Towards Persistence-Based Reconstruction in Euclidean Spaces

Frédéric Chazal, Steve Oudot

Towards Persistence-Based Reconstruction in Euclidean Spaces

Frédéric Chazal
INRIA Futurs
Parc Orsay Université
4, rue Jacques Monod - Bât. P
91893 ORSAY Cedex, France
frederic.chazal@inria.fr

Steve Y. Oudot
INRIA Futurs
Parc Orsay Université
4, rue Jacques Monod - Bât. P
91893 ORSAY Cedex, France
steve.oudot@inria.fr

Abstract

Manifold reconstruction has been extensively studied among the computational geometry community for the last decade or so, especially in two and three dimensions. Recently, significant improvements were made in higher dimensions, leading to new methods to reconstruct large classes of compact subsets of Euclidean space $\mathbb{R}^d$. However, the complexities of these methods scale up exponentially with d, which makes them impractical in medium or high dimensions, even for handling low-dimensional submanifolds.

In this paper, we introduce a novel approach that stands in-between reconstruction and topological estimation, and whose complexity scales up with the intrinsic dimension of the data. Specifically, our algorithm combines two paradigms: greedy refinement, and topological persistence. It builds a set of landmarks iteratively, while maintaining nested pairs of complexes, whose images in $\mathbb{R}^d$ lie close to the data, and whose persistent homology eventually coincides with the one of the underlying shape. When the data points are sufficiently densely sampled from a smooth m-submanifold of $\mathbb{R}^d$, our method retrieves the homology of the submanifold in time at most $c(m)n^5$, where n is the size of the input and $c(m)$ is a constant depending solely on m. It can also provably well handle a wide range of compact subsets of $\mathbb{R}^d$, though with worse complexities.

Along the way to proving the correctness of our algorithm, we obtain new results on Čech, Rips, and witness complex filtrations in Euclidean spaces. Specifically, we show how previous results on unions of balls can be transposed to Čech filtrations. Moreover, we propose a simple framework for studying the properties of filtrations that are intertwined with the Čech filtration, among which are the Rips and witness complex filtrations. Finally, we investigate further on witness complexes and quantify a conjecture of Carlsson and de Silva, which states that witness complex filtrations should have cleaner persistence barcodes than Čech or Rips filtrations, at least on smooth submanifolds of Euclidean spaces.

1 Introduction

The problem of reconstructing unknown structures from finite collections of data samples is ubiquitous in the Sciences, where it has many different variants, depending on the nature of the data and on the targeted application. In the last decade or so, the computational geometry community has gained a lot of interest in manifold reconstruction, where the goal is to reconstruct submanifolds of Euclidean spaces from point clouds. In particular, efficient solutions have been proposed in dimensions two and three, based on the use of the Delaunay triangulation – see [8] for a survey. In these

methods, the unknown manifold is approximated by a simplicial complex that is extracted from the full-dimensional Delaunay triangulation of the input point cloud. The success of this approach is explained by the fact that, not only does it behave well on practical examples, but the quality of its output is guaranteed by a sound theoretical framework. Indeed, the extracted complex is usually shown to be equal, or at least close, to the so-called *restricted Delaunay triangulation*, a particular subset of the Delaunay triangulation whose approximation power is well-understood on smooth or Lipschitz curves and surfaces [1, 2, 6]. Unfortunately, the size of the Delaunay triangulation grows too fast with the dimension of the ambient space for the approach to be still tractable in high-dimensional spaces [33].

