

---

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

---

**Quentin Mérigot**  
Université Paris-Sud

**Alex Delalande**  
Université Paris-Sud & Inria Saclay

**Frédéric Chazal**  
Inria Saclay

## Abstract

This work studies an explicit embedding of the set of probability measures into a Hilbert space, defined using optimal transport maps from a reference probability density. This embedding linearizes to some extent the 2-Wasserstein space and is shown to be bi-Hölder continuous. It enables the direct use of generic supervised and unsupervised learning algorithms on measure data consistently w.r.t. the Wasserstein geometry.

## 1 INTRODUCTION

Numerous problems involve the comparison of point clouds, i.e. sets of points that lie in a metric space and for which the spatial distribution is of interest. Seeing the point clouds as discrete probability measures in a metric space, it is natural to compare them with Wasserstein distances defined by the optimal transport theory (Villani, 2003). These distances have indeed been successfully used in a variety of applications in machine learning (Canas & Rosasco, 2012; Arjovsky et al., 2017; Gordaliza et al., 2019; Genevay et al., 2018; Flamary et al., 2018; Alaux et al., 2018) and in statistics (Weed & Berthet, 2019; Cazelles, Seguy, Bigot, Cuturi, & Papadakis, 2017; Bigot, Cazelles, & Papadakis, 2019; Ramdas, Garcia, & Cuturi, 2015). In the discrete setting, many efficient algorithms have been proposed to compute or approximate the Wasserstein distances, such as Sinkhorn-Knopp’s algorithm – see (Peyré & Cuturi, 2019) and references therein. However efficient these algorithms are, they still represent high computational costs when dealing with large databases of point clouds and they do not allow for the direct use of machine learning algorithms based on

the Wasserstein geometry. In this work, we leverage the semi-discrete formulation of OT to build explicit embeddings of point clouds in  $\mathbb{R}^d$  (seen as probability measures) into a Hilbert space. This linear embedding allows one to directly apply supervised and unsupervised learning methods on point clouds datasets consistently with the Wasserstein geometry, thus alleviating the non-Hilbertian nature of Wasserstein spaces in dimensions greater than 2 (see Section 8.3 in (Peyré & Cuturi, 2019)).

### 1.1 Optimal Transport and Monge Maps

Let  $\mathcal{X}, \mathcal{Y}$  be two compact and convex subsets of  $\mathbb{R}^d$ . Let  $\rho$  be a probability density on  $\mathcal{X}$  and  $\mu$  be a probability measure on  $\mathcal{Y}$ . We consider the squared Euclidean cost  $c(x, y) := \|x - y\|_2^2$  for all  $x, y \in \mathbb{R}^d$ . Monge’s optimal transport problem consists in minimizing the transport cost over all transport maps between  $\rho$  and  $\mu$ , that is

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \rho(x) d(x) \mid T : \mathcal{X} \rightarrow \mathcal{Y}, T_{\#}\rho = \mu \right\}, \quad (1)$$

where  $T_{\#}\rho$  is the *pushforward measure*, defined by

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

By the work of Brenier (Brenier, 1991), this problem admits a solution that is uniquely defined as the gradient  $T = \nabla\phi$  of a convex function  $\phi$  on  $\mathcal{X}$  referred to in what follows as a *Brenier potential*. Here, we will refer to the map  $T$  as the *Monge map*. In this work, the source probability density  $\rho \in \mathcal{P}(\mathcal{X})$  is fixed once for all.

**Definition 1.1** (Monge embedding). Given any probability measure  $\mu$  on  $\mathcal{Y}$ , we denote  $T_{\mu}$  the solution of the optimal transport problem (1) between  $\rho$  and  $\mu$ . We call *Monge embedding* the mapping

$$\begin{aligned} \mathcal{P}(\mathcal{Y}) &\rightarrow L^2(\rho, \mathbb{R}^d) \\ \mu &\mapsto T_{\mu}, \end{aligned} \quad (2)$$

where  $\mathcal{P}(\mathcal{Y})$  is the set of probability measures over  $\mathcal{Y}$ .

An attractive feature of the Monge embedding is that the map  $T_{\mu}$  can be efficiently computed when  $\mu$  is

finitely supported on  $\mathbb{R}^2$  or  $\mathbb{R}^3$  (Kitagawa, Mérigot, & Thibert, 2019). In higher dimension, it can also be estimated using stochastic optimization methods (Genevay, Cuturi, Peyré, & Bach, 2016). One could also approximate the source measure with a discrete measure and resort to map estimation methods for discrete OT (Perrot et al., 2016; Flamary et al., 2019; Paty et al., 2019).

**Notations.** Given a measure  $\mu$ , we denote  $T_\mu$  the Monge map,  $\phi_\mu$  a (convex) Brenier potential so that  $T_\mu = \nabla \phi_\mu$ , and  $\psi_\mu$  on  $\mathcal{Y}$  the Legendre transform of  $\phi_\mu$ :

$$\psi_\mu(y) = \max_{x \in \mathcal{X}} \langle x | y \rangle - \phi_\mu(x). \quad (3)$$

**Remark 1.1** (Uniqueness and Estimates). The two potentials  $(\phi_\mu, \psi_\mu)$  are closely related to the *Kantorovich potentials* associated to the optimal transport problem (1). In our setting, where the support of  $\rho$  is the whole domain, these potentials are unique up to addition of a constant (Santambrogio, 2015, Proposition 7.18) that we fix by enforcing  $\int \psi_\mu d\mu = 0$ .

## 1.2 Contributions

Our main interest in this work is the regularity properties of the Monge embedding (2), or equivalently the stability of the optimal transport maps in terms of the target measure. Our main theorem shows that the Monge map is a bi-Hölder embedding of  $\mathcal{P}(\mathcal{Y})$  endowed with the Wasserstein distance  $W_p$  (defined in equation (5)) into the Hilbert space  $L^2(\rho, \mathbb{R}^d)$ . More importantly, we show that the Hölder exponent does not depend on the ambient dimension  $d$ .

**Theorem** (Theorem 3.1). *Let  $\rho$  be the Lebesgue measure on a compact convex subset  $\mathcal{X}$  of  $\mathbb{R}^d$  with unit volume, and let  $\mathcal{Y}$  be a compact convex set. Then, for all  $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ , and all  $p \geq 1$ ,*

$$W_2(\mu, \nu) \leq \|T_\mu - T_\nu\|_{L^2(\rho)} \leq CW_p(\mu, \nu)^{\frac{2}{15}},$$

where the constant  $C$  depends on  $d$ ,  $\mathcal{X}$  and  $\mathcal{Y}$ .

The upper bound of this theorem should be compared to Theorem 2.2 (similar to a result of Ambrosio reported in (Gigli, 2011)), which shows a  $\frac{1}{2}$ -Hölder behaviour under a very strong regularity assumption on  $T_\mu$ , and to Corollary 2.4 (from Berman, see(Berman, 2018)), which holds without assumption on  $\mu, \nu$ , but whose exponent scales exponentially badly with the dimension  $d$ . We conclude the article by illustrations of the behavior of this embedding, and we showcase a few applications.

