Witnessed k-Distance

Extended abstract

Leonidas Guibas
Department of Computer Science Stanford University
guibas@cs.stanford.edu

Quentin Mérigot
Laboratoire Jean Kuntzmann Université Grenoble 1 and CNRS
quentin.merigot@imag.fr

Dmitriy Morozov
Departments of Computer Science and Mathematics
Stanford University
dmitriy@mrzv.org

ABSTRACT

Distance function to a compact set plays a central role in several areas of computational geometry. Methods that rely on it are robust to the perturbations of the data by the Hausdorff noise, but fail in the presence of outliers. The recently introduced *distance to a measure* offers a solution by extending the distance function framework to reasoning about the geometry of probability measures, while maintaining theoretical guarantees about the quality of the inferred information. A combinatorial explosion hinders working with distance to a measure as an ordinary power distance function. In this paper, we analyze an approximation scheme that keeps the representation linear in the size of the input, while maintaining the guarantees on the inference quality close to those for the exact but costly representation.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling; I.5.1 [**Pattern Recognition**]: Models—*Geometric*

General Terms

Algorithms, Theory

Keywords

geometric inference, power distance, computational topology

1. INTRODUCTION

The problem of recovering the geometry and topology of compact sets from finite point samples has seen several important developments in the previous decade. Homeomorphic surface reconstruction algorithms have been proposed to deal with surfaces in $\mathbb{R}^3$ sampled without noise [1] and with moderate Hausdorff (local) noise [11]. In the case of submanifolds of a higher dimensional Euclidean space [17], or even for more general compact subsets [4], it is also possible, at least in principle, to compute the homotopy type from a Hausdorff sampling. If one is only interested in the homology of the underlying space, the theory of persistent homology [13] applied to Rips graphs provides an algorithmically tractable way to estimate the Betti numbers from a finite Hausdorff sampling [6].

All of these constructions share a common feature: they estimate the geometry of the underlying space by a union of balls of some radius r centered at the data points P. A different way to interpret this union is as the r-sublevel set of the *distance function* to P, $\mathrm{d}_P : x \mapsto \min_{p\in P} \|x - p\|$. Distance functions capture the geometry of their defining sets, and they are stable to Hausdorff perturbations of those sets, making them well-suited for reconstruction results. However, they are also extremely sensitive to the presence of outliers (i.e. data points that lie far from the underlying set); all reconstruction techniques that rely on them fail even in presence of a single outlier.

To counter this problem, Chazal, Cohen-Steiner, and Mérigot [5] developed a notion of *distance function to a probability measure* that retains the properties of the (usual) distance important for geometric inference. Instead of assuming an underlying compact set that is sampled by the points, they assume an underlying probability measure μ from which the point sample P is drawn. The distance function d_{μ,m_0} to the measure μ depends on a mass parameter $m_0 \in (0, 1)$. This parameter acts as a smoothing term: a smaller m_0 captures the geometry of the support better, while a larger m_0 leads to better stability at the price of precision. The crucial feature of the function d_{μ,m_0} is its stability to the perturbations of the measure μ under the Wasserstein distance, defined in Section 2.2. For instance, the Wasserstein distance between the underlying measure μ and the uniform probability measure on the point set P can be small even if P contains some outliers. When this happens, the stability result ensures that distance function $\mathrm{d}_{\mathbf{1}_P,m_0}$ to the uniform probability measure $\mathbf{1}_P$ on P retains the geometric information contained in the underlying measure μ and its support.

Computing with distance functions to measures. In this article we address the computational issues related to this new notion. If P is a subset of $\mathbb{R}^d$ containing N points, and $m_0 = k/N$, we will denote the distance function to the uniform measure on P by $\mathrm{d}_{P,k}$. As observed in [5], the value of $\mathrm{d}_{P,k}$ at a given point x is easy to compute: it is the square root of the average squared distance from the point x to its k nearest neighbors in P. However, most inference methods require a way to represent the function, or more precisely its

SCG'11, June 13–15, 2011, Paris, France.

sublevel sets, globally. It turns out that the distance function $\mathrm{d}_{P,k}$ can be rewritten as a minimum

$$\mathrm{d}^2_{P,k}(x) = \min_{\bar{c}} \|x - \bar{c}\|^2 - w_{\bar{c}}, \qquad (1)$$

where $\bar{c}$ ranges over the set of barycenters of k points in P (see Section 3). Computational geometry provides a rich toolbox to represent sublevel sets of such functions, for example, via weighted α-complexes [12].

The difficulty in applying these methods is that to get an equality in (1) the minimum number of barycenters to store is the same as the number of order-k Voronoi sites of P, making this representation unusable even for modest input sizes. The solution that we propose is to construct an approximation of the distance function $\mathrm{d}_{P,k}$, defined by the same equation as (1), but with $\bar{c}$ ranging over a smaller subset of barycenters. In this article, we study the quality of approximation given by a *linear-sized* subset: the *witnessed barycenters* defined as the barycenters of any k points in P whose order-k Voronoi cell contains at least one of the sample points. The algorithmic simplicity of the scheme is appealing: we only have to find the $k-1$ nearest neighbors for each input point. We denote by $\mathrm{d}^{\mathrm{w}}_{P,k}$ and call *witnessed k-distance* the function defined by Equation (1), where $\bar{c}$ ranges over the witnessed barycenters.

Contributions. Our goal is to give conditions on the point cloud P under which the witnessed k-distance $\mathrm{d}^{\mathrm{w}}_{P,k}$ provides a good uniform approximation of the distance to measure $\mathrm{d}_{P,k}$. We first give a general multiplicative bound on the error produced by this approximation. However, most of our paper (Sections 4 and 5) analyzes the uniform approximation error, when P is a set of independent samples from a measure concentrated near a lower-dimensional subset of the Euclidean space. The following is a prototypical example for our setting, although the analysis we propose allows for a wider range of problems. Note that some of the common settings in the literature either fit directly into this example, or in its logic: the mixture of Gaussians [10] and off-manifold Gaussian noise in normal directions [16] are two examples.

(H1) We assume that the "ground truth" is an unknown probability measure μ whose *dimension is bounded* by a constant $\ell \ll d$. Practically, this means that μ is concentrated on a compact set $K \subseteq \mathbb{R}$ whose dimension is at most ℓ, and that its mass distribution shouldn't "forget" any part of K (see Definition 3). As an example μ could be the uniform measure on a smooth compact ℓ-dimensional submanifold K, or on a finite union of such submanifolds.

This hypothesis ensures that the distance to the measure μ is close to the distance to the support K of μ, and lets us recover information about K. Our first result (Witnessed Bound Theorem 2) states that if the uniform measure to a point cloud P is a good Wasserstein-approximation of μ, then the witnessed k-distance to P provides a good approximation of the distance to the underlying compact set K. The bound we obtain is only a constant times worse than the bound for the exact k-distance.

