Generalized Sliced Wasserstein Distances

Soheil Kolouri, Kimia Nadjahi, Umut Şimşekli, Roland Badeau, Gustavo K. Rohde

Generalized Sliced Wasserstein Distances

Soheil Kolouri [1] **Kimia Nadjahi** [2] **Umut Şimşekli** [2] **Roland Badeau** [2] **Gustavo K. Rohde** [3]

Abstract

The Wasserstein distance and its variations, e.g., the sliced-Wasserstein (SW) distance, have recently drawn tremendous amount of attention from the machine learning community. The SW distance, specifically, was shown to have similar flavor to that of the Wasserstein distance, while being much simpler to compute, and is used in various applications including generative modeling and supervised learning. In this paper, we first clarify the mathematical connections between the SW distance and the Radon transform. We then utilize the generalized Radon transform to define a new family of distances for probability measures, which we call generalized sliced-Wasserstein (GSW) distances. We provide the conditions under which a GSW is a distance. We then show that, similarly to the SW distance, the GSW distance can be extended to a max-GSW distance. Finally, we compare the numerical performance of the proposed distances between probability measures on several generative modeling tasks, including sliced-Wasserstein flows and sliced-Wasserstein auto-encoders.

1. Introduction

Emerging from the heart of Optimal Transportation (OT) theory, the Wasserstein distance (Villani, 2008) forms a metric between two probability measures and has attracted abundant attention in data sciences and machine learning due to its nice theoretical properties and applications on ubiquitous domains (Solomon et al., 2014; Frogner et al., 2015; Montavon et al., 2016; Kolouri et al., 2017; Courty et al., 2017; Peyré & Cuturi, 2018; Schmitz et al., 2018), especially in implicit generative modeling such as OT-based generative adversarial networks (GAN) and variational auto-encoders (Arjovsky et al., 2017; Bousquet et al., 2017; Gulrajani et al., 2017; Tolstikhin et al., 2018).

While OT brings new perspectives and principled ways to formalize problems, the OT-based methods usually suffer from high computational complexity. The Wasserstein distance is often the computational bottleneck as it turns out that computing it between multi-dimensional measures is numerically intractable in general. This important computational burden is a major limiting factor in the application of OT distances to large-scale data analysis. Recently, several numerical methods have been proposed to speed-up the evaluation of the Wasserstein distance. For instance, entropic regularization techniques (Cuturi, 2013; Cuturi & Peyré, 2015; Solomon et al., 2015) provide a fast approximation to the Wasserstein distance by regularizing the original OT problem with an entropy term. Other notable contributions towards computational methods for OT include multi-scale and sparse approximation approaches (Oberman & Ruan, 2015; Schmitzer, 2016), and Newton-based schemes for semi-discrete OT (Lévy, 2015; Kitagawa et al., 2016).

There are some special favorable cases where solving the OT problem is easy and reasonably cheap. In particular, the Wasserstein distance for one-dimensional probability densities has a closed-form formula and can be efficiently approximated. This nice property motivates the use of the sliced-Wasserstein distance (Bonneel et al., 2015), an alternative OT distance which is obtained by computing infinitely many *linear projections* of the high-dimensional distribution to one-dimensional distributions and then computing the average of the Wasserstein distance between these one-dimensional representations. While having similar theoretical properties (Bonnotte, 2013), the sliced-Wasserstein distance has significantly lower computational requirements than the classical Wasserstein distance. Therefore, it has recently attracted ample attention and successfully been applied to a variety of practical tasks (Bonneel et al., 2015; Kolouri et al., 2016; Carriere et al., 2017; Karras et al., 2017; Şimşekli et al., 2018; Deshpande et al., 2018; Kolouri et al., 2018; 2019).

As we will detail in the next sections, the linear projection process used in the sliced-Wasserstein distance is closely related to the Radon transform, which is widely used in tomography (Radon, 1917; Helgason, 2011). In other words, the sliced-Wasserstein distance is calculated via linear slicing of the probability distributions. The linear nature of

[1]HRL Laboratories, LLC., Malibu, CA, USA [2]Télécom ParisTech, Paris, France [3]University of Virginia Charlottesville, VA, USA. Correspondence to: Soheil Kolouri <skolouri@hrl.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

these projections creates an important computational bottleneck. This is due to the fact that, in very high dimensional settings, the data often lives in a thin manifold and the required number of randomly chosen linear projections grows very quickly in order to be able to capture the structure of the data distribution (Şimşekli et al., 2018). Therefore, if the number of required projections could be reduced, it would result in a significant performance improvement in sliced-Wasserstein computations.

Contributions. In this paper, we address the aforementioned computational issues of the sliced-Wasserstein distance and for the first time, we extend the linear slicing to *non-linear* slicing of probability measures. Our main contributions are summarized as follows:

- We dive deep in the mathematics of the *generalized* Radon transform (Beylkin, 1984) and extend the definition of the sliced-Wasserstein distance to an entire class of distances, which we call the generalized sliced-Wasserstein (GSW) distance. We prove that replacing the linear projections with *polynomial* projections will still yield a valid distance metric and we then identify general conditions under which GSW is a proper metric.
- We then show that, instead of using infinitely many projections as required by GSW, we can still define a valid distance metric by using a *single* projection, as long as the projection gives the maximal distance in the projected space. We aptly call this distance the max-GSW distance. The max-GSW distance vastly reduces the computational cost induced by the projection operations; however, it comes with an additional cost since it requires optimization over the space of projectors.
- Due to their inherent non-linearity, the GSW and max-GSW distances are expected to capture the complex structure of high dimensional distributions by using much less projections, which will reduce the overall computational burden in a significant amount. We verify this fact in our experiments, where we illustrate the superior performance of the proposed distances in both synthetic and real-data settings.

2. Background on Optimal Transport distances

We review in this section the preliminary concepts and formulations needed to develop our framework, namely the p-Wasserstein distance, the Radon transform and the sliced p-Wasserstein distance. In what follows, we denote by $P_p(\Omega)$ the set of Borel probability measures with finite p'th moment defined on a given metric space (Ω, d) and by $\mu \in P_p(X)$ and $\nu \in P_p(Y)$ probability measures defined on $X, Y \subseteq \Omega$ with corresponding probability density functions I_μ and I_ν, *i.e.* $d\mu(x) = I_\mu(x)dx$ and $d\nu(y) = I_\nu(y)dy$.

