Computational Geometry 58 (2016) 70–96

Contents lists available at ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

Efficient and robust persistent homology for measures

Mickaël Buchet *, Frédéric Chazal, Steve Y. Oudot, Donald R. Sheehy

ARTICLE INFO

Article history:
Received 27 May 2014
Received in revised form 11 August 2015
Accepted 3 June 2016
Available online 7 July 2016

Keywords:
Persistent homology
Topological data analysis
Distance to a measure
Power distance
Sparse rips filtration

ABSTRACT

A new paradigm for point cloud data analysis has emerged recently, where point clouds are no longer treated as mere compact sets but rather as empirical measures. A notion of distance to such measures has been defined and shown to be stable with respect to perturbations of the measure. This distance can easily be computed pointwise in the case of a point cloud, but its sublevel-sets, which carry the geometric information about the measure, remain hard to compute or approximate. This makes it challenging to adapt many powerful techniques based on the Euclidean distance to a point cloud to the more general setting of the distance to a measure on a metric space.

We propose an efficient and reliable scheme to approximate the topological structure of the family of sublevel-sets of the distance to a measure. We obtain an algorithm for approximating the persistent homology of the distance to an empirical measure that works in arbitrary metric spaces. Precise quality and complexity guarantees are given with a discussion on the behavior of our approach in practice.

1. Introduction

Given a sample of points P from a metric space $\mathbb{X}$, the distance function d_P maps each $x \in \mathbb{X}$ to the distance from x to the nearest point of P. The related fields of geometric inference and topological data analysis have provided a host of theorems about what information can be extracted from the distance function, with a particular focus on discovering and quantifying intrinsic properties of the shape underlying a data set [4,19]. The flagship tool in topological data analysis is persistent homology and the most common goal is to apply the persistence algorithm to distance functions, either in Euclidean space or in metric spaces [1,14,23]. From the very beginning, this line of research encountered two major challenges. First, distance functions are very sensitive to noise and outliers (Fig. 1 left). Second, the representations of the sublevel sets of a distance function become prohibitively large even for moderately sized data. These two challenges led to two distinct research directions. First, the distance to the data set was replaced with a distance to a measure induced by that data set [5]. The resulting theory is provably more robust to outliers, but the sublevel sets become even more complex to represent (Fig. 1 center). Towards more efficient representations, several advances in *sparse filtrations* have led to linear-size constructions [12,20,21], but all of these methods exploit the specific structure of the distance function and do not obviously generalize. In this paper, we bring these two research directions together by showing how to combine the robustness of the distance to a measure, with the efficiency of sparse filtrations.

* Corresponding author.
E-mail addresses: mickael.buchet@m4x.org (M. Buchet), frederic.chazal@inria.fr (F. Chazal), steve.oudot@inria.fr (S.Y. Oudot), don.r.sheehy@gmail.com (D.R. Sheehy).

http://dx.doi.org/10.1016/j.comgeo.2016.07.001
0925-7721/

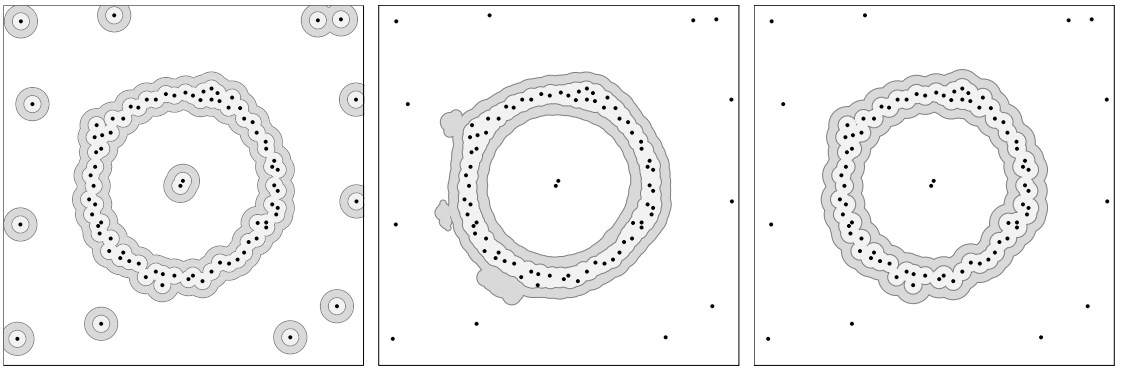


Fig. 1. From left to right, two sublevel sets for d_P, $d_{\mu_P,m}$, and $d^P_{\mu_P,m}$ with $m = \frac{3}{|P|}$. The first is too sensitive to noise and outliers. The second is smoother, but substantially more difficult to compute. The third is our approximation, which is robust to noise, efficient to compute, and compact to represent.

Contributions:

1. **A Generalization of the Wasserstein stability and persistence stability of the distance to a measure for triangulable metric spaces.**
2. **A general method for approximating the sublevel sets of the distance to a measure by a union of balls.** Our method uses $O(n)$ balls for inputs of n samples. Known methods for representing the exact sublevel sets can require $n^{\Theta(d)}$ balls. Existing approximations using a linear number of balls are only applicable in Euclidean space [15].
3. **A linear size approximation to the weighted Rips filtration.** For intrinsically low-dimensional metric spaces, we construct a filtration of size $O(n)$ that achieves a guaranteed quality approximation. Specifically, if the doubling dimension of the metric is d then the size complexity is $2^{O(dk)}n$ if one considers simplices up to dimension k (see Definition 2.1 for the formal definition of doubling dimension). This is a significant improvement over the full weighted Rips filtration, which has size 2^n in general or size n^{k+1} if one considers only simplices up to dimension k. It also has the advantage that the sparsification is independent of the weights. Thus, the (geo)metric preprocessing phase can be reused for any weighting of the points. If one attempted to use previous methods directly, this preprocessing phase would have to be repeated for each set of weights. This is especially useful if one is interested in several different weight functions such as when approximating the distance to a measure for several different values of the mass parameter.
4. **An effective implementation with experimental results.**

Overview of the paper Originally, the distance to a measure was introduced to capture information about both scale and density in a Euclidean point cloud. We extend the distance to a measure to any metric space $\mathbb{X}$. We write $\bar{B}(x,r)$ to denote the closed ball with center x and radius r. The distance to a measure is then defined as follows.

Definition 1.1. Let μ be a probability measure on a metric space $\mathbb{X}$ and let $m \in]0,1]$ be a mass parameter. We define the distance $d_{\mu,m}$ to the measure μ as

$$d_{\mu,m} : x \in \mathbb{X} \mapsto \sqrt{\frac{1}{m}\int_0^m \delta_{\mu,l}(x)^2 dl},$$

where $\delta_{\mu,m}$ is defined as

$$\delta_{\mu,m} : x \in \mathbb{X} \mapsto \inf\{r > 0 \mid \mu(\bar{B}(x,r)) > m\}.$$

The distance to a measure has interesting inference and stability results in the Euclidean setting [5]. That is, the sublevel sets of the function can be used to infer the topology of the support of the underlying distribution (inference), and also, the output for similar inputs will be similar (stability). In Section 3, we extend these stability results to any metric space. The results about the stability of persistence diagrams apply to any triangulable metric space, i.e. metric spaces homeomorphic to a locally finite simplicial complex (the persistence diagram may not exist for non-triangulable metric spaces).

We then give a new way to approximate the distance to a measure. Using a sampling of the support of a measure, we are able to compute accurately the sublevel sets of the distance to a measure in any metric space, using power distances. We show in Section 4.1 that these functions have adequate stability and approximation properties. Then, in Section 4.2, we give the practical implications for computing persistence diagram for finite samples.

The *witnessed k-distance* is another approach to approximating the distance to a measure proposed in [15]. This approach works only in Euclidean spaces as it relies on the existence of barycenters of points. The analysis links the quality of the

approximation to the underlying topological structure. In this paper, we look at bounds independent of intrinsic geometry. When restricted to the Euclidean setting in section 4.3, our method improves the approximation bounds from [15]. The new bounds match the quality of approximation achieved by our method of Section 4.1, which has the added advantage that it is valid in any metric space.

In Section 5, we introduce the *weighted Rips complex*. Given a parameter, the sublevel set of a power distance associated with this parameter is a union of balls. Generalizing the Vietoris–Rips complex, we define the weighted Rips complex as the clique complex whose 1-skeleton is the same as the one of the nerve of this union of balls. The induced filtration has important stability properties and can be used to approximate persistence diagrams.

Unfortunately, the weighted Rips filtration is too large to construct in full for large instances. This problem already exists with the usual Rips filtration. Sparsifying schemes have been recently proposed in [12,21]. Extending the approach used in [21], we construct a sparse approximation that has linear size in the number of points (Section 6). This can be used to approximate persistence diagrams even for high dimensional inputs if the data is intrinsically low dimensional. As we show in Section 6, there are very simple examples where the input metric is intrinsically low-dimensional and yet the weight function can cause the weighted distance function to be high-dimensional. Our approach has the advantage over previous methods in that the size complexity will only depend on the dimension of the input metric, rather than the dimension of points under the weighted distance.

The combination of these approaches makes it possible to use the distance to a measure to infer topology on real instances. In Section 7, we illustrate the theory with some examples and results from an implementation.

2. Background

In this paper, we consider a metric space $\mathbb{X}$ with distance $d_{\mathbb{X}}(\cdot,\cdot)$. In a slight abuse of notation, we also write $d_{\mathbb{X}}$ to denote the distance between a point and a set defined as $d_{\mathbb{X}}(x,P)=\inf_{p\in P} d_{\mathbb{X}}(x,p)$. The Hausdorff distance between two sets P and Q will be denoted $d_H(P,Q)$. We write $B(x,r)$ for the open ball of center x and radius r in $d_{\mathbb{X}}$, and we write $\bar{B}(x,r)$ for the corresponding closed ball.

Metric spaces and doubling dimension For metric spaces that are not embedded in Euclidean space, the doubling dimension gives a useful way to describe the intrinsic dimension of the metric space by bounding the size of certain covers of subsets. Formally it is defined as follows.

Definition 2.1. The *doubling constant* $\lambda_{\mathbb{X}}$ of a metric space $\mathbb{X}$ is the maximum over all balls $B(x,r)$ with $x\in\mathbb{X}$ of the minimum number of balls of radius $r/2$ required to cover $B(x,r)$. The *doubling dimension* is defined to be $\log_2(\lambda_{\mathbb{X}})$.

Wasserstein distance

To compare measures, we use the Wasserstein distance, also called the earth-mover distance. Intuitively, it is the minimal cost to move all the mass from one measure to another. To state the formal definition we first introduce some notation.

Given a measure μ on a metric space $\mathbb{X}$, we write $\mathfrak{B}(\mathbb{X})$ to denote the set of all Borel subsets of $\mathbb{X}$. Given $A\in\mathfrak{B}(\mathbb{X})$, we define the *mass of* A as $\mu(A)$. Similarly $\mu(\mathbb{X})$ is called the *total mass* of μ. We write $\mathrm{Supp}(\mu)$ for the support of the measure μ.

Definition 2.2. Let μ and ν be positive measures with the same total mass on a metric space $\mathbb{X}$. A *transport plan* between μ and ν is a measure π on $\mathbb{X}\times\mathbb{X}$ such that for all $A,B\in\mathfrak{B}(\mathbb{X})$,

$$\pi(A\times\mathbb{X})=\mu(A) \text{ and } \pi(\mathbb{X}\times B)=\nu(B).$$

We denote by $\Pi(\mu,\nu)$ the set of all transport plans between μ and ν. The pth order cost of the transport plan π is defined as

$$C_p(\pi)=\left(\int_{\mathbb{X}\times\mathbb{X}} d_{\mathbb{X}}(x,y)^p d\pi(x,y)\right)^{\frac{1}{p}}.$$

The Wasserstein distance between μ and ν is the minimum cost over all transport plans.

Definition 2.3. Let μ and ν be positive measures with the same total mass on a metric space $\mathbb{X}$. The *Wasserstein distance* of order p between μ and ν is defined as

$$W_p(\mu,\nu)=\min_{\pi\in\Pi(\mu,\nu)}\left(\int_{\mathbb{X}\times\mathbb{X}} d_{\mathbb{X}}(x,y)^p d\pi(x,y)\right)^{\frac{1}{p}}.$$

The Wasserstein distance is finite if both probability measures have finite p-moments, which is always the case for measures with compact support.

Persistence theory

A *filtration* $F = \{F_\alpha\}_{\alpha\in\mathbb{R}}$ is a sequence of spaces such that $F_\alpha \subseteq F_\beta$ whenever $\alpha \leq \beta$. Persistence theory studies the evolution of the homology of the sets F_α for α ranging from $-\infty$ to $+\infty$. More precisely, the filtration induces a family of vector spaces connected by linear maps at the homology level, called a *persistence module*. More generally, a persistence module is a pair $\mathbb{V} = (\{V_\alpha\}, \{v_\alpha^\beta\})$ where each V_α is a vector space and v_α^β is a linear map $V_\alpha \to V_\beta$ such that $v_\beta^\gamma \circ v_\alpha^\beta = v_\alpha^\gamma$ for all $\alpha \leq \beta \leq \gamma$ and v_α^α is the identity. A persistence module is said to be *q-tame* if v_α^β has finite rank for every $\alpha < \beta$. A filtration is said to be q-tame if its corresponding persistence module is q-tame. The algebraic structure of a q-tame persistence module $\mathbb{U}$ can be described and visualized by the *persistence diagram* $\mathrm{Dgm}(\mathbb{U})$, a multiset of points in the plane. If $\mathbb{U}$ comes from a filtration $\{F_\alpha\}$, a point (α, β) in $\mathrm{Dgm}(\mathbb{U})$ indicates a nontrivial homology class that exists in the filtration between the parameter values α and β.

We overload notation and write $\mathrm{Dgm}(\{F_\alpha\})$ to denote the persistence diagram of the persistence module defined by the filtration $\{F_\alpha\}$. Moreover, for a real-valued function f, we write $\mathrm{Dgm}(f)$ to denote $\mathrm{Dgm}(\{f^{-1}(]-\infty, \alpha])\})$, the persistence diagram of the sublevel sets filtration of f. For an introduction to persistent homology, the reader is directed to [6,13].

Bottleneck distance

We put a metric on the space of persistence diagrams as follows. First, a partial matching M between diagrams D and E is a subset of $D \times E$ in which each element of $D \cup E$ appears in at most one pair. The bottleneck cost of M is $\max_{(d,e)\in M} \|d - e\|_\infty$. We say M is an ϵ-matching if the bottleneck cost is ϵ and every (α, β) in D or E with $|\beta - \alpha| \geq 2\epsilon$ is matched. The *bottleneck distance* between D and E is defined as

$$d_B(D, E) = \inf\{\epsilon \mid \text{there exists an } \epsilon\text{-matching between } D \text{ and } E\}.$$

It is often useful to look at persistence diagrams on a logarithmic scale, because the distance does no longer depend on the scale at which the object is seen. The *log-bottleneck distance*, denoted $d_B^{\log}$ is the bottleneck distance between diagrams after the change of coordinates $(\alpha, \beta) \mapsto (\ln\alpha, \ln\beta)$.

Filtration interleaving

One way to prove that two persistence diagrams are close is to prove that the filtrations inducing them are interleaved. Two filtrations $\{U_\alpha\}_{\alpha\in\mathbb{R}}$ and $\{V_\alpha\}_{\alpha\in\mathbb{R}}$ are said to be *ϵ-interleaved* if for any α,

$$U_\alpha \subseteq V_{\alpha+\epsilon} \subseteq U_{\alpha+2\epsilon}.$$

The following classic result [2,6,10] about stability of persistence diagrams says that interleaved filtrations yield similar persistence diagrams.