(H2) The second assumption is that we are not sampling directly from μ, but through a noisy channel. We model this by considering that our measurements come from a measure ν, which is obtained by adding noise to μ. For instance, ν could be the result of the convolution of μ with a Gaussian distribution $\mathcal{N}(0, d^{-1}\sigma^2\mathbf{I})$ whose variance is σ^2. More generally, ν can be any measure such that the *Wasserstein distance* from μ to ν is at most σ. This generalization allows, in particular, to consider noise models that are not translation-invariant.

(H3) Finally, we suppose that our input data set $P \subseteq \mathbb{R}^d$ consists of N points drawn independently from the noisy measure ν. Denote with $\mathbf{1}_P$ the uniform measure on P.

These two hypothesis allow us to control the Wasserstein distance between μ and $\mathbf{1}_P$ with high probability. We assume that the point cloud P is gathered following the three hypothesis above. Our second result states that the witnessed k-distance to P provides a good approximation of the distance to the compact set K with high probability, as soon as the amount of noise σ is low enough and the number of points N is large enough.

Approximation Theorem (Theorem 4). Let P be a set of N points drawn according to the three hypothesis **(H1)**-**(H3)**, let $k \in \{1, \ldots, N\}$ and $m_0 = k/N$. Then, the error bound

$$\|\mathrm{d}^{\mathrm{w}}_{P,k} - \mathrm{d}_K\|_\infty \le 54 m_0^{-1/2}\sigma + 24 m_0^{1/\ell}\alpha_\mu^{-1/\ell}$$

holds with probability at least

$$1 - \gamma_\mu \exp(-\beta_\mu N \max(\sigma^{2+2\ell}, \sigma^4) - \ell \ln(\sigma))$$

where the constants β_μ and γ_μ depend only on μ.

Outline. The relevant background appears in Section 2. We present our approximation scheme together with a general bound of its quality in Section 3. We analyze its approximation quality for measures concentrated on low-dimensional subsets of the Euclidean space in Section 4. The convergence of the uniform measure on a point cloud sampled from a measure of low complexity appears in Section 5 and leads to our main result. We illustrate the utility of the bound with an example and a topological inference statement in our final Section 6.

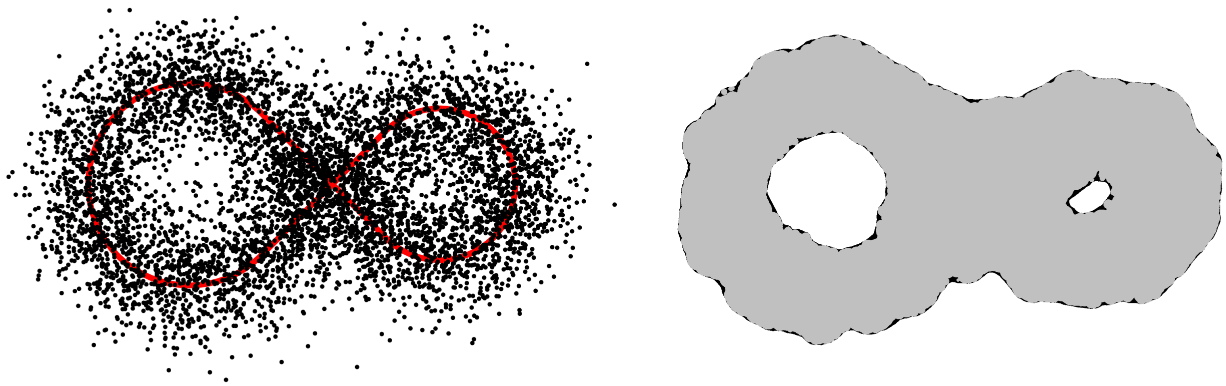
2. BACKGROUND

We begin by reviewing the relevant background.

2.1 Measure

Let us briefly recap the few concepts of measure theory that we use. A *non-negative measure* μ on the space $\mathbb{R}^d$ is a map from (Borel) subsets of $\mathbb{R}^d$ to a non-negative numbers, which is *additive* in the sense that $\mu(\cup_{i\in\mathcal{N}} B_i) = \sum_i \mu(B_i)$ whenever (B_i) is a countable family of disjoint (Borel) subsets. The *total mass* of a measure μ is $\mathrm{mass}(\mu) := \mu(\mathbb{R}^d)$. A measure μ with unit total mass is called a *probability measure*. The *support* of a measure μ, denoted by $\mathrm{spt}(\mu)$ is the smallest closed set whose complement has zero measure. The *expectation* or *mean* of μ is the point $\mathbb{E}(\mu) = \int_{\mathbb{R}^d} x \mathrm{d}\mu(x)$; the variance of μ is the number $\sigma^2_\mu = \int_{\mathbb{R}^d} \|x - \mathbb{E}(\mu)\|^2 \mathrm{d}\mu(x)$.

Although the results we present are often more general, the typical probability measures we have in mind are of two kinds: (i) the uniform probability measure defined by



(a) Data

(b) Sublevel sets

Figure 1: (a) 6000 points sampled from a sideways figure 8 (in red), with circle radii $R_1 = \sqrt{2}$ and $R_2 = \sqrt{9/8}$. The points are sampled from the uniform measure on the figure-8, convolved with the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ where $\sigma = .45$. (b) r-sublevel sets of the witnessed (in gray) and exact (additional points in black) k-distances with mass parameter $m_0 = 50/6000$, and $r = .239$.

the volume form of a lower-dimensional submanifold of the ambient space and (ii) discrete probability measures that are obtained through noisy sampling of probability measures of the previous kind. For any finite set P with N points, denote by $\mathbf{1}_P$ the uniform measure supported on P, i.e. the sum of Dirac masses centered at $p \in P$ with weight $1/N$.

2.2 Wasserstein distance

A natural way to quantify the distance between two measures is the *Wasserstein distance*. This distance measures the L^2-cost of transporting the mass of the first measure onto the second one. A general study of this notion and its relation to the problem of optimal transport appear in [18]. We first give the general definition and then explain its interpretation when one of the two measures has finite support.

A *transport plan* between two measures μ and ν with the same total mass is a measure π on the product space $\mathbb{R}^d \times \mathbb{R}^d$ such that for every subsets A, B of $\mathbb{R}^d$, $\pi(A \times \mathbb{R}^d) = \mu(A)$ and $\pi(\mathbb{R}^d \times B) = \nu(B)$. Intuitively, $\pi(A \times B)$ represents the amount of mass of μ contained in A that will be transported to B by π. The *cost* of this transport plan is given by

$$c(\pi) := \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) \right)^{1/2}$$

Finally, the *Wasserstein distance* between μ and ν is the minimum cost of a transport plan between these measures.

