Generalized Sliced-Wasserstein Distances

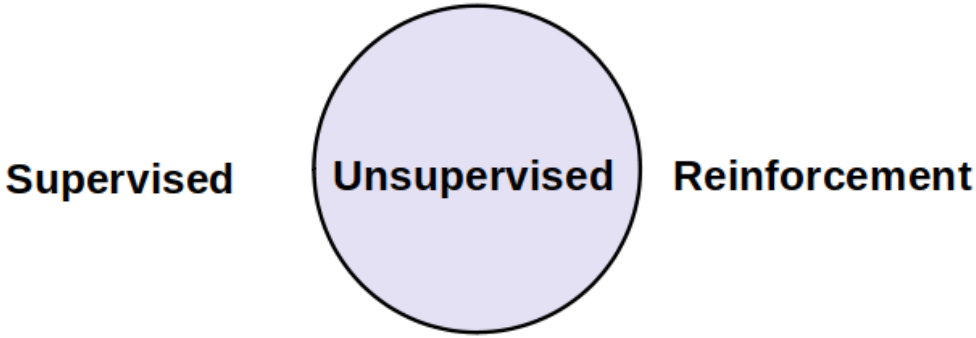
A new family of optimal transport metrics.
Neurips 2019

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

[1] HRL Laboratories, [2] Télécom Paris, [3] University of Virginia

Generative Models

Unsupervised Learning



Learning useful information from a dataset of *unlabeled* samples

"Humans build a model of the world through predictive unsupervised learning." – Yann Le Cun

"One of the main challenges for AI remains unsupervised learning, at which humans are much better than machines." – Yoshua Bengio

"What I cannot create, I cannot understand." – Richard Feynman

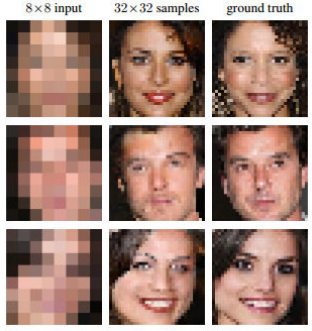
The Generative Approach

One of the most promising approaches toward unsupervised learning.

Learns the data distribution in an unsupervised manner.

Why?

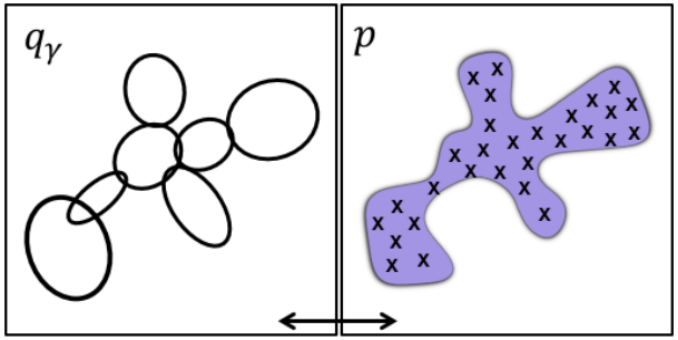
Missing data
NVIDIA, 2018



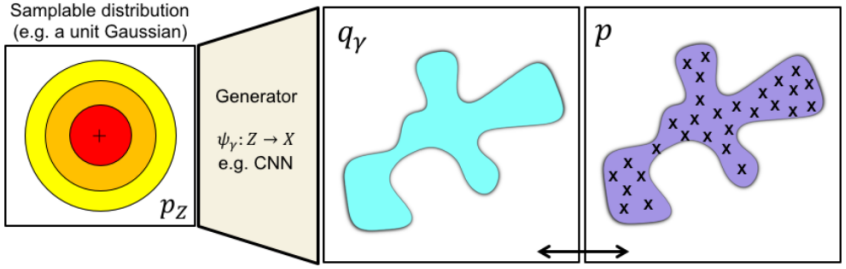
Super resolution
Dahl et al., CVPR 2017

Domain transfer
Zhu et al., ICCV 2017

Examples of Generative Models



Gaussian Mixture Models



Deep Generative Modeling (e.g., GANs)

Measuring the *dissimilarity between probability distributions* is at the heart of generative models.