Theorem 2.4. *Let U and V be two q-tame and ϵ-interleaved filtrations. Then, the persistence diagrams of these filtrations are ϵ-close in bottleneck distance, i.e.,*

$$d_B(\mathrm{Dgm}(U), \mathrm{Dgm}(V)) \leq \epsilon.$$

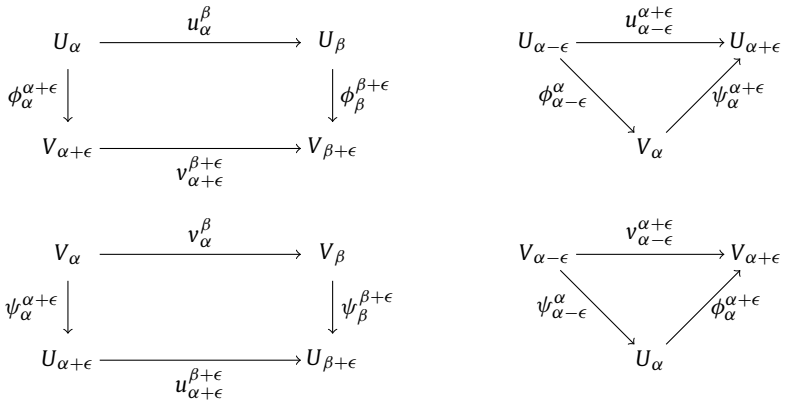
We work with the persistence theory on functions, which means studying the persistence of *the sublevel sets filtration* defined as $\{f^{-1}(]-\infty, \alpha])\}_{\alpha\in\mathbb{R}}$ for any real-valued function. To simplify notation, we write $\mathrm{Dgm}(f)$ to denote the persistence diagram of the sublevel sets filtration of f.

Persistence module interleaving

The notion of interleaving can be extended to persistence modules as seen in [7]. Given two persistence modules $\mathbb{U} = (\{U_\alpha\}, \{u_\alpha^\beta\})$ and $\mathbb{V} = (\{V_\alpha\}, \{v_\alpha^\beta\})$ and a real $\epsilon > 0$, an *ϵ-homomorphism* from $\mathbb{U}$ to $\mathbb{V}$ is a collection of linear maps $\Phi = \{\phi_\alpha\}$ such that for all $\alpha < \beta$, $v_{\alpha+\epsilon}^{\beta+\epsilon} \circ \phi_\alpha = \phi_\beta \circ u_\alpha^\beta$. Two ϵ-homomorphisms Φ from $\mathbb{U}$ to $\mathbb{V}$ and Ψ from $\mathbb{V}$ to $\mathbb{W}$ can be composed to build a 2ϵ-homomorphism $\Psi\Phi$ from $\mathbb{U}$ to $\mathbb{W}$ whose linear maps are obtained by composing the linear maps of Φ and Ψ. Among ϵ-homomorphisms from $\mathbb{U} \to \mathbb{U}$, one has a particular role. The *ϵ-shift map* $1_{\mathbb{U}}^\epsilon$ is the collection of maps $u_\alpha^{\alpha+\epsilon}$ given in the persistence module $\mathbb{U}$. We use it to define the interleaving of two persistence modules as follows.

Definition 2.5. Let $\mathbb{U}$ and $\mathbb{V}$ be two q-tame persistence modules. $\mathbb{U}$ and $\mathbb{V}$ are *ϵ-interleaved* if there exists ϵ-homomorphisms $\Phi : \mathbb{U} \to \mathbb{V}$ and $\Psi : \mathbb{V} \to \mathbb{U}$ such that $\Phi\Psi = 1_{\mathbb{V}}^{2\epsilon}$ and $\Psi\Phi = 1_{\mathbb{U}}^{2\epsilon}$.

Note that the definition is equivalent to the commutativity of the following diagrams for any $\alpha < \beta$, where $\Phi = \{\phi_\alpha\}$ and $\Psi = \{\psi_\alpha\}$.



The following theorem is an algebraic analog of Theorem 2.4. The proof can be found in [6].

Theorem 2.6. *Let $\mathbb{U}$ and $\mathbb{V}$ be two q-tame and ϵ-interleaved persistence modules. Then,*

$$d_B(\mathrm{Dgm}(\mathbb{U}), \mathrm{Dgm}(\mathbb{V})) \leq \epsilon.$$

Contiguous simplicial maps

Let X and Y be simplicial complexes. A *simplicial map* $f : X \to Y$ is a map between the corresponding vertex sets so that for every simplex $\sigma \in X$, $f(\sigma) = \bigcup_{p\in\sigma} f(p)$ is a simplex in Y. Two simplicial maps f and g are *contiguous* if $\sigma \in X$ implies that $f(\sigma) \cup g(\sigma) \in Y$. If two simplicial maps are contiguous, then they induce the same homomorphism at the homology level [18, Chapter 1].

A *clique complex* is a simplicial complex whose simplices are the cliques of a graph. Many of the simplicial complexes considered in this paper are clique complexes. We will use the following simple lemma to construct contiguous simplicial maps between clique complexes.

Lemma 2.7. *Let X and Y be clique complexes and let f and g be two functions from the vertex set of X to the vertex set of Y. If for every edge $(p, q) \in X$, the tetrahedron $\{f(p), g(p), f(q), g(q)\}$ is in Y, then f and g induce contiguous simplicial maps from X to Y.*

Proof. Let σ be a simplex of X. Every pair in $f(\sigma) \cup g(\sigma)$ is of the form $(f(p), f(q))$, $(f(p), g(q))$, or $(g(p), g(q))$ for some vertices p and q in σ. Since $(p, q) \in \sigma$, the tetrahedron hypothesis of the lemma implies that all of these pairs are edges of Y. Thus, $f(\sigma) \cup g(\sigma)$ is a simplex in Y because Y is a clique complex. Moreover, $f(\sigma) \in Y$ and $g(\sigma) \in Y$ because simplices are closed under taking subsets. Therefore, f and g are indeed contiguous simplicial maps as desired. □

3. Persistence and stability of the distance to a measure in a metric space

In this section, we prove that, if we have two close probability measures, then the persistence diagrams of the sublevel sets filtration of their distance to measure functions are close. The result applies to *triangulable* metric spaces, i.e., those that are homeomorphic to a locally finite simplicial complex. The persistence diagrams considered in this paper are well defined in this class of spaces. In particular, every compact Riemannian manifold is triangulable.

If the persistence diagram is to be meaningful, one might expect that it is stable with respect to perturbations in the underlying measure. The following theorem shows that this is indeed the case. Two measures that are close in the quadratic Wasserstein distance, W_2 yield persistence diagrams that are close in bottleneck distance, d_B (see [22, Sec. 7.1]).

Theorem 3.1. *Let μ and ν be two probability measures on a triangulable metric space $\mathbb{X}$ and let m be a mass parameter. Then* $\mathrm{Dgm}(d_{\mu,m})$ *and* $\mathrm{Dgm}(d_{\nu,m})$ *are well-defined and*

$$d_B(\mathrm{Dgm}(d_{\mu,m}), \mathrm{Dgm}(d_{\nu,m})) \leq \frac{1}{\sqrt{m}} W_2(\mu, \nu).$$

To prove this theorem, we first show that the distance to measure functions are stable with respect to the Wasserstein distance. Then, we prove that their diagrams are well-defined and are close using Theorem 2.4. This result provides the same bound as the existing Euclidean space result. Technicalities exist for the well-definition of persistence diagrams which required the metric space to be triangulable.

3.1. Wasserstein stability

A measure ν is a *submeasure* of a measure μ if for every $B \in \mathfrak{B}(\mathbb{X})$, $\nu(B) \leq \mu(B)$. Let $\mathrm{Sub}_m(\mu)$ be the set of all submeasures of μ, which have a total mass m.

The distance to a measure μ at point a x can be expressed as the Wasserstein distance between two measures, the Dirac mass δ_x on x and a submeasure of μ of mass m. Using this view, we generalize the stability result from [5] as follows.

Proposition 3.2. *Let μ be a probability measure on a metric space $\mathbb{X}$, and let $m \in]0,1]$ be a mass parameter. Then,*

$$d_{\mu,m}(x) = \min_{\nu \in \mathrm{Sub}_m(\mu)} \frac{1}{\sqrt{m}} W_2(m\delta_x, \nu).$$

Given $x \in \mathbb{X}$ and $m > 0$, let $\mathcal{R}_{\mu,m}(x)$ be the set of the submeasures of μ with total mass m whose support is contained in the closed ball $\bar{B}(x, \delta_{\mu,m}(x))$ and whose restriction to the open ball $B(x, \delta_{\mu,m(x)})$ coincides with μ. The proof shows that $\mathcal{R}_{\mu,m}(x)$ is exactly the set of minimizers of Proposition 3.2.

In order to prove this theorem we need to introduce a few definitions. The *cumulative function* $F_\nu : \mathbb{R}^+ \to \mathbb{R}$ of a measure ν on $\mathbb{R}^+$ is the non-decreasing function defined by $F_\nu(y) = \nu([0, y))$. Its *generalized inverse* $F_\nu^{-1} : m \mapsto \inf\{t \in \mathbb{R} \mid F_\nu(t) > m\}$ is left-continuous.

Proof. If ν is a measure of total mass m on $\mathbb{X}$ then there exists only one transport plan between ν and the Dirac mass $m\delta_x$. It transports every point of $\mathbb{X}$ to x. Hence we get

$$W_2(m\delta_x, \nu)^2 = \int_{\mathbb{X}} d_{\mathbb{X}}(h, x)^2 \, d\nu(h).$$

Let $d_x : \mathbb{X} \to \mathbb{R}$ denote the distance function to the point x and let ν_x be the pushforward of ν by the distance function to x. That is, for any subset I of $\mathbb{R}$, $\nu_x(I) = \nu(d_x^{-1}(I))$. Note that $F_{\nu_x}^{-1}(m) = \delta_{\nu,m}(x)$. Using the change of variable formula and the definition of the cumulative function we get:

$$\int_{\mathbb{X}} d_{\mathbb{X}}(h, x)^2 d\nu(h) = \int_{\mathbb{R}^+} t^2 d\nu_x(t) = \int_0^m F_{\nu_x}^{-1}(l)^2 dl.$$

Suppose further that ν is a submeasure of μ, then $F_{\nu_x}(t) \leq F_{\mu_x}(t)$ for all $t > 0$. So, $F_{\nu_x}^{-1}(l) \geq F_{\mu_x}^{-1}(l)$ for all $l > 0$, and thus,

$$W_2(m\delta_x, \nu)^2 \geq \int_0^m F_{\mu_x}^{-1}(l)^2 dl = \int_0^m \delta_{\mu,l}(x)^2 dl = m d_{\mu,m}(x)^2. \tag{1}$$

This inequality implies that $d_{\mu,m}(x)$ is smaller than $\frac{1}{\sqrt{m}} W_2(m\delta_x, \nu)$ for any $\nu \in \mathrm{Sub}_m(\mu)$.

Consider the case when the inequality in (1) is tight. Such a case happens when for almost every $l \leq m$, $F_{\nu_x}^{-1}(l) = F_{\mu_x}^{-1}(l)$. Since these functions are increasing and left-continuous, equality must hold for every such l. By the definition of the pushforward, this implies that $\nu(\bar{B}(x, \delta_{\mu,m}(x))) = m$, i.e., all the mass of ν is contained in the closed ball $\bar{B}(x, \delta_{\mu,m}(x))$, and that $\nu(B(x, \delta_{x,\mu}(m))) = \mu(B(x, \delta_{x,\mu}(m)))$. Because ν is a submeasure of μ this is true if and only if ν is in the set $\mathcal{R}_{\mu,m}(x)$ described before the proof. Thus $\mathcal{R}_{\mu,m}(x)$ is exactly the set of submeasures $\nu \in \mathrm{Sub}_m(\mu)$ such that $d_{\mu,m}(x) = \frac{1}{\sqrt{m}} W_2(m\delta_x, \nu)$.

To conclude the proof we need only show that there exists at least one measure $\mu_{x,m}$ in the set $\mathcal{R}_{\mu,m}(x)$. If $\mu(\bar{B}(x, \delta_{\mu,m}(x))) = m$, then $\mu_{x,m} = \mu|_{\bar{B}(x,\delta_{\mu,m}(x))}$ is an obvious choice. The only difficulty is when the boundary $\partial B(x, \delta_{\mu,m}(x))$ of the ball has too much mass. In this case we uniformly rescale the mass contained in the bounding sphere such that the measure $\mu_{x,m}$ has total mass m. More precisely we let:

$$\mu_{x,m} = \mu|_{B(x,\delta_{\mu,m}(x))} + (m - \mu(B(x, \delta_{\mu,m}(x)))) \frac{\mu|_{\partial B(x,\delta_{\mu,m}(x))}}{\mu(\partial B(x, \delta_{\mu,m}(x)))}.$$

We hence have $\frac{1}{\sqrt{m}} W_2(m\delta_x, \mu_{x,m}) = d_{\mu,m}(x)$. □

From this result, we have the following Wasserstein stability guarantee for the distance to a measure.

Theorem 3.3. *Let μ and ν be two probability measures on a metric space $\mathbb{X}$ and let $m \in]0,1]$ be a mass parameter. Then:*

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

Proof. Using Proposition 3.2, we get that $\sqrt{m}\, d_{\mu,m}(x) = W_2(m\delta_x, \mu_{x,m})$, where $\mu_{x,m} \in \mathcal{R}_{\mu,m}(x)$. Let π be an optimal transport plan between μ and ν, i.e., a transport plan between μ and ν such that

$$\int_{\mathbb{X}\times\mathbb{X}} d_{\mathbb{X}}(x,y)^2 d\pi(x,y) = W_2(\mu,\nu)^2.$$

Let us consider the submeasure $\mu_{x,m}$ of μ. Then there exists $\tilde{\pi}$ a submeasure of π that transports $\mu_{x,m}$ to a submeasure $\tilde{\nu}$ of ν. We get that:

$$W_2(\mu_{x,m}, \tilde{\nu}) \leq W_2(\mu,\nu).$$

Using Proposition 3.2 again, we get that for any $x \in \mathbb{X}$, $\sqrt{m}\, d_{\nu,m}(x) \leq W_2(m\delta_x, \tilde{\nu})$. Thus,

$$\begin{aligned}\sqrt{m}\, d_{\nu,m}(x) &\leq W_2(m\delta_x, \tilde{\nu}) \leq W_2(m\delta_x, \mu_{x,m}) + W_2(\tilde{\nu}, \mu_{x,m}) \\ &\leq \sqrt{m}\, d_{\mu,m}(x) + W_2(\mu,\nu).\end{aligned}$$

The roles of μ and ν can be reversed to conclude the proof. □

Another consequence of Proposition 3.2 is that $d_{\mu,m}$ is 1-Lipschitz with respect to x.

Proposition 3.4. *Let μ be a probability measure on a metric space $\mathbb{X}$ and let $m \in]0,1]$ be a mass parameter. Then $d_{\mu,m}$ is 1-Lipschitz.*

Proof. Let x and y be two points of $\mathbb{X}$. Using Proposition 3.2, there exists a submeasure $\mu_{x,m}$ of μ such that $d_{\mu,m}(x) = \frac{1}{\sqrt{m}} W_2(m\delta_x, \mu_{x,m})$. The same proposition applied to y gives $d_{\mu,m}(y) \leq \frac{1}{\sqrt{m}} W_2(m\delta_y, \mu_{x,m})$. Knowing that $W_2(m\delta_x, m\delta_y) = \sqrt{m}\, d_{\mathbb{X}}(x,y)$, we can conclude that $d_{\mu,m}(y) \leq d_{\mu,m}(x) + d_{\mathbb{X}}(x,y)$. The choice of x and y is arbitrary, so by symmetry, $d_{\mu,m}(x) \leq d_{\mu,m}(y) + d_{\mathbb{X}}(x,y)$. Therefore, $d_{\mu,m}$ is 1-Lipschitz. □

3.2. Persistence

For persistence diagrams of sublevel sets filtrations of distance to measure functions to be well-defined, we need to prove that they are q-tame.

