Flowing Datasets with Wasserstein over Wasserstein Gradient Flows

Clément Bonet[1]

Joint work with Christophe Vauthier[2] and Anna Korba[1]

[1]ENSAE, CREST, Institut Polytechnique de Paris
[2]Université Paris-Saclay, Laboratoire de Mathématique d'Orsay

BIRS Workshop
01/07/2025

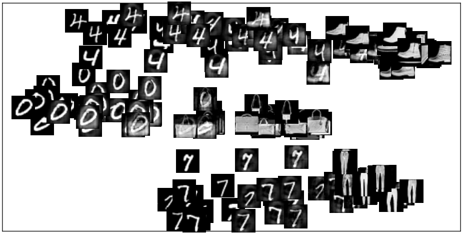
Motivations

Labeled dataset: $\mathcal{D} = \big((x_i, y_i)\big)_{i=1}^{n}$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
Typically: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \ldots, C\}$,

Motivations

Labeled dataset: $\mathcal{D} = \big((x_i, y_i)\big)_{i=1}^{n}$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
Typically: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \dots, C\}$,

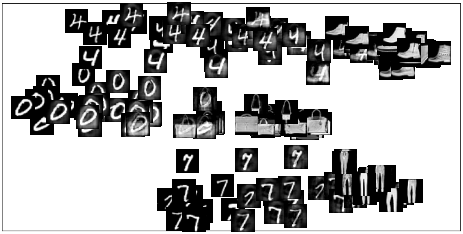
Goal: Generate samples from $\mathcal{D}$ respecting the structure of the dataset



Motivations

Labeled dataset: $\mathcal{D} = \left((x_i, y_i)\right)_{i=1}^{n}$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
Typically: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \ldots, C\}$,

Goal: Generate samples from $\mathcal{D}$ respecting the structure of the dataset



Applications:

- Domain adaptation (Courty et al., 2016)
- Transfer learning (Alvarez-Melis and Fusi, 2021; Hua et al., 2023)
- Dataset distillation (Wang et al., 2018)

OTDD (Alvarez-Melis and Fusi, 2020)

- $\mathcal{D}_1 : \mu_1 = \frac{1}{m}\sum_{i=1}^{m} \delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \ldots, C\})$,
  $\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^{m} \delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \ldots, C\})$
  C: number of classes, n: number of sample in each class, $m = nC$

Question: how to compare datasets $\mathcal{D}_1$ and $\mathcal{D}_2$?

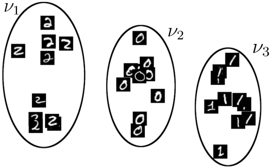
OTDD (Alvarez-Melis and Fusi, 2020)

- $\mathcal{D}_1 : \mu_1 = \frac{1}{m}\sum_{i=1}^m \delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$,
  $\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^m \delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$
  C: number of classes, n: number of sample in each class, $m = nC$

Question: how to compare datasets $\mathcal{D}_1$ and $\mathcal{D}_2$?

Solution of Alvarez-Melis and Fusi (2020):

- Embed a label (a class) in $\mathcal{P}(\mathbb{R}^d)$ as $c \mapsto \nu_c^k = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^k} \mathbb{1}_{\{y_i^k = c\}}$ for $k = 1, 2$



$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m}\sum_{i=1}^m \delta_{(x_i^k, \nu_{y_i^k}^k)} \in \mathcal{P}\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$

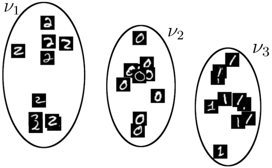
OTDD (Alvarez-Melis and Fusi, 2020)

- $\mathcal{D}_1 : \mu_1 = \frac{1}{m}\sum_{i=1}^m \delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$,
  $\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^m \delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$
  C: number of classes, n: number of sample in each class, $m = nC$

Question: how to compare datasets $\mathcal{D}_1$ and $\mathcal{D}_2$?

Solution of Alvarez-Melis and Fusi (2020):

- Embed a label (a class) in $\mathcal{P}(\mathbb{R}^d)$ as $c \mapsto \nu_c^k = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^k} \mathbb{1}_{\{y_i^k = c\}}$ for $k = 1, 2$



$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m}\sum_{i=1}^m \delta_{(x_i^k, \nu_{y_i^k}^k)} \in \mathcal{P}\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$

- Cost: $d\big((x, y), (x', y')\big)^2 = \|x - x'\|_2^2 + \mathrm{W}_2^2(\nu_y, \nu_{y'})$
- Optimal transport distance: $O\big(C^2 n^3 \log n + n^3 C^3 \log(nC)\big)$

$$\mathrm{OTDD}(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \int d\big((x, y), (x', y')\big)^2 \, \mathrm{d}\gamma\big((x, y), (x', y')\big).$$

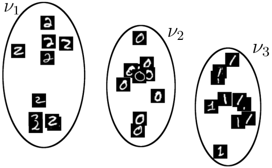
OTDD (Alvarez-Melis and Fusi, 2020)

- $\mathcal{D}_1 : \mu_1 = \frac{1}{m}\sum_{i=1}^m \delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$,
  $\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^m \delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$
  C: number of classes, n: number of sample in each class, $m = nC$

Question: how to compare datasets $\mathcal{D}_1$ and $\mathcal{D}_2$?