2.1. Wasserstein Distance

The p-Wasserstein distance for $p \in [1, \infty)$ between μ and ν is defined as the optimal mass transportation (OMT) problem (Villani, 2008) with cost function $c(x, y) = d^p(x, y)$, such that:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times Y} d^p(x, y) d\gamma(x, y) \right)^{\frac{1}{p}}, \quad (1)$$

where $\Gamma(\mu, \nu)$ is the set of all transportation plans $\gamma \in \Gamma(\mu, \nu)$ such that:

$$\begin{array}{ll} \gamma(A \times Y) = \mu(A) & \text{for any Borel subset } A \subseteq X \\ \gamma(X \times B) = \nu(B) & \text{for any Borel subset } B \subseteq Y \end{array}.$$

Due to Brenier's theorem (Brenier, 1991), for absolutely continuous probability measures μ and ν (with respect to the Lebesgue measure), the p-Wasserstein distance can be equivalently obtained from

$$W_p(\mu, \nu) = \left(\inf_{f \in MP(\mu,\nu)} \int_X d^p(x, f(x)) d\mu(x) \right)^{\frac{1}{p}} \quad (2)$$

where $MP(\mu, \nu) = \{f : X \to Y \mid f_{\#}\mu = \nu\}$ and $f_{\#}\mu$ represents the pushforward of measure μ, characterized as

$$\int_{f^{-1}(A)} d\mu(x) = \int_A d\nu(y) \text{ for any Borel subset } A \subseteq Y.$$

Note that in most engineering and computer science applications, Ω is a compact subset of $\mathbb{R}^d$ and $d(x, y) = |x - y|$ is the Euclidean distance. By abuse of notation, we will use $W_p(\mu, \nu)$ and $W_p(I_\mu, I_\nu)$ interchangeably.

One-dimensional distributions: The case of one-dimensional continuous probability measures is specifically interesting as the p-Wasserstein distance has a closed-form solution. More precisely, for one-dimensional probability measures, there exists a unique monotonically increasing transport map that pushes one measure to another. Let $F_\mu(x) = \mu((-\infty, x]) = \int_{-\infty}^{x} I_0(\tau)d\tau$ be the cumulative distribution function (CDF) for I_μ and define F_ν to be the CDF of I_ν. The transport map is then uniquely defined as $f(x) = F_\nu^{-1}(F_\mu(x))$ and, consequently, the p-Wasserstein distance has an analytical form given as follows:

$$\begin{aligned} W_p(\mu, \nu) &= \left(\int_X d^p(x, F_\nu^{-1}(F_\mu(x))) d\mu(x) \right)^{\frac{1}{p}} \\ &= \left(\int_0^1 d^p(F_\mu^{-1}(z), F_\nu^{-1}(z)) dz \right)^{\frac{1}{p}} \quad (3) \end{aligned}$$

where Eq. (3) results from the change of variable $F_\mu(x) = z$. It should be noted that for empirical distributions, Eq. (3) is calculated by simply sorting the samples from the two

distributions and calculating the average $d^p(\cdot,\cdot)$ between the sorted samples. This requires only $O(M)$ operations at best and $O(M \log M)$ at worst, where M is the number of samples drawn from each distribution. See Kolouri et al. (2019) for more details. The closed-form solution of the p-Wasserstein distance for one-dimensional distributions is an attractive property that gives rise to the sliced-Wasserstein (SW) distance. Next, we review the Radon transform, which enables the definition of the SW distance.

2.2. Radon Transform

The standard Radon transform, denoted by $\mathcal{R}$, maps a function $I \in L^1(\mathbb{R}^d)$, where

$$L^1(\mathbb{R}^d) \triangleq \{I : \mathbb{R}^d \to \mathbb{R} \mid \int_{\mathbb{R}^d} |I(x)|dx < \infty\},$$

to the infinite set of its integrals over the hyperplanes of $\mathbb{R}^d$ and is defined as

$$\mathcal{R}I(t,\theta) := \int_{\mathbb{R}^d} I(x)\delta(t - \langle x, \theta\rangle)dx, \tag{4}$$

for $(t,\theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ stands for the d-dimensional unit sphere, $\delta(\cdot)$ the one-dimensional Dirac delta function, and $\langle\cdot,\cdot\rangle$ the Euclidean inner-product. Note that $\mathcal{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d-1})$. Each hyperplane can be written as:

$$H(t,\theta) = \{x \in \mathbb{R}^d \mid \langle x, \theta\rangle = t\}, \tag{5}$$

which alternatively can be interpreted as a level set of the function $g \in L^1(\mathbb{R}^d \times \mathbb{S}^{d-1})$ defined as $g(x,\theta) = \langle x, \theta\rangle$. For a fixed θ, the integrals over all hyperplanes orthogonal to θ define a continuous function $\mathcal{R}I(\cdot,\theta) : \mathbb{R} \to \mathbb{R}$ which is a projection (or a slice) of I.

The Radon transform is a linear bijection (Natterer, 1986; Helgason, 2011) and its inverse $\mathcal{R}^{-1}$ is defined as:

$$\begin{aligned} I(x) &= \mathcal{R}^{-1}(\mathcal{R}I(t,\theta)) \\ &= \int_{\mathbb{S}^{d-1}} (\mathcal{R}I(\langle x,\theta\rangle,\theta) * \eta(\langle x,\theta\rangle)d\theta \end{aligned} \tag{6}$$

where $\eta(\cdot)$ is a one-dimensional high-pass filter with corresponding Fourier transform $\mathcal{F}\eta(\omega) = c|\omega|^{d-1}$, which appears due to the Fourier slice theorem (Helgason, 2011), and '$*$' is the convolution operator. The above definition of the inverse Radon transform is also known as the filtered back-projection method, which is extensively used in image reconstruction in the biomedical imaging community. Intuitively each one-dimensional projection (or slice) $\mathcal{R}I(\cdot,\theta)$ is first filtered via a high-pass filter and then smeared back into $\mathbb{R}^d$ along $H(\cdot,\theta)$ to approximate I. The summation of all smeared approximations then reconstructs I. Note that in practice, acquiring an infinite number of projections is not feasible, therefore the integration in the filtered back-projection formulation is replaced with a finite summation over projections (*i.e.*, a Monte-Carlo approximation).

[Gustavo] As with the other paper, I think this section does not clarify anything mathematical, and could probably be moved towards the end, close to a computational section.