**Remark 1.2.** Similarly to (Wang et al., 2013, Eq. (3)) or (Ambrosio, Gigli, & Savaré, 2008, §10.2), one can define a distance by the formula

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

and our main results reads as a bi-Hölder equivalence between this distance and the 2-Wasserstein distance:

$$W_2(\mu, \nu) \leq W_{2,\rho}(\mu, \nu) \leq CW_2(\mu, \nu)^{\frac{2}{15}}. \quad (4)$$

## 1.3 Related Work in Statistics and Learning

The same construction (2) was introduced in (Wang et al., 2013) in the context of pattern recognition in images, where the problem of computing a distance matrix based on transportation metrics over a possibly large dataset of images is tackled. The approach proposed in (Wang et al., 2013) computes a reference image as a mean image (for the 2-Wasserstein distance) of the whole dataset and then computes the OT maps between this reference image  $\rho$  and each image  $\mu_i$  of the training set. Distances between images are then defined based on Euclidean distances between these maps.

The geometric idea comes from a Riemannian interpretation of the Wasserstein geometry (Otto, 2001; Ambrosio et al., 2008). In this interpretation, the tangent space to  $\mathcal{P}(\mathbb{R}^d)$  at  $\rho$  is included in  $L^2(\rho, \mathbb{R}^d)$ . The optimal transport map  $T_{\mu_i}$  between  $\rho$  and  $\mu_i$  can then be regarded as the vector in the tangent space at  $\rho$  which supports the Wasserstein geodesic from  $\rho$  to  $\mu_i$ . Thus Monge's embedding sends any probability measure  $\mu_i$  in the (curved) manifold  $\mathcal{P}(\mathbb{R}^d)$  to a vector  $T_{\mu_i}$  belonging to the linear space  $L^2(\rho, \mathbb{R}^d)$ , which retains some of the geometry of the space. In the Riemannian language again, the map  $\mu \mapsto T_\mu$  would be called a *logarithm*, i.e. the inverse of the Riemannian exponential map. This establishes a connection between this idea and similar strategies used to extend statistical inference notions (such as PCA) on manifold-valued data, e.g. (Fletcher et al., 2004; Cazelles et al., 2017).

The work in (Chernozhukov et al., 2017) also proposes to use OT maps in a statistical context to overcome the lack of a canonical ordering in  $\mathbb{R}^d$  for  $d > 1$ . Notions of vector-quantile, vector-ranks and depth are defined based on the transport maps (and their inverses) between a reference measure defined as the uniform distribution on the unit hyperball and the  $d$ -dimensional samples of interest.

Monge maps are also studied in (Hütter & Rigollet, 2019) where an estimator for such maps between population distributions is proposed when only samples from the distributions of interest are available. Minimax estimation rates for (very) smooth transport maps in general dimension are given and the proposed estimator is shown to achieve near minimax optimality.

## 2 KNOWN PROPERTIES OF THE MONGE EMBEDDING

From now on, we fix two compact convex subsets  $\mathcal{X}, \mathcal{Y}$  of  $\mathbb{R}^d$ , and we fix once and for all a probability density  $\rho$  on  $\mathcal{X}$ . We also denote  $M_{\mathcal{X}} \geq 0$  the smallest positive real such that  $\mathcal{X} \subset B(0, M_{\mathcal{X}})$

A first obvious property of the embedding  $\mu \mapsto T_\mu$  is its injectivity: if  $\mu$  and  $\nu$  are measures on  $\mathcal{Y}$  such that  $T_\mu = T_\nu$ , then  $(T_\mu)_\# \rho = \mu = (T_\nu)_\# \rho = \nu$ . This injectivity ensures that the Monge embedding preserves the discriminative information about the measures it embeds. A stronger formulation of this injectivity property can be made using the Wasserstein distance.

**Definition 2.1** (Wasserstein distance). The Wasserstein distance of exponent  $p$  between  $\mu, \nu \in \mathcal{P}(\mathcal{Y})$  is defined by

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{Y} \times \mathcal{Y}} \|y - y'\| d\gamma(y, y'), \quad (5)$$

where  $\Pi(\mu, \nu) = \{\gamma \in \mathcal{M}(\mathcal{Y} \times \mathcal{Y}) \mid \forall A \subset \mathcal{Y}, \gamma(A \times \mathcal{Y}) = \mu(A), \gamma(\mathcal{Y} \times A) = \nu(A)\}$ .

**Remark 2.1.** Jensen's inequality gives  $W_1 \leq W_p$ , showing that  $W_1$  is the weakest Wasserstein distance.

**Proposition 2.1.** *The following properties hold:*

(i) *The Monge embedding is reverse-Lipschitz:*

$$\forall \mu, \nu \in \mathcal{P}(\mathcal{Y}), \quad W_2(\mu, \nu) \leq \|T_\mu - T_\nu\|_{L^2(\rho)}.$$

(ii) *The Monge embedding is continuous.*

(iii) *The Monge embedding is in general not better than  $\frac{1}{2}$ -Hölder.*

The proof of this proposition is given in the supplementary material. Note that the general continuity result (ii) is not quantitative. Our goal in the next two sections is to study the Hölder continuity of the Monge map embedding with respect to the 1-Wasserstein distance (which is the weakest Wasserstein distance) and to the total variation distance between measures.

### 2.1 Hölder-continuity Near a Regular Measure

We state a first result, which is a slight variant of a known stability result due to Ambrosio and reported in (Gigli, 2011). While (Gigli, 2011) shows a local  $1/2$ -Hölder behaviour for regular enough source and target measures along a curve in the 2-Wasserstein space, we show the same Hölder behaviour near a probability measure  $\mu$  whose Monge map  $T_\mu$  is Lipschitz continuous, but with respect to the 1-Wasserstein distance.

**Theorem 2.2.** *Let  $\mu, \nu$  be two probability measures over  $\mathcal{Y}$  and assume that  $T_\mu$  is  $K$ -Lipschitz. Then,*

$$\|T_\mu - T_\nu\|_{L^2(\rho)} \leq 2\sqrt{M_{\mathcal{X}} K} W_1(\mu, \nu)^{1/2}.$$

The proof of this theorem follows from simple arguments from convex analysis and Kantorovich-Rubinstein's duality theorem and is given in the supplementary material.

### 2.2 Dimension-dependent Hölder Continuity

Here we assume that  $\rho \equiv 1$  on a compact convex set  $\mathcal{X}$  with unit volume. With no assumption on the embedded measures  $\mu$  and  $\nu$ , another Hölder-continuity result for Monge's embedding can be derived from the following theorem of Berman (Berman, 2018).

