Efficient and Robust Persistent Homology for Measures

Mickaël Buchet, Frédéric Chazal, Steve Yann Oudot, Donald R. Sheehy

Efficient and Robust Persistent Homology for Measures

Mickaël Buchet* Frédéric Chazal† Steve Y. Oudot‡ Donald R. Sheehy§

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 [5, 21]. 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 [2, 16, 25]. 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 [6]. 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 [13, 22, 23], 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.

Contributions:

1. A Generalization of the Wasserstein stability and persistence stability of the distance to a measure for triangulable metric spaces. We describe the setting of these stability results and state the main theorems in this extended abstract; the proofs may be found in the full version [1].

2. A general method for approximating the sublevel sets of the distance to a measure by a union of balls (Fig. 1 right). 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, where d is the dimension of the ambient space.

3. A Generalization of the Vietoris-Rips filtration to weighted point sets called the weighted Rips filtration. This is the first construction for computing approximations to the distance to a measure in non-Euclidean metrics. Independently, this filtration comes with stability properties that make it useable for other applications in topological data analysis such as shape signatures [4].

4. 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. This is a significant improvement over the full weighted

*Inria Saclay Île-de-France, mickael.buchet@inria.fr.
†Inria Saclay Île-de-France, frederic.chazal@inria.fr.
‡Inria Saclay Île-de-France, steve.oudot@inria.fr.
§University of Connecticut, don.r.sheehy@gmail.com.

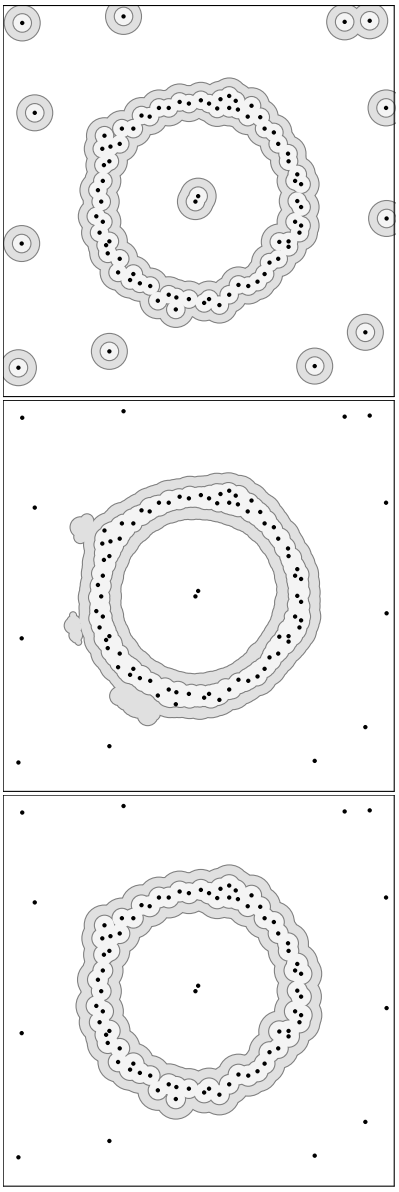


Figure 1: From top to bottom, two sublevel sets for d_P, $d_{\mu_P,m}$, and $d^P_{\mu_P,m}$ with $m = \frac{3}{|P|}$. The first is too sensitive to noise and outliers. The second is smoother, but substantially more difficult to compute. The third is our approximation, which is robust to noise, efficient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

Rips filtration, which has size 2^n in general or size $\binom{n}{d+1}$ if one considers only simplices up to dimension d.

Related Work. A filtration of the power distance to a weighted point set arises naturally in α-shapes [14]. Cazals et al. studied the persistent homology of filtration from unions of balls of varying radii using the so-called *conformal α-shape filtration* [3]. Similar to our work, the conformal α-shape attempts to put an alternative filtration on a complex that describes the distance function, the Delaunay triangulation in their case. That approach is limited to Euclidean space and suffers from the complexity blowup of the Delaunay triangulation.

The *witnessed k-distance* is another approach to approximating the distance to a measure proposed in [17]. 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, our method improves the approximation bounds obtained in [17] without any hypotheses on the intrinsic geometry.

2 Background

In this paper, we consider a metric space $\mathbb{X}$ with the 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 is 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.

The distance to a measure is defined as follows.

Definition 2.1. *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 $\delta_{\mu,l}$ is defined as

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

For a finite point set $P \subset \mathbb{X}$, the *empirical measure* μ_P is the normalized sum of Dirac measures δ_p:

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

The distance to an empirical measure has a simpler

description as

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

where $k = m|P|$ is assumed to be an integer and S ranges over the k nearest neighbors of x in P.

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 $\alpha < \beta$. 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 [7, 15].

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

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

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 $d_B^{\ln}$ is the bottleneck distance between diagrams after the change of coordinates $(\alpha,\beta) \mapsto (\ln\alpha, \ln\beta)$.

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 $\mathbb{U}$. These are used to define the algebraic stability of persistence modules as follows.

