Quantitative Stability of Optimal Transport Maps and Linearization of the 2-Wasserstein Space

Alex Delalande[*, +]
with Quentin Mérigot[*] and Frédéric Chazal[+]

Laboratoire de Mathématiques d'Orsay, Université Paris-Sud[*]
DataShape team, INRIA Saclay[+]

AISTATS 2020

Introduction

- Numerous problems involve the **comparison of point clouds/probability measures** (e.g. in astronomy, shape recognition, image processing, generative modelling, large-scale learning, etc).

Figure 1: Point cloud comparison may appear in astronomy, 3D shape recognition or color transfer (from [Yu et al., 2012, Wu et al., 2015, Paty et al., 2019]).

Introduction

- Numerous problems involve the **comparison of point clouds/probability measures** (e.g. in astronomy, shape recognition, image processing, generative modelling, large-scale learning, etc).

- **Wasserstein distances**, provided by Optimal Transport (OT), are natural to perform these comparisons.

Introduction

- Numerous problems involve the **comparison of point clouds/probability measures** (e.g. in astronomy, shape recognition, image processing, generative modelling, large-scale learning, etc).

- **Wasserstein distances**, provided by Optimal Transport (OT), are natural to perform these comparisons.

Wasserstein distances.

- $\mathcal{X}$, $\mathcal{Y}$ compact and convex subsets of $\mathbb{R}^d$.
- α, β probability measures on $\mathcal{X}$, $\mathcal{Y}$ respectively.

$$\mathrm{W}_p^p(\alpha, \beta) := \min_{\pi} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} ||x - y||^p \mathrm{d}\pi(x, y) \mid \pi \in \mathcal{U}(\alpha, \beta) \right\},$$

where $\mathcal{U}(\alpha, \beta) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid (P_{\mathcal{X}})_{\#}\pi = \alpha \text{ and } (P_{\mathcal{Y}})_{\#}\pi = \beta\}$ with $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$.

Introduction

- Numerous problems involve the **comparison of point clouds/probability measures** (e.g. in astronomy, shape recognition, image processing, generative modelling, large-scale learning, etc).

- **Wasserstein distances**, provided by Optimal Transport (OT), are natural to perform these comparisons.

  Wasserstein distances.

  - $\mathcal{X}$, $\mathcal{Y}$ compact and convex subsets of $\mathbb{R}^d$.
  - α, β probability measures on $\mathcal{X}$, $\mathcal{Y}$ respectively.

  $$\mathrm{W}_p^p(\alpha, \beta) := \min_{\pi} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} ||x-y||^p \mathrm{d}\pi(x,y) \mid \pi \in \mathcal{U}(\alpha,\beta) \right\},$$

  where $\mathcal{U}(\alpha,\beta) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \mid (P_{\mathcal{X}})_{\#}\pi = \alpha \text{ and } (P_{\mathcal{Y}})_{\#}\pi = \beta\}$ with $P_{\mathcal{X}}(x,y) = x$ and $P_{\mathcal{Y}}(x,y) = y$.

  They enable notions such as:

  $\mathrm{W}_p($  ,  $) \leq \mathrm{W}_p($  ,  $)$

Introduction

- **In practice:** for $\alpha_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\beta_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$, computing $\mathrm{W}_p^p(\alpha_n, \beta_m)$ corresponds to solving a **LP problem**.

Introduction

- **In practice:** for $\alpha_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\beta_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, computing $\mathrm{W}_p^p(\alpha_n, \beta_m)$ corresponds to solving a **LP problem**.

  Several algorithms, e.g.

Algorithm	Complexity
Network Simplex	$O(n^3 \log(n)^2)$
Auction	$O(n^3)$
Sinkhorn (τ-*approximate OT*) [Peyré and Cuturi, 2019]	$O(n^2 \log(n)\tau^{-3})$

Introduction

- **In practice:** for $\alpha_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\beta_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$, computing $\mathrm{W}_p^p(\alpha_n, \beta_m)$ corresponds to solving a **LP problem**.

  Several algorithms, e.g.

Algorithm	Complexity
Network Simplex	$O(n^3 \log(n)^2)$
Auction	$O(n^3)$
Sinkhorn (τ-*approximate OT*) [Peyré and Cuturi, 2019]	$O(n^2 \log(n)\tau^{-3})$

  1 solving of OT $\implies$ high computational costs.

