Comparing and Flowing Datasets with Wasserstein over Wasserstein Gradient Flows

Clément Bonet[1]

Joint works with Christophe Vauthier[3], Anna Korba[2], Elsa Cazelles[4], Lucas Drumetz[5] and Nicolas Courty[6]

[1]Ecole Polytechnique, CMAP, Institut Polytechnique de Paris
[2]ENSAE, CREST, Institut Polytechnique de Paris
[3]Université Paris-Saclay, Laboratoire de Mathématique d'Orsay
[4]CNRS, Université de Toulouse, IRIT
[5]IMT Atlantique, Lab-STICC
[6]Université Bretagne Sud, IRISA

Séminaire Palaisien
06/01/2026

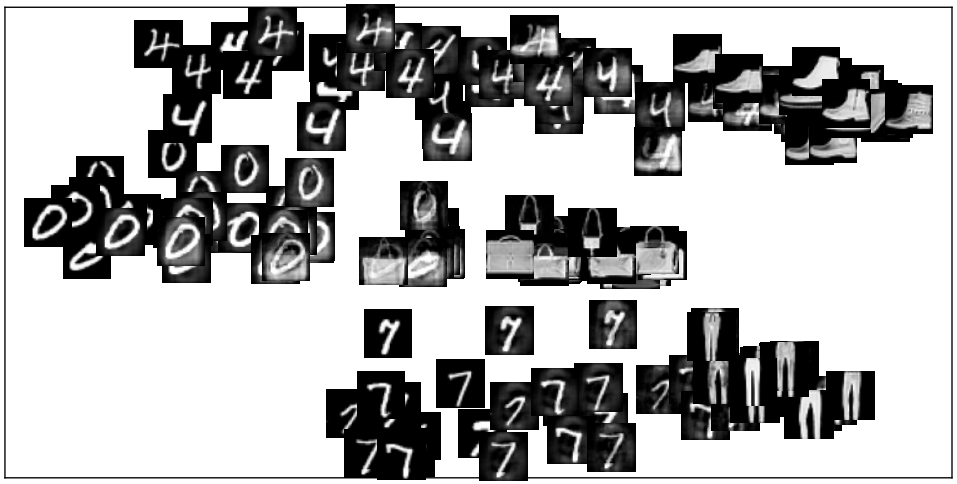
Motivations

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

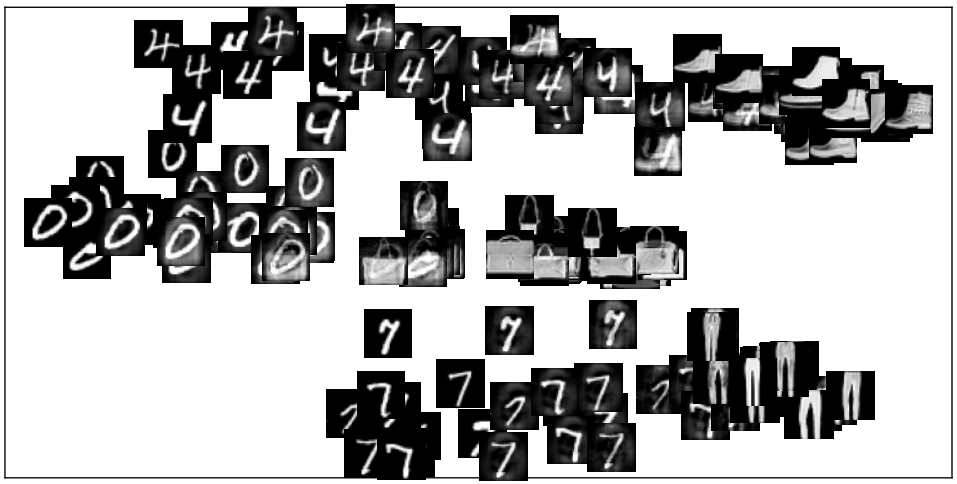
Goal: Generate samples from $\mathcal{D}$ respecting the structure of the dataset



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:

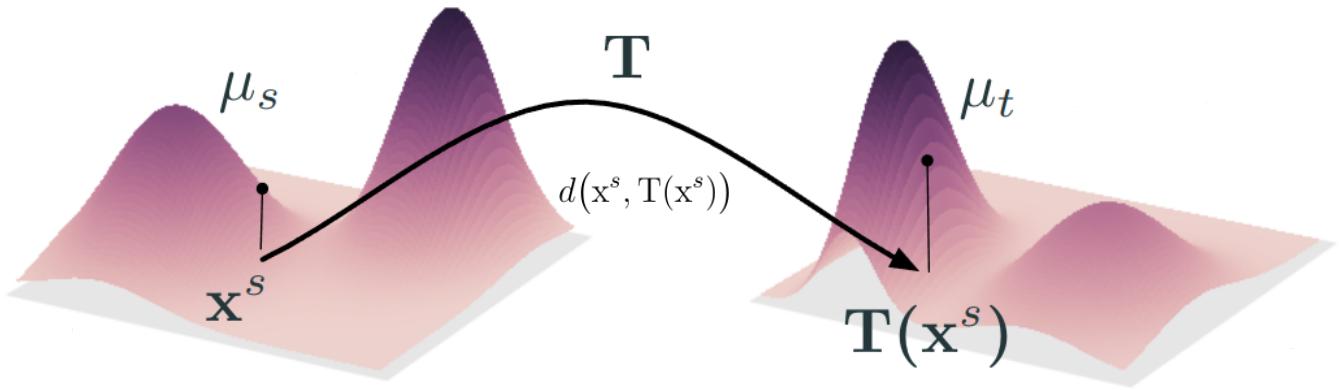
- Domain adaptation (Courty et al., 2016)
- Transfer learning (Alvarez-Melis and Fusi, 2021; Hua et al., 2023)
- Dataset distillation (Wang et al., 2018)