Radon transform of empirical PDFs: The Radon transform of I_μ simply follows Equation (4), where $\mathcal{R}I_\mu(\cdot,\theta)$ is a one-dimensional marginal distribution of I_μ. However, in most machine learning applications we do not have access to the distribution I_μ but to its samples, x_n. Kernel density estimation could be used in such scenarios to approximate I_μ from its samples,

$$I_\mu(x) \approx \frac{1}{N}\sum_{n=1}^{N} \phi(x - x_n)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^+$ is a density kernel where $\int_{\mathbb{R}^d} \phi(x)dx = 1$ (e.g. Gaussian kernel). The Radon transform of I_μ can then be approximated from its samples:

$$\mathcal{R}I_\mu(t,\theta) \approx \frac{1}{N}\sum_{n=1}^{N} \mathcal{R}\phi(t - x_n \cdot \theta, \theta)$$

Note that certain density kernels have analytic Radon transformation. For instance when $\phi(x) = \delta(x)$ the Radon transform $\mathcal{R}\phi(t,\theta) = \delta(t)$. Similarly for Gaussian kernels and when $\phi(x) = \mathcal{N}_d(0_d, \sigma^2 I_{d\times d})$ the Radon transform is equal to $\mathcal{R}\phi(t,\theta) = \mathcal{N}_1(0,\sigma^2)$. Moreover, given the high-dimensional nature of the problem estimating density I in $\mathbb{R}^d$ requires large number of samples, however, the projections of I, $\mathcal{R}I(.,\theta)$, are one dimensional and therefore it may not be critical to have large number of samples to estimate these one-dimensional densities.

2.3. SW and Max-SW Distances

The idea behind the sliced p-Wasserstein distance is to first obtain a family of one-dimensional representations for a higher-dimensional probability distribution through linear projections (via Radon transform), and then calculate the distance between two input distributions as a functional on the p-Wasserstein distance of their one-dimensional representations (*i.e.*, the one-dimensional marginal distributions). In that sense, the distance is obtained by solving several one-dimensional optimal transport problems, which have closed-form solutions. More precisely, the sliced p-Wasserstein distance between I_μ and I_ν is defined as

$$SW_p(I_\mu, I_\nu) = \left(\int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}I_\mu(.,\theta), \mathcal{R}I_\nu(.,\theta))d\theta\right)^{\frac{1}{p}} \tag{7}$$

The sliced p-Wasserstein distance as defined above is positive, symmetric, and it satisfies coincidence axiom and

the triangle inequality and hence it is a true metric (Bonnotte, 2013; Kolouri et al., 2016). Calculation of the sliced-Wasserstein distance requires an integration over the unit sphere in $\mathbb{R}^d$, i.e., $\mathbb{S}^{d-1}$. In practice, this integration is approximated by using a simple Monte Carlo scheme that draws uniform samples from $\mathbb{S}^{d-1}$ and replaces the integral with a finite-sample average,

$$SW_p(I_\mu, I_\nu) \approx \left(\frac{1}{|\Theta|}\sum_{\theta_l \in \Theta} W_p^p(\mathcal{R}I_\mu(\cdot;\theta_l), \mathcal{R}I_\nu(\cdot;\theta_l))\right)^{\frac{1}{p}} \tag{8}$$

The sliced p-Wasserstein distance is especially useful when one only has access to samples of a high-dimensional PDFs and kernel density estimation is required to estimate I_μ. One dimensional kernel density estimation of PDF slices is a much simpler task compared to direct estimation of I from its samples. The catch however, is that as the dimensionality grows one requires larger number of projections to estimate I from $\mathcal{R}I(.,\theta)$. In short, if a reasonably smooth two-dimensional distribution can be approximated by its L projections, then one would require $\mathcal{O}(L^{d-1})$ number of projection to approximate a similarly smooth d-dimensional distribution (for $d \geq 2$).

To further clarify this, Let $I_\mu = \mathcal{N}(0, I_d)$ and $I_\nu = \mathcal{N}(x_0, I_d)$ be two Gaussian densities with identity covariances in the d-dimensional space, where $x_0 \in \mathbb{R}^d$. The slices of these distributions then become one-dimensional Gaussians of the form, $\mathcal{R}I_\mu(\cdot,\theta) = \mathcal{N}(0,1)$ and $\mathcal{R}I_\mu(\cdot,\theta) = \mathcal{N}(\theta \cdot x_0, 1)$. It is therefore clear to see that $W_2(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta))$ achieves its maximum value when $\theta = \frac{x_0}{\|x_0\|_2}$, and it is zero for θs that are orthogonal to x_0. On the other hand, we know that randomly drawn vectors from the unit sphere are more likely to be nearly orthogonal in high-dimensions. More rigorously, the following inequality holds, $Pr(|\theta \cdot \frac{x_0}{\|x_0\|_2}| < \epsilon) > 1 - e^{(-d\epsilon^2)}$, which implies that for higher dimensions, d, the majority of sampled θs would be nearly orthogonal to x_0 and therefore $W_2(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)) \approx 0$ with high probability.

One remedy for the 'projection complexity' of the SW distance is to avoid uniform sampling of the unit sphere, and pick samples, θs, that contain discriminant information between I_μ and I_ν. This idea was for instance used in (Deshpande et al., 2018), where the authors first calculate a linear discriminant subspace and then measure the empirical SW distance by setting θs to be the discriminant components of the subspace. A similarly flavored but less heuristic approach is to use the max-SW distance, which is an alternative distance between I_μ and I_ν, and is defined as:

$$\text{max-}SW_p(I_\mu, I_\nu) = \sup_{\theta \in \mathbb{S}^{d-1}} W_p(\mathcal{R}I_\mu(.,\theta), \mathcal{R}I_\nu(.,\theta)) \tag{9}$$

Given that W_p is a true distance, it is trivial to show that (9) is also a distance: it is symmetric, it satisfies the Triangle Inequality, and the Coincidence Axiom holds. Later on, we will prove these properties for the generalized sliced-Wasserstein distance, which contains the SW distance as a special case.