  To get the distance matrix of k point clouds, $O(k^2)$ OT problems to solve: can be **prohibitive** for large values of k.

Introduction

- **In practice:** solving numerous OT problems can be **computationally prohibitive**.

  Wasserstein spaces are **curved**, hence **non linear**.

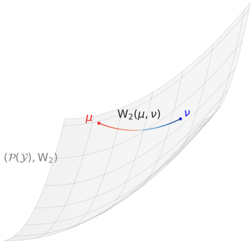


Figure 2: $(\mathcal{P}_2, \mathrm{W}_2)$ is curved.

No "closed-form" notions of sum or mean in $(\mathcal{P}_p, \mathrm{W}_p)$ (e.g. Wasserstein barycenters are defined as arg min's).

Introduction

- **In practice:** solving numerous OT problems can be **computationally prohibitive**.

  Wasserstein spaces are **curved**, hence **non linear**.

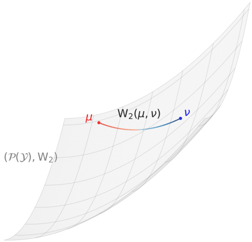


Figure 3: $(\mathcal{P}_2, \mathrm{W}_2)$ is curved.

No "closed-form" notions of sum or mean in $(\mathcal{P}_p, \mathrm{W}_p)$ (e.g. Wasserstein barycenters are defined as arg min's).
In ML, can be interesting to have an **Hilbertian structure**: impossible in Wasserstein spaces.

Introduction

- **In practice:** solving numerous OT problems can be **computationally prohibitive**. Wasserstein spaces are **curved** and **not Hilbertian**.

- **Here: we propose a measure embedding into a Hilbert space that conserves some of the (unregularized) Wasserstein geometry**.

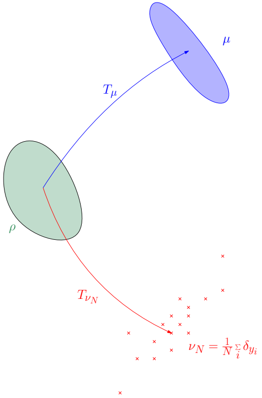
Proposed workaround: Monge embedding

Let ρ be a **fixed a.c. measure** on $\mathcal{X}$.

By [Brenier, 1991], for any $\mu \in \mathcal{P}(\mathcal{Y})$,

$$\min_{\pi} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} ||x-y||^2 \mathrm{d}\pi(x,y) \mid \pi \in \mathcal{U}(\rho,\mu) \right\}$$
$$=$$
$$\min_{T} \left\{ \int_{\mathcal{X}} ||x - T(x)||^2 \mathrm{d}\rho(x) \mid T : \mathcal{X} \to \mathcal{Y}, T_{\#}\rho = \mu \right\},$$

where $T_{\#}\rho$ is s.t. $\forall Y \subseteq \mathcal{Y}, \quad T_{\#}\rho(Y) = \rho(T^{-1}(Y))$.



Proposed workaround: Monge embedding

Let ρ be a **fixed a.c. measure** on $\mathcal{X}$.

By [Brenier, 1991], for any $\mu \in \mathcal{P}(\mathcal{Y})$,

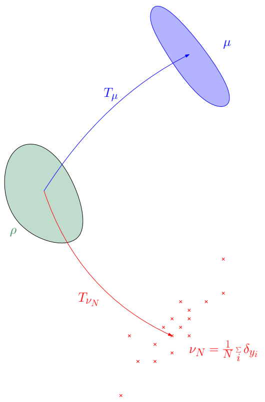
$$\min_{\pi} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} ||x-y||^2 \mathrm{d}\pi(x,y) \mid \pi \in \mathcal{U}(\rho,\mu) \right\}$$
$$=$$
$$\min_{T} \left\{ \int_{\mathcal{X}} ||x - T(x)||^2 \mathrm{d}\rho(x) \mid T : \mathcal{X} \to \mathcal{Y}, T_{\#}\rho = \mu \right\},$$

where $T_{\#}\rho$ is s.t. $\forall Y \subseteq \mathcal{Y}, \quad T_{\#}\rho(Y) = \rho(T^{-1}(Y))$.