Table of Contents

Optimal Transport

Optimal Transport methods

- **Compare probability distributions**
- Leverage the **geometry** of the underlying space X



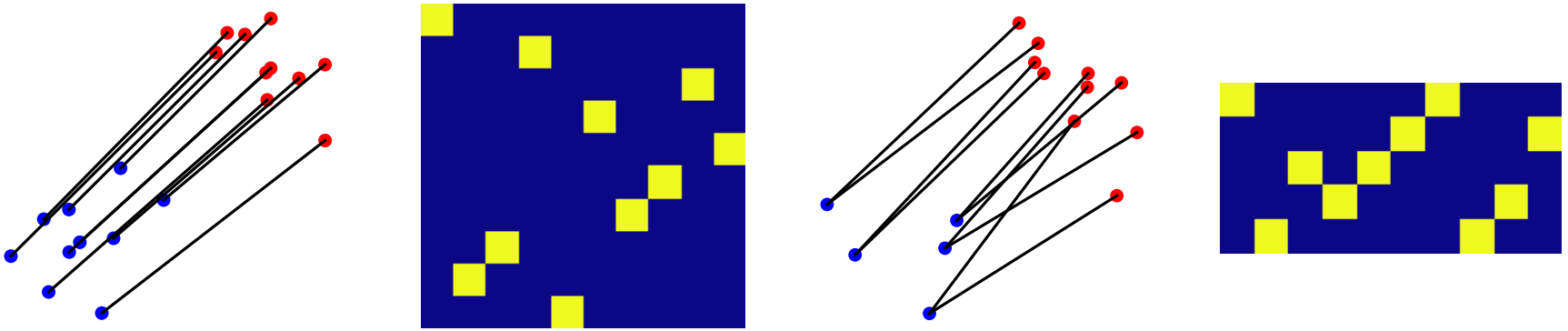
The Kantorovich Problem

Kantorovich Problem

Let $\mu, \nu \in \mathcal{P}_2(\mathrm{X})$, $c : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$,

$$\mathrm{OT}_c(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int c(x, y) \, \mathrm{d}\gamma(x, y),$$

$$\Pi(\mu, \nu) = \left\{\gamma \in \mathcal{P}(\mathrm{X} \times \mathrm{X}), \ \forall A \subset \mathrm{X}, \ \gamma(A \times \mathrm{X}) = \mu(A), \ \gamma(\mathrm{X} \times A) = \nu(A)\right\}$$



The Wasserstein Distance

Wasserstein Distance

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $c(x, y) = \|x - y\|_2^2$ for all $x, y \in \mathbb{R}^d$,

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

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 $O(n^3 \log n)$

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 $O(n^3 \log n)$

Sample Complexity (Boissard and Le Gouic, 2014)

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $x_1, \dots, x_n \sim \mu$, $y_1, \dots, y_n \sim \nu$, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$,

$$\mathbb{E}\left[\left|\mathrm{W}_2(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{W}_2(\mu, \nu)\right|\right] = O(n^{-1/d})$$

$\rightarrow$ **curse of dimensionality**

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

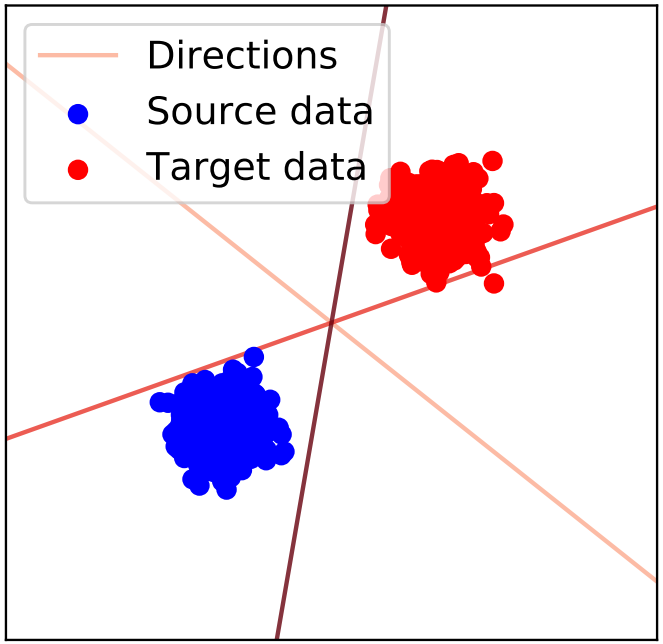
$$\mathrm{W}_2^2(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^2 \mathrm{d}u = \left\|F_\mu^{-1} - F_\nu^{-1}\right\|^2_{L^2([0,1])}$$

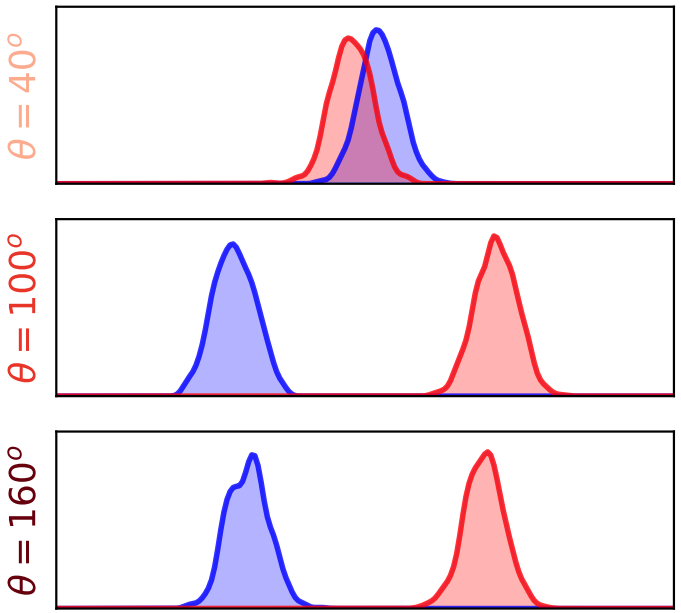
Let $x_1 < \cdots < x_n,\ y_1 < \cdots < y_n$, $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$,

