Flowing Datasets with Wasserstein over Wasserstein Gradient Flows

Clément Bonet[1]

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

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

GT CalVa
10/02/2025

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: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \ldots, C\}$

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

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

Applications:

- Domain adaptation (Courty et al., 2016)
- Transfer learning (Alvarez-Melis and Fusi, 2021; Hua et al., 2023)
- Dataset distillation (Wang et al., 2018)
- Conditional generative modeling (Chemseddine et al., 2024)

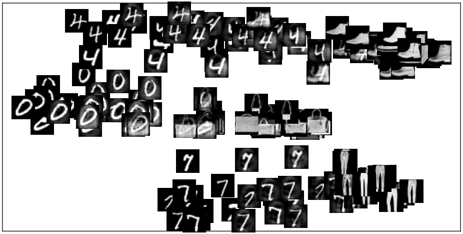


Table of Contents

OTDD (Alvarez-Melis and Fusi, 2020)

- $\mathcal{D}_1 : \mu_1 = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$
- $\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^m \delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C'\})$
- A priori: no relation between labels of $\mathcal{D}_1$ and $\mathcal{D}_2$

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

OTDD (Alvarez-Melis and Fusi, 2020)

- $\mathcal{D}_1 : \mu_1 = \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1,\dots,C\})$
- $\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^{m}\delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1,\dots,C'\})$
- A priori: no relation between labels of $\mathcal{D}_1$ and $\mathcal{D}_2$

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

Solution of Alvarez-Melis and Fusi (2020):

- Embed labels in $\mathcal{P}(\mathbb{R}^d)$ as $c \mapsto \nu_c^k = \frac{1}{n_c}\sum_{i=1}^{n}\delta_{x_i^k}\mathbb{1}_{\{y_i^k = c\}}$
- Cost: $d\big((x,y),(x',y')\big)^2 = \|x - x'\|_2^2 + \mathrm{W}_2^2(\nu_y, \nu_{y'})$
- Optimal transport distance:

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

To reduce computational burden $\rightarrow \nu_y \approx \mathcal{N}(m_y, \Sigma_y)$

Alternatives to OTDD

- MMD on $\mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}^2 \times S_2^{++}(\mathbb{R})\big)$ (Hua et al., 2023)

Alternatives to OTDD

- MMD on $\mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^2 \times S_2^{++}(\mathbb{R})\right)$ (Hua et al., 2023)
- Wasserstein task embedding (Liu et al., 2025)

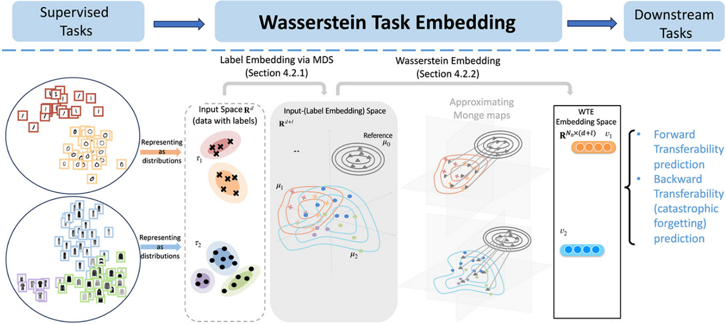
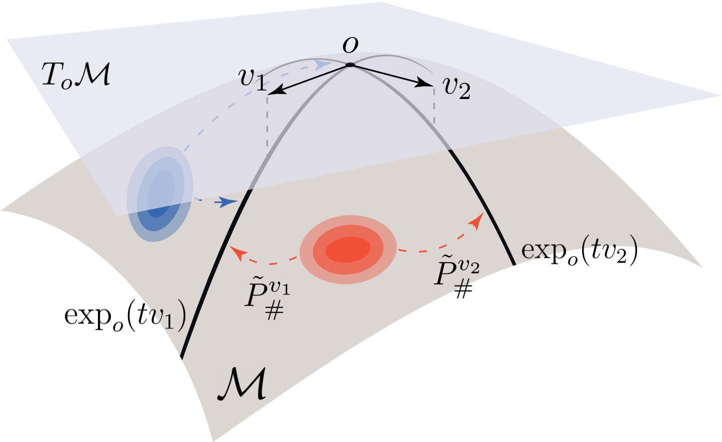