Recently, significant steps were made towards a full understanding of the potential and limitations of the restricted Delaunay triangulation on smooth manifolds [14, 35]. In parallel, new sampling theories were developped, such as the critical point theory for distance functions [9], which provides sufficient conditions for the topology of a shape $X \subset \mathbb{R}^d$ to be captured by the offsets of a point cloud L lying at small Hausdorff distance. These advances lay the foundations of a new theoretical framework for the reconstruction of smooth submanifolds [11, 34], and more generally of large classes of compact subsets of $\mathbb{R}^d$ [9, 10, 12]. Combined with the introduction of more lightweight data structures, such as the *witness complex* [16], they have lead to new reconstruction techniques in arbitrary Euclidean spaces [4], whose outputs can be guaranteed under mild sampling conditions, and whose complexities can be orders of magnitude below the one of the classical Delaunay-based approach. For instance, on a data set with n points in $\mathbb{R}^d$, the algorithm of [4] runs in time $2^{O(d^2)}n^2$, whereas the size of the Delaunay triangulation can be of the order of $n^{\lceil \frac{d}{2} \rceil}$. Unfortunately, $2^{O(d^2)}n^2$ still remains too large for these new methods to be practical, even when the data points lie on or near a very low-dimensional submanifold.

A weaker yet similarly difficult version of the reconstruction paradigm is topological estimation, where the goal is not to exhibit a data structure that faithfully approximates the underlying shape X, but simply to infer the topological invariants of X from an input point cloud L. This problem has received a lot of attention in the recent years, and it finds applications in a number of areas of Science, such as sensor networks [19], statistical analysis [7], or dynamical systems [32, 36]. A classical approach to learning the homology of X consists in building a nested sequence of spaces $\mathcal{K}^0 \subseteq \mathcal{K}^1 \subseteq \cdots \subseteq \mathcal{K}^m$, and in studying the persistence of homology classes throughout this sequence. In particular, it has been independently proved in [12] and [15] that the persistent homology of the sequence defined by the α-offsets of a point cloud L coincides with the homology of the underlying shape X, under sampling conditions that are milder than the ones of [9]. Specifically, if the Hausdorff distance between L and X is less than ε, for some small enough ε, then, for all $\alpha \geq \varepsilon$, the canonical inclusion map $L^\alpha \hookrightarrow L^{\alpha+2\varepsilon}$ induces homomorphisms between homology groups, whose images are isomorphic to the homology groups of X. Combined with the structure theorem of [38], which states that the persistent homology of the sequence $\{L^\alpha\}_{\alpha \geq 0}$ is fully described by a finite set of intervals, called a *persistence barcode* or a *persistence diagram* — see Figure 1 (left), the above result means that the homology of X can be deduced from this barcode, simply by removing the intervals of length less than 2ε, which are therefore viewed as topological noise.

From an algorithmic point of view, the persistent homology of a nested sequence of simplicial complexes (called a *filtration*) can be efficiently computed using the persistence algorithm [22, 38]. Among the many filtrations that can be built on top of a point set L, the α-shape enables to reliably recover the homology of the underlying space X, since it is known to be a deformation retract of L^α [21]. However, this property is useless in high dimensions, since computing the α-shape requires

to build the full-dimensional Delaunay triangulation. It is therefore appealing to consider other filtrations that are easy to compute in arbitrary dimensions, such as the Rips and witness complex filtrations. Nevertheless, to the best of our knowledge, there currently exists no equivalent of the result of [12, 15] for such filtrations. In this paper, we produce such a result, not only for Rips and witness complexes, but more generally for any filtration that is intertwined with the Čech filtration. Recall that, for all $\alpha > 0$, the Čech complex $\mathcal{C}^\alpha(L)$ is the nerve of the union of the open balls of same radius α about the points of L, *i.e.* the nerve of L^α. It follows from the nerve theorem [31, Cor. 4G.3] that $\mathcal{C}^\alpha(L)$ and L^α are homotopy equivalent. However, despite the result of [12, 15], this is not sufficient to prove that the persistent homology of $\mathcal{C}^\alpha(L) \hookrightarrow \mathcal{C}^{\alpha+2\varepsilon}(L)$ coincides with the homology of X, mainly because it is not clear whether the homotopy equivalences $\mathcal{C}^\alpha(L) \to L^\alpha$ and $\mathcal{C}^{\alpha+2\varepsilon}(L) \to L^{\alpha+2\varepsilon}$ provided by the nerve theorem commute with the canonical inclusions $\mathcal{C}^\alpha(L) \hookrightarrow \mathcal{C}^{\alpha+2\varepsilon}(L)$ and $L^\alpha \hookrightarrow L^{\alpha+2\varepsilon}$. Using standard arguments of algebraic topology, we prove that there exist some homotopy equivalences that do commute with the canonical inclusions, at least at homology and homotopy levels. This enables us to extend the result of [12, 15] to the Čech filtration, and from there to the Rips and witness complex filtrations.