Proposition 3.5. *Let $\mathbb{X}$ be a triangulable metric space, let μ be a probability measure on $\mathbb{X}$, and let $m \in]0,1]$ be a mass parameter. Then, the sublevel sets filtration of $d_{\mu,m}$ is q-tame.*

Proof. According to Proposition 3.4 $d_{\mu,m}$ is 1-Lipschitz and thus continuous. Also, $d_{\mu,m}$ is nonnegative by definition. Moreover, $d_{\mu,m}$ is proper, i.e., the preimage of any compact set is compact. As the function is nonnegative and continuous, it suffices to show that any sublevel set $d_{\mu,m}^{-1}([0,\alpha])$ is compact.

Suppose for contradiction that for a fixed $\alpha > 0$, $d_{\mu,m}^{-1}([0,\alpha])$ is not compact. Then there exists a sequence $(x_i)_{i>0}$ of points of $d_{\mu,m}^{-1}([0,\alpha])$ such that $d_{\mathbb{X}}(x_0, x_n) \to \infty$ when $n \to \infty$. Hence we can extract a sub-sequence $(x_{\phi(i)})_{i>0}$ such that for any i and j, $\bar{B}(x_{\phi(i)}, \sqrt{2}\alpha) \cap \bar{B}(x_{\phi(j)}, \sqrt{2}\alpha) = \emptyset$. Let us remark that $\mu(\bar{B}(x_{\phi(i)}, \sqrt{2}\alpha)) \geq \frac{m}{2}$. So,

$$d_{\mu,m}(x_{\phi(i)})^2 = \frac{1}{m}\int_0^m \delta_{\mu,l}(x_{\phi(i)})^2 dl \leq \alpha^2.$$

The function $\delta_{\mu,l}(x_{\phi(i)})$ is nonnegative and increasing with l and therefore $\frac{m}{2}\delta_{\mu,\frac{m}{2}}(x_{\phi(i)})^2 \leq m\alpha^2$. Using the definition of $\delta_{\mu,m}$, this implies that $\mu(\bar{B}(x_{\phi(i)}, \sqrt{2}\alpha)) \geq \frac{m}{2}$. Measures are countably additive, so

$$\mu(\mathbb{X}) \geq \sum_{i>0} \mu(\bar{B}(x_{\phi(i)}, \sqrt{2}\alpha)) \geq \sum_{i>0} \frac{m}{2} = \infty.$$

However, μ is a probability measure and therefore $\mu(\mathbb{X}) = 1$. This contradiction implies that $d_{\mu,m}^{-1}([0,\alpha])$ is compact.

As $\mathbb{X}$ is triangulable, there exists a homeomorphism h from $\mathbb{X}$ to a locally finite simplicial complex C. Then for any $\alpha > 0$, we can restrict the simplicial complex C to a finite simplicial complex C_α that contains $h(d_{\mu,m}^{-1}([0,\alpha]))$ as $d_{\mu,m}^{-1}([0,\alpha])$ is compact. The function $d_{\mu,m} \circ h^{-1}|_{C_\alpha}$ is continuous on C_α. Thus its sublevel sets filtration is q-tame by Theorem 2.22 of [6].

The construction extends to any α and therefore the sublevel sets filtration of $d_{\mu,m} \circ h^{-1}$ is q-tame. Furthermore, homology is preserved by homeomorphisms and thus we can say that the sublevel sets filtration of $d_{\mu,m}$ is q-tame. □

Theorem 3.1 is now obtained by combining Theorem 2.4 and Proposition 3.5.

Proof of Theorem 3.1. Theorem 3.3 guarantees that:

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

The sublevel sets filtrations are therefore interleaved since for all $\alpha \in \mathbb{R}$,

$$d_{\mu,m}^{-1}(]-\infty, \alpha]) \subseteq d_{\nu,m}^{-1}(]-\infty, \alpha + \frac{1}{\sqrt{m}} W_2(\mu, \nu)]) \subseteq d_{\mu,m}^{-1}(]-\infty, \alpha + \frac{2}{\sqrt{m}} W_2(\mu, \nu)]).$$

Therefore, applying Theorem 2.4 gives

$$d_B(Dgm(d_{\mu,m}), Dgm(d_{\nu,m})) \leq \frac{1}{\sqrt{m}} W_2(\mu, \nu). \qquad \square$$

4. Approximating the distance to a measure

Computing the persistence diagram of the sublevel sets filtration of $d_{\mu,m}$ requires knowing the sublevel sets. They are not generally easy to compute. We propose an approximation paradigm for $d_{\mu,m}$ that replaces the sublevel sets by a union of balls. The approach works in any metric space and yields equivalent guarantees as the witnessed k-distance approach used in [15] for Euclidean spaces.

4.1. Power distances

Definition 4.1. Given a metric space $\mathbb{X}$, a set P and a function $w : P \to \mathbb{R}$, we define the *power distance* f associated with (P, w) as

$$f(x) = \sqrt{\min_{p \in P} d_{\mathbb{X}}(p, x)^2 + w_p^2}, \tag{2}$$

where w_p is the value of w at the point p.

The function w can be defined on a superset of P. Moreover, the sublevel set $f^{-1}(]-\infty, \alpha])$ is the union of the closed balls centered on the points p of P with radius $r_p(\alpha) = \sqrt{\alpha^2 - w_p^2}$. By convention, we assume that the ball is empty when the radius is imaginary.

Stability

Power distances are stable under small perturbations of the points.

The following lemma states a result about inclusions between balls. It allows another stability result on power distances (Proposition 4.3) and will be useful for studying the stability of the weighted Rips filtration in Section 5.

Lemma 4.2. *Let $p, q \in \mathbb{X}$ be such points such that $d_{\mathbb{X}}(p, q) \leq \epsilon$, and let $w : \mathbb{X} \to \mathbb{R}$ be a t-Lipschitz function. For all $\alpha \geq w_p$,*

$$r_p(\alpha) + \epsilon \leq r_q(\alpha + \sqrt{1 + t^2}\, \epsilon).$$

Proof. First, observe that $r_p(\alpha)$ can be bounded as follows.

$$\begin{aligned} r_p(\alpha)^2 = \alpha^2 - w_p^2 &\leq \alpha^2 - w_p^2 + (t\alpha - \sqrt{1 + t^2}\, w_p)^2 \\ &= (\sqrt{1 + t^2}\, \alpha - t w_p)^2. \end{aligned}$$

Next, we relate r_p and r_q as follows.

$$\begin{aligned} (r_p(\alpha) + \epsilon)^2 &= \alpha^2 - w_p^2 + 2\epsilon\sqrt{\alpha^2 - w_p^2} + \epsilon^2 \\ &\leq \alpha^2 - w_p^2 + 2\epsilon(\sqrt{1 + t^2}\, \alpha - t w_p) + \epsilon^2 \\ &= (\alpha + \sqrt{1 + t^2}\, \epsilon)^2 - (w_p + t\epsilon)^2 \\ &\leq (\alpha + \sqrt{1 + t^2}\, \epsilon)^2 - w_q^2 \\ &= r_q(\alpha + \sqrt{1 + t^2}\, \epsilon)^2. \end{aligned}$$

The requirement that $\alpha \geq w_p$ allows us to take the square root of both sides of the inequality since both will be nonnegative. $\square$

As a consequence, we obtain the following.

Proposition 4.3. *Let $\mathbb{X}$ be a metric space and let $w : \mathbb{X} \to \mathbb{R}$ be a function. Let P and Q be two compact subsets of $\mathbb{X}$. Let f_P and f_Q be the power distances associated with (P, w) and (Q, w). If w is t-Lipschitz, then*

$$\|f_P - f_Q\|_\infty \leq \sqrt{1+t^2}\, d_H(P, Q).$$

Proof. Let x be any point of $\mathbb{X}$. There exists $p \in P$ such that $x \in \bar{B}(p, r_p(f_P(x)))$. There also exists $q \in Q$ such that $d_{\mathbb{X}}(p, q) \leq d_H(P, Q)$. By Lemma 4.2 and the triangle inequality, $x \in \bar{B}(q, r_q(f_P(x) + \sqrt{1+t^2}\, d_H(P, Q)))$. Thus, $f_Q(x) \leq f_P(x) + \sqrt{1+t^2}\, d_H(P, Q)$. P and Q are interchangeable therefore $\|f_Q - f_P\|_\infty \leq \sqrt{1+t^2}\, d_H(P, Q)$. □

Remark that this bound is tight. If we replace t by 0, we have $\|f_Q - f_P\|_\infty = d_H(P, Q)$.

Approximation

To approximate the distance to a probability measure μ, we introduce the following function.

Definition 4.4. Let μ be a probability measure on a metric space $\mathbb{X}$ and let $m \in]0, 1]$ be a mass parameter. Given a subset P of $\mathbb{X}$, we define $d^P_{\mu,m}$ as the power distance associated with $(P, d_{\mu,m})$.

$$d^P_{\mu,m}(x) = \sqrt{\min_{p \in P} d_{\mathbb{X}}(p, x)^2 + d_{\mu,m}(p)^2}$$

That is, the weight of each point is its distance to the empirical measure. If P is close to $\mathrm{Supp}(\mu)$, we obtain an approximation of $d_{\mu,m}$.

Theorem 4.5. *Let μ be a probability measure on a metric space $\mathbb{X}$ and let $m \in]0, 1]$ be a mass parameter. Let P be a subset of $\mathbb{X}$. If $d_H(P, \mathrm{Supp}(\mu)) \leq \epsilon$, then*

$$\frac{1}{\sqrt{2}} d_{\mu,m} \leq d^P_{\mu,m} \leq \sqrt{5}\,(d_{\mu,m} + \epsilon).$$

A multiplicative approximation implies a multiplicative interleaving of the sublevel sets filtrations that becomes an additive interleaving on a logarithmic scale. Theorem 2.4 thus guarantees that the persistence diagrams are close in the bottleneck distance on a logarithmic scale.

Proof. Let x be a point of $\mathbb{X}$. Using the previous notations we get

$$d_{\mu,m}(x)^2 = \frac{1}{m} \int_{\mathbb{X}} d_{\mathbb{X}}(y, x)^2 \mu_{x,m}(y) \mathrm{d}y.$$

Let us now fix a point $p \in \mathrm{Supp}(\mu)$. Since $\mu_{p,m}$ is a submeasure of μ of total mass m,

$$\begin{aligned}
d_{\mu,m}(x)^2 &= \frac{1}{m} \int_{\mathbb{X}} d_{\mathbb{X}}(y, x)^2 \mu_{x,m}(y) \mathrm{d}y \\
&\leq \frac{1}{m} \int_{\mathbb{X}} d_{\mathbb{X}}(y, x)^2 \mu_{p,m}(y) \mathrm{d}y \\
&\leq \frac{1}{m} \int_{\mathbb{X}} ((d_{\mathbb{X}}(y, p) + d_{\mathbb{X}}(p, x))^2) \mu_{p,m}(y) \mathrm{d}y \\
&\leq d_{\mathbb{X}}(p, x)^2 \frac{2}{m} \int_{\mathbb{X}} \mu_{p,m}(y) \mathrm{d}y + \frac{2}{m} \int_{\mathbb{X}} d_{\mathbb{X}}(y, p)^2 \mu_{p,m}(y) \mathrm{d}y \\
&= 2(d_{\mathbb{X}}(p, x)^2 + d_{\mu,m}(p)^2).
\end{aligned}$$

The third inequality follows from the triangle inequality and the relation $(a+b)^2 \leq 2(a^2 + b^2)$.

As the above inequality holds for any point p in P we can conclude that

$$d_{\mu,m}(x) \leq \sqrt{2}\, d^P_{\mu,m}(x).$$

To show the other inequality, let p be a point of P. Then by definition we get:

$$\begin{aligned} d^P_{\mu,m}(x)^2 &\leq d_{\mathbb{X}}(x,p)^2 + d_{\mu,m}(p)^2 \\ &\leq d_{\mathbb{X}}(x,p)^2 + \frac{1}{m}\int_{\mathbb{X}} d_{\mathbb{X}}(p,y)^2 \mu_{x,m}(y)\mathrm{d}y \\ &\leq d_{\mathbb{X}}(x,p)^2 + \frac{1}{m}\int_{\mathbb{X}} (d_{\mathbb{X}}(p,x)+d_{\mathbb{X}}(x,y))^2 \mu_{x,m}(y)\mathrm{d}y \\ &\leq 3\, d_{\mathbb{X}}(x,p)^2 + 2\, d_{\mu,m}(x)^2. \end{aligned}$$

By the definition of the distance to a measure, $d_{\mathbb{X}}(x, \mathrm{Supp}(\mu)) \leq d_{\mu,m}(x)$. Consequently, there exists a point $p \in P$ such that $d_{\mathbb{X}}(x,p) \leq d_{\mu,m}(x) + \epsilon$. Hence,

$$d^P_{\mu,m}(x)^2 \leq 5(d_{\mu,m}(x)+\epsilon)^2. \qquad \square$$

4.2. Measures with finite support

We now assume that data is given as a finite set of points P in a metric space $\mathbb{X}$. We define the following measure to study the point set P.

Definition 4.6. Given a finite point set P in a metric space $\mathbb{X}$, the *empirical measure* μ_P on P is defined as a normalized sum of Dirac measures:

$$\mu_P = \frac{1}{|P|}\sum_{p\in P}\delta_p.$$

Let x be a point of $\mathbb{X}$. We introduce the parameter $k = m|P|$. To simplify the exposition we will assume that k is an integer. See Remark 1 for the generalization.

We reorder the points of P such that $P = (p_1(x), \cdots, p_{|P|}(x))$ and

$$d_{\mathbb{X}}(x, p_1(x)) \leq \cdots \leq d_{\mathbb{X}}(x, p_{|P|}(x)). \tag{3}$$

If two points are at the same distance of x, we order them arbitrarily. We define the set

$$NN^P_k(x) = \{p_1(x), \cdots, p_k(x)\}$$

and call it the set of k^{th} nearest neighbors of x. The set Λ^P_k consists of all k-tuples of points of P.

Lemma 4.7. *Let P be a finite point set in a metric space $\mathbb{X}$ then for any $x \in \mathbb{X}$:*

$$d_{\mu_P,m}(x) = \sqrt{\min_{S\in\Lambda^P_k}\frac{1}{k}\sum_{p\in S} d_{\mathbb{X}}(p,x)^2} = \sqrt{\frac{1}{k}\sum_{p\in NN^P_k(x)} d_{\mathbb{X}}(p,x)^2}.$$

Proof. Since μ_P has finite support, all its submeasures also have finite support.

$$\mathrm{Sub}_m(\mu_P) = \left\{ \sum_{p\in P}\lambda_p\delta_p \mid \forall p\in P,\ 0\leq\lambda_p\leq\frac{1}{|P|} \text{ and } \sum_{p\in P}\lambda_p = m \right\}$$

Let $\nu = \sum_{p\in P}\lambda_p\delta_p$ be an element of $\mathrm{Sub}_m(\mu_P)$.

$$W_2(m\delta_x, \nu)^2 = \sum_{p\in P}\lambda_p d_{\mathbb{X}}(x,p)^2$$

Combined with the relation (3), we get

$$S_x = \sum_{p\in NN^P_k(x)}\delta_p \in \mathrm{argmin}_{\nu\in\mathrm{Sub}_m(\mu_P)} W_2(m\delta_x,\nu).$$

As $S_x \in \Lambda^P_k$, we are done. $\square$

The distance to the empirical measure, $d_{\mu_P,m}$, is thus defined as a lower envelope of quadratic functions. It is generally costly if not impossible to compute its sublevel sets.

