Linearization of the 2-Wasserstein space and stability of Optimal Transport maps

Quentin Mérigot[(1)], Alex Delalande[(1),(2)], Frédéric Chazal[(2)]

[(1)]Laboratoire de Mathématiques d'Orsay, Université Paris-Sud. [(2)]INRIA Saclay, DataShape team.

Overview

Wasserstein distances:

- Give a geometry to spaces of probability measures.
- Are defined using Optimal Transport (OT) theory.

Problem motivation:

- Wasserstein spaces are curved and Wasserstein distances are not Hilbertian: generic Machine Learning algorithms cannot readily work on measure data using the Wasserstein geometry.
- The comparison of k measures involve the resolution of $\frac{k(k-1)}{2}$ OT problems which can be prohibitive.

Our contributions:

- Propose a measure embedding into a Hilbert space that can be computed efficiently in practice.
- Show that this embedding induces a distance that is bi-Hölder equivalent to the 2-Wasserstein distance.
- Illustrate the behavior and applications of the embedding on toy examples.

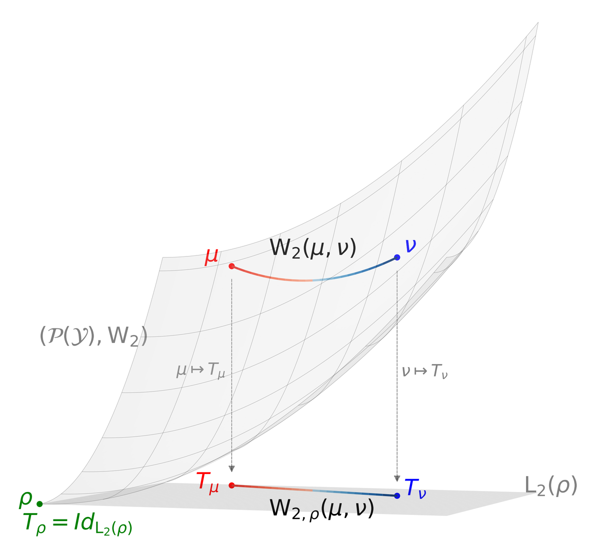
I. 2-Wasserstein distances and spaces

For $\mathcal{Y} \subset \mathbb{R}^d$ **compact and convex**, the 2-Wasserstein distance is defined for $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ as

$$\boxed{\mathrm{W}_2^2(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} \|y-y'\|^2 \mathrm{d}\gamma(y,y'),}$$

with $\Pi(\mu,\nu) := \{\gamma \in \mathcal{M}(\mathcal{Y}\times\mathcal{Y}) \mid \forall A \subset \mathcal{Y}, \gamma(A\times\mathcal{Y}) = \mu(A), \gamma(\mathcal{Y}\times A) = \nu(A)\}$ the matchings between μ and ν.

The **2-Wasserstein space** $(\mathcal{P}(\mathcal{Y}), \mathrm{W}_2)$ is a **curved metric space**: $\mathrm{W}_2(\mu,\nu)$ is actually the length of the shortest curve (geodesic) connecting μ and ν.



d defined on a $\mathcal{Z}\times\mathcal{Z}$ is an **Hilbertian distance** if there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{Z} \to \mathcal{H}$ s.t. $\forall z, z'$ in $\mathcal{Z}$, $d(z,z') = ||\phi(z) - \phi(z')||_{\mathcal{H}}$.

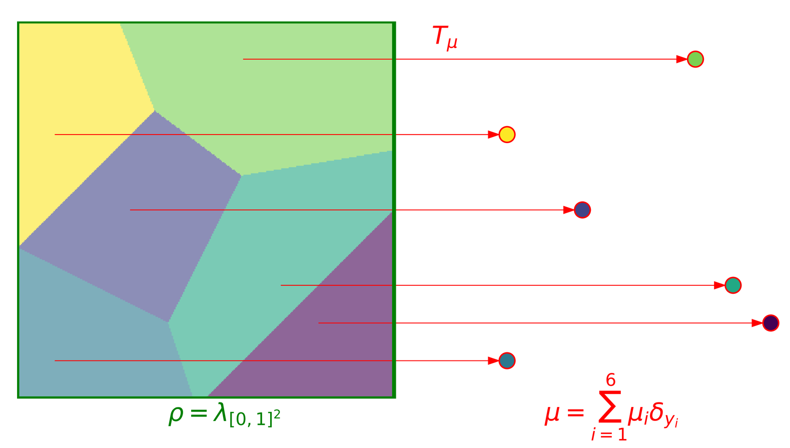
Wasserstein distances are not Hilbertian when $d \geq 2$.

