DISTRIBUTIONAL SLICED-WASSERSTEIN AND APPLICATIONS TO GENERATIVE MODELING

Khai Nguyen
VinAI Research, Vietnam
v.khainb@vinai.io

Nhat Ho*
University of Texas, Austin
VinAI Research, Vietnam
minhnhat@utexas.edu

Tung Pham
VinAI Research, Vietnam
v.tungph4@vinai.io

Hung Bui
VinAI Research, Vietnam
v.hungbh1@vinai.io

ABSTRACT

Sliced-Wasserstein distance (SW) and its variant, Max Sliced-Wasserstein distance (Max-SW), have been used widely in the recent years due to their fast computation and scalability even when the probability measures lie in a very high dimensional space. However, SW requires many unnecessary projection samples to approximate its value while Max-SW only uses the most important projection, which ignores the information of other useful directions. In order to account for these weaknesses, we propose a novel distance, named *Distributional Sliced-Wasserstein* distance (DSW), that finds an *optimal* distribution over projections that can balance between exploring distinctive projecting directions and the informativeness of projections themselves. We show that the DSW is a generalization of Max-SW, and it can be computed efficiently by searching for the optimal push-forward measure over a set of probability measures over the unit sphere satisfying certain regularizing constraints that favor distinct directions. Finally, we conduct extensive experiments with large-scale datasets to demonstrate the favorable performances of the proposed distances over the previous sliced-based distances in generative modeling applications.

1 INTRODUCTION

Optimal transport (OT) is a classical problem in mathematics and operation research. Due to its appealing theoretical properties and flexibility in practical applications, it has recently become an important tool in the machine learning and statistics community; see for example, (Courty et al., 2017; Arjovsky et al., 2017; Tolstikhin et al., 2018; Gulrajani et al., 2017) and references therein. The main usage of OT is to provide a distance named Wasserstein distance, to measure the discrepancy between two probability distributions. However, that distance suffers from expensive computational complexity, which is the main obstacle to using OT in practical applications.

There have been two main approaches to overcome the high computational complexity problem: either approximate the value of OT or apply the OT adaptively to specific situations. The first approach was initiated by (Cuturi, 2013) using an entropic regularizer to speed up the computation of the OT (Sinkhorn, 1967; Knight, 2008). The entropic regularization approach has demonstrated its usefulness in several application domains (Courty et al., 2014; Genevay et al., 2018; Bunne et al., 2019). Along this direction, several works proposed efficient algorithms for solving the entropic OT (Altschuler et al., 2017; Lin et al., 2019b;a) as well as methods to stabilize these algorithms (Chizat et al., 2018; Peyré & Cuturi, 2019; Chizat et al., 2018; Schmitzer, 2019). However, these algorithms have complexities of the order $\mathcal{O}(k^2)$, where k is the number of supports. It is expensive when we need to compute the OT repeatedly, especially in learning the data distribution.

*Nhat Ho finished the work during the time he worked at VinAI Research in the summer of 2020.

The second approach, known as "slicing", takes a rather different perspective. It leverages two key ideas: the OT closed-form expression for two distributions in one-dimensional space, and the transformation of a distribution into a set of projected one-dimensional distributions by the Radon transform (RT) (Helgason, 2010). The popular proposal along this direction is Sliced-Wasserstein (SW) distance (Bonneel et al., 2015), which samples the projecting directions uniformly over a unit sphere in the data ambient space and takes the expectation of the resulting one-dimensional OT distance. The SW distance hence requires a significantly lower computation cost than the original Wasserstein distance and is more scalable than the first approach. Due to its solid statistical guarantees and efficient computation, the SW distance has been successfully applied to a variety of practical tasks (Deshpande et al., 2018; Liutkus et al., 2019; Kolouri et al., 2018; Wu et al., 2019; Deshpande et al., 2019) where it has been shown to have comparative performances to other distances and divergences between probability distributions. However, there is an inevitable bottleneck of computing the SW distance. Specifically, the expectation with respect to the uniform distribution of projections in SW is intractable to compute; therefore, the Monte Carlo method is employed to approximate it. Nevertheless, drawing from a uniform distribution of directions in high-dimension can result in an overwhelming number of irrelevant directions, especially when the actual data lies in a low-dimensional manifold. Hence, SW typically needs to have a large number of samples to yield an accurate estimation of the discrepancy. Alternatively, in the other extreme, Max Sliced-Wasserstein (Max-SW) distance (Deshpande et al., 2019) uses only one important direction to distinguish the probability distributions. However, other potentially relevant directions are ignored in Max-SW. Therefore, Max-SW can miss some important differences between the two distributions in high dimension. We note that the linear projections in the Radon transform can be replaced by non-linear projections resulting in the generalized sliced-Wasserstein distance and its variants (Beylkin, 1984; Kolouri et al., 2019).