A solution T_μ **always exists** and is uniquely defined as the gradient $T_\mu = \nabla\phi_\mu$ of a convex function ϕ_μ that minimizes the *Kantorovich functional*

$$\mathcal{K}(\phi_\mu) = \int_{\mathcal{X}} \phi_\mu \mathrm{d}\rho + \int_{\mathcal{Y}} \psi_\mu \mathrm{d}\mu,$$

where $\psi_\mu = \phi_\mu^*$ is the Legendre transform of ϕ_μ.



Proposed workaround: Monge embedding

Let ρ be a **fixed a.c. measure** on $\mathcal{X}$.

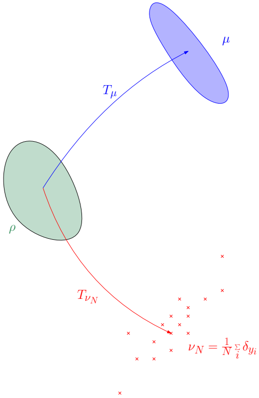
By [Brenier, 1991], for any $\mu \in \mathcal{P}(\mathcal{Y})$,

$$\boxed{\begin{gathered}\min_{\pi} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} ||x-y||^2 \mathrm{d}\pi(x,y) \mid \pi \in \mathcal{U}(\rho,\mu) \right\} \\ = \\ \min_{T} \left\{ \int_{\mathcal{X}} ||x - T(x)||^2 \mathrm{d}\rho(x) \mid T : \mathcal{X} \to \mathcal{Y}, T_{\#}\rho = \mu \right\}, \\ \text{where } T_{\#}\rho \text{ is s.t. } \forall Y \subseteq \mathcal{Y}, \quad T_{\#}\rho(Y) = \rho(T^{-1}(Y)).\end{gathered}}$$

A solution T_μ **always exists** and is uniquely defined as the gradient $T_\mu = \nabla\phi_\mu$ of a convex function ϕ_μ that minimizes the *Kantorovich functional*

$$\mathcal{K}(\phi_\mu) = \int_{\mathcal{X}} \phi_\mu \mathrm{d}\rho + \int_{\mathcal{Y}} \psi_\mu \mathrm{d}\mu,$$

where $\psi_\mu = \phi_\mu^*$ is the Legendre transform of ϕ_μ.



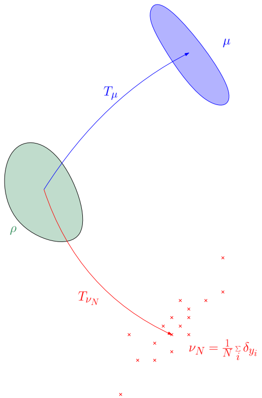
Definition (Monge embedding)

The **Monge embedding** is the mapping

$$\mathcal{P}(\mathcal{Y}) \to \mathrm{L}^2(\rho, \mathbb{R}^d),$$
$$\mu \mapsto T_\mu.$$

Proposed workaround: Monge embedding

Let ρ be a **fixed a.c. measure** on $\mathcal{X}$.



Definition (Monge embedding)

For all $\mu \in \mathcal{P}(\mathcal{Y})$, denote T_μ the solution of Monge's OT problem between ρ and μ for the squarred Euclidean cost.

The **Monge embedding** is the mapping

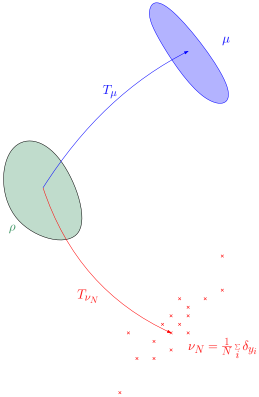
$$\mathcal{P}(\mathcal{Y}) \to \mathrm{L}^2(\rho, \mathbb{R}^d),$$
$$\mu \mapsto T_\mu.$$

Remarks

- When μ is discrete: semi-discrete OT. Efficiently solved in low dimensions with second-order methods [Kitagawa et al., 2019] and in higher dimensions with stochastic optimization methods [Genevay et al., 2016].
- $\mathrm{L}^2(\rho)$-Distance matrix of k point-clouds $\implies$ k OT problems to solve + $O(k^2)$ distance computations on Hilbertian/Euclidean data.