However, we can directly use the approximation presented in Section 4.1. Using P in Definition 4.4 and Theorem 4.5, we get the following.

Corollary 4.8. *Let P be a finite point set of a metric space $\mathbb{X}$ and $m \in]0,1]$ be a mass parameter. Then,*

$$\frac{1}{\sqrt{2}} d_{\mu_P,m} \leq d^P_{\mu_P,m} \leq \sqrt{5}\, d_{\mu_P,m}.$$

The multiplicative approximation gives a closeness result between persistence diagrams on a logarithmic scale.

Corollary 4.9. *Let P be a finite point set of a triangulable metric space $\mathbb{X}$ and $m \in]0,1]$ be a mass parameter. Then,*

$$d_B^{\log}(\mathrm{Dgm}(d_{\mu_P,m}), \mathrm{Dgm}(d^P_{\mu_P,m})) \leq \ln(\sqrt{5}).$$

Proof. Corollary 4.8 implies that

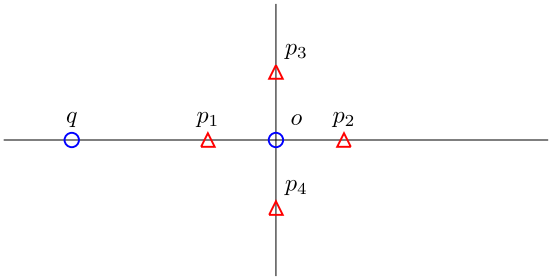
$$\ln(d_{\mu_P,m}) - \ln(\sqrt{2}) \leq \ln(d^P_{\mu_P,m}) \leq \ln(\sqrt{5}) + \ln(d_{\mu_P,m}).$$

The sublevel sets of $\ln(d_{\mu_P,m})$ and $\ln(d^P_{\mu_P,m})$ are thus $\ln(\sqrt{5})$-interleaved and Theorem 2.4 applies. □

Moreover, these bounds cannot be improved.

Proposition 4.10. *The bounds of Corollary 4.8 are tight.*

Proof. We are looking for a worst case scenario where inequalities become equalities for at least one point. We consider the space $\mathbb{R}^d$ with the L_1-norm, denoted $|\cdot|$. For any fixed dimension d, we build the set of $2d$ points whose coordinates have the form $(0,\cdots,0,\pm 1,0,\cdots,0)$. These points are marked by triangles in the following drawing in dimension 2.



We fix $k = 2d$ and we study $d_{\mu_P,m}$ and $d^P_{\mu_P,m}$ at points $q(-3,0\cdots,0)$ and o. First we compute the value of $d_{\mu_P,m}(p_i)$ for any i:

$$d_{\mu_P,m}(p_i)^2 = \frac{1}{2d}\sum_{j=1}^{2d} |p_j - p_i|^2 = 4\frac{2d-1}{2d} = 4 - \frac{2}{d}$$

Now we compute the value of $d_{\mu_P,m}$ at q and o:

$$d_{\mu_P,m}(o)^2 = \frac{1}{2d}\sum_{i=1}^{2d} |p_i - o|^2 = 1$$

$$d_{\mu_P,m}(q)^2 = \frac{1}{2d}\sum_{i=1}^{2d} |p_i - q|^2 = \frac{1}{2d}(|p_1 - q|^2 + (2d-1)16) = 16 - \frac{6}{d}$$

All the points p_i have the same value for $d_{\mu_P,m}$. It is easy to compute $d^P_{\mu_P,m}$ at q and o

$$d^P_{\mu_P,m}(o)^2 = d_{\mu_P,m}(p_1)^2 + |p_1 - o|^2 = 5 - \frac{2}{d}$$

$$d^P_{\mu_P,m}(q)^2 = d_{\mu_P,m}(p_1)^2 + |p_1 - q|^2 = 8 - \frac{2}{d}$$

When d increases, the ratio $\frac{d^P_{\mu_P,m}(o)}{d_{\mu_P,m}(o)}$ tends to $\sqrt{5}$, while $\frac{d^P_{\mu_P,m}(q)}{d_{\mu_P,m}(q)}$ tends to $\frac{1}{\sqrt{2}}$. Thus, the bounds of Corollary 4.8 are reached at the limit for the same data set, although at two different points. □

Remark 1. If k is not an integer, it suffices to do the same construction with a careful weighting of the point $p_{\lceil k \rceil}$. The results stay exactly the same after replacing k by $\lceil k \rceil$.

4.3. Euclidean case

We consider the standard Euclidean space $\mathbb{R}^d$ with the L_2-norm. Considering the finite point set P and its empirical measure in $\mathbb{R}$, we are able to express the distance to the empirical measure $d_{\mu_P,m}$ as a power distance. This restricted settings allows us to improve the bounds of Corollary 4.8 as follows.

Theorem 4.11. *Let P be a finite point set in $\mathbb{R}^d$ and let $m \in]0, 1]$ be a mass parameter. Then the following relation is tight.*

$$\frac{1}{\sqrt{2}} d_{\mu_P,m} \leq d^P_{\mu_P,m} \leq \sqrt{3}\, d_{\mu_P,m}.$$

Moreover, it implies a relation between persistence diagrams:

$$d_B^{\log}(\mathrm{Dgm}(d_{\mu_P,m}), \mathrm{Dgm}(d^P_{\mu_P,m})) \leq \ln(\sqrt{3}).$$

We first present a way to express the distance to a measure as a power distance to the set of all barycenters of k-tuples of P. Then we prove Theorem 4.11 before comparing it with the previous approximation, called the witnessed k-distance proposed in [15]. We improve the bounds on the witnessed k-distance and show that the quality of the approximation is the same for both functions.

4.3.1. Power distance expression of $d_{\mu_P,m}$

For a fixed integer k, the barycenter associated with a point x is the barycenter of its k-nearest neighbors. It is also the center of the cell of the k^{th}-order Voronoi diagram that contains x.

Definition 4.12. For a point set P in $\mathbb{R}^d$ and an integer $k \leq |P|$, the *barycenter associated with x* is

$$\mathrm{bar}(x) = \frac{1}{k} \sum_{p \in NN^P_k(x)} p.$$

Any subset of k elements from P is uniquely associated with a barycenter. We identify the two objects and define a cell energy that describes how clustered the points are.

Definition 4.13. Let P be a point set of $\mathbb{R}^d$ and let $k \leq |P|$. Given $S \in \Lambda^P_k$, we fix $q = \frac{1}{k} \sum_{p \in S} p$ and define the cell energy as

$$E^C(q) = \frac{1}{k} \sum_{p \in S} \|p - q\|^2.$$

Notice that the set S is not necessarily the set $NN^P_k(q)$ and that $E^C(q) \geq d_{\mu_P,m}(q)^2$. We can now write $d_{\mu_P,m}$ in the following form.

Lemma 4.14. *Let P be a finite point set of $\mathbb{R}^d$ let $m \in]0, 1]$ be a mass parameter. For any $x \in \mathbb{R}^d$,*

$$d_{\mu_P,m}(x) = \sqrt{\min_{y \in \mathbb{R}^d} E^C(\mathrm{bar}(y)) + \|\mathrm{bar}(y) - x\|^2} = \sqrt{E^C(\mathrm{bar}(x)) + \|\mathrm{bar}(x) - x\|^2}.$$

Proof. Fix $S \in \Lambda^P_k$ and write $q = \frac{1}{k} \sum_{p \in S} p$. We adapt Lemma 4.7 to the Euclidean setting to get

$$\frac{1}{k} \sum_{p \in S} \|p - x\|^2 = E^C(q) + \|q - x\|^2.$$

This requires the inner product as follows.

$$A = \frac{1}{k}\sum_{p\in S}\|p-x\|^2 = \frac{1}{k}\sum_{p\in S}\Big(\|p-q\|^2 + \|q-x\|^2 + 2\langle p-q|q-x\rangle\Big)$$
$$= E^C(q) + \|q-x\|^2 + 2\langle q-q|q-x\rangle.$$

Lemma 4.7 guarantees that

$$d_{\mu_P,m}(x) = \sqrt{\min_{S\in\Lambda_k^P}\frac{1}{k}\sum_{p\in S}\|p-x\|^2} = \sqrt{\frac{1}{k}\sum_{p\in NN_k^P(x)}\|p-x\|^2},$$

and thus,

$$d_{\mu_P,m}(x) = \sqrt{\min_{S\in\Lambda_k^P} E^C(q) + \|q-x\|} = \sqrt{E^C(\mathrm{bar}(x)) + \|x-\mathrm{bar}(x)\|^2}. \qquad \square$$

In Euclidean space, it is possible to compute the sublevel sets of $d_{\mu_P,m}$ exactly. The function is a power distance and its sublevel sets are unions of balls. However, the complexity problem pointed out in section 4.2 is still valid. The number of balls required to describe a sublevel set is $\Omega(k^{\lceil\frac{d+1}{2}\rceil}n^{\lfloor\frac{d+1}{2}\rfloor})$ [9].

4.3.2. Proof of Theorem 4.11

Proof. The first inequality is exactly the same as the one from Theorem 4.5. For the second inequality, let x be a point in $\mathbb{R}^d$, and let p be a point of P. Thus,

$$d^P_{\mu_P,m}(x)^2 \le d_{\mu_P,m}(p)^2 + \|p-x\|^2.$$

Using Lemma 4.14, we get,

$$d^P_{\mu_P,m}(x)^2 \le E^C(\mathrm{bar}(x)) + \|p-\mathrm{bar}(x)\|^2 + \|p-x\|^2,$$

and with the inner product, this becomes

$$d^P_{\mu_P,m}(x)^2 \le E^C(\mathrm{bar}(x)) + \|x-\mathrm{bar}(x)\|^2 + 2\|p-x\|^2 + 2 < x-\mathrm{bar}(x)|p-x >$$
$$= d_{\mu_P,m}(x)^2 + 2\|p-x\|^2 + 2 < x-\mathrm{bar}(x)|p-x > .$$

Note that

$$2 < \mathrm{bar}(x)-x|x-p >= \|\mathrm{bar}(x)-p\|^2 - \|\mathrm{bar}(x)-x\|^2 - \|x-p\|^2.$$

Then we can write the following relation.

$$d^P_{\mu_P,m}(x)^2 \le d_{\mu_P,m}(x)^2 + \|p-x\|^2 + \|\mathrm{bar}(x)-p\|^2 - \|x-\mathrm{bar}(x)\|^2.$$

This relation holds for any point of P. In particular it holds for any of the k nearest neighbors of x. If we take the average over the k nearest neighbors of x and eliminate the negative term $-\|x-\mathrm{bar}(x)\|^2$, we obtain

$$d^P_{\mu_P,m}(x)^2 \le d_{\mu_P,m}(x)^2 + \frac{1}{k}\sum_{p\in NN_k^P(x)}\|p-x\|^2 + \frac{1}{k}\sum_{p\in NN_k^P(x)}\|\mathrm{bar}(x)-p\|^2.$$

Using the definitions of the cell energy and of the distance to the measure, we can write:

$$d^P_{\mu_P,m}(x)^2 \le d_{\mu_P,m}(x)^2 + d_{\mu_P,m}(x)^2 + E^C(\mathrm{bar}(x))$$

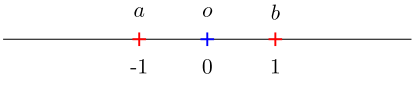
where $E^C(\mathrm{bar}(x)) \le d_{\mu_P,m}(x)^2$. We conclude that

$$d^P_{\mu_P,m}(x) \le \sqrt{3}\, d_{\mu_P,m}(x).$$

The relation between persistence diagrams is exactly as in the proof of Corollary 4.9.

Tightness

The tight example is the point set P of two points a and b on the real line with coordinates 1 and -1.



Fix the mass parameter m equal to 1 so that $k = 2$. It follows that

$$d_{\mu_P,m}(a) = d_{\mu_P,m}(b) = \sqrt{\frac{1}{2}\|b-a\|^2} = \sqrt{2},$$

and

$$d_{\mu_P,m}(o) = \sqrt{\frac{1}{2}\|o-b\|^2 + \|o-a\|^2} = 1.$$

We now compute the last interesting value:

$$d^P_{\mu_P,m}(o)^2 = d_{\mu_P,m}(a)^2 + \|a-o\|^2 = 3.$$

We can thus conclude that $d^P_{\mu_P,m}(o) = \sqrt{3}\, d_{\mu_P,m}(o)$. □

4.3.3. Comparison with witnessed k-distance

Another way of approximating $d_{\mu_P,m}$ was proposed in [15]. Taking advantage of the power distance expression of $d_{\mu_P,m}$, it reduced the set of barycenters to consider. Selecting only the barycenter which are associated with the k nearest neighbors of a point of P gives a set of size at most $|P|$.

Definition 4.15. Let P be a finite point set of $\mathbb{R}^d$ and let $m \in]0,1]$ be a mass parameter. The *witnessed k-distance* is defined as

$$d^W_{\mu_P,m}(x) = \sqrt{\min_{p\in P} E^C(\mathrm{bar}(p)) + \|\mathrm{bar}(p) - x\|^2}.$$

A bound on the quality of the approximation was given in Lemma 3.3 of [15]. We improve this bound and prove it to be at least as good as our approximation. We are not able to prove the tightness of this bound. However, we can give a lower bound on the precision. Using $d^P_{\mu_P,m}$ will not improve the results compared to the witnessed k-distance but will not downgrade the quality either. Moreover it can be used in a more general setting as we do not need the existence of the barycenters.

Theorem 4.16. *Let P be a finite point set of $\mathbb{R}^d$ and let $m \in]0,1]$ be a mass parameter. Then,*

$$d_{\mu_P,m} \le d^W_{\mu_P,m} \le \sqrt{6}\, d_{\mu_P,m}.$$

The previous version of this theorem used a 3 instead of the $\sqrt{6}$.

Proof. The first inequality is obtained by noticing that $d^W_{\mu_P,m}$ is a minimum over a smaller set than $d_{\mu_P,m}$. We thus get $d_{\mu_P,m} \le d^W_{\mu_P,m}$.

Let x be a point in $\mathbb{R}^d$. Thus for any $p \in P$,

$$\begin{aligned} d^W_{\mu,m}(x)^2 &\le E^C(\mathrm{bar}(p)) + \|\mathrm{bar}(p) - x\|^2 \\ &\le E^C(\mathrm{bar}(p)) + \|\mathrm{bar}(p) - p\|^2 + \|p-x\|^2 + 2\langle \mathrm{bar}(p) - p | p - x\rangle \\ &\le d_{\mu,m}(p)^2 + 2\|p-x\|^2 + \|\mathrm{bar}(p) - p\|^2 \\ &\le 2(d_{\mu,m}(p)^2 + \|p-x\|^2). \end{aligned}$$

Thus, $d^W_{\mu,m}(x) \le \sqrt{2} d^P_{\mu,m}(x)$ and using Theorem 4.11 we can conclude that:

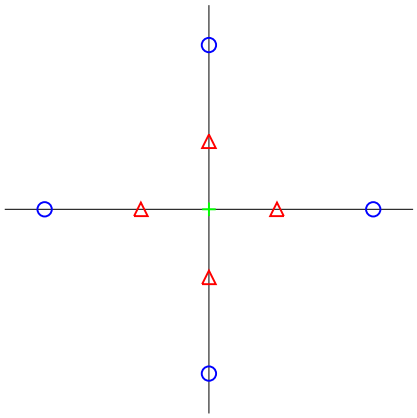
$$d^W_{\mu,m}(x) \le \sqrt{2}\, d^P_{\mu,m}(x) \le \sqrt{6}\, d_{\mu,m}(x). \qquad \square$$

Tightness The tightness of the lower bound is obvious as it suffices to take $k = 1$ to get an equality between $d_{\mu_P,m}$ and $d^W_{\mu_P,m}$.