Solution of Alvarez-Melis and Fusi (2020):

- Embed a label (a class) in $\mathbb{R}^p \times S_p^{++}(\mathbb{R})$ as $c \mapsto \nu_c^k \approx \mathcal{N}(m_c^k, \Sigma_c^k)$ for $k = 1, 2$



$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m}\sum_{i=1}^m \delta_{(x_i^k, m_{y_i^k}^k, \Sigma_{y_i^k}^k)} \in \mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}^p \times S_p^{++}(\mathbb{R})\big)$

- Cost: $d\big((x,y),(x',y')\big)^2 = \|x - x'\|_2^2 + \mathrm{BW}_2^2(\nu_y, \nu_{y'})$
- Optimal transport distance: approximated in $O\big(C^2 d^3 + n^2 C^2 \log(nC)/\varepsilon^2\big)$

$$\mathrm{OTDD}_\varepsilon(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \int d\big((x,y),(x',y')\big)^2 \, \mathrm{d}\gamma\big((x,y),(x',y')\big) + \varepsilon \mathcal{H}(\gamma).$$

Table of Contents

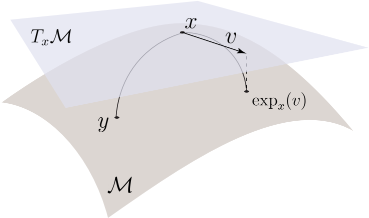
Riemannian Manifolds

Definition

A Riemannian manifold $\mathcal{M}$ of dimension p is a space that behaves locally as a linear space diffeomorphic to $\mathbb{R}^p$.

Properties:

- To any $x \in \mathcal{M}$, associate a tangent space $T_x\mathcal{M}$ with a smooth inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$.
- Geodesic between x and y: shortest path minimizing the length $\mathcal{L}$
- Geodesic distance: $d(x, y) = \inf_\gamma \; \mathcal{L}(\gamma)$
- Exponential map: $\forall x \in \mathcal{M}, \; \exp_x : T_x\mathcal{M} \to \mathcal{M}$, inverse $\log_x : \mathcal{M} \to T_x\mathcal{M}$



For $\mathcal{M} = \mathbb{R}^d$: $d(x, y) = \|x - y\|_2$, $\exp_x(v) = x + v$, $\log_x(y) = y - x$

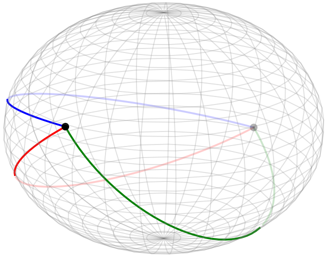
Riemannian Manifolds

Definition

A Riemannian manifold $\mathcal{M}$ of dimension p is a space that behaves locally as a linear space diffeomorphic to $\mathbb{R}^p$.

Properties:

- To any $x \in \mathcal{M}$, associate a tangent space $T_x\mathcal{M}$ with a smooth inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$.
- Geodesic starting between x and y: $\forall t \in [0, 1],\ \gamma(t) = \exp_x\big(t \log_x(y)\big)$
- Geodesic distance: $d(x, y) = \inf_\gamma\ \mathcal{L}(\gamma)$
- Exponential map: $\forall x \in \mathcal{M},\ \exp_x : T_x\mathcal{M} \to \mathcal{M}$, inverse $\log_x : \mathcal{M} \to T_x\mathcal{M}$



Wasserstein Geometry

Let $\mathcal{M}$ be a (compact connected) Riemannian manifold, $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ the geodesic distance.

Definition (Wasserstein distance)

Let $\mu, \nu \in \mathcal{P}_2(\mathcal{M})$ and denote by $\Pi(\mu, \nu)$ the set of coupling between μ, ν. Then, the Wasserstein distance is

$$\mathrm{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int d(x, y)^2 \, \mathrm{d}\gamma(x, y).$$

Properties:

- W_2 distance, $(\mathcal{P}_2(\mathcal{M}), \mathrm{W}_2)$: Wasserstein space
- **Riemannian structure**

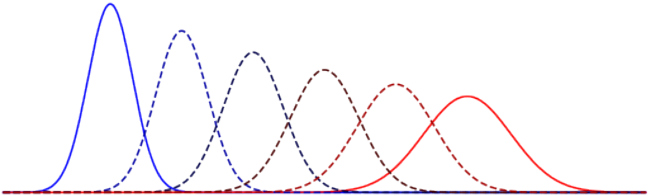
Riemannian Structure of the Wasserstein Space

Let $T\mathcal{M} = \{(x,v),\ x \in \mathcal{M},\ v \in T_x\mathcal{M}\}$, $\pi^{\mathcal{M}}((x,v)) = x$, $\pi^v((x,v)) = v$.

$$\exp_\mu^{-1}(\nu) = \left\{\gamma \in \mathcal{P}_2(T\mathcal{M}),\ \pi^{\mathcal{M}}_\#\gamma = \mu,\ \exp_\#\gamma = \nu,\ \int \|v\|_x^2 \mathrm{d}\gamma(x,v) = \mathrm{W}_2^2(\mu,\nu)\right\}$$

- Geodesics between $\mu, \nu \in \mathcal{P}_2(\mathcal{M})$,
  - If log defined μ-a.e.: $\forall t \in [0,1],\ \mu_t = \left(\exp_{\pi^1}(t \log_{\pi^1} \circ \pi^2)\right)_\# \tilde{\gamma}$, $\tilde{\gamma} \in \Pi_o(\mu,\nu)$
  - In general: $\forall t \in [0,1],\ \mu_t = \left(\exp_{\pi^{\mathcal{M}}} \circ (t\pi^v)\right)_\# \gamma$, $\gamma \in \exp_\mu^{-1}(\nu)$ (Gigli, 2011)

    $\rightarrow$ **precise which geodesic was chosen to move the mass**



For $\mathcal{M} = \mathbb{R}^d$:

- In general: $\mu_t = \left((1-t)\pi^1 + t\pi^2\right)_\#\gamma = \left(\pi^1 + t(\pi^2 - \pi^1)\right)_\#\gamma$, $\gamma \in \Pi_o(\mu,\nu)$

Riemannian Structure of the Wasserstein Space

Let $T\mathcal{M} = \{(x,v),\ x\in\mathcal{M},\ v\in T_x\mathcal{M}\}$, $\pi^{\mathcal{M}}((x,v)) = x$, $\pi^v((x,v)) = v$.

$$\exp_\mu^{-1}(\nu) = \left\{\gamma\in\mathcal{P}_2(T\mathcal{M}),\ \pi^{\mathcal{M}}_\#\gamma = \mu,\ \exp_\#\gamma = \nu,\ \int \|v\|_x^2 \mathrm{d}\gamma(x,v) = \mathrm{W}_2^2(\mu,\nu)\right\}$$

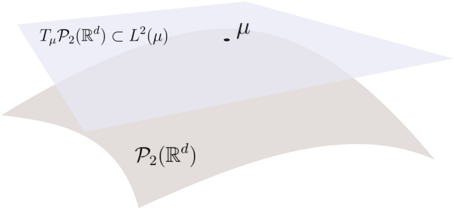
- Geodesics between $\mu,\nu\in\mathcal{P}_2(\mathcal{M})$,
  - If $\log$ defined μ-a.e.: $\forall t\in[0,1],\ \mu_t = \big(\exp_{\pi^1}(t\log_{\pi^1}\circ\pi^2)\big)_\#\tilde{\gamma}$, $\tilde{\gamma}\in\Pi_o(\mu,\nu)$
  - In general: $\forall t\in[0,1],\ \mu_t = \big(\exp_{\pi^{\mathcal{M}}}\circ(t\pi^v)\big)_\#\gamma$, $\gamma\in\exp_\mu^{-1}(\nu)$ (Gigli, 2011)

    $\rightarrow$ **precise which geodesic was chosen to move the mass**

- Tangent space at $\mu\in\mathcal{P}_{2,\mathrm{ac}}(\mathcal{M})$ (Ambrosio et al., 2008; Erbar, 2010):

$$T_\mu\mathcal{P}_2(\mathcal{M}) = \overline{\{\nabla\psi,\ \psi\in C_c^\infty(\mathcal{M})\}} \subset L^2(\mu, T\mathcal{M}),$$

where $L^2(\mu,T\mathcal{M}) = \{f\in\mathcal{M}\to T\mathcal{M},\ \int\|f(x)\|_2^2\ \mathrm{d}\mu(x) < \infty\}$.



Wasserstein Gradient (Ambrosio et al., 2008; Erbar, 2010)

Definition (Wasserstein gradient)

Let $\mu \in \mathcal{P}_2(\mathcal{M})$. $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) \in L^2(\mu, T\mathcal{M})$ is a Wasserstein gradient of $\mathcal{F}$ at μ if for any $\nu \in \mathcal{P}_2(\mathcal{M})$ and any $\gamma \in \exp_\mu^{-1}(\nu)$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), v\rangle_x \, \mathrm{d}\gamma(x,v) + o\big(\mathrm{W}_2(\mu,\nu)\big).$$

If such a gradient exists, then we say that $\mathcal{F}$ is W_2-differentiable at μ.

Wasserstein Gradient (Ambrosio et al., 2008; Erbar, 2010)

Definition (Wasserstein gradient)

Let $\mu \in \mathcal{P}_2(\mathcal{M})$. $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) \in L^2(\mu, T\mathcal{M})$ is a Wasserstein gradient of $\mathcal{F}$ at μ if for any $\nu \in \mathcal{P}_2(\mathcal{M})$ and any $\gamma \in \exp_\mu^{-1}(\nu)$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), v\rangle_x \, \mathrm{d}\gamma(x,v) + o\big(\mathrm{W}_2(\mu,\nu)\big).$$

If such a gradient exists, then we say that $\mathcal{F}$ is W_2-differentiable at μ.

Properties:

- There is a unique gradient in $T_\mu\mathcal{P}_2(\mathcal{M})$
- Differential are strong, *i.e.* for any $\gamma \in \mathcal{P}(T\mathcal{M})$ s.t. $\pi^{\mathcal{M}}_\#\gamma = \mu$, $\exp_\#\gamma = \nu$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), v\rangle_x \, \mathrm{d}\gamma(x,v) + o\left(\sqrt{\int \|v\|_x^2 \, \mathrm{d}\gamma(x,v)}\right)$$

In particular, for $\gamma = (\mathrm{Id}, \exp\circ \mathrm{T})_\#\mu$,

$$\mathcal{F}\big((\exp\circ\mathrm{T})_\#\mu\big) = \mathcal{F}(\mu) + \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu), \mathrm{T}\rangle_{L^2(\mu,T\mathcal{M})} + o(\|\mathrm{T}\|_{L^2(\mu,T\mathcal{M})})$$

Wasserstein Gradient

Example of functionals

- Potential energies $\mathcal{V}(\mu) = \int V \mathrm{d}\mu$: For V differentiable and smooth,

$$\nabla_{\mathrm{W}_2}\mathcal{V}(\mu) = \nabla V$$

- Interaction energies $\mathcal{W}(\mu) = \iint W(x,y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$: For W differentiable and smooth,

$$\nabla_{\mathrm{W}_2}\mathcal{W}(\mu)(x) = \int \big(\nabla_1 W(x,\cdot) + \nabla_2 W(\cdot,x)\big)\ \mathrm{d}\mu$$

