# Report for CSE 5339 2018 — (OTMLSA) Optimal Transport in Machine Learning and Shape Analysis

## Distance to Measure

Zhengchao Wan

**Abstract**

The distance between point and measure was introduced by Chazal, Cohen-Steiner and Mérigot, which is robust with respect to noise and outliers and can be used in the offset method to infer geometry and topology knowledge of point clouds. To distinguish samples from different metric measure spaces, DTM-signature was introduced by Brécheteau to build an asymptotic statistical test, which is of level $\alpha$ for chosen small parameter $\alpha$.

## 1 Distance to the Measure

Data is often of the form of a point cloud sampled from an unknown Euclidean space. Here is a basic problem regarding geometric inference of data:

**Question 1.1** *Given a noisy point cloud approximation $C$ of a compact set $K \subset \mathbb{R}^d$, how can we recover geometric and topological informations about $K$, such as its curvature, boundaries, Betti numbers, etc. knowing only the point cloud $C$?*

One idea to retrieve information of a point cloud is to consider the **$R$-offset** of the point cloud - that is the union of balls of radius $R$ whose center lie in the point cloud.

This offset makes good estimation of the topology, normal cones, and curvature measures of the underlying object, shown in previous literature.

The main analyzing tool used is a notion of **distance function**: for a compact $K \subset \mathbb{R}^d$, we define

$$\begin{aligned} d_K : \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto \operatorname{dist}(x, K) \end{aligned}$$

This function is good with the following properties, which are essential to geometric inference:

1. $d_K$ is 1-Lipschitz.
2. $d_K^2$ is 1-semiconcave.
3. $\|d_K - d_{K'}\|_\infty \leqslant d_H(K, K')$.

Unfortunately, offset-based methods do not work well at all in the presence of outliers. For example, the number of connected components will be overestimated if one adds just a single data point far from the original point cloud. This problem arises from the fact that Hausdorff distance is sensitive to outliers.

To reduce the influence by outliers, it's natural to introduce measure on metric spaces and consider the Wasserstein distance. In [2], the authors replaced the distance function to a set $K$ by a **distance function to a measure**. Inspired by the formula $d_K(x) = \min_{y \in K} \|x - y\| =$

$\min\{r > 0 : B(x,r) \cap K \neq \varnothing\}$, a notion of **pseudo-distance function** arises when given a measure $\mu$ on $\mathbb{R}^d$:

$$\delta_{\mu,m} : x \in \mathbb{R}^d \mapsto \inf\{r > 0; \mu(\bar{B}(x,r)) > m\},$$

which is 1-Lipschitz but not semi-concave. With the help of this pseudo-distance function, we arrive at the definition of **distance to a measure**:

**Definition 1.2 (Distance to a Measure)** *For any measure $\mu$ with finite second moment and a positive mass parameter $m_0 > 0$, the distance function to measure (DTM) $\mu$ is defined by the formula:*

$$d^2_{\mu,m_0} : \mathbb{R}^n \to \mathbb{R}, x \mapsto \frac{1}{m_0}\int_0^{m_0} \delta_{\mu,m}(x)^2 dm.$$

**Example 1.3** *Let $C = \{p_1, \cdots, p_n\}$ be a point cloud and $\mu_C = \frac{1}{n}\sum_i \delta_{p_i}$. Then function $\delta_{\mu_C,m_0}$ with $m_0 = k/n$ evaluated at $x \in \mathbb{R}^d$ equal to the distance between $x$ and its $k$th nearest neighbor in $C$.*

*Given $S \subset C$ with $|S| = k$, define $\mathrm{Vor}_C(S) = \{x \in \mathbb{R}^d : \forall p_i \notin S, d(x,p_i) > d(x,S).\}$, which means its elements take $S$ as their $k$ first nearest neighbors in $C$.*

$$\forall x \in \mathrm{Vor}_C(S), d^2_{\mu_C,\frac{k}{n}}(x) = \frac{n}{k}\sum_{p\in S}\|x-p\|^2.$$

Here is a duality like formulation of DTM, which lies in the heart of proof of stability theorems.