However, we do not know if the upper bound is tight. The bound $\sqrt{6}$ can not be improved more than to $1+\sqrt{2}$, whose value is greater than $\sqrt{5.82}$.

Let us introduce the following example in $\mathbb{R}^d$. We fix $k = 2d$ and $0 < \epsilon < \sqrt{2}$. The point cloud P consists of $4d^2$ points located at the coordinates $(0, \cdots, 0, \alpha, 0, \cdots, 0)$ with multiplicity 1 when $\alpha = 1$ or $\alpha = -1$ and multiplicity $2d-1$ when $\alpha = 1 + \sqrt{2} - \epsilon$ or $\alpha = \epsilon - 1 - \sqrt{2}$.

The following figure is its representation in dimension 2 where the triangles have multiplicity 1 and the circles have multiplicity 3.



The points are placed such that the k nearest neighbors of any triangle are itself and the $k-1$ points located at the nearest circle. These k nearest neighbors are also the ones from the circles.

Let us now take a look to the value of the functions at the origin o. Each of the k nearest neighbors of o are at distance exactly 1 from o. This allows us to conclude that:

$$d_{\mu_P,m}(o) = 1.$$

The construction induced that the structure is perfectly symmetric and the set of barycenters W we consider in the witnessed k-distance contains exactly $2d$ points. These points are located at the coordinates $(0,\cdots,0,\alpha,0,\cdots,0)$ where $\alpha = 1 + \frac{2d-1}{2d}(\sqrt{2}-\epsilon)$ or the opposite.

Let b be a member of W. Thus we can compute its cell energy:

$$\begin{aligned} E^C(b) &= \frac{1}{2d}\left[\left(\frac{2d-1}{2d}(\sqrt{2}-\epsilon)\right)^2 + (2d-1)\left(\frac{1}{2d}(\sqrt{2}-\epsilon)\right)^2\right] \\ &= \frac{2d-1}{(2d)^3}\left[(2d-1)(\sqrt{2}-\epsilon)^2 + (\sqrt{2}-\epsilon)^2\right] \\ &= \frac{2d-1}{(2d)^2}(\sqrt{2}-\epsilon)^2. \end{aligned}$$

All of the points of W are located at the same distance to o. Thus, the witnessed k-distance at the point o is

$$\begin{aligned} d^W_{\mu_P,m}(o)^2 &= E^C(b) + \left(1 + \frac{2d-1}{2d}(\sqrt{2}-\epsilon)\right)^2 \\ &= \frac{2d-1}{(2d)^2}(\sqrt{2}-\epsilon)^2 + 1 + \frac{2d-1}{d}(\sqrt{2}-\epsilon) + \frac{(2d-1)^2}{(2d)^2}(\sqrt{2}-\epsilon)^2 \\ &= \frac{1}{2d} + \frac{2d-1}{2d}\left(1 + 2(\sqrt{2}-\epsilon) + (\sqrt{2}-\epsilon)^2\right) \\ &= \frac{1}{2d} + \frac{2d-1}{2d}(1+\sqrt{2}-\epsilon)^2. \end{aligned}$$

Since we can take ϵ as small as we want and make the dimension grow, this relation assures us that we cannot find a better constant than $1+\sqrt{2}$ in Theorem 4.16.

5. The weighted Rips filtration

Given a weighted set (P, w) and the associated power distance f (as in (2)), one can introduce a generalization of the Rips filtration that is adapted to the weighted setting as has been done in [15]. This construction allows us to approximate the persistence diagram of $d_{\mu,m}$ in some cases. Moreover, we show that it is stable with respect to perturbation of the underlying sample (Theorem 5.6) and that it gives a guaranteed approximation to the persistence diagram of the distance to an empirical measure (Theorem 5.7). Furthermore, it has an interest of its own as it is stable for close weighted sets and can therefore be used as a shape signature.

Let us consider the sublevel set $f^{-1}(]-\infty, \alpha])$. It is the union of the balls centered on the points p of P with radius $r_p(\alpha) = \sqrt{\alpha^2 - w_p^2}$. By convention, we consider that the ball is empty when the radius is imaginary. We can define the nerve of this union:

Definition 5.1. Let (P, w) be a weighted set in a metric space $\mathbb{X}$, then the *weighted Čech complex* $C_\alpha(P, w)$ for parameter α is defined as the union of simplices σ such that $\bigcap_{p\in\sigma} B(p, r_p(\alpha)) \neq 0$.

However, the Čech complex can be difficult to compute due the problem of testing if a collection of metric balls has a common intersection. Instead, we define a weighted version of the Rips complex, which only requires distance computations.

Definition 5.2. For a weighted set (P, w) in a metric space $\mathbb{X}$, the *weighted Rips complex* $R_\alpha(P, w)$ for a parameter α is the maximal simplicial complex whose 1-skeleton has an edge for each pair (p, q) such that $d_{\mathbb{X}}(p, q) < r_p(\alpha) + r_q(\alpha)$. The *weighted Rips filtration* is the sequence $\{R_\alpha(P, w)\}$ for all $\alpha \geq 0$.

Remark that if all weights are equal to 0, we are in the classical case of balls with equal radii. We use the weighted Rips filtration to approximate the weighted Čech filtration thanks to the following interleaving. For simplicity, the notation (P, w) indicating the point set P with weights w is omitted in the notation.

Proposition 5.3. *If (P, w) is a weighted set on a metric space $\mathbb{X}$, then for all $\alpha \in \mathbb{R}$:*

$$C_\alpha \subseteq R_\alpha \subseteq C_{2\alpha}.$$

Proof. Let α be a real number. The first inclusion is obtained by the definition of the weighted Rips complex that gives $C_\alpha \subseteq R_\alpha$.

For the other inclusion, let σ be a simplex of R_α. We fix p_0 to be the point of σ with the greatest weight. This implies especially that for any $p \in P$, $r_p(\alpha) \geq r_{p_0}(\alpha)$.

Since $\sigma \in R_\alpha$, we get that, for all p and q in P, we have $d_{\mathbb{X}}(p, q) \leq r_p(\alpha) + r_q(\alpha)$ with both radius real. To prove that $\sigma \in C_{2\alpha}$ we need to prove that:

$$\bigcap_{p\in\sigma} B(p, r_p(2\alpha)) \neq 0.$$

It will suffice to prove that p_0 belongs to this intersection. For each $p \in \sigma$:

$$d_{\mathbb{X}}(p, p_0) \leq r_p(\alpha) + r_{p_0}(\alpha) \leq 2\, r_p(\alpha) = \sqrt{(2\alpha)^2 - 4w_p^2} \leq r_p(2\alpha). \qquad \square$$

Stability

The persistence diagram of a weighted Rips filtration $\{R_\alpha(P, w)\}$ is stable under small perturbations of the set P. It can thus be used in applications like signatures in the spirit of [3].

Speaking of the persistence diagram of a weighted Rips filtration requires that the filtration is q-tame. This is always the case when the set P is compact as shown in the following proposition.

Proposition 5.4. *Let P be a subset of a metric space $\mathbb{X}$ and let $w : \mathbb{X} \to \mathbb{R}$ be a function. If P is compact, then $\{R_\alpha(P, w)\}_{\alpha\in\mathbb{R}}$ is q-tame.*

This will be deduced from the following technical lemma.

Lemma 5.5. *Let P, Q be two subsets of a metric space $\mathbb{X}$ and let $w : \mathbb{X} \to \mathbb{R}$ be a t-Lipschitz function. Then $\mathrm{H}_*(\{R_\alpha(P, w)\})$ and $\mathrm{H}_*(\{R_\alpha(Q, w)\})$ are ϵ-interleaved for $\epsilon = (1 + t)d_H(P, Q)$.*

Proof. We need to show that the there exists ϵ-homomorphisms π_{P*} and π_{Q*} such that $\pi_{P*}\pi_{Q*} = 1^{2\epsilon}_{\mathrm{H}_*(R_\alpha(P,w))}$ and $\pi_{Q*}\pi_{P*} = 1^{2\epsilon}_{\mathrm{H}_*(R_\alpha(Q,w))}$.

To do so, we need three steps. First, we build simplicial maps $R_\alpha(P, w) \to R_{\alpha+\epsilon}(Q, w)$ and $R_\alpha(Q, w) \to R_{\alpha+\epsilon}(P, w)$ for every α. Then, we show that these simplicial maps induce ϵ-homomorphisms. Finally, we show that the simplicial maps are contiguous and thus the two persistence modules are ϵ-interleaved.

The simplicial maps $i^\beta_\alpha : R_\alpha(P, w) \to R_\beta(P, w)$ and $j^\beta_\alpha : R_\alpha(Q, w) \to R_\alpha(Q, w)$ for $\alpha < \beta$ are induced by the canonical inclusion. We consider two maps $\pi_P : Q \to P$ and $\pi_Q : P \to Q$ such that $d_{\mathbb{X}}(p, \pi_Q(p)) \leq d_H(P, Q)$ and $d_{\mathbb{X}}(q, \pi_P(q)) \leq d_H(P, Q)$ for any $p \in P$ and $q \in Q$. By definition of the Hausdorff distances, such maps always exist. Let us show that these maps induce simplicial maps.

Let us consider the function π_P and let us fix $\alpha > 0$. Let (q', q'') be an edge of $R_\alpha(Q, w)$. It means that $B(q', r_{q'}(\alpha)) \cap B(q'', r_{q''}(\alpha)) \neq \emptyset$. Lemma 4.2 implies that $B(q, r_q(\alpha)) \subset B(\pi_P(q), r_{\pi_P(q)}(\alpha + (1+t)d_H(P, Q)))$ for any $q \in Q$. Thus, $(\pi_P(q'), \pi_(q''))$ is an edge of $R_{\alpha+\epsilon}(P, w)$ because:

$$B(\pi_P(q'), r_{q'}(\alpha+\epsilon)) \cap B(\pi_P(q''), r_{q''}(\alpha+\epsilon)) \supset B(q', r_{q'}(\alpha)) \cap B(q'', r_{q''}(\alpha)) \neq \emptyset.$$

As $R_\alpha(P, w)$ is a clique complex for any α, this is sufficient to prove that π_P induce a family of simplicial maps $\{{\pi_P}_\alpha^{\alpha+\epsilon}\}$. The roles of P and Q are symmetric. Therefore, the result holds for π_Q as well.

Furthermore π_P induces an ϵ-homomorphism π_{P*} at the homology level. For any $\alpha < \beta$, $i_{\alpha+\epsilon}^{\beta+\epsilon} \circ {\pi_P}_\alpha^{\alpha+\epsilon} = {\pi_P}_\beta^{\beta+\epsilon} \circ j_\alpha^\beta$ because the maps $i_{\alpha+\epsilon}^{\beta+\epsilon}$ and j_α^ϵ are induced by the canonical inclusion while the two others simplicial maps are induced by the same map $\pi_P : Q \to P$. Hence the two compositions are the same map and thus guarantees that π_{P*} is an ϵ-homomorphism. Again, this results can be applied to π_Q to get an ϵ-homomorphism π_{Q*}.

To prove that $\pi_{P*}\pi_{Q*} = 1^{2\epsilon}_{\mathrm{H}_*(R_\alpha(P,w))}$, we prove that ${\pi_P}_\alpha^{\alpha+\epsilon} \circ {\pi_Q}_{\alpha-\epsilon}^{\alpha}$ and $i_{\alpha-\epsilon}^{\alpha+\epsilon}$ are contiguous for any α.

Let us fix α and let (p, p') be an edge of $R_{\alpha-\epsilon}(P, w)$. By definition, $B(p, r_p(\alpha-\epsilon)) \cap B(p', r_{p'}(\alpha-\epsilon)) \neq \emptyset$. Moreover, using Lemma 4.2 we get:

$$B(p, r_p(\alpha-\epsilon)) \subset B(\pi_Q(p), r_{\pi_Q(p)}(\alpha)) \subset B(\pi_P \circ \pi_Q(p), r_{\pi_P \circ \pi_Q(p)}(\alpha+\epsilon)).$$

The same holds for p' and thus:

$$B(p, r_p(\alpha+\epsilon)) \cap B(\pi_P \circ \pi_Q(p), r_{\pi_P \circ \pi_Q(p)}(\alpha+\epsilon)) \cap B(p', r_{p'}(\alpha+\epsilon)) \cap B(\pi_P \circ \pi_Q(p'), r_{\pi_P \circ \pi_Q(p')}(\alpha+\epsilon)) \neq \emptyset.$$

Thus the tetrahedron $\{i_{\alpha-\epsilon}^{\alpha+\epsilon}(p), i_{\alpha-\epsilon}^{\alpha+\epsilon}(p'), {\pi_P}_\alpha^{\alpha+\epsilon} \circ {\pi_Q}_{\alpha-\epsilon}^{\alpha}(p), {\pi_P}_\alpha^{\alpha+\epsilon} \circ {\pi_Q}_{\alpha-\epsilon}^{\alpha}(p')\}$ is in $C_{\alpha+\epsilon}(P, w) \subset R_{\alpha+\epsilon}(P, w)$. Lemma 2.7 guarantees that ${\pi_P}_\alpha^{\alpha+\epsilon} \circ {\pi_Q}_{\alpha-\epsilon}^{\alpha}$ and $i_{\alpha-\epsilon}^{\alpha+\epsilon}$ are contiguous.

From before, $\{{\pi_P}_\alpha^{\alpha+\epsilon} \circ {\pi_Q}_{\alpha-\epsilon}^{\alpha}\}$ induces the 2ϵ-homomorphism $\pi_{P*}\pi_{Q*}$. By definition, $\{i_{\alpha-\epsilon}^{\alpha+\epsilon}\}$ induces $1^{2\epsilon}_{\mathrm{H}_*(R_\alpha(P,w))}$. By contiguity of the simplicial maps, we have equality of the 2ϵ-homomorphisms and therefore $\pi_{P*}\pi_{Q*} = 1^{2\epsilon}_{\mathrm{H}_*(R_\alpha(P,w))}$.

By symmetry of the roles of P and Q, $\{R_\alpha(P, w)\}$ and $\{R_\alpha(Q, w)\}$ are ϵ-interleaved. □

Proof of Proposition 5.4. We will show that, for any $\epsilon > 0$, one can build a finite persistence module which is ϵ-interleaved with the persistence module of $\{R_\alpha(P, w)\}$. A finite persistence module is a fortiori locally finite and Theorem 4.19 of [6] induces the q-tameness of $\{R_\alpha(P, w)\}$.

Let us fix $\epsilon > 0$. P is compact. As a consequence, there exists a finite point set Q of P such that $d_H(P, Q) \leq \frac{\epsilon}{1+t}$. The persistence module of $\{R_\alpha(Q, w)\}$ is finite and therefore locally finite. Moreover, using Lemma 5.5, $\{R_\alpha(Q, w)\}$ and $\{R_\alpha(P, w)\}$ are ϵ-interleaved. Hence Theorem 4.19 of [6] induces the q-tameness of $\{R_\alpha(P, w)\}$. □

Notice that the simplicial maps π_P and π_Q are not necessarily uniquely defined. However, if π_P and π'_P are two maps verifying the construction property, then the induced simplicial maps are contiguous and therefore the induced homomorphisms are identical.

