# The Wasserstein Transform

**Facundo Mémoli** [*1,2] **Zane Smith** [*3] **Zhengchao Wan** [*1]

## Abstract

We introduce the Wasserstein transform, a method for enhancing and denoising datasets defined on general metric spaces. The construction draws inspiration from Optimal Transportation ideas. We establish the stability of our method under data perturbation and, when the dataset is assumed to be Euclidean, we also exhibit a precise connection between the Wasserstein transform and the mean shift family of algorithms. We then use this connection to prove that mean shift also inherits stability under perturbations. We study the performance of the Wasserstein transform method on different datasets as a preprocessing step prior to clustering and classification tasks.

## 1. Introduction

Optimal transport (OT) is concerned with finding cost efficient ways of deforming a given source probability distribution into a target distribution (Villani, 2003; 2008; Santambrogio, 2015). In recent years, ideas from OT have found applications in machine learning and data analysis in general. Applications range from image equalization (Delon, 2004), shape interpolation (Solomon et al., 2015), image/shape (Solomon et al., 2016; Rubner et al., 1998) and document classification (Kusner et al., 2015; Rolet et al., 2016), semi-supervised learning (Solomon et al., 2014), to population analysis of Gaussian processes (Mallasto & Feragen, 2017) and domain adaptation (Courty et al., 2017).

In line with previous applications of OT, we represent datasets as probability measures on an ambient metric space. We introduce the so called *Wasserstein transform* (WT) which takes this input dataset and alters its interpoint distance information in order to both enhance features, such as clusters, present in the data, and to denoise the data. As a theoretical contribution, we prove the stability of our construction to perturbations in the input data (i.e. changes in the input probability measure).

We also interpret our proposed feature enhancing method as both a generalization and a strengthening of Mean Shift (MS) (Cheng, 1995; Fukunaga & Hostetler, 1975) which can operate on general metric spaces. Although mean shift has been generalized to data living on Riemannian manifolds (Subbarao & Meer, 2009; Shamir et al., 2006), our interpretation departs from the ones in those papers in that we do not attempt to estimate a local mean or median of the data but, instead, we use the local density of points to iteratively directly adjust the distance function on the metric space. We do this without appealing to any intermediate embedding into a Euclidean space. As a further contribution, through this connection between the WT and MS, we are able to prove that MS is stable to data perturbations. We are not aware of any extant results in the literature that address this type of stability for MS methods.

Our experiments show that the Wasserstein transform is effective in both denoising and resolving the well known *chaining effect* that affects linkage based clustering methods. Furthermore, we compared the perfomance of our method with mean shift on the MNIST dataset (LeCun et al., 1998) and on Grassmannian manifold data (Cetingul & Vidal, 2009).

## 2. Optimal Transport Concepts

Given a compact metric space $(X, d_X)$ one of the fundamental concepts of OT (Villani, 2003) is the so called Wasserstein distance on the set of all probability measures $\mathcal{P}(X)$ on $X$. The $\ell^1$-Wassertein distance $d_{W,1}(\alpha, \beta)$ between probability measures $\alpha, \beta \in \mathcal{P}(X)$ is obtained by solving the following linear optimization problem:

$$d_{W,1}(\alpha,\beta) := \inf_{\mu \in \Pi(\alpha,\beta)} \iint_{X\times X} d_X(x,x')\, d\mu(x \times x'),$$

where $\Pi(\alpha, \beta)$ is the set of all *couplings* between the probability measures $\alpha$ and $\beta$: namely, $\mu$ in $\Pi(\alpha, \beta)$ is a probability measure on $X \times X$ whose marginals are $\alpha$ and $\beta$, respectively.

[*]Equal contribution [1]Department of Mathematics, The Ohio State University, Ohio, USA [2]Department of Computer Science and Engineering, University of Minnesota, Minnesota, USA [3]Department of Computer Science and Engineering, The Ohio State University, Ohio, USA. Correspondence to: Facundo Mémoli <memoli@math.osu.edu>, Zane Smith <smit9474@umn.edu>, Zhengchao Wan <wan.252@osu.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

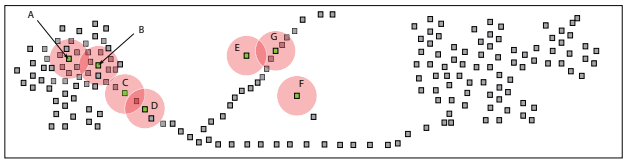


*Figure 1.* In this illustration $\alpha$ is the empirical probability measure associated to the point cloud $X$ shown in the figure, and $d_X$ is the Euclidean distance. With the truncation kernel, the Wasserstein transform $\mathbf{W}_\varepsilon$ will calculate the dissimilarity (via $d_{W,1}$) of the $\varepsilon$-neighborhoods (shown as light red disks) corresponding to all pairs of points to produce a new distance $d_\alpha^{(\varepsilon)}$ on $X$. For example, for the pair of left most points, $A$ and $B$, their respective $\varepsilon$-neighborhoods are not only similar, but also the distance between these regions is small so $d_\alpha^{(\varepsilon)}(A,B)$ will be small too. Something similar is true for the pair $C$ and $D$. In contrast, despite the fact that the points $B$ and $C$ are very close to eachother, their $\varepsilon$-neighborhoods are structurally different: the neighborhood of $B$ is essentially 2-dimensional whereas that of $C$ is 1-dimensional. This will result in $d_\alpha^{(\varepsilon)}(B,C)$ being large. Similarly, since the $\varepsilon$-neighborhood of $E$ is 0-dimensional and that of $G$ is 1-dimensional, despite being very close to each other $d_\alpha^{(\varepsilon)}(E,G)$ will be large. Finally, $d_\alpha^{(\varepsilon)}(E,F)$ will equal the ground distance between $E$ and $F$ since their respective neighborhoods consist of a single point (cf. Remark 2.1).