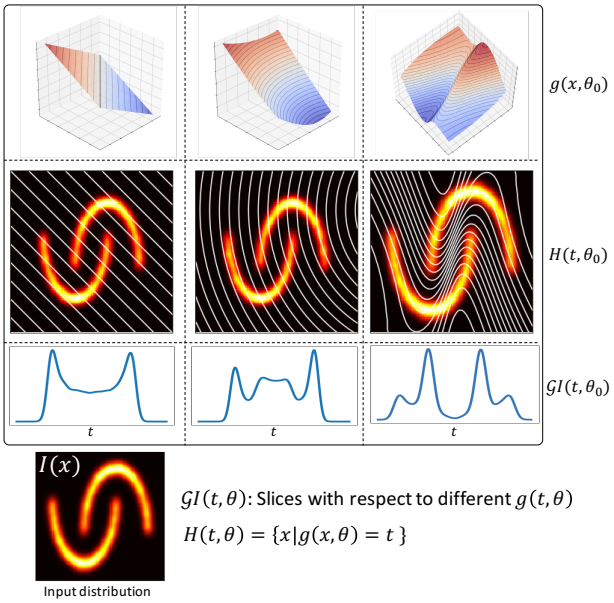

Figure 1. Visualizing the slicing process for classic and generalized Radon transform for the Half Moon distribution. The slices $\mathcal{G}I(t,\theta)$ follow Equation (10).

2.4. Generalized Radon transform

The Generalized Radon transform (GRT) extends the original idea of the classic Radon transform introduced by Radon (1917) from integration over hyperplanes of $\mathbb{R}^d$ to integration over hypersurfaces, i.e. $(d-1)$-dimensional manifolds (Beylkin, 1984; Denisyuk, 1994; Ehrenpreis, 2003; Gel'fand et al., 1969; Kuchment, 2006; Homan & Zhou, 2017). GRT has various applications, including Thermoacoustic Tomography, where the hypersurfaces are spheres, and Electrical Impedance Tomography (EIT), where integration over hyperbolic surfaces appear.

We introduce a function $g(x,\theta)$ defined on $\mathcal{X} \times (\mathbb{R}^n \backslash \{0\})$ where $\mathcal{X}$ is a domain in $\mathbb{R}^d$. We say that g is a *defining function* when it satisfies the following conditions:

H1. *$g(x,\theta)$ is a real-valued C^∞ function on $\mathcal{X} \times (\mathbb{R}^n \backslash \{0\})$*

H2. *$g(x,\theta)$ is homogeneous of degree one in θ, i.e.*

$$\forall \lambda \in \mathbb{R},\ g(x, \lambda\theta) = \lambda g(x,\theta)$$

H3. *g is non-degenerate in the sense that $d_x g(x,\theta) \neq 0$ in $\mathcal{X} \times \mathbb{R}^n \backslash \{0\}$*

H4. *The mixed Hessian of g is strictly positive, i.e.*

$$det\left(\frac{\partial^2 g}{\partial x^i \partial \theta^j}\right) > 0$$

Then, the generalized Radon transform of $I \in L^1(\mathbb{R}^d)$ is the integration of I over hypersurfaces characterized by the level sets of g, which are denoted by

$$H_{t,\theta} = \{x \in \mathcal{X} \mid g(x,\theta) = t\}.$$

In other words, the GRT of I is defined as:

$$\mathcal{G}I(t,\theta) = \int_{\mathbb{R}^d} I(x)\delta(t - g(x,\theta))dx \tag{10}$$

where g is a defining function.

Note that the standard Radon transform is a special case of the GRT with $g(x,\theta) = x \cdot \theta$. Figure 1 visualizes the slicing process for classic and generalized Radon transform for the Half Moon distribution as input.

3. GSW and Max-GSW Distances

Following the definition of the SW distance in Equation (7), we define the generalized sliced p-Wasserstein distance (GSW) using the generalized Radon transform as

$$GSW_p(I_\mu, I_\nu) = \left(\int_{\Omega_\theta} W_p^p(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta))d\theta\right)^{\frac{1}{p}} \tag{11}$$

and subsequently,

$$\text{max-}GSW_p(I_\mu, I_\nu) = \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)) \tag{12}$$

Proposition 1. *The generalized sliced p-Wasserstein distance and the maximum generalized sliced p-Wasserstein distance are, indeed, distances over $\mathcal{P}_p(\Omega)$ if and only if the generalized Radon transform is injective.*

Proof. The non-negativity and symmetry are direct consequences from the fact that the Wasserstein distance is a metric (Villani, 2008): see supplementary material.

We prove the triangle inequality for GSW_p and max-GSW_p. Let μ_1, μ_2 and μ_3 in $\mathcal{P}_p(\Omega)$. Since the Wasserstein distance satisfies the triangle inequality, we have, for all $\theta \in \Omega_\theta$,

$$\begin{aligned} W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta), \mathcal{G}\mathcal{I}_{\mu_3}(\cdot,\theta)) \leq\ & W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta), \mathcal{G}\mathcal{I}_{\mu_2}(\cdot,\theta)) \\ & + W_p(\mathcal{G}\mathcal{I}_{\mu_2}(\cdot,\theta), \mathcal{G}\mathcal{I}_{\mu_3}(\cdot,\theta)) \end{aligned}$$

Therefore, we can write:

$$\begin{aligned} & GSW_p(I_{\mu_1}, I_{\mu_3}) \\ &= \left(\int_{\Omega_\theta} W_p^p(\mathcal{G}I_{\mu_1}(\cdot,\theta), \mathcal{G}I_{\mu_3}(\cdot,\theta))d\theta\right)^{1/p} \\ &\leq \left(\int_{\Omega_\theta} \big(W_p(\mathcal{G}I_{\mu_1}(\cdot,\theta), \mathcal{G}I_{\mu_2}(\cdot,\theta)) \right. \\ &\quad \left. + W_p(\mathcal{G}I_{\mu_2}(\cdot,\theta), \mathcal{G}I_{\mu_3}(\cdot,\theta))\big)^p d\theta\right)^{1/p} \\ &\leq \left(\int_{\Omega_\theta} W_p^p(\mathcal{G}I_{\mu_1}(\cdot,\theta), \mathcal{G}I_{\mu_2}(\cdot,\theta))d\theta\right)^{1/p} \\ &\quad + \left(\int_{\Omega_\theta} W_p^p(\mathcal{G}I_{\mu_2}(\cdot,\theta), \mathcal{G}I_{\mu_3}(\cdot,\theta))d\theta\right)^{1/p} \qquad (13) \\ &\leq GSW_p(I_{\mu_1}, I_{\mu_2}) + GSW_p(I_{\mu_2}, I_{\mu_3}) \end{aligned}$$

where inequality (13) follows from the application of the Minkowski inequality in $L^p(\Omega_\theta)$. We conclude that GSW_p satisfies the triangle inequality.