Theorem 2.1. (Stability of Persistence Modules [7]) *Let $\mathbb{U}$ and $\mathbb{V}$ be two q-tame persistence modules. If there exist 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 $\Psi\Phi = 1_{\mathbb{U}}^{2\epsilon}$, then*

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

Theorem 2.1 is the algebraic generalization of previous work on stability of persistence diagrams [11]. One consequence of the theorem is that for filtrations $\{F_\alpha\}$ and $\{G_\alpha\}$,

$$F_{\alpha/c} \subseteq G_\alpha \subseteq F_{c\alpha} \text{ for all } \alpha \text{ implies}$$

$$d_B^{\ln}(\mathrm{Dgm}(\{F_\alpha\}), \mathrm{Dgm}(\{G_\alpha\})) \leq \ln c.$$

Note that we pass to the log-bottleneck distance because the filtrations are interleaved multiplicatively. Similarly, for functions f and g such that $f/c \leq g \leq cf$, we have $d_B^{\ln}(\mathrm{Dgm}(f), \mathrm{Dgm}(g)) \leq \ln c$.

Because we want discrete objects to compute with, many of the topological spaces that appear in this paper are simplicial complexes. A *simplicial complex* K is a collection of subsets of a vertex set P that is closed under taking subsets, i.e. $\sigma \in K$ and $\tau \subset \sigma$ imply $\tau \in K$. Let X and Y be simplicial complexes. A *simplicial map* $f : X \to Y$ is a map between the corresponding vertex sets such 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 [20, Chapter 1].

3 Persistence and Stability of the Distance to a Measure in a Metric Space

To compare the persistence diagram computed from a finite sample to that of the distance to the underlying measure on the (perhaps infinite) metric space requires that the distance to the measure is sufficiently tame so that its persistence diagram is well-defined. For this, we require the underlying metric space to be *triangulable*, i.e., homeomorphic to a locally finite simplicial complex. In particular, the following theorem gives conditions under which there exists a ground truth to compare to, thus making it possible to speak coherently about approximation.

THEOREM 3.1. *Let μ be a probability measure on a metric space $\mathbb{X}$. Then, $d_{\mu,m}$ is 1-Lipschitz, and if $\mathbb{X}$ is triangulable, then* $\mathrm{Dgm}(d_{\mu,m})$ *is well-defined for any mass parameter $m \in]0,1]$.*

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 [24, Sec. 7.1]).

THEOREM 3.2. *Let μ and ν be probability measures on a triangulable metric space $\mathbb{X}$. For all $m \in]0,1]$,*

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

The proofs of Theorems 3.1 and 3.2 are given in the full version [1]. The techniques are similar to those used in previous work in the Euclidean setting [6, 7].

4 Approximating the Distance to a Measure

To compute the persistence diagram of the sublevel sets filtration of $d_{\mu,m}$, one must represent 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 [17] for Euclidean space.

Given a metric space $\mathbb{X}$, a finite set P and a function $w : P \to \mathbb{R}$, the *power distance* f associated with (P,w) is

$$f(x) = \sqrt{\min_{p\in P} d_{\mathbb{X}}(p,x)^2 + w_p^2}, \tag{4.1}$$

where w_p is the value of w at the point p. The function w can be defined on a superset of P. Remark that we restrict ourselves to the case of positive weights. This is sufficient for our purpose and avoids technicalities. 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 the ball is empty when the radius is imaginary.

We introduce the parameter $k = m|P|$. To simplify the exposition we assume that k is an integer. In Euclidean space, the distance to an empirical measure is a power distance:

$$d_{\mu_P,m}(x) = \sqrt{\min_q \|x-q\|^2 + w_q^2}, \tag{4.2}$$

where q ranges over all barycenters of k-tuples $S \subseteq P$ and $w_q^2 = \frac{1}{k}\sum_{p\in S}\|p-q\|^2$.

In Euclidean space, it is possible to compute the sublevel sets of $d_{\mu_P,m}$ exactly. They are unions of balls, however $\Omega(k^{\lceil \frac{d+1}{2}\rceil} n^{\lfloor \frac{d+1}{2}\rfloor})$ balls may be required [10]. For any measure μ, we introduce the following approximation that has much smaller complexity, requiring only $O(n)$ balls.

$$d^P_{\mu,m}(x) = \sqrt{\min_{p\in P} d_{\mathbb{X}}(p,x)^2 + d_{\mu,m}(p)^2} \tag{4.3}$$

This is the power distance to the points of P with weights given by the distance to the measure. It only requires computing $d_{\mu,m}$ at a finite collection of points. It is easy to see that this approximation cannot be significantly smaller than the distance to the true measure, as shown in the following lemma which provides the lower bound for both Theorem 4.1 and Theorem 4.2.