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

$\rightarrow O(n \log n)$

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:

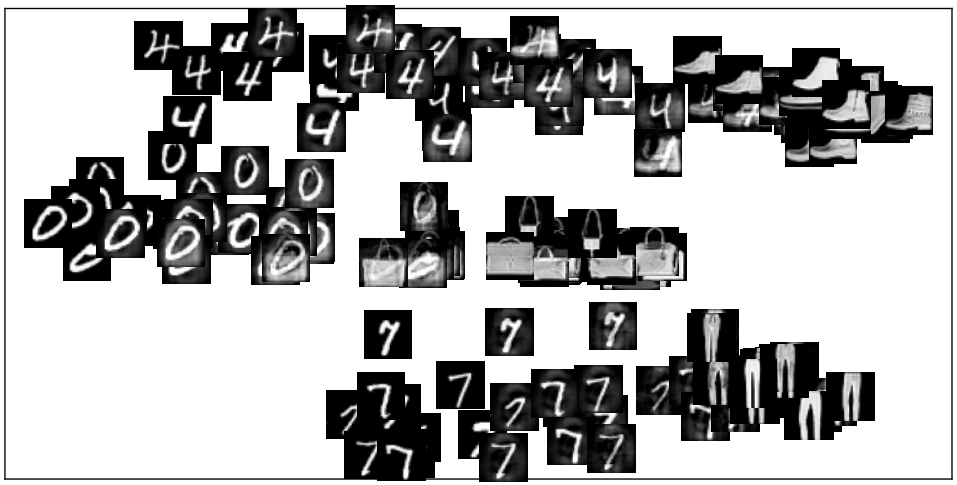
- Computational complexity: $O(Ln \log n + Lnd)$
- Sample complexity: independent of the dimension (Nadjahi et al., 2020)
- SW_2 distance (Bonnotte, 2013)
- Topologically equivalent to the Wasserstein distance (Nadjahi et al., 2019), *i.e.* $\lim_{n \to \infty} \mathrm{SW}_2^2(\mu_n, \mu) = 0 \iff \lim_{n \to \infty} \mathrm{W}_2^2(\mu_n, \mu) = 0.$

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, $m = nC$

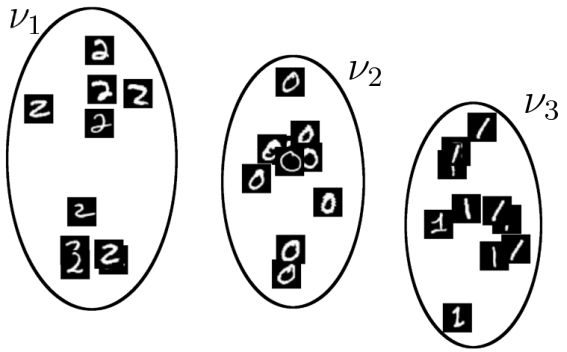
Question: how to compare datasets $\mathcal{D}_1$ and $\mathcal{D}_2$?



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 $k = 1, 2$

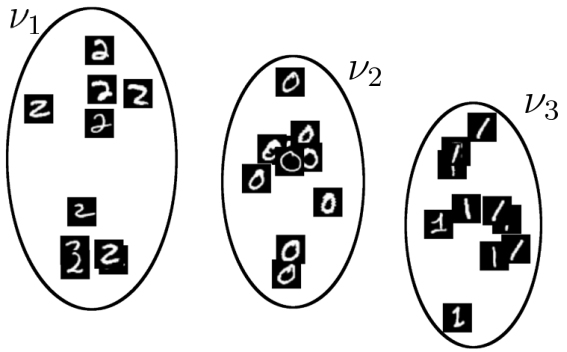


$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m} \sum_{i=1}^{m} \delta_{(x_i^k, \nu^k_{y_i^k})} \in \mathcal{P}\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$

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 $k = 1, 2$



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

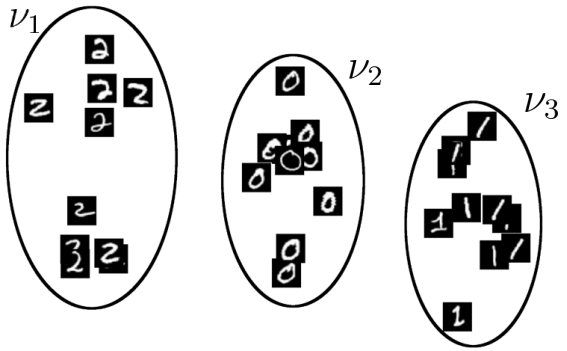
- Cost: $d\big((x,y),(x',y')\big)^2 = \|x - x'\|_2^2 + \mathrm{W}_2^2(\nu_y, \nu_{y'})$
- **Optimal transport distance**: $O\big(C^2 n^3 \log n + n^3 C^3 \log(nC)\big)$