Wasserstein Gradient

Example of functionals

- Potential energies $\mathcal{V}(\mu) = \int V \mathrm{d}\mu$: For V differentiable and smooth,

$$\nabla_{\mathrm{W}_2}\mathcal{V}(\mu) = \nabla V$$

- Interaction energies $\mathcal{W}(\mu) = \iint W(x,y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$: For W differentiable and smooth,

$$\nabla_{\mathrm{W}_2}\mathcal{W}(\mu)(x) = \int \big(\nabla_1 W(x,\cdot) + \nabla_2 W(\cdot,x)\big)\ \mathrm{d}\mu$$

Example of discrepancy: **Maximum Mean Discrepancy** (MMD) (Arbel et al., 2019)

$$\begin{aligned}\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}_k^2(\mu,\nu) &= \iint k(x,y)\ \mathrm{d}(\mu-\nu)(x)\mathrm{d}(\mu-\nu)(y) \\ &= \mathcal{V}(\mu) + \mathcal{W}(\mu) + \mathrm{cst},\end{aligned}$$

with k positive definite kernel, and:

$$\mathcal{V}(\mu) = \int V \mathrm{d}\mu, \quad V(x) = -\int k(x,y)\mathrm{d}\nu(y), \quad \mathcal{W}(\mu) = \frac{1}{2}\iint k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y)$$

Wasserstein Gradient Descent

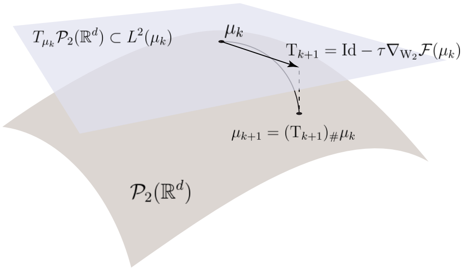
Time discretization of the flow (Riemannian Wasserstein Gradient Descent):

$$\forall k \geq 0,\ \mu_{k+1} = \exp_{\mu_k}\big(-\tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big) = \big(\exp_{\mathrm{Id}}(-\tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k))\big)_{\#}\mu_k$$

Particle approximation: $\mu_k^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^k}$,

$$\forall i \in \{1,\dots,n\},\ x_i^{k+1} = \exp_{x_i^k}\big(-\tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k^n)(x_i^k)\big)$$

On $\mathbb{R}^d$: $x_i^{k+1} = x_i^k - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k^n)(x_i^k)$



Flowing Datasets (previous works)

Let $\mu, \nu \in \mathcal{P}_2\big(\mathbb{R}^d \times \mathbb{R}^p \times S_p^{++}(\mathbb{R})\big)$, $p \leq d$.

Goal: $\min_\mu \ \mathcal{F}(\mu)$

Choice of $\mathcal{F}$:

- (Alvarez-Melis and Fusi, 2021): $\mathcal{F}(\mu) := \mathrm{OTDD}(\mu, \nu)$
- (Hua et al., 2023): $\mathcal{F}(\mu) := \frac{1}{2}\mathrm{MMD}_k^2(\mu, \nu)$ with kernel

$$k\big((x, m, \Sigma), (x', m', \Sigma')\big) = e^{-\|x-x'\|_2^2/h_x} e^{-\|m-m'\|_2^2/h_m} e^{-\|\Sigma-\Sigma'\|_2^2/h_\Sigma}$$

Several strategies:

- Wasserstein gradient flow on features + update the C Gaussian
- Wasserstein gradient flow on $\mathbb{R}^d \times \mathbb{R}^p \times S_p^{++}(\mathbb{R})$, *i.e.*,

$$\mu_{k+1} = \exp_{\mu_k}\big(-\tau \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big),$$

  where $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big((x, m, \Sigma)\big) \in \mathbb{R}^d \times \mathbb{R}^p \times S_p(\mathbb{R})$.

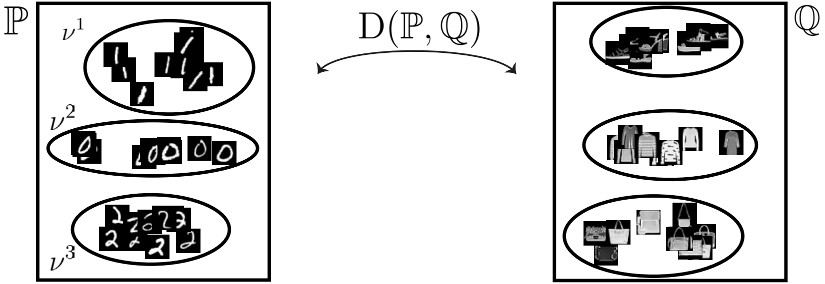
Drawbacks:

- OTDD costly + non differentiable (require entropic approximation)
- Both require lots of hyperparameters to tune

Contributions

- Model datasets as $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\nu^c} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ where $\nu^c = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^c}$
- Flow a dataset $\mathbb{P}$ towards $\mathbb{Q}$ by minimizing a discrepancy D on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$

$\rightarrow$ minimization problem on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$



Example

$\mathrm{D}(\mathbb{P},\mathbb{Q}) = \mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})$ with $K : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}$

Table of Contents

Wasserstein over Wasserstein Distance (WoW)

Definition (WoW distance)

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$ and denote by $\Pi(\mathbb{P}, \mathbb{Q})$ the set of coupling between $\mathbb{P}, \mathbb{Q}$. Then, the WoW distance is

$$\mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q}) = \inf_{\Gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \int \mathrm{W}_2^2(\mu, \nu)\ \mathrm{d}\Gamma(\mu, \nu).$$

Properties:

- $\mathrm{W}_{\mathrm{W}_2}$ distance, $\big(\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big), \mathrm{W}_{\mathrm{W}_2}\big)$: WoW space
- Brenier-McCann's theorem: Let $\mathbb{P}_0$ a reference measure satisfying suitable assumptions (no atom, satisfies an IPP, see (Dello Schiavo, 2020)).
  If $\mathbb{P} \ll \mathbb{P}_0$, then there exists a unique T s.t. $\mathrm{T}_\# \mathbb{P} = \mathbb{Q}$ (Emami and Pass, 2025).
- **Riemannian structure**

Geodesics

On $\mathcal{P}_2(\mathcal{M})$: $\mu_t = \big((1-t)\pi^1 + t\pi^2\big)_{\#}\gamma$, $\gamma \in \Pi_o(\mu,\nu)$
$\rightarrow$ use $\exp^{-1}$

Geodesics

On $\mathcal{P}_2(\mathcal{M})$: $\mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\#\gamma$, $\gamma \in \Pi_o(\mu,\nu)$
$\rightarrow$ use $\exp^{-1}$

Let $\gamma \in \mathcal{P}_2(T\mathcal{M})$. Define $\varphi^{\mathcal{M}}(\gamma) = \pi^{\mathcal{M}}_\#\gamma$, $\varphi^{\exp}(\gamma) = \exp_\#\gamma$ and $\varphi^v(\gamma) = \pi^v_\#\gamma$.

For any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$,

$$\exp_{\mathbb{P}}^{-1}(\mathbb{Q}) = \Big\{\mathbb{\Gamma} \in \mathcal{P}_2\big(\mathcal{P}_2(T\mathcal{M})\big),\ \varphi^{\mathcal{M}}_\#\mathbb{\Gamma} = \mathbb{P},\ \varphi^{\exp}_\#\mathbb{\Gamma} = \mathbb{Q},\ \iint \|v\|_x^2 \mathrm{d}\gamma(x,v)\mathrm{d}\mathbb{\Gamma}(\gamma) = \mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P},\mathbb{Q})\Big\}.$$

Properties

- $\mathbb{\Gamma} \mapsto (\varphi^{\mathcal{M}}, \varphi^{\exp})_\#\mathbb{\Gamma}$ is a surjective map from $\exp_{\mathbb{P}}^{-1}(\mathbb{Q})$ to $\Pi_o(\mathbb{P},\mathbb{Q})$
- Geodesic between $\mathbb{P}$ and $\mathbb{Q}$: $\forall t \in [0,1]$, $\mathbb{P}_t = \big(\exp_{\varphi^{\mathcal{M}}} \circ (t\varphi^v)\big)_\#\mathbb{\Gamma}$ with $\mathbb{\Gamma} \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$

Tangent Space

Definition (Cylinder (von Renesse and Sturm, 2009))

$\mathcal{F} : \mathcal{P}_2(\mathcal{M}) \to \mathbb{R} \in \mathrm{Cyl}\big(\mathcal{P}_2(\mathcal{M})\big)$ is a cylinder if there exists $k \geq 0$, $F \in C_c^\infty(\mathbb{R}^k)$ and $V_1, \dots, V_k \in C_c^\infty(\mathcal{M})$ such that, for all $\mu \in \mathcal{P}_2(\mathcal{M})$,

$$\mathcal{F}(\mu) = F\left(\int V_1 \mathrm{d}\mu, \dots, \int V_k \mathrm{d}\mu\right).$$

Tangent space at $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$:

$$T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big) = \overline{\{\nabla_{\mathrm{W}_2}\varphi,\ \varphi \in \mathrm{Cyl}\big(\mathcal{P}_2(\mathcal{M})\big)\}}^{L^2(\mathbb{P})}.$$

Let $(\mathbb{P}_t)_{t\in I}$ be an absolutely continuous curve on $\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. Then, for a.e. $t \in I$, there exists $v_t \in T_{\mathbb{P}_t}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$ such that $\|v_t\|_{L^2(\mathbb{P}_t, T\mathcal{P}_2(\mathcal{M}))} \leq |\mathbb{P}'|(t)$ and for all $\varphi \in \mathrm{Cyl}(I \times \mathcal{P}_2(\mathcal{M}))$,

$$\iint \big(\partial_t \varphi_t(\mu) + \langle \nabla_{\mathrm{W}_2}\varphi_t(\mu), v_t(\mu)\rangle_{L^2(\mu)}\big)\, \mathrm{d}\mathbb{P}_t(\mu)\mathrm{d}t = 0.$$

WoW Gradient

Definition (WoW gradient)

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}) \in L^2(\mathbb{P}, T\mathcal{P}_2(\mathcal{M}))$ is a WoW gradient of $\mathbb{F}$ at $\mathbb{P}$ if for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$ and any $\mathbb{\Gamma} \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P})(\pi^{\mathcal{M}}_{\#}\gamma)(x), v\rangle_x \, \mathrm{d}\gamma(x,v)\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P},\mathbb{Q})\big).$$

If such a gradient exists, then we say that $\mathbb{F}$ is $\mathrm{W}_{\mathrm{W}_2}$-differentiable at $\mathbb{P}$.

WoW Gradient

Definition (WoW gradient)

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. $\nabla_{\mathrm{W_{W_2}}}\mathbb{F}(\mathbb{P}) \in L^2(\mathbb{P}, T\mathcal{P}_2(\mathcal{M}))$ is a WoW gradient of $\mathbb{F}$ at $\mathbb{P}$ if for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$ and any $\Gamma \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W_{W_2}}}\mathbb{F}(\mathbb{P})(\pi^{\mathcal{M}}_{\#}\gamma)(x), v\rangle_x \, \mathrm{d}\gamma(x,v)\Gamma(\gamma) + o\big(\mathrm{W_{W_2}}(\mathbb{P},\mathbb{Q})\big).$$

If such a gradient exists, then we say that $\mathbb{F}$ is $\mathrm{W_{W_2}}$-differentiable at $\mathbb{P}$.