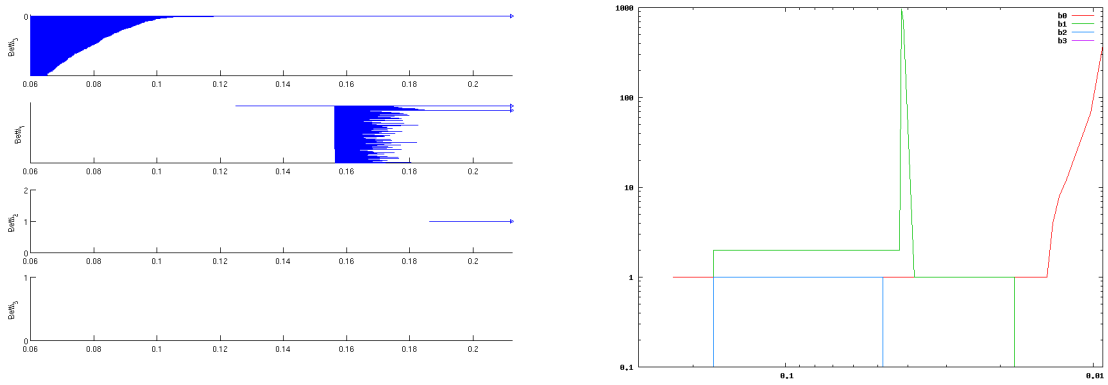

Figure 1: Results obtained from a set W of $10,000$ points sampled uniformly at random from a helical curve drawn on the 2d torus $(u, v) \mapsto \frac{1}{2}(\cos 2\pi u, \sin 2\pi u, \cos 2\pi v, \sin 2\pi v)$ in $\mathbb{R}^4$ — see [30]. Left: persistence barcode of the Rips filtration, built over a set of 900 carefully-chosen landmarks. Right: result of our algorithm, applied blindly to the input W. Both methods highlight the two underlying structures: curve and torus.

Another common concern in topological data analysis is the size of the vertex set on top of which a filtration is built. In many practical situations indeed, the point cloud W given as input samples the underlying shape very finely. In such situations, it makes sense to build the filtration on top of a small subset L of landmarks, to avoid a waste of computational resources. However, building a filtration on top of the sparse landmark set L instead of the dense point cloud W can result in a significant degradation in the quality of the persistence barcode. This is true in particular with the Čech and Rips filtrations, whose barcodes can have topological noise of amplitude depending directly on the density of L. The introduction of the witness complex filtration appeared as an elengant way of solving this issue [18]. The witness complex of L relative to W, or $\mathcal{C}_W(L)$ for short, can be viewed as a relaxed version of the Delaunay triangulation of L, in which the points of $W \setminus L$ are used to drive the construction of the complex [16]. Due to its special nature, which takes advantage of the points of $W \setminus L$, and due to its close relationship with the restricted Delaunay

triangulation, the witness complex filtration is likely to give persistence barcodes whose topological noise depends on the density of W rather than on the one of L, as conjectured in [18]. We prove in the paper that this statement is only true to some extent, namely: whenever the points of W are sufficiently densely sampled from some smooth submanifold of $\mathbb{R}^d$, the topological noise in the barcode can be arbitrarily small compared to the density of L. Nevertheless, it cannot depend solely on the density of W. This shows that the witness complex filtration does provide cleaner persistence barcodes than Čech or Rips filtrations, but maybe not as clean as expected.

