Flowing Datasets with Wasserstein over Wasserstein Gradient Flows

Clément Bonet[1]

Joint work with Christophe Vauthier[2] and Anna Korba[3]

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

5th IMPMS
11/06/2026

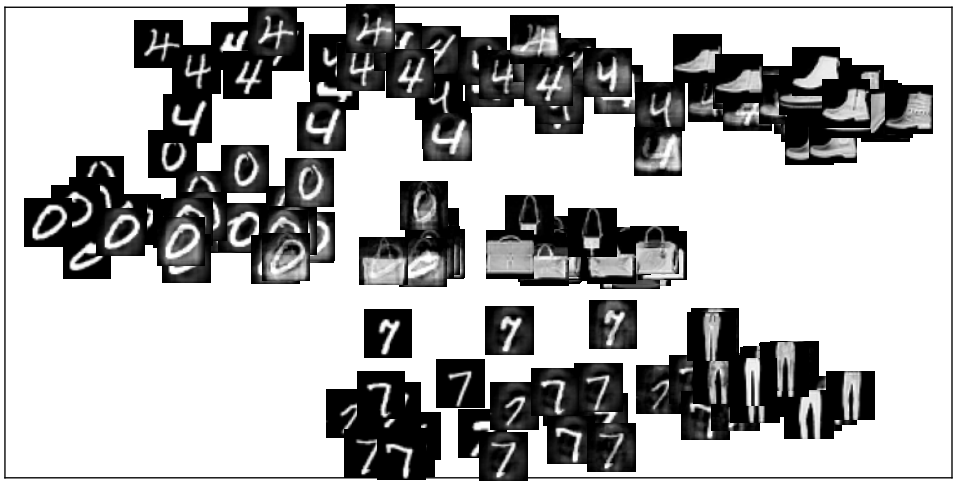
Motivations

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

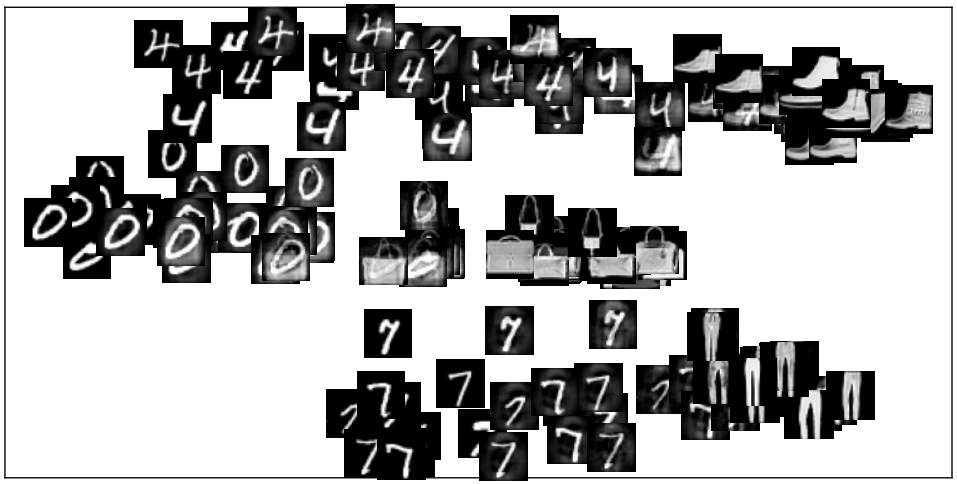
Goal: Generate samples from $\mathcal{D}$ respecting the structure of the dataset



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)

Table of Contents

The Wasserstein Distance

Wasserstein Distance

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

with $\Pi(\mu, \nu)$ set of couplings between μ and ν.

Properties:

- W_2 distance
- $\mathrm{W}_2(\delta_x, \delta_y) = \|x - y\|_2$
- Metrizes the weak convergence
- $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ has a Riemannian structure
  $\rightarrow$ Geodesics, Gradients...

Solving the OT Problem

Let $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$,

$$\mathrm{W}_2^2(\mu, \nu) = \min_{P \in \mathbb{R}_+^{n \times n},\ P\mathbb{1}_n = \alpha,\ P^T \mathbb{1}_n = \beta} \langle C, P \rangle_F \quad \text{with} \quad C = \left(\|x_i - y_j\|_2^2 \right)_{i,j}$$

Solving the OT Problem

Let $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$,

$$\mathrm{W}_2^2(\mu, \nu) = \min_{P \in \mathbb{R}_+^{n \times n},\ P\mathbb{1}_n = \alpha,\ P^T \mathbb{1}_n = \beta} \langle C, P \rangle_F \quad \text{with} \quad C = \left(\|x_i - y_j\|_2^2 \right)_{i,j}$$