Properties:

- There is at most one element in $\partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$
- Under assumptions on $\mathbb{P}$ and $\mathcal{M}$, existence of $\xi \in \partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$
- If $\xi \in \partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. Then ξ is a strong differential of $\mathbb{F}$ at $\mathbb{P}$, *i.e.*, for $\Gamma = \mathcal{P}_2\big(\mathcal{P}_2(T\mathcal{M})\big)$ s.t. $\phi^{\mathcal{M}}_{\#}\Gamma = \mathbb{P}$, $\phi^{\exp}_{\#}\Gamma := \mathbb{Q}$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \int \langle \xi(\pi^{\mathcal{M}}_{\#}\gamma)(x), v\rangle_x \mathrm{d}\gamma(x,v)\mathrm{d}\Gamma(\gamma) + o\left(\sqrt{\iint \|v\|_x^2 \, \mathrm{d}\gamma(x,v)\mathrm{d}\Gamma(\gamma)}\right).$$

WoW Gradient

Definition (WoW gradient)

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}) \in L^2(\mathbb{P}, T\mathcal{P}_2(\mathcal{M}))$ is a WoW gradient of $\mathbb{F}$ at $\mathbb{P}$ if for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$ and any $\Gamma \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P})(\pi^{\mathcal{M}}_{\#}\gamma)(x), v\rangle_x \ \mathrm{d}\gamma(x,v)\Gamma(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P},\mathbb{Q})\big).$$

If such a gradient exists, then we say that $\mathbb{F}$ is $\mathrm{W}_{\mathrm{W}_2}$-differentiable at $\mathbb{P}$.

Example of functionals

- Potential energies $\mathbb{V}(\mathbb{P}) = \int \mathcal{F}(\mu)\mathrm{d}\mathbb{P}(\mu)$: For $\mathcal{F} : \mathcal{P}_2(\mathcal{M}) \to \mathbb{R}$ differentiable and smooth,

$$\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{V}(\mathbb{P}) = \nabla_{\mathrm{W}_2}\mathcal{F}$$

- Interaction energies $\mathbb{W}(\mathbb{P}) = \iint \mathcal{W}(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu)$: For $\mathcal{W}$ differentiable and smooth,

$$\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{W}(\mathbb{P})(\mu) = \int \big(\nabla_1\mathcal{W}(\mu,\cdot) + \nabla_2\mathcal{W}(\cdot,\mu)\big)\ \mathrm{d}\mathbb{P}$$

WoW Gradient Descent

Forward scheme:

$$\forall k \geq 0,\ \mathbb{P}_{k+1} = \exp_{\mathbb{P}_k}\big(-\tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)\big)$$

WoW Gradient Descent

Forward scheme:

$$\forall k \geq 0,\ \mathbb{P}_{k+1} = \exp_{\mathbb{P}_k}\big(-\tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)\big)$$

In practice: For $\mathbb{P}_k = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu_k^{c,n}}$ with $\mu_k^{c,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{i,k}^c} \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\forall k \geq 0,\ \text{particle (image) } i,\ \text{class } c,\ x_{i,k+1}^c = x_{i,k}^c - \tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)(\mu_k^{c,n})(x_{i,k}^c).$$

$\mathbb{P}_k$: inter-class interaction, $\mu_k^{c,n}$: intra-class interaction, $x_{i,k}^c$ image

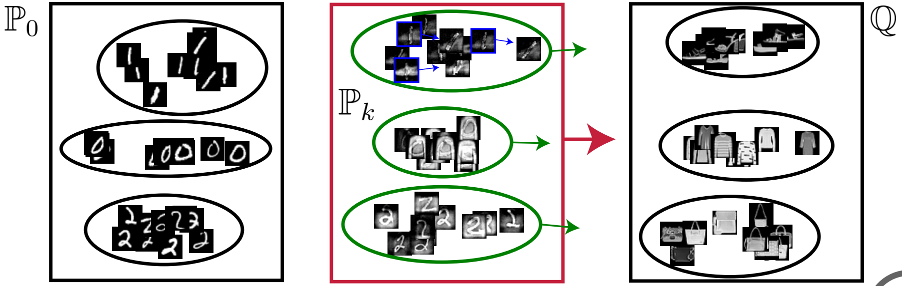


Table of Contents

Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

where $\begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu,\nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 L n \log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

Sliced-Wasserstein distance (Rabin et al., 2011; Bonneel et al., 2015):

$$\mathrm{SW}_2^2(\mu,\nu) = \int_{S^{d-1}} \mathrm{W}_2^2(P_\#^\theta \mu, P_\#^\theta \nu)\mathrm{d}\sigma(\theta) \approx \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{W}_2^2(P_\#^{\theta_\ell}\mu, P_\#^{\theta_\ell}\nu),$$

with $S^{d-1} = \{\theta \in \mathbb{R}^d,\ \|\theta\|_2 = 1\}$, $P^\theta(x) = \langle x, \theta\rangle$, $\theta_1, \dots, \theta_L \sim \sigma = \mathrm{Unif}(S^{d-1})$.