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

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 $k = 1, 2$



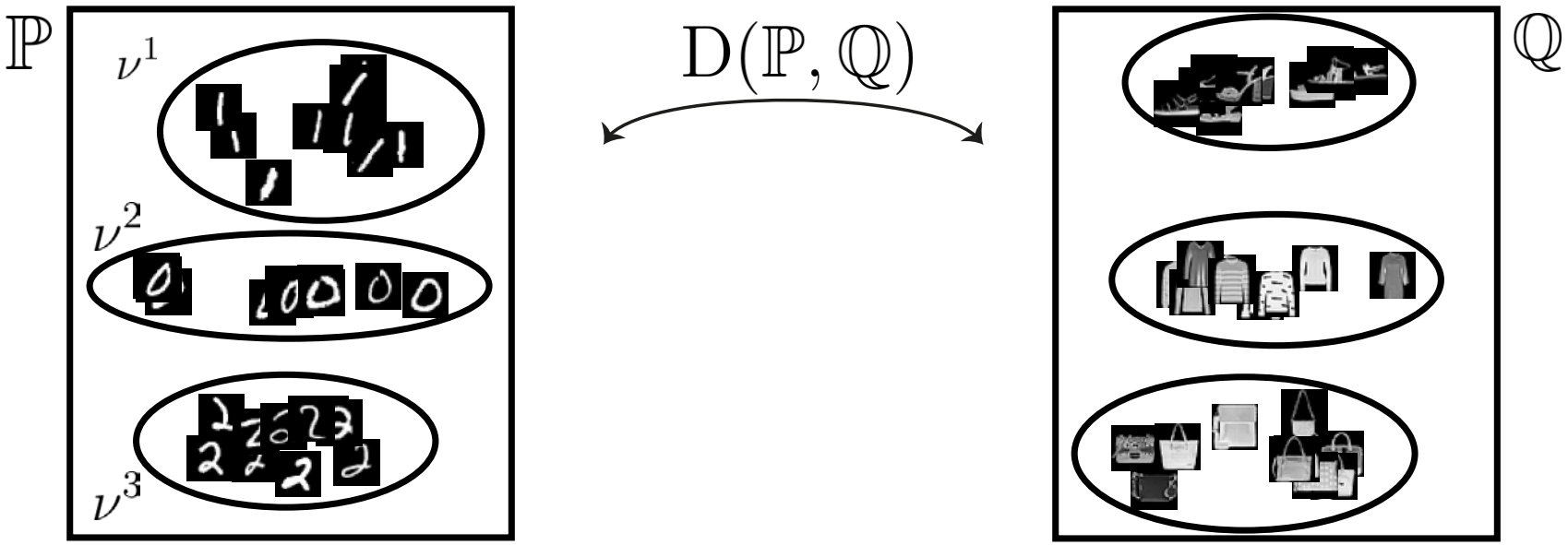
$\rightarrow \mathcal{D}_k : \mu_k = \frac{1}{m} \sum_{i=1}^m \delta_{(x_i^k, m^k_{y_i^k}, \Sigma^k_{y_i^k})} \in \mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}^p \times S_p^{++}(\mathbb{R})\big)$

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

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

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 $\mathcal{P}(\mathbb{R}^d)$
- Sliced on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$



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

Maximum Mean Discrepancy

Let $\mu, \nu \in \mathcal{P}(X)$, $k : X \times X \to \mathbb{R}$ a positive definite kernel.

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

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

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

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

Sliced on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$

- Sliced-Wasserstein on $\mathcal{P}_2(X \times Y)$ (Nguyen and Ho, 2024):
  - Define 2 projections $P^\theta : X \to \mathbb{R}$, $Q^\phi : Y \to \mathbb{R}$
  - For $\alpha \in S^1$, define

$$\forall (x, y) \in X \times Y,\ P^{\alpha,\theta,\phi}(x, y) = \alpha_1 P^\theta(x) + \alpha_2 Q^\phi(y)$$

  - For $\mu, \nu \in \mathcal{P}_2(X \times Y)$,

$$\mathrm{SW}_2^2(\mu, \nu) = \int \mathrm{W}_2^2(P^{\alpha,\theta,\phi}_{\#}\mu, P^{\alpha,\theta,\phi}_{\#}\nu)\ \mathrm{d}\lambda(\alpha, \theta, \phi)$$

Sliced on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$

- Sliced-Wasserstein on $\mathcal{P}_2(X \times Y)$ (Nguyen and Ho, 2024):
  - Define 2 projections $P^\theta : X \to \mathbb{R}$, $Q^\phi : Y \to \mathbb{R}$
  - For $\alpha \in S^1$, define

$$\forall (x, y) \in X \times Y,\ P^{\alpha,\theta,\phi}(x, y) = \alpha_1 P^\theta(x) + \alpha_2 Q^\phi(y)$$

  - For $\mu, \nu \in \mathcal{P}_2(X \times Y)$,

$$\mathrm{SW}_2^2(\mu, \nu) = \int \mathrm{W}_2^2(P^{\alpha,\theta,\phi}_{\#}\mu, P^{\alpha,\theta,\phi}_{\#}\nu)\ \mathrm{d}\lambda(\alpha, \theta, \phi)$$

- Sliced-Wasserstein on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$ (SOTDD) (Nguyen et al., 2025)
  - For a label $y \in \{1, \dots, C\}$, define $\varphi(y) = \frac{1}{n}\sum_{i=1}^m \delta_{x_i} \mathbb{1}_{\{y_i = y\}}$
  - Use for $\alpha \in S^k$,

$$P^{\alpha,\theta,\lambda}(x, y) = \alpha_1 P^\theta(x) + \sum_{i=1}^{k} \alpha_{i+1} \mathcal{M}^{\lambda_i}\left(P^\theta_{\#}\varphi(y)\right),$$

with $P^\theta : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{M}^\lambda : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ the moment transform projection.