Taking advantage of the above theoretical results on Rips and witness complexes, we propose a novel approach to reconstruction that stands somewhere in-between the classical reconstruction and topological estimation paradigms. Our algorithm is a variant of the method of [4, 30] that combines greedy refinement and topological persistence. Specifically, given an input point cloud W, the algorithm builds a subset L of landmarks iteratively, and in the meantime it maintains a nested pair of simplicial complexes (which happen to be Rips or witness complexes) and computes its persistent Betti numbers. The outcome of the algorithm is the sequence of nested pairs maintained throughout the process, or rather the diagram of evolution of their persistent Betti numbers. Using this diagram, a user or software agent can determine a relevant scale at which to process the data. It is then easy to rebuild the corresponding set of landmarks, as well as its nested pair of complexes. Note that our method does not completely solve the classical reconstruction problem, since it does not exhibit an embedded complex that is close to X topologically and geometrically. Nevertheless, it comes with theoretical guarantees, it is easily implementable, and above all it has reasonable complexity. Indeed, in the case where the input point cloud is sampled from a smooth submanifold X of $\mathbb{R}^d$, we show that the complexity of our algorithm is bounded by $c(m)n^5$, where $c(m)$ is a quantity depending solely on the intrinsic dimension m of X, while n is the size of the input. To the best of our knowledge, this is the first provably-good topological estimation or reconstruction method whose complexity scales up with the intrinsic dimension of the manifold. In the case where X is a more general compact set in $\mathbb{R}^d$, our complexity bound becomes $c(d)n^5$.

The paper is organized as follows: after introducing the Čech, Rips, and witness complex filtrations in Section 2, we prove our structural results in Sections 3 and 4, focusing on the general case of compact subsets of $\mathbb{R}^d$ in Section 3, and more specifically on the case of smooth submanifolds of $\mathbb{R}^d$ in Section 4. Finally, we present our algorithm and its analysis in Section 5.

2 Various complexes and their relationships

The definitions, results and proofs of this section hold in any arbitrary metric space. However, for the sake of consistency with the rest of the paper, we state them in the particular case of $\mathbb{R}^d$, endowed with the Euclidean norm $\|p\| = \sqrt{\sum_{i=1}^{d} p_i^2}$. As a consequence, our bounds are not the tightest possible for the Euclidean case, but they are for the general metric case. Using specific properties of Euclidean spaces, it is indeed possible to work out somewhat tighter bounds, but at the price of a loss of simplicity in the statements.

For any compact set $X \subset \mathbb{R}^d$, we call $\mathrm{diam}(X)$ the diameter of X, and $\mathrm{diam}_{\mathrm{CC}}(X)$ the *component-wise diameter* of X, defined by: $\mathrm{diam}_{\mathrm{CC}}(X) = \inf_i \mathrm{diam}(X_i)$, where the X_i are the path-connected components of X. Finally, given two compact sets X, Y in $\mathbb{R}^d$, we call $\mathrm{d}_{\mathcal{H}}(X, Y)$ their Hausdorff distance.

Čech complex. Given a finite set L of points of $\mathbb{R}^d$ and a positive number α, we call L^α the union of the open balls of radius α centered at the points of L: $L^\alpha = \bigcup_{x \in L} B(x, \alpha)$. This definition makes sense only for $\alpha > 0$, since for $\alpha = 0$ we get $L^\alpha = \emptyset$. We also denote by $\{L^\alpha\}$ the open cover of L^α formed by the open balls of radius α centered at the points of L. The Čech complex of L of parameter α, or $\mathcal{C}^\alpha(L)$ for short, is the *nerve* of this cover, *i.e.* it is the abstract simplicial complex whose vertex set is L, and such that, for all $k \in \mathbb{N}$ and all $x_0, \cdots, x_k \in L$, $[x_0, \cdots, x_k]$ is a k-simplex of $\mathcal{C}^\alpha(L)$ if and only if $B(x_0, \alpha) \cap \cdots \cap B(x_k, \alpha) \neq \emptyset$.