Proposed workaround: Monge embedding

Let ρ be a **fixed a.c. measure** on $\mathcal{X}$.



Definition (Monge embedding)

For all $\mu \in \mathcal{P}(\mathcal{Y})$, denote T_μ the solution of Monge's OT problem between ρ and μ for the squarred Euclidean cost.

The **Monge embedding** is the mapping

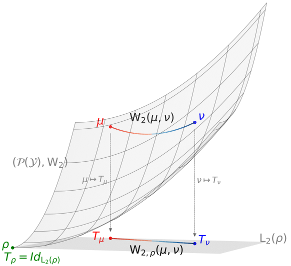
$$\mathcal{P}(\mathcal{Y}) \to \mathrm{L}^2(\rho, \mathbb{R}^d),$$
$$\mu \mapsto T_\mu.$$

Main result: bi-Hölder behavior of $\mu \mapsto \mathbf{T}_\mu$

$$\forall \mu, \nu \in \mathcal{P}(\mathcal{Y}), \quad \boxed{\mathrm{W}_2(\mu,\nu) \leq \| T_\mu - T_\nu \|_{\mathrm{L}^2(\rho)} \leq C_{d,\mathcal{X},\mathcal{Y}} \mathrm{W}_2(\mu,\nu)^{2/15}.}$$

Geometric interpretation of the Monge embedding

Monge embedding as logarithm map (similar construction and interpretation in the *Linear Optimal Transportation Framework* of [Wang et al., 2013])



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

	Riemannian geometry	Optimal transport
Point	$x \in M$	$\mu \in \mathcal{P}_2(\mathbb{R}^d)$
Geodesic distance	$d_g(x, y)$	$\mathrm{W}_2(\mu, \nu)$
Tangent space	$\mathcal{T}_\rho M$	$\mathcal{T}_\rho \mathcal{P}_2(\mathbb{R}^d) \approx \mathrm{L}^2(\rho)$
Inverse exponential map	$\exp^{-1}(x) \in \mathcal{T}_\rho M$	$T_\mu \in \mathrm{L}^2(\rho)$
Distance in tangent space	$\left\|\exp^{-1}(x) - \exp^{-1}(y)\right\|_{g(\rho)}$	$\left\|T_\mu - T_\nu\right\|_{\mathrm{L}^2(\rho)}$

What amount of the Wasserstein geometry is preserved by the embedding $\mu \mapsto T_\mu$?

Immediate properties of the Monge embedding

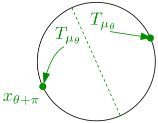
$\mu \mapsto T_\mu$ is discriminative

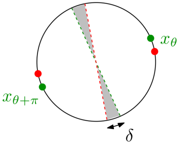
- $\mu \mapsto T_\mu$ is **injective**
  (by definition of the push-forward operator).

- $\mu \mapsto T_\mu$ is **reverse-Lipschitz**: $\boxed{||T_\mu - T_\nu||_{\mathrm{L}^2(\rho)} \geq \mathrm{W}_2(\mu, \nu)}$
  ($\gamma := (T_\mu, T_\nu)_\# \rho$ defines an admissible coupling between μ and ν).

$\mu \mapsto T_\mu$ is **not better than $\frac{1}{2}$-Hölder**

- Take $\rho := \frac{1}{\pi}\mathrm{Leb}_{B(0,1)}$ on $\mathbb{R}^2$ and $\mu_\theta := \frac{\delta_{x_\theta} + \delta_{x_{\theta+\pi}}}{2}$ with $x_\theta = (\cos(\theta), \sin(\theta))$. Then $||T_{\mu_\theta} - T_{\mu_{\theta+\delta}}||^2_{\mathrm{L}^2(\rho)} \geq C\delta$ while $\mathrm{W}_2(\mu_\theta, \mu_{\theta+\delta}) \leq C\delta$.