Computational Complexity (Pele and Werman, 2009)

Numerical computation: **Linear program** in $O(n^3 \log n)$

1D OT Problem

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$,

- Cumulative distribution function:

$$\forall t \in \mathbb{R},\ F_\mu(t) = \mu(]-\infty, t]) = \int \mathbb{1}_{]-\infty,t]}(x)\ \mathrm{d}\mu(x)$$

- Quantile function:

$$\forall u \in [0,1],\ F_\mu^{-1}(u) = \inf\left\{x \in \mathbb{R},\ F_\mu(x) \geq u\right\}$$

1D Wasserstein Distance

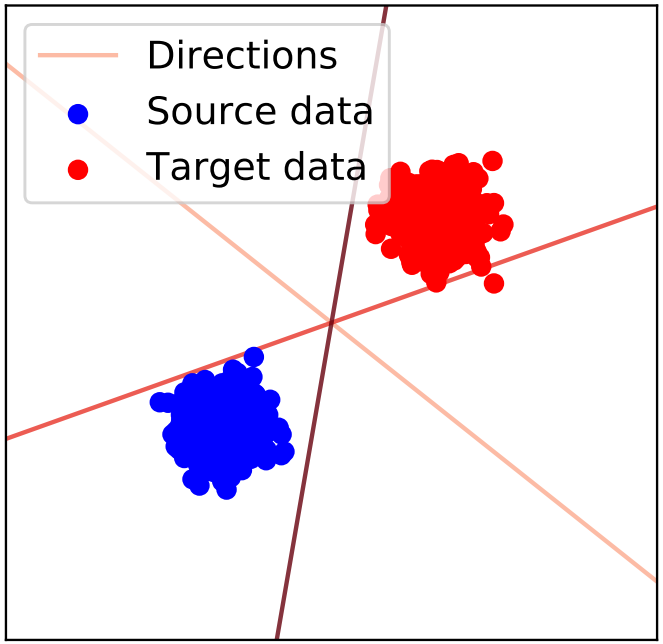
$$\mathrm{W}_2^2(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^2 \mathrm{d}u = \left\|F_\mu^{-1} - F_\nu^{-1}\right\|^2_{L^2([0,1])}$$

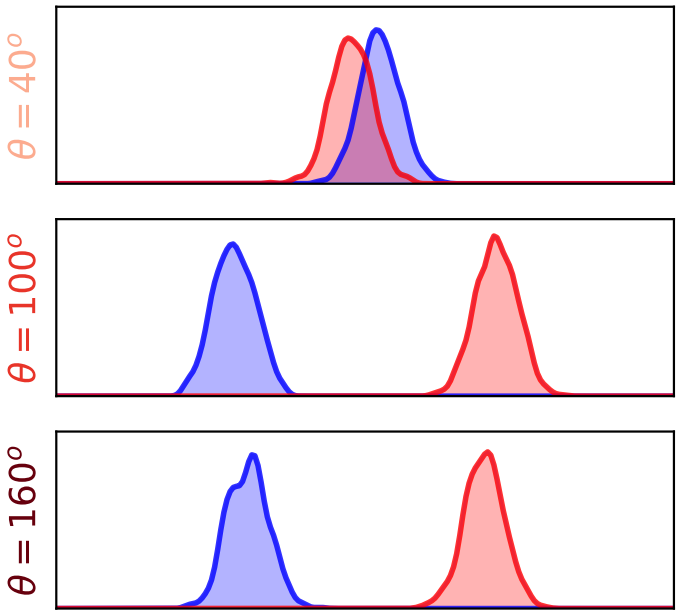
Let $x_1 < \cdots < x_n,\ y_1 < \cdots < y_n$, $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$,

$$\mathrm{W}_2^2(\mu,\nu) = \frac{1}{n}\sum_{i=1}^n (x_i - y_i)^2$$

$\rightarrow O(n \log n)$

Sliced-Wasserstein Distance





Definition (Sliced-Wasserstein (Rabin et al., 2011))

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{SW}_2^2(\mu, \nu) = \int_{S^{d-1}} \mathrm{W}_2^2(P^\theta_\# \mu, P^\theta_\# \nu) \, \mathrm{d}\lambda(\theta),$$

where $P^\theta(x) = \langle x, \theta \rangle$, λ uniform measure on S^{d-1}.

Properties of the Sliced-Wasserstein Distance

Let $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$.

Approximation via Monte-Carlo:

$$\widehat{\mathrm{SW}}_{2,L}^2(\mu, \nu) = \frac{1}{L} \sum_{\ell=1}^{L} \mathrm{W}_2^2(P_{\#}^{\theta_\ell} \mu, P_{\#}^{\theta_\ell} \nu),$$

$\theta_1, \dots, \theta_L \sim \lambda$.