Rips complex. Given a finite set $L \subset \mathbb{R}^d$ and a positive number α, the Rips complex of L of parameter α, or $\mathcal{R}^\alpha(L)$ for short, is the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points of L which are pairwise within Euclidean distance α of one another. The Rips complex is closely related to the Čech complex, as stated in the following standard lemma, whose proof is recalled for completeness:

Lemma 2.1 *For all finite set $L \subset \mathbb{R}^d$ and all $\alpha > 0$, we have: $\mathcal{C}^{\frac{\alpha}{2}}(L) \subseteq \mathcal{R}^\alpha(L) \subseteq \mathcal{C}^\alpha(L)$.*

Proof. The proof is standard. Let $[x_0, \cdots, x_k]$ be an arbitrary k-simplex of $\mathcal{C}^{\frac{\alpha}{2}}(L)$. The Euclidean balls of same radius $\frac{\alpha}{2}$ centered at the x_i have a non-empty common intersection in $\mathbb{R}^d$. Let p be a point in the intersection. We then have: $\forall 0 \leq i, j \leq k$, $\|x_i - x_j\| \leq \|x_i - p\| + \|p - x_j\| \leq \alpha$. This implies that $[x_0, \cdots, x_k]$ is a simplex of $\mathcal{R}^\alpha(L)$, which proves the first inclusion of the lemma.

Let now $[x_0, \cdots, x_k]$ be an arbitrary k-simplex of $\mathcal{R}^\alpha(L)$. We have $\|x_0 - x_i\| \leq \alpha$ for all $i = 0, \cdots, k$. This means that x_0 belongs to all the Euclidean balls $B(x_i, \alpha)$, which therefore have a non-empty common intersection in $\mathbb{R}^d$. It follows that $[x_0, \cdots, x_k]$ is a simplex of $\mathcal{C}^\alpha(L)$, which proves the second inclusion of the lemma. □

Witness complex. Let L be a finite subset of $\mathbb{R}^d$, referred to as the landmark set, and let W be another (possibly infinite) subset of $\mathbb{R}^d$, identified as the witness set. Let also $\alpha \in [0, \infty)$.

- Given a point $w \in W$ and a k-simplex σ with vertices in L, w is an *α-witness* of σ (or, equivalently, *w α-witnesses σ*) if the vertices of σ lie within Euclidean distance $(\mathrm{d}_k(w) + \alpha)$ of w, where $\mathrm{d}_k(w)$ denotes the Euclidean distance between w and its $(k+1)$th nearest landmark in the Euclidean metric.
- The *α-witness complex* of L *relative to* W, or $\mathcal{C}_W^\alpha(L)$ for short, is the maximum abstract simplicial complex, with vertices in L, whose faces are α-witnessed by points of W.

When $\alpha = 0$, the α-witness complex coincides with the standard witness complex $\mathcal{C}_W(L)$, introduced in [17]. The α-witness complex is also closely related to the Čech complex, though the relationship is a bit more subtle than in the case of the Rips complex:

Lemma 2.2 *Let $L, W \subseteq \mathbb{R}^d$ be such that L is finite. If every point of L lies within Euclidean distance l of W, then for all $\alpha > l$ we have: $\mathcal{C}^{\frac{\alpha-l}{2}}(L) \subseteq \mathcal{C}_W^\alpha(L)$. In addition, if the Euclidean distance from any point of W to its second nearest neighbor in L is at most l', then for all $\alpha > 0$ we have: $\mathcal{C}_W^\alpha(L) \subseteq \mathcal{C}^{2(\alpha+l')}(L)$.*