Sliced on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$

- Sliced-Wasserstein on $\mathcal{P}_2(X \times Y)$ (Nguyen and Ho, 2024):
  - Define 2 projections $P^\theta : X \to \mathbb{R}$, $Q^\phi : Y \to \mathbb{R}$
  - For $\alpha \in S^1$, define

$$\forall (x, y) \in X \times Y,\ P^{\alpha,\theta,\phi}(x, y) = \alpha_1 P^\theta(x) + \alpha_2 Q^\phi(y)$$

  - For $\mu, \nu \in \mathcal{P}_2(X \times Y)$,

$$\mathrm{SW}_2^2(\mu, \nu) = \int \mathrm{W}_2^2(P^{\alpha,\theta,\phi}_{\#}\mu, P^{\alpha,\theta,\phi}_{\#}\nu)\ \mathrm{d}\lambda(\alpha, \theta, \phi)$$

- Sliced-Wasserstein on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$ (SOTDD) (Nguyen et al., 2025)
  - For a label $y \in \{1, \dots, C\}$, define $\varphi(y) = \frac{1}{n}\sum_{i=1}^m \delta_{x_i} \mathbb{1}_{\{y_i = y\}}$
  - Use for $\alpha \in S^k$,

$$P^{\alpha,\theta,\lambda}(x, y) = \alpha_1 P^\theta(x) + \sum_{i=1}^{k} \alpha_{i+1} \mathcal{M}^{\lambda_i}\big(P^\theta_{\#}\varphi(y)\big),$$

  with $P^\theta : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{M}^\lambda : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ the moment transform projection.

$\to$ **Use the Busemann function B^μ for projecting distributions on $\mathbb{R}$**

Busemann Function

Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $t \in \mathbb{R}_+ \mapsto \mu_t$ a geodesic ray starting from μ_0

Busemann function in $\left(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2\right)$: $B^\mu : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^d),\ B^\mu(\nu) = \lim_{t \to +\infty} \mathrm{W}_2(\mu_t, \nu) - \kappa_\mu t,$$

with $\kappa_\mu = \mathrm{W}_2(\mu_0, \mu_1)$.

$\to$ Generalization of linear projections in Euclidean spaces

For $x_0, v \in \mathbb{R}^d$, $x_t = x_0 + t(v - x_0)$,

$$\begin{aligned} \forall x \in \mathbb{R}^d,\ B^v(x) &= \lim_{t \to +\infty} \|x - x_t\|_2 - t\|v - x_0\|_2 \\ &= -\left\langle x - x_0, \frac{v - x_0}{\|v - x_0\|_2} \right\rangle \end{aligned}$$

Busemann Function

Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $t \in \mathbb{R}_+ \mapsto \mu_t$ a geodesic ray starting from μ_0

Busemann function in $\left(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2\right)$: $B^\mu : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^d),\ B^\mu(\nu) = \lim_{t \to +\infty} \mathrm{W}_2(\mu_t, \nu) - \kappa_\mu t,$$

with $\kappa_\mu = \mathrm{W}_2(\mu_0, \mu_1)$.

$\to$ Generalization of linear projections in Euclidean spaces

In (Bonet et al., 2025a):

- **Existence** and **characterization** of geodesic rays on $\mathcal{P}_2(\mathbb{R}^d)$
- **Closed-form in 1D**:

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}),\ B^\mu(\nu) = -\langle F_1^{-1} - F_0^{-1}, F_\nu^{-1} - F_0^{-1} \rangle_{L^2([0,1])}.$$

- **Closed-form in Gaussian case**: For $\mu_i = \mathcal{N}(m_i, \Sigma_i),\ \nu = \mathcal{N}(m, \Sigma)$,

$$B^\mu(\nu) = -\langle m_1 - m_0, m - m_0 \rangle + \mathrm{tr}\left(\Sigma_0(A - I_d)\right) - \mathrm{tr}\left((\Sigma^{\frac{1}{2}}(\Sigma_0 - \Sigma_0 A - A\Sigma_0 + \Sigma_1)\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\right),$$

where $A = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}$.

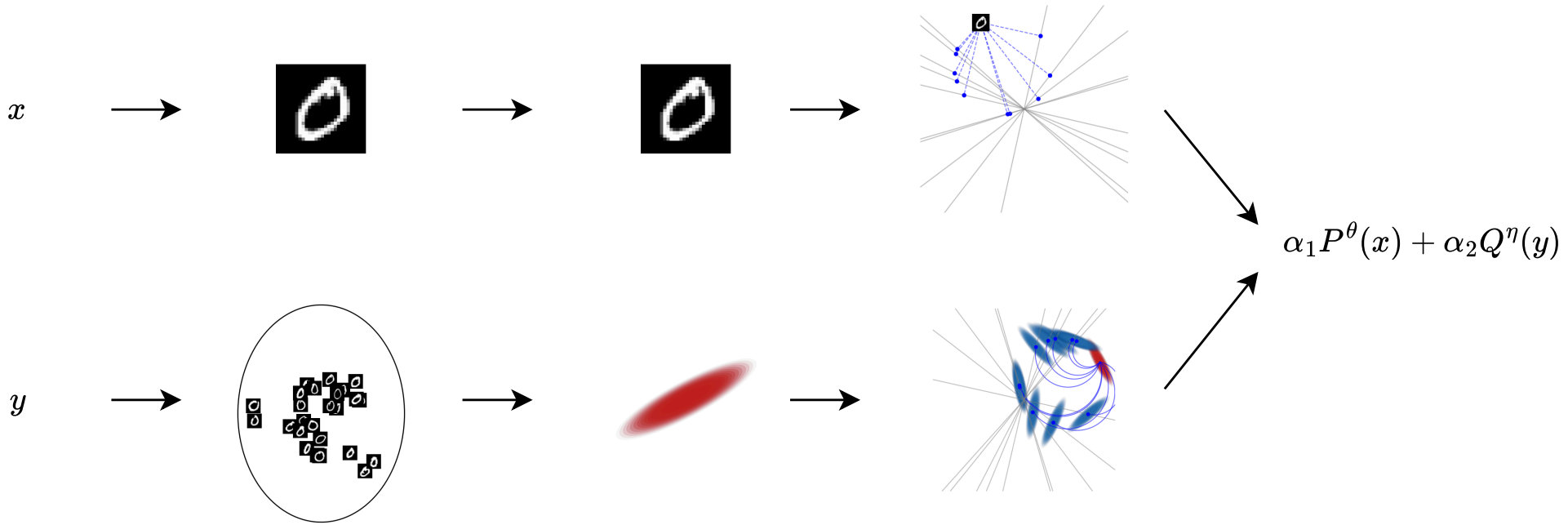
Slicing Datasets with Busemann on Gaussian