**Theorem 2.3** ((Berman, 2018) Proposition 3.4). *For any measures  $\mu$  and  $\nu$  in  $\mathcal{P}(\mathcal{Y})$ ,*

$$\|\nabla \psi_\mu - \nabla \psi_\nu\|_{L^2(\mathcal{Y})}^2 \leq C \left( \int_{\mathcal{Y}} (\psi_\nu - \psi_\mu) d(\mu - \nu) \right)^{\frac{1}{2d-1}},$$

where  $C$  depends only on  $\rho$ ,  $\mathcal{X}$  and  $\mathcal{Y}$ .

We deduce a global Hölder-continuity result for the Monge embedding (2), but with a Hölder exponent that depends on the ambient dimension  $d$ . The proof of this result is in the supplementary material.

**Corollary 2.4.** *For any measures  $\mu$  and  $\nu$  in  $\mathcal{P}(\mathcal{Y})$ ,*

$$\|T_\mu - T_\nu\|_{L^2(\rho)} \leq C W_1(\mu, \nu)^{\frac{1}{2(d-1)(d+2)}},$$

where  $C$  depends only on  $\rho$ ,  $\mathcal{X}$  and  $\mathcal{Y}$

## 3 DIMENSION-INDEPENDENT HÖLDER-CONTINUITY OF THE MONGE EMBEDDING

This section is devoted to a global stability result for the Monge map embedding. We again require that the source measure is the Lebesgue measure  $\rho \equiv 1$  on some compact convex domain  $\mathcal{X}$  with unit volume. Unlike Theorem 2.2, this stability result does not make any regularity assumption on the measures  $\mu, \nu$ . In addition, the Hölder exponent does not depend on the ambient dimension, unlike Corollary 2.4 of the previous section. This dimension-independent stability thus ensures that the use of our embedding will not amplify the curse of dimensionality to which OT-based metrics are subject. We also report a stability of  $\mu \mapsto T_\mu$  with respect to the total variation (TV) distance. This distance is much stronger than the Wasserstein distance, but the exponent is slightly better.

**Theorem 3.1.** *The following inequalities hold for all probability measures  $\mu, \nu$  on a bounded set  $\mathcal{Y}$*

$$\|T_\nu - T_\mu\|_{L^2(\mathcal{X})} \leq C \|\nu - \mu\|_{\text{TV}}^{1/5},$$

$$\|T_\nu - T_\mu\|_{L^2(\mathcal{X})} \leq C W_1(\mu, \nu)^{2/15},$$

and the constants only depend on  $d, \mathcal{X}$  and  $\mathcal{Y}$ .

**Remark 3.1** (Non-optimality). The Hölder-exponent  $\frac{2}{15}$  comes up from our proof, but we see no reason why it should be the optimal exponent. Combining Theorem 3.1 with Proposition 2.1.(iii), we see that the best exponent belongs to the range  $[\frac{2}{15}, \frac{1}{2}]$ .

**Remark 3.2** (Brenier embedding). Instead of working with the optimal transport maps  $T_\mu$ , one could also directly work with the Brenier potentials  $\phi_\mu$ . Our proof also shows Hölder-continuity of the map  $\mu \in \mathcal{P}(\mathcal{Y}) \mapsto \phi_\mu \in L^2(\mathcal{X})$ , with slightly improved exponents. The exponent would be  $1/3$  with respect to the Wasserstein distance and  $2/9$  with respect to the TV distance.

We will establish Theorem 3.1 in the case where both measures  $\mu^0$  and  $\mu^1$  are supported on the same set, which is finite. The general case follows from a simple density argument, summarized in the following lemma, whose proof is in the supplementary material.

**Lemma 3.2.** *Given any  $\mu^0, \mu^1 \in \mathcal{P}(\mathcal{Y})$ , there exists sequences  $(\mu_N^k)_{N \geq 1}$  such that*

- $\mu_N^0$  and  $\mu_N^1$  have the same support, which is finite,
- $\limsup_{N \rightarrow +\infty} \|\mu_N^0 - \mu_N^1\|_{\text{TV}} \leq \|\mu^0 - \mu^1\|_{\text{TV}}$ ,
- $\lim_{N \rightarrow +\infty} W_1(\mu_N^0, \mu_N^1) = W_1(\mu^0, \mu^1)$ ,
- $\lim_{N \rightarrow +\infty} \|T_{\mu_N^1} - T_{\mu_N^0}\|_{L^2(\rho)} = \|T_{\mu^1} - T_{\mu^0}\|_{L^2(\rho)}$ .

### 3.1 Semi-discrete Optimal Transport

When the target probability measure  $\mu$  is discrete, i.e.  $\mu = \sum_{i=1}^N \mu_i \delta_{y_i}$ , we call the optimal transport problem between  $\rho$  and  $\mu$  *semi-discrete*. For the quadratic cost, the dual problem can be written as (e.g. (Hütter & Rigollet, 2019, Eq. (2.6))):

$$\begin{aligned} & \min_{\psi} \int_{\mathcal{X}} \psi^* d\rho + \int_{\mathcal{Y}} \psi d\mu \\ &= \min_{\psi} \sum_{i=1}^N \int_{V_i(\psi)} (\langle x | y_i \rangle - \psi(y_i)) d\rho(x) + \sum_{i=1}^N \mu_i \psi(y_i), \end{aligned} \quad (6)$$

where the minimum is taken among functions  $\psi$  on  $\{y_1, \dots, y_N\}$ . To simplify notations, we will often conflate the function  $\psi$  with the vector  $\psi \in \mathbb{R}^N$  defined by  $\psi_i = \psi(y_i)$ . The function  $\psi$  is referred to as a (dual) *potential* and defines a partition of the domain  $\mathcal{X}$  into so-called Laguerre cells, described for all  $1 \leq i \leq N$  by

$$V_i(\psi) = \{x \in \mathcal{X} \mid \forall j, \psi_j \geq \psi_i + \langle y_j - y_i | x \rangle\}$$

By Theorem 1.1 in (Kitagawa et al., 2019) (see also (Aurenhammer, Hoffmann, & Aronov, 1998)), a potential  $\psi$  solves the dual problem (6) if and only if

$$\forall i \in \{1, \dots, N\}, \quad \int_{V_i(\psi)} \rho(x) dx = \mu_i,$$

The optimal potential  $\psi$  in (6) defines a Monge map  $T : \mathcal{X} \rightarrow \mathcal{Y}$  that is piecewise constant, sending each point  $x$  in  $V_i(\psi)$  to  $y_i$ . Alternatively, one can define  $T = \nabla \phi$  where  $\phi = \psi^*$  is the Legendre transform of  $\psi$ .