**Remark 2.1** (Wasserstein distance between Dirac measures)**.** *A simple but important remark (Villani, 2003) is that for points $x, x' \in X$, if one considers the Dirac measures supported at those points (which will be probability measures), $\delta_x$ and $\delta_{x'}$, then the Wasserstein distance between these Dirac measures equals the ground distance: $d_{W,1}(\delta_x, \delta_{x'}) = d_X(x,x')$.*

**Remark 2.2** (A lower bound on Euclidean spaces)**.** *It is known (Rubner et al., 1998) that in Euclidean space $\mathbb{R}^d$, $\|\text{mean}(\alpha) - \text{mean}(\beta)\| \leq d_{W,1}(\alpha,\beta)$ for any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$. In words, in Euclidean spaces, the Wasserstein distance between two probability measures is bounded below by the Euclidean distance between their respective means, which is compatible with the fact that $\alpha$ and $\beta$ can certainly have the same means but can still be quite different as measures. In Section 3.4, this simple fact will help elucidating a relationship between MS and WT on Euclidean spaces.*

## 3. The Wasserstein Transform

Given a compact metric space $(X, d_X)$, we introduce a subset $\mathcal{P}_f(X)$ of $\mathcal{P}(X)$, which consists of those probability measures on $X$ with full support: the support $\text{supp}(\alpha)$ of a probability measure $\alpha$ is the largest closed subset such that every open neighborhood of a point in $\text{supp}(\alpha)$ has positive measure. Given an ambient metric space $X = (X, d_X)$, we interpret a given probability measure $\alpha \in \mathcal{P}_f(X)$ as the *data*. For example, given point cloud $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ one could choose $\alpha$ to be the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. The ambient space distance between data points (in this case the Euclidean distance) is not always directly useful, and by absorbing information about the spatial density of data points, the Wasserstein transform introduced below produces a new metric on the data points which can be used in applications to reveal and concentrate interesting features present but not apparent in the initial presentation of the data. The essential idea behind the Wasserstein transform is to first capture local information of the data and then induce a new distance function between pairs of points based on the dissimilarity between their respective neighborhoods. Localization operators are gadgets that capture these neighborhoods.

### 3.1. Localization Operators

One can always regard a point in a metric space as a Dirac measure supported at that point. More generally, a point in a metric space can be replaced by any reasonable probability measure which includes information about the neighborhood of the point – this leads to the notion of *localization operators* for probability measures.

**Definition 1.** *Let $(X, d_X)$ be a metric space – referred to as the* ambient metric space*. A localization operator $L$ is a map from $\mathcal{P}_f(X)$ to Markov kernels over $X$, i.e., given $\alpha \in \mathcal{P}_f(X)$, $L$ produces $L(\alpha) = (X, m_\alpha^L(\cdot))$, where for every $x \in X$, $m_\alpha^L(x)$ is a probability measure on $X$. We refer to $m_\alpha^L(x)$ as the localized measure at $x$.*

The following are two simple extreme examples. (a) Given $\alpha$ in $\mathcal{P}_f(X)$, let $m_\alpha^L(x) \equiv \alpha, \forall x \in X$, which assigns to all points in $X$ the probability measure $\alpha$. This is a trivial example in that it does not localize the measure $\alpha$ at all. (b) For any $\alpha$ in $\mathcal{P}_f(X)$, let $m_\alpha^L(x) = \delta_x, \forall x \in X$. This is a legitimate localization operator but it does not retain any information from $\alpha$. We will see some useful choices of localization operators in the next couple sections.

### 3.2. The Wasserstein Transform

After specifying a localization operator $L$ and given $\alpha \in \mathcal{P}_f(X)$, one associates each point $x$ in $X$ with a probability measure $m_\alpha^L(x)$, and then obtains a new metric space by considering the Wasserstein distance between each pair of these localized measures.

**Definition 2** (The Wasserstein transform)**.** *Let $(X, d_X)$ be a given ambient metric space and let $\alpha \in \mathcal{P}_f(X)$. Given a localization operator $L$, the* Wasserstein transform $\mathbf{W}_L$ *applied to $\alpha$ gives the distance function $d_\alpha^L$ on $X$ defined by*

$$d_\alpha^L(x,x') := d_{W,1}\left(m_\alpha^L(x), m_\alpha^L(x')\right), \forall x, x' \in X.$$

By $\mathbf{W}_L(\alpha)$ we will denote the (pseudo) metric space $(X, d_\alpha^L)$. Even if in this paper we consider only the $\ell^1$-Wasserstein transform, it is possible to formulate a similar transform using the notion of $\ell^p$-Wasserstein distance.

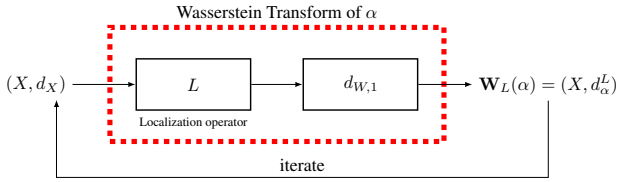


*Figure 2.* Iterating the Wasserstein transform: iteratively change the metric based on density of points as indicated by $\alpha$.

**Remark 3.1** (Iterating the Wasserstein transform)**.** *The Wasserstein transform can be iterated any desired number of times with the purpose of successively enhancing features and/or reducing noise. See Figure 2. After applying the Wasserstein transform once to $\alpha \in \mathcal{P}_f(X)$, the ambient metric space $(X, d_X)$ is transformed into $(X, d_\alpha^L)$. Then we can apply the Wasserstein transform again to $\alpha$ on the ambient space $(X, d_\alpha^L)$ etc. This fact is useful in applications such as clustering; see Section 5.*

### 3.3. Local Truncations

We now concentrate on a particular type of localization operator which we call *local truncation*. Given $\alpha \in \mathcal{P}_f(X)$ and a *scale parameter* $\varepsilon > 0$, consider for each $x \in X$ the probability measure

$$m_\alpha^{(\varepsilon)}(x) := \frac{\alpha|_{B_\varepsilon(x)}}{\alpha(B_\varepsilon(x))},$$

arising from restricting $\alpha$ to the closed ball $B_\varepsilon(x)$ and then renormalizing to obtain a new probability measure. In other words, for each set $A \subset X$, the measure of that set is $m_\alpha^{(\varepsilon)}(x)(A) = \frac{\alpha(B_\varepsilon(x) \cap A)}{\alpha(B_\varepsilon(x))}$. When $X$ is finite, $X = \{x_1, \ldots, x_n\}$, and $\alpha$ is its empirical measure, this formula becomes

$$m_\alpha^{(\varepsilon)}(x)(A) = \frac{\#\{i \,|\, x_i \in A \text{ and } d_X(x_i, x) \leq \varepsilon\}}{\#\{i \,|\, d_X(x_i, x) \leq \varepsilon\}}.$$

We denote the resulting Wasserstein transform by $\mathbf{W}_\varepsilon$, and in this case, for each $\alpha$, the new metric produced by $\mathbf{W}_\varepsilon(\alpha)$ will be denoted as $d_\alpha^{(\varepsilon)}$. See Figure 1 for an intuitive explanation.