- Gaussian SW kernel: $K(\mu,\nu) = e^{-\mathrm{SW}_2^2(\mu,\nu)/h}$ (Kolouri et al., 2016)
- Riesz SW kernel: $K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu)$

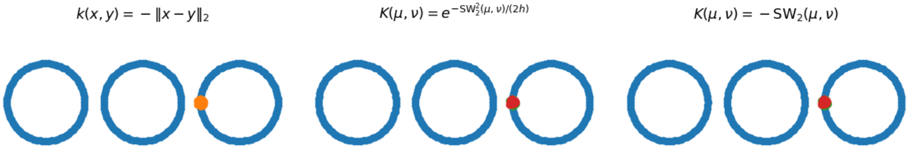
Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

where $\begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu, \nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu, \nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 Ln \log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Goal: $\min_{\mathbb{P}}\ \mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q})$, where $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^{c,n}}$, $\nu^{c,n}$ ring.



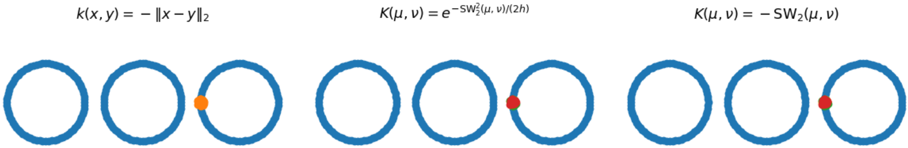
Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

where $\begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu,\nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 Ln \log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

Goal: $\min_{\mathbb{P}}\ \mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})$, where $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3}\delta_{\nu^{c,n}}$, $\nu^{c,n}$ ring.



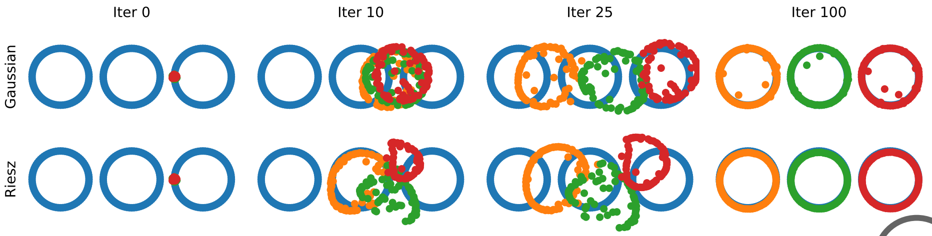
Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

$$\text{where} \quad \begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu,\nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 L n \log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

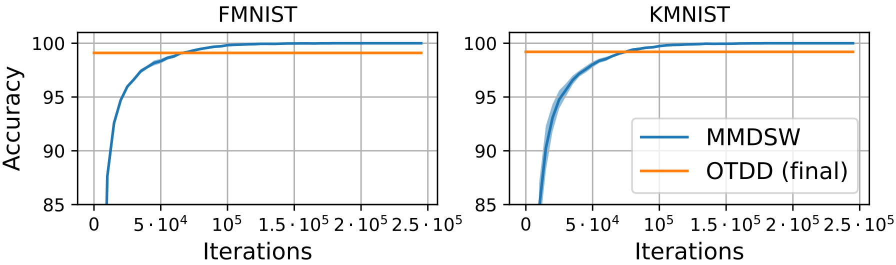
Goal: $\min_{\mathbb{P}}\ \mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q})$, where $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^{c,n}}$, $\nu^{c,n}$ ring.



"Domain Adaptation"

Setting:

1. Pretrain a classifier on $\mathbb{Q}$ =MNIST
2. Flow starting from $\mathbb{P}_0$ =Fashion MNIST (**Left**) or from $\mathbb{P}_0$ =KMNIST (**Right**) by minimizing $\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})$ with $K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu)$
3. Measure accuracy on $\mathbb{P}_t$ (flowed data)

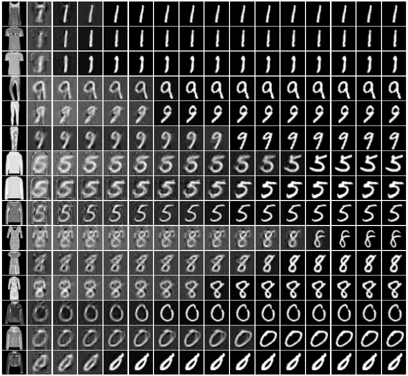


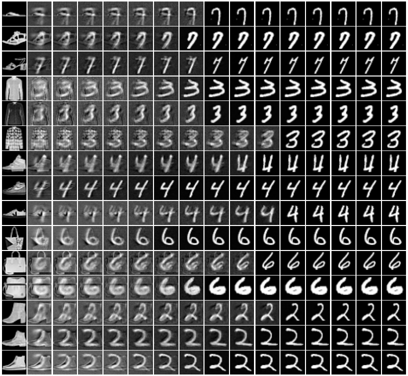
→ reach 100% accuracy

"Domain Adaptation"

Setting:

1. Pretrain a classifier on $\mathbb{Q}$ =MNIST
2. Flow starting from $\mathbb{P}_0$ =Fashion MNIST (**Left**) or from $\mathbb{P}_0$ =KMNIST (**Right**) by minimizing $\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q})$ with $K(\mu, \nu) = -\mathrm{SW}_2(\mu, \nu)$
3. Measure accuracy on $\mathbb{P}_t$ (flowed data)





Dataset distillation: synthesize a big dataset $\mathbb{Q} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\nu_c^n}$ with a small dataset $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu_c^k}$, k small

Transfer learning: augment a small dataset $\mathbb{Q} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\nu_c^k}$ with k small