Consider the special case where the measure ν is supported on a finite set P. This means that ν can be written as $\sum_{p \in P} \alpha_p \delta_p$, where δ_p is the unit Dirac mass at P. Moreover, $\sum_p \alpha_p$ must equal the total mass of μ. A transport plan π between μ and ν corresponds to a decomposition of μ into a sum of positive measures $\sum_{p \in P} \mu_p$ such that $\mathrm{mass}(\mu_p) = \alpha_p$. The squared cost of the plan defined by this decomposition is then

$$c(\pi) = \left(\sum_{p \in P} \left[\int_{\mathbb{R}^d} \|x - p\|^2 \mathrm{d}\mu_p(x) \right] \right)^{1/2}.$$

Wasserstein noise. Two properties of the Wasserstein distances are worth mentioning for our purpose. Together, they show that the Wasserstein noise and sampling model generalize the commonly used empirical sampling with Gaussian noise model:

- Consider a probability measure μ and $f : \mathbb{R}^d \to \mathbb{R}$ the density of a probability distribution centered at the origin, and denote by ν the result of the convolution of μ by f. Then, the Wasserstein distance between μ and ν is at most σ, where $\sigma^2 := \int_{\mathbb{R}^d} \|x\|^2 f(x) \mathrm{d}x$ is the variance of the probability distribution defined by f.

- Let P denote a set of N points drawn independently from a given measure ν. Then, the the Wasserstein distance $\mathrm{W}_2(\nu, \mathbf{1}_P)$ between ν and the uniform probability measure on P converges to zero as N grows to infinity with high probability. Examples of such asymptotic convergence results are common in statistics, e.g. [3] and references therein. In Proposition 3 below, we give a quantitative non-asymptotic result assuming that ν is low-dimensional **(H1)**.

Using the notation introduced in the two items above, one has $\limsup_{N \to +\infty} \mathrm{W}_2(\mu, \mathbf{1}_p) \leq \sigma$ with high probability. A more quantitative version of this statement can be found in Corollary 1.

2.3 Distance-to-measure and k-distance

In [5], the authors introduce a distance to a probability measure as a way to infer the geometry and topology of this measure in the same way the geometry and topology of a set is inferred from its distance function. Given a probability measure μ and a *mass parameter* $m_0 \in (0, 1)$, they define a distance function d_{μ, m_0} which captures the properties of the usual distance function to a compact set that are used for geometric inference.

DEFINITION 1. For any point x in $\mathbb{R}^d$, let $\delta_{\mu,m}(x)$ be the radius of the smallest ball centered at x that contains a mass at least m of the measure μ. The *distance to the measure* μ with parameter m_0 is defined by $\mathrm{d}_{\mu,m_0}(x) = m_0^{-1/2} \left(\int_{m=0}^{m_0} \delta_{\mu,m}(x)^2 \mathrm{d}m \right)^{1/2}$.

Given a point cloud P containing N points, the measure of interest is the uniform measure $\mathbf{1}_P$ on P. When m_0 is a fraction k/N of the number of points (where k is an integer), we call *k-distance* and denote by $\mathrm{d}_{P,k}$ the distance to the measure $\mathrm{d}_{\mathbf{1}_P,m_0}$. The value of $\mathrm{d}_{P,k}$ at a query point x is given by

$$\mathrm{d}^2_{P,k}(x) = \frac{1}{k} \sum_{p \in \mathrm{NN}^k_P(x)} \|x - p\|^2.$$

where $\mathrm{NN}^k_P(x) \subseteq P$ denotes the k nearest neighbors in P to the point $x \in \mathbb{R}^d$. (Note that while the k-th nearest neighbor itself might be ambiguous, on the boundary of an order-k Voronoi cell, the distance to the k-th nearest neighbor is always well defined, and so is $\mathrm{d}_{P,k}$.)

The most important property of the distance function d_{μ,m_0} is its stability, for a fixed m_0, under perturbations of the underlying measure μ. This property provides a bridge between the underlying (continuous) μ and the discrete measures $\mathbf{1}_P$. According to [5, Theorem 3.5], for any two probability measures μ and ν on $\mathbb{R}^d$,

$$\| \mathrm{d}_{\mu,m_0} - \mathrm{d}_{\nu,m_0} \|_\infty \le m_0^{-1/2} \, \mathrm{W}_2(\mu, \nu), \tag{2}$$

where $\mathrm{W}_2(\mu,\nu)$ denotes the Wasserstein distance between the two measures. The bound in this inequality depends on the choice of m_0, which acts as a smoothing parameter.

3. WITNESSED k-DISTANCE

In this section, we describe a simple scheme for approximating the distance to a uniform measure, together with a general error bound. The main contribution of our work, presented in Section 4, is the analysis of the quality of approximation given by this scheme when the input points come from a measure concentrated on a lower-dimensional subset of the Euclidean space.

3.1 k-Distance as a Power Distance

Given a set of points $U = \{u_1, \ldots, u_n\}$ in $\mathbb{R}^d$ with weights w_u for every $u \in U$, we call *power distance* to U the function pow_U obtained as the lower envelope of all the functions $x \mapsto \|u - x\|^2 - w_u$, where u ranges over U. By Proposition 3.1 in [5], we can express the square of any distance to a measure as a power distance with non-positive weights. The following proposition recalls this property of the k-distance $\mathrm{d}_{P,k}$.

PROPOSITION 1. *For any $P \subseteq \mathbb{R}^d$, denote by $\mathrm{Bary}^k(P)$ the set of barycenters of any subset of k points in P. Then*

$$\mathrm{d}^2_{P,k} = \min \left\{ \|x - \bar{c}\|^2 - w_{\bar{c}};\ \bar{c} \in \mathrm{Bary}^k(P) \right\}, \tag{3}$$

where the weight of a barycenter $\bar{c} = \frac{1}{k}\sum_i p_i$ is given by $w_{\bar{c}} := -\frac{1}{k}\sum_i \|\bar{c} - p\|^2$.

PROOF. For any subset C of k points in P, define

$$\delta^2_C(x) := \frac{1}{k} \sum_{p \in C} \|x - p\|^2$$

Denoting by $\bar{c}$ the barycenter of the points in C, an easy computation shows

$$\delta^2_C(x) = \frac{1}{k} \sum_{p \in C} \|x - p\|^2 = \|x - \bar{c}\|^2 - w_{\bar{c}}$$

where the weight is given by $w_{\bar{c}} = -\frac{1}{k}\sum_{p \in C} \|\bar{c} - p\|^2$. The proposition follows from the definition of the k-distance. □

In other words, the square of the k-distance function to P coincides exactly with the power distance to the set of barycenters $\mathrm{Bary}^k(P)$ with the weights defined above. From this expression, it follows that the sublevel sets of the k-distance $\mathrm{d}_{P,k}$ are finite unions of balls,

$$\mathrm{d}^{-1}_{P,k}([0,\rho]) = \bigcup_{c \in \mathrm{NN}^k_P(\mathbb{R}^d)} \mathrm{B}(\bar{c}, (\rho^2 + w_{\bar{c}})^{1/2}).$$

Therefore, ignoring the complexity issues, it is possible to compute the homotopy type of this sublevel set by considering the weighted alpha-shape of $\mathrm{Bary}^k(P)$ (introduced in [12]), which is a subcomplex of the regular triangulation of the set of weighted barycenters.