**Remark 3.2** (Behavior across scales)**.** *Notice that as $\varepsilon \to \infty$ one has $m_\alpha^{(\varepsilon)}(x) = \alpha$ for any $x \in X$. However, for $\varepsilon \to 0$, $m_\alpha^{(\varepsilon)}(x) \to \delta_x$. In words, $\varepsilon$ acts as a* localization *parameter: for small $\varepsilon$ the renormalized measures absorb local information, whereas for large values the renormalized measures for different points become indistinguishable. Thus we have the following for any $x, x'$ in $X$:*

*(1) as $\varepsilon \to 0$ one has $d_\alpha^{(\varepsilon)}(x, x') \to d_X(x, x')$; and*

*(2) as $\varepsilon \to \infty$ one has $d_\alpha^{(\varepsilon)}(x, x') \to 0$.*

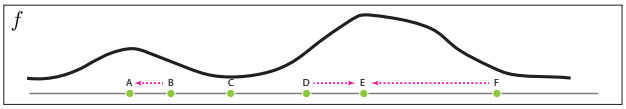


*Figure 3.* After applying one iteration of the Wasserstein transform, both the distance between $A, C$ and the distance between $C, E$ should remain almost the same since these are all critical points of $f$. According to the formula in Remark 3.3, since $f'$ has negative sign at $B$ and $B$ lies to the right of $A$, $B$ will be pushed towards $A$, while $D$ will be pushed away from $A$ since $f'(D) > 0$ and it lies to the right of $A$. Similarly both $D$ and $F$ are pushed towards $E$.

**Interpretation of $\mathbf{W}_\varepsilon(\alpha)$ on the real line.** Using the fact that the Wasserstein distance on $\mathbb{R}$ admits a closed form expression (Villani, 2003) we are able to prove the following Taylor expansion.

**Remark 3.3** (Taylor expansion for $d_\alpha^{(\varepsilon)}(x, x')$ )**.** *When $X$ is a subset of the real line, and the probability measure $\alpha$ has a density $f$, we have the asymptotic formula for $d_\alpha^{(\varepsilon)}(x, x')$ as $\varepsilon \to 0$: for $x' > x$ and $f(x), f(x') > 0$,*

$$d_\alpha^{(\varepsilon)}(x, x') = x' - x + \frac{1}{3}\left[\frac{f'(x')}{f(x')} - \frac{f'(x)}{f(x)}\right]\varepsilon^2 + O(\varepsilon^3).$$

*The interpretation is that after one iteration of the Wasserstein transform $\mathbf{W}_\varepsilon$ of $\alpha$, pairs of points $x$ and $x'$ on very dense areas (reflected by large values of $f(x)$ and $f(x')$) will be at roughly the same distance they were before applying the Wasserstein transform. However, if one of the points, say $x'$ is in a sparse area (i.e. $f(x')$ is small), then the Wasserstein transform will push it away from $x$. It is also interesting what happens when $x$ and $x'$ are both critical points of $f$: in that case the distance does not change (up to order $\varepsilon^2$). See Figure 3 for an illustration. See the supplementary document for a proof of this Taylor expansion.*

### 3.4. The Wasserstein Transform as a Generalization of Mean Shift to Any Metric Space

Mean Shift (Cheng, 1995; Fukunaga & Hostetler, 1975) is a clustering method for Euclidean data which operates by iteratively updating each data point until convergence according to a rule that moves points towards the mean/barycenter of their neighbors. More specifically, given a point cloud $X = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$, a kernel function $K : \mathbb{R}_+ \to \mathbb{R}_+$, and a scale parameter $\varepsilon > 0$, then in the $k$th iteration the $i$th point is shifted as follows: $x_i(0) = x_i$ and for $k \geq 0$,

$$x_i(k+1) = \frac{\sum_{j=1}^n K\left(\frac{\|x_j(k) - x_i(k)\|}{\varepsilon}\right) x_j(k)}{\sum_{j=1}^n K\left(\frac{\|x_j(k) - x_i(k)\|}{\varepsilon}\right)}. \quad (1)$$

The kernels of choice are the Gaussian kernel $K(t) = e^{-t^2/2}$, the Epanechnikov kernel $K(t) = \max\{1 - t, 0\}$, or

the truncation kernel $K(t)$ (which equals 1 if $t \in [0,1]$ and is zero otherwise).

To see how the Mean Shift method can be embedded in the framework of the Wasserstein Transform, let us firstly introduce a new type of localization operator. We assume that the ambient space $X$ is a convex compact subset of $\mathbb{R}^d$ endowed with Euclidean distance. Given any localization operator $L$, define a new localization operator $L^{\mathrm{ms}}$ as follows: for $\alpha \in \mathcal{P}_f(\mathbb{R}^d)$, and $x \in X$, $m_\alpha^{L^{\mathrm{ms}}}(x) := \delta_{\mathrm{mean}(m_\alpha^L(x))}$. In words, at a fixed point $x$, $L^{\mathrm{ms}}$ applied to a measure $\alpha$ at $x$ first localizes $\alpha$ via $L$ to obtain $m_\alpha^L(x)$, and then further localizes this measure by *only retaining information about its mean*. The fact that we can actually compute the mean (or barycenter) of a probability measure (and that this mean remains in $X$) is enabled by the assumption that the ambient space is (a convex) subset of Euclidean space.

Since by Remark 2.1, the Wasserstein distance between Dirac measures equals the ground distance between their support points, then, considering the Wasserstein transform $\mathbf{W}_{L^{\mathrm{ms}}}(\alpha)$ arising from $L^{\mathrm{ms}}$, we have for all $x, x' \in X$ that

$$d_\alpha^{L^{\mathrm{ms}}}(x,x') = \left\| \mathrm{mean}\left(m_\alpha^L(x)\right) - \mathrm{mean}\left(m_\alpha^L(x')\right) \right\|.$$

**The connection with mean shift.** Now, given any kernel function $K$ and $\varepsilon > 0$ as in the case of mean shift one obtains an associated *kernel based localization operator* $L_{K,\varepsilon}$ such that for any $x \in X$ and $A \subset X$,

$$m_\alpha^{L_{K,\varepsilon}}(x)(A) := \frac{\int_A K\left(\frac{\|x-x'\|}{\varepsilon}\right) d\alpha(x')}{\int_{\mathbb{R}^d} K\left(\frac{\|x-x'\|}{\varepsilon}\right) d\alpha(x')}.$$

Now, if for a point cloud $X = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$ we consider $\alpha$ to be the empirical measure induced by $X$, that is, $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, then, for the localization operator $m^{L_{K,\varepsilon}}$ defined above, we obtain, for $x \in X$, the following formula which agrees with the result (1) of applying one iteration of mean shift to the points in $X$:

$$\mathrm{mean}\left(m_\alpha^{L_{K,\varepsilon}}(x)\right) = \frac{\sum_{i=1}^n K\left(\frac{\|x-x_i\|}{\varepsilon}\right) x_i}{\sum_{i=1}^n K\left(\frac{\|x-x_i\|}{\varepsilon}\right)}.$$

Now, that the metric space $\mathbf{W}_{L_{K,\varepsilon}}(\alpha)$ contains the same information as the collection of mean shift points above follows from the classical distance geometry fact that any finite set in $\mathbb{R}^d$ can be reconstructed up to rigid transformations from its interpoint distance matrix (Blumenthal, 1953).

