and Riverraid.

F Sensitivity Analysis and Computational Cost

F.1 More results in Sensitivity Analysis

From Figure 10 (a), we can observe that if we gradually decline ε to 0, SinkhornDRL's performance tends to QR-DQN. Note that an overly small ε will lead to a trivial almost 0 $\mathcal{K}_{i,j}$ in Sinkhorn iteration in Algorithm 2, and will cause $\frac{1}{0}$ numerical instability issue for a_l and b_l in Line 5 of Algorithm 2. Due to this reason, the performance of SinkhornDRL with $\varepsilon = 0.1$ or 0.075 declines as the training proceeds, and eventually converges to the average return that QR-DQN achieves. In addition, we also conducted experiments on Seaquest, the similar result is also observed in Figure 11. The performance of SinkhornDRL is robust when $\varepsilon = 10, 100, 500$ and a small $\epsilon = 1$ tends to worsen the performance.

Moreover, for breakout, if we increase ε, the performance of SinkhornDRL tends to that of MMDDRL as suggested in Figure 10 (b). It is also noted that an overly large ε will let the $\mathcal{K}_{i,j}$ explode to ∞. This also leads to numerical instability issue in Sinkhorn iteration in Algorithm 2.

In summary, the trend of SinkhornDRL to close MMDDRL and QR-DQN if we increase or decrease ε, respectively, provides strong empirical evidence to demonstrate the theoretical relationships between Sinkhorn divergence and MMD / Wasserstein distance, although an overly large or small ε will lead to numerical instability issue.

F.2 Comparison with the Computational Cost

We evaluate the computational time every 10,000 iterations across the whole training process of all considered distributional RL algorithms and make a comparison in Figure 12. It suggests that SinkhornDRL indeed increases around 50% computation cost compared with QR-DQN and C51, but only slightly increases the the cost in contrast to MMDDRL on both Breakout and Qbert

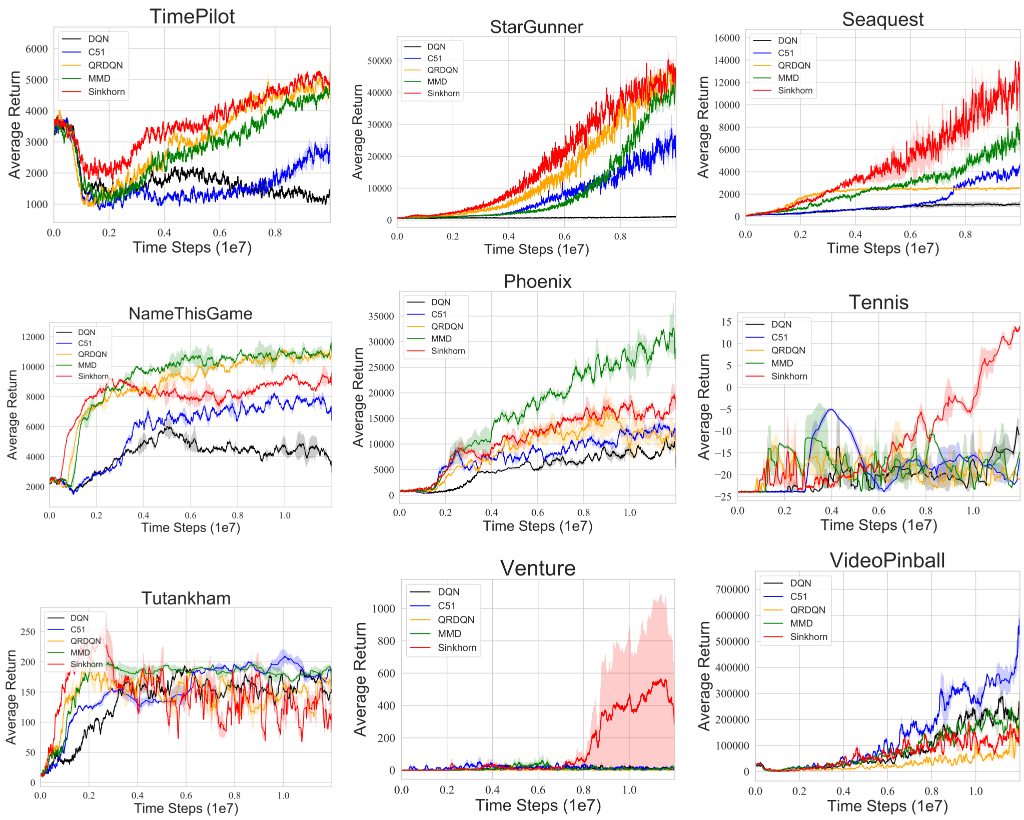


Figure 6: Performance of SinkhornDRL compared with DQN, C51, QRDQN and MMD on TimePilot, StarGuner, Seaquest, NameThisGame, Phoenix, Tennix, Tutankham, Venture and VideoPinball.