From the proof of Proposition 1, we also see that the only barycenters that actually play a role in (3) are the barycenters of k points of P whose order-k Voronoi cell is not empty. However, the dependence on the number of non-empty order-k Voronoi cells makes computation intractable even for moderately sized point clouds in the Euclidean space.

One way to avoid this difficulty is to replace the k-distance to P by an approximate k-distance, defined as in Equation (3), but where the minimum is taken over a smaller set of barycenters. The question is then: given a point set P, can we replace the set of barycenters Bary^k_P in the definition of k-distance by a small subset B while controlling the approximation error $\| \mathrm{pow}_B^{1/2} - \mathrm{d}_{P,k}\|_\infty$?

This approach is especially attractive since many geometric and topological inference methods using distance functions to compact sets or to measures continue to hold when one of the distance functions is replaced by a good approximation *in the class of power distances.*

3.2 Approximating by witnessed k-distance

In order to approach this question, we consider a subset of the supporting barycenters suggested by the input data which we call witnessed barycenters. The answer to the question is then essentially positive when the input point cloud P satisfies the hypotheses **(H1)**-**(H3)**.

DEFINITION 2. For every point x in P, the barycenter of x and its $(k-1)$ nearest neighbors in P is called a *witnessed k-barycenter.* Let $\mathrm{Bary}^k_{\mathrm{w}}(P)$ be the set of all such barycenters. We get one witnessed barycenter for every point x of the sampled point set, and define the *witnessed k-distance*,

$$\mathrm{d}^{\mathrm{w}}_{P,k} = \min\{\|x - \bar{c}\|^2 - w_{\bar{c}}; \bar{c} \in \mathrm{Bary}^k_{\mathrm{w}}(P)\}.$$

Computing the set of all witnessed barycenters of a point set P only requires finding the $k-1$ nearest neighbors of every point in P. This search problem has a long history in computational geometry [2, 7, 14], and now has several practical implementation.

General error bound. Because the distance functions we consider are defined by minima, and $\mathrm{Bary}^k_{\mathrm{w}}(P)$ is a subset of $\mathrm{Bary}^k(P)$, the witnessed k-distance is always greater than the exact k-distance. In the lemma below, we give a general multiplicative upper bound. This lemma does not assume any specific property for the input point set P. However, even such a coarse bound can be used to estimate Betti numbers of sublevel sets of $\mathrm{d}_{P,k}$, using arguments similar to those in [6].

LEMMA 1 (GENERAL BOUND). *For any finite point set $P \subseteq \mathbb{R}^d$ and $0 < k < |P|$, one has*

$$\mathrm{d}_{P,k} \leq \mathrm{d}^{\mathrm{w}}_{P,k} \leq (2+\sqrt{2})\,\mathrm{d}_{P,k}$$

PROOF. Let $y \in \mathbb{R}^d$ be a point, and $\bar{p}$ the barycenter associated to a cell that contains y. This translates into $\mathrm{d}_{P,k}(y) = \mathrm{d}_{\bar{p}}(y)$. In particular, $\|\bar{p} - y\| \leq \mathrm{d}_{P,k}(y)$ and $\sqrt{-w_{\bar{p}}} \leq \mathrm{d}_{P,k}(y)$.

Let us find a witnessed barycenter $\bar{q}$ that is close to $\bar{p}$. We know that $\bar{p}$ is the barycenters of k points $x_1, \dots, x_n$, and that $-w_{\bar{p}} = \frac{1}{k}\sum_{i=1}^k \|x_i - \bar{p}\|^2$. Consequently, there should exist an x_i such that $\|x_i - \bar{p}\| \leq \sqrt{-w_{\bar{p}}}$. Let $\bar{q}$ be the barycenter witnessed by x. Then,

$$\begin{aligned}\mathrm{d}^{\mathrm{w}}_{P,k}(y) &\leq \mathrm{d}_{\bar{q}}(y) \leq \mathrm{d}_{\bar{q}}(x) + \|x-y\| \\ &\leq \mathrm{d}_{\bar{p}}(x) + \|x - \bar{p}\| + \|\bar{p} - y\|\end{aligned}$$

Combining the inequality

$$\mathrm{d}_{\bar{p}}(x) = \left(\|x-\bar{p}\|^2 - w_{\bar{p}}\right)^{1/2} \leq \sqrt{2}\sqrt{-w_{\bar{p}}}$$

together with $\|x - \bar{p}\| \leq \sqrt{-w_{\bar{p}}}$, we get

$$\begin{aligned}\mathrm{d}^{\mathrm{w}}_{P,k}(y) &\leq (1+\sqrt{2})\sqrt{-w_{\bar{p}}} + \|\bar{p} - y\| \\ &\leq (2+\sqrt{2})\,\mathrm{d}_{P,k}(y)\end{aligned}$$ □

4. APPROXIMATION QUALITY

Let us recall briefly our hypothesis **(H1)**-**(H3)**. There is an ideal, well-conditioned measure μ on $\mathbb{R}^d$ supported on an unknown compact set K. We also have a noisy version of μ, that is another measure ν with $\mathrm{W}_2(\mu,\nu) \leq \sigma$, and we suppose that our data set P consists of N points independently sampled from ν. In this section we give conditions under which the witnessed k-distance to P provides a good approximation of the distance to the underlying set K.

4.1 Dimension of a measure

First, we make precise the main assumption **(H1)** on the underlying measure μ, which we use to bound the approximation error made when replacing the exact by the witnessed k-distance. We require μ to be low dimensional in the following sense.

DEFINITION 3. A measure μ on $\mathbb{R}^d$ is said to have *dimension at most* ℓ, which we denote by $\dim \mu \leq \ell$, if there is a positive constant α_μ such that the amount of mass contained in the ball $B(p,r)$ is at least $\alpha_\mu r^\ell$, for every point p in the support of μ and every r smaller than the diameter of this support.

The important assumption here is that the lower bound $\mu(\mathrm{B}(p,r)) \geq \alpha r^\ell$ should be true for some positive constant α and for r smaller than a given constant R. The choice of $R = \mathrm{diam}(\mathrm{spt}(\mu))$ provides a normalization of the constant α_μ and slightly simplifies the statements of the results.