**Remark 3.4** (The Wasserstein transform as a strengthening of mean shift)**.** *Note that in general, because of Remark 2.2 one has that whenever $X \subset \mathbb{R}^d$ is convex and compact, and $\alpha \in \mathcal{P}(\mathbb{R}^d)$, then for all $x, x' \in X$,*

$$\|\mathrm{mean}(m_\alpha^{(\varepsilon)}(x)) - \mathrm{mean}(m_\alpha^{(\varepsilon)}(x'))\| \leq d_\alpha^{(\varepsilon)}(x,x'),$$

*which indicates that the mean shift procedure provides a lower bound for the result of applying the Wasserstein transform to a dataset represented by $\alpha$. In other words, in a certain sense the Wasserstein transform retains, via $d_\alpha^{(\varepsilon)}$, more information about the dataset than mean shift.*

## 4. Stability Under Perturbations of $\alpha$

The goal of this section is to establish the stability of the Wasserstein transform $\mathbf{W}_\varepsilon(\alpha)$ under perturbations of the probability measure $\alpha$ representing the dataset.

As a byproduct of this, we will also obtain a novel stability result for mean shift. We are not aware of extant related stability results for MS in the literature.

As before we fix a compact metric space $(X, d_X)$ (the ambient space). Probability measures on $X$ are required to satisfy a mild doubling type condition.

**Definition 3.** *Given $\Lambda > 0$, we say that a Borel measure $\alpha$ on $X$ satisfies the $\Lambda$-doubling condition if for all $x \in \mathrm{supp}(\alpha), r_1 \geq r_2 > 0$ one has*

$$\frac{\alpha(B_{r_1}(x))}{\alpha(B_{r_2}(x))} \leq \left(\frac{r_1}{r_2}\right)^\Lambda.$$

**Remark 4.1.** *Suppose $\alpha \in \mathcal{P}_f(X)$ and $\mathrm{diam}(X) < D$. If $\alpha$ satisfies the $\Lambda$-doubling condition, then we have $\alpha(B_r(x)) \geq \psi_{\Lambda,D}(r)$, for all $x \in X$ and $r > 0$, where $\psi_{\Lambda,D}(r) := \min\left(1, \left(\frac{r}{D}\right)^\Lambda\right)$.*

*Proof.* Take $r_1 = D, r_2 = r$ in Definition 3, we have when $r \leq D$, $\frac{\alpha(B_D(x))}{\alpha(B_r(x))} \leq \left(\frac{D}{r}\right)^\Lambda$. Notice that $X = B_D(x)$, hence we have $\alpha(B_D(x)) = \alpha(X) = 1$. Therefore

$$\alpha(B_r(x)) \geq \left(\frac{r}{D}\right)^\Lambda \geq \min\left(1, \left(\frac{r}{D}\right)^\Lambda\right).$$

When $r > D$, obviously we have $\alpha(B_r(x)) = \alpha(X) = 1 \geq \min\left(1, \left(\frac{r}{D}\right)^\Lambda\right)$. □

**Setup and assumptions.** We assume that the diameter of $X$ satisfies $\mathrm{diam}(X) := \max_{x,x' \in X} d_X(x,x') < D$ for some $D > 0$. Additionally, we assume that two (fully supported) probability measures $\alpha$ and $\beta$ in $\mathcal{P}_f(X)$ are given and satisfy the doubling condition for some $\Lambda > 0$. Also, since our results below are for local truncations, we fix a scale parameter $\varepsilon > 0$.

Define $\Phi_{\Lambda,D,\varepsilon}(\eta) := \frac{\eta}{\psi_{\Lambda,D}(\varepsilon)} + \left[\left(1 + \frac{\eta}{\varepsilon}\right)^\Lambda - 1\right]$ for $\eta \geq 0$.

**Remark 4.2.** *Notice that $\Phi_{\Lambda,D,\varepsilon}(\eta)$ is an increasing function of $\eta$ and furhermore that $\lim_{\eta \to 0} \Phi_{\Lambda,D,\varepsilon}(\eta) = 0$.*

Then, we have the following stability result for the localization (via local truncations) of two different probability measures on the same ambient space. The stability is expressed in terms of the Wasserstein distance itself.

**Theorem 4.3** (Stability of local truncations)**.**

$$\begin{aligned}&\sup_{x\in X} d_{W,1}\left(m_\alpha^{(\varepsilon)}(x), m_\beta^{(\varepsilon)}(x)\right)\\ \leq&(1+2\varepsilon)\Phi_{\Lambda,D,\varepsilon}\left(\sqrt{d_{W,1}(\alpha,\beta)}\right).\end{aligned}$$

**Remark 4.4.** *We have also established a stability theorem for WT with Lipschitz kernels based localization operators in which no condition such as $\Lambda$-doubling condition is required. Assume $X$ is a compact metric space. If $K:\mathbb{R}_+\to\mathbb{R}_+$ is any $C$-Lipschitz kernel, then there exist constants $N>0$ and $M>0$ only depending on $K$ and $X$, such that for all probability measures $\alpha$ and $\beta$ on $X$, and all $x\in X$:*

$$d_{W,1}\big(m_\alpha^K(x), m_\beta^K(x)\big)\leq\frac{2C\,\mathrm{diam}(X)+M}{N}d_{W,1}(\alpha,\beta).$$

*Above, $m_\alpha^K(x)(A)=\frac{\int_A K(d(x,y)\,d\alpha(y)}{\int_X K(d(x,y)\,d\alpha(y)}$ for $A\subset X$. See (Mémoli et al., 2018) for more details.*

By Remark 4.2, Theorem 4.3 indicates that if $\alpha$ and $\beta$ are similar in terms of the Wasserstein distance, then *for every point* $x\in X$ the localized measures $m_\alpha^{(\varepsilon)}(x)$ and $m_\beta^{(\varepsilon)}(x)$ will also be similar. As a consequence of Theorem 4.3 we obtain the following two theorems:

**Theorem 4.5** (Stability of $d_\alpha^{(\varepsilon)}$)**.**

$$\begin{aligned}&\sup_{x,x'\in X}\|d_\alpha^{(\varepsilon)}(x,x')-d_\beta^{(\varepsilon)}(x,x')\|\\ \leq&2(1+2\varepsilon)\Phi_{\Lambda,D,\varepsilon}\left(\sqrt{d_{W,1}(\alpha,\beta)}\right).\end{aligned}$$

*Proof.* By applying the triangle inequality for the Wasserstein distance [1], we have for any $x,x'\in X$

$$\begin{aligned}&\|d_\alpha^{(\varepsilon)}(x,x')-d_\beta^{(\varepsilon)}(x,x')\|\\ =&\left\|d_{W,1}\left(m_\alpha^{(\varepsilon)}(x),m_\alpha^{(\varepsilon)}(x')\right)-d_{W,1}\left(m_\beta^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x')\right)\right\|\\ \leq&\left\|d_{W,1}\left(m_\alpha^{(\varepsilon)}(x),m_\alpha^{(\varepsilon)}(x')\right)-d_{W,1}\left(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x')\right)\right\|\\ +&\left\|d_{W,1}\left(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x')\right)-d_{W,1}\left(m_\beta^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x')\right)\right\|\\ \leq&d_{W,1}\left(m_\alpha^{(\varepsilon)}(x'),m_\beta^{(\varepsilon)}(x')\right)+d_{W,1}\left(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x)\right)\end{aligned}$$