The persistence diagrams of weighted Rips filtrations are related by the following:

Theorem 5.6. *Let P and Q be two compact subsets of a metric space $\mathbb{X}$. Let $w : \mathbb{X} \to \mathbb{R}$ be a t-Lipschitz function. Then,*

$$d_B(\mathrm{Dgm}(\{R_\alpha(P, w)\}), \mathrm{Dgm}(\{R_\alpha(Q, w)\})) \leq (1+t)d_H(P, Q).$$

Proof. P and Q are two compact sets and thus the diagrams are well-defined thanks to Proposition 5.4 that guarantees the q-tameness of the filtrations. Lemma 5.5 implies that $\mathrm{H}_*(\{R_\alpha(P, w)\})$ and $\mathrm{H}_*(\{R_\alpha(Q, w)\})$ are $(1+t)d_H(P, Q)$-interleaved. The relation between the persistence diagrams is then obtained by applying Theorem 2.6. □

Remark 2. When P and Q are two compact metric spaces, Theorem 5.6 can be extended using the notion of correspondence as in [7]. Notice that the correspondence has to induce bounded distortion on the weights as well as on the distances.

Approximation

To use the weighted Rips filtration to approximate the persistence diagram of the distance to a measure, we need to restrict the class of spaces considered. If the intersection of any finite number of balls in $\mathbb{X}$ is either contractible or empty, $\mathbb{X}$ is said to have the *good cover property*. Then the Čech complex has the same homology as the union of balls, of which it is the nerve, by the Nerve Theorem [16]. We can also compute the persistence diagram thanks to the Persistent Nerve Lemma [8]. We obtain an approximation of $\mathrm{Dgm}(d_{\mu_P,m})$ using the weighted Rips filtration.

Theorem 5.7. *Let $\mathbb{X}$ be a triangulable metric space with the good cover property and let P be a finite point set of $\mathbb{X}$, then on a logarithmic scale:*

$$d_B^{\log}(\mathrm{Dgm}(d_{\mu_P,m}), \mathrm{Dgm}(\{R_\alpha(P, d_{\mu_P,m})\})) \leq \ln(2\sqrt{5}).$$

Proof. Given that $\mathbb{X}$ is triangulable, we know that the sublevel sets filtration of $d_{\mu_P,m}$ is q-tame by Proposition 3.5. The persistence diagram $\mathrm{Dgm}(d_{\mu_P,m})$ is thus well-defined. Recall that $d_{\mu_P,m}$ is a 1-Lipschitz function (see Proposition 3.4). P is a compact subset of $\mathbb{X}$ and therefore $\mathrm{Dgm}(R_\alpha(P, d_{\mu_P,m}))$ is well-defined according to Proposition 5.4.

We approximate $d_{\mu_P,m}$ with $d^P_{\mu_P,m}$. The result of Theorem 4.5 gives us a $\sqrt{5}$ multiplicative interleaving. For any $\alpha \in \mathbb{R}$,

$$d_{\mu_P,m}(]-\infty,\alpha]) \subset d^P_{\mu_P,m}(]-\infty,\sqrt{2}\alpha]) \subset d_{\mu_P,m}(]-\infty,\sqrt{10}\, d^P_{\mu_P,m}]).$$

So, Theorem 2.4 implies

$$d_B^{\log}(\mathrm{Dgm}(d_{\mu_P,m}), \mathrm{Dgm}(d^P_{\mu_P,m})) \leq \ln(\sqrt{5}).$$

By the Persistent Nerve Lemma, the sublevel sets filtration of $d^P_{\mu_P,m}$ (a union of balls of increasing radii) has the same persistent homology as its nerve filtration. Thus, we can use weighted Rips filtration to approximate the persistence diagram:

$$d_B^{\log}(\mathrm{Dgm}(d^P_{\mu_P,m}), \mathrm{Dgm}(\{R_\alpha(P, d_{\mu_P,m})\}) \leq \ln(2).$$

The triangle inequality for the bottleneck distance gives the desired inequality. □

6. The sparse weighted Rips filtration

The weighted Rips filtration presented in the previous section has the desired approximation guarantees, but like the Rips filtration for unweighted points, it usually grows too large to be computed in full. In [21], it was shown how to construct a filtration $\{S_\alpha\}$ called the *sparse Rips filtration* that gives a provably good approximation to the Rips filtration and has size linear in the number of points for metrics with constant doubling dimension (see Section 6.1 for the construction). Specifically, for a user-defined parameter ε, the log-bottleneck distance between the persistence diagrams of the Sparse Rips filtration and the Rips filtration is at most ε. The goal of this section is to extend that result to weighted Rips filtrations.

The sparse Rips filtration cannot be used directly here, since the power distance does not induce a metric. Indeed, even the case of setting all weights to some large constant yields a persistence diagram that is far from the persistence diagram of the Rips filtration of any metric space. This follows because individual points in a Rips filtration appear at time zero, but this is not the case in the weighted Rips filtration.

Even if one were to construct a metric whose Rips filtration exactly matched that of the weighted Rips filtration, there are simple examples where that metric would necessarily have very high doubling dimension, making previous methods unsuitable. For example, consider a set of points in the unit interval $[0, 1]$, with a constant weight function that assigns a weight of 1 to every point. Although the points lie in a 1-dimensional space, the weighted distance function has doubling dimension $\log n$ because all of the points are within a weighted distance of 2, whereas every pair has weighted distance at least 1. So the doubling constant would be n and the doubling dimension would be $\log n$ despite that the input was 1-dimensional. Thus, any construction that depends on low doubling dimension will blowup when confronted with such weighted examples.

For the rest of this section, we fix a weighted point set P in a metric space $\mathbb{X}$, where the weight function $w : \mathbb{X} \to \mathbb{R}$ is t-Lipschitz, for some constant t. To simplify notation, we let R_α denote the weighted Rips complex $R_\alpha(P, w)$.

The *sparse weighted Rips filtration*, $\{T_\alpha\}$, is defined as

$$T_\alpha = S_\alpha \cap R_\alpha.$$

The (unweighted) sparse Rips filtration $\{S_\alpha\}$ captures the underlying metric space and the weighted Rips filtration $\{R_\alpha\}$ captures the structure of the sublevel sets of the power distance function. Computing $\{T_\alpha\}$ can be done efficiently by first computing $\{S_\alpha\}$ and then reordering the simplices according to the birth time in $\{R_\alpha\}$. This is equivalent to filtering the complex S_∞. Note that the sparsification depends only on the metric, and not on the weights. Thus, the same sparse Rips complex can be used as the underlying complex for different weight functions. We also simplify the construction of $\{S_\alpha\}$ by using a furthest point sampling instead of the more complex structure of net tree.

The technical challenge is to relate the persistence diagram of this new filtration to the persistence diagram of the weighted Rips filtration as in the following theorem.

Theorem 6.1. *Let (P, w), be a finite, weighted subset of a metric space $\mathbb{X}$ with t-Lipschitz weights. Let $\varepsilon < 1$ be a fixed constant used in the construction of the sparse weighted Rips filtration $\{T_\alpha\}$. Then,*

$$d_B^{\log}(\mathrm{Dgm}(\{T_\alpha\}), \mathrm{Dgm}(\{R_\alpha\})) \leq \ln\left(\frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}\right).$$

Since these filtrations are not interleaved, the only hope is to find an interleaving of the persistence modules, which requires finding suitable homomorphisms between the homology groups of the different filtrations. After detailing the construction of the sparse Rips filtration with the furthest point sampling, the rest of this section proves Theorem 6.1.

6.1. Sparse Rips complexes

Let $(p_1, \dots, p_n)$ be a greedy permutation of the points P in a finite metric space $\mathbb{X}$. That is, $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d_{\mathbb{X}}(p, P_{i-1})$, where $P_{i-1} = \{p_1, \dots, p_{i-1}\}$ is the $(i-1)$st prefix. We define the *insertion radius* λ_{p_i} of point p_i to be

$$\lambda_{p_i} = d_{\mathbb{X}}(p_i, P_{i-1}).$$

To avoid excessive superscripts, we write λ_i in place of λ_{p_i} when we know the index of p_i. We adopt the convention that $\lambda_1 = \infty$ and $\lambda_{n+1} = 0$. The greedy permutation has the nice property that each prefix P_i is a λ_i-*net* in the sense that

1. $d_{\mathbb{X}}(p, P_i) \leq \lambda_i$ for all $p \in P$.
2. $d_{\mathbb{X}}(p, q) \geq \lambda_i$ for all $p, q \in P_i$.

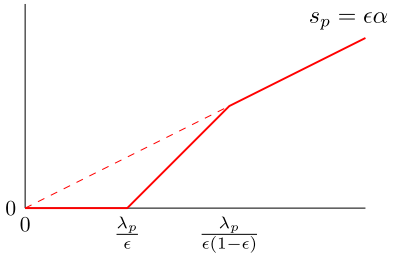
We extend these nets to an arbitrary parameter γ as

$$N_\gamma = \{p \in P \mid \lambda_p > \gamma\}.$$
$$\overline{N}_\gamma = \{p \in P \mid \lambda_p \geq \gamma\}.$$

Note that for all $p \in P$, $d_{\mathbb{X}}(p, N_\gamma) \leq \gamma$ and $d_{\mathbb{X}}(p, \overline{N}_\gamma) < \gamma$.

One way to get a sparse Rips-like filtration is to take a union of Rips complexes on the nets N_γ. However, this can add significant noise to the persistence diagram compared to the Rips filtrations. This noise can be diminished by a careful perturbation of the distance. For a point p, the perturbation varies with the scale and is defined as follows:

$$s_p(\alpha) = \begin{cases} 0 & \text{if } \alpha \leq \frac{\lambda_p}{\varepsilon} \\ \alpha - \frac{\lambda_p}{\varepsilon} & \text{if } \frac{\lambda_p}{\varepsilon} < \alpha < \frac{\lambda_p}{\varepsilon(1-\varepsilon)} \\ \varepsilon\alpha & \text{if } \frac{\lambda_p}{\varepsilon(1-\varepsilon)} \leq \alpha \end{cases}$$



Note that s_p is 1-Lipschitz. The resulting perturbed distance is defined as

$$\mathrm{f}_\alpha(p, q) = d_{\mathbb{X}}(p, q) + s_p(\alpha) + s_q(\alpha).$$

Definition 6.2. Given the nets N_γ and the distance function f_α, we define the *sparse Rips complex* at scale α as

$$Q_\alpha = \{\sigma \subset \overline{N}_{\varepsilon(1-\varepsilon)\alpha} \mid \forall p, q \in \sigma,\ \mathrm{f}_\alpha(p, q) \leq 2\alpha\}.$$

On its own, the sequence of complexes $\{Q_\alpha\}$ does not form a filtration. However, we can build a natural filtration by defining

Definition 6.3. The *sparse Rips filtration* is defined as:

$$S_\beta = \bigcup_{\alpha \leq \beta} Q_\alpha.$$

6.2. Projection onto nets

To relate sparse Rips complexes with Rips complexes, we build a collection of projections of the points onto the nets.

$$\pi_\alpha(p) = \begin{cases} p & \text{if } p \in N_{\varepsilon(1-\varepsilon)\alpha} \\ \operatorname{argmin}_{q \in N_{\varepsilon\alpha}} d_{\mathbb{X}}(p, q) & \text{otherwise} \end{cases}$$

For any scale α, the projection π_α maps the points of P to the net $N_{\varepsilon(1-\varepsilon)\alpha}$. Note that π_α is a retraction onto $N_{\varepsilon(1-\varepsilon)\alpha}$.

The following are the four main lemmas we will use with respect to the perturbed distance functions and projections. The projections will be used extensively to induce maps between simplicial complexes.

First, we prove that edges do not disappear as the filtration grows.

Lemma 6.4. *If* $\mathrm{f}_\alpha(p,q) \le 2\alpha \le 2\beta$ *then* $\mathrm{f}_\beta(p,q) \le 2\beta$.

Proof. The proof follows from the definitions f_α and f_β, the Lipschitz property of the perturbations s_p and s_q, and the hypothesis as follows.

$$\begin{aligned}\mathrm{f}_\beta(p,q) &= d_{\mathbb{X}}(p,q) + s_p(\beta) + s_q(\beta)\\ &\le d_{\mathbb{X}}(p,q) + s_p(\alpha) + (\beta-\alpha) + s_q(\alpha) + (\beta-\alpha)\\ &= \mathrm{f}_\alpha(p,q) + 2(\beta-\alpha)\\ &\le 2\alpha + 2(\beta-\alpha)\\ &= 2\beta. \qquad \square\end{aligned}$$

Next, we show that the distance between a point and its projection is at most the change in the perturbed distance.

Lemma 6.5. *For all* $q \in P$, $d_{\mathbb{X}}(q, \pi_\alpha(q)) \le s_q(\alpha) - s_{\pi_\alpha(q)}(\alpha)$, *and in particular,* $d_{\mathbb{X}}(q,\pi_\alpha(q)) \le \varepsilon\alpha$.

Proof. Both statements are trivial if $q \in N_{\varepsilon(1-\varepsilon)\alpha}$, because that would imply that $\pi_\alpha(q) = q$. So, we may assume that $\pi_\alpha(q)$ is the nearest point to q in $N_{\varepsilon\alpha}$ It follows that

$$d_{\mathbb{X}}(q,\pi_\alpha(q)) \le \varepsilon\alpha.$$

Moreover, $\lambda_q \le \varepsilon(1-\varepsilon)\alpha$, and thus $s_q(\alpha) = \varepsilon\alpha$. Also, since $\pi_\alpha(q) \in N_{\varepsilon\alpha}$, it must be that $\lambda_{\pi_\alpha(q)} > \varepsilon\alpha$ and so $s_{\pi_\alpha(q)} = 0$. Combining these statements, we get

$$d_{\mathbb{X}}(q,\pi_\alpha) \le \varepsilon\alpha = s_q(\alpha) - s_{\pi_\alpha(q)}(\alpha). \qquad \square$$

Now, we prove that replacing a point with its projection does not increase the perturbed distance.

Lemma 6.6. *For all* $p, q \in P$ *and all* $\alpha \ge 0$, $\mathrm{f}_\alpha(p, \pi_\alpha(q)) \le \mathrm{f}_\alpha(p,q)$. $\square$

Proof. The statement follows from the definition of f_α, the triangle inequality, and Lemma 6.5 as follows.

$$\begin{aligned}\mathrm{f}_\alpha(p,\pi_\alpha(q)) &= d_{\mathbb{X}}(p,\pi_\alpha(q)) + s_p(\alpha) + s_{\pi_\alpha(q)}(\alpha)\\ &\le d_{\mathbb{X}}(p,q) + d_{\mathbb{X}}(q,\pi_\alpha(q)) + s_p(\alpha) + s_{\pi_\alpha(q)}(\alpha)\\ &\le d_{\mathbb{X}}(p,q) + s_p(\alpha) + s_q(\alpha)\\ &= \mathrm{f}_\alpha(p,q). \qquad \square\end{aligned}$$

6.3. Sometimes the projections induce contiguous simplicial maps

In this section, we look at the maps between simplicial complexes that are induced by the projection functions π_α. We are most interested in the case when a pair of projections π_α and π_β induce contiguous simplicial maps between sparse Rips complexes (Lemma 6.9) or weighted Rips complexes (Lemma 6.10). First, we need a couple lemmas that describe the effect of different projections on the endpoints of an edge in sparse or weighted Rips complexes.