Properties:

- Computational complexity: $O(Ln \log n + Lnd)$
- SW_2 distance (Bonnotte, 2013)
- Hilbertian structure $\rightarrow$ can be used in kernel methods

Table of Contents

Labeled Datasets

$\mathcal{D}_1 : \mu_1 = \frac{1}{m}\sum_{i=1}^{m} \delta_{(x_i^1, y_i^1)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$,

$\mathcal{D}_2 : \mu_2 = \frac{1}{m}\sum_{j=1}^{m} \delta_{(x_j^2, y_j^2)} \in \mathcal{P}(\mathbb{R}^d \times \{1, \dots, C\})$

C: number of classes, n: number of sample in each class, $m = nC$

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

Labeled Datasets

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

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

Optimal transport distance:

- $d\big((x, y), (x', y')\big)^2 = \|x - x'\|_2^2 + c(y, y')$
- $c(y, y') =?$

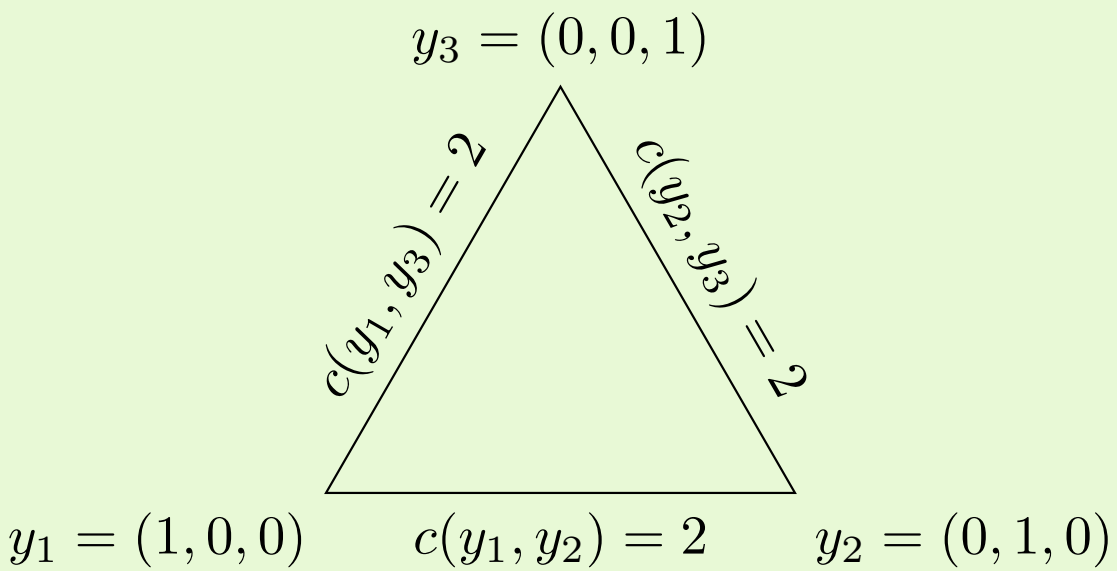
Labeled Datasets

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

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

- $d\big((x, y), (x', y')\big)^2 = \|x - x'\|_2^2 + c(y, y')$
- $c(y, y') = ?$

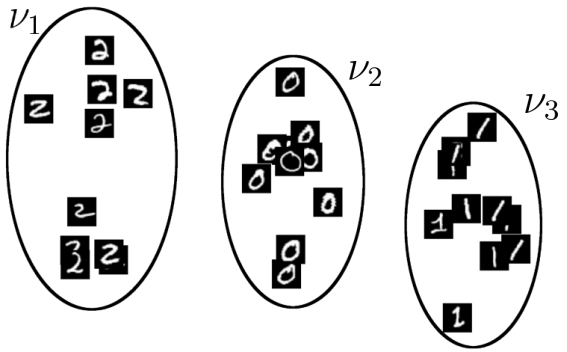
Example of bad choice: 1-hot encoding, $c(y, y') = 2\mathbb{1}_{\{y \neq y'\}}$



OTDD (Alvarez-Melis and Fusi, 2020)

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 $k = 1, 2$

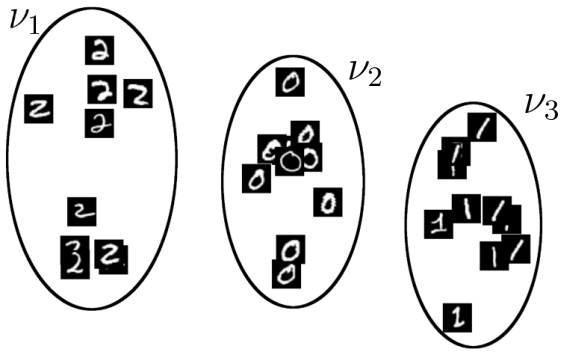


$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m} \sum_{i=1}^{m} \delta_{(x_i^k, \nu^k_{y_i^k})} \in \mathcal{P}\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$