### 3.2 Stability of Dual Potentials

In this section, we work in the semi-discrete setting, assuming that all measures are supported on a (fixed) set  $\{y_1, \dots, y_N\}$ . Given a potential  $\psi \in \mathbb{R}^N$ , we denote

$$\begin{aligned} G_i(\psi) &= \rho(V_i(\psi)), \\ G(\psi) &= (G_i(\psi))_{1 \leq i \leq N} \in \mathbb{R}^N, \\ \mu_\psi &= \sum_{1 \leq i \leq N} G_i(\psi) \delta_{y_i}. \end{aligned}$$

And we will consider the set  $S_+$  of potentials such that all cells  $V_i(\psi)$  carry some mass, defined by

$$S_+ = \{\psi \in \mathbb{R}^N \mid \forall i, G_i(\psi) > 0\}.$$

From Theorems 1.3 and 4.1 in (Kitagawa et al., 2019), we know that the map  $G$  is  $C^1$  on  $S_+$ .

**Lemma 3.3.** *Let  $\psi^0, \psi^1 \in S^+$  and consider the linear interpolant  $\psi^t = (1-t)\psi^0 + t\psi^1$ . Then,*

$$\forall i, \quad G_i(\psi^t)^{\frac{1}{d}} \geq (1-t)G_i(\psi^0)^{\frac{1}{d}} + tG_i(\psi^1)^{\frac{1}{d}}. \quad (7)$$

In particular,  $\psi^t \in S^+$ . Moreover,

$$\|G(\psi^t) - G(\psi^0)\|_1 \leq \|G(\psi^1) - G(\psi^0)\|_1, \quad (8)$$

$$\|G(\psi^t) - G(\psi^0)\|_1 \leq 2(1 - (1-t)^d). \quad (9)$$

This lemma follows from Brunn-Minkowski's inequality, and is reported in the supplementary material. By Theorem 1.3 in (Kitagawa et al., 2019), if  $\psi \in S_+$ , then  $G$  is  $C^1$  and its partial derivatives are given by

$$\begin{cases} \frac{\partial G_i}{\partial \psi_j}(\psi) = \frac{\mathcal{H}^{d-1}(V_i(\psi) \cap V_j(\psi))}{\|y_j - y_i\|} & \text{for } i \neq j, \\ \frac{\partial G_i}{\partial \psi_i}(\psi) = -\sum_{j \neq i} \frac{\partial G_i}{\partial \psi_j}(\psi) \end{cases}$$

where  $\mathcal{H}^{d-1}$  is the  $(d-1)$ -dimensional Hausdorff measure. The next proposition gives an explicit lower bound on the smallest non-zero eigenvalue of the opposite of the Jacobian matrix of the map  $G$ . Its proof follows from the stability analysis of finite volumes discretization of elliptic PDEs, see Lemma 3.7 in (Eymard, Gallouët, & Herbin, 2000), but it is also reported in the supplementary material.

**Proposition 3.4** (Discrete Poincaré-Wirtinger inequality). *Consider  $\psi \in S_+$  and  $v \in \mathbb{R}^N$ . Then,*

$$\langle v^2 | G(\psi) \rangle - \langle v | G(\psi) \rangle^2 \leq -C_{d,\mathcal{X},\mathcal{Y}} \langle DG(\psi)v | v \rangle,$$

where  $C_{d,\mathcal{X},\mathcal{Y}} = C(d) \operatorname{diam}(\mathcal{Y}) \operatorname{diam}(\mathcal{X})^{d+1}$  and  $DG$  is the Jacobian of  $G$ .

**Remark 3.3.** In particular,  $-DG(\psi)$  is semidefinite positive, since its smallest non-zero eigenvalue is greater than a variance. This can also be seen from the definition of  $DG(\psi)$  as a Laplacian matrix.

With these two results at hand, we show  $L^2$  stability of the dual potentials.

**Theorem 3.5.** *Let  $\psi^0, \psi^1$  be two potentials in  $S_+$ , satisfying  $\langle \psi^1 - \psi^0 | G(\psi^0) \rangle = 0$ . Then, with  $\mu^k = \mu_{\psi^k}$ ,*

$$\langle (\psi^1 - \psi^0)^2 | G(\psi^0) \rangle \leq C_{d,\mathcal{X},\mathcal{Y}} \|\mu^1 - \mu^0\|_{\text{TV}},$$

$$\langle (\psi^1 - \psi^0)^2 | G(\psi^0) \rangle \leq C_{d,\mathcal{X},\mathcal{Y}} W_1(\mu^1, \mu^0)^{\frac{2}{3}}. \quad (10)$$

*Proof.* In this proof,  $A \lesssim B$  means that  $A \leq CB$  for a constant  $C$  depending only on  $d$  and the diameters of  $\mathcal{X}$  and  $\mathcal{Y}$ . Denote  $\psi^t = (1-t)\psi^0 + t\psi^1$  and  $v = \psi^1 - \psi^0$ . By Taylor's formula,

$$\langle G(\psi^1) - G(\psi^0) | v \rangle = \int_0^1 \langle DG(\psi^t)v | v \rangle dt.$$

Moreover, proposition 3.4 gives

$$\langle v^2 | G(\psi^t) \rangle - \langle v | G(\psi^t) \rangle^2 \lesssim -\langle DG(\psi^t)v | v \rangle.$$

Let us restrict to  $t \in [0, \frac{1}{4}]$ . Then, by Eq. (7), one has

$$G_i(\psi^t) \geq (1-t)^d G_i(\psi^0) \gtrsim G_i(\psi^0).$$

Thus, on the interval  $t \in [0, \frac{1}{4}]$ ,

$$\langle v^2 | G(\psi^0) \rangle \lesssim \langle v^2 | G(\psi^t) \rangle.$$

On the other hand, using the assumption  $\langle v | G(\psi^0) \rangle = 0$  we get

$$\begin{aligned} |\langle v | G(\psi^t) \rangle| &= |\langle v | G(\psi^t) - G(\psi^0) \rangle| \\ &\leq \|v\|_\infty \|G(\psi^t) - G(\psi^0)\|_1 \\ &\lesssim \|G(\psi^t) - G(\psi^0)\|_1, \end{aligned}$$

where we used that  $v$  is  $2M_{\mathcal{X}}$ -Lipschitz (as a difference of dual potentials), and takes positive and negative values (since its scalar product with a constant sign function vanishes). Using the fact that  $v$  is  $2M_{\mathcal{X}}$ -Lipschitz and Kantorovich-Rubinstein's theorem, we also get

$$\begin{aligned} |\langle G(\psi^1) - G(\psi^0) | v \rangle| &\lesssim W_1(\mu^0, \mu^1) \\ &\lesssim \|\mu^0 - \mu^1\|_{\text{TV}} \\ &\lesssim \|G(\psi^0) - G(\psi^1)\|_1. \end{aligned}$$

Proposition 3.4 implies that  $\langle DG(\psi^t)v | v \rangle \leq 0$  for all  $t \in [0, 1]$ . We therefore get

$$\begin{aligned} \int_0^T (\langle v^2 | G(\psi^0) \rangle - \|G(\psi_t) - G(\psi_0)\|_1^2) dt \\ \lesssim W_1(\mu^0, \mu^1) \\ \lesssim \|G(\psi^0) - G(\psi^1)\|_1. \end{aligned} \quad (11)$$

Combining with (8) and  $T = \frac{1}{4}$  concludes the proof of the first stability result of this theorem, with respect to total variation.