Figure: Taken from https://www.vanderbilt.edu/valiant/2024/11/21/wasserstein-task-embedding-for-measuring-task-similarities/

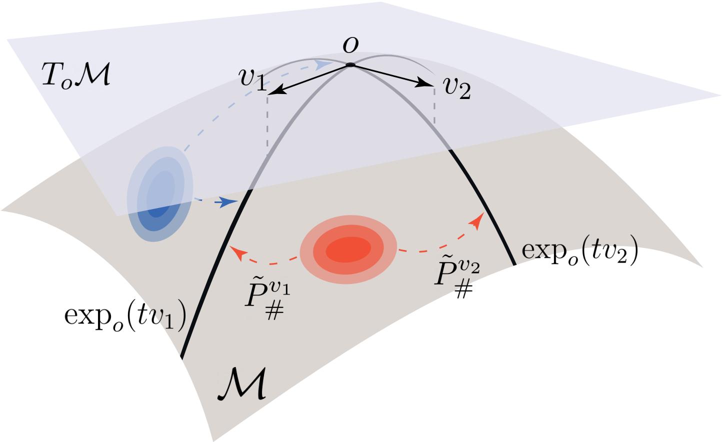
Alternatives to OTDD

- MMD on $\mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}^2 \times S_2^{++}(\mathbb{R})\big)$ (Hua et al., 2023)
- Wasserstein task embedding (Liu et al., 2025)
- Sliced-Wasserstein on $\mathbb{R}^d \times \mathbb{H}$ (Bonet et al., 2024; Nguyen and Ho, 2024)



Alternatives to OTDD

- MMD on $\mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}^2 \times S_2^{++}(\mathbb{R})\big)$ (Hua et al., 2023)
- Wasserstein task embedding (Liu et al., 2025)
- Sliced-Wasserstein on $\mathbb{R}^d \times \mathbb{H}$ (Bonet et al., 2024; Nguyen and Ho, 2024)



- Sliced-Wasserstein on $\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$ (Nguyen et al., 2025)

Table of Contents

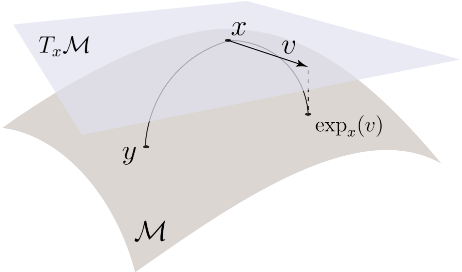
Riemannian Manifolds

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 $\log_x : \mathcal{M} \to T_x\mathcal{M}$



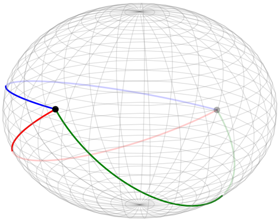
Riemannian Manifolds

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: $\forall t \in [0, 1],\ \gamma(t) = \exp_x \left(t \log_x(y) \right)$
- 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
- **Brenier-McCann's theorem**: If $\mu \ll \mathrm{Vol}$, then there exists a unique T_μ^ν s.t.

1. $(\mathrm{T}_\mu^\nu)_\# \mu = \nu$ ($\mathrm{T}_\# \mu(A) = \mu\big(\mathrm{T}^{-1}(A)\big)$ for all $A \subset \mathcal{M}$)
2. For all $x \in \mathcal{M}$, $\mathrm{T}_\mu^\nu(x) = \exp_x\big(-\nabla \varphi_{\mu,\nu}(x)\big)$, φ_μ^ν Kantorovich potential
3. $\mathrm{W}_2^2(\mu, \nu) = \int d\big(x, \mathrm{T}_\mu^\nu(x)\big)^2 \, \mathrm{d}\mu(x) = \int \|\nabla \varphi_\mu^\nu(x)\|_x^2 \, \mathrm{d}\mu(x)$

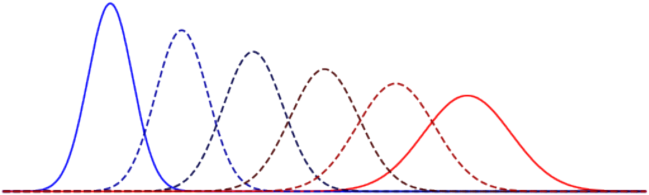
Reminder: For $\mathcal{M} = \mathbb{R}^d$, $d(x, y) = \|x - y\|_2$, $\exp_x(v) = x + v$, $\log_x(y) = y - x$.