II. Monge's embedding

For $\mathcal{X} \subset B(0, M_{\mathcal{X}}) \subset \mathbb{R}^d$ **compact and convex**, ρ a **fixed absolutely continuous measure** supported on $\mathcal{X}$, define **Monge's embedding of μ** as the mapping to the solution of Monge's OT problem between ρ and μ:

$$\boxed{\mu \mapsto \left(T_\mu := \arg\min_{T\in\mathcal{X}\to\mathcal{Y}} \int_{\mathcal{X}} \|x - T(x)\|^2 \rho(x)\mathrm{d}(x)\right),}$$

where T is a transport map between ρ and μ, i.e. $T_{\#}\rho = \mu$.



By Brenier, T_μ always exist and is uniquely defined as the gradient $T_\mu = \nabla\phi_\mu$ of a convex function $\phi_\mu : \mathcal{X} \to \mathbb{R}$.

Riemannian interpretation [4]:
$T_\mu \in \mathrm{L}^2(\rho)$ that includes the tangent space to $\mathcal{P}(\mathcal{Y})$ at ρ.
$\implies$ $\mu \mapsto T_\mu$ can be interpreted as the **inverse of an exponential map**.

III. Known properties of Monge's embedding

- The Monge's embedding is injective.
- Define

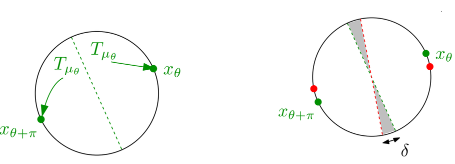
$$\boxed{\mathrm{W}_{2,\rho}(\mu,\nu) := \|T_\mu - T_\nu\|_{\mathrm{L}^2(\rho)},}$$

$\mathrm{W}_{2,\rho}$ **is an Hilbertian distance.**

- The Monge's embedding is reverse-Lipschitz:

$$\boxed{\mathrm{W}_2(\mu,\nu) \leq \mathrm{W}_{2,\rho}(\mu,\nu).}$$

- The Monge's embedding is in general not better than $\frac{1}{2}$-Hölder w.r.t. W_2.



$$||T_{\mu_\theta} - T_{\mu_{\theta+\delta}}||_{L^2(\rho)} \geq C\mathrm{W}_2(\mu_\theta, \mu_{\theta+\delta})^{1/2}$$

- **(Hölder-continuity near a regular measure)** For $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ such that T_μ is K-Lipschitz,

$$\boxed{\mathrm{W}_{2,\rho}(\mu,\nu) \leq 2\sqrt{M_{\mathcal{X}}K}\,\mathrm{W}_2(\mu,\nu)^{1/2}.}$$

- **(Dimension-dependent Hölder continuity)** (Corollary from [1]) For $\rho \equiv 1$ on $\mathcal{X}$ with unit volume, for any $\mu, \nu \in \mathcal{P}(\mathcal{Y})$,

$$\boxed{\mathrm{W}_{2,\rho}(\mu,\nu) \lesssim \mathrm{W}_2(\mu,\nu)^{\frac{1}{(d+2)2^{(d-1)}}}.}$$

($A \lesssim B$ means that $A \leq CB$ for a constant C depending only on d, the diameters of $\mathcal{X}$ and $\mathcal{Y}$, $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$)

IV. Dimension-independent Hölder-continuity of the Monge embedding

Theorem. For $\rho \equiv 1$ on $\mathcal{X}$ with unit volume, for any $\mu, \nu \in \mathcal{P}(\mathcal{Y})$,

$$\boxed{\mathrm{W}_{2,\rho}(\mu,\nu) \lesssim \mathrm{W}_2(\mu,\nu)^{2/15}.}$$

Sketch of Proof.

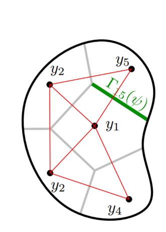
- Take μ and ν defined on $\{y_1, \dots, y_N\}$ (general case by a density argument) $\implies$ Semi-discrete OT.
- **Semi-discrete OT:**
  - **Kantorovich dual** (squared-Euclidean cost): $\boxed{(\mathrm{D}_\mu) = \min_{\psi:\mathcal{Y}\to\mathbb{R}} \int_{\mathcal{X}} \psi^* \mathrm{d}\rho + \int_{\mathcal{Y}} \psi \mathrm{d}\mu}$, with ψ^* the **Legendre transform** of ψ: $\forall x \in \mathcal{X}, \psi^*(x) = \max_{y\in\mathcal{Y}} \langle x|y\rangle - \psi(y)$. Actually, for ψ_μ solution of (D_μ): $\boxed{\phi_\mu = {\psi_\mu}^*}$ and $\boxed{T_\mu = \nabla\phi_\mu}$.
  - **Finitely-supported target:** when $\mu = \sum_{i=1}^N \mu_i \delta_{y_i}$, $\psi \equiv \boldsymbol{\psi} \in \mathbb{R}^N$ with $\boldsymbol{\psi}_i = \psi(y_i)$, finite-dimensional optimization problem:

$$\boxed{(\mathrm{D}) = \min_{\boldsymbol{\psi}\in\mathbb{R}^N} \mathcal{K}(\boldsymbol{\psi}) := \sum_{i=1}^N \int_{V_i(\boldsymbol{\psi})} (\langle x|y_i\rangle - \boldsymbol{\psi}_i)\mathrm{d}\rho(x) + \sum_{i=1}^N \mu_i \boldsymbol{\psi}_i,}$$

with $(V_i(\boldsymbol{\psi}))_i =$ **Laguerre cells**:
$\forall 1 \leq i \leq N, V_i(\boldsymbol{\psi}) = \{x \in \mathcal{X} \mid \forall j, \boldsymbol{\psi}_j \geq \boldsymbol{\psi}_i + \langle y_j - y_i | x\rangle\}$.

  - **Study of $\mathcal{K}$:** By [2], $\boxed{\nabla\mathcal{K}(\boldsymbol{\psi}) = \mu - G(\boldsymbol{\psi})}$ with $\boxed{G_i(\boldsymbol{\psi}) = \rho(V_i(\boldsymbol{\psi}))}$ and $\boxed{G(\boldsymbol{\psi}) = (G_i(\boldsymbol{\psi}))_{1\leq i\leq N} \in \mathbb{R}^N}$.
    Jacobian of G: Let $S_+ = \{\boldsymbol{\psi} \in \mathbb{R}^N \mid \forall i, G_i(\boldsymbol{\psi}) > 0\}$. On S_+, G is $\mathcal{C}^1$ and

$$\mathrm{D}G(\boldsymbol{\psi}) = \left(\frac{\partial G_i}{\partial \boldsymbol{\psi}_j}(\boldsymbol{\psi})\right)_{1\leq i,j\leq N} \text{ with } \begin{cases} \frac{\partial G_i}{\partial \boldsymbol{\psi}_j}(\boldsymbol{\psi}) = \frac{\mathrm{vol}^{d-1}(\Gamma_{ij})}{\|y_j - y_i\|} \text{ for } i \neq j \text{ and } \Gamma_{ij} = V_i(\boldsymbol{\psi}) \cap V_j(\boldsymbol{\psi}) \\ \frac{\partial G_i}{\partial \boldsymbol{\psi}_i}(\boldsymbol{\psi}) = -\sum_{j\neq i} \frac{\partial G_i}{\partial \boldsymbol{\psi}_j}(\boldsymbol{\psi}) \end{cases}$$



- **Discrete Poincaré-Wirtinger inequality** (from [3]): **local lower-bound of the smallest non-zero eigenvalue of** $\mathrm{D}G(\boldsymbol{\psi})$. For $\boldsymbol{\psi} \in S_+$ and $v \in \mathbb{R}^N$,

$$\boxed{\langle v^2 | G(\boldsymbol{\psi})\rangle_{\mathbb{R}^N} - \langle v | G(\boldsymbol{\psi})\rangle_{\mathbb{R}^N}^2 \lesssim \langle \mathrm{D}G(\boldsymbol{\psi})v|v\rangle_{\mathbb{R}^N}.}$$

- **Stability of dual potentials:**

$$\boxed{\|\psi_\nu - \psi_\mu\|_{L^2(\mu+\nu)} \lesssim \mathrm{W}_1(\mu,\nu)^{\frac{1}{3}}.}$$

- **Stability of OT maps:**

$$\boxed{\|\nabla\phi_\nu - \nabla\phi_\mu\|_{L^2(\rho)} \lesssim \mathrm{W}_1(\mu,\nu)^{\frac{2}{15}}.}$$

V. Numerical illustrations

Setting: $\mathcal{X} = \mathcal{Y} = [0,1]^2 \subset \mathbb{R}^d$. $\rho \equiv 1$ on $\mathcal{X}$.
Vectorization of Monge maps: with $\mathcal{X}_{s,t} = [\frac{s}{m}, \frac{s+1}{m}) \times [\frac{t}{m}, \frac{t+1}{m})$, define

$$\boxed{\boldsymbol{T}_{\boldsymbol{\mu}} := \left(\int_{\mathcal{X}_{s,t}} T_\mu \mathrm{d}\rho\right)_{1\leq s,t\leq m}.}$$