Let M be an ℓ-dimensional compact submanifold of $\mathbb{R}^d$, and $f : M \to \mathbb{R}$ a positive weight function on M with values bounded away from zero and infinity. Then, the dimension of the volume measure on M weighted by the function f is at most ℓ. A quantitative statement can be obtained using the Bishop-Günther comparison theorem; the bound depends on the maximum absolute sectional curvature of the manifold M (see e.g. Proposition 4.9 in [5]). Note that the positive lower bound on the density is really necessary. For instance, the dimension of the standard Gaussian distribution $\mathcal{N}(0,1)$ on the real line is not bounded by 1 — nor by any positive constant. (This fact follows since the density of this distribution decreases to zero faster than any polynomial as one moves away from the origin.)

It is easy to see that if m measures $\mu_1, \dots, \mu_m$ have dimension at most ℓ, then so does their sum. Consequently, if (M_j) is a finite family of compact submanifolds of $\mathbb{R}^d$ with dimensions (d_j), and μ_j is the volume measure on M_j weighted by a function bounded away from zero and infinity, the dimension of the sum $\mu = \sum_{j=1}^m \mu_j$ is at most $\max_j d_j$.

4.2 Bounds

In the remaining of this section, we bound the error between the witnessed k-distance $\mathrm{d}^{\mathrm{w}}_{P,k}$ and the (ordinary) distance d_K to the compact set K. We start from a proposition from [5] that bounds the error between the exact k-distance $\mathrm{d}_{P,k}$ and d_K:

THEOREM 1 (EXACT BOUND). *Let μ denote a probability measure with dimension at most ℓ, and supported on a set K. Consider the uniform measure $\mathbf{1}_P$ on a point cloud P, and set $m_0 = k/|P|$. Then*

$$\|\mathrm{d}_{P,k} - \mathrm{d}_K\|_\infty \leq m_0^{-1/2}\,\mathrm{W}_2(\mu, \mathbf{1}_P) + \alpha_\mu^{-1/\ell} m_0^{1/\ell}.$$

PROOF. Recall that $\mathrm{d}_{P,k} = \mathrm{d}_{\mathbf{1}_P, m_0}$. Using the triangle inequality and Equation (2), one has

$$\begin{aligned}\|\mathrm{d}_{\mathbf{1}_P,m_0} - \mathrm{d}_K\|_\infty &\leq \|\mathrm{d}_{\mu,m_0} - \mathrm{d}_{\mathbf{1}_P,m_0}\|_\infty + \|\mathrm{d}_{\mu,m_0} - \mathrm{d}_K\|_\infty \\ &\leq m_0^{-1/2}\,\mathrm{W}_2(\mu,\mathbf{1}_P) + \|\mathrm{d}_{\mu,m_0} - \mathrm{d}_K\|_\infty\end{aligned}$$

Then, from Lemma 4.7 in [5], $\|\mathrm{d}_{\mu,m_0} - \mathrm{d}_K\|_\infty \leq \alpha_\mu^{-1/\ell} m_0^{1/\ell}$, and the claim follows. □

In the main theorem of this section, the exact k-distance in the above bound is replaced by the witnessed k-distance.

THEOREM 2 (WITNESSED BOUND). *Let μ be a probability measure satisfying the dimension assumption and let K be its support. Consider the uniform measure $\mathbf{1}_P$ on a point cloud P, and set $m_0 = k/|P|$. Then,*

$$\|\mathrm{d}^{\mathrm{w}}_{P,k} - \mathrm{d}_K\|_\infty \leq 6 m_0^{-1/2}\,\mathrm{W}_2(\mu,\mathbf{1}_P) + 24 m_0^{1/\ell}\alpha_\mu^{-1/\ell}.$$

Observe that the error term given by this theorem is a constant factor times the bound in the previous theorem. Before proceeding with the proof, we prove an auxiliary lemma, which emphasizes that a measure ν, close to a measure μ satisfying an upper dimension bound (as in Definition 3), remains concentrated around the support of μ.

LEMMA 2 (CONCENTRATION). *Let μ be a probability measure satisfying the dimension assumption, and ν be another probability measure. Let m_0 be a mass parameter. Then, for every point p in the support of μ, $\nu(\mathrm{B}(p,\eta)) \geq m_0$, where $\eta = m_0^{-1/2}\,\mathrm{W}_2(\mu,\nu) + 4 m_0^{1/2+1/\ell}\alpha_\mu^{-1/\ell}$.*

PROOF. Let π be an optimal transport plan between ν and μ. For a fixed point p in the support of K, let r be the smallest radius such that $\mathrm{B}(p,r)$ contains at least $2m_0$ of mass μ. Consider now a submeasure μ' of μ of mass exactly $2m_0$ and whose support is contained in the ball $\mathrm{B}(p,r)$. This measure is obtained by transporting a submeasure ν' of ν by the optimal transport plan π. Our goal is to determine for what choice of η the ball $\mathrm{B}(p,\eta)$ contains a ν'-mass (and, therefore, a ν-mass) of at least m_0. We make use of the Chebyshev's inequality for ν' to bound the mass of ν' *outside* of the ball $\mathrm{B}(p,\eta)$:

$$\begin{aligned}\nu'(\mathbb{R}^d \setminus \mathrm{B}(p,\eta)) &= \nu'(\{x \in \mathbb{R}^d;\ \|x-p\| \geq \eta\}) \\ &\leq \frac{1}{\eta^2}\int \|x-p\|^2 \mathrm{d}\nu'\end{aligned} \tag{4}$$

Observe that the right hand term of this inequality is exactly the Wasserstein distance between μ' and the Dirac mass $2m_0\delta_p$. We bound it using the triangle inequality for the Wasserstein distance:

$$\begin{aligned}\int \|x-p\|^2 \mathrm{d}\nu' &= \mathrm{W}_2^2(\nu', 2m_0\delta_p) \\ &\leq (\mathrm{W}_2(\mu',\nu') + \mathrm{W}_2(\mu', 2m_0\delta_p))^2 \\ &\leq (\mathrm{W}_2(\mu,\nu) + 2m_0 r)^2\end{aligned} \tag{5}$$

Combining equations (4) and (5), we get:

$$\begin{aligned}\nu(\bar{\mathrm{B}}(p,\eta)) \geq \nu'(\bar{\mathrm{B}}(p,\eta)) &\geq \nu'(\mathbb{R}^d) - \nu'(\mathbb{R}^d \setminus \mathrm{B}(p,\eta)) \\ &\geq 2m_0 - \frac{(\mathrm{W}_2(\mu,\nu) + 2m_0 r)^2}{\eta^2}.\end{aligned}$$

By the lower bound on the dimension of μ, and the definition of the radius r, one has $r \leq (2m_0/\alpha_\mu)^{1/\ell}$. Hence, the ball $\bar{\mathrm{B}}(p,\eta)$ contains a mass of at least m_0 as soon as

$$\frac{(\mathrm{W}_2(\mu,\nu) + \alpha_\mu^{-1} 2^{1+1/\ell} m_0^{1+1/\ell})^2}{\eta^2} \leq m_0.$$