Riemannian Structure of the Wasserstein Space

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

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

- Geodesics between $\mu,\nu\in\mathcal{P}_2(\mathcal{M})$,
  - If $\mu \ll \mathrm{Vol}$: $\forall t\in[0,1],\ \mu_t = \big(\exp_{\mathrm{Id}}\circ(-t\nabla\varphi_{\mu,\nu})\big)_\#\mu$
  - 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)



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

- If $\mu\ll\mathrm{Leb}$, $\mu_t = \big((1-t)\mathrm{Id} + t\mathrm{T}_\mu^\nu\big)_\#\mu = \big(\mathrm{Id} + t(\mathrm{T}_\mu^\nu - \mathrm{Id})\big)_\#\mu = \big(\mathrm{Id} - t\nabla\varphi_{\mu,\nu}\big)_\#\mu$
- In general: $\mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\#\gamma = \big(\pi^1 + t(\pi^2-\pi^1)\big)_\#\gamma$, $\gamma\in\Pi_o(\mu,\nu)$

Riemannian Structure of the Wasserstein Space

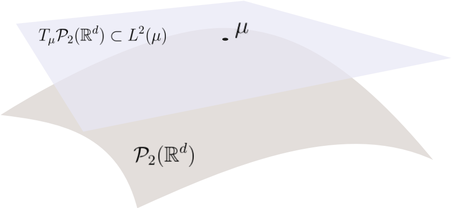
Let $T\mathcal{M} = \{(x,v),\ x\in\mathcal{M},\ v\in T_x\mathcal{M}\}$, $\pi^{\mathcal{M}}((x,v)) = x$, $\pi^v((x,v)) = v$.

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

- Geodesics between $\mu,\nu\in\mathcal{P}_2(\mathcal{M})$,
  - If $\mu \ll \mathrm{Vol}$: $\forall t\in[0,1],\ \mu_t = \big(\exp_{\mathrm{Id}}\circ(-t\nabla\varphi_{\mu,\nu})\big)_\#\mu$
  - 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)
- Tangent space at $\mu\in\mathcal{P}_2(\mathcal{M})$ (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

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 (Erbar, 2010, Lemma 3.2), *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)

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

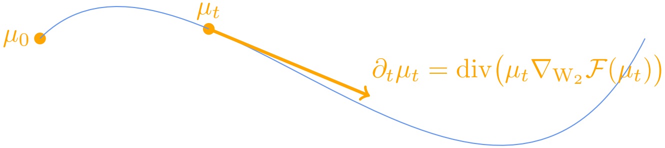
with k positive definite kernel, and:

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

Wasserstein Gradient Flows (Ambrosio et al., 2008)

Wasserstein gradient flow of $\mathcal{F}$: curve $t \mapsto \mu_t$ satisfying (weakly)

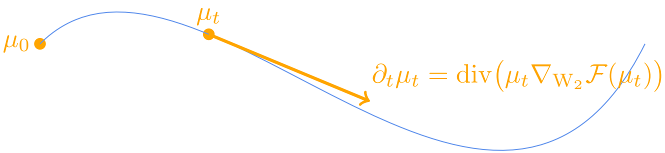
$$\partial_t \mu_t = \mathrm{div}\big(\mu_t \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_t)\big)$$



Wasserstein Gradient Flows (Ambrosio et al., 2008)

Wasserstein gradient flow of $\mathcal{F}$: curve $t \mapsto \mu_t$ satisfying (weakly)

$$\partial_t \mu_t = \mathrm{div}\big(\mu_t \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_t)\big)$$



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

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

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

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

Drawbacks:

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

Contributions

Model datasets as $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\nu_c^n} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ where $\nu_c^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^c}$
$\rightarrow$ require to minimize a functional on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$

Contributions

Model datasets as $\mathbb{P} = \frac{1}{C}\sum_{c=1}^C \delta_{\nu_c^n} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ where $\nu_c^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^c}$
$\to$ require to minimize a functional on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$