Apart from these main directions, there are also few proposals that try either to modify them or to combine the advantages of the above-mentioned approaches. In particular, Paty & Cuturi (2019) extended the idea of the max-sliced distance to the max-subspace distance by considering finding an optimal orthogonal subspace. However, this approach is computationally expensive, since it could not exploit the closed-form of the one-dimensional Wasserstein distance. Another approach named the Projected Wasserstein distance (PWD), which was proposed in (Rowland et al., 2019), uses sliced decomposition to find multiple one-dimension optimal transport maps. Then, it computes the average cost of those maps equally in the original dimension.

Our contributions. Our paper also follows the slicing approach. However, we address key friction in this general line of work: how to obtain a relatively small number of slices simultaneously to maintain the computational efficiency, but at the same time, cover the major differences between two high-dimensional distributions. We take a probabilistic view of slicing by using a probability measure on the unit sphere to represent how important each direction is. From this viewpoint, SW uses the uniform distribution while Max-SW searches for the best delta-Dirac distribution over the projections, both can be considered as special cases. In this paper, we propose to search for an optimal distribution of important directions. We regularize this distribution such that it prefers directions that are far away from one another, hence encouraging an efficient exploration of the space of directions. In the case of no regularization, our proposed method recovers max-(generalized) SW as a special case. In summary, our main contributions are two-fold:

1. First, we introduce a novel distance, named *Distributional Sliced-Wasserstein distance* (DSW), to account for the issues of previous sliced distances. Our main idea is to search for not just a single most important projection, but an *optimal* distribution over projections that could balance between an expansion of the area around important projections and the informativeness of projections themselves, i.e., how well they can distinguish the two target probability measures. We show that DSW is a proper metric in the probability space and possesses appealing statistical and computational properties as the previous sliced distances.

2. Second, we apply the DSW distance to generative modeling tasks based on the generative adversarial framework. The extensive experiments on real and large-scale datasets show that DSW distance significantly outperforms the SW and Max-SW distances under similar computational time on these tasks. Furthermore, the DSW distance helps model distribution converge to the data distribution faster and provides more realistic generated images than the SW and Max-SW distances.

Organization. The remainder of the paper is organized as follows. In Section 2, we provide backgrounds for Wasserstein distance and its slice-based versions. In Section 3, we propose distributional (generalized) sliced-Wasserstein distance and analyze some of its theoretical properties. Section 4 includes extensive experiment results followed by discussions in Section 5. Finally, we defer the proofs of key results and extra materials in the Appendices.

Notation. For any $\theta, \theta' \in \mathbb{R}^d$, $\cos(\theta, \theta') = \frac{\theta^\top \theta'}{\|\theta\|\|\theta'\|}$, where $\|.\|$ is ℓ_2 norm. For any $d \geq 2$, $\mathbb{S}^{d-1}$ denotes the unit sphere in d dimension in ℓ_2 norm . Furthermore, δ denotes the Dirac delta function, and $\langle \cdot, \cdot \rangle$ is the Euclidean inner-product. For any $p \geq 1$, $\mathbb{L}^p(\mathbb{R}^d)$ is the set of real-valued functions on $\mathbb{R}^d$ with finite p-th moment.

2 Background

In this section, we provide necessary backgrounds for the (generalized) Radon transform, the Wasserstein, and sliced-Wasserstein distances.

2.1 Wasserstein distance

We start with a formal definition of Wasserstein distance. For any $p \geq 1$, we define $\mathcal{P}_p(\mathbb{R}^d)$ as the set of Borel probability measures with finite p-th moment defined on a given metric space $(\mathbb{R}^d, \|.\|)$. For any probability measures μ, ν defined on $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, we denote their corresponding probability density functions as I_μ and I_ν. The Wasserstein distance of order p between μ and ν is given by (Villani, 2008; Peyré & Cuturi, 2019):

$$W_p(\mu, \nu) := \Big(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|^p d\pi(x, y) \Big)^{\frac{1}{p}},$$

where $\Pi(\mu, \nu)$ is a set of all transportation plans π such that the marginal distributions of π are μ and ν, respectively. In order to simplify the presentation, we abuse the notation by using both $W_p(\mu, \nu)$ and $W_p(I_\mu, I_\nu)$ interchangeably for the Wasserstein distance between μ and ν. When μ and ν are *one-dimension* measures, the Wasserstein distance between μ and ν has a closed-form expression $W_p(\mu, \nu) = (\int_0^1 |F_\mu^{-1}(z) - F_\nu^{-1}(z)|^p dz)^{1/p}$ where F_μ and F_ν are the cumulative distribution function (CDF) of I_μ and I_ν, respectively.

2.2 (Generalized) Radon transforms

Now, we review (generalized) Radon transform maps, which are key to the notion of (generalized) sliced-Wasserstein distance and its variants. The *Radon transform* (RT) maps a function $I \in \mathbb{L}^1(\mathbb{R}^d)$ to the space of functions defined over space of lines in $\mathbb{R}^d$. In particular, for any $t \in \mathbb{R}$ and direction $\theta \in \mathbb{S}^{d-1}$, the RT is defined as follows (Helgason, 2010) : $\mathcal{R}I(t, \theta) := \int_{\mathbb{R}^d} I(x)\delta(t - \langle x, \theta \rangle)dx$.