Therefore by taking supremum on both sides and invoking Theorem 4.3 we obtain the claim. □

**Theorem 4.6** (Stability of mean shift for local truncations)**.** *Assume that $(X,d_X)$ is a subspace of $\mathbb{R}^n$ with Euclidean distance. Then, for mean shift arising from local $\varepsilon$-truncations we have:*

$$\begin{aligned}&\sup_{x\in X}\left\|\mathrm{mean}(m_\alpha^{(\varepsilon)}(x))-\mathrm{mean}(m_\beta^{(\varepsilon)}(x))\right\|\\ \leq&(1+2\varepsilon)\,\Phi_{\Lambda,D,\varepsilon}\left(\sqrt{d_{W,1}(\alpha,\beta)}\right).\end{aligned}$$

*Proof.* By Remark 3.4 and Theorem 4.3 we have $\forall x\in X$,

$$\begin{aligned}&\left\|\mathrm{mean}\left(m_\alpha^{(\varepsilon)}(x)\right)-\mathrm{mean}\left(m_\beta^{(\varepsilon)}(x)\right)\right\|\\ \leq&d_{W,1}\left(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x)\right)\\ \leq&(1+2\varepsilon)\Phi_{\Lambda,D,\varepsilon}\big(\sqrt{d_{W,1}(\alpha,\beta)}\big).\end{aligned}$$

□

### 4.1. The Proof of Theorem 4.3

*Proof of Theorem 4.3.* To bound the Wasserstein distance between the localized measures associated to $\alpha$ and $\beta$, $d_{W,1}\big(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x)\big)$, it is more convenient to first analyze the Prokhorov distance (Gibbs & Su, 2002), and then convert the result to a Wasserstein distance bound by the lemma below. Recall that the Prokhorov distance $d_P(\alpha,\beta)$ between probability measures $\alpha$ and $\beta$ equals $\inf\{\delta>0:\alpha(A)\leq\beta(A^\delta)+\delta,\forall A\subset X\}$. Here $A^\delta$ is the $\delta$-fattening of $A$: the set of points in $X$ which are at distance less than $\delta$ from a point in $A$. Though seemingly asymmetric, $d_P$ is actually symmetric (Gibbs & Su, 2002).

**Lemma 4.7** (Theorem 2 of (Gibbs & Su, 2002))**.** *Given a metric space $(X,d_X)$ with bounded diameter, then $\forall\alpha,\beta\in\mathcal{P}_f(X)$, we have the following relation between the Wasserstein and Prokhorov distances:*

$$\big(d_P(\alpha,\beta)\big)^2\leq d_{W,1}(\alpha,\beta)\leq\big(1+\mathrm{diam}(X)\big)d_P(\alpha,\beta).$$

**Remark 4.8.** *If $\alpha$ and $\beta$ are not fully supported, then by restricting the metric $d$ to $S=\mathrm{supp}(\alpha)\cup\mathrm{supp}(\beta)\subset X$, the rightmost inequality above can be improved to $d_{W,1}(\alpha,\beta)\leq\big(1+\mathrm{diam}(S)\big)d_P(\alpha,\beta)$.*

**Claim 1.** *For any $x\in X$, we have $d_P\big(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x)\big)\leq\Phi_{\Lambda,D,\varepsilon}\big(d_P(\alpha,\beta)\big)$.*

*Proof of Claim 1.* Suppose $d_P(\alpha,\beta)<\eta$ for some $\eta>0$. Fix $x\in X$ and assume WLOG that $\beta(B_\varepsilon(x))\leq\alpha(B_\varepsilon(x))$. Then invoke the expression $d_P(m_\alpha^{(\varepsilon)}(x),m_\beta^{(\varepsilon)}(x))=\inf\{\delta>0:m_\alpha^{(\varepsilon)}(x)(A)\leq m_\beta^{(\varepsilon)}(x)(A^\delta)+\delta,\forall A\subset X\}$.

For any $A \subset X$ we have the following inclusions:

$$\begin{aligned}(A \cap B_\varepsilon(x))^\eta &\subset A^\eta \cap (B_\varepsilon(x))^\eta \subset A^\eta \cap B_{\varepsilon+\eta}(x)\\ &= A^\eta \cap \Big(B_\varepsilon(x) \cup \big(B_{\varepsilon+\eta}(x) \backslash B_\varepsilon(x)\big)\Big)\\ &\subset A^\eta \cap B_\varepsilon(x) \bigcup A^\eta \cap B_{\varepsilon+\eta}(x) \backslash B_\varepsilon(x)\\ &\subset A^\eta \cap B_\varepsilon(x) \bigcup B_{\varepsilon+\eta}(x) \backslash B_\varepsilon(x).\end{aligned}$$

Then by monotonicity of measure and the fact that $d_P(\alpha, \beta) < \eta$, we have

$$\begin{aligned}&m_\alpha^{(\varepsilon)}(x)(A) = \frac{\alpha(A \cap B_\varepsilon(x))}{\alpha(B_\varepsilon(x))}\\ \leq &\frac{\beta((A \cap B_\varepsilon(x))^\eta) + \eta}{\alpha(B_\varepsilon(x))} \leq \frac{\beta((A \cap B_\varepsilon(x))^\eta) + \eta}{\beta(B_\varepsilon(x))}\\ \leq &\frac{\beta(A^\eta \cap B_\varepsilon(x)) + \beta(B_\varepsilon^\eta(x) \backslash B_\varepsilon(x))}{\beta(B_\varepsilon(x))} + \frac{\eta}{\beta(B_\varepsilon(x))}\\ \leq &\frac{\beta(A^\eta \cap B_\varepsilon(x))}{\beta(B_\varepsilon(x))} + \frac{\beta(B_{\varepsilon+\eta}(x)) + \eta}{\beta(B_\varepsilon(x))} - 1\\ \leq &m_\beta^{(\varepsilon)}(x)(A^\eta) + \Big(1 + \frac{\eta}{\varepsilon}\Big)^\Lambda - 1 + \frac{\eta}{\beta(B_\varepsilon(x))}\\ \leq &m_\beta^{(\varepsilon)}(x)(A^\eta) + \xi,\end{aligned}$$

where $\xi := \Phi_{\Lambda,D}(\eta) = \big(1 + \frac{\eta}{\varepsilon}\big)^\Lambda - 1 + \frac{\eta}{\psi_{\Lambda,D}(\varepsilon)}$, and the last inequality follows from Remark 4.1. Note that since $\big(1 + \frac{\eta}{\varepsilon}\big)^\Lambda - 1 \geq 0$, and $\psi_{\Lambda,D}(\varepsilon) \leq 1$, then $\xi \geq \eta$. Thus, from the inequality above, and since $A^\eta \subset A^\xi$, then $m_\alpha^{(\varepsilon)}(x)(A) \leq m_\beta^{(\varepsilon)}(x)(A^\eta) + \xi \leq m_\beta^{(\varepsilon)}(x)(A^\xi) + \xi$. Therefore $d_P(m_\alpha^{(\varepsilon)}(x), m_\beta^{(\varepsilon)}(x)) \leq \xi = \Phi_{\Lambda,D}(\eta)$. Then by letting $\eta \to d_P(\alpha, \beta)$ we have $d_P(m_\alpha^{(\varepsilon)}(x), m_\beta^{(\varepsilon)}(x)) \leq \Phi_{\Lambda,D,\varepsilon}\big(d_P(\alpha, \beta)\big)$, where the RHS is independent of $x$, so the proof is done. □