Let $\theta^* = \arg\max_{\theta\in\Omega_\theta} W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta), \mathcal{G}\mathcal{I}_{\mu_3}(\cdot,\theta))$; then, we can write:

$$\begin{aligned} & \text{max-}GSW_p(I_{\mu_1}, I_{\mu_3}) \\ &= \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta), \mathcal{G}\mathcal{I}_{\mu_3}(\cdot,\theta)) \\ &= W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta^*), \mathcal{G}\mathcal{I}_{\mu_3}(\cdot,\theta^*)) \\ &\leq W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta^*), \mathcal{G}\mathcal{I}_{\mu_2}(\cdot,\theta^*)) \\ &\quad + W_p(\mathcal{G}\mathcal{I}_{\mu_2}(\cdot,\theta^*), \mathcal{G}\mathcal{I}_{\mu_3}(\cdot,\theta^*)) \\ &\leq \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}\mathcal{I}_{\mu_1}(\cdot,\theta), \mathcal{G}\mathcal{I}_{\mu_2}(\cdot,\theta)) \\ &\quad + \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_{\mu_2}(\cdot,\theta), \mathcal{G}I_{\mu_3}(\cdot,\theta)) \\ &\leq \text{max-}GSW_p(I_{\mu_1}, I_{\mu_2}) + \text{max-}GSW_p(I_{\mu_2}, I_{\mu_3}) \end{aligned}$$

So max-GSW_p satisfies the triangle inequality.

What remains to be proved is $GSW_p(I_\mu, I_\nu) = 0$ (or max-$GSW_p(I_\mu, I_\nu) = 0$) if and only if $\mu = \nu$. $GSW_p(I_\mu, I_\mu) = 0$ and max-$GSW_p(I_\mu, I_\mu) = 0$ follow directly from $W_p(\mu,\mu) = 0$ for any μ, and $GSW_p(I_\mu, I_\nu) = 0$ (or max-$GSW_p(I_\mu, I_\nu) = 0$) is equivalent to $\mathcal{G}I_\mu(\cdot,\theta) = \mathcal{G}I_\nu(\cdot,\theta)$ for almost all $\theta \in \Omega_\theta$. We conclude that GSW and max-GSW are distances if and only if $\mathcal{G}I_\mu(\cdot,\theta) = \mathcal{G}I_\nu(\cdot,\theta)$ implies $\mu = \nu$, *i.e.* the generalized Radon transform is injective.

□

Remark 1. *If the chosen generalized Radon transform is not injective, then we can only say that the GSW and max-GSW dissimilarity measures are pseudo-metrics: they still satisfy non-negativity, symmetry, the triangle inequality, and $GSW_p(I_\mu, I_\mu) = 0$ and max-$GSW_p(I_\mu, I_\mu) = 0$.*

Algorithm 1 GSW Distance

input $\{x_i \sim I_\mu\}_{i=1}^N$, $\{y_i \sim I_\nu\}_{i=1}^N$
Defining function $g(x, \theta)$, number of slices L, and p
Initialize $d = 0$
for $l = 1$ to L **do**
Sample θ_l from Ω_θ uniformly
Calculate $\{\hat{x}_i = g(x_i, \theta_l)\}_{i=1}^N$ and $\{\hat{y}_i = g(y_i, \theta_l)\}_{i=1}^N$
Sort $\hat{x}_i$ and $\hat{y}_i$ in ascending order such that $\hat{x}_{i[n]} \leq \hat{x}_{i[n+1]}$
$d = d + \frac{1}{L}\sum_{n=1}^N |\hat{x}_{i[n]} - \hat{y}_{i[n]}|^p$
end for
output $d^{\frac{1}{p}} \approx GSW_p(I_\mu, I_\nu)$

Algorithm 2 Max-GSW Distance

input $\{x_i \sim I_\mu\}_{i=1}^N$, $\{y_i \sim I_\nu\}_{i=1}^N$
Defining function $g(x, \theta)$, and p
Randomly initialize $\theta \in \Omega_\theta$
while θ is not converged **do**
Calculate $\{\hat{x}_i = g(x_i, \theta)\}_{i=1}^N$ and $\{\hat{y}_i = g(y_i, \theta_l)\}_{i=1}^N$
Sort $\hat{x}_i$ and $\hat{y}_i$ in ascending order such that $\hat{x}_{i[n]} \leq \hat{x}_{i[n+1]}$
$\theta = Proj_{\Omega_\theta}(ADAM(\nabla_\theta(\frac{1}{N}\sum_{n=1}^N |\hat{x}_{i[n]} - \hat{y}_{j[n]}|^p), \theta))$
end while
Sort $\hat{x}_i$ and $\hat{y}_i$ in ascending order
$d = \frac{1}{N}\sum_{n=1}^N |\hat{x}_{i[n]} - \hat{y}_{i[n]}|^p$
output $d^{\frac{1}{p}} \approx$ max-$GSW_p(I_\mu, I_\nu)$

3.1. Conditions for the injectivity of GRT

In the previous section, we have shown that the injectivity of GRT is crucial for GSW and max-GSW distances to be valid distances for probability measures. In this section, we will identify the conditions for the GRT to be injective.

The investigation of the sufficient and necessary conditions for showing GRT to be injective is a long-standing topic (Beylkin, 1984; Homan & Zhou, 2017; Uhlmann, 2003; Ehrenpreis, 2003)

More interestingly, one can also show that any homogeneous polynomial with an odd degree yields an injective GRT (Rouviere, 2015), i.e.

$$g(x, \theta) = \sum_{|\alpha|=m} \theta_\alpha x^\alpha, \tag{14}$$

where we use the multi-index notation, $\alpha \triangleq (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, $|\alpha| \triangleq \sum_{i=1}^d \alpha_i$, and $x^\alpha \triangleq \prod_{i=1}^d x_i^{\alpha_i}$. Here, the summation iterates over all possible multi-indices α, such that $|\alpha| = m$, where m denotes the degree of the polynomial and $\theta_\alpha \in \mathbb{R}$. We can observe that choosing $m = 1$ reduces to the linear case $x \cdot \theta$, since the set of the multi-indices with $|\alpha| = 1$ becomes $\{(\alpha_1, \ldots, \alpha_d); \alpha_i = 1, \text{for a single } i \in [\![1, d]\!], \text{ and } \alpha_j = 0, \quad \forall j \neq i\}$ and contains d elements.