The *generalized* Radon transform (GRT) (Beylkin, 1984) extends the original one from integration over hyperplanes of $\mathbb{R}^d$ to integration over hypersurfaces. In particular, it is defined as: $\mathcal{G}I(t, \theta) := \int_{\mathbb{R}^d} I(x)\delta(t - g(x, \theta))dx$, where $t \in \mathbb{R}$ and $\theta \in \Omega_\theta$. Here, Ω_θ is a compact subset of $\mathbb{R}^d$ and $g : \mathbb{R}^d \times \mathbb{S}^{d-1} \mapsto \mathbb{R}$ is a defining function (cf. Assumptions H1-H4 in (Kolouri et al., 2019) for the definition of defining function) inducing the hypersurfaces. When $g(x, \theta) = \langle x, \theta \rangle$ and $\Omega_\theta = \mathbb{S}^{d-1}$, the generalized Radon transform becomes the standard Radon transform.

2.3 (Generalized) sliced-Wasserstein distances

The sliced-Wasserstein distance (SW) between two probability measures μ and ν is defined as (Bonneel et al., 2015): $SW_p(\mu, \nu) := (\int_{\mathbb{S}^{d-1}} W_p^p\big(\mathcal{R}I_\mu(\cdot, \theta), \mathcal{R}I_\nu(\cdot, \theta)\big)d\theta)^{1/p}$. Similarly, the generalized sliced-Wasserstein distance (Kolouri et al., 2019) (GSW) is given by $\text{GSW}_p(\mu, \nu) := (\int_{\Omega_\theta} W_p^p\big(\mathcal{G}I_\mu(\cdot, \theta), \mathcal{G}I_\nu(\cdot, \theta)\big)d\theta)^{1/p}$, where Ω_θ is the compact set of feasible parameter. However, these integrals are usually intractable. Thus, they are often approximated by using Monte Carlo scheme to draw uniform samples $\{\theta_i\}_{i=1}^N$ from $\mathbb{S}^{d-1}$ and Ω_θ. In particular, $SW_p^p(\mu, \nu) \approx \frac{1}{N}\sum_{i=1}^N W_p^p\big(\mathcal{R}I_\mu(\cdot, \theta_i), \mathcal{R}I_\nu(\cdot, \theta_i)\big)$ and $\text{GSW}_p^p(\mu, \nu) \approx \frac{1}{N}\sum_{i=1}^N W_p^p\big(\mathcal{G}I_\mu(\cdot, \theta_i), \mathcal{G}I_\nu(\cdot, \theta_i)\big)$. In order to obtain a good approximation of (generalized) SW distances, N needs to be sufficiently large.

However, important directions are not distributed uniformly over the sphere. Thus, this approach will draw potentially many unimportant projections that are not only expensive but also greatly reduce the effect of the SW distance.

2.4 MAX (GENERALIZED) SLICED-WASSERSTEIN DISTANCES

An approach to using only informative directions is to simply take the best slice in discriminating two given probability distributions. That distance is max sliced-Wasserstein distance (Max-SW) (Deshpande et al., 2019), which is given by $\max SW_p(\mu,\nu) := \max_{\theta\in\mathbb{S}^{d-1}} W_p(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta))$. By combining this idea with non-linear projections from generalized Radon transform, we obtain max generalized sliced-Wasserstein distance (Max-GSW) (Kolouri et al., 2019). The formal definition of that distance is: $\max GSW_p(\mu,\nu) := \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_\mu(\cdot,\theta),\mathcal{G}I_\nu(\cdot,\theta))$. The (generalized) Max-SW distances focus on finding only the most important direction. Meanwhile, other informative directions play no role in the distance. Therefore, (generalized) Max-SW distances can ignore useful information about the structure of high dimensional probability measures.

3 DISTRIBUTIONAL SLICED-WASSERSTEIN DISTANCE

With the aim of improving the limitations of the previous sliced distances, we propose a novel distance, named *Distributional Sliced-Wasserstein distance* (DSW), that can search for not just a single but a distribution of important directions on the unit sphere. We prove that it is a well-defined metric and discuss its connection to the existing sliced-based distances in Section 3.1. Then, we provide a procedure to approximate DSW based on its dual form in Section 3.2.

3.1 DEFINITION AND METRICITY

We first start with a definition of distributional sliced-Wasserstein distance. We say $C > 0$ *admissible* if the set $\mathbb{M}_C$ of probability measures σ on $\mathbb{S}^{d-1}$ satisfying $\mathbb{E}_{\theta,\theta'\sim\sigma}\left[|\theta^\top\theta'|\right] \leq C$ is not empty.

