# Distributional Sliced-Wasserstein and Applications to Generative Modeling

**Khai Nguyen**[1], Nhat Ho[2], Tung Pham[1], Hung Bui[1]

[1]VinAI Research, Vietnam [2]University of Texas, Austin

VinAI RESEARCH

TEXAS
The University of Texas at Austin

# Wasserstein Distance

**Wasserstein distance** is a metric between two probability measures that follows Kantorovich formulation of Optimal Transport:

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

- $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ which is defined on a given metric space $(\mathbb{R}^d, ||.||)$
- $\Pi(\mu,\nu)$ is a set of all transportation plans $\pi$ such that the marginal distributions are $\mu, \nu$

**Pros:**

- Meaningful metric
- Able to work with empirical measures
- Stable and versatile

**Cons:**

- Suffer from the curse of dimensionality
- High computational complexity $\mathcal{O}(n^3 \log n)$ with $n$ is the number of supports of $\mu, \nu$ when they are empirical measures

# Slicing with Radon Transform

**Radon Transform** maps a function $I \in \mathbb{L}^1(\mathbb{R}^d)$ to a set to the space of functions defined over space of lines. For $\theta \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}$, the Radon Transform is defined as:

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

- $\delta$ is the Dirac delta function
- $\langle . , . \rangle$ is the inner product
- $\theta$ is called projecting direction

- With each value of $\theta$, Radon Transform gives a 1-d function on the real line.
- Radon Transform is injective
- Radon Transform can be extended to Generalized Radon Transform

# Sliced Wasserstein Distance

**Sliced Wasserstein distance** is a variant that leverages the closed-form advantage of Wasserstein distance in 1D by using slicing technique

$$SW_p(\mu,\nu) := \left(\mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[W_p^p\left(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\right)\right]\right)^{1/p}$$

- $\mathcal{U}(\mathbb{S}^{d-1})$ is the uniform distribution on the hypersphere of $d$ dimension
- $I_\mu$ is the cumulative distribution function of $\mu$

Since the expectation is intractable, Monte Carlo scheme is used

$$SW_p(\mu,\nu) \approx \left(\frac{1}{L}\sum_{l=1}^{L} W_p^p\left(\mathcal{R}I_\mu(\cdot,\theta_l),\mathcal{R}I_\nu(\cdot,\theta_l)\right)\right)^{1/p}$$

- L is the number of projections, $\{\theta\}_{i=1}^{L} \sim \mathcal{U}(\mathbb{S}^{d-1})$

When $\mu,\nu$ are empirical measures with $n$ support points, each 1D Wasserstein can be solved at time of $\mathcal{O}(Ln\log n)$ by **sorting** projected supports, and

**SW does not suffer from the curse of dimensionality**

# Max Sliced Wasserstein Distance

**Max sliced Wasserstein distance** is a variant of sliced Wasserstein that tries to find the "best" projecting direction on the unit hypersphere:

$$\text{max-}SW_p(\mu,\nu) := \max_{\theta\in\mathbb{S}^{d-1}} W_p\left(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\right)$$

- Still is a metric between probability measures
- Require optimization over the unit-sphere to compute

- Low sample-projections complexity
- Can find the best discriminative projection
- No curse of dimensionality

# Distributional Sliced Wasserstein Distance

We generalize the idea of slicing by using a **generic distribution** over the space of projecting directions (the unit hypersphere)

$$D_p(\mu,\nu) := \left(\mathbb{E}_{\theta\sim\sigma}\left[W_p^p\left(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\right)\right]\right)^{1/p}$$

- $\sigma$ is an arbitrary distribution over sphere

Which $\sigma$ is good? We need to guarantee $\sigma$ putting masses to informative directions

$$D_p(\mu,\nu) := \sup_{\sigma\in\mathcal{P}(\mathbb{S}^{d-1})}\left(\mathbb{E}_{\theta\sim\sigma}\left[W_p^p\left(\mathcal{R}I_\mu(\cdot,\theta),\mathcal{R}I_\nu(\cdot,\theta)\right)\right]\right)^{1/p}$$

- $\mathcal{P}(\mathbb{S}^{d-1})$ Is the space of all distributions over the unit-hypersphere

⇨ This formulation will gives $\sigma \to \delta_{\theta^*}$ which is the Dirac distribution on the **max** slice.

# Distributional Sliced Wasserstein Distance

⇨ Need a **regularization** to avoiding collapsing.