**Proposition 1.4** *1. DTM is the minimal cost of the following problem:*

$$d_{\mu,m_0}(x) = \min_{\tilde{\mu}}\left\{W_2\big(\delta_x, \frac{1}{m_0}\tilde{\mu}\big); \tilde{\mu}(\mathbb{R}^d) = m_0, \tilde{\mu} \leqslant \mu\right\}$$

*2. Denote the set of minimizers as $\mathcal{R}_{\mu,m_0}(x)$. Then for each $\tilde{\mu}_{x,m_0} \in \mathcal{R}_{\mu,m_0}(x)$,*

- $\mathrm{supp}(\tilde{\mu}_{x,m_0}) \subset \bar{B}(x, \delta_{\mu,m_0}(x))$;
- $\tilde{\mu}_{x,m_0}\big|_{B(x,\delta_{\mu,m_0}(x))} = \mu\big|_{B(x,\delta_{\mu,m_0}(x))}$;
- $\tilde{\mu}_{x,m_0} \leqslant \mu$.

*3. For any $\tilde{\mu}_{x,m_0} \in \mathcal{R}_{\mu,m_0}(x)$,*

$$d^2_{\mu,m_0}(x) = \frac{1}{m_0}\int_{h\in\mathbb{R}^d}\|h-x\|^2 d\tilde{\mu}_{x,m_0} = W_2^2\Big(\delta_x, \frac{1}{m_0}\tilde{\mu}_{x,m_0}\Big).$$

**Remark 1.5** *According to this proposition, we can regard DTM as the Wasserstein distance between point mass measure and localized measure in a ball with radius related to the pseudo-distance function.*

The following two theorems show that DTM is as good as the distance function to the set. These properties play a key role in the geometric and topological inference of data using the offset method.

**Theorem 1.6 (Regularity)** *1. $d^2_{\mu,m_0}$ is semiconcave, which means $\|x\|^2 - d^2_{\mu,m_0}$ is convex;*

*2. $d^2_{\mu,m_0}$ is differentiable at a point $x$ iff $\mathrm{supp}(\mu) \cap \partial B(x, \delta_{\mu,m_0}(x))$ contains at most 1 point;*

*3. $d_{\mu,m_0}$ is 1-Lipschitz.*

**Theorem 1.7 (DTM stability theorem)** *If $\mu, \nu$ are two probability measures on $\mathbb{R}^d$ and $m_0 > 0$, then*

$$\|d_{\mu,m_0} - d_{\nu,m_0}\|_\infty \leqslant \frac{1}{\sqrt{m_0}}W_2(\mu,\nu).$$

## 2 Offset Reconstruction

**Lemma 2.1** *If $\mu$ is a compactly-supported measure, then $d_S$ is the uniform limit of $d_{\mu,m_0}$ as $m_0$ converges to 0, where $S = \mathrm{supp}(\mu)$, i.e.,*

$$\lim_{m_0 \to 0} \|d_{\mu,m_0} - d_S\|_\infty = 0.$$

**Remark 2.2** *If $\mu$ has **dimension at most** $k > 0$, i.e. $\mu(B(x,\epsilon)) \geqslant C\epsilon^k, \forall x \in S$ when $\epsilon$ is small, then we can control the convergence speed:*

$$\|d_{\mu,m_0} - d_S\|_\infty = O(m_0^{1/k}).$$

If $\mu$ is a probability measure of dimension at most $k > 0$ with compact support $K \subset \mathbb{R}^d$, and $\mu'$ is another probability measure, one has

$$\begin{aligned} \left\|d_K - d_{\mu',m_0}\right\|_\infty &\leqslant \|d_K - d_{\mu,m_0}\|_\infty + \left\|d_{\mu,m_0} - d_{\mu',m_0}\right\|_\infty \\ &\leqslant O(m_0^{1/k}) + \frac{1}{\sqrt{m_0}} W_2(\mu,\mu'). \end{aligned}$$

Here we regard $K$ as the real underlying metric space and $\mu'$ as a noised version of it. Then we see that DTM is stable with respect to the noise.