LEMMA 4.1. *Let μ be a probability measure on a metric space $\mathbb{X}$ and let $m \in]0,1]$ be a mass parameter. If P is a nonempty subset of $\mathbb{X}$, then $d_{\mu,m} \leq \sqrt{2}\, d^P_{\mu,m}$.*

Proof. Let x be any point of $\mathbb{X}$. By (4.3), there exists $p \in P$ such that $d^P_{\mu,m}(x)^2 = d_{\mathbb{X}}(p,x)^2 + d_{\mu,m}(p)^2$. Since $d_{\mu,m}$ is 1-Lipschitz (Theorem 3.1), we get

$$\begin{aligned} d_{\mu,m}(x)^2 &\leq (d_{\mathbb{X}}(p,x) + d_{\mu,m}(p))^2 \\ &\leq 2(d_{\mathbb{X}}(p,x)^2 + d_{\mu,m}(p)^2) = 2\, d^P_{\mu,m}(x)^2. \end{aligned}$$

The approximation $d^P_{\mu,m}$ yields the following guarantees for approximating the distance to an empirical measure in Euclidean space.

THEOREM 4.1. *Let P be a finite point set in $\mathbb{R}^d$ and let $m \in]0,1]$ be a mass parameter. Then,*

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

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

Proof. The lower bound follows from Lemma 4.1 when applied to the empirical measure μ_P. For the upper bound, let x be a point in $\mathbb{R}^d$, and let S be the k nearest neighbors of x in P (breaking ties arbitrarily). Let $\bar{x}$ be the barycenter of S with squared weight $w_{\bar{x}}^2 = \frac{1}{k}\sum_{q\in S}\|\bar{x}-q\|^2$. In Euclidean space, the barycenter minimizes the sum of squared distances to the points in S, so

$$w_{\bar{x}}^2 = \frac{1}{k}\sum_{q\in S}\|\bar{x}-q\|^2 \leq \frac{1}{k}\sum_{q\in S}\|x-q\|^2 = d_{\mu_P,m}(x)^2. \tag{4.4}$$

It follows from the definitions of $d^P_{\mu_P,m}$ and $d_{\mu_P,m}$ that for all $p \in P$,

$$d^P_{\mu_P,m}(x)^2 \leq \|x-p\|^2 + d_{\mu_P,m}(p)^2 \leq \|x-p\|^2 + \|p-\bar{x}\|^2 + w_{\bar{x}}^2. \tag{4.5}$$

The inequalities above come from replacing minimizations in (4.3) and (4.2) respectively with particular values. Since (4.5) holds for all $p \in P$, we can average the values on the right hand side over all values of $q \in S$ to get a new bound as follows.

$$\begin{aligned} d^P_{\mu_P,m}(x)^2 &\leq \frac{1}{k}\sum_{q\in S}\left(\|x-q\|^2 + \|q-\bar{x}\|^2 + w_{\bar{x}}^2\right) \\ &\leq 3\; d_{\mu_P,m}(x)^2. \end{aligned}$$

The last inequality follows from (4.4) and proves the desired upper bound.

The multiplicative interleaving of the functions implies an additive interleaving of the persistence modules of the sublevel sets filtrations on the log scale. So, the relation between persistence diagrams follows from Theorem 2.1.

The *witnessed k-distances* approach of Guibas et al. gives another way to approximate $d_{\mu_P,m}$ [17]. The bounds in Theorem 4.1 are tighter than those given in Lemma 3.3 of [17]. In the full version of this paper, we give a new analysis of witnessed k-distances to show equally tight bounds for the approximation. However, the real power of our approach is apparent when considering non-Euclidean metrics. The witnessed k-distance cannot be defined in these cases because it relies on the existence of barycenters. Our approximation gives the following guarantee.

THEOREM 4.2. *Let P be a finite point set of a metric space $\mathbb{X}$ and let $m \in]0,1]$ be a mass parameter. Then,*

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

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

Proof. The lower bound is implied by Lemma 4.1. Let $x \in \mathbb{X}$ be any point and let p be its nearest neighbor in P. Using the definition of $d^P_{\mu_P,m}$ and the Lipschitz property of $d_{\mu_P,m}$ (Theorem 3.1), we get the following.

$$\begin{aligned} d^P_{\mu_P,m}(x)^2 &\leq d_{\mathbb{X}}(x,p)^2 + d_{\mu_P,m}(p)^2 \\ &\leq d_{\mathbb{X}}(x,p)^2 + (d_{\mu_P,m}(x) + d_{\mathbb{X}}(x,p))^2 \\ &\leq 3\; d_{\mathbb{X}}(x,p)^2 + 2\; d_{\mu_P,m}(x)^2 \end{aligned}$$

Since $d_{\mathbb{X}}(x,p) \leq d_{\mu_P,m}(x)$, it follows that $d^P_{\mu_P,m}(x)^2 \leq 5\; d_{\mu_P,m}(x)^2$ as desired. The multiplicative interleaving of the functions implies an additive interleaving of the persistence modules of the sublevel sets filtrations on the log scale. So, the relation between persistence diagrams follows from Theorem 2.1.

The bounds proved in Theorems 4.1 and 4.2 are tight. In the full version of this paper, we give some examples of points sets in metric spaces where some points achieve the lower bound and others achieve the upper bound [1]. The different factors in the upper and lower bounds imply that one can improve slightly the persistence diagram approximation to $\ln(\sqrt[4]{10})$ by scaling $d^P_{\mu_P,m}$ by $\frac{\sqrt[4]{10}}{\sqrt{2}}$. The effect of scaling on the log-scale persistence diagram is just a shift along the direction of the diagonal.