OTDD (Alvarez-Melis and Fusi, 2020)

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 $k = 1, 2$



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

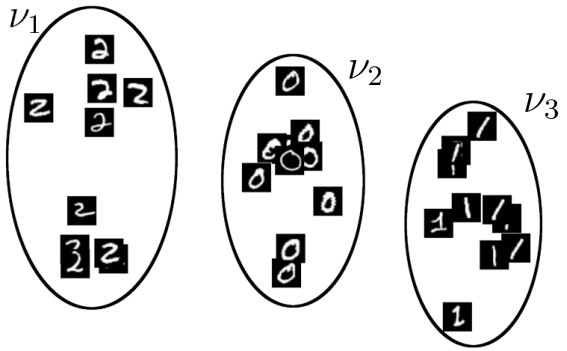
- Cost: $d\big((x,y),(x',y')\big)^2 = \|x - x'\|_2^2 + \mathrm{W}_2^2(\nu_y, \nu_{y'})$
- **Optimal transport distance**: $O\big(C^2 n^3 \log n + n^3 C^3 \log(nC)\big)$

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

OTDD (Alvarez-Melis and Fusi, 2020)

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 $k = 1, 2$



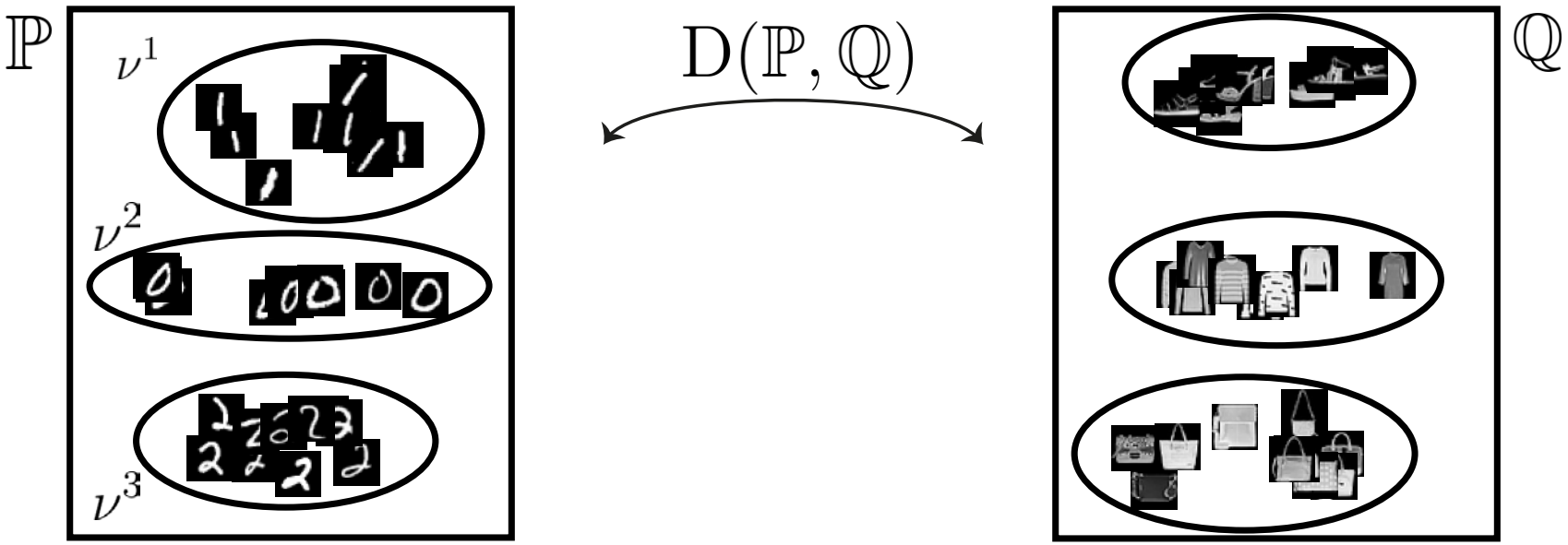
$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m} \sum_{i=1}^m \delta_{(x_i^k, m^k_{y_i^k}, \Sigma^k_{y_i^k})} \in \mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}^p \times S_p^{++}(\mathbb{R})\big)$

- 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 $O\big(C^2 p^3 + n^2 C^2 \log(nC)/\varepsilon^2\big)$