Definition 1. *Given two probability measures μ and ν on $\mathbb{R}^d$ with finite p-th moments where $p \geq 1$ and an admissible regularizing constant $C > 0$. The* distributional sliced-Wasserstein distance *(DSW) of order p between μ and ν is given by:*

$$DSW_p(\mu,\nu;C) := \sup_{\sigma\in\mathbb{M}_C}\left(\mathbb{E}_{\theta\sim\sigma}\left[W_p^p(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta))\right]\right)^{\frac{1}{p}}, \tag{1}$$

where $\mathcal{R}$ is the Radon transform operator.

The DSW aims to find the optimal probability measure of slices on the unit sphere $\mathbb{S}^{d-1}$. Note that, the Max-SW distance is equivalent to searching for the best Dirac measure on a single point in $\mathbb{S}^{d-1}$, which puts all weights in only one direction. Meanwhile, the uniform measure in the formulation of SW distance distributes the same weights in all directions. Indeed, the uniform and Dirac measures are two special cases, because they view that either all directions are equally important or only one direction is important. That view is too restricted if the data actually lie on low dimensional space. Thus, we aim to find a probability measure which concentrates only on areas around important directions. Furthermore, we do not want these directions to lie in only one small area, because under the orthogonal projection of RT, their corresponding one-dimensional distributions will become similar. In order to achieve this, we search for an optimal measure σ that satisfies the regularization constraint $\mathbb{E}_{\theta,\theta'\sim\sigma}[|\theta^\top\theta'|] \leq C$. By Cauchy-Schwarz inequality, C is no greater than 1, thus $\mathbb{M}_1$ contains all probability measures on the unit sphere. Optimizing over $\mathbb{M}_1$ simply returns the best Dirac measure corresponding to the Max-SW distance. When C is small, the constraint forces the measure σ to distribute more weights to directions that are far from each other (in terms of their angles). Thus, a small appropriate value of C will help to balance between the distinctiveness and informativeness of these targeted directions. For further discussion about C, see Appendix B.1.

Next, we show that DSW is a well-defined metric on the probability space.

Theorem 1. *For any $p \geq 1$ and admissible $C > 0$, $DSW_p(\cdot,\cdot;C)$ is a well-defined metric in the space of Borel probability measures with finite p-th moment. In particular, it is non-negative, symmetric, identity, and satisfies the triangle inequality.*

The proof of Theorem 1 is in Appendix A.1. Our next result establishes the topological equivalence between DSW distance and (max)-sliced Wasserstein and Wasserstein distances.

Theorem 2. *For any $p \geq 1$ and admissible $C > 0$, the following holds*

(a) $DSW_p(\mu, \nu; C) \leq maxSW_p(\mu, \nu) \leq W_p(\mu, \nu)$.

(b) If $C \geq 1/d$, we have $DSW_p(\mu, \nu; C) \geq \left(\frac{1}{d}\right)^{1/p} maxSW_p(\mu, \nu) \geq \left(\frac{1}{d}\right)^{1/p} SW_p(\mu, \nu)$.

As a consequence, when $p \geq 1$ and $C \geq 1/d$, $DSW_p(\cdot, \cdot; C)$, SW_p, $maxSW_p$, and W_p are topologically equivalent, namely, the convergence of probability measures under $DSW_p(\cdot, \cdot; C)$ implies the convergence of these measures under other metrics and vice versa.

The proof of Theorem 2 is in Appendix A.2. As a consequence of Theorem 2, the statistical error of estimating the unknown distribution based on the empirical distribution of n i.i.d data under DSW distance is $C_d \cdot n^{-1/2}$ with high probability where C_d is some universal constant depending on dimension d (see Appendix B.3). Therefore, as other sliced-based Wasserstein distances, the DSW distance does not suffer from the curse of dimensionality.

3.2 Computation of DSW distance

Direct computation of DSW distance is challenging. Hence we consider a dual form of DSW distance and a reparametrization of σ as follows.

Definition 2. *For any $p \geq 1$ and admissible $C > 0$, there exists a non-negative constant λ_C depending on C such that the dual form of DSW distance takes the following form*

$$DSW_p^*(\mu, \nu; C) = \sup_{\sigma \in \mathbb{M}} \left\{ \left(\mathbb{E}_{\theta \sim \sigma} \left[W_p^p(\mathcal{R}I_\mu(\cdot, \theta), \mathcal{R}I_\nu(\cdot, \theta)) \right] \right)^{\frac{1}{p}} - \lambda_C \mathbb{E}_{\theta, \theta' \sim \sigma} \left[|\theta^\top \theta'| \right] \right\} + \lambda_C C,$$

where $\mathbb{M}$ denotes the space of all probability measures on the unit sphere $\mathbb{S}^{d-1}$.