To get the second stability result, with respect to the Wasserstein distance, we use Lemma 3.3–(9), which gives for  $t \in [0, T]$ ,

$$\|G(\psi^t) - G(\psi^0)\|_1 \leq 2(1 - (1-t)^d) \lesssim T.$$

Combining this inequality with Eq. (11) gives for  $T \leq \frac{1}{4}$

$$\langle v^2 | G(\psi^0) \rangle \lesssim \frac{1}{T} W_1(\mu^0, \mu^1) + T^2.$$

If  $W_1(\mu^0, \mu^1)^{\frac{1}{3}} \leq \frac{1}{4}$ , we take  $T = W_1(\mu^0, \mu^1)^{\frac{1}{3}}$  to obtain the desired inequality (10). On the other hand, if  $W_1(\mu^0, \mu^1)^{\frac{1}{3}} \geq \frac{1}{4}$ , then taking  $T = \frac{1}{4}$  we get

$$\begin{aligned} \langle v^2 | G(\psi^0) \rangle &\lesssim W_1(\mu^0, \mu^1) \\ &= D \frac{W_1(\mu, \nu)}{D} \leq D \left( \frac{W_1(\mu, \nu)}{D} \right)^{2/3}, \end{aligned}$$

with  $D := \max_{\mu, \nu \in \mathcal{P}(\mathcal{Y})} W_1(\mu, \nu) \leq \text{diam}(\mathcal{Y})$  thus also proving (10) in that case.  $\square$

### 3.3 Proof of Theorem 3.1

We need a result from (Chazal et al., 2017), providing an upper bound on the  $L^2$  norm between gradients of convex functions.

**Proposition 3.6** ((Chazal et al., 2017) Theorem 22). *Let  $f$  and  $g$  be convex functions on a bounded convex set  $\mathcal{X}$ , then*

$$\|\nabla f - \nabla g\|_{L^2} \leq 2C_{\mathcal{X}} \|f - g\|_\infty^{1/2} (\|\nabla f\|_\infty^{1/2} + \|\nabla g\|_\infty^{1/2}),$$

where  $C_{\mathcal{X}}$  depends only on  $\mathcal{X}$ .

We now prove Theorem 3.1 assuming that the measures  $\mu^0, \mu^1$  are supported on a finite set  $\{y_1, \dots, y_N\}$ , which implies the general case thanks to Lemma 3.2. In this proof, we will freely use notation introduced from the previous section. Let  $\psi^0, \psi^1 \in S_+$  be the dual potentials associated to the optimal transport from  $\rho$  to  $\mu^0$  and  $\mu^1$  respectively. Adding a constant to  $\psi^0$  if necessary we assume that  $\langle \psi^0 - \psi^1 | G(\psi^0) \rangle = 0$ . The stability of potentials (Theorem 3.5) implies that

$$\|\psi^0 - \psi^1\|_{L^2(\mu^0)}^2 \lesssim \varepsilon \quad (12)$$

with  $\varepsilon = \|\mu^0 - \mu^1\|_{\text{TV}}$  or  $\varepsilon = W_1(\mu^0, \mu^1)^{\frac{2}{3}}$ .

Defining  $\tilde{\psi}^0 = \psi^0 - \langle \psi^0 - \psi^1 | G(\psi^1) \rangle$  and  $\tilde{\psi}^1 = \psi^1$ , we get  $\langle \tilde{\psi}^0 - \tilde{\psi}^1 | G(\psi^1) \rangle = 0$ . Applying Theorem 3.5, but

switching the role of 0 and 1 we get

$$\begin{aligned} & \|\psi^0 - \psi^1\|_{L^2(\mu^1)} \\ & \leq |\langle \psi^0 - \psi^1 | G(\psi^1) \rangle| + \|\tilde{\psi}^0 - \tilde{\psi}^1\|_{L^2(\mu^1)} \\ & = |\langle \psi^0 - \psi^1 | G(\psi^1) - G(\psi^0) \rangle| + \|\tilde{\psi}^0 - \tilde{\psi}^1\|_{L^2(\mu^1)} \\ & \lesssim W_1(\mu^0, \mu^1) + \sqrt{\varepsilon} \lesssim \sqrt{\varepsilon}, \end{aligned} \quad (13)$$

where we use the same rescaling of the end of the proof of Theorem 3.5 if  $W_1(\mu^0, \mu^1) \geq 1$ . In practice, these  $L^2$  estimates are not sufficient to conclude, and we need to translate them into a  $L^\infty$  estimate in order to apply Proposition 3.6. For this purpose, we consider  $\alpha \in (0, 1)$ , and we define

$$\mathcal{Y}_\alpha = \{y \in \mathcal{Y} \mid |\psi^0(y) - \psi^1(y)| \leq \varepsilon^\alpha\}. \quad (14)$$

By Chebyshev's inequality, we deduce from (12)–(13) that for  $k \in \{0, 1\}$ ,

$$\varepsilon^{2\alpha} \mu^k(\mathcal{Y} \setminus \mathcal{Y}_\alpha) \leq \|\psi^0 - \psi^1\|_{L^2(\mu^k)}^2 \lesssim \varepsilon,$$

which gives

$$1 - \mu^k(\mathcal{Y}_\alpha) \lesssim \varepsilon^{1-2\alpha}.$$

We construct the Legendre transform of the functions  $\psi^k$  on the whole set  $\mathcal{Y} = \{y_1, \dots, y_N\}$ , and of the restrictions of  $\psi^k$  to the set  $\mathcal{Y}_\alpha$ :

$$\phi^k(x) = \max_i \langle x | y_i \rangle - \psi^k(y_i), \quad (15)$$

$$\phi^{k,\alpha}(x) = \max_{y \in \mathcal{Y}_\alpha} \langle x | y \rangle - \psi^k(y). \quad (16)$$

Comparing Eqs. (15) and (16), one sees that  $\phi^{k,\alpha} \leq \phi^k$ . Moreover, if  $x$  belongs to the Laguerre cell  $V_i(\psi^k)$  for some  $y_i \in \mathcal{Y}_\alpha$ , one can check that the maximums in the definition of  $\phi^k(x)$  and  $\phi^{k,\alpha}(x)$  are both attained at the point  $y_i$ . This implies that  $\phi^k \equiv \phi^{k,\alpha}$  on the set

$$\mathcal{X}_\alpha^k = \bigcup_{y_i \in \mathcal{Y}_\alpha} V_i(\psi^k).$$

Note also that this set  $\mathcal{X}_\alpha^k$  is "large", in the sense that

$$\begin{aligned} 1 - \rho(\mathcal{X}_\alpha^k) &= 1 - \sum_{y_i \in \mathcal{Y}_\alpha} \rho(V_i(\psi^k)) \\ &= 1 - \mu^k(\mathcal{Y}_\alpha) \lesssim \varepsilon^{1-2\alpha}. \end{aligned}$$