The following reconstruction theorem tells that the topological information is preserved by DTM. First define $\alpha$-**reach** of $K$, $\alpha \in (0,1]$ as $r_\alpha(K) = \inf\{d_K(x) > 0 : \|\nabla_x d_K\| \leqslant \alpha\}$.

**Theorem 2.3** *Suppose $\mu$ has dimension at most $k$ with compact support $K \subset \mathbb{R}^d$ such that $r_\alpha(K) > 0$ for some $\alpha$. For any $0 < \eta < r_\alpha(K)$, $\exists m_1 = m_1(\mu,\alpha,\eta) > 0$ and $C = C(m_1) > 0$ such that: for any $m_0 < m_1$ and $\mu'$ satisfying $W_2(\mu,\mu') < C\sqrt{m_0}$, $d^{-1}_{\mu',m_0}([0,\eta])$ is **homotopy equivalent** to the offset $d_K^{-1}([0,r])$ for $0 < r < r_\alpha(K)$.*

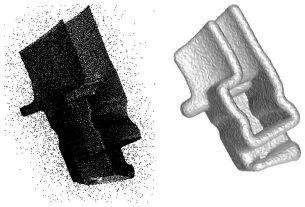

Figure 1: On the left, a point cloud sampled on a mechanical part to which 10% of outliers have been added- the outliers are uniformly distributed in a box enclosing the original point cloud. On the right, the reconstruction of an isosurface of the distance function $d_{\mu_C,m_0}$ to the uniform probability measure on this point cloud.

## 3 DTM signature

In previous sections, we only consider about different measures on $\mathbb{R}^d$, however we can generalize the notion of DTM to general metric measure spaces without obstruction. In [1], the author tried to answer the following question: how to determine that two $N$-samples are from the same underlying metric measure space?

With the help of DTM constructed before, an asymptotic statistical test was introduced by using the DTM-signature.

**Definition 3.1 (DTM-signature)** *The **DTM-signature** associated to some mm-space $(X, \delta, \mu)$, denoted $d_{\mu,m}(\mu)$, is the distribution of the real valued random variable $d_{\mu,m}(Y)$ where $Y$ is some random variable of law $\mu$.*

**Theorem 3.2 (Stability of DTM-signature)** *Given two mm-spaces $(X, \delta_X, \mu), (Y, \delta_Y, \nu)$, we have*

$$W_1(d_{\mu,m}(\mu), d_{\nu,m}(\nu)) \leqslant \frac{1}{m} GW_1(X, Y).$$

**Proposition 3.3** *If $(X, \delta_X, \mu), (Y, \delta_Y, \nu)$ are embedded into some metric space $(Z, \delta)$, then we can upper bound $W_1(d_{\mu,m}(\mu), d_{\nu,m}(\nu))$ by*

$$W_1(\mu, \nu) + \min\{\|d_{\mu,m} - d_{\nu,m}\|_{\infty,\mathrm{supp}(\mu)}, \|d_{\mu,m} - d_{\nu,m}\|_{\infty,\mathrm{supp}(\nu)}\},$$

*and more generally by $(1 + \frac{1}{m})W_1(\mu, \nu)$.*

DTM-signature is not discriminative in general, however under some conditions, DTM-signature is discriminative:

**Proposition 3.4 (Discriminative example)** *Let $(O, \|\|_2, \mu_O), (O', \|\|_2, \mu_{O'})$ be two mm-spaces, for $O, O'$ two non-empty bounded open subset of $\mathbb{R}^d$ satisfying $O = (\bar{O})^\circ$ and $O = (\bar{O'})^\circ$, $\mu_O, \mu_{O'}$ uniform measures. A lower bound for $W_1(d_{\mu_O,m}(\mu_O), d_{\mu_{O'},m}(\mu_{O'}))$ is given by $C|\mathrm{Leb}_d(O)^{\frac{1}{d}} - \mathrm{Leb}_d(O')^{\frac{1}{d}}|$, where $C$ depends on $m, \epsilon, O, O', d$.*

# 4 Statistical test

Given two $N$-samples from the mm-spaces $(X, \delta, \mu), (Y, \gamma, \nu)$, we want to build a algorithm using these two samples to test the null hypothesis:

$$H_0 \text{ "two mm-spaces } X, Y \text{ are isomorphic"},$$

against its alternative:

$$H_1 \text{ "two mm-spaces } X, Y \text{ are not isomorphic"},$$

The test proposed is based on the fact that DTM-signature associated to two isomorphic mm-spaces are equal, which leads to $W_1(d_{\mu,m}(\mu), d_{\nu,m}(\nu)) = 0$.