Immediate properties of the Monge embedding

Theorem ($\frac{1}{2}$-Hölder continuity near a regular measure, similar to a result of [Gigli, 2011])

Let $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ and assume that T_μ is K-Lipschitz. Then,

$$\boxed{\|T_\mu - T_\nu\|_{\mathrm{L}^2(\rho)} \leq 2\sqrt{M_{\mathcal{X}} K}\mathrm{W}_1(\mu, \nu)^{1/2},}$$

where $M_{\mathcal{X}}$ is s.t. $\mathcal{X} \subset B(0, M_{\mathcal{X}})$.

Very strong hypothesis on μ.

Theorem (General Hölder-continuity as a corollary of [Berman, 2018], Proposition 3.4)

If $\rho \equiv 1$ on $\mathcal{X}$ with $|\mathcal{X}| = 1$, then for any measures μ and ν in $\mathcal{P}(\mathcal{Y})$,

$$\boxed{\|T_\mu - T_\nu\|_{\mathrm{L}^2(\rho)} \leq C_{d,\mathcal{X},\mathcal{Y}} \mathrm{W}_1(\mu, \nu)^{\frac{1}{(d+2)2^{(d-1)}}}.}$$

High dependence on the ambient dimension. Optimal exponent?

Stability of the Monge embedding

Theorem (Dimension-independent Hölder continuity)

If $\rho \equiv 1$ on $\mathcal{X}$ **convex** with $|\mathcal{X}| = 1$, then for any measures μ and ν in $\mathcal{P}(\mathcal{Y})$,

$$\boxed{\|T_\nu - T_\mu\|_{\mathrm{L}^2(\mathcal{X})} \leq C_{d,\mathcal{X},\mathcal{Y}} \mathrm{W}_1(\mu,\nu)^{2/15}.}$$

Remarks

- Hölder exponent independent of the ambient dimension.
- No hypothesis on μ and ν except that they are compactly supported.
- No regularization.
- **Optimality:** best exponent belongs to $[\frac{2}{15}, \frac{1}{2}]$.

Stability of the Monge embedding

Theorem (Dimension-independent Hölder continuity)

If $\rho \equiv 1$ on $\mathcal{X}$ with $|\mathcal{X}| = 1$, then for any measures μ and ν in $\mathcal{P}(\mathcal{Y})$,

$$\boxed{\|T_\nu - T_\mu\|_{\mathrm{L}^2(\mathcal{X})} \leq C_{d,\mathcal{X},\mathcal{Y}} \mathrm{W}_1(\mu, \nu)^{2/15}.}$$

Remarks

- Hölder exponent independent of the ambient dimension.
- No hypothesis on μ and ν except that they are compactly supported.
- No regularization.
- **Optimality:** best exponent belongs to $[\frac{2}{15}, \frac{1}{2}]$.
- **Proof ingredients:**
  - Discrete μ and ν (general case by density).

Stability of the Monge embedding

Theorem (Dimension-independent Hölder continuity)

If $\rho \equiv 1$ on $\mathcal{X}$ with $|\mathcal{X}| = 1$, then for any measures μ and ν in $\mathcal{P}(\mathcal{Y})$,

$$\boxed{\|T_\nu - T_\mu\|_{\mathrm{L}^2(\mathcal{X})} \leq C_{d,\mathcal{X},\mathcal{Y}} \mathrm{W}_1(\mu,\nu)^{2/15}.}$$

Remarks

- Hölder exponent independent of the ambient dimension.
- No hypothesis on μ and ν except that they are compactly supported.
- No regularization.
- **Optimality:** best exponent belongs to $[\frac{2}{15}, \frac{1}{2}]$.
- **Proof ingredients:**
  - Discrete μ and ν (general case by density).
  - A Discrete Poincaré-Wirtinger inequality $\implies$ local estimate of the strong convexity of the Kantorovich functional (*non trivial because of the lack of regularization*). Gives:

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

Stability of the Monge embedding

Theorem (Dimension-independent Hölder continuity)

If $\rho \equiv 1$ on $\mathcal{X}$ with $|\mathcal{X}| = 1$, then for any measures μ and ν in $\mathcal{P}(\mathcal{Y})$,

$$\boxed{\|T_\nu - T_\mu\|_{\mathrm{L}^2(\mathcal{X})} \leq C_{d,\mathcal{X},\mathcal{Y}} \mathrm{W}_1(\mu,\nu)^{2/15}.}$$