We now finish the proof of Theorem 4.3. Since $\mathrm{supp}\big(m_\alpha^{(\varepsilon)}(x)\big)$ and $\mathrm{supp}\big(m_\beta^{(\varepsilon)}(x)\big)$ are both contained in $B_\varepsilon(x)$ and $\mathrm{diam}(B_\varepsilon(x)) \leq 2\varepsilon$, we have from Remark 4.8 that $d_{W,1}\Big(m_\alpha^{(\varepsilon)}(x), m_\beta^{(\varepsilon)}(x)\Big) \leq (1 + 2\varepsilon)\, d_P\Big(m_\alpha^{(\varepsilon)}(x), m_\beta^{(\varepsilon)}(x)\Big)$. Now, from this inequality, by Claim 1 above we in turn obtain $d_{W,1}\big(m_\alpha^{(\varepsilon)}(x), m_\beta^{(\varepsilon)}(x)\big) \leq (1 + 2\varepsilon)\, \Phi_{\Lambda,D,\varepsilon}\left(d_P(\alpha, \beta)\right)$. Finally, since $\Phi_{\Lambda,D,\varepsilon}(\eta)$ is an increasing function of $\eta$, by Lemma 4.7 we obtain the statement of the theorem. □

## 5. Implementation and Experiments

In the case of the WT arising from local truncations, $\mathbf{W}_\varepsilon$, for each pair of points $x, x' \in X$, the computation of $d_\alpha^{(\varepsilon)}(x, x') = d_{W,1}(m_\alpha^{(\varepsilon)}(x), m_\alpha^{(\varepsilon)}(x'))$ only requires knowledge of the rectangular chunk of $d_X$ consisting of those points in $B_\varepsilon(x) \times B_\varepsilon(x')$ and, as such, the size of each instance of $d_{W,1}$ can be controlled by choosing $\varepsilon$ to be sufficiently small. The solution of the associated Kantorovich optimal transport problem was carried via entropic regularization (Cuturi, 2013; Genevay et al., 2016; Peyré et al., 2017) using the Sinkhorn code from (Peyre, 2017). The computation of the matrix $\big(\!\big(d_\alpha^{(\varepsilon)}(x, x')\big)\!\big)_{x,x' \in X}$ is an eminently parallelizable task. In our implementation we ran this on a 24 core server via Matlab's parallel computing toolbox. In all of our experiments we used the implementation `sinhorn_log` from (Peyre, 2017) with `options.niter` = 2, `epsilon` = 0.05, and `options.tau` = 0.

**Ameliorating the chaining effect.** In this application we considered the case of clustering two well defined disk shaped blobs (each containing 100 points) connected by a thin trail consisting of 30 points. This is a standard scenario in which standard single linkage hierarchical clustering fails to detect the existence of two clusters due to the so called *chaining effect*. However, successive applications of the Wasserstein transform $\mathbf{W}_\varepsilon$ (corresponding to local truncations) consistently improve the quality of the dendrograms. See Figure 4 for results. See Figure 5 for a study of the effects of increasing $\varepsilon$ and the number of iterations on this dataset. As already suggested by the interpretation in Figure 1, $\varepsilon$-neighborhoods of points in the interior of the dumbbell are essentially two dimensional, whereas $\varepsilon$-neighborhoods of points on the chain are one dimensional – this means that their Wasserstein distance will be quite large, thus having the desired effect of separating the clusters in the sense of the distance $d_\alpha^{(\varepsilon)}$.

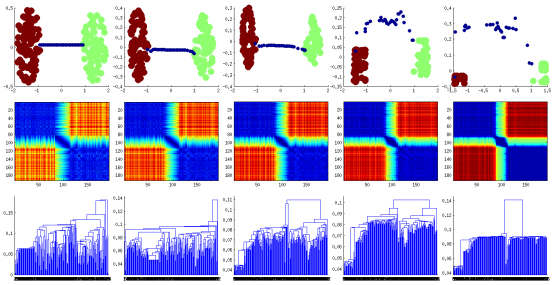

*Figure 4.* **(Chaining effect and WT.)** Top left: A dumbbell shape consisting of two disk shaped blobs each with 100 points and separated by a thin chain of 30 points in the plane with Euclidean distance. The diameter of the initial shape was approximately 4. From left to right: 0,1, 2, 3, and 4, iterations of $\mathbf{W}_\varepsilon$ for $\varepsilon = 0.3$. The top row shows MDS (multi-dimensional scaling) plots of the successive metric spaces thus obtained (color is by class: first blob, chain, and second blob), the middle row shows their distance matrices (ordered so that first we see the points in one blob, then the points on the connecting chain, and then the points of the second blob. The third row shows the corresponding single linkage dendrograms. Notice how the the MDS plot/distance matrices/dendrograms at iteration 5 exhibit clearly defined clusters.

*Table 1.* **(Classification results over the MNIST database.)** Entries show the number of incorrectly classified images over 5K test images (with 5K training images). The best performance of `dT` (tangent distance) is for k=1 and k=3. In both cases WT provides a better performance than `dT`. There is no major difference between the performance of the exact calculation of WT and the one via Sinkhorn, however these take place for different values of knn. Notice that in this dataset the best performances of MS and WT are similar (although they happen for different values of knn).

| knn | 1 | 2 | 3 | 4 | 5 | 10 | 20 | 50 | 100 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|
| `WT-Exact-1` | 121 | 144 | 121 | 122 | **116** | 135 | 181 | 258 | 345 | 1231 |
| `WT-Sinkhorn-1` | 125 | 145 | 118 | 121 | **114** | 142 | 183 | 259 | 342 | 1182 |
| `MS-1` | **117** | 131 | 128 | 133 | 133 | 157 | 199 | 285 | 371 | 1223 |
| `dT` | **133** | 166 | **130** | 141 | 145 | 176 | 219 | 324 | 435 | 1198 |