games. We argue that this additional computational burden can be tolerant in view of the significant outperformance of SinkhornDRL in a large amount of environments.

In addition, we also find that the number of Sinkhorn iterations L is negligible to the computation cost, while an overly large samples N, e.g., 500, will lead to a large computational burden as illustrated in Figure 13. This can be intuitively explained as the computation complexity of the cost function $c_{i,j}$ is $\mathcal{O}(N^2)$ in SinkhornDRL, which is particularly heavy in computation if N is large enough.

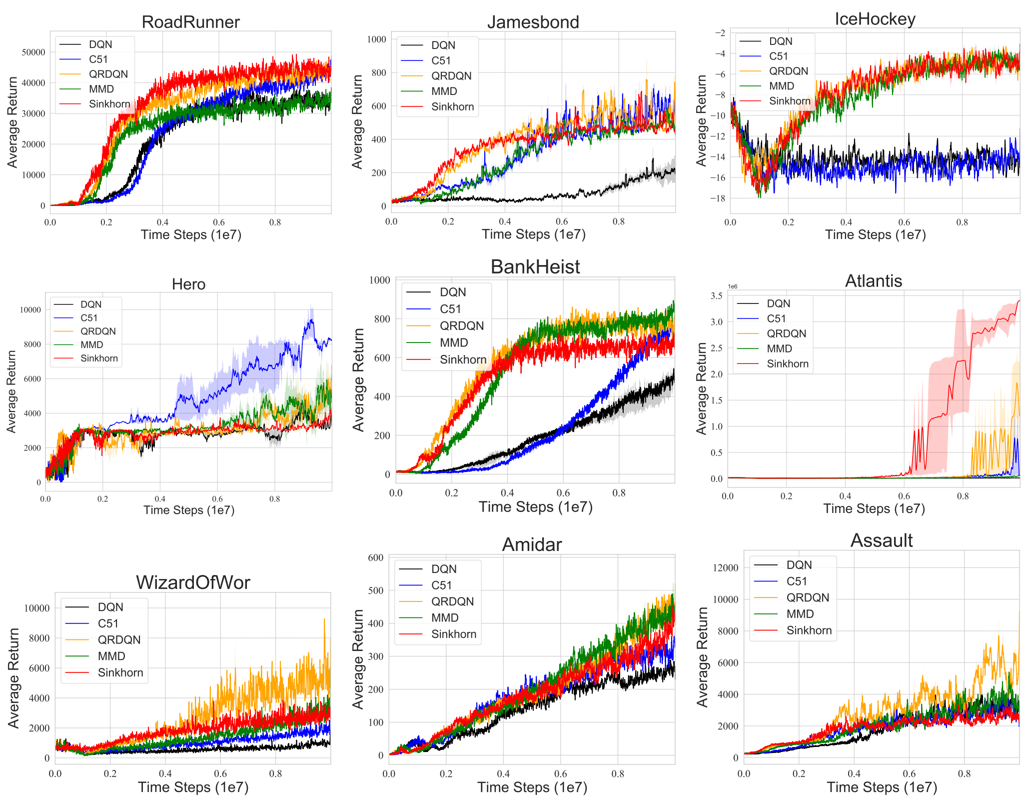


Figure 7: Performance of SinkhornDRL compared with DQN, C51, QRDQN and MMD on RoadRunner, Jamesbond, IceHockey, Hero, BankHeist, Atlantis, WizardOfWor, Amidar and Assault.