Distances between Probability Measures

Examples of distances

Let μ and ν be probability measures with finite p'th moment, defined on $\mathsf{X} \subset \mathbb{R}^d$, i.e.,

$$\text{for some } x_0 \in \mathsf{X}, \int_{\mathsf{X}} \|x - x_0\| \mathrm{d}\mu(x) < +\infty$$

with corresponding densities l_μ and l_ν.

How "far" is μ from ν?

L_p-metrics, $p \geq 1$:	$\left(\int_{\mathsf{X}} \lvert l_\mu(x) - l_\nu(x)\rvert^p \, \mathrm{d}x\right)^{1/p}$
Hellinger distance:	$\left(\frac{1}{2} \int_{\mathsf{X}} \left(\sqrt{l_\mu(x)} - \sqrt{l_\nu(x)}\right)^2 \mathrm{d}x\right)^{1/2}$
Kullback–Leibler divergence:	$\int_{\mathsf{X}} l_\mu(x) \log\left(\frac{l_\mu(x)}{l_\nu(x)}\right) \mathrm{d}x$

Examples of distances

Let μ and ν be probability measures with finite p'th moment, defined on $\mathsf{X} \subset \mathbb{R}^d$. Suppose X is endowed with a distance d.

Wasserstein distance of order p.

$$\mathbf{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{X}} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathsf{X} \times \mathsf{X}$ with marginals μ (resp. ν) for the first (resp. second) variable.

Link with Optimal Transport

A solution γ is called the *transport plan* between μ and ν.

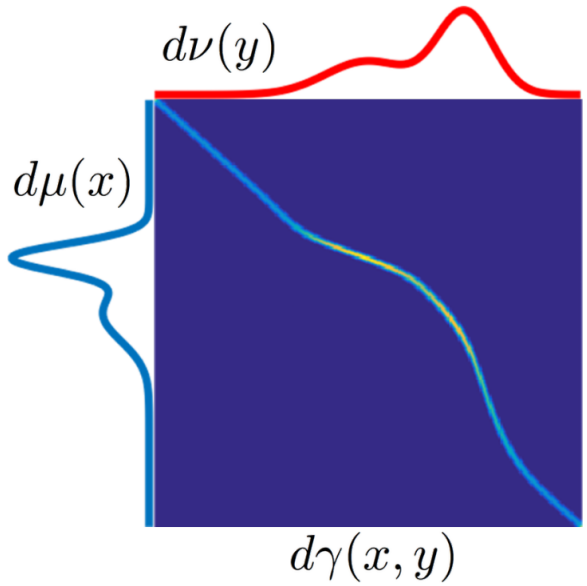
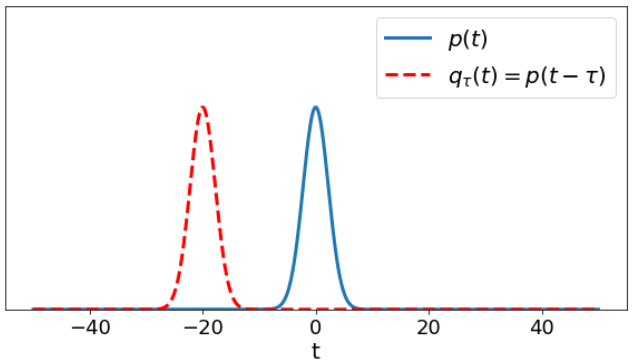
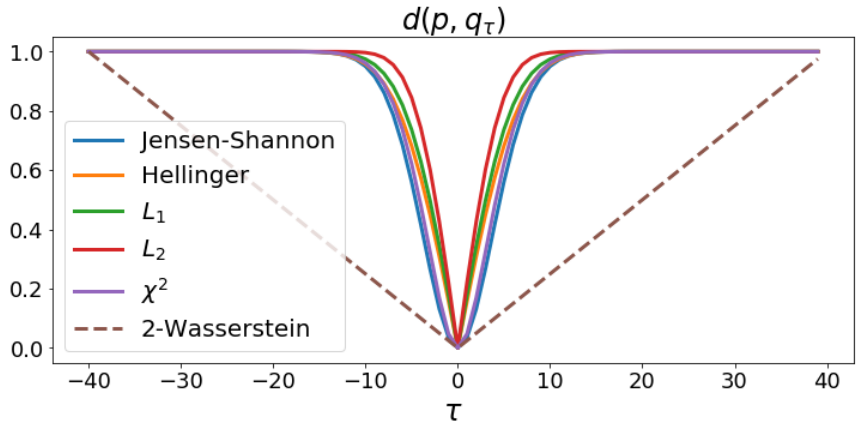


Illustration of optimal transport between two probability measures in the continuous case.