*Table 2.* **(Classification results over the Grassmann manifold dataset: error rates.)** The best overall performance (0.9753) was obtained by 3-iterations of WT with $\varepsilon = 0.9$. For all $\varepsilon$ except 1.2 WT has better performance than directly using the manifold distance (which is the baseline). The best performance (1.8148) of MS corresponds to the Gaussian kernel for $\varepsilon = 0.8$. The best performance of WT for that value of $\varepsilon$ (1.0039) is the one corresponding to the 3rd iteration. For $\varepsilon = 0.9$ the best performance of MS (1.9155) is worse than the baseline 1.7903 whereas the performance of WT is still better (0.9751). Interestingly, whereas in some cases successive iterations of WT improve its performance, with these parameter choices, in no case did further iterations of MS improve performance. For $\varepsilon = 1.0$ and 1.1 the only method that performs better than the baseline is one iteration of WT. For $\varepsilon = 1.2$ all methods performed below the baseline. For all values of $\varepsilon$ the performance of MS with the truncation kernel was not competitive neither with that of MS with Gaussian kernel, nor with that of WT. See Figure 8 regarding the choice of the range of $\varepsilon$.

| $\mathbf{N = 1.0}$ | dist | trunc1 | trunc2 | trunc3 | gauss1 | gauss2 | gauss3 | wass1 | wass2 | wass3 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\varepsilon = 0.8$ | **1.7903** | **13.1189** | 22.7249 | 27.1865 | **1.8148** | 8.3543 | 31.5804 | **1.4912** | **1.1150** | **1.0039** |
| $\varepsilon = 0.9$ | **1.7903** | 15.7174 | 28.9120 | 51.0221 | **1.9155** | 13.2684 | 44.0966 | **1.2031** | **1.3344** | **0.9753** |
| $\varepsilon = 1.0$ | 1.7903 | 19.3057 | 43.2942 | 58.2056 | 2.1883 | 15.7550 | 54.4154 | 1.2328 | 2.1466 | 8.7605 |
| $\varepsilon = 1.1$ | 1.7903 | 25.8416 | 58.0784 | 75.3460 | 2.3203 | 23.3293 | 65.4943 | 1.7193 | 7.3658 | 75.2454 |
| $\varepsilon = 1.2$ | 1.7903 | 28.2444 | 68.2164 | 71.4105 | 2.4537 | 33.3808 | 74.4915 | 2.5633 | 65.2158 | 75.2193 |

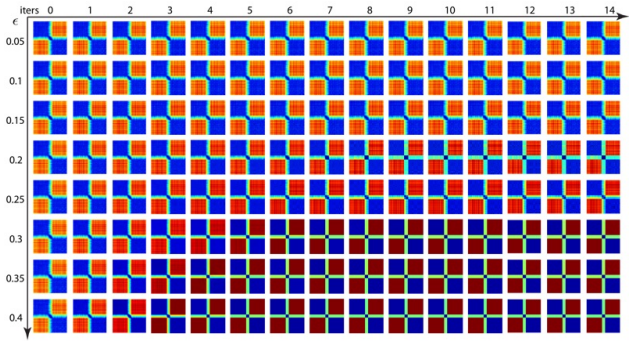


*Figure 5.* **(Chaining effect: varying $\varepsilon$ and number of iterations of WT.)** In this figure we computed 14 different iterations of the dumbbell dataset for $\varepsilon = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35,$ and $0.4$. Notice how distance matrices corresponding to the lower right corner show a very well defined block structure indicative of the presence two large clusters (the blobs) and a smaller one (the points originally corresponding to the chain) .

**Denoising of a circle.** In this example we study a dataset consisting of 800 points uniformly spaced on a circle with radius 1 and centered at the origin in $\mathbb{R}^2$. This circle is heavily corrupted by 1200 outliers chosen uniformly at random in the unit square $[-1, 1] \times [-1, 1]$. This type of preprocessing may help in applications where one wishes to detect voids/loops in data which may signal periodicity (Perea, 2016; Emrani et al., 2014). We compare the performance of

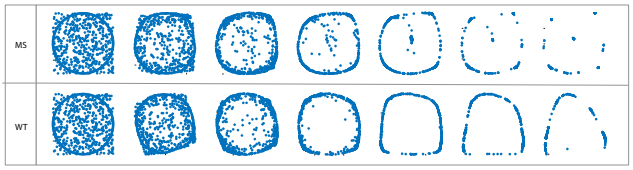


*Figure 6.* **(Denoising of a circle: several iterations of mean shift vs. $\mathbf{W}_\varepsilon$.)** The top row shows the result of applying mean shift with the truncation kernel; the bottom row shows 2D MDS plots of the results obtained from applying the local truncation Wasserstein transform $\mathbf{W}_\varepsilon$. In each case $\varepsilon$ was chosen to be 0.3 relative to the diameter at each iteration. The first column shows the initial dataset which is the same for both cases. From left to right we show increasing number of iterations. Notice how $\mathbf{W}_\varepsilon$ is able to better resolve the shape of the circle; in particular, it is better at displacing interior points towards the high density area around the circle. This feature indicates that $\mathbf{W}_\varepsilon$ can be useful as a preprocessing step for persistent homology calculations (Perea, 2016) or before applying nonlinear dimensionality reduction or manifolds learning techniques to a dataset.

$\mathbf{W}_\varepsilon$ with MS (with the same kernel and same parameter $\varepsilon$) through 6 succesive iterations. See Figure 6.

**Experiment on the MNIST dataset.** In this example we performed knn classification on 5K test images (using 5K training images) from the MNIST dataset. We used both deskewing and the tangent distance `dT` as explained in (LeCun et al., 1998). We tested 3 different methods via one

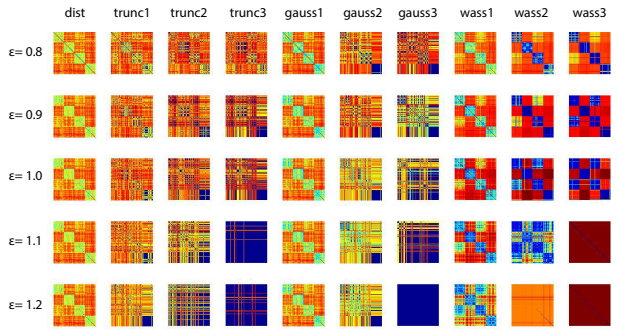


*Figure 7.* **(Grassmann manifold experiment: distance matrices.)** This figure shows the distance matrices (for one realization of the 400 random points in $\mathcal{G}_{10,6}$) corresponding to the rows and columns from Table 2. From this figure, it is evident that WT can help accentuate clustering by moving apart points in different clusters and by concentrating similar points. It is apparent that when the parameter $\varepsilon$ is large, all three methods perform poorly, furthermore, their performance degrades with further iterations (columns 4, 7, and 10). Figure 8 provides one possible explanation.