Contributions:

- Endow $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ with $\mathrm{W}_{\mathrm{W}_2}$
- Study differential structure of $\big(\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big), \mathrm{W}_{\mathrm{W}_2}\big)$
- Develop gradient flows on this space

Contributions

Model datasets as $\mathbb{P} = \frac{1}{C}\sum_{c=1}^C \delta_{\nu_c^n} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ where $\nu_c^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^c}$
$\rightarrow$ require to minimize a functional on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$

Contributions:

- Endow $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ with $\mathrm{W}_{\mathrm{W}_2}$
- Study differential structure of $\big(\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big), \mathrm{W}_{\mathrm{W}_2}\big)$
- Develop gradient flows on this space

Applications:

$$\min_{\mathbb{P}\in\mathcal{P}(\mathcal{P}(\mathbb{R}^d))} \ \mathbb{F}(\mathbb{P})$$

where $\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})$ for $\mathbb{Q}\in\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ a target dataset, and K a (positive definite kernel) on $\mathcal{P}_2(\mathbb{R}^d)$.

Example

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

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

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

Properties:

- $\mathrm{W}_{\mathrm{W}_2}$ distance, $(\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big), \mathrm{W}_{\mathrm{W}_2})$: WoW space
- **Brenier-McCann's theorem**: Let $\mathbb{P}_0$ a reference measure satisfying suitable assumptions (no atom, satisfies an IPP, see (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).

Geodesics

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

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

Properties

- $\mathbb{\Gamma} \mapsto (\varphi^{\mathcal{M}}, \varphi^{\exp})_{\#}\mathbb{\Gamma}$ is a surjective map from $\exp_{\mathbb{P}}^{-1}(\mathbb{Q})$ to $\Pi_o(\mathbb{P},\mathbb{Q})$
- If $\mathbb{P} \ll \mathbb{P}_0$, $\mathbb{\Gamma} = (\mu \mapsto (\mathrm{Id}, -\nabla\varphi_{\mu,\mathrm{T}(\mu)})_{\#}\mu)_{\#}\mathbb{P} \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$ is unique

Geodesic between $\mathbb{P}$ and $\mathbb{Q}$:

- If $\mathbb{P} \ll \mathbb{P}_0$, $\forall t \in [0,1]$, $\mathbb{P}_t = \big(\exp_{\mathrm{Id}} \circ (-t\nabla\varphi_{\mathrm{Id},\mathrm{T}})\big)_{\#}\mathbb{P}$
- In general, $\forall t \in [0,1]$, $\mathbb{P}_t = \big(\exp_{\varphi^{\mathcal{M}}} \circ (t\varphi^{v})\big)_{\#}\mathbb{\Gamma}$

Tangent Space

Definition (Cylinder)

$\mathcal{F} : \mathcal{P}_2(\mathcal{M}) \to \mathbb{R} \in \mathrm{Cyl}\big(\mathcal{P}_2(\mathcal{M})\big)$ 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 $\mu \in \mathcal{P}_2(\mathcal{M})$,

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

Definition (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 $\Gamma \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$,

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

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

Properties:

- If $\mathbb{P} \ll \mathbb{P}_0$, then 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 additional 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 $\mathbb{P} \ll \mathbb{P}_0$ and $\xi \in \partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. Then ξ is a strong subdifferential of $\mathbb{F}$ at $\mathbb{P}$, *i.e.*, for $\Psi \in L^2(\mathbb{P})$, $\Gamma = (\mathrm{Id}, \Psi)_{\#}\mathbb{P}$ and $\mathbb{Q} := \varphi^{\exp}_{\#}\Gamma$,

$$\mathbb{F}(\mathbb{Q}) \geq \mathbb{F}(\mathbb{P}) + \int \langle \xi(\mu), \Psi(\mu)\rangle_{L^2(\mu)}\mathrm{d}\mathbb{P}(\mu) + o(\|\Psi\|_{L^2(\mathbb{P})}).$$