With Gaussian approximation:

- Define $\Xi(\mu) = \mathcal{N}\big(m(\mu), \Sigma(\mu)\big)$
- For all $y \in \{1, \dots, C\}$, $\varphi(y) = \frac{1}{n} \sum_{i=1}^{m} \delta_{x_i} \mathbb{1}_{\{y_i = y\}}$

$$Q^{\eta}(y) = B^{\eta}\big(\Xi(\varphi(y))\big),$$

with η a geodesic ray on $BW(\mathbb{R}^d)$



Computational Complexity: $O\big(LCd^3 + LnC(\log(nC) + d) + d^2Cn\big)$

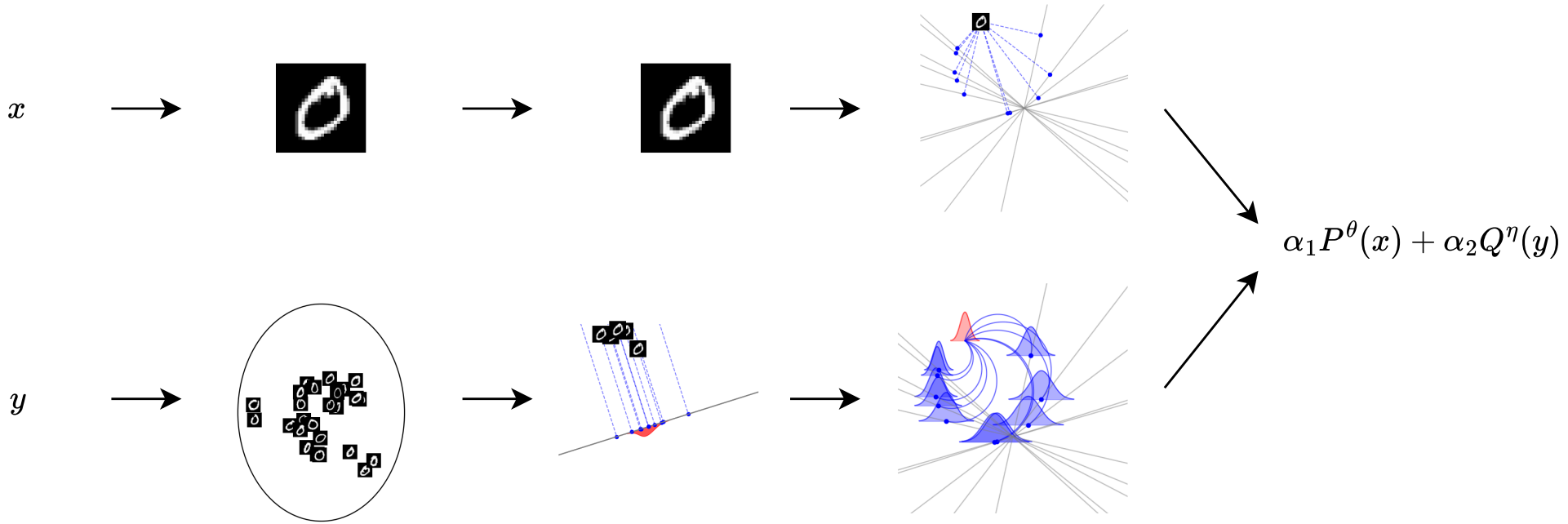
Slicing Datasets with Busemann in 1D

With 1D Projection:

- For all $y \in \{1, \dots, C\}$, $\varphi(y) = \frac{1}{n} \sum_{i=1}^{m} \delta_{x_i} \mathbb{1}_{\{y_i = y\}}$

$$Q^{\eta,\theta}(y) = B^{\eta}\big(P^{\theta}_{\#}\varphi(y)\big),$$

with η a geodesic ray on $\mathcal{P}_2(\mathbb{R})$



Computation Complexity: $O\big(LnC(\log(nC) + d)\big)$

Correlation vs OTDD

Goal: Measure correlation between sliced distances and OTDD

$\rightarrow$ Compare randomly sampled subdatasets + Spearman and Pearson correlations