Let $\theta, \theta'$ are two vectors on the unit-hypersphere $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d; ||\theta||_2^2 = 1\}$

$|\theta^\top \theta'|$ measures the “positive” angle between two vectors

The regularization:

$$\mathbb{E}_{\theta,\theta' \sim \sigma}[|\theta^\top \theta'|] \leq C$$

The final definition of **distributional sliced Wasserstein distance** (DSW):

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

- $\mathbb{M}_C := \{\sigma \in \mathcal{P}(\mathbb{S}^{d-1}) | \mathbb{E}_{\theta,\theta' \sim \sigma}[|\theta^\top \theta'| \leq C]\}$

# Distributional Sliced Wasserstein Distance

DSW is:

- A valid metric between two probability measures since it satisfies non-negativity, symmetry, triangle inequality and identity.
- The generalization of max-SW ( $C = 1$ )
- A sliced distance that does not suffer from the curse of dimensionality since

$$\mathbb{E}\left[DSW_p\left(\mu_n, \mu; C\right)\right] \leq c\sqrt{\frac{d \log n}{n}}$$

  - $\mu$ is supported on a compact subset in $\mathbb{R}^d$, $\mu_n$ is the n-supports empirical measure of $\mu$
  - $c > 0$ is some universal constant

- $DSW_p(\mu, \nu; C) \leq \max SW_p(\mu, \nu) \leq W_p(\mu, \nu)$
- If $C \geq 1/d$, $DSW_p(\mu, \nu; C) \geq \left(\frac{1}{d}\right)^{1/p} \max SW_p(\mu, \nu) \geq \left(\frac{1}{d}\right)^{1/p} SW_p(\mu, \nu)$

⇨ The convergence of probability measures under DSW implies the convergence of these measures under max-SW, Wasserstein distance and vice versa

# Computation of DSW

Dual form of DSW:

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

- $\lambda_C$ is the Lagrange multiplier
- Each value of $C$ has a corresponding optimal $\lambda_C$
- $\lambda_C \to 0 \leftrightarrow C \to 1 \leftrightarrow DSW \to \text{max-}SW$

Let $f : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ be a Borel measurable function

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

- $\mathcal{F}$ a class of all Borel measurable functions from $\mathbb{S}^{d-1}$ to $\mathbb{S}^{d-1}$
- We can limit the function space by parametrizing $f$ with some parameters $\phi$ (e.g. a neural net)

# Computation of DSW

To solve the optimization problem, we use stochastic gradient estimation

$$\nabla_\phi DSW_p^*(\mu,\nu;C) \approx \nabla_\phi \Big( (\textstyle\sum_{l=1}^{L}[W_p^p(\mathcal{R}I_\mu(\cdot,f_\phi(\theta_l)),\mathcal{R}I_\nu(\cdot,f_\phi(\theta_l))])^{1/p} - \lambda_C \sum_{l=1}^{L}\sum_{l'}^{L}[|f_\phi(\theta_l)^\top f_\phi(\theta_{l'})|] \Big)$$

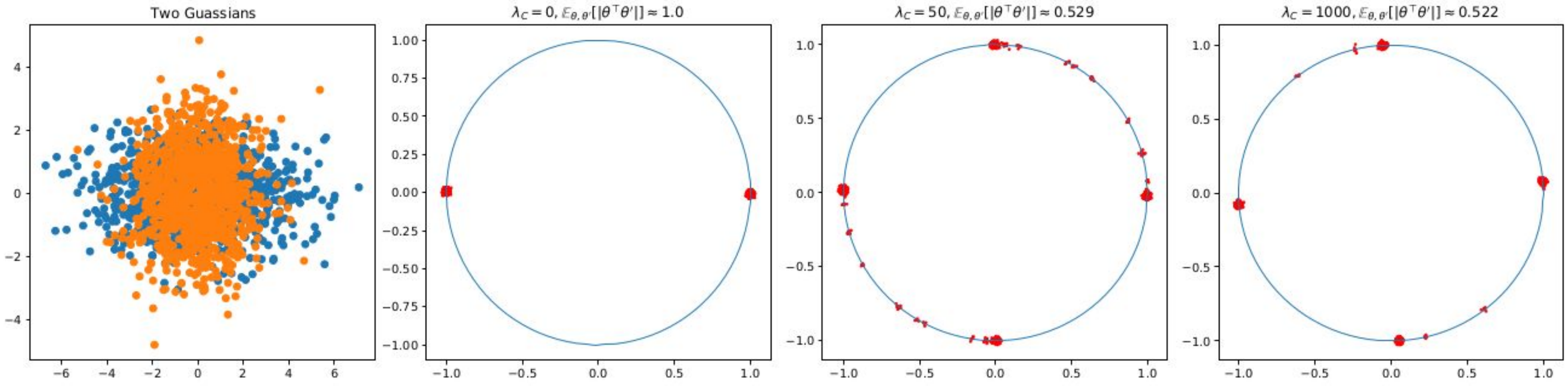
- $\{\theta\}_{i=1}^{L} \sim \mathcal{U}(\mathbb{S}^{d-1})$; L is called the number of projections