This will be true, in particular, if η is larger than

$$\mathrm{W}_2(\mu,\nu) m_0^{-1/2} + 4\alpha_\mu^{-1/\ell} m_0^{1/2+1/\ell}. \qquad \square$$

Proof of the Witnessed Bound Theorem. Since the witnessed k-distance is a minimum over fewer barycenters, it is larger than the real k-distance. Using this fact and the Exact Bound Theorem one gets the lower bound:

$$\mathrm{d}^{\mathrm{w}}_{P,k} \geq \mathrm{d}_{P,k} \geq \mathrm{d}_K - m_0^{-1/2}\,\mathrm{W}_2(\mu, \mathbf{1}_P) + \alpha_\mu^{-1/\ell} m_0^{1/\ell}$$

For the upper bound, if we set η as in Lemma 2, for every point p in K, the ball $\mathrm{B}(p,\eta)$ contains at least k points in P. Consider one of these points x_1; its $(k-1)$ nearest neighbors $x_2,\dots,x_k$ in P cannot be at a distance greater than 2η from x_1. Hence, the points $x_1,\dots,x_k$ belong to the ball $\mathrm{B}(p,3\eta)$ and so does their barycenter. This shows that the set W of witnessed barycenters, obtained by this construction, is a 3η-covering of K, that is $\mathrm{d}_W \leq \mathrm{d}_K + 3\eta$. Since the weight of any barycenter in W is at most 3η, we get $\mathrm{d}^{\mathrm{w}}_{P,k} \leq \mathrm{d}_W + 3\eta$. To sum up,

$$\mathrm{d}^{\mathrm{w}}_{P,k} \leq \mathrm{d}_W + 3\eta \leq \mathrm{d}_K + 6\eta$$

Replacing η by its value from the Concentration Lemma concludes the proof. $\square$

5. CONVERGENCE UNDER EMPIRICAL SAMPLING

One term remains moot in the bound in Theorem 2, namely the Wasserstein distance $\mathrm{W}_2(\mu, \mathbf{1}_P)$. In this section, we analyze its convergence. The rate depends on the complexity of the measure μ, defined below. The moral of this section is that if a measure can be well approximated with few points, then it is also well approximated by random sampling.

Definition 4. The *complexity* of a probability measure μ at a scale $\varepsilon > 0$ is the minimum cardinality of a finitely supported probability measure ν which ε-approximates μ in the Wasserstein sense, i.e. such that $\mathrm{W}_2(\mu,\nu) \leq \varepsilon$. We denote this number by $\mathcal{N}_\mu(\varepsilon)$.

Observe that this notion is very close to the *ε-covering number* of a compact set K, denoted by $\mathcal{N}_K(\varepsilon)$, which counts the minimum number of balls of radius ε needed to cover K. It's worth noting that if measures μ and ν are close — as are the measure μ and its noisy approximation ν in the previous section — and μ has low complexity, then so does the measure ν. The following lemma shows that measures satisfying the dimension assumption have low complexity. Its proof follows from a classical covering argument, that can be found e.g. in Proposition 4.1 of [15].

Lemma 3 (Dimension-Complexity). *Let K be the support of a measure μ with $\dim\mu \leq \ell$. Then,*

(i) for every positive ε, $\mathcal{N}_K(\varepsilon) \leq \alpha_\mu/\varepsilon^\ell$. Said otherwise, the upper box-counting dimension *of K is bounded: $\dim(K) := \limsup_{\varepsilon\to 0} \log(\mathcal{N}_K(\varepsilon))/\log(1/\varepsilon) \leq \ell$.*

(ii) for every positive ε, $\mathcal{N}_\mu(\varepsilon) \leq \alpha_\mu 5^\ell/\varepsilon^\ell$.

Theorem 3 (Convergence). *Let μ be a probability measure on $\mathbb{R}^d$ whose support has diameter at most D, and let P be a set of N points independently drawn from the measure μ. Then, $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{P}(\mathrm{W}_2(\mathbf{1}_P,\mu) \leq 4\varepsilon) \geq 1 &- \mathcal{N}_\mu(\varepsilon)\exp(-2N\varepsilon^2/(D\mathcal{N}_\mu(\varepsilon))^2) \\ &- \exp(-2N\varepsilon^4/D^2)\end{aligned}$$

Proof. Let n be a fixed integer, and ε be the minimum Wasserstein distance between μ and a measure $\bar\mu$ supported on (at most) n points. Let S be the support of the optimal measure $\bar\mu$, so that $\bar\mu$ can be decomposed as $\sum_{s\in S}\alpha_s\delta_s$ ($\alpha_s \geq 0$). Let π be an optimal transport plan between μ and $\bar\mu$; this is equivalent to finding a decomposition of μ as a sum of n non-negative measures $(\pi_s)_{s\in S}$ such that $\mathrm{mass}(\pi_s) = \alpha_s$, and

$$\sum_{s\in S}\int \|x-s\|^2 \mathrm{d}\pi_s(x) = \varepsilon^2 = \mathrm{W}_2(\mu,\bar\mu)^2$$

Drawing a random point X from the measure μ amounts to (i) choosing a random point s in the set S (with probability α_s) and (ii) drawing a random point X following the distribution π_s. Given N independent points $X_1,\dots,X_N$ drawn from the measure μ, denote by $I_{s,N}$ the proportion of the (X_i) for which the point s was selected in step (i). Hoeffding's inequality allows to easily quantify how far the proportion $I_{s,N}$ deviates from α_s: $\mathbb{P}(|I_{s,N} - \alpha_s| \geq \delta) \leq \exp(-2N\delta^2)$. Combining these inequalities for every point s and using the union bound yields

$$\mathbb{P}\left(\sum_{s\in S}|I_{s,N} - \alpha_s| \leq \delta\right) \geq 1 - n\exp(-2N\delta^2/n^2).$$

For every point s, denote by $\tilde\pi_s$ the distribution of the distances to s in the submeasure π_s, i.e. the measure on the real line defined by $\tilde\pi_s(I) := \pi_s(\{x \in \mathbb{R}^d; \|x-s\| \in I\})$ for every interval I. Define $\tilde\mu$ as the sum of the $\tilde\pi_s$; by the change of variable formula one has

$$\int_{\mathbb{R}} t^2 \mathrm{d}\tilde\mu(t) = \sum_s \int_{\mathbb{R}} t^2 \mathrm{d}\tilde\pi_s = \sum_s \int_{\mathbb{R}^d}\|x-s\|^2\mathrm{d}\pi_s = \varepsilon^2$$