4. Numerical implementation

Here we briefly review the numerical method used in calculating GSW and max-GSW.

4.1. Numerical implementation of GSW_p

Let $\{x_i \sim I_\mu\}_{i=1}^N$ and $\{y_j \sim I_\nu\}_{j=1}^N$ be samples from distributions I_μ, and I_ν, and let $g(\cdot, \theta)$ be the defining function. Following the work of Kolouri et al. (2019), the Wasserstein distance between one-dimensional distributions $\mathcal{G}I_\mu(\cdot, \theta)$ and $\mathcal{G}I_\nu(\cdot, \theta)$ can be calculated from sorting their samples and calculating the ℓ_p distance between the sorted samples. In other words, the GSW distance between I_μ and I_ν can be approximated from their samples as follows:

$$GSW_p(I_\mu, I_\nu) \approx \left(\frac{1}{|\Theta|}\sum_{\theta_l \in \Theta}\sum_{n=1}^N |g(x_{i[n]}, \theta) - g(y_{j[n]}, \theta)|^p\right)^{\frac{1}{p}}$$

where $i[m]$ and $j[n]$ are the indices of sorted $\{g(x_i, \theta)\}_{i=1}^N$ and $\{g(y_j, \theta)\}_{j=1}^N$. The algorithm to calculate the GSW distance is shown in Algorithm 1.

4.2. Numerical implementation of max-GSW_p

To calculate max-GSW_p we perform an EM like optimization scheme, where: a) for a fixed θ, $g(x_i, \theta)$ and $g(y_i, \theta)$ are first sorted to calculate the W_p distance, and b) update θ using:

$$\theta = Proj_{\Omega_\theta}(ADAM(\nabla_\theta(\frac{1}{N}\sum_{n=1}^N |g(x_{i[n]}, \theta) - g(y_{j[n]}, \theta)|^p), \theta))$$

where we use ADAM optimizer to update θ, and $Proj_{\Omega_\theta}(\cdot)$ projects θ onto Ω_θ. For instance, when $\theta \in \mathbb{S}^{n-1}$, $Proj_{\Omega_\theta}(\theta) = \frac{\theta}{\|\theta\|}$. Algorithm 2 shows the algorithm to calculate max-GSW_p. Here we note that the described optimization to find the optimal θ optimizes the actual W_p, as opposed to the heuristic approaches proposed in (Deshpande et al., 2018; Kolouri et al., 2019), which calculate the pseudo-optimal slice via perceptrons or via penalized linear discriminant analysis (Wang et al., 2011). Finally, after convergence is achieved the max-GSW_pdistance is approximated as:

$$\text{max-}GSW_p(I_\mu, I_\nu) \approx \left(\frac{1}{N}\sum_{n=1}^N |g(x_{i[n]}, \theta^*) - g(y_{j[n]}, \theta^*)|^p\right)^{\frac{1}{p}}.$$

5. Experiments

5.1. Generalized Sliced-Wasserstein Flows

For our first experiment, in order to demonstrate the effect of the choice of the GSW distance in its purest form, we consider the following problem,

$$\min_\mu GSW(\mu, \nu) \tag{15}$$

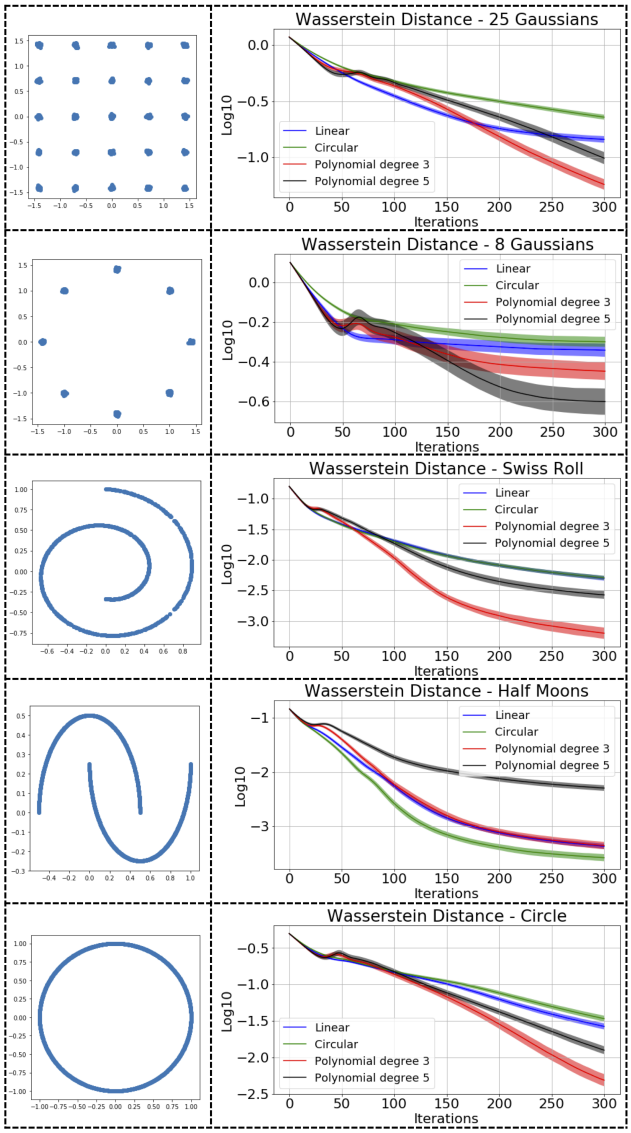

Figure 2. Log 2-Wasserstein distance measured between the source and target distributions as a function of number of iterations for five classic target distributions.

where ν is a target distribution, and μ is the source distribution, which is initialized to be the normal distribution. The optimization is then solved iteratively via,

$$\partial_t \mu_t = -\nabla GSW(\mu_t, \nu), \quad \mu_0 = \mathcal{N}(0, 1)$$

Here, we used 5 well-known distributions as the target distribution, namely the 25-Gaussians, 8-Gaussians, Swiss Roll, Half Moons, and the Circle distributions. For the GSW we used linear (i.e., SW distance), circular, homogeneous polynomial of degree 3, and homogeneous polynomial of degree 5. We used the exact same optimization scheme for all methods, used $L = 10$ random projections in each iteration,

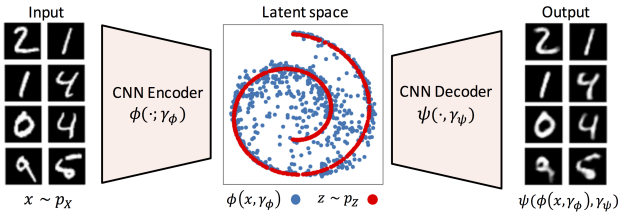

Figure 3. Schematic of the SWAE architecture. The distribution of the embedded data in the latent space is enforced to follow a prior samplable distribution p_Z.

and measured the 2-Wasserstein distance between μ_t and ν, at each iteration of the optimization (via solving a linear programming at each step). We repeated each experiment 100 times and report the mean and standard deviation of the 2-Wasserstein distance for all five target datasets in Figure 2. While the choice of the defining function $g(\cdot, \theta)$ is data dependent, it can be seen that the homogeneous polynomial of degree 3 is among the top two performers for all datasets.