The gradients  $\nabla \phi^{k,\alpha}$  and  $\nabla \phi^k$  are bounded by  $\text{diam}(\mathcal{Y})$  (by Eqs. (15) and (16)) and they coincide on the "large" set  $\mathcal{X}_\alpha^k$ . This directly implies that they are close in  $L^2$  norm:

$$\begin{aligned} & \|\nabla \phi^{k,\alpha} - \nabla \phi^k\|_{L^2(\mathcal{X})} \\ & = \|\nabla \phi^{k,\alpha} - \nabla \phi^k\|_{L^2(\mathcal{X} \setminus \mathcal{X}_\alpha^k)} \\ & \leq (1 - \rho(\mathcal{X}_\alpha^k))(\|\nabla \phi^{k,\alpha}\|_\infty + \|\nabla \phi^k\|_\infty) \\ & \lesssim \varepsilon^{1-2\alpha}. \end{aligned} \quad (17)$$

On the other hand, by definition of  $\mathcal{Y}_\alpha$  (see Eq. (14)), the functions  $\psi^0$  and  $\psi^1$  are uniformly close on the set  $\mathcal{Y}_\alpha$ . This implies that the Legendre transforms  $\phi^{0,\alpha}$  and  $\phi^{1,\alpha}$ , defined in (16), are also close. Indeed,

$$\begin{aligned} \phi^{0,\alpha}(x) &= \max_{y \in \mathcal{Y}_\alpha} \langle x | y \rangle - \psi^0(x) \\ &\leq \max_{y \in \mathcal{Y}_\alpha} \langle x | y \rangle - \psi^1(x) + \varepsilon^\alpha \\ &= \phi^{1,\alpha}(x) + \varepsilon^\alpha, \end{aligned}$$

thus giving by symmetry

$$\|\phi^{1,\alpha} - \phi^{0,\alpha}\|_\infty \leq \varepsilon^\alpha.$$

Combining this inequality with Proposition 3.6, we obtain

$$\begin{aligned} & \|\nabla \phi^{1,\alpha} - \nabla \phi^{0,\alpha}\|_{L^2(\mathcal{X})} \\ & \lesssim 2(\|\nabla \phi^{0,\alpha}\|_\infty + \|\nabla \phi^{1,\alpha}\|_\infty)^{1/2} \|\phi^{1,\alpha} - \phi^{0,\alpha}\|_\infty^{1/2} \\ & \lesssim \varepsilon^{\frac{\alpha}{2}}. \end{aligned} \quad (18)$$

Using the triangle inequality and the two previous estimations (17)–(18), we obtain

$$\begin{aligned} & \|\nabla \phi^1 - \nabla \phi^0\|_{L^2(\mathcal{X})} \\ & \leq \|\nabla \phi^1 - \nabla \phi^{1,\alpha}\|_{L^2(\mathcal{X})} + \|\nabla \phi^{1,\alpha} - \nabla \phi^{0,\alpha}\|_{L^2(\mathcal{X})} \\ & \quad + \|\nabla \phi^{0,\alpha} - \nabla \phi^0\|_{L^2(\mathcal{X})} \\ & \lesssim \varepsilon^{1-2\alpha} + \varepsilon^{\alpha/2}. \end{aligned}$$

The best exponent is obtained when  $1 - 2\alpha = \alpha/2$  i.e.  $1 = 5\alpha/2$ ,  $\alpha = 2/5$ , giving

$$\|\nabla \phi^1 - \nabla \phi^0\|_{L^2(\mathcal{X})} \lesssim \varepsilon^{\frac{1}{5}},$$

which gives the desired estimates by replacing  $\varepsilon$  with its two possible values (see Eq. 12).

## 4 EXPERIMENTS

We conclude by illustrating<sup>1</sup> our theoretical findings on the Monge map embeddings and rapidly mentioning potential use of these embeddings in machine learning. In what follows, we consider that  $d = 2$  and that  $\rho$  is the Lebesgue measure on the unit square  $\mathcal{X} = [0, 1] \times [0, 1]$ . Additionally, we only consider discrete measures  $\mu$  and  $\nu$  on  $\mathcal{Y} \subset [0, 1] \times [0, 1]$ , for which algorithms readily give estimates of  $W_p(\mu, \nu)$  and of  $T_\mu$  or  $T_\nu$ :  $W_p(\mu, \nu)$  is computed exactly with the network simplex algorithm implemented in the Python Optimal Transport library (Flamary & Courty, 2017) while  $T_\mu$  and  $T_\nu$  are approximated with a damped Newton's algorithm (Kitagawa et al., 2019). The embeddings  $T_\mu$  are infinite dimensional objects that are approximated by their block

<sup>1</sup>[https://github.com/alex-delalande/stability\\_ot\\_maps\\_and\\_linearization\\_wasserstein\\_space](https://github.com/alex-delalande/stability_ot_maps_and_linearization_wasserstein_space)

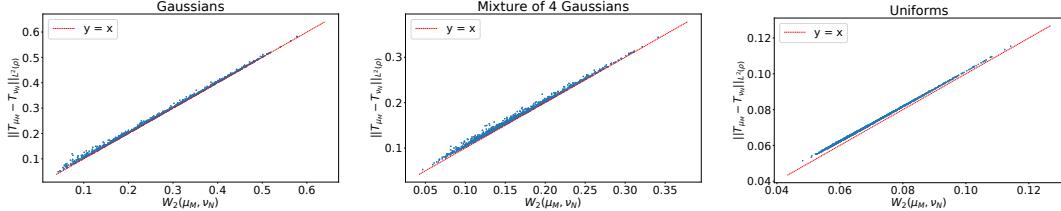


Figure 1:  $W_{2,\rho}$  vs.  $W_2$  between point clouds sampled from Gaussian, Mixture of 4 Gaussian and Uniform distributions.  $W_2$  being approximated with entropic regularization, we may have  $W_2 \geq W_{2,\rho}$  on certain points.