Given a random point X_i sampled from μ, denote by Y_i Euclidean distance between the point X_i and the point s chosen in step (i). By construction, the distribution of Y_i is given by the measure $\tilde{\mu}$; using the Hoeffding inequality again one gets

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i^2 \geq (\varepsilon+\eta)^2\right) \leq 1-\exp(-2N\eta^2\varepsilon^2/D^2).$$

In order to conclude, we need to define a transport plan from the empirical measure $\mathbf{1}_P = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}$ to the finite measure $\bar{\mu}$. To achieve this, we order the points (X_i) by increasing distance Y_i; then transport every Dirac mass $\frac{1}{N}\delta_{X_i}$ to the corresponding point s in S until s is "full", i.e. the mass α_s is reached. The squared cost of this transport operation is at most $\frac{1}{N}\sum_{i=1}^{N} Y_i^2$. Then distribute the remaining mass among the s points in any way; the cost of this step is at most D times $\sum_{s\in S}|I_{s,N}-\alpha_s|$. The total cost of this transport plan is the sum of these two costs. From what we have shown above, setting $\eta=\varepsilon$ and $\delta=\varepsilon/D$, one gets

$$\begin{aligned}\mathbb{P}(\mathrm{W}_2(\mathbf{1}_P,\mu)\leq 4\varepsilon) \geq 1 &- n\exp(-2N\varepsilon^2/(Dn)^2)\\ &-\exp(-2N\varepsilon^4/D^2)\end{aligned}$$

□

As a consequence of the Dimension-Complexity Lemma 3 and of the Convergence Theorem 3, any measure μ satisfying an upper bound on its dimension is well approximated by empirical sampling. A result similar to the Convergence Theorem follows when the samples are drawn not from the original measure μ, but from a "noisy" approximation ν which need not be compactly supported:

Corollary 1 (Noisy Convergence). *Let μ,ν be two probability measures on $\mathbb{R}^d$ with $\mathrm{W}_2(\mu,\nu)=\sigma$, and P be a set of N points drawn independently from the measure ν. Then,*

$$\begin{aligned}\mathbb{P}(\mathrm{W}_2(\mathbf{1}_P,\mu)\leq 9\sigma) \geq 1 &- \mathcal{N}_\mu(\sigma)\exp(-8N\sigma^2/(D\mathcal{N}_\mu(\sigma))^2)\\ &-\exp(-32N\sigma^4/D^2).\end{aligned}$$

Proof. One only needs to apply the previous Convergence Theorem to the measures ν and $\mathbf{1}_P$:

$$\begin{aligned}\mathbb{P}(\mathrm{W}_2(\nu,\mathbf{1}_P)\leq 4\varepsilon) \geq 1 &- \mathcal{N}_\mu(\varepsilon)\exp(-2N\varepsilon^2/(D\mathcal{N}_\nu(\varepsilon))^2)\\ &-\exp(-2N\varepsilon^4/D^2) \qquad (6)\end{aligned}$$

Set $\varepsilon=2\sigma$ and recall that by definition $\mathcal{N}_\nu(2\sigma)\leq\mathcal{N}_\mu(\sigma)$. Then, using $\mathrm{W}_2(\mathbf{1}_P,\mu)\leq\mathrm{W}_2(\mathbf{1}_P,\nu)+\sigma$ one has

$$\mathbb{P}(\mathrm{W}_2(\mathbf{1}_P,\mu)\leq 9\sigma)\geq\mathbb{P}(\mathrm{W}_2(\mathbf{1}_P,\nu)\leq 8\sigma)$$

We conclude by using Eq. (6) with $\varepsilon=2\sigma$. □

It is now possible to combine Theorem 2 (Witnessed Bound), Corollary 1 (Noisy Convergence) and Lemma 3 (Dimension-Complexity) to get the following probabilistic statement.

Theorem 4 (Approximation). *Suppose that μ is a measure satisfying the dimension assumption, supported on a set K of diameter D, and ν a noisy approximation of μ, i.e. $\mathrm{W}_2(\mu,\nu)\leq\sigma$. Let P be a set of N points independently sampled from ν. Then, the inequality*

$$\|\mathrm{d}^{\mathrm{w}}_{P,k}-\mathrm{d}_K\|_\infty \leq 54m_0^{-1/2}\sigma+24m_0^{1/\ell}\alpha_\mu^{-1/\ell}$$

holds with probability at least

$$1-\gamma_\mu\exp(-\beta_\mu N\max(\sigma^{2+2\ell},\sigma^4)-\ell\ln(\sigma)),$$

where $\beta_\mu=\frac{1}{D^2}\max\left[\frac{8}{(\alpha_\mu 5^\ell)^2},32\right]$ *and* $\gamma_\mu=1+\alpha_\mu 5^\ell$.

Proof. Thanks to the Witnessed Bound Theorem and the Noisy Convergence Corollary, the inequality holds with probability at least:

$$1-\mathcal{N}_\mu(\sigma)\exp(-8N\sigma^2/(D\mathcal{N}_\mu(\sigma))^2)-\exp(-32N\sigma^4/D^2)$$

We use Lemma 3 to lower bound the covering number $\mathcal{N}_\mu(\sigma)$ by $\alpha_\mu 5^\ell/\sigma^\ell$. Hence, the previous expression is bounded from below by

$$\begin{aligned}1-\alpha_\mu 5^\ell &\exp(-8N\sigma^{2+2\ell}/(D\alpha_\mu 5^\ell)^2\\ &-\ell\ln(\sigma))-\exp(-32N\sigma^4/D^2)\\ &\geq 1-\gamma_\mu\exp(-\beta_\mu N\max(\sigma^{2+2\ell},\sigma^4)-\ell\ln(\sigma))\end{aligned}$$

where $\gamma_\mu=1+\alpha_\mu 5^\ell$ and $\beta_\mu=\frac{1}{D^2}\max\left[\frac{8}{(\alpha_\mu 5^\ell)^2},32\right]$, as stated in the theorem. □

6. DISCUSSION

We illustrate the utility of the bound in the Witnessed Bound Theorem by example and an inference statement. Figure 1 shows 6000 points drawn from the uniform distribution on a sideways figure-8 (in red), convolved with a Gaussian distribution. The ordinary distance function to the point set has no hope of recovering geometric information out of these points since both loops of the figure-8 are filled in. On the right, we show the sublevel sets of the distance to the uniform measure on the point set, both the witnessed k-distance and the exact k-distance. Both functions recover the topology of figure-8, the bits missing from the witnessed k-distance smooth out the boundary of the sublevel set, but do not affect the image at large.