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

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 $\nu^c = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^c}$
  $\rightarrow$ MMD with positive definite kernel on $\mathcal{P}(\mathbb{R}^d)$
- Flow a dataset $\mathbb{P}$ towards $\mathbb{Q}$ by minimizing a discrepancy D on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$



MMD on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ (Bonet et al., 2025)

Maximum Mean Discrepancy

Let $\mu, \nu \in \mathcal{P}(X)$, $k : X \times X \to \mathbb{R}$ a positive definite kernel, *i.e.* for all $x_1, \dots, x_n \in X$, $a_1, \dots, a_n \in \mathbb{R}$, $\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$.

$$\mathrm{MMD}_k^2(\mu, \nu) = \iint k(x, y)\ \mathrm{d}(\mu - \nu)(x)\mathrm{d}(\mu - \nu)(y)$$

MMD on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ (Bonet et al., 2025)

Maximum Mean Discrepancy

Let $\mu, \nu \in \mathcal{P}(X)$, $k : X \times X \to \mathbb{R}$ a positive definite kernel, *i.e.* for all $x_1, \dots, x_n \in X$, $a_1, \dots, a_n \in \mathbb{R}$, $\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$.

$$\mathrm{MMD}_k^2(\mu, \nu) = \iint k(x, y)\ \mathrm{d}(\mu - \nu)(x)\mathrm{d}(\mu - \nu)(y)$$

Positive Definite Kernels on $\mathcal{P}_2(\mathbb{R}^d)$

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $h > 0$,

- Gaussian SW kernel: $K(\mu, \nu) = e^{-\mathrm{SW}_2^2(\mu,\nu)/(2h)}$
  $\to$ Positive definite (Kolouri et al., 2016)
- Riesz SW kernel: $K(\mu, \nu) = -\mathrm{SW}_2(\mu, \nu)$
  $\to$ Conditionally positive definite (*i.e.* holds for $\sum_{i=1}^{n} a_i = 0$)

Complexity: $O\big(C^2 Ln(\log n + d)\big)$

Table of Contents

Wasserstein over Wasserstein Distance (WoW)

Definition (WoW distance)

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

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

Wasserstein over Wasserstein Distance (WoW)

Definition (WoW distance)

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ 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, $(\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big), \mathrm{W}_{\mathrm{W}_2})$: WoW space
- **Riemannian structure**
- Can be rewritten (Bonet et al., 2025; Pinzi and Savaré, 2025):

$$\mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q}) = \inf_{\mathbb{\Gamma} \in \Lambda(\mathbb{P}, \mathbb{Q})} \iint \|y - x\|_2^2\, \mathrm{d}\gamma(x, y) \mathrm{d}\mathbb{\Gamma}(\gamma),$$

where $\Lambda(\mathbb{P}, \mathbb{Q}) = \{\mathbb{\Gamma} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)\big),\ \phi^1_{\#}\mathbb{\Gamma} = \mathbb{P},\ \phi^2_{\#}\mathbb{\Gamma} = \mathbb{Q}\}$,
$\phi^1(\gamma) = \pi^1_{\#}\gamma$ and $\phi^2(\gamma) = \pi^2_{\#}\gamma$ for $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$.

Tangent Space

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

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

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

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

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

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

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

WoW Gradient

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$.

Definition (WoW gradient)

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

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P})(\pi^1_{\#}\gamma)(x), y - x\rangle \, \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\big).$$

WoW Gradient

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$.

Definition (WoW gradient)

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

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P})(\pi^1_\#\gamma)(x), y - x\rangle \, \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\big).$$

Properties:

- There is at most one element in $\partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$
- Under assumptions on $\mathbb{P}$ and $\mathbb{R}^d$, existence of $\xi \in \partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$
- If $\xi \in \partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. Then ξ is a strong differential of $\mathbb{F}$ at $\mathbb{P}$, *i.e.*, for $\mathbb{\Gamma} \in \Lambda(\mathbb{P}, \mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \int \langle \xi(\pi^1_\#\gamma)(x), y - x\rangle \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\left(\sqrt{\iint \|y - x\|_2^2 \, \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma)}\right).$$

WoW Gradient

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$.