Comparison of distances





$\frac{\partial d(p,q_\tau)}{\partial \tau} \approx 0$ whenever p and q_τ are non-overlapping.

The Wasserstein distance captures the underlying geometry of the space and is suitable for learning.

Limitations of Optimal Transport

Computing the Wasserstein distance is expensive.

OT amounts to solving a large-scale linear program.

Computational complexity: $O\left(n^3 \log(n)\right)$, where n is the number of data samples.

Curse of dimensionality.

The error made when approximating the Wasserstein distance from samples grows *exponentially fast* with the dimension of the space.

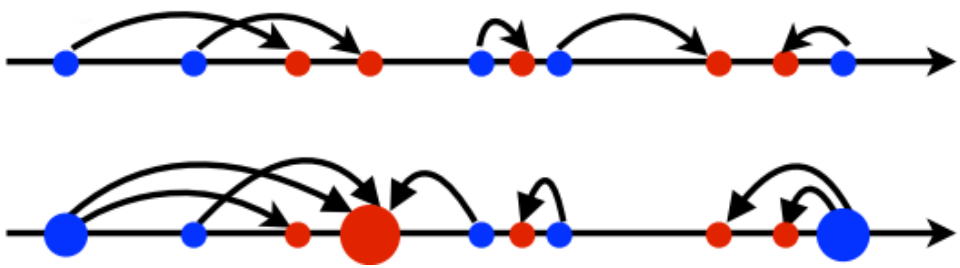
Sliced-Wasserstein Distance

Wasserstein distance in 1D.

If $X \subset \mathbb{R}$,

$$\mathbf{W}_p(\mu,\nu) = \left(\int_0^1 d\left(F_\mu^{-1}(t), F_\nu^{-1}(t)\right)^p \mathrm{d}t \right)^{1/p},$$

where F_μ^{-1} and F_ν^{-1} denote the quantile functions of μ and ν resp.



For empirical distributions: simply compute the average distance between the sorted samples. (Source: Computational Optimal Transport, G. Peyré and M. Cuturi)

Radon transform

Consider a function $h \in L^1(\mathbb{R}^d)$ where

$$L^1(\mathbb{R}^d) = \{h : \mathbb{R}^d \to \mathbb{R} \ / \int_{\mathbb{R}^d} |h(x)| \mathrm{d}x < \infty\}.$$

The standard Radon transform, $\mathcal{R}$, maps h to the infinite set of its *integrals over the hyperplanes* of $\mathbb{R}^d$:

$$\forall (t, \theta) \in \mathbb{R} \times \mathbb{S}^{d-1}, \ \mathcal{R}h(t, \theta) = \int_{\mathbb{R}^d} h(x) \delta(t - \langle x, \theta \rangle) \mathrm{d}x,$$

where,

$\mathbb{S}^{d-1} \subset \mathbb{R}^d$: the d-dimensional unit sphere,

$\delta(\cdot)$: the Dirac delta function in 1D,

$\langle \cdot, \cdot \rangle$: the Euclidean inner-product.

Radon transform

Each hyperplane can be written as:

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

(A level set of $g(x, \theta) = \langle x, \theta \rangle$)

Fix $\theta \in \mathbb{S}^{d-1}$. The integrals over all hyperplanes orthogonal to θ define a continuous function $\mathcal{R}h(\cdot, \theta) : \mathbb{R} \to \mathbb{R}$ which is a *"slice"* of h.

Sliced-Wasserstein distance

Motivation.

Define an alternative optimal transport distance which does not suffer from a computational burden.

Idea.

Utilize the closed-form formula of the Wasserstein distance in 1D:

1. Obtain representations in 1D for a higher-dimensional distribution through linear projections (via the Radon transform),
2. Calculate the distance between two input distributions as a functional on the Wasserstein distance of their 1D representations.