Dataset distillation

Dataset	k	$\psi^\theta = \mathcal{A}^\omega = \mathrm{Id}$		Baselines	
		DM	MMDSW	Random	Full data
MNIST	1	$61.1_{\pm 6.5}$	$\mathbf{66.5}_{\pm 5.5}$	$55.8_{\pm 2.0}$	99.4
	10	$88.2_{\pm 2.8}$	$\mathbf{93.2}_{\pm 0.7}$	$92.2_{\pm 1.1}$	
	50	$95.9_{\pm 0.9}$	$97.0_{\pm 0.2}$	$97.6_{\pm 0.2}$	
FMNIST	1	$54.4_{\pm 3.2}$	$\mathbf{60.0}_{\pm 4.1}$	$49.0_{\pm 7.5}$	92.4
	10	$74.6_{\pm 1.0}$	$\mathbf{76.7}_{\pm 1.0}$	$75.3_{\pm 0.7}$	
	50	$81.3_{\pm 0.5}$	$\mathbf{84.2}_{\pm 0.1}$	$83.2_{\pm 0.2}$	

Transfer learning

Dataset	k	Train on $\mathbb{Q}$	MMDSW	OTDD	(Hua et al., 2023)
M to F	1	$26.0_{\pm 5.3}$	$\mathbf{40.5}_{\pm 4.7}$	$30.5_{\pm 4.2}$	$36.4_{\pm 3.3}$
	5	$38.5_{\pm 6.7}$	$61.5_{\pm 4.6}$	$59.7_{\pm 1.8}$	$\mathbf{62.7}_{\pm 1.1}$
	10	$53.9_{\pm 7.9}$	$65.4_{\pm 1.5}$	$64.0_{\pm 1.4}$	$\mathbf{66.2}_{\pm 1.0}$
	100	$71.1_{\pm 1.5}$	$\mathbf{74.7}_{\pm 0.8}$	-	$73.5_{\pm 0.7}$
M to K	1	$18.4_{\pm 3.1}$	$\mathbf{20.9}_{\pm 2.0}$	$18.8_{\pm 2.1}$	$19.4_{\pm 1.9}$
	5	$25.9_{\pm 4.0}$	$37.4_{\pm 2.2}$	$31.3_{\pm 1.4}$	$\mathbf{39.0}_{\pm 1.0}$
	10	$30.9_{\pm 4.6}$	$\mathbf{44.7}_{\pm 1.8}$	$34.1_{\pm 0.9}$	$44.1_{\pm 1.2}$
	100	$60.1_{\pm 1.1}$	$\mathbf{66.8}_{\pm 0.8}$	$66.3_{\pm 0.9}$	$62.4_{\pm 1.2}$

Conclusion

Conclusion:

- Differential structure over the Wasserstein over Wasserstein Space
- Wasserstein over Wasserstein Gradient Flows
- Implementation on the MMD
- Application to image datasets (Dataset distillation, Transfer learning...)

Perspectives:

- Use other positive definite kernels for the MMD (Bachoc et al., 2023; Kachaiev and Recanatesi, 2024)
- Minimize other functionals (Catalano and Lavenant, 2024)
- Theoretical convergence

Conclusion

Conclusion:

- Differential structure over the Wasserstein over Wasserstein Space
- Wasserstein over Wasserstein Gradient Flows
- Implementation on the MMD
- Application to image datasets (Dataset distillation, Transfer learning...)

Perspectives:

- Use other positive definite kernels for the MMD (Bachoc et al., 2023; Kachaiev and Recanatesi, 2024)
- Minimize other functionals (Catalano and Lavenant, 2024)
- Theoretical convergence

Thank you for your attention!

David Alvarez-Melis and Nicolo Fusi. Geometric Dataset Distances via Optimal Transport. *Advances in Neural Information Processing Systems*, 33: 21428–21439, 2020.

David Alvarez-Melis and Nicolò Fusi. Dataset Dynamics via Gradient Flows in Probability Space. In *International conference on machine learning*, pages 219–230. PMLR, 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum Mean Discrepancy Gradient Flow. *Advances in Neural Information Processing Systems*, 32, 2019.

François Bachoc, Louis Béthune, Alberto Gonzalez-Sanz, and Jean-Michel Loubes. Gaussian Processes on Distributions based on Regularized Optimal Transport. In *International Conference on Artificial Intelligence and Statistics*, pages 4986–5010. PMLR, 2023.

References II

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Marta Catalano and Hugo Lavenant. Hierarchical Integral Probability Metrics: A distance on random probability measures with low sample complexity. *arXiv preprint arXiv:2402.00423*, 2024.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Lorenzo Dello Schiavo. A Rademacher-type theorem on L2-Wasserstein spaces over closed Riemannian manifolds. *Journal of Functional Analysis*, 278(6): 108397, 2020.

Pedram Emami and Brendan Pass. Optimal transport with optimal transport cost: the Monge–Kantorovich problem on Wasserstein spaces. *Calculus of Variations and Partial Differential Equations*, 64(2):43, 2025.

Matthias Erbar. The heat equation on manifolds as a gradient flow in the wasserstein space. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pages 1–23, 2010.

Nicola Gigli. On the inverse implication of Brenier-McCann theorems and the structure of $(P_2(M), W_2)$. *Methods and Applications of Analysis*, 18(2): 127–158, 2011.

Xinru Hua, Truyen Nguyen, Tam Le, Jose Blanchet, and Viet Anh Nguyen. Dynamic Flows on Curved Space Generated by Labeled Data. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3803–3811. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.

Oleksii Kachaiev and Stefano Recanatesi. Learning to Embed Distributions via Maximum Kernel Entropy. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein Kernels for Probability Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011.

Max-K von Renesse and Karl-Theodor Sturm. Entropic measure and wasserstein diffusion. 2009.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset Distillation. *arXiv preprint arXiv:1811.10959*, 2018.