Inference. Suppose that we are in the conditions of the Approximation Theorem, but additionally we assume that the support K of the original measure μ has a *weak feature size* larger than R. This means that the distance function d_K has no critical value in $[0,R]$, and implies that all the offsets $K^r=\mathrm{d}_K^{-1}[0,r]$ of K are homotopy equivalent for $r\in(0,R)$. Suppose again that we have drawn a set P of N points from a Wasserstein approximation ν of μ, such that $\mathrm{W}_2(\mu,\nu)\leq\sigma$. From the Approximation Theorem, we have

$$\|\mathrm{d}^{\mathrm{w}}_{P,k}-\mathrm{d}_K\|_\infty\leq e(m_0):=54m_0^{-1/2}\sigma+24m_0^{1/\ell}\alpha_\mu^{-1/\ell}$$

with high probability as N goes to infinity. Then, the standard argument [8] shows that the Betti numbers of the compact set K can be inferred from the function $\mathrm{d}^{\mathrm{w}}_{P,k}$, which is defined only from the point sample P, as long as $e(m_0)$ is less than $R/4$. Indeed, denoting by K^r and P^r the r-sublevel sets of the functions d_K and $\mathrm{d}^{\mathrm{w}}_{P,k}$, the sequence of inclusions

$$K^0\subseteq P^{e(m_0)}\subseteq K^{2e(m_0)}\subseteq P^{3e(m_0)}\subseteq K^{4e(m_0)}$$

holds with high probability. By assumption the function d_K has no critical values in the range $(0,4e(m_0))\subseteq(0,R)$. Therefore, the rank of the image on the homology induced by inclusion $\mathsf{H}(P^{e(m_0)})\to\mathsf{H}(P^{3e(m_0)})$ is equal to the Betti numbers of the set K. In the language of persistent homology [13], the persistent Betti numbers $\beta^{(e(m_0),3e(m_0))}$ of the function $\mathrm{d}^{\mathrm{w}}_{P,k}$ are equal to the Betti numbers of the set K.

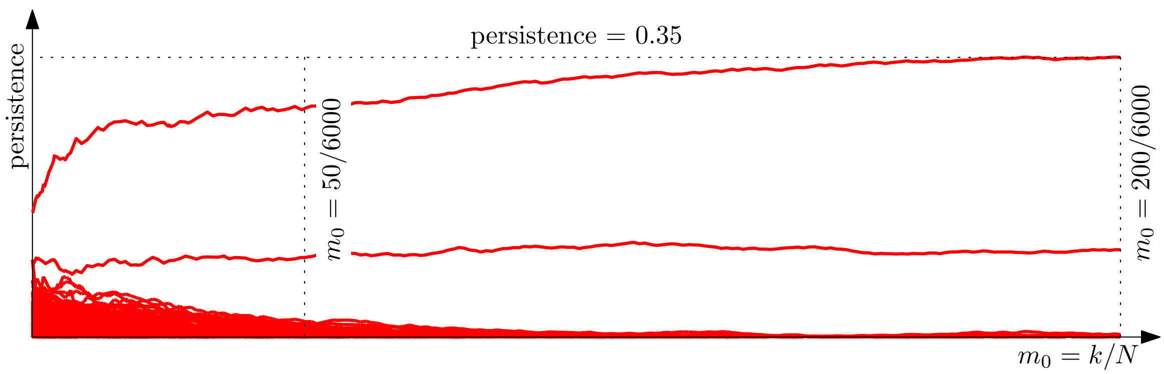


Figure 2: (PL-approximation of the) 1-dimensional persistence vineyard of the witnessed k-distance function. Topological features of the space, obscured by noise for low values of m_0, stand out as we increase the mass parameter.

Choice of the mass parameter. This language also suggests a strategy for choosing a mass parameter m_0 for the distance to a measure, a question that has not been addressed by the original paper [5]. For every mass parameter m_0, the p-dimensional *persistence diagram* $\mathrm{Pers}_p(\mathrm{d}_{\mu,m_0})$ is a set of points $\{(b_i(m_0), d_i(m_0))\}_i$ in the extended plane $(\mathbb{R} \cup \{\infty\})^2$. Each of these points represents a homology class of dimension p in the sublevel sets of d_{μ,m_0}; $b_i(m_0)$ and $d_i(m_0)$ are the values at which it is born and dies. Since the distance to measure $\mathrm{d}_{\mathbf{1}_P,m_0}$ depends continuously on m_0, by [8] so do its persistence diagrams. Thus, one can use the algorithm in [9] to track their evolution. Figure 2 illustrates such a construction for the point set in Figure 1 and the witnessed k-distance. It displays the evolution of the persistence $(d_1(m_0) - b_1(m_0))$ of each of the 1-dimensional homology classes as m_0 varies, thus highlighting the choices of the mass parameter that lead to the presence of the two prominent classes (corresponding to the two loops of the figure-8).

Acknowledgement

This work has been partly supported by a grant from the French ANR, ANR-09-BLAN-0331-01, NSF grants FODAVA 0808515, CCF 1011228, and NSF/NIH grant 0900700. The second author would also like to acknowledge the support of the Fields Institute during the redaction of this article.

7. REFERENCES

[1] N. Amenta and M. Bern. Surface reconstruction by Voronoi filtering. *Discrete and Computational Geometry*, 22(4):481–504, 1999.

[2] S. Arya and D. Mount. Computational geometry: proximity and location. *Handbook of Data Structures and Applications*, pages 63.1–63.22, 2005.

[3] F. Bolley, A. Guillin, and C. Villani. Quantitative Concentration Inequalities for Empirical Measures on Non-compact Spaces. *Probability Theory and Related Fields*, 137(3):541–593, 2007.

[4] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in Euclidean space. *Discrete and Computational Geometry*, 41(3):461–479, 2009.

[5] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. Preprint (INRIA RR-6930 v2), 2010.

[6] F. Chazal and S. Oudot. Towards persistence-based reconstruction in Euclidean spaces. *Proceedings of the ACM Symposium on Computational Geometry*, pages 232–241, 2008.

[7] K. Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, pages 15–59, 2006.

[8] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

[9] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. In *Proceedings of the ACM Symposium on Computational Geometry*, pages 119–126, 2006.

[10] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, page 634, 1999.

[11] T. Dey and S. Goswami. Provable surface reconstruction from noisy samples. *Computational Geometry*, 35(1-2):124–141, 2006.

[12] H. Edelsbrunner. The union of balls and its dual shape. *Discrete and Computational Geometry*, 13:415–440, 1995.

[13] H. Edelsbrunner and J. Harer. Persistent homology — a survey. *Surveys on Discrete and Computational Geometry. Twenty Years Later*, pages 257–282, 2008.

[14] P. Indyk. Nearest neighbors in high-dimensional spaces. *Handbook of Discrete and Computational Geometry*, pages 877–892, 2004.

[15] B. Kloeckner. Approximation by finitely supported measures. Preprint (arXiv:1003.1035), 2010.

[16] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. Preprint, 2008.

[17] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1):419–441, 2008.

[18] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.