5 The Weighted Rips Filtration

Given a weighted set (P,w) and the associated power distance f (as in (4.1)), one can introduce a generalization of the Rips filtration that is adapted to the weighted setting as has been done in [17]. 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 5.1) and that it gives a guaranteed approximation to the persistence diagram of the distance to an empirical measure (Theorem 5.2).

Recall that the sublevel set $f^{-1}(]-\infty,\alpha])$ is a union of balls of different radii centered on the points P. The nerve of this collection of balls is the *weighted Čech complex*, a simplicial complex composed of all subsets σ of P such that $\bigcap_{p\in\sigma} B(p, r_p(\alpha)) \neq 0$. For a wide class of metric spaces, this weighted Čech complex will have the same homology as the union of balls. However, computing the Čech complex requires testing if a collection of metric balls has a common intersection, which may be difficult. Instead, we use a weighted version of the Rips complex that only requires distance computations.

DEFINITION 5.1. *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 $\alpha \geq 0$.*

As shown in [17], the weighted Rips and Čech complexes share many properties with their unweighted analogues. For example, as in the unweighted case,

$$C_\alpha(P,w) \subseteq R_\alpha(P,w) \subseteq C_{2\alpha}(P,w) \tag{5.6}$$

for all $\alpha \geq 0$. See [1, 17] for a proof.

The first theorem we prove about weighted distance functions shows that as long as the weights do not vary too wildly, two Hausdorff-close samples yield weighted Rips filtrations with similar persistence diagrams. This

provides a useful stability property for weighted Rips filtrations built on samples.

THEOREM 5.1. *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,* $\mathrm{Dgm}(\{R_\alpha(P,w)\})$ *and* $\mathrm{Dgm}(\{R_\alpha(Q,w)\})$ *are well-defined and*

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

Proof. [Proof Sketch] Theorem 2.1 implies that it suffices to find ϵ-homomorphisms between $\mathrm{H}_*\{R_\alpha(P,w)\}$ and $\mathrm{H}_*\{R_\alpha(Q,w)\}$ for $\epsilon = (1+t)d_H(P,Q)$ that commute appropriately with the ϵ-shifts. Following the pattern established in [8], we construct these homomorphisms by first defining projections from P and Q to nearest neighbors in Q and P respectively. The main work of the proof is to show that any such projections induce simplicial maps $R_\alpha(P,w) \to R_{\alpha+\epsilon}(Q,w)$ and $R_\alpha(Q,w) \to R_{\alpha+\epsilon}(P,w)$ and that these simplicial maps are contiguous with the corresponding inclusion maps. The full details may be found in [1].

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 [18]. This equivalence is extended to filtrations by the Persistent Nerve Lemma [9].

The following Theorem is the main result of this section and gives a guarantee on the approximation of $\mathrm{Dgm}(d_{\mu_P,m})$ using the weighted Rips filtration.

THEOREM 5.2. *Let $\mathbb{X}$ be a triangulable metric space with the good cover property and let P be a finite subset of $\mathbb{X}$. Then,* $d_B^{\ln}(\mathrm{Dgm}(d_{\mu_P,m}), \mathrm{Dgm}(\{R_\alpha(P, d_{\mu_P,m})\})) \leq \ln(2\sqrt{5})$.

Proof. Since $\mathbb{X}$ is triangulable, Theorem 4.2 implies that

$$d_B^{\ln}(\mathrm{Dgm}(d_{\mu_P,m}), \mathrm{Dgm}(d^P_{\mu_P,m})) \leq \ln(\sqrt{5}). \tag{5.7}$$

The sublevel sets of $d^P_{\mu_P,m}$ are the unions of balls centered at points of P with weights given by $d_{\mu_P,m}$. The weighted Čech complex is the nerve of this set of balls, so the Persistent Nerve Lemma and the good cover property of $\mathbb{X}$ imply that

$$\mathrm{Dgm}(d^P_{\mu_P,m}) = \mathrm{Dgm}(\{C_\alpha(P, d_{\mu_P,m})\}). \tag{5.8}$$

The multiplicative interleaving of the weighted Čech and weighted Rips filtrations given in (5.6) implies an additive interleaving of their persistence modules on the log scale. Thus, Theorem 2.1 implies that

$$d_B^{\ln}(\mathrm{Dgm}(\{C_\alpha(P, d_{\mu_P,m})\}), \mathrm{Dgm}(\{R_\alpha(P, d_{\mu_P,m})\})) \leq \ln(2). \tag{5.9}$$

The result now follows from (5.7), (5.8), (5.9), and the triangle inequality for $d_B^{\ln}$.

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 [23], 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 construction cannot be applied directly here, since the weighted distance does not induce a metric. Moreover, if all weights are equal to some large constant and the points are on a circle, their pairwise distances will all be equal. The sparsification technique will thus be utterly inefficient. As we will show, this difficulty can be overcome as long as the weights are Lipschitz with respect to the input metric.