Definition (WoW gradient)

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

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P})(\pi^1_\# \gamma)(x), y - x \rangle \, \mathrm{d}\gamma(x,y) \mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\big).$$

Example of functionals

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

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

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

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

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

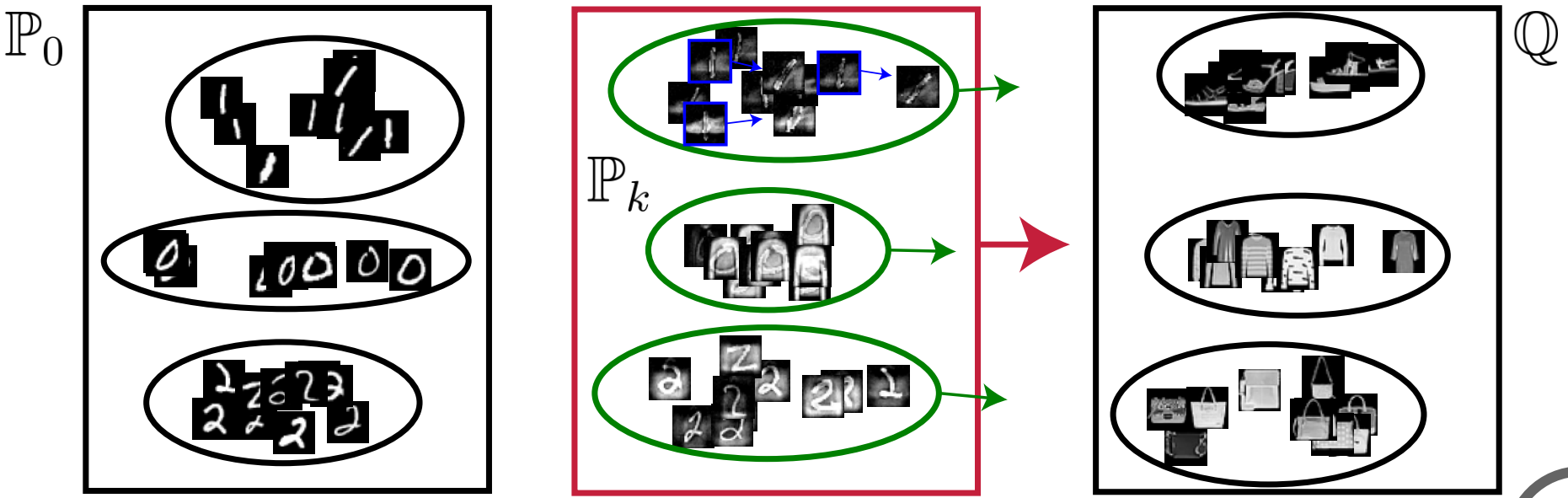


Table of Contents

Synthetic Data

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

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

- 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.

$k(x,y) = -\|x-y\|_2$ $K(\mu,\nu) = e^{-\mathrm{SW}_2^2(\mu,\nu)/(2h)}$ $K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu)$