WoW Gradient

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

Conjecture:
$$\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}) = \nabla_{\mathrm{W}_2}\frac{\delta\mathbb{F}}{\delta\mathbb{P}}(\mathbb{P}),$$
where the first variation $\frac{\delta\mathbb{F}}{\delta\mathbb{P}}(\mathbb{P}):\mathcal{P}_2(\mathcal{M}) \to \mathbb{R}$ at $\mathbb{P}$ is defined as the unique function (up to a constant) satisfying
$$\lim_{\varepsilon\to 0}\ \frac{\mathbb{F}(\mathbb{P}+\varepsilon\chi) - \mathbb{F}(\mathbb{P})}{\varepsilon} = \int \frac{\delta\mathbb{F}}{\delta\mathbb{P}}(\mathbb{P})\ \mathrm{d}\chi,$$
where $\int \mathrm{d}\chi = 0$ and $\mathbb{P}+\varepsilon\chi \in \mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$ for ε small.

WoW Gradient Flow

Discretizations:

- JKO scheme (Jordan et al., 1998):

$$\mathbb{P}_{k+1} = \underset{\mathbb{P}\in\mathcal{P}_2(\mathcal{P}_2(\mathcal{M}))}{\mathrm{argmin}} \ \frac{1}{2\tau}\mathrm{W}_{\mathrm{W}_2}(\mathbb{P},\mathbb{P}_k)^2 + \mathbb{F}(\mathbb{P})$$

$\rightarrow$ converges to the WoW gradient flow when $\tau \to 0$.

- Forward scheme:

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

At the distribution level: $\mu_{k+1} = \exp_{\mu_k}\big(-\tau\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}_k)(\mu_k)\big)$ where $\mu_k\sim\mathbb{P}_k$.

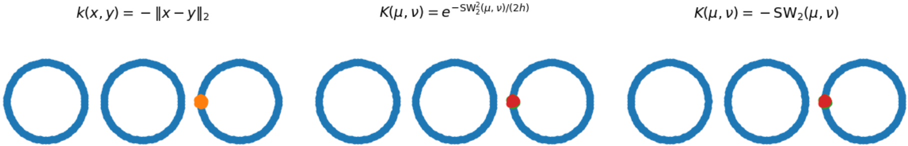
In practice: For $\mathbb{P}_k = \frac{1}{C}\sum_{c=1}^C \delta_{\mu_k^{c,n}}$ with $\mu_k^{c,n} = \frac{1}{n}\sum_{i=1}^n \delta_{x_k^{i,c}}$:

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

Table of Contents

Synthetic Data

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



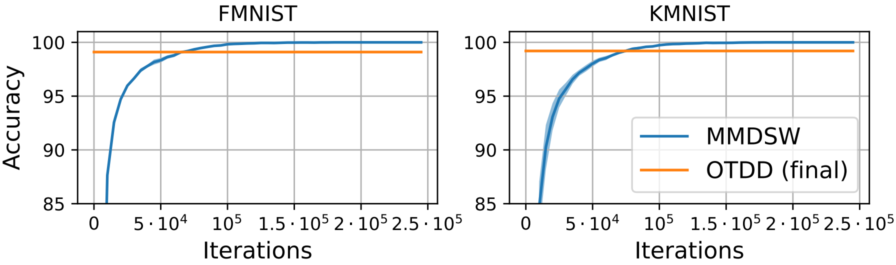
Kernels considered:

- Gaussian SW kernel: $K(\mu,\nu) = e^{-\mathrm{SW}_2^2(\mu,\nu)/(2h)}$ (h=0.05)
- Riesz SW kernel: $K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu)$
- Riesz kernel on $\mathbb{R}^d$: $k(x,y) = -\|x-y\|_2$

Domain Adaptation

Setting:

1. Pretrain a classifier on MNIST $\mathbb{Q}$
2. Flow other dataset to MNIST 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 flowed data