Sliced-Wasserstein distance

Let μ and ν be probability measures with finite p'th moment, defined on $\mathsf{X} \subset \mathbb{R}^d$, with corresponding densities I_μ and I_ν.

Sliced-Wasserstein distance of order p.

$$\mathbf{SW}_p(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\big)\mathrm{d}\theta \right)^{1/p}$$

The SW-distance is, indeed, a distance [1, 3]

Limitation of Sliced-Wasserstein distance

In practice, we use a Monte-Carlo approximation where samples $\{\theta_l\}_l$ are uniformly drawn on $\mathbb{S}^{d-1}$.

But in high-dimension, there is a high chance that, for any randomly sampled θ, $\mathbf{W}_p\big(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\big) \approx 0$.

Maximum Sliced-Wasserstein distance

Sliced-Wasserstein distance of order p.

$$\mathbf{SW}_p(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\big)\mathrm{d}\theta\right)^{1/p}$$

Maximum Sliced-Wasserstein distance of order p. [2]

$$\text{max-}\mathbf{SW}_p(\mu,\nu) = \max_{\theta\in\mathbb{S}^{d-1}} \mathbf{W}_p\big(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\big)$$

Maximum Sliced-Wasserstein distance

Sliced-Wasserstein distance of order p.

$$\mathbf{SW}_p(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\big)\mathrm{d}\theta\right)^{1/p}$$

Maximum Sliced-Wasserstein distance of order p. [2]

$$\text{max-}\mathbf{SW}_p(\mu,\nu) = \max_{\theta\in\mathbb{S}^{d-1}} \mathbf{W}_p\big(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\big)$$

What happens if we use *nonlinear* projections?

Generalized Sliced-Wasserstein Distances

Generalized Radon transform

The Generalized Radon transform (GRT, $\mathcal{G}$) integrates the data distribution over general hyper-surfaces (instead of hyperplanes) $H(t,\theta) = \{x \in \mathbb{R}^d \mid g(x,\theta) = t\}$, i.e. :

$$\forall (t,\theta) \in \mathbb{R} \times \Omega_\theta,\ \mathcal{G}h(t,\theta) = \int_{\mathbb{R}^d} h(x)\delta(t - g(x,\theta))\mathrm{d}x$$

where Ω_θ is a compact set of feasible parameters for $\theta \mapsto g(\cdot,\theta)$.

Generalized Radon transform

g is the *defining function* and must satisfy:

H1. g 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.*,

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

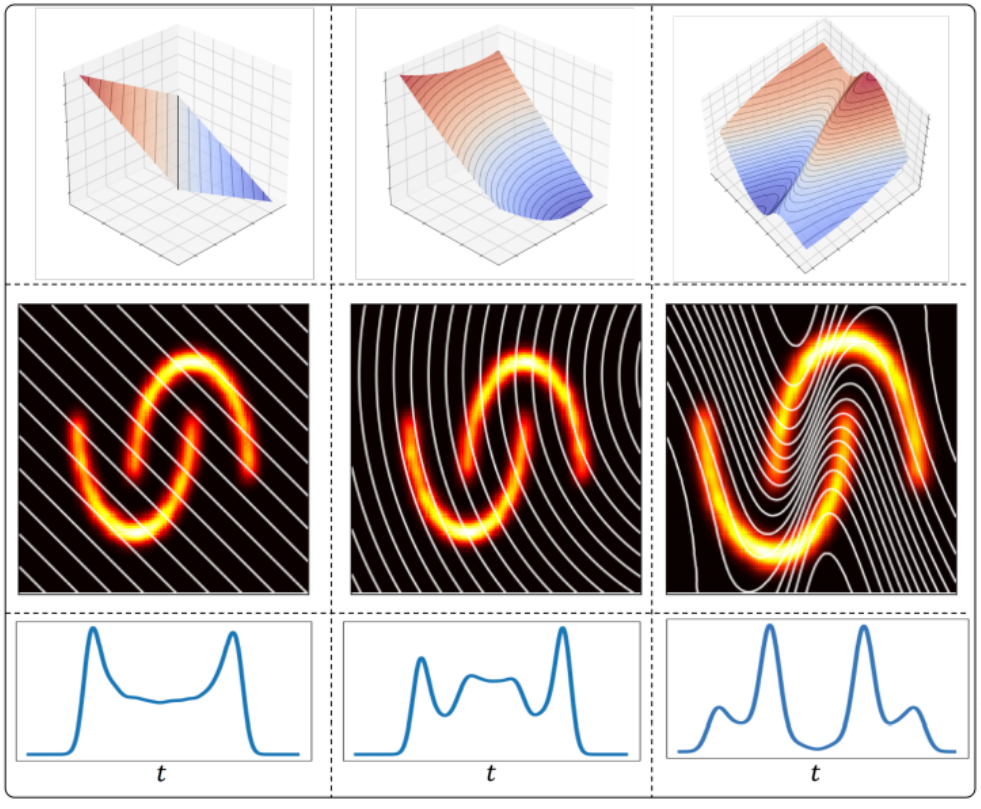
H3. g is non-degenerate in the sense that,