Lemma 6.7. *Let* α, β, γ, *and* i *be such that* $\frac{\lambda_{i+1}}{\varepsilon(1-\varepsilon)} \le \alpha \le \beta \le \gamma \le \frac{\lambda_i}{\varepsilon(1-\varepsilon)}$. *If an edge* (p,q) *is in* Q_ρ *for some* $\rho \le \gamma$ *then the edge* $(\pi_\alpha(p), \pi_\beta(q)) \in Q_\gamma$.

Proof. First, it is easy to check that the conditions on α, β, γ, and i imply that $\pi_\alpha(p)$ and $\pi_\beta(q)$ are in $\overline{N}_{\varepsilon(1-\varepsilon)\gamma}$, which is the vertex set of Q_γ. So, it will suffice to prove that $\mathrm{f}_\gamma(\pi_\alpha(p), \pi_\beta(q)) \le 2\gamma$. Next we consider three cases depending on the value of ρ in relation to α and β.

Case 1: If $\alpha, \beta \le \rho$ then $\pi_\alpha(p) = p$ and $\pi_\beta(q) = q$. So, using Lemma 6.4 and the assumption $\rho \le \gamma$, we see that $\mathrm{f}_\gamma(\pi_\alpha(p), \pi_\beta(q)) = \mathrm{f}_\gamma(p,q) \le 2\gamma$.

Case 2: If $\alpha \le \rho < \beta$ then $\pi_\alpha(p) = p$ and Lemma 6.4 implies that $\mathrm{f}_\beta(p,q) \le 2\beta$.

$$\begin{aligned}
\mathrm{f}_\gamma(\pi_\alpha(p),\pi_\beta(q)) &= \mathrm{f}_\gamma(p,\pi_\beta(q)) \\
&\leq \mathrm{f}_\beta(p,\pi_\beta(q)) + 2(\gamma-\beta) \\
&\leq \mathrm{f}_\beta(p,q) + 2(\gamma-\beta) \\
&\leq 2\gamma.
\end{aligned}$$

Case 3: If $\rho < \alpha, \beta$ then Lemma 6.4 implies that $\mathrm{f}_\alpha(p,q) \leq 2\alpha$.

$$\begin{aligned}
\mathrm{f}_\gamma(\pi_\alpha(p),\pi_\beta(q)) &\leq \mathrm{f}_\beta(\pi_\alpha(p),\pi_\beta(q)) + 2(\gamma-\beta) \\
&\leq \mathrm{f}_\beta(\pi_\alpha(p),q) + 2(\gamma-\beta) \\
&\leq \mathrm{f}_\alpha(\pi_\alpha(p),q) + 2(\gamma-\beta) + 2(\beta-\alpha) \\
&\leq \mathrm{f}_\alpha(p,q) + 2(\gamma-\beta) + 2(\beta-\alpha) \\
&\leq 2\gamma. \qquad \square
\end{aligned}$$

Lemma 6.8. *Let (p,q) be an edge of R_δ with $\alpha, \beta \leq \frac{\delta}{1+\varepsilon}$, then $(\pi_\alpha(p),\pi_\beta(q)) \in R_{\kappa\delta}$, where $\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}$.*

Proof. First, note that the projection functions satisfy the following inequalities.

$$\begin{aligned}
d_{\mathbb{X}}(p,\pi_\alpha(p)) &\leq \varepsilon\alpha \leq \frac{\varepsilon\delta}{1-\varepsilon} \\
d_{\mathbb{X}}(q,\pi_\beta(q)) &\leq \varepsilon\beta \leq \frac{\varepsilon\delta}{1-\varepsilon}
\end{aligned}$$

So, by applying the triangle inequality, the definition of an edge in R_δ, and Lemma 4.2, we get the following.

$$\begin{aligned}
d_{\mathbb{X}}(\pi_\alpha(p),\pi_\beta(q)) &\leq d_{\mathbb{X}}(p,q) + \frac{2\varepsilon\delta}{1-\varepsilon} \\
&\leq \left(r_p(\delta) + \frac{\varepsilon\delta}{1-\varepsilon}\right) + \left(r_q(\delta) + \frac{\varepsilon\delta}{1-\varepsilon}\right) \\
&\leq \left(r_p\left(\frac{\delta}{1-\varepsilon}\right) + \frac{\varepsilon\delta}{1-\varepsilon}\right) + \left(r_q\left(\frac{\delta}{1-\varepsilon}\right) + \frac{\varepsilon\delta}{1-\varepsilon}\right) \\
&\leq r_{\pi_\alpha(p)}(\kappa\delta) + r_{\pi_\beta(q)}(\kappa\delta).
\end{aligned}$$

This is precisely the necessary condition to guarantee that $(\pi_\alpha(p),\pi_\alpha(q)) \in R_{\kappa\delta}$ as desired. $\square$

The following two lemmas follow easily from repeated application of the preceding lemmas.

Lemma 6.9. *Two projections π_α and π_β induce contiguous simplicial maps from $Q_\rho \to Q_\beta$ whenever $\rho \leq \beta$ and there exists i so that $\frac{\lambda_{i+1}}{\varepsilon(1-\varepsilon)} \leq \alpha \leq \beta \leq \frac{\lambda_i}{\varepsilon(1-\varepsilon)}$.*

Proof. Let us fix $\rho \leq \beta$ and take (p,q) an edge from Q_ρ. Given that Q_ρ and Q_β are cliques complexes, we can get the result from Lemma 2.7 if we show that the tetrahedron $\{\pi_\alpha(p),\pi_\alpha(q),\pi_\beta(p),\pi_\beta(q)\}$ is in Q_β. We only need to prove that all edges of the tetrahedron belongs to Q_β.

We apply Lemma 6.7, while replacing γ by β and β by α. Thus we obtain $(\pi_\alpha(p),\pi_\alpha(q)) \in Q_\beta$. Let us repeat this operation with $\alpha = \beta = \gamma$ thus we get $(\pi_\beta(p),\pi_\beta(q)) \in Q_\beta$. The last two edges are given by replacing γ by β and choosing correctly the role of p and q. $\square$

Lemma 6.10. *Two projections π_α and π_β induce contiguous simplicial maps from $R_\delta \to R_{\kappa\delta}$, where $\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}$ whenever $\alpha, \beta \leq \frac{\delta}{1-\varepsilon}$.*

Proof. The previous proof can be applied to get the result, while replacing Lemma 6.7 by Lemma 6.8. $\square$

6.4. Sparse filtrations and power distance functions

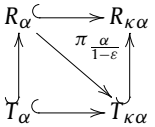
We define a sparse filtration that gives a good approximation to the weighted Rips filtration $\{R_\alpha\}$ in terms of persistent homology. It is simply the intersection of the weighted Rips complex and the union of sparse Rips complexes at different scales.

$$T_\alpha = R_\alpha \cap S_\alpha.$$

Our main goal is to show that the filtration $\{T_\alpha\}$ has a persistence diagram that is similar to that of $\{R_\alpha\}$. To do this we will demonstrate a multiplicative interleaving between these filtrations, where the interleaving constant is

$$\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}.$$

Specifically, we show that for all $\alpha \geq 0$, the following diagram commutes at the homology level.



We first need to check that the projection $\pi_{\frac{\alpha}{1-\varepsilon}}$ indeed induces a simplicial map from R_δ to $T_{\kappa\delta}$.

Lemma 6.11. *For all $\alpha > 0$, the projection $\pi_{\frac{\alpha}{1-\varepsilon}}$ induces a simplicial map from $R_\alpha \to T_{\kappa\alpha}$, where $\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}$.*

Proof. We show that for each edge $(p,q) \in R_\alpha$, there is a corresponding edge $(\pi_{\frac{\alpha}{1-\varepsilon}}(p), \pi_{\frac{\alpha}{1-\varepsilon}}(q)) \in R_{\kappa\alpha} \cap Q_{\frac{\alpha}{1-\varepsilon}}$. Since the latter complex is a clique complex, this will imply that for all $\sigma \in R_\alpha$, we have $\pi_{\frac{\alpha}{1-\varepsilon}}(\sigma) \in R_{\kappa\alpha} \cap Q_{\frac{\alpha}{1-\varepsilon}} \subseteq T_{\kappa\alpha}$ as desired. First, $(\pi_{\frac{\alpha}{1-\varepsilon}}(p), \pi_{\frac{\alpha}{1-\varepsilon}}(q)) \in R_{\kappa\alpha}$ as a direct consequence of Lemma 6.10.

Next, we need to show that $(\pi_{\frac{\alpha}{1-\varepsilon}}(p), \pi_{\frac{\alpha}{1-\varepsilon}}(q)) \in Q_{\frac{\alpha}{1-\varepsilon}}$. It suffices to show that $\mathrm{f}_{\frac{\alpha}{1-\varepsilon}}(\pi_{\frac{\alpha}{1-\varepsilon}}(p), \pi_{\frac{\alpha}{1-\varepsilon}}(q)) \leq \frac{2\alpha}{1-\varepsilon}$.

$$\begin{aligned}
\mathrm{f}_{\frac{\alpha}{1-\varepsilon}}(\pi_{\frac{\alpha}{1-\varepsilon}}(p), \pi_{\frac{\alpha}{1-\varepsilon}}(q)) &\leq \mathrm{f}_{\frac{\alpha}{1-\varepsilon}}(p,q) \\
&= d_{\mathbb{X}}(p,q) + s_p(\frac{\alpha}{1-\varepsilon}) + s_q(\frac{\alpha}{1-\varepsilon}) \\
&\leq d_{\mathbb{X}}(p,q) + \frac{2\varepsilon\alpha}{1-\varepsilon} \\
&\leq 2\alpha + \frac{2\varepsilon\alpha}{1-\varepsilon} \\
&= \frac{2\alpha}{1-\varepsilon} \qquad \square
\end{aligned}$$

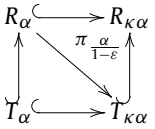
Now, we give conditions for when two projections induce contiguous simplicial maps between the sparse weighted Rips complexes T_δ and $T_{\kappa\delta}$.

Lemma 6.12. *Two projections π_α and π_β induce contiguous simplicial maps from $T_\delta \to T_{\kappa\delta}$, where $\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}$ whenever $\alpha, \beta \leq \frac{\delta}{1-\varepsilon}$ and there exists i so that $\frac{\lambda_{i+1}}{\varepsilon(1-\varepsilon)} \leq \alpha \leq \beta \leq \frac{\lambda_i}{\varepsilon(1-\varepsilon)}$.*

Proof. We simply observe that for any $\sigma \in T_\delta$, $\sigma \in Q_\rho$ for some $\rho \leq \delta$. If $\rho \leq \beta$ then Lemma 6.9 implies $\pi_\alpha(\sigma) \cup \pi_\beta(\sigma) \in Q_\beta$. Otherwise $\pi_\alpha(\sigma) \cup \pi_\beta(\sigma) = \sigma \in Q_\rho$. So in either case, we have $\pi_\alpha(\sigma) \cup \pi_\beta(\sigma) \in S_{\delta\gamma}$. Now, by Lemma 6.10, we have that $\pi_\alpha(\sigma) \cup \pi_\beta(\sigma) \in R_{\kappa\delta}$. So, we have that $\pi_\alpha(\sigma) \cup \pi_\beta(\sigma) \in R_{\kappa\delta} \cap S_{\kappa\delta} = T_{\kappa\delta}$ as desired. $\square$

We can now give the proof of the interleaving which will imply the desired approximation of the persistent homology.

Lemma 6.13. *For all $\alpha > 0$, the following diagram commutes the homology level.*



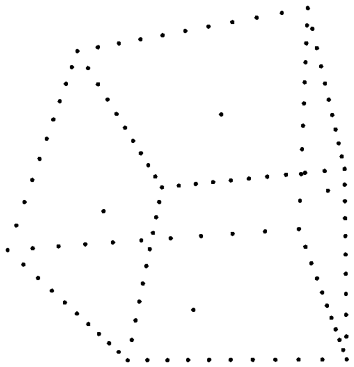


Fig. 2. Skeleton of a cube with outliers.

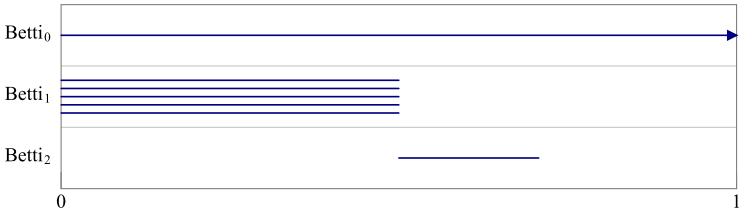


Fig. 3. Persistence diagram of a cube skeleton without noise.

Proof. By Lemma 6.10, the projection $\pi_{\frac{\alpha}{1-\varepsilon}}$ and the inclusion π_0 are contiguous and thus produce identical homomorphisms at the homology level. For the lower triangle it will suffice to show that the homomorphism induced by $\pi_{\frac{\alpha}{1-\varepsilon}}$ commutes with the one induced by the inclusion π_0. Let $\phi_i = \pi_{\frac{\lambda_i}{1-\varepsilon}}$ for $i = 1, \dots, n+1$. Now, Lemma 6.12 implies that ϕ_i and ϕ_{i+1} are contiguous. So, choosing k such that $\lambda_k \leq \varepsilon\alpha < \lambda_{k-1}$, we can apply Lemma 6.12 repeatedly to conclude that

$$\pi_{0\star} = \phi_{n+1\star} = \phi_{n\star} = \cdots = \phi_{k\star} = \pi_{\frac{\alpha}{1-\varepsilon}\star}. \qquad \square$$

7. Numerical illustration

In this section, we illustrate our results from three different perspectives: the quality of the approximation, the stability of the diagrams with respect to noise, and the size of the filtration after sparsification.

We used the ANN library [17] for the k-nearest neighbors search and code from Zomorodian following [23] for the persistence. The topology of the union of balls is acquired through the α-shapes implementation from the CGAL library [11].

Datasets For the first two parts, we consider the set of points in $\mathbb{R}^3$ obtained by sampling regularly the skeleton of the unit cube with 116 points. Then we add four noise points in the center of four of its faces such that two opposite faces are empty (Fig. 2).

We would like to compute the persistence diagram of the skeleton of the cube. We write this diagram $\mathrm{Dgm}(Skel)$. It contains five homology classes in dimension 1 and one in dimension 2, and it has the barcode representation given in Fig. 3.

For sparsification, we use a slightly bigger dataset composed of 10000 points regularly distributed on a curve rolled around a torus. The point set is shown on Fig. 4.

Approximation We work from now on with a mass parameter m such that $k = mn = 5$. The persistence diagram of $d_{\mu_P,m}$ is given in Fig. 5:

The diagrams obtained with our various approximations have very similar looks. We only show the one obtained with the sparse Rips filtration with a parameter $\epsilon = 0.5$ in Fig. 6.

To compare diagrams, we use the bottleneck distances between the diagrams. Fig. 7 shows the distance matrix between the various diagrams, while Fig. 8 shows some bottleneck distances between persistence diagrams of different dimensions. Note that $\mathrm{Dgm}(d_P)$ corresponds to the diagram obtained by using the distance function to the point cloud.

The largest difference is between $\mathrm{Dgm}(Skel)$ and $\mathrm{Dgm}(d_{\mu_P,m})$. This is partly due to an effect of shifting while using the distance to a measure. After this initial shift, the distances are small compared to the theoretical bounds. Notice that the different steps of the approximation do not have the same effect on all dimensions.