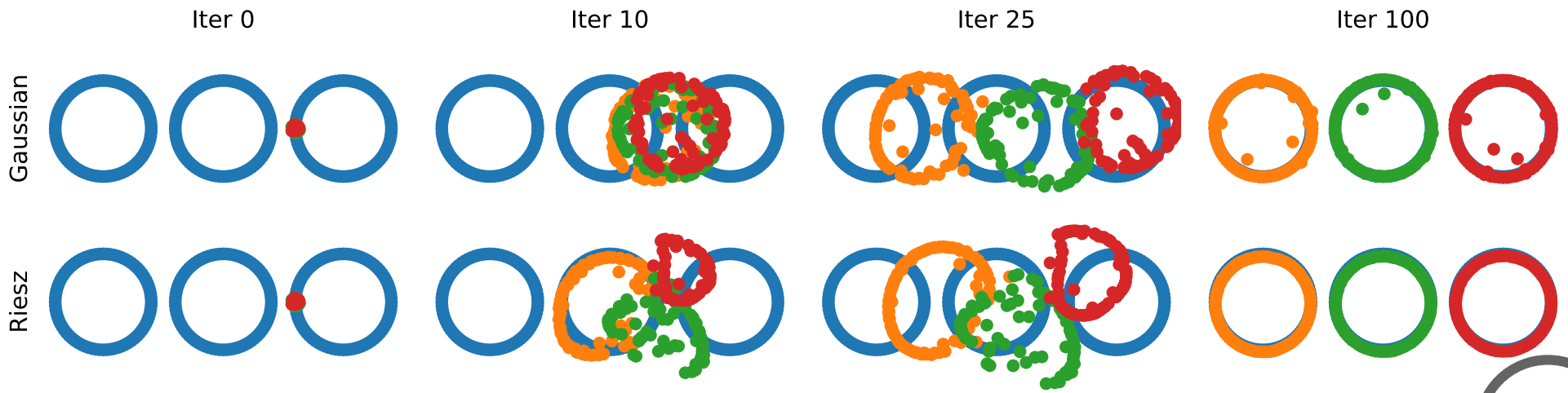
Synthetic Data

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

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

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

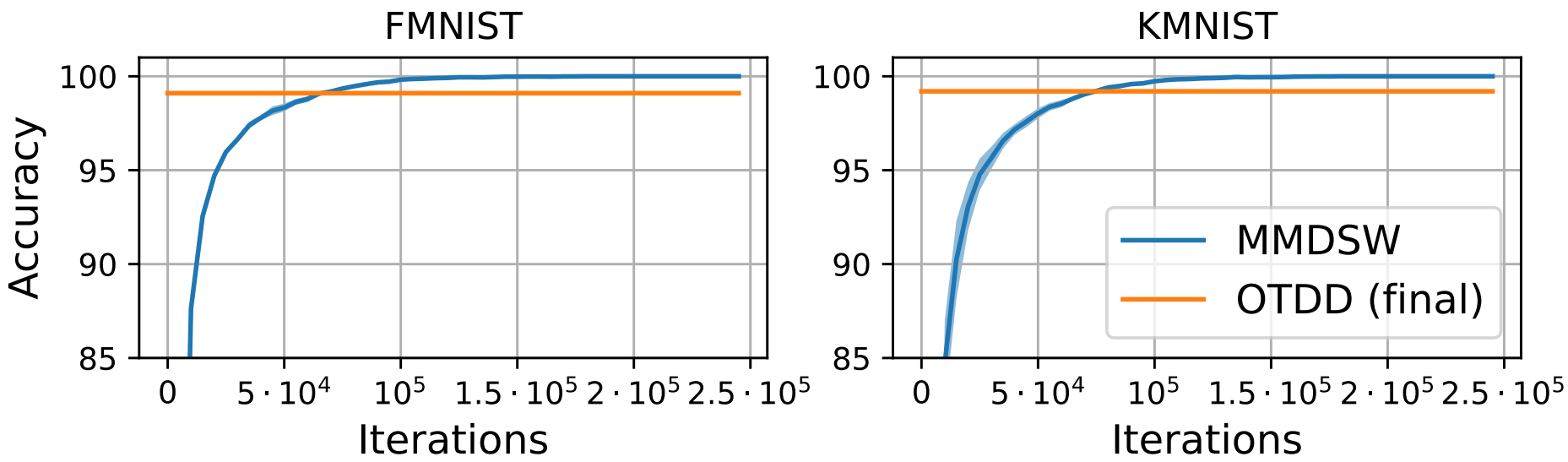
Goal: $\min_{\mathbb{P}} \ \mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q})$, where $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^{c,n}}$, $\nu^{c,n}$ ring.



"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)