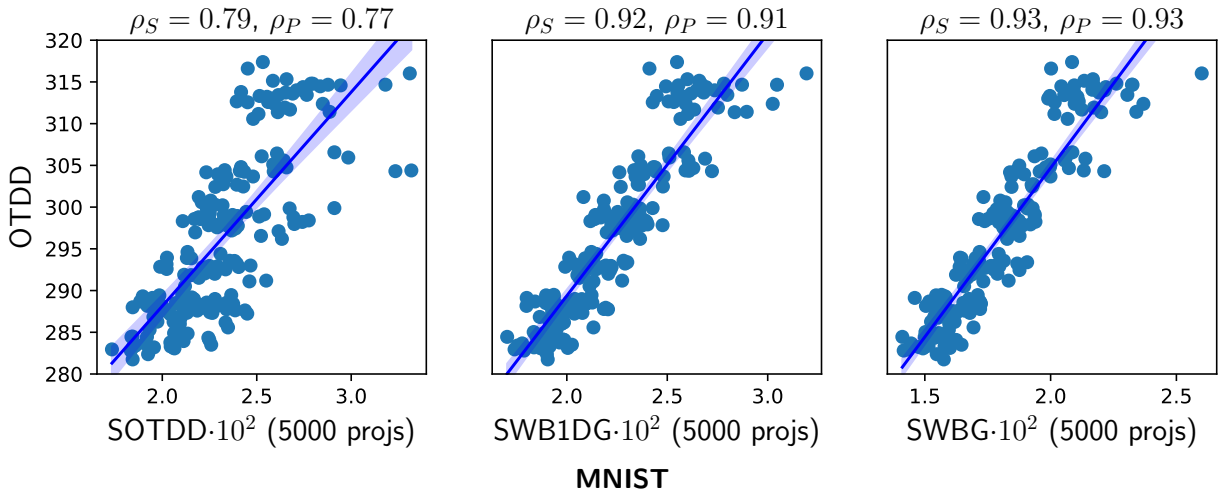


Table of Contents

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

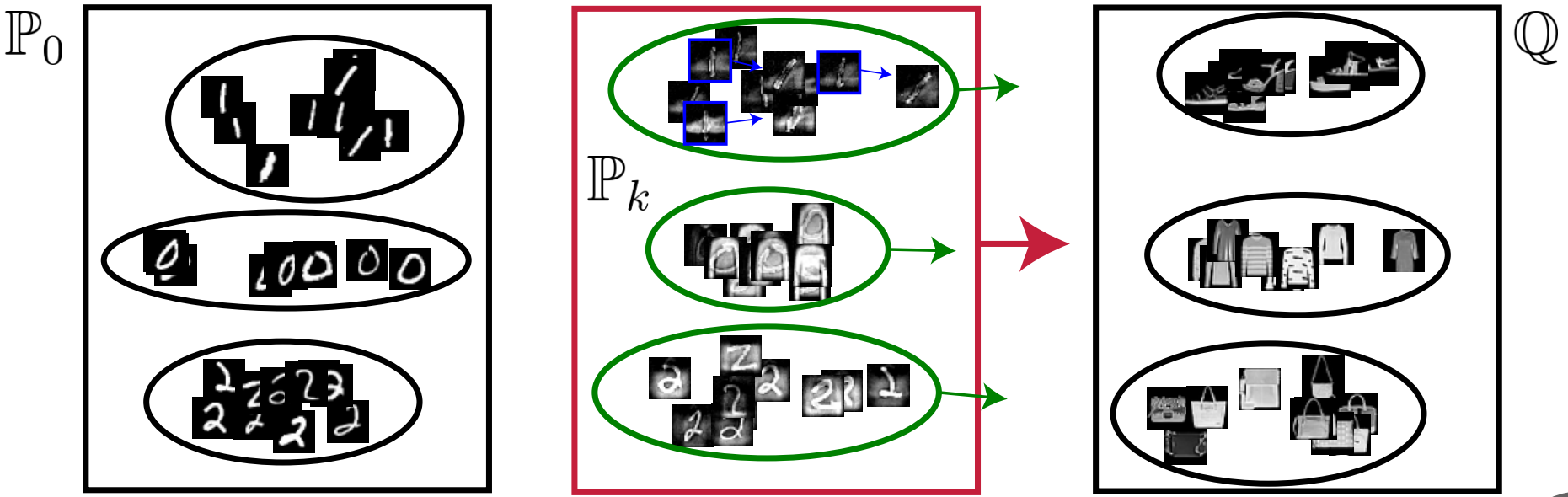
Goal: minimize $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$

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

$\nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)(\mu_k^{c,n})(x_{i,k}^c) = nC[\nabla F(\mathbf{x})]_{i,c}$ with $F(\mathbf{x}) = \mathbb{F}(\mathbb{P}_k)$, $\mathbf{x} = (x_{i,k}^c)_{i,c}$



Synthetic Data

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

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

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

Let $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^c}$, ν^c ring

$$\min_{\mathbb{P}} \ \mathbb{F}(\mathbb{P}) = D(\mathbb{P}, \mathbb{Q})$$

SOTDD SWB1DG SWBG

Synthetic Data

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

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

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

Let $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^c}$, ν^c ring

$$\min_{\mathbb{P}} \ \mathbb{F}(\mathbb{P}) = D(\mathbb{P}, \mathbb{Q})$$

SOTDD SWB1DG SWBG

Synthetic Data

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

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

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

Let $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^c}$, ν^c ring

$$\min_{\mathbb{P}} \ \mathbb{F}(\mathbb{P}) = D(\mathbb{P}, \mathbb{Q})$$

$k(x, y) = -\|x - y\|_2$

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

$K(\mu, \nu) = -\mathrm{SW}_2(\mu, \nu)$

Synthetic Data

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

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

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

Let $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^c}$, ν^c ring

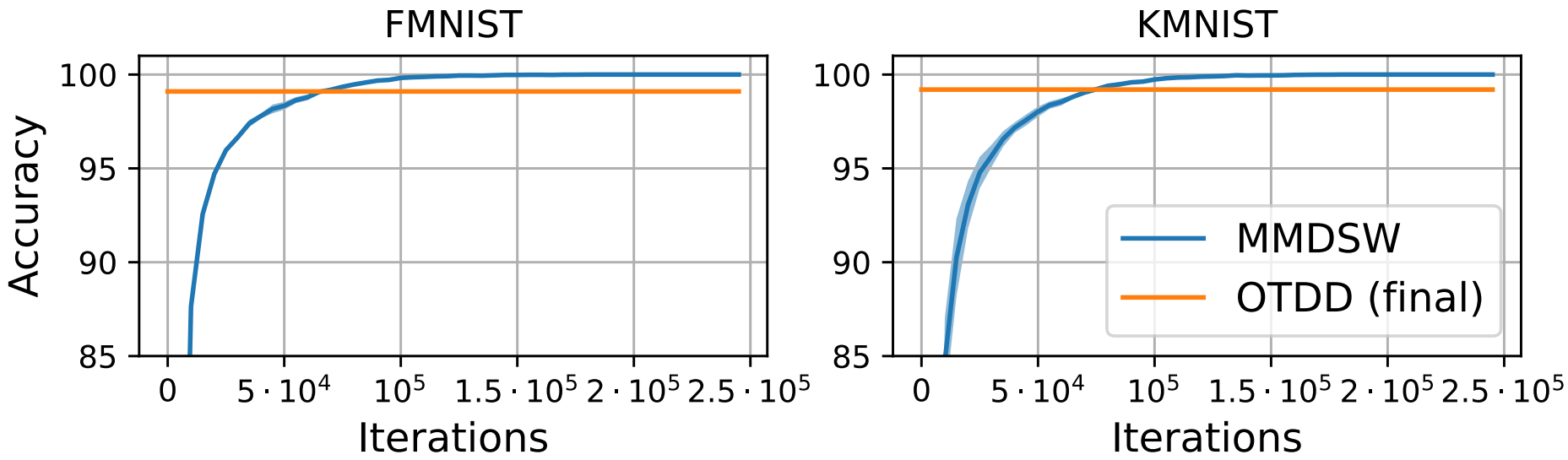
$$\min_{\mathbb{P}} \ \mathbb{F}(\mathbb{P}) = D(\mathbb{P}, \mathbb{Q})$$