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,*

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

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 $(p_1, \ldots, p_n)$ be a furthest point sampling of the points P in a finite metric space $\mathbb{X}$. That is, $p_i = \mathrm{argmax}_{p \in P \setminus P_{i-1}} d_{\mathbb{X}}(p, P_{i-1})$, where $P_{i-1} = \{p_1, \ldots, p_{i-1}\}$ with p_1 chosen arbitrarily. 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 furthest point sampling has the nice property that each prefix P_i is a λ_i-*net* in the sense that $d_{\mathbb{X}}(p, P_i) \leq \lambda_i$ for all $p \in P$, and $d_{\mathbb{X}}(p, q) \geq \lambda_i$ for all $p, q \in P_i$. We extend these nets to an arbitrary parameter γ by defining

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

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 $N_{\varepsilon\gamma}$. However, this can add significant noise to the persistence diagram compared to the Rips filtration. This noise can be reduced to ε on the log-scale by a careful perturbation of the distance, where $\varepsilon < 1$ is a user-provided parameter. 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 6.1. *Given the nets $\{N_\gamma\}$, and distance functions f_α, the* sparse Rips complex *at scale α is*

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

and the sparse Rips filtration *is the sequence of spaces $\{S_\beta\}_{\beta \geq 0}$, where $S_\beta = \bigcup_{\alpha \leq \beta} Q_\alpha$.*

6.2 Projection onto nets and the induced simplicial maps The following projection functions provide our main tool for defining maps between complexes.

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

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

A basic property of the perturbed distances is that replacing a point with its projection, does not increase the perturbed distance to the other points.

LEMMA 6.1. *For all $p, q \in P$ and all $\alpha \geq 0$, $\mathrm{f}_\alpha(p, \pi_\alpha(q)) \leq \mathrm{f}_\alpha(p, q)$.*

Proof. See [23, Lemma 7] or [1] for a proof.

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 6.2) or weighted Rips complexes (Lemma 6.4).

LEMMA 6.2. *Two projections π_α and π_β induce contiguous simplicial maps $Q_\rho \to Q_\beta$ 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. The proof for this variant of the sparse Rips complex can be found in the full version [1], though it is virtually identical to [23, Lemma 9].

To prove the analogous result for sparse weighted Rips complexes, we first need a lemma that describes the effect of different projections on the endpoints of an edge.

LEMMA 6.3. *Let (p, q) be an edge of R_δ with $\alpha, \beta \leq \frac{\delta}{1-\varepsilon}$, then $(\pi_\alpha(p), \pi_\beta(q)) \in R_{\kappa\delta}$ and $(\pi_\alpha(p), \pi_\beta(p)) \in R_{\kappa\delta}$, where $\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}$.*

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

$$d_{\mathbb{X}}(p, \pi_\alpha(p)) \leq \varepsilon\alpha \leq \frac{\varepsilon\delta}{1-\varepsilon}$$

$$d_{\mathbb{X}}(q, \pi_\beta(q)) \leq \varepsilon\beta \leq \frac{\varepsilon\delta}{1-\varepsilon}$$

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

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

The last inequality follows from the fact that w is t-Lipschitz (a full proof can be found in [1]). The bound above is precisely the condition necessary to guarantee that $(\pi_\alpha(p),\pi_\beta(q)) \in R_{\kappa\delta}$ as desired. The proof is symmetric to show $(\pi_\alpha(p),\pi_\beta(p)) \in R_{\kappa\delta}$ after recalling that $r_p \geq 0$.

LEMMA 6.4. *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. Fix α, β, and δ so that $\alpha,\beta \leq \frac{\delta}{1-\varepsilon}$. Let (p,q) be an edge from R_δ. Lemma 6.3 implies that all edges of the tetrahedron $\{\pi_\alpha(p),\pi_\alpha(q),\pi_\beta(p),\pi_\beta(q)\}$ are in $R_{\kappa\delta}$. It follows that for any simplex $\sigma \in R_\delta$, every edge of $\pi_\alpha(\sigma)\cup\pi_\beta(\sigma)$ is in $R_{\kappa\delta}$. The definition of the weighted Rips complex implies that every clique is a simplex, so $\pi_\alpha(\sigma)\cup\pi_\beta(\sigma) \in R_{\kappa\delta}$. Thus, π_α and π_β induce contiguous simplicial maps from $R_\delta \to R_{\kappa\delta}$ as desired.

All of the homomorphisms in the persistence module interleaving will be induced by projections. We first need to check that the projection $\pi_{\frac{\alpha}{1-\varepsilon}}$ induces a simplicial map from R_δ to $T_{\kappa\delta}$, where $\kappa = \frac{1+\sqrt{1+t^2}\,\varepsilon}{1-\varepsilon}$.

LEMMA 6.5. *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. It will suffice to 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 $\pi_{\frac{\alpha}{1-\varepsilon}}(\sigma) \in R_{\kappa\alpha} \cap Q_{\frac{\alpha}{1-\varepsilon}} \subseteq T_{\kappa\alpha}$ for all $\sigma \in R_\alpha$ as desired.