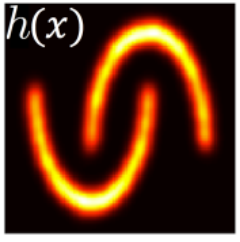
$$\forall (x, \theta) \in \mathcal{X} \times \mathbb{R}^n \backslash \{0\},\ \frac{\partial g}{\partial x}(x, \theta) \neq 0$$

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

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

Generalized Radon transform





Input distribution

$\mathcal{G}h(t,\theta)$: Slices with respect to different $g(t,\theta)$

$H(t,\theta) = \{x | g(x,\theta) = t\}$

Generalized Sliced-Wasserstein distances

Let μ and ν be probability measures with finite p'th moment, defined on $\mathsf{X} \subset \mathbb{R}^d$, with corresponding densities I_μ and I_ν.

Generalized Sliced-Wasserstein distance of order p.

$$\textbf{GSW}_p(\mu,\nu) = \left(\int_{\Omega_\theta} \mathbf{W}_p^p\big(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)\big)\mathrm{d}\theta\right)^{1/p}$$

Maximum Generalized Sliced-Wasserstein distance of order p.

$$\text{max-}\textbf{GSW}_p(\mu,\nu) = \max_{\theta\in\Omega_\theta} \mathbf{W}_p\big(\mathcal{G}I_\mu(\cdot,\theta), \mathcal{G}I_\nu(\cdot,\theta)\big)$$