All diagrams obtained by the different approximations are closer to $\mathrm{Dgm}(Skel)$ than the persistence diagram of the distance to the point cloud, $\mathrm{Dgm}(d_P)$ given in Fig. 9. For inference purposes, one crucial parameter is the *signal-to-noise*

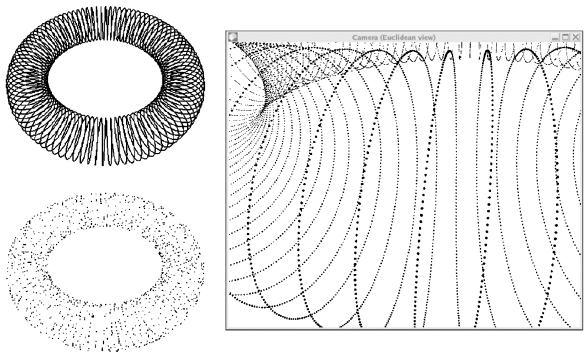


Fig. 4. Spiral on a torus.

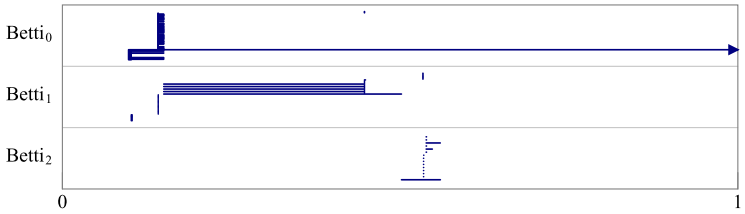


Fig. 5. Dgm($d_{\mu_P,m}$) for the cube skeleton with outliers with $k = 5$.

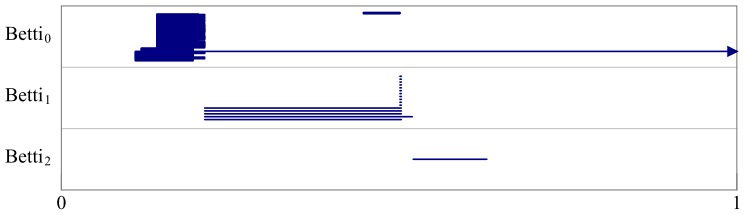


Fig. 6. Dgm($\{T_\alpha\}$) for the cube skeleton with outliers with $k = 5$ and $\epsilon = .5$.

	Dgm($Skel$)	Dgm($d_{\mu_P,m}$)	Dgm($d^P_{\mu_P,m}$)	Dgm(R_α)	Dgm(T_α)	Dgm(d_P)
Dgm($Skel$)	0	.1528	.1473	.1473	.1817	.25
Dgm($d_{\mu_P,m}$)	.1528	0	.09872	.0865	.1183	.2543
Dgm($d^P_{\mu_P,m}$)	.1473	.09872	0	.0459	.1084	.2642
Dgm(R_α)	.1473	.0865	.0459	0	.1128	.2598
Dgm(T_α)	.1817	.1183	.1084	.1128	0	.2484
Dgm(d_P)	.25	.2543	.2642	.2598	.2484	0

Fig. 7. Matrix of distances for the bottleneck distance.

Dgm(A)	Dgm(B)	dim 0	dim 1	dim 2
Dgm($Skel$)	Dgm($d_{\mu_P,m}$)	.05202	.1528	.1495
Dgm($d_{\mu_P,m}$)	Dgm($d^P_{\mu_P,m}$)	.09872	.0195	.0972
Dgm($d^P_{\mu_P,m}$)	Dgm($R_\alpha(P, d_{\mu_P,m})$)	.0007	.0044	.0459
Dgm($R_\alpha(P, d_{\mu_P,m})$)	Dgm($T_\alpha(P, d_{\mu_P,m})$)	.0872	.1128	.0026
Dgm($Skel$)	Dgm($d^P_{\mu_P,m}$)	.0405	.1473	.0982
Dgm($Skel$)	Dgm($T_\alpha(P, d_{\mu_P,m})$)	.1026	.1817	.098
Dgm($Skel$)	Dgm(d_P)	.25	.2071	.1481

Fig. 8. Bottleneck distances between diagrams.

ratio. We define it as the ratio between the smallest lifespan of topological feature we aim to infer and the longest lifespan of noise features. A ratio of 1 corresponds to a signal that is not differentiable from the noise and ∞ corresponds to a noiseless diagram. In our example, only the dimensions 1 and 2 are relevant as the dimension 0 diagram corresponding to connected components has only one relevant feature and its lifespan is infinite. Results are listed in Fig. 10.

Signal-to-noise ratios are clearly better than the one of Dgm(d_P). Some of the approximation steps improve the ratio. This is due to two phenomena.

Fig. 9. Dgm(d_P) for the cube skeleton with outliers.

Diagram	dim 1	dim 2
Dgm($Skel$)	∞	∞
Dgm($d_{\mu_P,m}$)	247	2.74
Dgm($d^P_{\mu_P,m}$)	69.8	43
Dgm($R_\alpha(P, d_{\mu_P,m})$)	∞	∞
Dgm($T_\alpha(P, d_{\mu_P,m})$)	132	∞
Dgm(d_P)	5.66	1

Fig. 10. Signal to noise ratios.

Standard deviation	.05	.1	.5
d_B in dimension 1	.1469	.2261	.2722
d_B in dimension 2	.047	.0914	.1046

Fig. 11. d_B between Dgm($\{T_\alpha\}$) with and without Gaussian noise.

Standard deviation	0	.05	.1	.5
Ratio in dimension 1	132	8.27	3.17	1.04
Ratio in dimension 2	∞	∞	100.2	∞

Fig. 12. Signal to noise ratio of Dgm($\{T_\alpha\}$) depending on noise intensity.

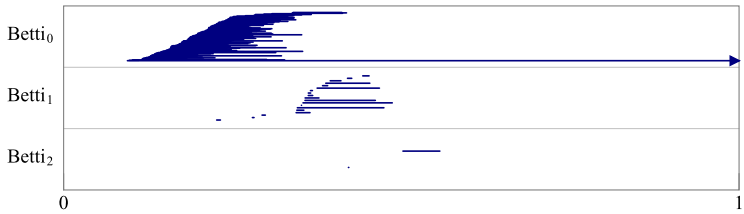


Fig. 13. Persistence diagram of $\{T_\alpha\}$ with $k = 5$, $\epsilon = 0.5$ and a Gaussian noise with standard deviation 0.1.

When one goes from $d_{\mu_P,m}$ to $d^P_{\mu_P,m}$, the filtration eliminates the cells of the kth order Voronoi diagram that are far from the point cloud. These cells induce local minima that produce noise features in the diagrams. Removing them cleans parts of the diagram. The same phenomenon happens with the witnessed k-distance previously mentioned.

Using the Rips filtration instead of the Čech also reduces some noise. It eliminates artifacts from simplices that are introduced and almost immediately killed in the Čech complex due to balls that intersect pairwise but have no common intersection.

Stability

The weighted Rips filtration is stable with respect to noise. We illustrate this by studying the effect of an isotropic noise on our skeleton of a cube. We consider three different standard deviations for our noise. Fig. 11 shows the bottleneck distances between the persistence diagram of the sparse weighted Rips structure with the Gaussian noise and the one without Gaussian noise.

Unsurprisingly, the bottleneck distance is increasing with the standard deviation of the noise. The signal-to-noise ratio shown in Fig. 12 is more interesting.

Inferring correctly the homology of the cube skeleton is possible with standard deviation 0.05 and 0.1. Fig. 13 shows the persistence diagram obtained with a standard deviation of 0.1. The ∞ in the 0.5 case in dimension 2 is not relevant as there is no noise but the feature is too small compared to the rest of the diagram as shown in Fig. 14. Note that 0.5 corresponds to half of the side of the cube, and thus, it is logical to be unable to retrieve any useful information.

Some structure appears even with standard deviation as large as 0.5. The three bigger features in dimension 1 are relevant. However, we miss two elements and it is difficult to decide where to draw the frontier between relevant and irrelevant features.

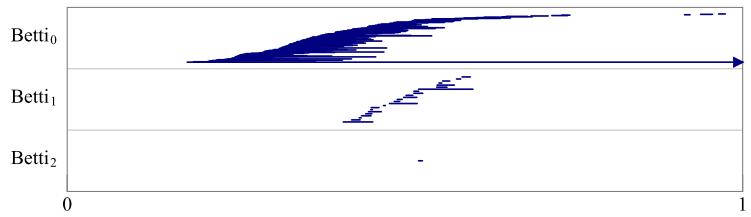


Fig. 14. Persistence diagram of $\{T_\alpha\}$ with $k = 5$, $\epsilon = .5$ and a Gaussian noise with standard deviation .5.

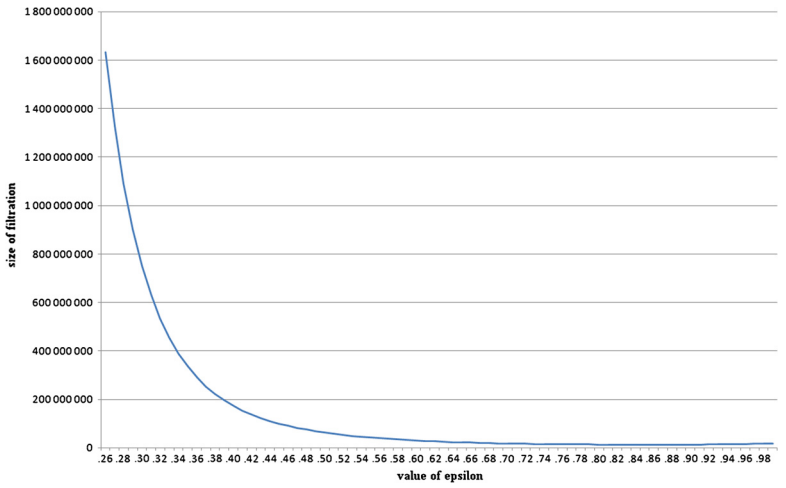


Fig. 15. Size of the filtration depending on ϵ for the spiral with a Gaussian noise of standard deviation .05.

Sparsification efficiency

We introduced sparsification in Section 6.4 to reduce the size of the Rips filtration. The method introduced a new parameter ϵ, and the size of the filtration depends heavily on ϵ. The evolution of the size of the filtration depending on the parameter ϵ is given in Fig. 15.

The minimum size is reached around $\epsilon = .86$. This minimum depends on the structure of the dataset. For example, considering a set of points uniformly sampled in a square, we obtain filtrations with monotonously decreasing size. While theoretical results depend on the doubling dimension of the ambient space, experimental results strongly suggests that it depends on the intrinsic doubling dimension, which can vary depending on the scale.

The filtration size is nearly constant after a rapid decrease. In this example, the size is of order 10^7 simplices for an input of 10^5 vertices. Computing persistent homology is tractable for any value in this range. Structure in the data helps reduce the complexity of the sparse filtration.

8. Conclusion

In this paper, we generalize several aspects of the existing theory on the persistent homology of distances to measures from Euclidean space to general metric spaces. Then, we showed how to efficiently approximate the sublevel sets of these distance functions with a linear number of metric balls. We gave a detailed analysis of the tightness of this approximation.

We then showed how to give a sparse filtration that gives a guaranteed close approximation to the persistent homology of the distance to the measure. This last construction was given in the more general context of power distances. Thus, we have given a way to efficiently compute the persistent homology of the sublevel set filtration of any power distance function built on points in metric space of low doubling dimension. Since power distances can be used to approximate many different kinds of functions, we expect that this technique will find many more uses in the future.

A different perspective on our approach is that we use the sparse Rips filtration analogously to how one might use a grid in Euclidean space. It provides a structure over which one can go on to study many different functions.

Lastly, we showed that this approach can be made practical, by providing some experimental results and analysis.

Acknowledgements

The authors acknowledge the support of the ANR TopData (ANR-13-BS01-0008) and the ERC Grant GUDHI. The authors also wish to thank the anonymous referees for their helpful comments that contributed to the improvement of the paper.

References

[1] Gunnar Carlsson, Topology and data, Bull. Am. Math. Soc. 46 (2009) 255–308.
[2] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, Steve Y. Oudot, Proximity of persistence modules and their diagrams, in: Proceedings of the 25th Annual Symposium on Computational Geometry, ACM, 2009, pp. 237–246.
[3] Frédéric Chazal, David Cohen-Steiner, Leonidas J. Guibas, Facundo Mémoli, Steve Y. Oudot, Gromov–Hausdorff stable signatures for shapes using persistence, Computer Graphics Forum, vol. 28, Wiley Online Library, 2009, pp. 1393–1403.
[4] Frédéric Chazal, David Cohen-Steiner, André Lieutier, A sampling theory for compact sets in Euclidean space, Discrete Comput. Geom. 41 (3) (2009) 461–479.
[5] Frédéric Chazal, David Cohen-Steiner, Quentin Mérigot, Geometric inference for probability measures, Found. Comput. Math. 11 (6) (2011) 733–751.
[6] Frédéric Chazal, Vin de Silva, Marc Glisse, Steve Oudot, The structure and stability of persistence modules, preprint, arXiv:1207.3674, 2012.
[7] Frédéric Chazal, Vin de Silva, Steve Oudot, Persistence stability for geometric complexes, preprint, arXiv:1207.3885, 2012.
[8] Frédéric Chazal, Steve Y. Oudot, Towards persistence-based reconstruction in Euclidean spaces, in: Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry, ACM, 2008, pp. 232–241.
[9] Kenneth L. Clarkson, Peter W. Shor, Applications of random sampling in computational geometry, II, Discrete Comput. Geom. 4 (1) (1989) 387–421.
[10] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, Stability of persistence diagrams, Discrete Comput. Geom. 37 (1) (2007) 103–120.
[11] Tran Kai Frank Da, Sébastien Loriot, Mariette Yvinec, 3D alpha shapes, in: CGAL User and Reference Manual. CGAL Editorial Board, 4.2 edition, 2013, http://www.cgal.org/Manual/4.2/doc_html/cgal_manual/packages.html#Pkg:AlphaShapes3.
[12] Tamal K. Dey, Fengtao Fan, Yusu Wang, Computing topological persistence for simplicial maps, preprint, arXiv:1208.5018, 2012.
[13] Herbert Edelsbrunner, John L. Harer, Computational Topology: An Introduction, American Mathematical Soc., 2010.
[14] Herbert Edelsbrunner, David Letscher, Afra Zomorodian, Topological persistence and simplification, in: Proceedings of 41st Annual Symposium on Foundations of Computer Science, 2000, IEEE, 2000, pp. 454–463.
[15] Leonidas Guibas, Dmitriy Morozov, Quentin Mérigot, Witnessed k-distance, Discrete Comput. Geom. 49 (1) (2013) 22–45.
[16] Allen Hatcher, Algebraic Topology, Cambridge University Press, 2002.
[17] David M. Mount, Sunil Arya, ANN: Library for Approximate Nearest Neighbour Searching, 1998.
[18] James R. Munkres, Elements of Algebraic Topology, Addison–Wesley, 1984.
[19] Partha Niyogi, Stephen Smale, Shmuel Weinberger, Finding the homology of submanifolds with high confidence from random samples, Discrete Comput. Geom. 39 (1–3) (2008) 419–441.
[20] Steve Y. Oudot, Donald R. Sheehy, Zigzag zoology: Rips zigzags for homology inference, in: Proceedings of the 29th Annual Symposium on Computational Geometry, 2013, pp. 387–396.
[21] Donald R. Sheehy, Linear-size approximations to the Vietoris–Rips filtration, Discrete Comput. Geom. 49 (4) (2013) 778–796.
[22] C. Villani, Topics in Optimal Transportation, American Mathematical Society, 2003.
[23] Afra Zomorodian, Gunnar Carlsson, Computing persistent homology, Discrete Comput. Geom. 33 (2) (2005) 249–274.