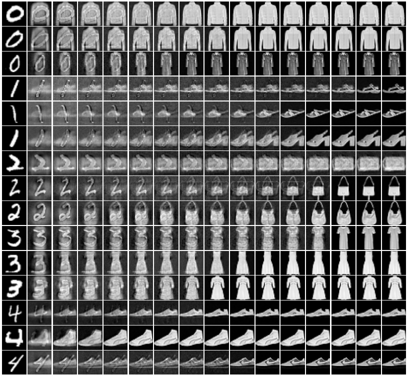


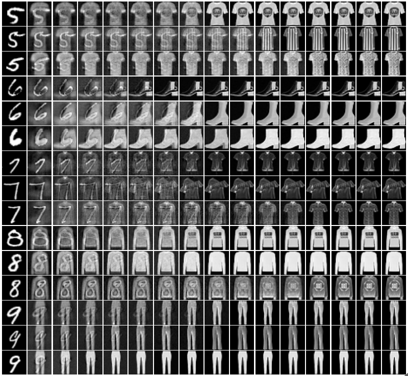
$\rightarrow$ reach 100% accuracy

Domain Adaptation

Setting:

1. Pretrain a classifier on MNIST $\mathbb{Q}$
2. Flow other dataset to MNIST 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 flowed data





Dataset Distillation (Wang et al., 2018)

Let $\mathcal{A}^\omega : \mathbb{R}^d \to \mathbb{R}^d$ be some data augmentation (*e.g.* rotation, cropping...), $\psi^\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$ with $d' \ll d$ a randomly initialized neural network used to embed the data, $\varphi^{\theta,\omega}(\mu) = \psi^\theta_\# \mathcal{A}^\omega_\# \mu$.

Goal: synthesize big dataset $\mathbb{Q} = \frac{1}{C} \sum_{c=1}^C \delta_{\nu_c}$

- Distribution Matching (DM) (Zhao and Bilen, 2023):

$$\mathcal{F}\big((\mu_c)_c\big) = \mathbb{E}_{\theta,\omega}\left[\sum_{c=1}^C \mathrm{MMD}_k^2\big(\psi^\theta_\# \mathcal{A}^\omega_\# \mu_c, \psi^\theta_\# \mathcal{A}^\omega_\# \nu_c\big)\right],$$

  with linear kernel $k(x,y) = \langle x, y\rangle$.

- Ours:

$$\tilde{\mathbb{F}}(\mathbb{P}) = \mathbb{E}_{\theta,\omega}\left[\mathrm{MMD}_K^2(\varphi^{\theta,\omega}_\# \mathbb{P}, \varphi^{\theta,\omega}_\# \mathbb{Q})\right],$$

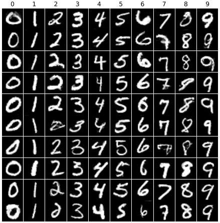
  with $K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^C \delta_{\mu_c^k}$.

Results Dataset Distillation

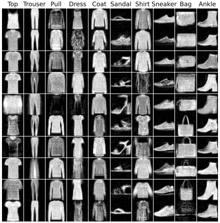
Table: Accuracy of the classifier trained on synthetic datasets with $k \in \{1, 10, 50\}$ synthetic images by class.

Dataset	k	$\psi^\theta = \mathcal{A}^\omega = \mathrm{Id}$		$\psi^\theta = \mathrm{Id}$		$\mathcal{A}^w = \mathrm{Id}$		$\mathcal{A}^w + \psi^\theta$		Baselines	
		DM	MMDSW	DM	MMDSW	DM	MMDSW	DM	MMDSW	Random	Full data
MNIST	1	$61.1_{\pm 6.5}$	**$66.5_{\pm 5.5}$**	-	**$66.8_{\pm 5.3}$**	**$87.8_{\pm 0.6}$**	$60.3_{\pm 3.4}$	**$87.7_{\pm 0.5}$**	$60.9_{\pm 3.3}$	$55.8_{\pm 2.0}$	99.4
	10	$88.2_{\pm 2.8}$	**$93.2_{\pm 0.7}$**	$88.7_{\pm 3.3}$	**$93.8_{\pm 0.7}$**	**$97.0_{\pm 0.1}$**	$96.4_{\pm 0.2}$	**$97.0_{\pm 0.1}$**	$96.4_{\pm 0.3}$	$92.2_{\pm 1.1}$	
	50	$95.9_{\pm 0.9}$	$97.0_{\pm 0.2}$	$95.3_{\pm 1.4}$	$97.5_{\pm 0.1}$	**$98.4_{\pm 0.1}$**	**$98.4_{\pm 0.1}$**	**$98.4_{\pm 0.1}$**	**$98.4_{\pm 0.1}$**	$97.6_{\pm 0.2}$	
FMNIST	1	$54.4_{\pm 3.2}$	**$60.0_{\pm 4.1}$**	-	**$60.6_{\pm 3.6}$**	$58.7_{\pm 0.4}$	**$60.9_{\pm 2.6}$**	$58.7_{\pm 0.5}$	**$60.8_{\pm 2.2}$**	$49.0_{\pm 7.5}$	92.4
	10	$74.6_{\pm 1.0}$	**$76.7_{\pm 1.0}$**	$74.7_{\pm 0.8}$	**$76.6_{\pm 1.1}$**	**$81.2_{\pm 2.3}$**	$78.0_{\pm 0.9}$	**$82.5_{\pm 0.3}$**	$78.9_{\pm 1.2}$	$75.3_{\pm 0.7}$	
	50	$81.3_{\pm 0.5}$	**$84.2_{\pm 0.1}$**	$81.4_{\pm 1.0}$	**$85.0_{\pm 0.2}$**	**$87.6_{\pm 0.2}$**	**$87.6_{\pm 0.2}$**	$87.5_{\pm 0.1}$	**$87.6_{\pm 0.2}$**	$83.2_{\pm 0.2}$	