By the Lagrangian duality theory, $\text{DSW}_p(\mu, \nu; C) \geq \text{DSW}_p^*(\mu, \nu; C)$ for any $p \geq 1$ and admissible $C > 0$. In Definition 2, the set $\mathbb{M}_C$ disappears and λ_C plays the tuning role for the regularized term $\mathbb{E}_{\theta, \theta' \sim \sigma}\left[|\theta^\top \theta'|\right]$. When λ_C is large, $\mathbb{E}_{\theta, \theta' \sim \sigma}\left[|\theta^\top \theta'|\right]$ needs to be small, meaning that C is small. When λ_C is small, the value of $\mathbb{E}_{\theta, \theta' \sim \sigma}\left[|\theta^\top \theta'|\right]$ becomes less important, i.e., C is large.

For reparametrizing the measure σ, we use a pushforward of uniform measure on the unit sphere through some measurable function f. In particular, let f be a Borel measurable function from $\mathbb{S}^{d-1}$ to $\mathbb{S}^{d-1}$. For any Borel set $A \subset \mathbb{S}^{d-1}$, we define $\sigma(A) = \sigma^{d-1}(f^{-1}(A))$, where σ^{d-1} is the uniform probability measure on $\mathbb{S}^{d-1}$. Then for any Borel measurable function $g : \mathbb{S}^{d-1} \to \mathbb{R}$, we have $\int_{\theta \sim \sigma} g(\theta) d\sigma(\theta) = \int_{\theta \sim \sigma^{d-1}} (g \circ f)(\theta) d\sigma^{d-1}(\theta)$. Therefore, we obtain the equivalent dual form of DSW as follows:

$$\text{DSW}_p^*(\mu, \nu; C) = \sup_{f \in \mathcal{F}} \left\{ \left(\mathbb{E}_{\theta \sim \sigma^{d-1}} \left[W_p^p\big(\mathcal{R}I_\mu(\cdot, f(\theta)), \mathcal{R}I_\nu(\cdot, f(\theta))\big) \right] \right)^{1/p} - \lambda_C \mathbb{E}_{\theta, \theta' \sim \sigma^{d-1}} \left[\left| f(\theta)^\top f(\theta') \right| \right] \right\} + \lambda_C C := \sup_{f \in \mathcal{F}} \text{DS}(f), \tag{2}$$

where $\mathcal{F}$ is a class of all Borel measurable functions from $\mathbb{S}^{d-1}$ to $\mathbb{S}^{d-1}$.

Finding the optimal f: We parameterize f in the dual form (2) by using a deep neural network with parameter ϕ, defined as f_ϕ. Then, we estimate the gradient of the objective function $\text{DS}(f_\phi)$ in equation (2) with respect to ϕ and use stochastic gradient ascent algorithm to update ϕ. Since there are expectations over uniform distribution in the gradient of $\text{DS}(f_\phi)$, we use the Monte Carlo method to approximate these expectations. Note that, we can use the fixed point from the stochastic ascent algorithm to approximate the dual value of DSW in equation (2). A detailed argument for this point is in Appendix B.2. Finally, in generative model applications with DSW being the loss function, we only need to use the gradient of the function $\text{DS}(.)$ to update the parameters of interest. Therefore, we can treat λ_C as a regularized parameter and tune it to find suitable value in these applications.

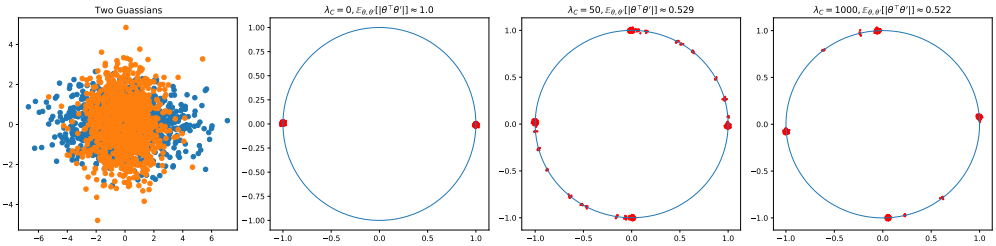


Figure 1: Empirical behavior of optimal measure σ, approximated by 1000 samples, on a circle for different values of λ_C (the constant in the dual form of DSW in Definition 2) when μ and ν are bivariate Gaussian distributions sharing the same eigenvectors. When $\lambda_C = 0$, $C = 1$. When λ_C increases, C becomes small.

Illustration of the roles of λ_C and C: To illustrate the roles of λ_C and C in finding optimal distribution σ, we conduct a simple experiment on two Gaussian distributions with zero means and covariance matrices given by $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$. The experiment optimizes the empirical form of Definition 2 with different choices of λ_C. The results are shown in Figure 1 with the reported value of λ_C and $\mathbb{E}_{\theta,\theta' \sim \sigma}\left[|\theta^\top \theta'|\right]$. For $\lambda_C = 0$, the obtained distribution concentrates only on one direction. When $\lambda_C = 50$, optimal σ distributes more weights to other directions on the circle. When $\lambda_C = 1000$, optimal σ is close to the discrete distribution concentrated on two eigenvectors of the covariance matrices, which are the main directions differentiating the two Gaussian distributions.