After finding the optimal $\phi^*$, we can approximate the value of DSW by:

$$DSW_p(\mu,\nu;C) \approx (\textstyle\sum_{l=1}^{L}[W_p^p(\mathcal{R}I_\mu(\cdot,f_{\phi^*}(\theta_l)),\mathcal{R}I_\nu(\cdot,f_{\phi^*}(\theta_l))])^{1/p}$$

- L is called the number of projections

# Experiments




Approximate the distribution over directions by 1000 samples

- When $\lambda_C = 0$, all samples are the "max" direction
- When $\lambda_C$ is large enough, "best" orthogonal directions are found

# Experiments

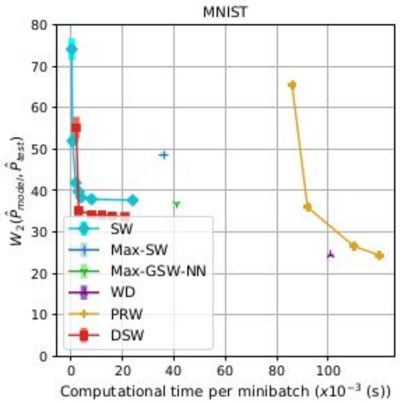
Generative model with minimum expected distance estimator:

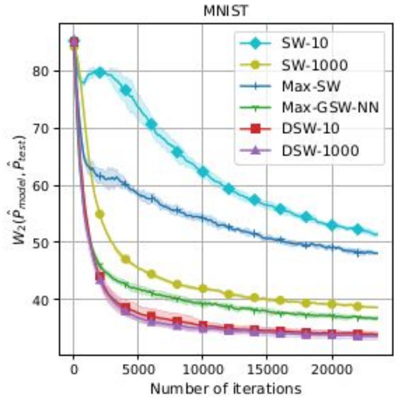
$$\hat{\theta}_{n,m} = \arg\min_{\theta \in \Theta} \mathbb{E}\left[\text{DSW}_p\left(\hat{\mu}_n, \hat{\mu}_{\theta,m}\right) \mid X_{1:n}\right]$$

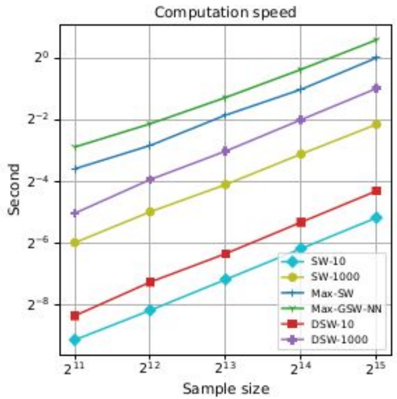
- $\Theta$ is the parameter space
- $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical data distribution
- $\hat{\mu}_{\theta,m} = \frac{1}{m}\sum_{i=1}^m \delta_{Y_i}$ is the empirical distribution that obtained by i.i.d sampling from $\mu_\theta$ which is often $T_\theta \sharp \epsilon$ (e.g. $\epsilon := \mathcal{N}(0, I)$ and $T_\theta$ is a neural net)

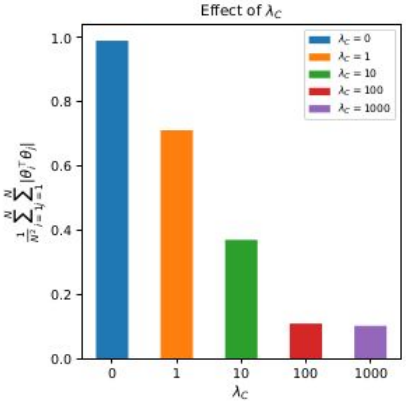
- In practice, we set $m = n =$ size of mini-batches
- The neural net architecture is chosen based on the dataset