Remarks

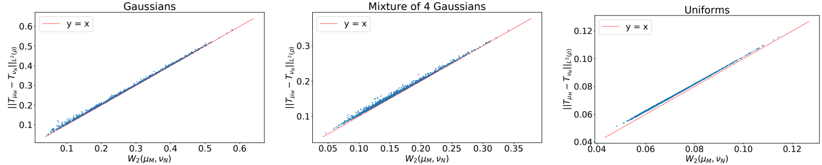
- Hölder exponent independent of the ambient dimension.
- No hypothesis on μ and ν except that they are compactly supported.
- No regularization.
- **Optimality:** best exponent belongs to $[\frac{2}{15}, \frac{1}{2}]$.
- **Proof ingredients:**
  - Discrete μ and ν (general case by density).
  - A Discrete Poincaré-Wirtinger inequality $\implies$ local estimate of the strong convexity of the Kantorovich functional (*non trivial because of the lack of regularization*).
  - An inverse Poincaré-Wirtinger inequality gives:

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

Numerical illustrations

- Let $\mathcal{X} = \mathcal{Y} = [0,1]^2$, $\rho \equiv 1$ on $\mathcal{X}$.
- **Project $\mathrm{L}^2(\rho, \mathbb{R}^2)$ onto a finite dimensional space.**

Distance approximation: $\mathrm{W}_{2,\rho}$ vs. W_2

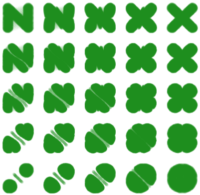


$\mathrm{W}_{2,\rho}$ vs. W_2 between point clouds sampled from a Gaussian, a Mixture of 4 Gaussian and a uniform distribution.

Numerical illustrations

Wasserstein barycenter [Agueh and Carlier, 2011] approximation

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



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

Push-forwards of the 20 centroids after clustering of the Monge embeddings of the MNIST training set.

Conclusion

$\mu \mapsto \mathbf{T}_\mu$ is an **Hilbert space embedding** that:

- Linearizes to some extent the 2-Wasserstein space.
- Is bi-Hölder continuous w.r.t. W_2.
- Allows for the direct use of generic ML algorithms on measure data thanks to the linearity and Hilbertian structure of $\mathrm{L}^2(\rho)$.

Future work:

- Not compactly supported target measures.
- More general source measures.
- Statistical properties: concentration and sample complexity of the defined distances.
- Applications: compact encoding of $T_\mu \in \mathrm{L}^2(\rho)$ that scales well to high dimensions.

References I

Agueh, M. and Carlier, G. (2011).
Barycenters in the Wasserstein space.
SIAM Journal on Mathematical Analysis, 43(2):904–924.

Berman, R. J. (2018).
Convergence rates for discretized monge-ampère equations and quantitative stability of optimal transport.
arXiv preprint 1803.00785.

Brenier, Y. (1991).
Polar factorization and monotone rearrangement of vector-valued functions.
Communications on Pure and Applied Mathematics, 44(4):375–417.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016).
Stochastic optimization for large-scale optimal transport.
In *Advances in neural information processing systems*, pages 3440–3448.

Gigli, N. (2011).
On hölder continuity-in-time of the optimal transport map towards measures along a curve.
Proceedings of the Edinburgh Mathematical Society, 54(2):401–409.

References II

Kitagawa, J., Mérigot, Q., and Thibert, B. (2019).
Convergence of a newton algorithm for semi-discrete optimal transport.
Journal of the European Mathematical Society.

Paty, F.-P., d'Aspremont, A., and Cuturi, M. (2019).
Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport.

Peyré, G. and Cuturi, M. (2019).
Computational optimal transport.
Foundations and Trends in Machine Learning, 11(5-6):355–607.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. (2013).
A linear optimal transportation framework for quantifying and visualizing variations in sets of images.
Int. J. Comput. Vision, 101(2):254–269.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. (2015).
3d shapenets: A deep representation for volumetric shapes.
In *CVPR*, pages 1912–1920. IEEE Computer Society.

References III

Yu, L., Efstathiou, K., Isenberg, P., and Isenberg, T. (2012).
Efficient structure-aware selection techniques for 3d point cloud visualizations with 2dof input.
IEEE Transactions on Visualization and Computer Graphics, 18(12):2245–2254.