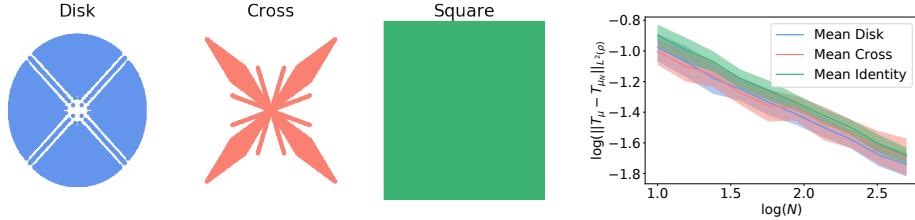


Figure 2: (Left) Target measures, push-forwards of the maps  $T_k = \nabla \phi_k$  where  $\phi_{\text{Disk}}(x, y) := 0.25(x+y) + 0.07(|x+y|^{3/2} + |x-y|^{3/2})$ ,  $\phi_{\text{Cross}}(x, y) := 0.5(x+y) + 0.04 \max(4(x+y-1)^2 + 0.5(2x-1)^2 + 0.5(2y-1)^2, 4(x-y)^2 + 0.5(2x-1)^2 + 0.5(2y-1)^2)$  and  $\phi_{\text{Square}}(x, y) := 0.5(x^2 + y^2)$ . (Right) Sampling distance  $\log(\|T_\mu - T_{\mu_N}\|_{L^2(\rho)})$ .

approximation over a uniform block partition on  $\mathcal{X}$ : for  $m$  a positive integer defining the blocks side  $\frac{1}{m}$ , the blocks are defined by  $\mathcal{X}_{s,t} = [\frac{s-1}{m}, \frac{s}{m}] \times [\frac{t-1}{m}, \frac{t}{m}]$  for  $s, t \in \{1, \dots, m\}$  and  $T_\mu$  is approximated by the vector  $\mathbf{T}_\mu := (\int_{\mathcal{X}_{s,t}} T_\mu d\rho)_{s,t \in \{1, \dots, m\}}$  of size  $2m^2$ . We can notice that our stability results on the maps  $T_\mu \in L^2(\rho)$  can be directly applied to the vectors  $\mathbf{T}_\mu$ . Indeed these vectors correspond to the projections of  $T_\mu$  on a subspace of  $L^2(\rho)$  of piece-wise constant functions on  $\mathcal{X}$ : as a projection this mapping is 1-Lipschitz, which allows to write  $\|\mathbf{T}_\mu - \mathbf{T}_\nu\|_2 \leq \|T_\mu - T_\nu\|_{L^2(\rho)}$ .

**Remark 4.1.** We can note that in dimension  $d$ , the approximation  $\mathbf{T}_\mu$  is of size  $dm^d$ : this limits the use of this approximation to small values of  $d$ . Lighter representations of the map  $T_\mu$  could however be considered, in particular one could leverage the fact that  $T_\mu$  is piece-wise constant or that it is defined properly by the dual potential  $\psi_\mu$  that can be seen as a vector of a size equal to the number of points in the support of  $\mu$ .

#### 4.1 Distance Approximation

We first compare  $W_{2,\rho}(\mu, \nu) = \|T_\mu - T_\nu\|_{L^2(\rho)}$  against  $W_2(\mu, \nu)$  in specific settings to illustrate Equation (4). We consider three different settings corresponding to three different families of distributions. In each setting, 50 point clouds of 300 points are sampled, each from a random distribution that belongs to the given family, and pairwise  $W_2$  and  $W_{2,\rho}$  distances on the 50 point clouds are computed. The distances  $\|T_\mu - T_\nu\|_{L^2(\rho)}$  are approximated with  $\|\mathbf{T}_\mu - \mathbf{T}_\nu\|_2$  with  $m = 200$ .

The three families of distributions we consider are: Gaussian, Mixture of 4 Gaussians and Uniform. Note that for each point cloud sampling in the two first settings the parameters of the sampled distribution are selected randomly. We report in Figure 1 the comparisons between  $W_{2,\rho}$  and  $W_2$ . We observe that  $W_{2,\rho}$  has a behavior very close from the one of  $W_2$ .

#### 4.2 Sampling Approximation

In practice, the population distribution  $\mu$  is often unknown and one can only access to samples  $(x_i)_{i=1, \dots, N}$  from this distribution, yielding the empirical distribution  $\mu_N = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ . One can thus wonder how well  $T_{\mu_N}$  represents  $T_\mu$  in function of the number of samples  $N$ . We illustrate the sampling approximation of  $T_{\mu_N}$  by observing the quantity  $\|T_\mu - T_{\mu_N}\|_{L^2(\rho)}$  as a function of  $N$  in again 3 different settings where the "ground truth map"  $T_\mu$  is prescribed. The 3 maps are chosen as gradients of convex functions and transport the unit square to measures resembling a disk, a cross and a square (Figure 2). For the different values of  $N$  the experiments are repeated 25 times and the standard deviations define the shaded areas surrounding the curves.

In a more statistical context, we observe in Figure 3 the same quantities when the target measures are a Gaussian, a Mixture of 4 Gaussians and the uniform distribution on  $\mathcal{X}$ . Since the "ground truth" maps  $T_\mu$  are unknown in these case, we approximate them with the map  $T_{\mu_M}$  for  $M = 10000$ .

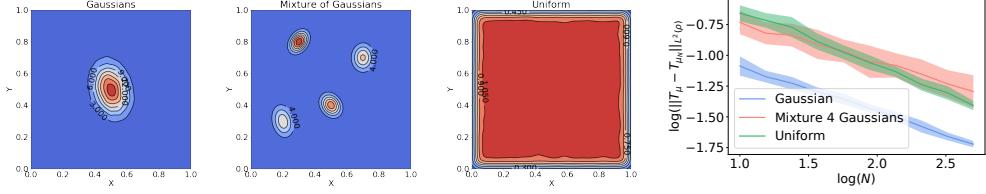


Figure 3: From left to right: densities of the sampled Gaussian, Mixture of 4 Gaussians and Uniform distributions and sampling distance  $\|T_\mu - T_{\mu_N}\|_{L^2(\rho)}$  as a function of  $N$



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

### 4.3 Barycenter Approximation and Clustering

Computing means and barycenters is often necessary in unsupervised learning contexts. For point cloud data, the Wasserstein distance is a natural choice to define such barycenters. For  $(\mu_s)_{s=1,\dots,S}$   $S$  discrete probability measures (corresponding to  $S$  point clouds), the barycenter of  $(\mu_s)$  with non-negative weights  $(\lambda_s)_{s=1,\dots,S}$  is the solution of the following minimization problem:

$$\min_{\mu} \sum_{s=1}^S \lambda_s W_2^2(\mu, \mu_s).$$

This problem does not have an explicit solution and its solution needs to be computed every time the weights are changed. Using transport maps from a reference measure  $\rho$ , it is natural to consider

$$\mu = \left( \sum_{s=1}^S \lambda_s T_{\mu_s} \right) \#$$

as the barycenter of the  $(\mu_s)$ , and one can indeed check that this  $\mu$  minimizes  $\sum_s \lambda_s \|T_\mu - T_{\mu_s}\|_{L^2(\rho)}^2$ . We illustrate this idea with the computation of barycenters of 4 point clouds in Figure 4. Again, operations are performed on the vectorized Monge maps  $T_\mu$ .