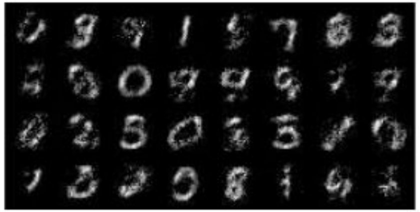
# Experiments



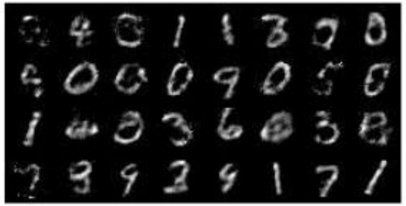



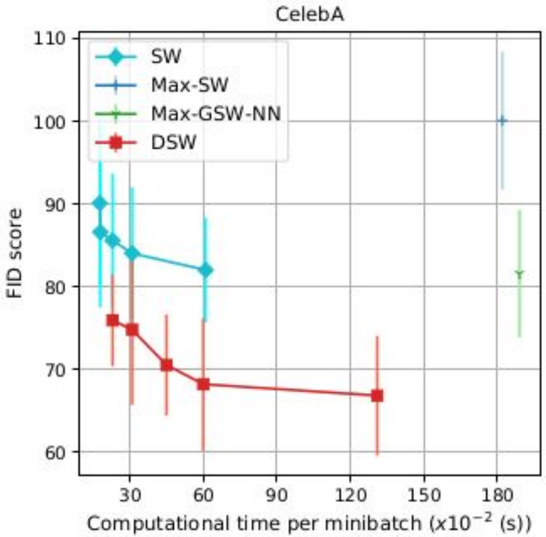


Max-SW

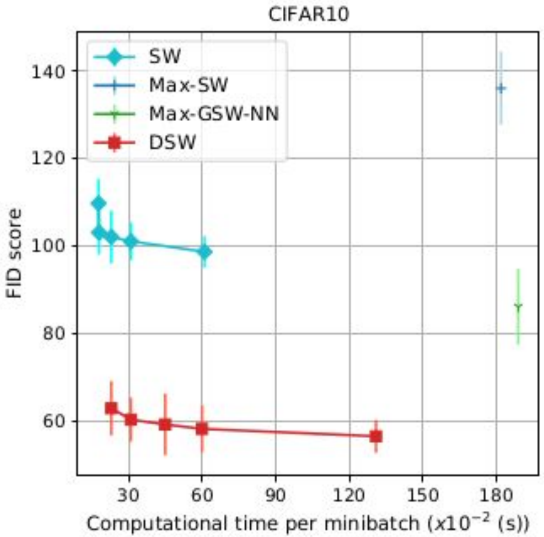

Max-GSW-NN

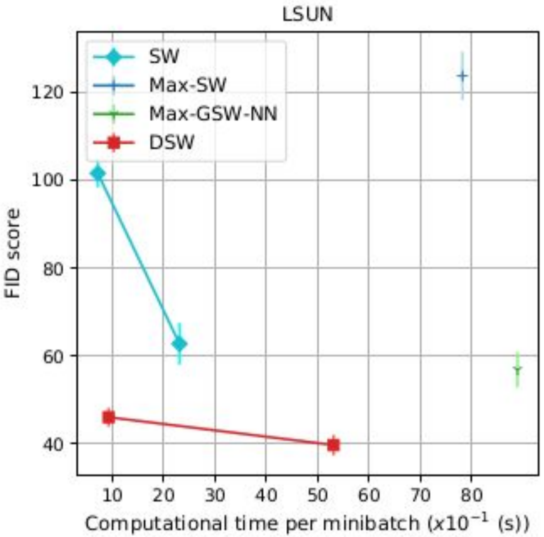

SW L=1000


DSW L=1000

# Experiments




Max-SW

Max-GSW-NN

SW L=10000

DSW L=10000

# Summary

- Introducing a new distance between probability measures - distributional sliced Wasserstein

- Theoretical analysis (metricity, connections to existing sliced optimal transport  distances, curse of dimensionality)

- Extensions with non-linear projecting

- Experimental results on generative modeling task to show the favorable performance of the new distance

Email: v.khainb@vinai.io