The idea is described as follows. Given two $N$-samples from the mm-spaces $(X, \delta, \mu), (Y, \gamma, \nu)$, choose randomly two $n$-samples from them respectively, which gives four empirical measures, $\hat{\mu}_n, \hat{\mu}_N, \hat{\nu}_n, \hat{\nu}_N$. We need to consider the following statistic: $T_{N,n,m}(\mu, \nu) = \sqrt{n} W_1(d_{\hat{\mu}_N,m}(\hat{\mu}_n), d_{\hat{\nu}_N,m}(\hat{\nu}_n))$. Denote the law of $T_{N,n,m}(\mu, \nu)$ as $\mathcal{L}_{N,n,m}(\mu, \nu)$.

**Lemma 4.1** *If two mm-spaces are isomorphic, then $\mathcal{L}_{N,n,m}(\mu, \nu) = \mathcal{L}_{N,n,m}(\nu, \nu) = \mathcal{L}_{N,n,m}(\mu, \mu) = \frac{1}{2}\mathcal{L}_{N,n,m}(\mu, \mu) + \frac{1}{2}\mathcal{L}_{N,n,m}(\nu, \nu)$.*

**Remark 4.2** *$\frac{1}{2}\mathcal{L}_{N,n,m}(\mu, \mu) + \frac{1}{2}\mathcal{L}_{N,n,m}(\nu, \nu)$ is the distribution of $ZT_{N,n,m}(\mu, \mu) + (1-Z)T_{N,n,m}(\nu, \nu)$, where $Z$ is another independent random variable with Bernoulli distribution.*

According to the lemma, we shall use $\frac{1}{2}\mathcal{L}_{N,n,m}(\mu, \mu) + \frac{1}{2}\mathcal{L}_{N,n,m}(\nu, \nu)$ to approximate $\mathcal{L}_{N,n,m}(\mu, \nu)$. The $\alpha$-quantile $q_{\alpha,N,n}$ of $\frac{1}{2}\mathcal{L}_{N,n,m}(\mu, \mu) + \frac{1}{2}\mathcal{L}_{N,n,m}(\nu, \nu)$ will be approximated by the $\alpha$-quantile $\hat{q}_{\alpha,N,n}$ of $\frac{1}{2}\mathcal{L}^*_{N,n,m}(\hat{\mu}_N, \hat{\mu}_N) + \frac{1}{2}\mathcal{L}^*_{N,n,m}(\hat{\nu}_N, \hat{\nu}_N)$, where $\alpha$ is a chosen small number. Here $\mathcal{L}^*_{N,n,m}(\hat{\mu}_N, \hat{\mu}_N)$ stands for the distribution of $T_{N,n,m}(\hat{\mu}_N, \hat{\mu}_N) = \sqrt{n} W_1(d_{\hat{\mu}_N,m}(\mu^*_n), d_{\hat{\mu}_N,m}(\mu'^*{}_n))$ conditionally to $\hat{\mu}_N$, where $\mu^*_n$ and $\mu'^*{}_n$ are two independent $n$-samples of law $\hat{\mu}_N$, which can be simulated via bootstrap method. In the end, we will deal with the **test**:

$$\phi_N = 1_{T_{N,n,m}(\mu,\nu) \geqslant \hat{q}_{\alpha,N,n}}.$$

Here is a description of the algorithm for the statistical test:

Algorithm 1: Test Procedure

**Input** : $P$ and $Q$ $N$-samples from $\mu$ (respectively $\nu$), $N$, $n$, $m$, $\alpha$, $N_{MC}$ even ;
`# Compute` $T$ `the test statistic`
Take $P'$ a random subset of $P$ of size $n$ ;
Take $Q'$ a random subset of $Q$ of size $n$ ;
$T \leftarrow \sqrt{n} W_1(\mathrm{d}_{\mathbb{1}_{P,m}}(\mathbb{1}_{P'}), \mathrm{d}_{\mathbb{1}_{Q,m}}(\mathbb{1}_{Q'}))$ ;
`# Compute` *boot* `a` $N_{MC}$`-sample from the bootstrap law`
$dtmP \leftarrow (\mathrm{d}_{\mathbb{1}_{P,m}}(x))_{x \in P}$ ;
$dtmQ \leftarrow (\mathrm{d}_{\mathbb{1}_{Q,m}}(x))_{x \in Q}$ ;
Let *boot* be empty ;
**for** j **in** $1 \,..\, \lfloor N_{MC}/2 \rfloor$ :
  Let $dtmP_1$ and $dtmP_2$ be two independent $n$-samples from $\mathbb{1}_{dtmP}$ ;
  Let $dtmQ_1$ and $dtmQ_2$ be two independent $n$-samples from $\mathbb{1}_{dtmQ}$ ;
  Add $\sqrt{n} W_1(\mathbb{1}_{dtmP_1}, \mathbb{1}_{dtmP_2})$ and $\sqrt{n} W_1(\mathbb{1}_{dtmQ_1}, \mathbb{1}_{dtmQ_2})$ to *boot* ;
`# Compute` *qalph*`, the` $\alpha$`-quantile of` *boot*
Let *qalph* be the $\lfloor N_{MC} - N_{MC} \times \alpha \rfloor$th smallest element of *boot* ;
**Output** : $(T \geq qalph)$

**Proposition 4.3 (Asymptotic level)** *For properly chosen $n$ depending on $N$, for example, $N = cn^\rho$, with $\rho > \frac{\max\{d,2\}}{2}$, test is of asymptotic level $\alpha$, i.e.*

$$\limsup_{N \to \infty} \mathbb{P}_{(\mu,\nu) \in H_0}(\phi_N = 1) \leqslant \alpha.$$

**Example 4.4 (Numerical example)** *$\mu_v$: distribution of $(R\sin(vR) + 0.03M, R\cos(vR) + 0.03M')$ with $R, M, M'$ independent variables; $M$ and $M'$ from the standard normal distribution and $R$ uniform on $(0, 1)$.*

*Sample $N = 2000$ points from two measures (with different $v$), choose $\alpha = 0.05, m = 0.05, n = 20, N_{MC} = 1000$.*

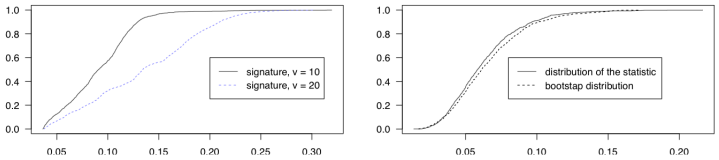


Figure 2: Left: DTM-signature estimates. Right: Bootstrap validity, $v = 10$.

| v | 15 | 20 | 30 | 40 | 100 |
|---|---|---|---|---|---|
| type I error **DTM** | 0.050 | 0.049 | 0.051 | 0.044 | 0.051 |
| power **DTM** | 0.525 | 0.884 | 0.987 | 0.977 | 0.985 |
| power **KS** | 0.768 | 0.402 | 0.465 | 0.414 | 0.422 |

Figure 3: Type 1 error and power approximations by repeating 1000 times. KS represents Kolmogorov Smirnov test.

# References


[1] C. Brécheteau. "The DTM-signature for a geometric comparison of metric-measure spaces from samples". `https://arxiv.org/abs/1702.02838`.

[2] Chazal et al. "Geometric Inference for Measures based on Distance Functions". `https://hal.inria.fr/inria-00383685/file/RR-6930v2.pdf`.