Distance approximation:

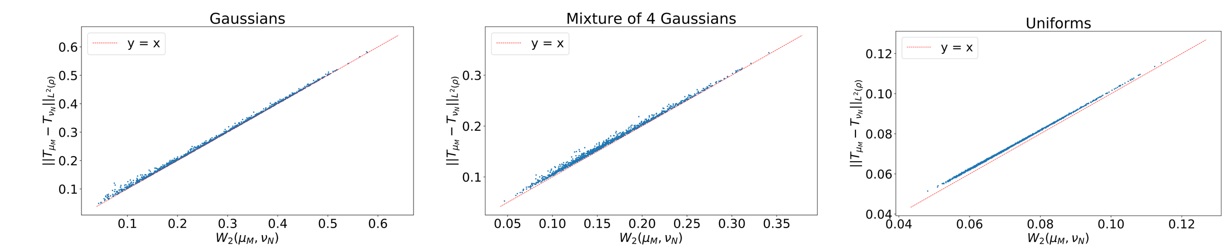


Figure: $\mathrm{W}_2(\mu,\nu)$ vs. $\|\boldsymbol{T}_{\boldsymbol{\mu}} - \boldsymbol{T}_{\boldsymbol{\nu}}\|_2$ $(\leq \|T_\mu - T_\nu\|_{\mathrm{L}^2(\rho)})$ between point clouds sampled from a Gaussian, a Mixture of 4 Gaussian and a Uniform distribution.

Sampling approximation:

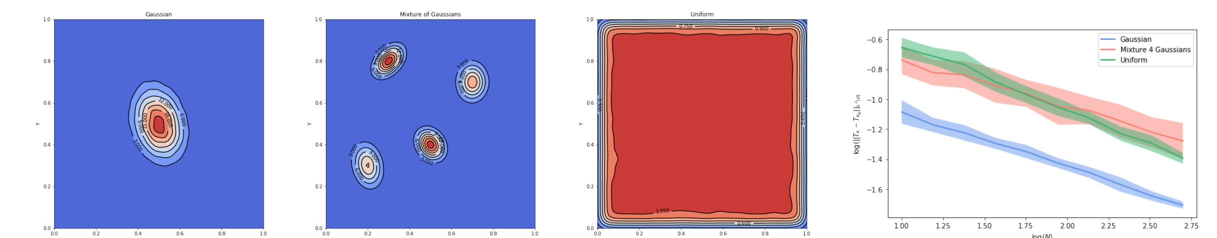


Figure: (1–3) Densities of the sampled targets. (4) Sampling distances $\|T_\mu - T_{\mu_N}\|_{\mathrm{L}^2(\rho)}$ as a function of N

Barycenter approximation:
Approximate $\mathrm{argmin}_\mu \sum_{s=1}^S \lambda_s \mathrm{W}_2^2(\mu, \mu_s)$ with $\mu = \left(\sum_{s=1}^S \lambda_s T_{\mu_s}\right)_{\#}\rho$.

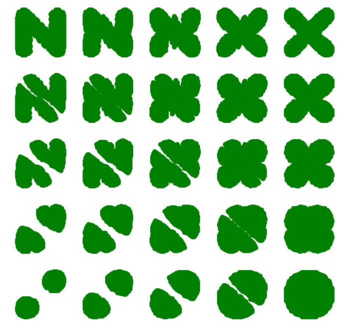


Figure: Barycenters of 4 point clouds. Weights $(\lambda_s)_s$ are bilinear w.r.t. the corners of the square.

Figure: Push-forwards of the 20 centroids after clustering with K-Means++ the Monge map embeddings of the MNIST training set.

References

[1] Robert J. Berman
Convergence rates for discretized Monge-Ampère equations and quantitative stability of Optimal Transport.
arXiv preprint 1803.00785, 2018.

[2] Kitagawa, Jun and Mérigot, Quentin and Thibert, Boris.
Convergence of a Newton algorithm for semi-discrete optimal transport.
In *Journal of the European Mathematical Society*, 2019.

[3] Eymard, Robert and Gallouët, Thierry and Herbin, Raphaèle.
Finite volume methods.
In *Handbook of numerical analysis*, Volume 7, pages 713–1018, Elsevier, 2000.

[4] Wang, Wei and Slepčev, Dejan and Basu, Saurav and Ozolek, John A. and Rohde, Gustavo K.
A Linear Optimal Transportation Framework for Quantifying and Visualizing Variations in Sets of Images.
In . J. Comput. Vision, 2013.