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

"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)

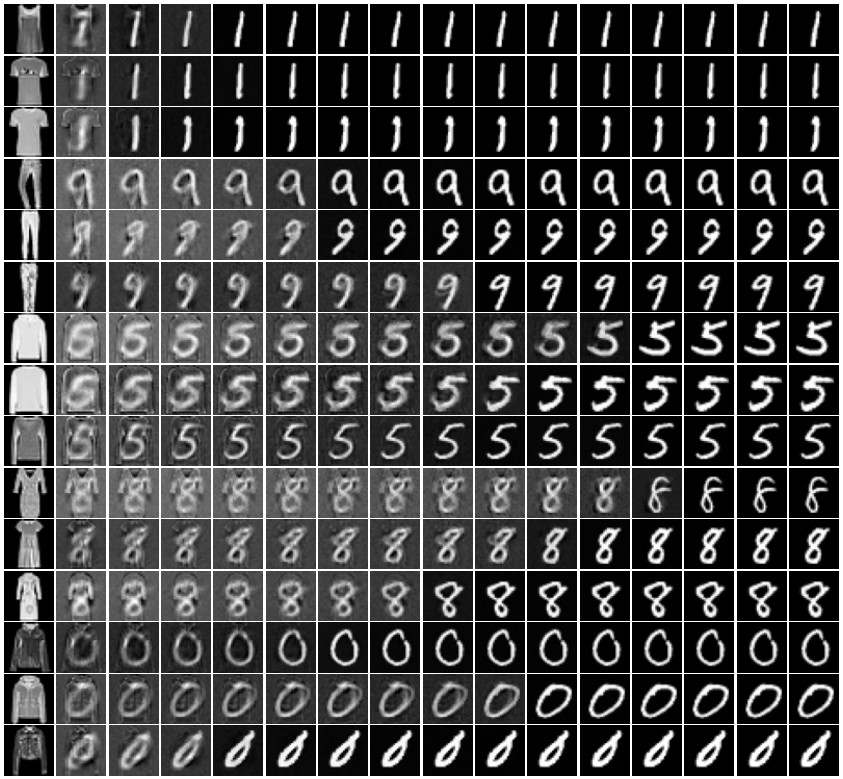


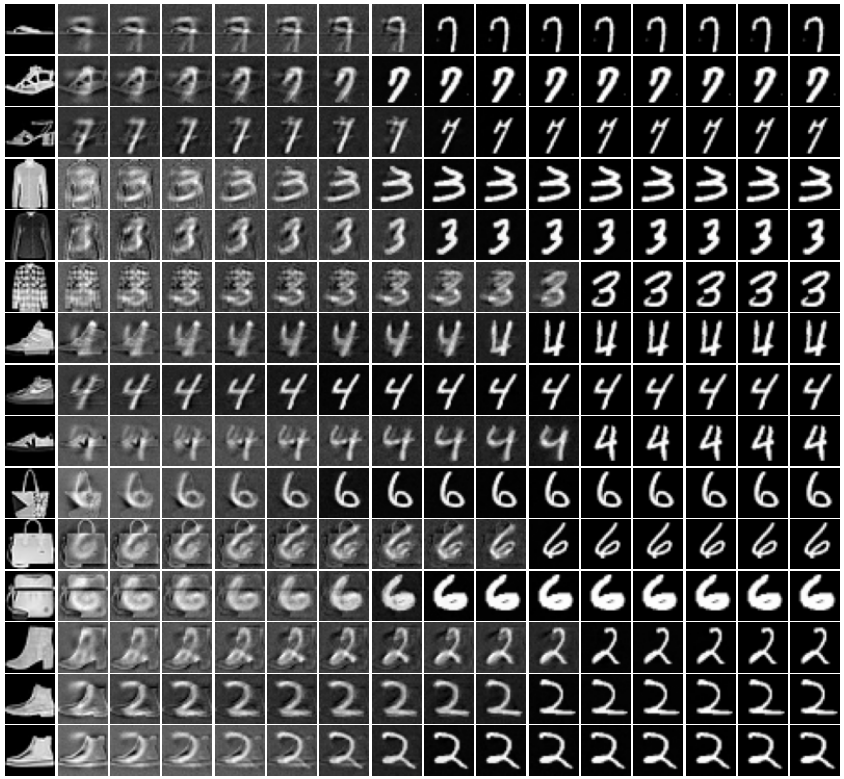
$\rightarrow$ reach 100% accuracy

"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:

- Efficient between Labeled Datasets
- Wasserstein over Wasserstein Gradient Flows
- Implementation on the MMD and Sliced Busemann Wasserstein
- 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:

- Efficient between Labeled Datasets
- Wasserstein over Wasserstein Gradient Flows
- Implementation on the MMD and Sliced Busemann Wasserstein
- 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!

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, Elsa Cazelles, Lucas Drumetz, and Nicolas Courty. Busemann Functions in the Wasserstein Space: Existence, Closed-Forms, and Applications to Slicing. *arXiv preprint arXiv:2510.04579*, 2025a.

References II

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

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.

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.

References IV

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.

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.

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