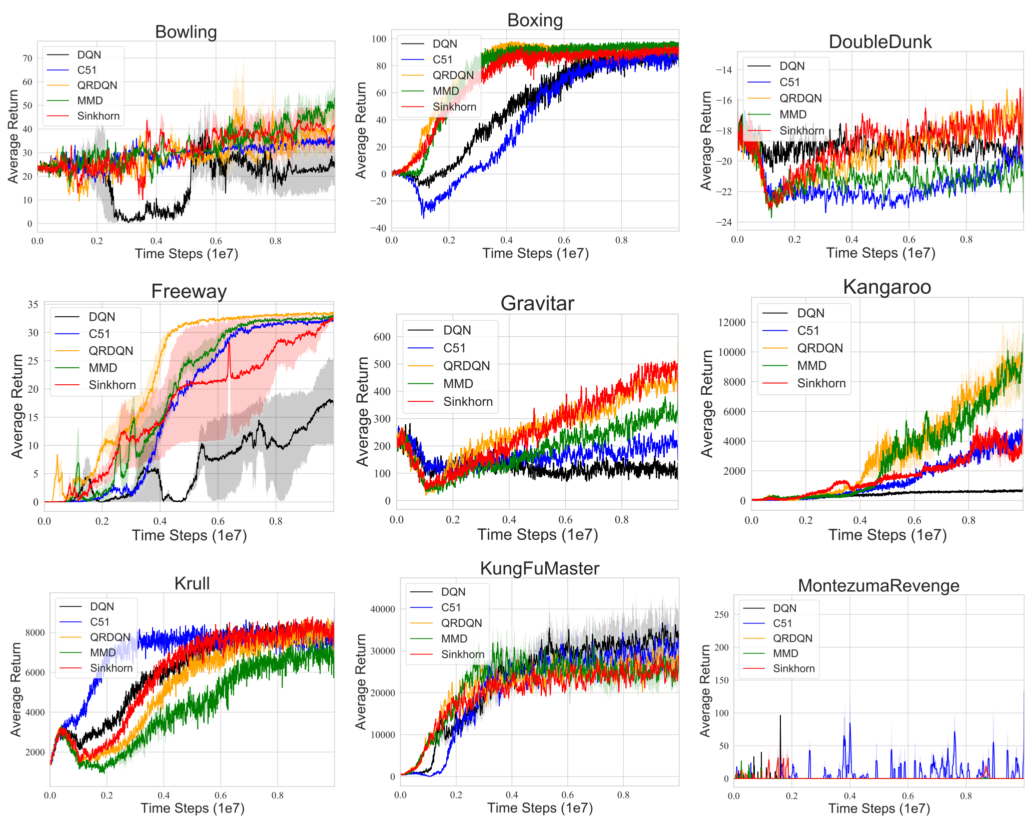


Figure 8: Performance of SinkhornDRL compared with DQN, C51, QRDQN and MMD on Bowling, Boxing, DoubleDunk, Freeway, Gravitar, Kangaroo, Krull, KunFuMaster and MontezumaRevenge.