Lemma 6.1 and the definitions of $\mathrm{f}_{\frac{\alpha}{1-\varepsilon}}$, s_p, s_q, and R_α imply

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

Thus, $(\pi_{\frac{\alpha}{1-\varepsilon}}(p),\pi_{\frac{\alpha}{1-\varepsilon}}(q)) \in Q_{\frac{\alpha}{1-\varepsilon}}$. Lemma 6.4 implies that $(\pi_{\frac{\alpha}{1-\varepsilon}}(p),\pi_{\frac{\alpha}{1-\varepsilon}}(q)) \in R_{\kappa\alpha}$. So, we conclude that indeed $(\pi_{\frac{\alpha}{1-\varepsilon}}(p),\pi_{\frac{\alpha}{1-\varepsilon}}(q)) \in R_{\kappa\alpha} \cap Q_{\frac{\alpha}{1-\varepsilon}}$

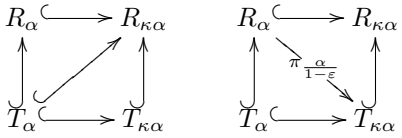
Now, we give conditions for when two projections induce contiguous simplicial maps between the sparse weighted Rips complexes T_δ and $T_{\kappa\delta}$.

LEMMA 6.6. *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 6.2 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 $\pi_\alpha(\sigma)\cup\pi_\beta(\sigma) \in S_{\kappa\delta}$. Now, by Lemma 6.4, 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.

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

Proof. [Proof of Theorem 6.1] To prove a multiplicative κ-interleaving between $\{R_\alpha\}$ and $\{T_\alpha\}$, it suffices to provide homomorphisms $\mathrm{H}_*(R_\alpha) \to \mathrm{H}_*(T_{\kappa\alpha})$ and $\mathrm{H}_*(T_\alpha) \to \mathrm{H}_*(R_{\kappa\alpha})$ for all $\alpha \geq 0$. In both cases, the homomorphisms will be induced by projections. Since $T_\alpha \subseteq R_{\kappa\alpha}$, we consider the homomorphism $\mathrm{H}_*(T_\alpha) \to \mathrm{H}_*(R_{\kappa\alpha})$ induced by the inclusion π_0. The other homomorphism is $\mathrm{H}_*(R_\alpha) \to \mathrm{H}_*(T_{\kappa\alpha})$ induced by the projection $\pi_{\frac{\alpha}{1-\varepsilon}}$, which is a simplicial map on the complexes as shown in Lemma 6.5. It will suffice to prove that for all $\alpha > 0$, the following diagrams commute at the homology level.



The left diagram commutes because all maps are inclusions and therefore it also commutes at the homology level. For the right diagram, the upper triangle commutes at the homology level by Lemma 6.4 and the observation that the π_0 is the inclusion since contiguous maps induce the same homomorphism 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 that produced by the inclusion π_0. Let $\phi_i = \pi_{\frac{\lambda_i}{1-\varepsilon}}$ for $i = 1,\ldots,n+1$. The subscript $*$ indicates the homomorphism induced by a simplicial map at the homology level. Now, Lemma 6.6 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 6.6 repeatedly to conclude that

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

7 Concluding Remarks on an Implementation

As a proof of concept, we have implemented the weighted Rips and sparse Rips filtrations of Sections 5-6, and we have tested them against manufactured data. Our C++ implementation works in Euclidean spaces and uses the ANN library [19] for proximity queries. Persistence barcodes are computed using A. Zomorodian's implementation of the persistence algorithm [25]. Our experimental results illustrate the three main selling points of our approach:

- the quality of the output when approximating the persistence diagram of the distance to a measure,
- the stability of the output with respect to non-local perturbations of the input, and
- the scalability of the approach, made possible by the control over the size of the sparse Rips filtration.

Several input point clouds have been tested, with different sets of parameters, to assess the relevance of these observations. However, no real-life data has been considered yet, and further investigations along this line are needed to fully validate the (sparse) weighted Rips filtrations as practical tools. Details on our data sets, results and timings can be found in the full version of the paper [1].

In this section, we illustrate our results 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 [19] for the k-nearest neighbors search and code from Zomorodian following [25] for the persistence. The topology of the union of balls is acquired through the α-shapes implementation from the CGAL library [12].

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.

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 Figure 3.

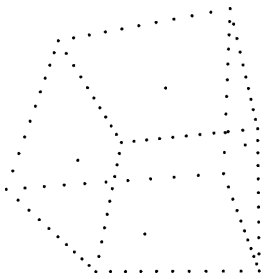


Figure 2: Skeleton of a cube with outliers

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 Figure 4.

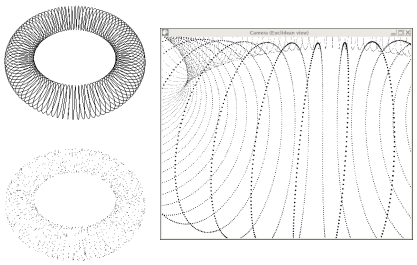


Figure 4: Spiral on a torus

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 Figure 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 Figure 6.