These barycenters are in general not equal to their Wasserstein counterparts but they seem to retain the geometric information contained in the point clouds. This idea can be used to extend unsupervised learning



Figure 5: Push-forwards of the 20 centroids after clustering of the Monge map embeddings of the MNIST training set.

algorithms such as  $k$ -Means to family of point clouds. As a toy example, we perform a clustering on the images of the MNIST dataset (LeCun & Cortes, 2010). We convert the 60,000 images of the training set into point clouds of  $\mathcal{X} = [0, 1]^2$  using a simple thresholding on the pixels intensity and we compute for each point cloud its Monge map embedding. We then perform a clustering with the  $k$ -means++ algorithm (Arthur & Vassilvitskii, 2007) on the vectorized Monge maps, looking for  $k = 20$  clusters. Figure 5 shows the push-forwards of the 20 centroids in  $L^2(\rho, \mathbb{R}^d)$ .

## 5 CONCLUSION

We have shown that measures can readily be embedded explicitly in a Hilbert space by their optimal transport map between an arbitrary reference measure and themselves. These embeddings are shown to be injective and bi-Hölder continuous w.r.t the Wasserstein distance. They enable the definition of distances between measures and the use of generic machine learning algorithms in a computationally tractable framework. Future work will focus on the extension of the stability theorem to more general sources and costs, to the improvement of the Hölder exponent and to statistical properties of transport plans, including concentration bounds and sample complexity of the distance they define.

**Acknowledgements** The first author warmly thanks Clément Cancès for pointing Lemma 3.7 in (Eymard et al., 2000) and Robert Berman for discussions related to the topic of this article, and acknowledges the support of the Agence Nationale de la Recherche through the project MAGA (ANR-16-CE40-0014).

## References

- Alaux, J., Grave, E., Cuturi, M., & Joulin, A. (2018). Unsupervised hyperalignment for multilingual word embeddings. *CoRR, abs/1811.01124*. Retrieved from <http://arxiv.org/abs/1811.01124>
- Ambrosio, L., Gigli, N., & Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017, 06–11 Aug). Wasserstein generative adversarial networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 214–223). International Convention Centre, Sydney, Australia: PMLR. Retrieved from <http://proceedings.mlr.press/v70/arjovsky17a.html>
- Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035).
- Aurenhammer, F., Hoffmann, F., & Aronov, B. (1998). Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1), 61–76.
- Berman, R. J. (2018). *Convergence rates for discretized monge-ampère equations and quantitative stability of optimal transport*. (arXiv preprint 1803.00785)
- Bigot, J., Cazelles, E., & Papadakis, N. (2019, February). *Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications*. Retrieved from <https://hal.archives-ouvertes.fr/hal-01647869> (working paper or preprint)
- Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4), 375–417.
- Canas, G., & Rosasco, L. (2012, 09). Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 4.
- Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., & Papadakis, N. (2017, 08). Log-pca versus geodesic pca of histograms in the wasserstein space. *SIAM Journal on Scientific Computing*, 40. doi: 10.1137/17M1143459
- Chazal, F., Cohen-Steiner, D., Lieutier, A., Mérigot, Q., & Thibert, B. (2017). Inference of curvature using tubular neighborhoods. In *Modern Approaches to Discrete Curvature* (p. 133–158). Springer. Retrieved from <https://hal.archives-ouvertes.fr/hal-01425558>
- Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45(1), 223–256. Retrieved from <https://doi.org/10.1214/16-AOS1450> doi: 10.1214/16-AOS1450
- Eymard, R., Gallouët, T., & Herbin, R. (2000). Finite volume methods. *Handbook of numerical analysis*, 7, 713–1018.
- Flamary, R., & Courty, N. (2017). *Pot python optimal transport library*. Retrieved from <https://github.com/rflamary/POT>
- Flamary, R., Cuturi, M., Courty, N., & Rakotomamonjy, A. (2018, December). Wasserstein Discriminant Analysis. *Machine Learning*, 107(12), 1923–1945. Retrieved from <https://hal.archives-ouvertes.fr/hal-01377528> doi: 10.1007/s10994-018-5717-1
- Flamary, R., Lounici, K., & Ferrari, A. (2019). Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation..
- Fletcher, T. P., Lu, C., Pizer, S. M., & Joshi, S. (2004). Principal geodesic analysis for the study of non-linear statistics of shape. *IEEE transactions on medical imaging*, 23(8), 995–1005.
- Genevay, A., Cuturi, M., Peyré, G., & Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems* (pp. 3440–3448).
- Genevay, A., Peyré, G., & Cuturi, M. (2018, 09–11 Apr). Learning generative models with sinkhorn divergences. In A. Storkey & F. Perez-Cruz (Eds.), *Proceedings of the twenty-first international conference on artificial intelligence and statistics* (Vol. 84, pp. 1608–1617). Playa Blanca, Lanzarote, Canary Islands: PMLR. Retrieved from <http://proceedings.mlr.press/v84/genevay18a.html>
- 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.
- Gordaliza, P., Barrio, E. D., Fabrice, G., & Loubes, J.-M. (2019, 09–15 Jun). Obtaining fairness using optimal transport theory. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 2357–2365). Long Beach, California, USA: PMLR. Retrieved from <http://proceedings.mlr.press/v97/gordaliza19a.html>
- Hütter, J.-C., & Rigollet, P. (2019). *Minimax rates of estimation for smooth optimal transport maps*. (arXiv preprint 1905.05828)
- Kitagawa, J., Mérigot, Q., & Thibert, B. (2019). Convergence of a newton algorithm for semi-discrete optimal transport. *Journal of the European Math-*

- ematical Society.*
- LeCun, Y., & Cortes, C. (2010). MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>. Retrieved 2016-01-14 14:24:11, from <http://yann.lecun.com/exdb/mnist/>
- Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26, 101–174.
- Paty, F.-P., D'Aspremont, A., & Cuturi, M. (2019, October). *Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport*. Retrieved from <https://hal.archives-ouvertes.fr/hal-02340371> (working paper or preprint)
- Perrot, M., Courty, N., Flamary, R., & Habrard, A. (2016). Mapping estimation for discrete optimal transport. In *Advances in neural information processing systems (neurips)* (pp. 4197–4205).
- Peyré, G., & Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 355-607. Retrieved from <http://dx.doi.org/10.1561/2200000073> doi: 10.1561/2200000073
- Ramdas, A., Garcia, N., & Cuturi, M. (2015, 09). On wasserstein two sample testing and related families of nonparametric tests. *Entropy*, 19. doi: 10.3390/e19020047
- Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkhäuser, NY*, 55, 58–63.
- Villani, C. (2003). *Topics in optimal transportation*. American Mathematical Soc.
- Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., & Rohde, G. K. (2013, January). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vision*, 101(2), 254–269. Retrieved from <http://dx.doi.org/10.1007/s11263-012-0566-z> doi: 10.1007/s11263-012-0566-z
- Weed, J., & Berthet, Q. (2019). *Estimation of smooth densities in wasserstein distance*.