Proof. Let $[x_0, \cdots, x_k]$ be a k-simplex of $\mathcal{C}^{\frac{\alpha-l}{2}}(L)$. This means that $\bigcap_{i=0}^k B(x_i, \frac{\alpha-l}{2}) \neq \emptyset$, and as a result, that $\|x_0 - x_i\| \leq \alpha - l$ for all $i = 0, \cdots, k$. Let w be a point of W closest to x_0 in the Euclidean metric. By the hypothesis of the lemma, we have $\|w - x_0\| \leq l$, therefore $x_0, \cdots, x_k$ lie within Euclidean distance α of w. Since the Euclidean distances from w to its nearest points of L

are non-negative, w is an α-witness of $[x_0, \cdots, x_k]$ and of all its faces. As a result, $[x_0, \cdots, x_k]$ is a simplex of $\mathcal{C}^{\alpha}_{W}(L)$.

Consider now a k-simplex $[x_0, \cdots, x_k]$ of $\mathcal{C}^{\alpha}_{W}(L)$. If $k = 0$, then the simplex is a vertex $[x_0]$, and therefore it belongs to $\mathcal{C}^{\alpha'}(L)$ for all $\alpha' > 0$. Assume now that $k \geq 1$. Edges $[x_0, x_1], \cdots, [x_0, x_k]$ belong also to $\mathcal{C}^{\alpha}_{W}(L)$, hence they are α-witnessed by points of W. Let $w_i \in W$ be an α-witness of $[x_0, x_i]$. Distances $\|w_i - x_0\|$ and $\|w_i - x_i\|$ are bounded from above by $\mathrm{d}_2(w_i) + \alpha$, where $\mathrm{d}_2(w_i)$ is the Euclidean distance from w_i to its second nearest point of L, which by assumption is at most l'. It follows that $\|x_0 - x_i\| \leq \|x_0 - w_i\| + \|w_i - x_i\| \leq 2\alpha + 2l'$. Since this is true for all $i = 0, \cdots, k$, we conclude that x_0 belongs to the intersection $\bigcap_{i=0}^{k} B(x_i, 2(\alpha + l'))$, which is therefore non-empty. As a result, $[x_0, \cdots, x_k]$ is a simplex of $\mathcal{C}^{2(\alpha+l')}(L)$. $\square$

Corollary 2.3 *Let X be a compact subset of $\mathbb{R}^d$, and let $L \subseteq W \subseteq \mathbb{R}^d$ be such that L is finite. Assume that $\mathrm{d}_{\mathcal{H}}(X, W) \leq \delta$ and that $\mathrm{d}_{\mathcal{H}}(W, L) \leq \varepsilon$, with $\varepsilon + \delta < \frac{1}{4}\,\mathrm{diam}_{\mathrm{CC}}(X)$. Then, for all $\alpha > \varepsilon$, we have: $\mathcal{C}^{\frac{\alpha-\varepsilon}{2}}(L) \subseteq \mathcal{C}^{\alpha}_{W}(L) \subseteq \mathcal{C}^{2\alpha+6(\varepsilon+\delta)}(L)$. In particular, if $\delta \leq \varepsilon < \frac{1}{8}\,\mathrm{diam}_{\mathrm{CC}}(X)$, then, for all $\alpha \geq 2\varepsilon$ we have: $\mathcal{C}^{\frac{\alpha}{4}}(L) \subseteq \mathcal{C}^{\alpha}_{W}(L) \subseteq \mathcal{C}^{8\alpha}(L)$.*

Proof. Since $\mathrm{d}_{\mathcal{H}}(W, L) \leq \varepsilon$, every point of L lies within Euclidean distance ε of W. As a result, the first inclusion of Lemma 2.2 holds with $l = \varepsilon$, that is: $\mathcal{C}^{\frac{\alpha-\varepsilon}{2}}(L) \subseteq \mathcal{C}^{\alpha}_{W}(L)$.