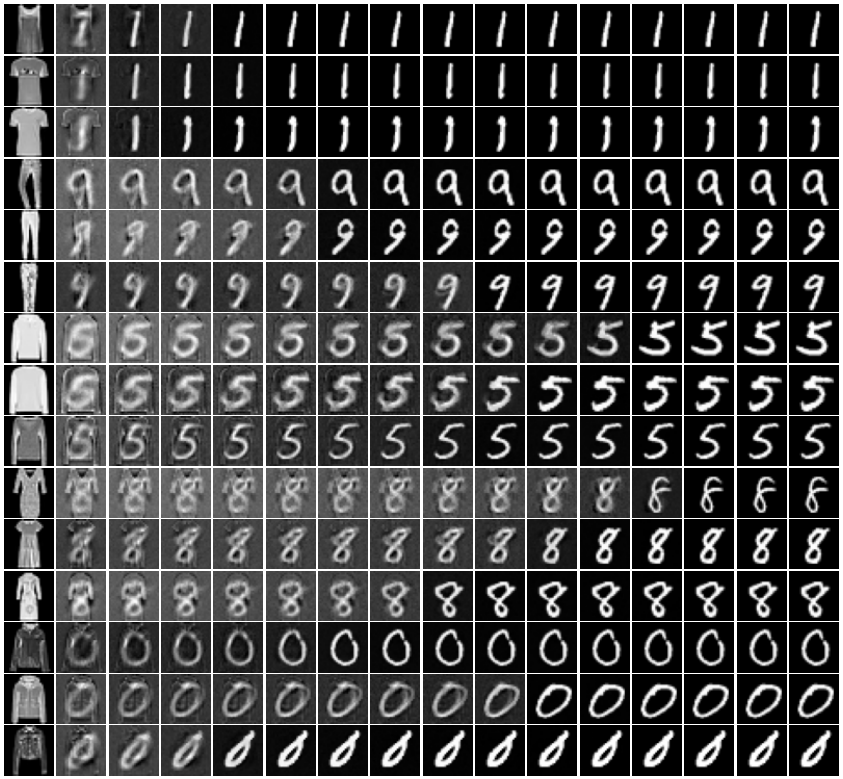


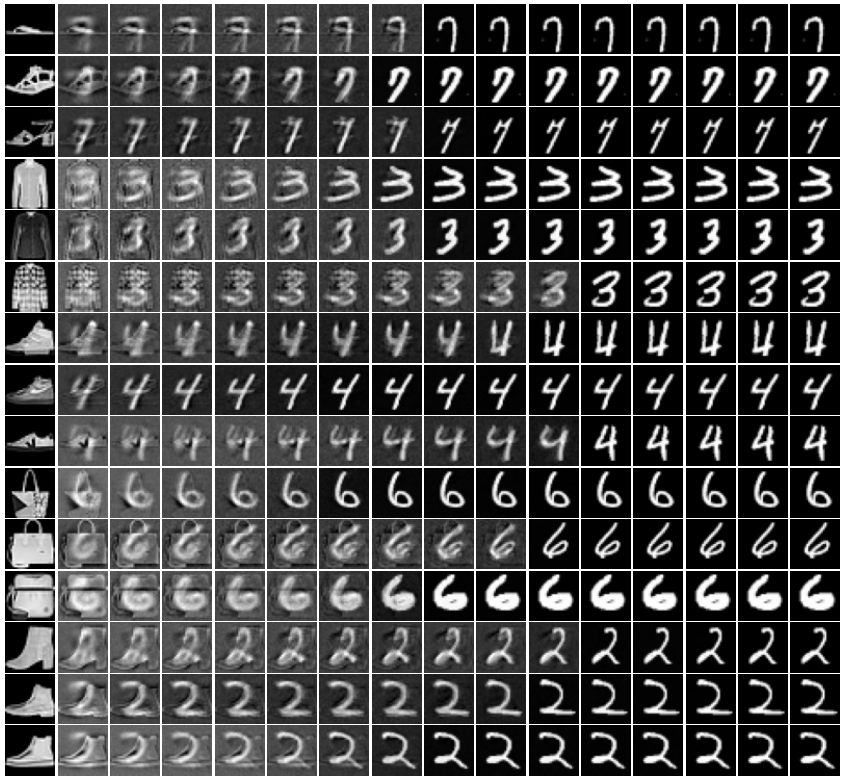
$\rightarrow$ reach 100% accuracy

"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)





Applications

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	$61.1_{\pm 6.5}$	**66.5**$_{\pm 5.5}$	$55.8_{\pm 2.0}$	99.4
	10	$88.2_{\pm 2.8}$	**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	$54.4_{\pm 3.2}$	**60.0**$_{\pm 4.1}$	$49.0_{\pm 7.5}$	92.4
	10	$74.6_{\pm 1.0}$	**76.7**$_{\pm 1.0}$	$75.3_{\pm 0.7}$	
	50	$81.3_{\pm 0.5}$	**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	$26.0_{\pm 5.3}$	**40.5**$_{\pm 4.7}$	$30.5_{\pm 4.2}$	$36.4_{\pm 3.3}$
	5	$38.5_{\pm 6.7}$	$61.5_{\pm 4.6}$	$59.7_{\pm 1.8}$	**62.7**$_{\pm 1.1}$
	10	$53.9_{\pm 7.9}$	$65.4_{\pm 1.5}$	$64.0_{\pm 1.4}$	**66.2**$_{\pm 1.0}$
	100	$71.1_{\pm 1.5}$	**74.7**$_{\pm 0.8}$	-	$73.5_{\pm 0.7}$
M to K	1	$18.4_{\pm 3.1}$	**20.9**$_{\pm 2.0}$	$18.8_{\pm 2.1}$	$19.4_{\pm 1.9}$
	5	$25.9_{\pm 4.0}$	$37.4_{\pm 2.2}$	$31.3_{\pm 1.4}$	**39.0**$_{\pm 1.0}$
	10	$30.9_{\pm 4.6}$	**44.7**$_{\pm 1.8}$	$34.1_{\pm 0.9}$	$44.1_{\pm 1.2}$
	100	$60.1_{\pm 1.1}$	**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; Bonet et al., 2026)
- 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; Bonet et al., 2026)
- 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.

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.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing Datasets with Wasserstein over Wasserstein Gradient Flows. In *International Conference on Machine Learning*. PMLR, 2025.

References II

Clément Bonet, Elsa Cazelles, Lucas Drumetz, and Nicolas Courty. Busemann Functions in the Wasserstein Space: Existence, Closed-Forms, and Applications to Slicing. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

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

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.

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

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.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33: 20802–20812, 2020.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

References IV

Alessandro Pinzi and Giuseppe Savaré. Totally convex functions, l^2-optimal transport for laws of random measures, and solution to the monge problem. *arXiv preprint arXiv:2509.01768*, 2025.

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.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As the
unprocessed data that should have been added to the final page this e
has been added to receive it.
If you rerun the document (without altering it) this surplus page will g
because LaTeX now knows how many pages to expect for this documen