To compare diagrams, we use the bottleneck distances between the diagrams. Figure 7 shows the distance matrix between the various diagrams, while Figure 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 distance 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 Figure 9. For inference purposes, one crucial parame-

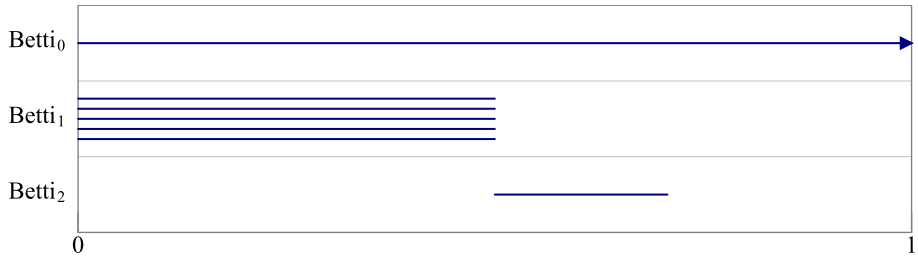


Figure 3: Persistence diagram of a cube skeleton without noise

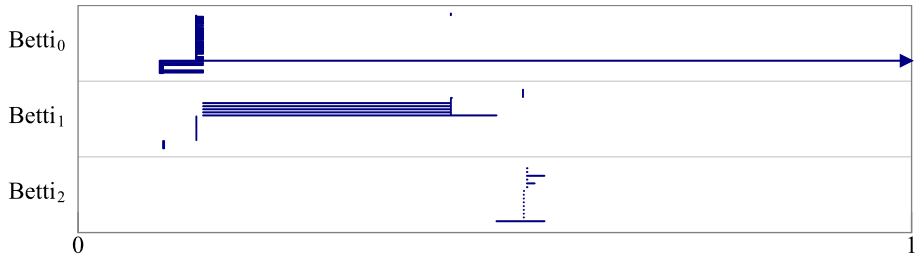


Figure 5: Dgm($d_{\mu_P,m}$) for the cube skeleton with outliers with $k = 5$

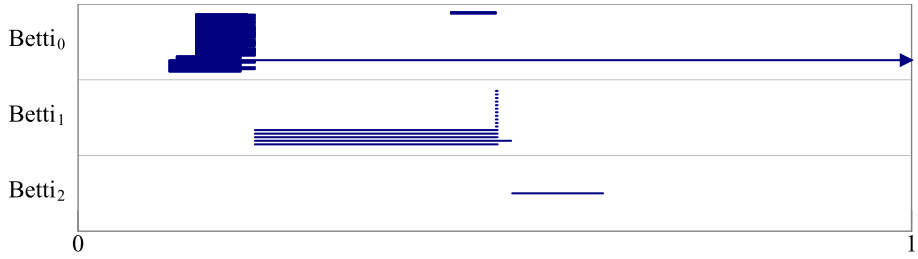


Figure 6: Dgm($\{T_\alpha\}$) for the cube skeleton with outliers with $k = 5$ and $\epsilon = .5$

	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

Figure 7: Matrix of distances for the bottleneck distance

ter is the *signal-to-noise 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 Figure 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.

When one goes from $d_{\mu_P,m}$ to $d^P_{\mu_P,m}$, the filtration eliminates the cells of the k^{th} order Voronoi diagram that are far from the point cloud. These cells induce

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

Figure 8: Bottleneck distances between diagrams

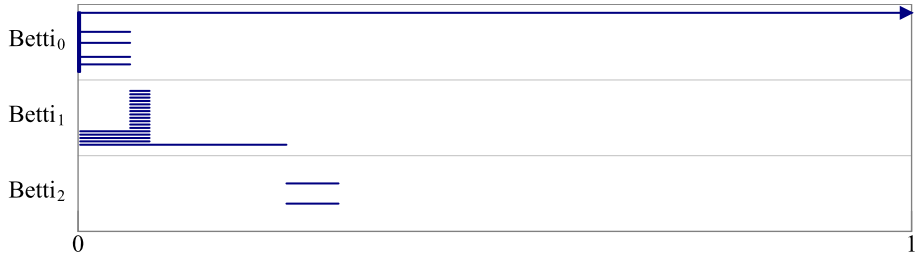


Figure 9: $\mathrm{Dgm}(d_P)$ for the cube skeleton with outliers

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

Figure 10: Signal to noise ratios

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 perviously 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. Figure 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 standard deviation of the noise. The signal-to-noise ratio shown in Figure 12 is more interesting.

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

Figure 11: d_b between $\mathrm{Dgm}(\{T_\alpha\})$ 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	∞

Figure 12: Signal to noise ratio of $\mathrm{Dgm}(\{T_\alpha\})$ depending on noise intensity

Inferring correctly the homology of the cube skeleton is possible with standard deviation 0.05 and 0.1. Figure 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 Figure 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.