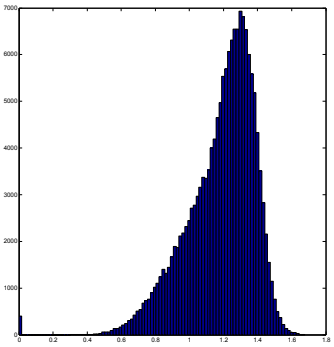

*Figure 8.* **(Grassmann manifold experiment: choice of $\varepsilon$.)** The figure shows the histogram of the manifold distance matrix on $\mathcal{G}_{10,6}$ corresponding to one realization of the 400 random matrices. Whereas $\varepsilon = 0.8$ seems to be small enough so that any ball $B_\varepsilon(x)$ around a point $x$ of the dataset captures only local information, values of $\varepsilon \geq 1.2$ would induce balls $B_\varepsilon(x)$ each of which cover a very large portion of the dataset thus failing to be local. This helps explain the poor performance of both MS (truncation and Gaussian) and WT when $\varepsilon = 1.0, 1.1$ and $1.2$ in Table 2.

iteration of each method. `WT-Exact-1` means the exact computation of the WT (via a standard linear programming algorithm for solving each OT problem (Rubner, 1998)), `WT-Sinkhorn-1` means computation via the Sinkhorn algorithm, whereas `MS-1` refers to mean shift (with the truncation kernel). In each of the three cases the $\varepsilon$ parameter (defining the neighborhood) was chosen to be the same and equal to $0.075$ of the maximal value of the tangent distance matrix corresponding to the 10K points under consideration. The classification was done for different values of knn. In each case, we partitioned the data into the same 5K training points and 5K test points. See Table 1 for the classification results, which indicate that the best performance for MS is comparable to the best performance of WT (for both the exact computation and Sinkhorn). We now present results on a dataset with a more intricate underlying geometric structure.

**Experiment on a Grassmann manifold data.** We tested WT on the synthetic dataset employed in (Cetingul & Vidal, 2009). In our experiments we generated 400 matrices from the Grassmann Manifold $\mathcal{G}_{10,6}$ (Absil et al., 2004; Hüper et al., 2010) as follows: we first generated 4 randomly selected (well separated) loci. We then randomly perturbed each loci 100 times. This was done following Section 4.1 of (Cetingul & Vidal, 2009) by corrupting the angles determining each loci by uniform random noise of width $N = 1$ and mean zero. We then randomly split the set into 100 test matrices and 300 train matrices and estimate the error on a 3nn classification task. The error is averaged over 10000 random selections of the test and train sets. The manifold $\mathcal{G}_{10,6}$ comes equipped with a certain distance which we simply refer to as the "manifold distance" (see the supplementary document). For each point, its 3nns are determined by four different metrics: the manifold distance, the manifold distance after MS with the truncation kernel, the manifold distance after MS with the Gaussian kernel, and the WT of the manifold distance. Both the WT and MS with truncation kernel require a parameter $\varepsilon$ determining the truncation width. The Gaussian kernel requires a standard deviation parameter which we set to 2/3 the $\varepsilon$-value of used for $\mathbf{W}_\varepsilon$. We then repeated the above for the parameter $\varepsilon$ ranging over $\{0.8, 0.9, 1.0, 1.1, 1.2\}$. See Table 2 and Figures 7 and 8. For details about computational techniques and mathematics on the Grassmannian data, see the supplementary document.

## 6. Conclusions

We introduced the Wasserstein transform as a method that takes a dataset represented by a distance matrix and a probability measure and iteratively alters the distance matrix with the goal of enhancing features and/or removing noise. We established the stability of our method under data perturbations, and exhibited a connection with the MS algorithm which permits establishing the stability of MS as well. We validated the ability of WT to reduce the chaining effect and to denoise data on synthetic datasets. We applied WT to classification tasks on MNIST and Grassmann manifold datasets and showed that WT performed at least as well as MS, despite being applicable in wider settings. For future work, it seems of interest to investigate the theoretical behaviour of the iterated WT, e.g., its connection with Ricci/gradient flows. It also seems of interest to study the experimental performance of versions of the WT based on $\ell^p$-Wasserstein distances for $p > 1$ and/or other localization operators.

**Acknowledgements** We acknowledge funding from these sources: NSF AF 1526513, NSF DMS 1723003, NSF DMS 1547357, and NSF CCF 1740761.

## References


Absil, P.-A., Mahony, R., and Sepulchre, R. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80 (2):199–220, 2004.

Blumenthal, L. M. *Theory and applications of distance geometry*. Oxford: Clarendon Press, 1953.

Cetingul, H. E. and Vidal, R. Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds. In *CVPR 2009.*, pp. 1896–1902. IEEE, 2009.

Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pp. 2292–2300, 2013.

Delon, J. Midway image equalization. *Journal of Mathematical Imaging and Vision*, 21(2):119–134, 2004.

Emrani, S., Gentimis, T., and Krim, H. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE SP Letters*, 21(4):459–463, 2014.

Fukunaga, K. and Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf.*, 21(1):32–40, 1975.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *NIPS*, pp. 3440–3448, 2016.

Gibbs, A. L. and Su, F. E. On choosing and bounding probability metrics. *International statistical review*, 70 (3):419–435, 2002.

Hüper, K., Helmke, U., and Herzberg, S. On the computation of means on Grassmann manifolds. In *Proc. Int. Symp. MTNS*, volume 19, pp. 2439–2441, 2010.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *ICML*, pp. 957–966, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mallasto, A. and Feragen, A. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *NIPS*, pp. 5660–5670, 2017.

Mémoli, F., Smith, Z., and Wan, Z. The Wasserstein transform. *arXiv preprint arXiv:1810.07793*, 2018.

Perea, J. A. Persistent homology of toroidal sliding window embeddings. In *ICASSP 2016*, pp. 6435–6439. IEEE, 2016.

Peyre, G. Sinkhorn code. https://github.com/gpeyre/2017-MCOM-unbalanced-ot, 2017.

Peyré, G., Cuturi, M., et al. Computational optimal transport. Technical report, 2017.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics*, pp. 630–638, 2016.

Rubner, Y. Code for the earth movers distance. http://robotics.stanford.edu/~rubner/emd/default.htm, 1998.

Rubner, Y., Tomasi, C., and Guibas, L. J. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pp. 59–66. IEEE, 1998.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, pp. 99–102, 2015.

Shamir, A., Shapira, L., and Cohen-Or, D. Mesh analysis using geodesic mean-shift. *The Visual Computer*, 22(2): 99–108, 2006.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. Wasserstein propagation for semi-supervised learning. In *ICML*, pp. 306–314, 2014.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM TOG*, 34(4):66, 2015.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM TOG*, 35(4):72, 2016.

Subbarao, R. and Meer, P. Nonlinear mean shift over Riemannian manifolds. *IJCV*, 84(1):1, 2009.

Villani, C. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.