Furthermore, we note that we avoided reporting the max-GSW_p results in Figure 2 to avoid clutter and confusion. The results for max-GSW_p for the same experiment is included in the supplementary material.

5.2. Generative Modeling via Auto-Encoders

Here we demonstrate the application of the GSW and max-GSW distances in generative modeling. We specifically use the recently proposed Sliced-Wasserstein Auto-Encoder (SWAE) (Kolouri et al., 2019) framework, which penalizes the distribution of the encoded data in the latent space of the auto-encoder to follow a prior samplable distribution, p_Z. More precisely, let $\{x_n \sim p_X\}_{n=1}^N$ be i.i.d. samples from p_X, $\phi(x, \gamma_\phi) : \mathcal{X} \to \mathcal{Z}$ and $\psi(z, \gamma_\psi) : \mathcal{Z} \to \mathcal{X}$ be the parametric encoder and decoder (e.g., CNNs) with parameters γ_ϕ and γ_ψ, respectively. Then SWAE's objective function (Kolouri et al., 2019) is defined as:

$$\min_{\gamma_\phi, \gamma_\psi} \mathbb{E}_x[c(x, \psi(\phi(x, \gamma_\phi), \gamma_\psi))] + \lambda SW(p_{\phi(x,\gamma_\phi)}, p_Z) \tag{16}$$

where λ is the regularizer coefficient for matching the encoded distribution to p_Z. Here, we substitute the SW distance in Equation 16 with GSW and max-GSW distances. Specifically, we encode the MNIST dataset (LeCun et al., 1998) into the encoder's latent space and enforce the distribution of the embedded data to follow a prior distribution, e.g., the Swiss Roll distribution as shown in Figure 3, while we simultaneously enforce the encoded features to be decodable to the original input images.

Similar to the previous section, we ran the optimization in Equation (16) with the GSW distances, with linear, circular,

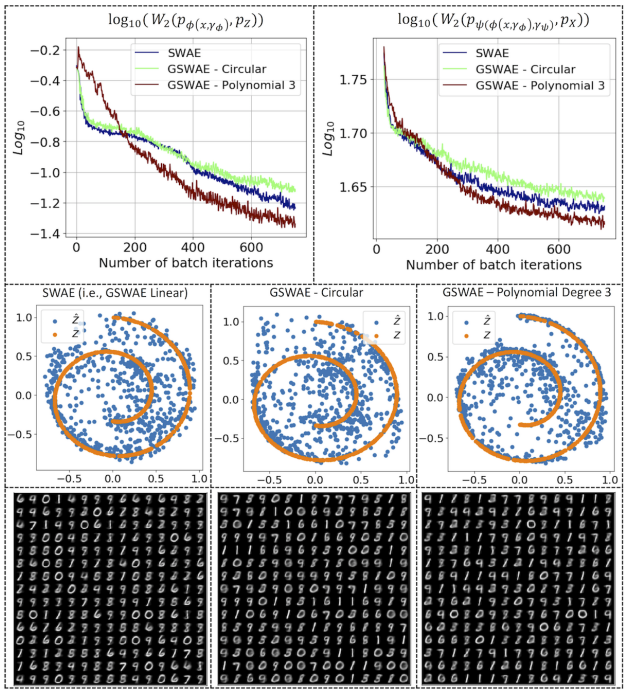

Figure 4. The $2-$Wasserstein distance between p_Z and $p_{\phi(x,\gamma_\phi)}$ and between p_X and $p_{\psi(\phi(x,\gamma_\phi),\gamma_\psi)}$ at different batch iterations for SWAE and GSWAE with circular and polynomial of degree 3 defining functions.

and homogeneous polynomial of degree 3. In each iteration, we measured the 2-Wasserstein distance between the embedded distribution and the prior distribution, $W_2(p_{\phi(x,\gamma_\phi)}, p_Z)$, and also between the data distribution and the distribution of the reconstructed samples, $W_2(p_{\psi(\phi(x,\gamma_\phi),\gamma_\psi)}, p_X)$. Each experiment was repeated 50 times and the average Wasserstein distances are reported in Figure 4. Figure 4, middle row, shows samples from p_Z and $\phi(x, \gamma_\phi)$ for $x \sim p_X$, and the last row shows decoded random samples, $\psi(z, \gamma_\psi)$ for $z \sim p_Z$.

6. Conclusion

In this paper, we generalized the definition of the celebrated sliced-Wasserstein distances to the generalized sliced Wasserstein distance.

Kimia: As future work: mention the estimation of g with a neural network.

References

Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

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

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

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

Bousquet, Olivier, Gelly, Sylvain, Tolstikhin, Ilya, Simon-Gabriel, Carl-Johann, and Schoelkopf, Bernhard. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

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

Carriere, Mathieu, Cuturi, Marco, and Oudot, Steve. Sliced wasserstein kernel for persistence diagrams. In *ICML 2017-Thirty-fourth International Conference on Machine Learning*, pp. 1–10, 2017.

Courty, Nicolas, Flamary, Rémi, Tuia, Devis, and Rakotomamonjy, Alain. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cuturi, Marco and Peyré, Gabriel. A Smoothed Dual Approach for Variational Wasserstein Problems. *SIAM Journal on Imaging Sciences*, December 2015. URL https://hal.archives-ouvertes.fr/hal-01188954.

Denisyuk, AS. Inversion of the generalized radon transform. *Translations of the American Mathematical Society-Series 2*, 162:19–32, 1994.

Deshpande, Ishan, Zhang, Ziyu, and Schwing, Alexander. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3483–3491, 2018.