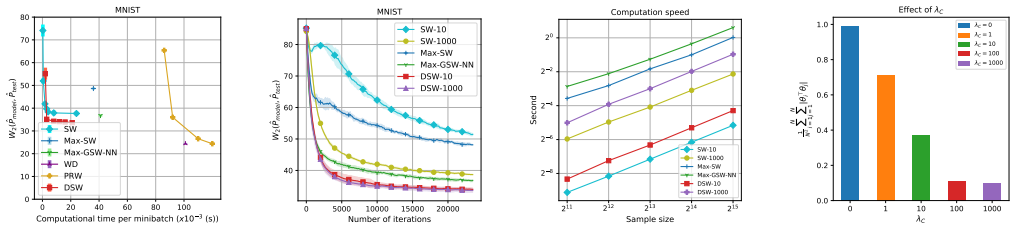
Extension of DSW and comparison of DSW to Max-GSW-NN: Similar to SW, we extend DSW to distributional generalized sliced Wasserstein (DGSW) by using the non-linear projecting operator via GRT. The definition of the DGSW and its properties are in Appendix C. Finally, in Appendix E.1, we show the distinction of the DSW to Max-GSW-NN (Kolouri et al., 2019) when the neural network defining function in Max-GSW-NN is $g(x, \theta) = \langle x, f(\theta) \rangle$ where $f : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$.

4 Experiments

In this section, we conduct extensive experiments comparing the performance in both generative quality and computational speed of the proposed DSW distance with other sliced-based distances, namely the SW, Max-SW, Max-GSW-NN (Kolouri et al., 2019) and projected robust subspace Wasserstein (PRW) (Paty & Cuturi, 2019; Lin et al., 2020) using the minimum expected distance estimator (MEDE) (Bernton et al., 2019) on MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky, 2009), CelebA (Liu et al., 2015) and LSUN (Yu et al., 2015) datasets. The details of the MEDE framework are described in Appendix D. We would like to note that the wall-clock timing of different methods may be subject to the differences in the hyperparameter settings and software implementations of different methods. On MNIST dataset, we train generative models with different distances and then evaluate their performances by comparing Wasserstein-2 distances between 10000 random generated images and all images from the MNIST test set. Due to the very large size of other datasets, e.g., 3 million images in LSUN, it is expensive to compute empirical Wasserstein-2 distance as its complexity is of order $\mathcal{O}(k^2 \log k)$ where k is the number of support points. Therefore, after we train generative models, we use FID score (Heusel et al., 2017) to evaluate the generative quality of these generators. The FID score is calculated from 10000 random generated images and all training samples using precomputed statistics in (Heusel et al., 2017). Finally, for λ_C in DSW (see Definition 2), it is chosen in the set $\{1, 10, 100, 1000\}$ such that its Wasserstein-2 (FID score) (between 10000 random generated images and all images from corresponding validation set) is the lowest among the four values. Detailed experiment settings are in Appendix G. Finally, we also apply the DSW into color transfer task (Rabin et al., 2010; 2014; Bonneel et al., 2015; Perrot et al., 2016) in Appendix F, where we find that DSW also performs better than SW and Max-SW in this task.

4.1 Results on MNIST

Generative quality and computational speed: We report the performance of the learned generative models for MNIST in Figure 2(a). To plot this figure, we vary the number of projections $N \in$



(a) (b) (c) (d)

Figure 2: (a) Comparison between DSW, SW, Max-SW, Max-GSW-NN, PRW and WD based on execution time and performance. Here, each dot of SW and DSW corresponds to the number of projections chosen in $\{1, 10, 10^2, 5 \times 10^2, 10^3, 5 \times 10^3, 10^4\}$. Each dot of PRW corresponds to the dimension of the subspace chosen in $\{2, 5, 10, 50\}$; (b) Comparison between SW, DSW, Max-SW and Max-GSW-NN based on Wasserstein-2 distance between distributions of learned model and test set over iterations; (c) Computation speed of distances based on the number of minibatch's samples (log-log scale); (d) Effect of λ_C on the mean of absolute values of pairwise cosine similarity between 10 random directions from the found distribution σ of DSW.

$\{1, 10, 10^2, 5 \times 10^2, 10^3, 5 \times 10^3, 10^4\}$ for the SW, and $N \in \{1, 10, 10^2, 5 \times 10^2, 10^3, 5 \times 10^3\}$ for the DSW. Then we measure the computational time per minibatch and the Wasserstein-2 score of the learned generators for each N. We plot the Wasserstein-2 score and computational time of Max-SW and Max-GSW-NN in their standard settings (Kolouri et al., 2019). Except for the regime with very fast but low-quality learned models, DSW is better than all the existing slice-based baselines in terms of both model quality and computational speed. Moreover, DSW can learn good models with very few projections, e.g., DSW-10 achieves better model quality than Max-GSW-NN and Max-SW and is one order-of-magnitude faster than these sliced distances. Finally, with a similar computational time, a learned generator by DSW has the Wasserstein-2 score that is roughly 10% lower than the one got from SW. For the qualitative comparison between these distances, we show random generated images from their generative models in Figure 7 in Appendix E.1. We observe that generated images from DSW are sharper and easier to classify into numbers than those from other baseline distances.