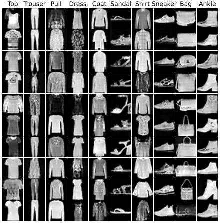
$\psi^\theta = \mathrm{Id}$



$\mathcal{A}^\omega + \psi^\theta$



$\psi^\theta = \mathrm{Id}$



$\mathcal{A}^\omega + \psi^\theta$

Transfer Learning

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

Table: Accuracy of classifier on augmented datasets for $k \in \{1, 10, 10, 100\}$. M refers to MNIST, F to Fashion MNIST, K to KMNIST and U to USPS.

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	$60.1_{\pm 1.1}$	$\mathbf{66.8}_{\pm 0.8}$	$66.3_{\pm 0.9}$	$62.4_{\pm 1.2}$
M to U	1	$32.4_{\pm 7.9}$	$37.4_{\pm 6.1}$	$\mathbf{39.5}_{\pm 7.9}$	$35.0_{\pm 5.6}$
	5	$51.4_{\pm 9.8}$	$73.0_{\pm 1.0}$	$\mathbf{73.3}_{\pm 1.4}$	$69.6_{\pm 1.3}$
	10	$60.3_{\pm 10.1}$	$\mathbf{77.2}_{\pm 1.2}$	$72.7_{\pm 2.7}$	$75.6_{\pm 1.2}$
	100	$87.5_{\pm 0.7}$	$\mathbf{89.7}_{\pm 0.4}$	-	$88.1_{\pm 0.6}$

Conclusion

Conclusion:

- Differential structure over the Wasserstein over Wasserstein Space
- Wasserstein over Wasserstein Gradient Flows
- Implementation on the MMD
- Application to Dataset Distillation and Transfer Learning

Perspectives:

- Use other positive definite kernels for the MMD
- Minimize other functionals
- Theoretical convergence

Conclusion

Conclusion:

- Differential structure over the Wasserstein over Wasserstein Space
- Wasserstein over Wasserstein Gradient Flows
- Implementation on the MMD
- Application to Dataset Distillation and Transfer Learning

Perspectives:

- Use other positive definite kernels for the MMD
- Minimize other functionals
- Theoretical convergence

Thank you for your attention!

References I

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.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds. *arXiv preprint arXiv:2403.06560*, 2024.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional Wasserstein Distances with Applications in Bayesian OT Flow Matching. *arXiv preprint arXiv:2403.18705*, 2024.

References II

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.

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.

References III

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM journal on mathematical analysis*, 29(1): 1–17, 1998.

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.

Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. Wasserstein Task Embedding for Measuring Task Similarities. *Neural Networks*, 181:106796, 2025.

Khai Nguyen and Nhat Ho. Hierarchical Hybrid Sliced Wasserstein: A Scalable Metric for Heterogeneous Joint Distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Khai Nguyen, Hai Nguyen, Tuan Pham, and Nhat Ho. Lightspeed geometric dataset distance via sliced optimal transport, 2025.

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

References IV

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

Bo Zhao and Hakan Bilen. Dataset Condensation with Distribution Matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023.