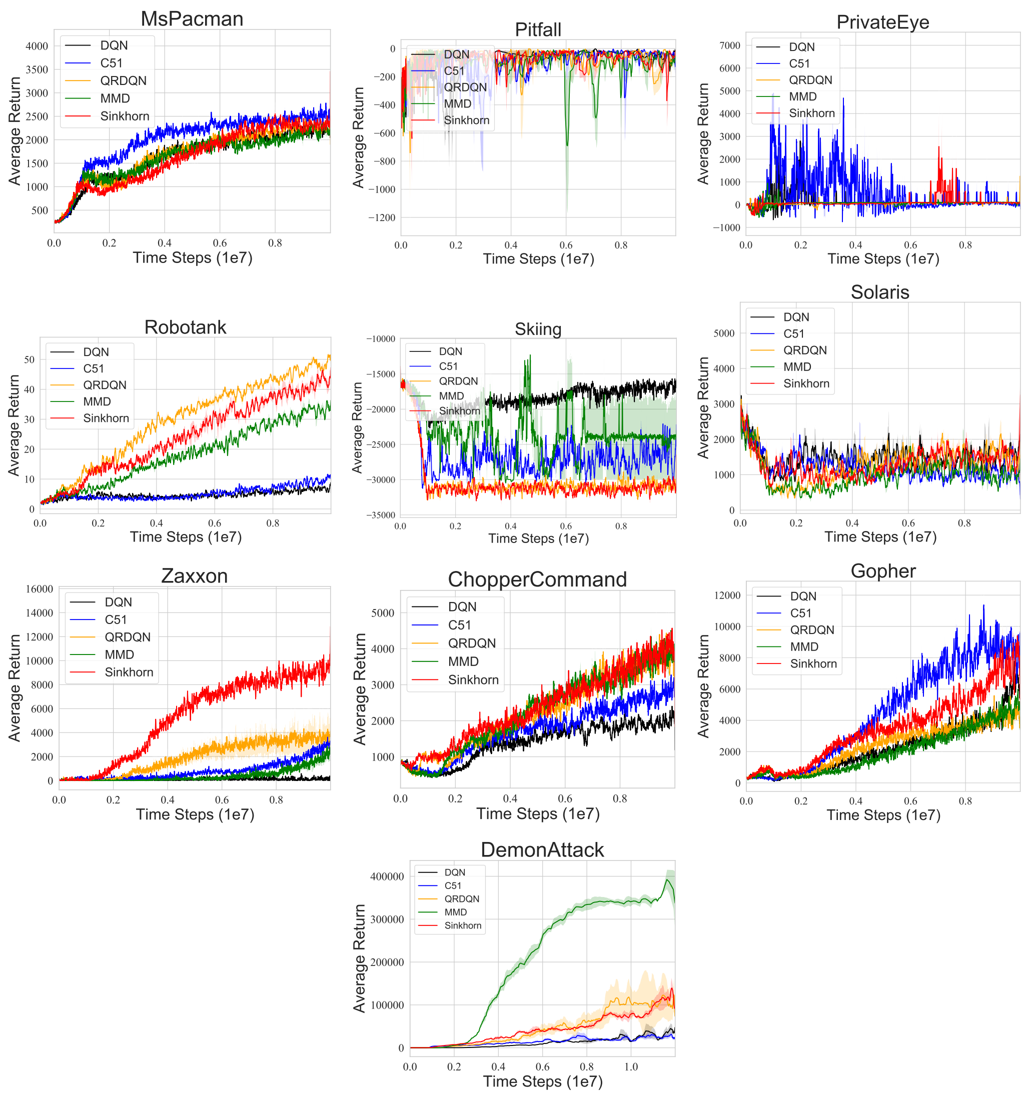
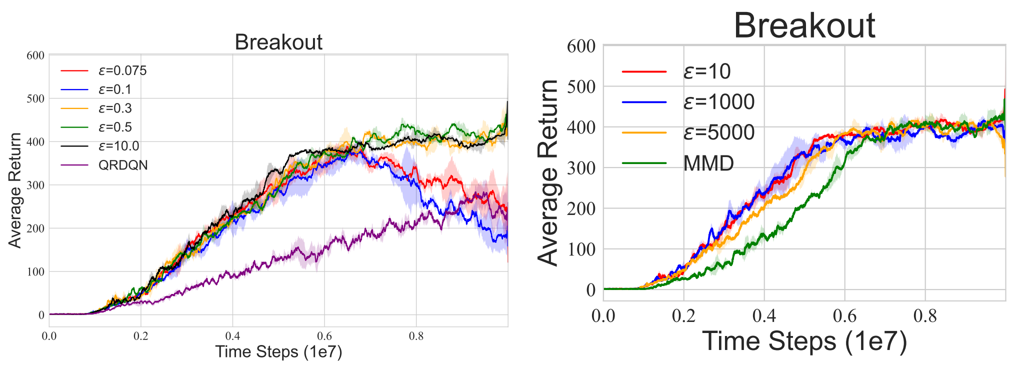


Figure 9: Performance of SinkhornDRL compared with DQN, C51, QRDQN and MMD on MsPacman, Pitfall, PrivateEye, Robotank, Skiing, Solaris, Zaxxon, ChopperCommand, Gopher and DemonAttack.



(a) Small ε in SinkrhornDRL vs QRDQN (b) Large ε in SinkrhornDRL vs MMDDRL

Figure 10: (Left) Sensitivity analysis w.r.t. a small level of ε SinkhornDRL to compare with QR-DQN that approximates Wasserstein distance on Breakout. (Right) Sensitivity analysis w.r.t. a large level of ε SinkhornDRL algorithm to compare with MMDDRL on Breakout. All learning curves are reported over 2 seeds.