Comparison with projected robust subspace Wasserstein (PRW) and Wasserstein distance: In Figure 2(a), we plot the Wasserstein-2 score and computational time of Wasserstein distance (WD) and PRW, where the subspace dimension of PRW varies in the range $\{2, 5, 10, 50\}$. PRW is able to improve upon the model quality of slice-based methods including DSW, however at the cost of being an order of magnitude slower than DSW with 10 projections (DSW-10). We observe that DSW-10 obtains a better Wasserstein-2 score than PRW with 5-dimensional subspace, while its corresponding computational time is 30 times faster than that of PRW-5. Using 50 dimension, PRW's Wasserstein-2 score improves about 29% to that of DSW-10 but the computational cost is also around 40 times slower. The model trained by WD gives good Wasserstein-2 score; however, it is computational expensive (about 40 times slower than DSW-10). The main computational advantage of DSW comes from the exact calculation of Wasserstein distance in one-dimension. The visual comparison between PRW, WD and DSW based on their generated images is in Figure 12 in Appendix E.2.

Convergence behavior: Figure 2(b) shows that DSW learns better models at a faster speed of convergence than other baseline distances with a very small number of projections, e.g., DSW-10 is the second lowest curve compared to curves from other sliced-based distances.

Scalability over sample size of minibatch: Results in Figure 2(c) show that DSW has a computational complexity of the order $\mathcal{O}(k \log k)$, which is similar to those of other sliced-based distances, where k is the number of samples per batch.

Effect of the regularization parameter λ_C: For each value of $\lambda_C \in \{1, 10, 100, 1000\}$, we find the optimal distribution σ of DSW with $N = 10$ projections, and then calculate $A_N = \frac{1}{N^2}\sum_{i,j=1}^{N} |\theta_i^\top \theta_j|$, an approximation of the regularized term $\mathbb{E}_{\theta,\theta' \sim \sigma}\left[|\theta^\top \theta'|\right]$ in the dual form of DSW in equation (2), where $\{\theta_i\}_{i=1}^N \sim \sigma$. The results are shown in Figure 2(d). We observe that when λ_C increases, A_N goes down. When $\lambda_C = 0$, i.e., no regularization, A_N gets close to 1, meaning that all projected directions collapse to one direction. When $\lambda_C = 1000$, A_N is close to 0.1, suggesting that all projected directions are nearly orthogonal.

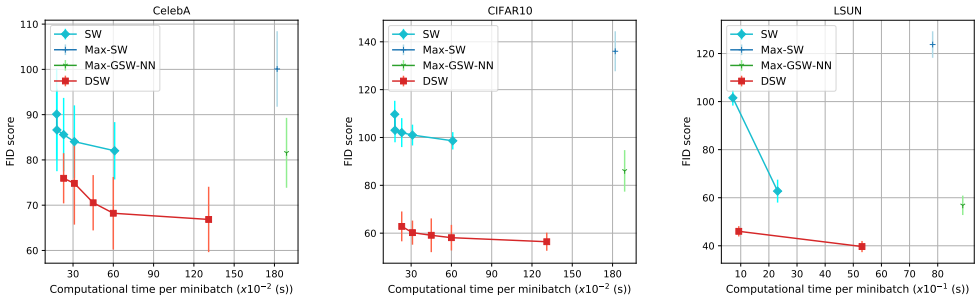


Figure 3: Comparison between DSW, SW, Max-SW and Max-GSW-NN in terms of execution time and performance. Here, each dot of SW and DSW corresponds to the number of projections chosen in $\{10^2, 5 \times 10^2, 10^3, 5 \times 10^3, 10^4\}$. We set the minibatch size be 512 on CelebA and CIFAR, and be 4096 on LSUN.

Additional experiments: We also investigate how the number of gradient-steps used for updating distribution of directions σ, and how the size of minibatches affects the quality of DSW (see Appendix E.1). The results show that an increasing number of gradient steps to update σ leads to better performance of DSW but also slows down the computation speed. Furthermore, we carry out experiments with DGSW, an extension of DSW to non-linear projections, and test the new proposed distances in training encoder-generator models on MNIST using joint contrastive inference (JCI) in Appendices E.1 and E.3. The description of these models is in Appendix D.

4.2 Results on Large-scale Datasets

Next, we conduct large-scale experiments on a range of more realistic image datasets. We train generative models using CIFAR10, CelebA, and LSUN datasets (all these datasets are rescaled to 64x64 resolution). When working with high dimensional distributions, Deshpande et al. (2018) proposed a trick to improve the quality of the generator by learning a feature function which maps data to a new feature space that is more manageable in size. When the feature function is fixed, the generator is trained to match the distribution of features. When the generator is fixed, the feature function tries to tease apart the data empirical features from the generated feature distribution. For the experiments in this section, we use the same technique with DSW and all other baseline distances.