Ehrenpreis, Leon. *The universality of the Radon transform*. Oxford University Press on Demand, 2003.

Frogner, Charlie, Zhang, Chiyuan, Mobahi, Hossein, Araya, Mauricio, and Poggio, Tomaso A. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Gel'fand, Israel M, Graev, Mark Iosifovich, and Shapiro, Z Ya. Differential forms and integral geometry. *Functional Analysis and its Applications*, 3(2):101–114, 1969.

Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Helgason, Sigurdur. The radon transform on rn. In *Integral Geometry and Radon Transforms*, pp. 1–62. Springer, 2011.

Homan, Andrew and Zhou, Hanming. Injectivity and stability for a generic class of generalized radon transforms. *The Journal of Geometric Analysis*, 27(2):1515–1529, 2017.

Karras, Tero, Aila, Timo, Laine, Samuli, and Lehtinen, Jaakko. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Kitagawa, Jun, Mérigot, Quentin, and Thibert, Boris. Convergence of a newton algorithm for semi-discrete optimal transport. *arXiv preprint arXiv:1603.05579*, 2016.

Kolouri, Soheil, Zou, Yang, and Rohde, Gustavo K. Sliced-Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4876–4884, 2016.

Kolouri, Soheil, Park, Se Rim, Thorpe, Matthew, Slepcev, Dejan, and Rohde, Gustavo K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

Kolouri, Soheil, Rohde, Gustavo K., and Hoffmann, Heiko. Sliced wasserstein distance for learning gaussian mixture models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Kolouri, Soheil, Pope, Phillip E., Martin, Charles E., and Rohde, Gustavo K. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1xaJn05FQ.

Kuchment, Peter. Generalized transforms of radon type and their applications. In *Proceedings of Symposia in Applied Mathematics*, volume 63, pp. 67, 2006.

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

Lévy, Bruno. A numerical algorithm for L_2 semi-discrete optimal transport in 3D. *ESAIM Math. Model. Numer. Anal.*, 49(6):1693–1715, 2015. ISSN 0764-583X. doi: 10.1051/m2an/2015055. URL http://dx.doi.org/10.1051/m2an/2015055.

Montavon, Grégoire, Müller, Klaus-Robert, and Cuturi, Marco. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pp. 3718–3726, 2016.

Natterer, Frank. *The mathematics of computerized tomography*, volume 32. Siam, 1986.

Oberman, Adam M and Ruan, Yuanlong. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

Peyré, Gabriel and Cuturi, Marco. Computational optimal transport. *arXiv preprint arXiv:1803.00567*, 2018.

Radon, Johann. Uber die bestimmug von funktionen durch ihre integralwerte laengs geweisser mannigfaltigkeiten. *Berichte Saechsishe Acad. Wissenschaft. Math. Phys., Klass*, 69:262, 1917.

Rouviere, Francois. Nonlinear radon and fourier transforms. https://math.unice.fr/~frou/recherche/Nonlinear%20RadonW.pdf, 2015.

Schmitz, Morgan A, Heitz, Matthieu, Bonneel, Nicolas, Ngole, Fred, Coeurjolly, David, Cuturi, Marco, Peyré, Gabriel, and Starck, Jean-Luc. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Schmitzer, Bernhard. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, Oct 2016. ISSN 1573-7683. doi: 10.1007/s10851-016-0653-9. URL https://doi.org/10.1007/s10851-016-0653-9.

Şimşekli, Umut, Liutkus, Antoine, Majewski, Szymon, and Durmus, Alain. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. *arXiv preprint arXiv:1806.08141*, 2018.

Solomon, Justin, Rustamov, Raif, Guibas, Leonidas, and Butscher, Adrian. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pp. 306–314, 2014.

Solomon, Justin, De Goes, Fernando, Peyré, Gabriel, Cuturi, Marco, Butscher, Adrian, Nguyen, Andy, Du, Tao, and Guibas, Leonidas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

Tolstikhin, Ilya, Bousquet, Olivier, Gelly, Sylvain, and Schoelkopf, Bernhard. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Uhlmann, Gunther. *Inside out: inverse problems and applications*, volume 47. Cambridge University Press, 2003.

Villani, Cédric. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wang, Wei, Mo, Yilin, Ozolek, John A, and Rohde, Gustavo K. Penalized fisher discriminant analysis and its application to image-based morphometry. *Pattern recognition letters*, 32(15):2128–2135, 2011.

7. Supplementary material

7.1. Non-negativity and Symmetry for GSW

In this section, we show that the GSW and max-GSW dissimilarity measures satisfy non-negativity and symmetry. Let μ, ν in $\mathcal{P}_p(\Omega)$.

7.1.1. Non-negativity

We use the non-negativity of the p-Wasserstein distance, *i.e.* $W_p(\mu,\nu) \geq 0$ for any μ, ν in $\mathcal{P}_p(\Omega)$, to show that GSW and max-GSW are non-negative as well:

$$\begin{aligned} GSW_p(I_\mu, I_\nu) &= \left(\int_{\Omega_\theta} W_p^p(\mathcal{G}I_\mu(.,\theta), \mathcal{G}I_\nu(.,\theta))d\theta\right)^{\frac{1}{p}} \\ &\geq \left(\int_{\Omega_\theta} (0)^p d\theta\right)^{\frac{1}{p}} = 0 \end{aligned}$$

$$\begin{aligned} \text{max-}GSW_p(I_\mu, I_\nu) &= \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)) \\ &= W_p(\mathcal{G}I_\mu(\cdot,\theta^*), \mathcal{G}I_\nu(\cdot,\theta^*)) \\ &\geq 0 \end{aligned}$$

where $\theta^* = \arg\max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta))$.

7.1.2. Symmetry

Since the p-Wasserstein distance is symmetric, we have, for any μ, ν in $\mathcal{P}_p(\Omega)$, $W_p(\mu,\nu) = W_p(\nu,\mu)$. In particular, we can write

$$\forall\theta \in \Omega_\theta,\ W_p(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)) = W_p(\mathcal{G}I_\nu(\cdot,\theta), \mathcal{G}I_\mu(\cdot,\theta)) \tag{17}$$

and

$$\max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)) = \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}I_\nu(\cdot,\theta), \mathcal{G}I_\mu(\cdot,\theta)) \tag{18}$$

The symmetry of the GSW and max-GSW dissimilarity measures are direct consequences of Equations (17) and (18) respectively.