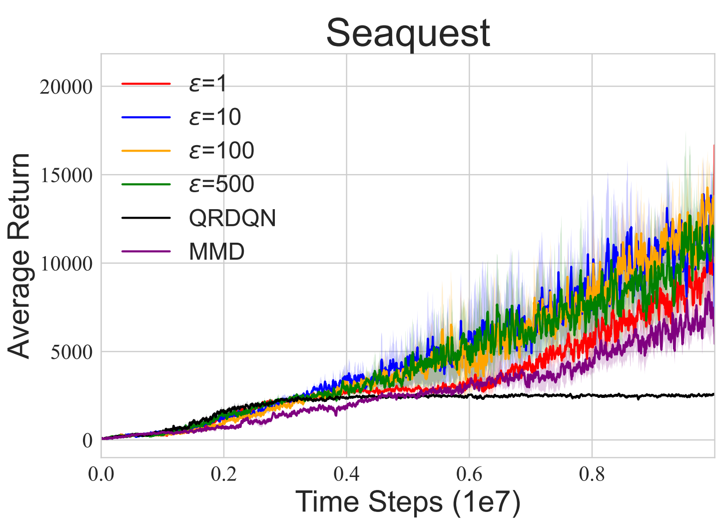


Figure 11: Sensitivity analysis w.r.t. ε SinkhornDRL to compare with QR-DQN and MMD on Seaquest. All learning curves are reported over **3** seeds.

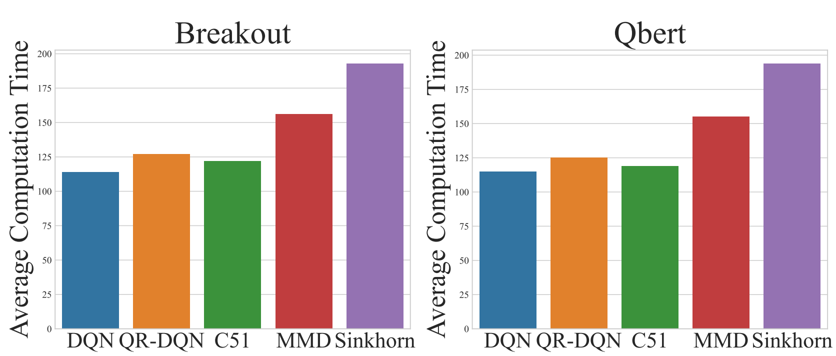


Figure 12: Average computational cost per 10,000 iterations of all considered distributional RL algorithm, where we select $\varepsilon = 10$, $L = 10$ and number of samples $N = 200$ in SinkhornDRL algorithm.

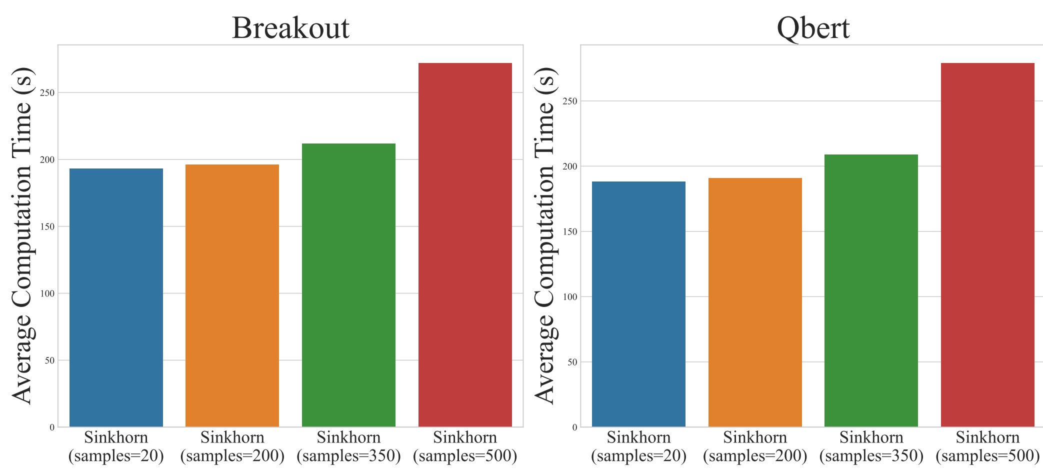


Figure 13: Average computational cost per 10,000 iterations of SinkhornDRL algorithm over different samples.