We compare DSW with SW, Max-SW, and Max-GSW-NN in both generative quality (FID score) and computational time in Figure 3. We could not compare DSW with PRW on the large-scale datasets since PRW is computationally expensive to train to obtain good generated images. On CelebA and CIFAR10, we let N, the number of projections of both DSW and SW, vary in the set $\{10^2, 5 \times 10^2, 10^3, 5 \times 10^3, 10^4\}$. For LSUN, since it takes considerably longer time to train each model, we only vary N in the set $\{10^2, 10^4\}$. On all these large datasets, DSW outperforms all the other baselines in both FID score of the learned model and computational efficiency. The gap of FID scores between DSW and other methods is especially large on CIFAR10 and LSUN. For example, on CIFAR10, with the same computational time, FID scores of DSW are always lower than those of SW about 20 units. On LSUN, with 100 projections, DSW can achieve an FID score of 46 while SW with 10000 projections still has a worse FID score of over 60. It is interesting to note that on these high-dimensional datasets, Max-SW performs rather poorly: it obtains the highest FID scores among all distances while requires heavy computation. Max-GSW-NN has better FID scores than (Max)-SW; however, it is still worse than DSW and while being slower. This is consistent with the intuition that as the number of dimension of the data grows, the use of a single important slice in Max-SW becomes a less efficient approximation. DSW, on the other hand, is able to make use of more important slices, and at the same time avoids SW's inefficiency of uniform slice-sampling.

Generated images from CelebA, CIFAR10 and LSUN are deferred to Appendix E.1. Comparing to other sliced-based Wasserstein distances, generated samples obtained from the DSW's generative model are also more visually realistic. Further experiments to compare DGSW with GSW, Max-GSW, and Max-GSW-NN are also given in the Appendix E.1. Based on these experiments, we can conclude that the distributional approach also improves the generative quality of non-linear slicing distances.

5 CONCLUSION

In this paper, we have presented the novel distributional sliced-Wasserstein (DSW) distances between two probability measures. Our main idea is to search for the best distribution of important directions while regularizing towards orthogonal directions. We prove that they are well-defined metrics and provide their theoretical and computational properties. We compare our proposed distances to other sliced-based distances in a variety of generative modeling tasks, including estimating generative models and jointly estimating both generators and inference models. Extensive experiments demonstrate that our new distances yield significantly better models and convergence behaviors during training than the previous sliced-based distances. One important future direction is to investigate theoretically the optimal choice of the regularization parameter λ_C such that the DSW distance can capture all the important directions that can distinguish two target probability measures well.

REFERENCES

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in neural information processing systems*, pp. 1964–1974, 2017.

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel AJ van Gerven, and Eric Maris. Wasserstein variational inference. In *Advances in Neural Information Processing Systems*, pp. 2473–2482, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. *arXiv preprint arXiv:1912.08247*, 2019.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

Gregory Beylkin. The inversion problem and applications of the generalized Radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984.

Nicolas Bonneel and David Coeurjolly. Spot: sliced partial optimal transport. *ACM Transactions on Graphics (TOG)*, 38(4):1–13, 2019.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015.

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *International Conference on Machine Learning*, 2019.

Xiongjie Chen, Yongxin Yang, and Yunpeng Li. Augmented sliced Wasserstein distances. *arXiv preprint arXiv:2006.08812*, 2020.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–289. Springer, 2014.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 3730–3739, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017. URL https://pythonot.github.io/.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein Gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Sigurdur Helgason. *Integral geometry and Radon transforms*. Springer Science & Business Media, 2010.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Philip A Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in pytorch, 2020.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 261–272, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Bing Li, Shaoli Wang, et al. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102:997–1008, 2007.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–328, 1991.

Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pp. 3982–3991, 2019a.

Tianyi Lin, Nhat Ho, and Michael I Jordan. On the acceleration of the Sinkhorn and Greenkhorn algorithms for optimal transport. *arXiv preprint arXiv:1906.01437*, 2019b.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael I. Jordan. Projection robust Wasserstein distance and Riemannian optimization. *arXiv preprint arXiv:2006.07458*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pp. 4104–4113, 2019.

Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. In *Advances in Neural Information Processing Systems*, pp. 8118–8129, 2019.

Boris Muzellec and Marco Cuturi. Subspace detours: Building transport plans that are optimal on subspace projections. In *Advances in Neural Information Processing Systems*, pp. 6917–6928, 2019.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, pp. 250–260, 2019.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International Conference on Machine Learning*, pp. 5072–5081, 2019.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29:4197–4205, 2016.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Julien Rabin, Julie Delon, and Yann Gousseau. Regularization of transportation maps for color and contrast transfer. In *2010 IEEE International Conference on Image Processing*, pp. 1933–1936. IEEE, 2010.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4852–4856. IEEE, 2014.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of Wasserstein distances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 186–195, 2019.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

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

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced Wasserstein generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722, 2019.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.