Generalized Sliced-Wasserstein distance

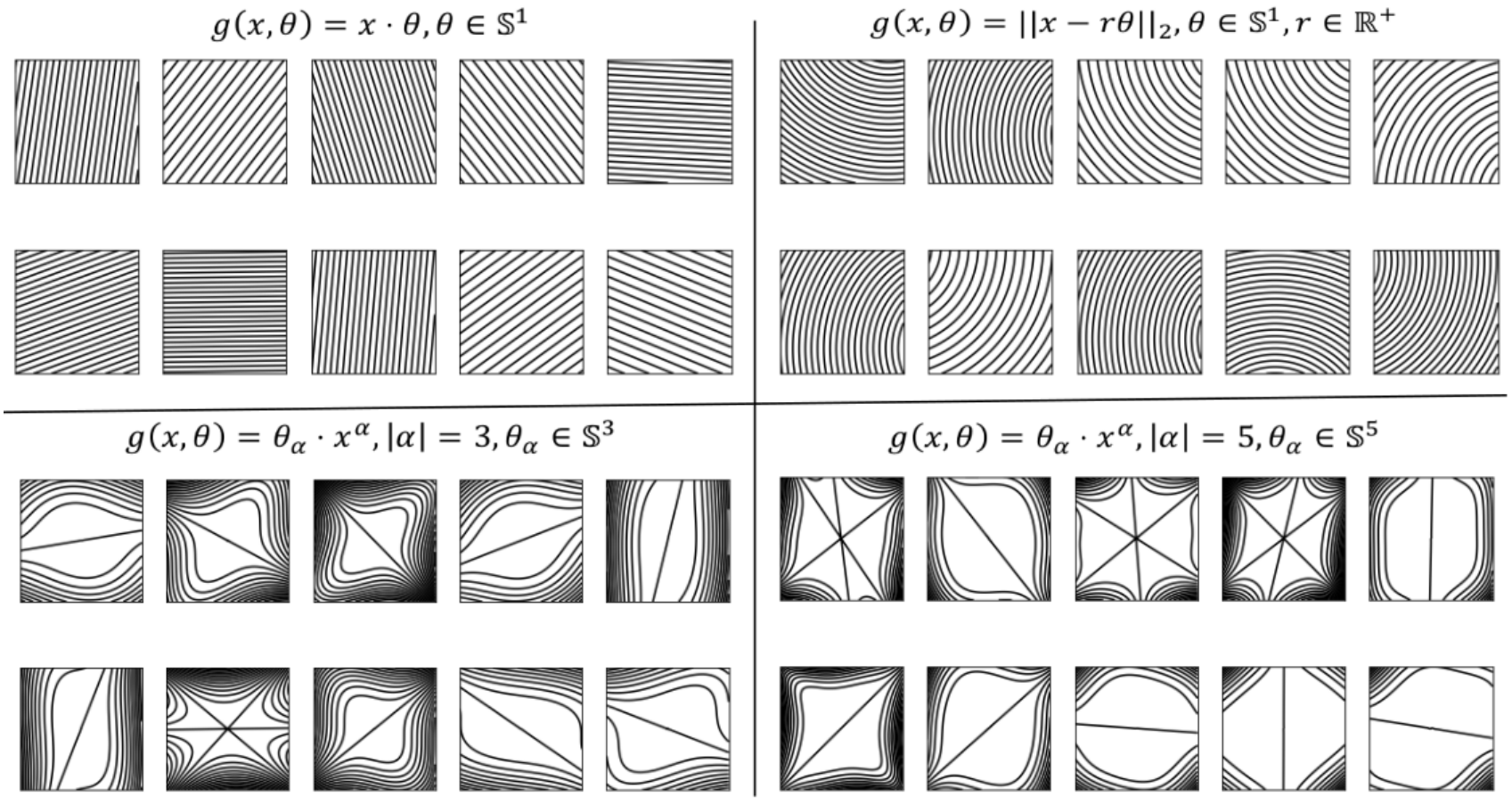
Proposition

The Generalized Sliced-Wasserstein and Maximum Generalized Sliced-Wasserstein distances of order p are, indeed, distances if and only if the Generalized Radon transform is injective.

If the GRT is not injective, then GSW and max-GSW are *pseudo-metrics*.

Choice of the defining function g

There exists a family of defining functions that provide injective GRT (e.g., homogeneous polynomials with odd degree).



Randomly generated integration hypersurfaces (curves in 2D)

Choice of the defining function g

The memory complexity of polynomial projections grows exponentially with the data dimension and the polynomial degree.

Alternative solution: define g as a neural network. We propose a multi-layer fully connected network with *leaky ReLU* activations.

With a neural network-based defining function, minimizing the max-GSW distance between two distributions is analogical to the adversarial learning.

Experiments

GSW flows

Consider the following problem:

$$\min_{\mu} \ \mathbf{GSW}_p(\mu, \nu),$$

where ν is a target distribution and μ is the source distribution, initialized to a normal distribution.

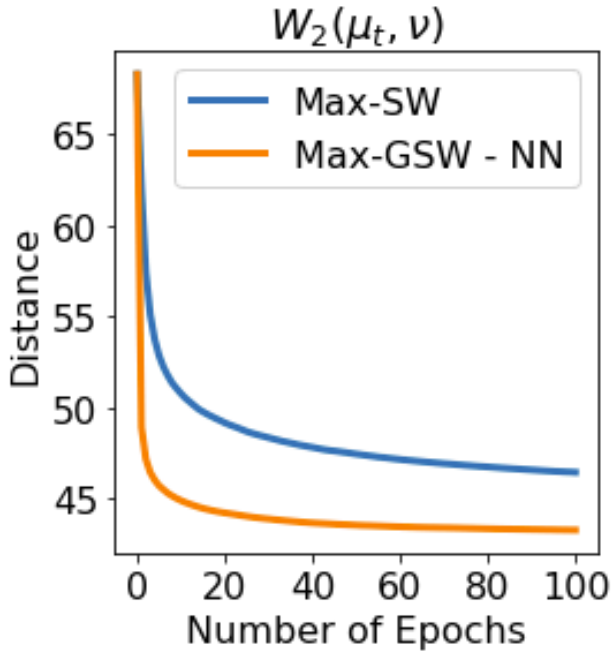
The optimization is then solved iteratively via:

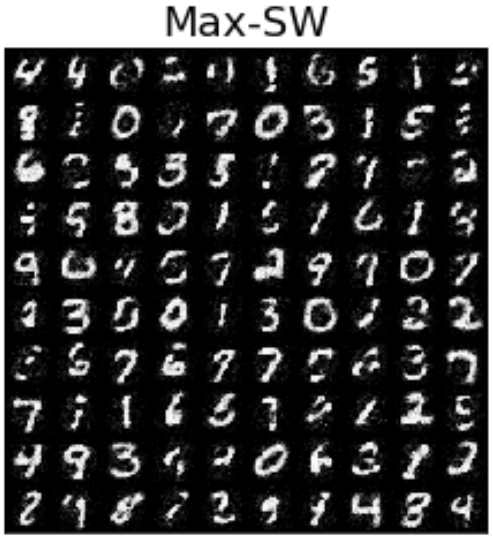
$$\partial_t \mu_t = -\nabla \mathbf{GSW}_p(\mu_t, \nu), \quad \mu_0 = \mathcal{N}(0, 1).$$

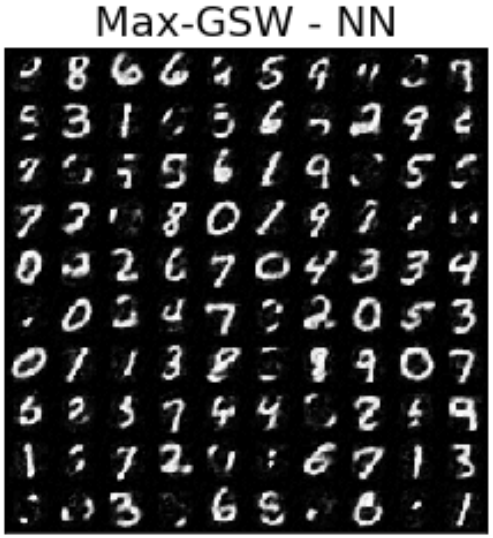
GSW flows: Synthetical datasets

(See videos in the supplementary files.)

GSW flows: Hand-written digit recognition

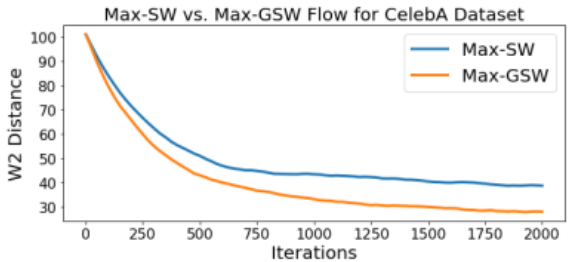


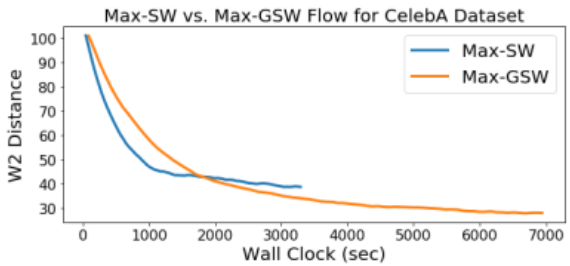




Wasserstein distance of order 2 between source and target distributions for the MNIST dataset.

GSW flows: Face generation





Wasserstein distance of order 2 between source and target distributions for the CelebA dataset.

Conclusion

Summary

- We introduced a new family of optimal transport metrics for probability measures that generalizes the Sliced-Wasserstein distance.
- We provided theoretical conditions that yield GSW and max-GSW to be, indeed, distances.
- We empirically demonstrated the superior performance of GSW and max-GSW over SW in various generative modeling applications.

References i


N. Bonnotte.
Unidimensional and evolution methods for optimal transportation.
PhD thesis, Université Paris 11, France, 2013.


I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. Schwing.
Max-sliced wasserstein distance and its use for gans.
In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.


S. Kolouri, Y. Zou, and G. K. Rohde.
Sliced-Wasserstein kernels for probability distributions.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2016.