Now, for every point $w \in W$, there is a point $p \in L$ such that $\|w - p\| \leq \varepsilon$. Moreover, there is a point $x \in X$ such that $\|w - x\| \leq \delta$, since we assumed that $\mathrm{d}_{\mathcal{H}}(X, W) \leq \delta$. Let X_x be the path-connected component of X that contains x. Take an arbitrary value $\lambda \in \big(0, \frac{1}{2}\,\mathrm{diam}_{\mathrm{CC}}(X) - 2(\varepsilon + \delta)\big)$, and consider the open ball $B(w, 2(\varepsilon + \delta) + \lambda)$. This ball clearly intersects X_x, since it contains x. Furthermore, X_x is not contained entirely in the ball, since otherwise we would have: $\mathrm{diam}_{\mathrm{CC}}(X) \leq \mathrm{diam}(X_x) \leq 4(\varepsilon + \delta) + 2\lambda$, hereby contradicting the fact that $\lambda < \frac{1}{2}\,\mathrm{diam}_{\mathrm{CC}}(X) - 2(\varepsilon + \delta)$. Hence, there is a point $y \in X$ lying on the bounding sphere of $B(w, 2(\varepsilon + \delta) + \lambda)$. Let $q \in L$ be closest to y. We have $\|y - q\| \leq \varepsilon + \delta$, since our hypothesis implies that $\mathrm{d}_{\mathcal{H}}(X, L) \leq \mathrm{d}_{\mathcal{H}}(X, W) + \mathrm{d}_{\mathcal{H}}(W, L) \leq \delta + \varepsilon$. It follows then from the triangle inequality that $\|p - q\| \geq \|w - y\| - \|w - p\| - \|y - q\| \geq 2(\varepsilon + \delta) + \lambda - (\varepsilon + \delta) - (\varepsilon + \delta) = \lambda > 0$. Thus, q is different from p, and therefore the ball $B(w, 3(\varepsilon + \delta) + \lambda)$ contains at least two points of L. Since this is true for arbitrarily small values of λ, the Euclidean distance from w to its second nearest neighbor in L is at most $3(\varepsilon + \delta)$. It follows that the second inclusion of Lemma 2.2 holds with $l' = 3(\varepsilon + \delta)$, that is: $\mathcal{C}^{\alpha}_{W}(L) \subseteq \mathcal{C}^{2(\alpha+3(\varepsilon+\delta))}(L)$. $\square$

As mentioned at the head of the section, slightly tighter bounds can be worked out using specific properties of Euclidean spaces. For the case of the Rips complex, this was done by de Silva and Ghrist [19, 27]. Their approach can be combined with ours in the case of the witness complex.

3 Structural properties of filtrations over compact subsets of $\mathbb{R}^d$

Throughout this section, we use classical concepts of algebraic topology, such as homotopy equivalences, deformation retracts, or singular homology. We refer the reader to [31] for a good introduction to these concepts.

Given a compact set $X \subset \mathbb{R}^d$, we denote by d_X the *distance function* defined by $\mathrm{d}_X(x) = \inf\{\|x - y\| : y \in X\}$. Although d_X is not differentiable, it is possible to define a notion of critical point for distance functions and we denote by $\mathrm{wfs}(X)$ the *weak feature size* of X, defined as the smallest positive critical value of the distance function to X [10]. We do not explicitly use the

notion of critical value in the following, but only its relationship with the topology of the *offsets* $X^\alpha = \{x \in \mathbb{R}^d : \mathrm{d}_X(x) \leq \alpha\}$, stressed in the following result from [29]:

Lemma 3.1 (Isotopy Lemma) *If $0 < \alpha < \alpha'$ are such that there is no critical value of d_X in the closed interval $[\alpha, \alpha']$, then X^α and $X^{\alpha'}$ are homeomorphic (and even isotopic), and $X^{\alpha'}$ deformation retracts onto X^α.*

In particular the hypothesis of the lemma is satisfied when $0 < \alpha_1 < \alpha_2 < \mathrm{wfs}(X)$. In other words, all the offsets of X have the same topology in the interval $(0, \mathrm{wfs}(X))$.