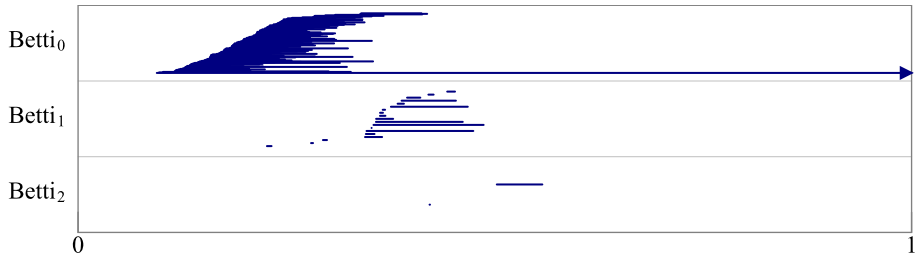


Figure 13: Persistence diagram of $\{T_\alpha\}$ with $k = 5$, $\epsilon = 0.5$ and a Gaussian noise with standard deviation 0.1

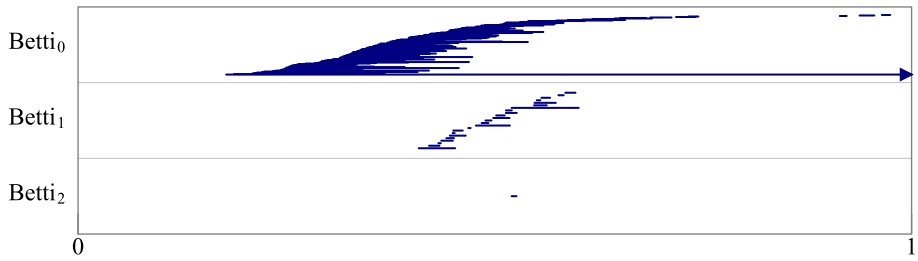


Figure 14: Persistence diagram of $\{T_\alpha\}$ with $k = 5$, $\epsilon = .5$ and a Gaussian noise with standard deviation .5

Sparsification efficiency

We introduced sparsification in Section 6 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 Figure 15.

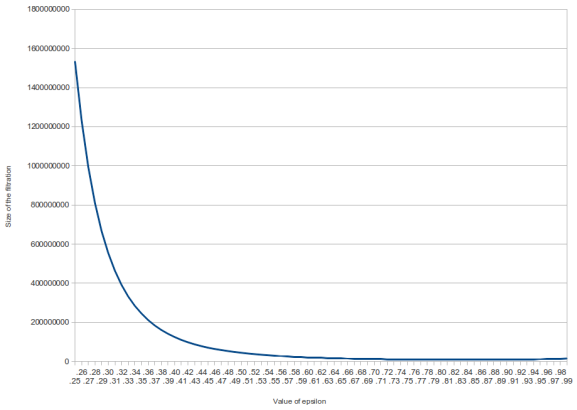


Figure 15: Size of the filtration depending on ϵ for the spiral

The minimum size is reached around $\epsilon = .83$. This minimum depends on the structure of the dataset. For example, considering a set of points uniformly sampled in a square, we obtain decreasing size of the filtration.

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.

Acknowledgements

The authors acknowledge the support of the ANR TopData (ANR-13-BS01-0008) and the ERC Grant GUDHI.

References

[1] Mickaël Buchet, Frédéric Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust persistent homology for measures. *arXiv preprint arXiv:1306.0039v2*, 2014.

[2] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308, 2009.

[3] Frederic Cazals, Joachim Giesen, Mark Pauly, and Afra Zomorodian. The conformal alpha shape filtration. *The Visual Computer*, 22(8):531–540, 2006.

[4] Frédéric Chazal, David Cohen-Steiner, Leonidas J. Guibas, Facundo Mémoli, and Steve Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library, 2009.

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

[6] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

[7] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

[8] Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *arXiv preprint arXiv:1207.3885*, 2012.

[9] Frédéric Chazal and Steve Y. Oudot. Towards persistence-based reconstruction in Euclidean spaces. In *Proceedings of the twenty-fourth Annual Symposium on Computational Geometry*, pages 232–241. ACM, 2008.

[10] Kenneth L. Clarkson and Peter W. Shor. Applications of random sampling in computational geometry, II. *Discrete & Computational Geometry*, 4(1):387–421, 1989.

[11] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

[12] Tran Kai Frank Da, Sébastien Loriot, and Mariette Yvinec. 3D alpha shapes. In *CGAL User and Reference Manual*. CGAL Editorial Board, 4.2 edition, 2013.

[13] Tamal K. Dey, Fengtao Fan, and Yusu Wang. Computing topological persistence for simplicial maps. *arXiv preprint arXiv:1208.5018*, 2012.

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

[15] Herbert Edelsbrunner and John L. Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[16] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 454–463. IEEE, 2000.

[17] Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. Witnessed k-distance. *Discrete & Computational Geometry*, 49(1):22–45, 2013.

[18] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[19] David M. Mount and Sunil Arya. ANN: Library for approximate nearest neighbour searching. 1998.

[20] James R. Munkres. *Elements of Algebraic Topology*. Addison-Wesley, 1984.

[21] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[22] Steve Y. Oudot and Donald R. Sheehy. Zigzag zoology: Rips zigzags for homology inference. In *Proceedings of the 29th annual Symposium on Computational Geometry*, pages 387–396, 2013.

[23] Donald R. Sheehy. Linear-size approximations to the Vietoris-Rips filtration. *Discrete & Computational Geometry*, 49(4):778–796, 2013.

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

[25] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.