3.1 Results on homology

We use singular homology with coefficients in an arbitrary field – omitted in our notations. In the following, we repeatedly make use of the following standard result of linear algebra:

Lemma 3.2 (Sandwich Lemma) *Consider the following sequence of homomorphisms between finite-dimensional vector spaces over a same field: $A \to B \to C \to D \to E \to F$. Assume that* $\mathrm{rank}\,(A \to F) = \mathrm{rank}\,(C \to D)$. *Then, this quantity also equals the rank of $B \to E$. In the same way, if $A \to B \to C \to E \to F$ is a sequence of homomorphisms such that* $\mathrm{rank}\,(A \to F) = \dim C$, *then* $\mathrm{rank}\,(B \to E) = \dim C$.

Proof. Observe that, for any sequence of homomorphisms $F \xrightarrow{f} G \xrightarrow{g} H$, we have $\mathrm{rank}\,(g \circ f) \leq \min\{\mathrm{rank}\, f, \mathrm{rank}\, g\}$. Applying this fact to maps $A \to F$, $B \to E$, and $C \to D$, which are nested in the sequence of the lemma, we get: $\mathrm{rank}\,(A \to F) \leq \mathrm{rank}\,(B \to E) \leq \mathrm{rank}\,(C \to D)$, which proves the first statement of the lemma. As for the second statement, it is obtained from the first one by letting $D = C$ and taking $C \to D$ to be the identity map. □

3.1.1 Čech filtration

Since the Čech complex is the nerve of a union of balls, its topological invariants can be read from the structure of its dual union. It turns out that unions of balls have been extensively studied in the past [9, 12, 15]. Our analysis relies particularly on the following result, which is an easy extension of Theorem 4.7 of [12]:

Lemma 3.3 *Let X be a compact set and L a finite set in $\mathbb{R}^d$, such that $\mathrm{d}_{\mathcal{H}}(X, L) < \varepsilon$ for some $\varepsilon < \frac{1}{4}\,\mathrm{wfs}(X)$. Then, for all $\alpha, \alpha' \in [\varepsilon, \mathrm{wfs}(X) - \varepsilon]$ such that $\alpha' - \alpha \geq 2\varepsilon$, and for all $\lambda \in (0, \mathrm{wfs}(X))$, we have: $\forall k \in \mathbb{N}$, $H_k(X^\lambda) \cong \mathrm{im}\; i_*$, where $i_* : H_k(L^\alpha) \to H_k(L^{\alpha'})$ is the homomorphism between homology groups induced by the canonical inclusion $i : L^\alpha \hookrightarrow L^{\alpha'}$. Given an arbitrary point $x_0 \in X$, the same conclusion holds for homotopy groups with base-point x_0.*

Proof. We can assume without loss of generality that $\varepsilon < \alpha < \alpha' - 2\varepsilon < \mathrm{wfs}(X) - 3\varepsilon$, since otherwise we can replace ε by any $\varepsilon' \in (d_H(X, L), \varepsilon)$. From the hypothesis we deduce the following sequence of inclusions:

$$X^{\alpha-\varepsilon} \hookrightarrow L^\alpha \hookrightarrow X^{\alpha+\varepsilon} \hookrightarrow L^{\alpha'} \hookrightarrow X^{\alpha'+\varepsilon} \tag{1}$$

By the Isotopy Lemma 3.1, for all $0 < \beta < \beta' < \mathrm{wfs}(X)$, the canonical inclusion $X^\beta \hookrightarrow X^{\beta'}$ is a homotopy equivalence. As a consequence, Eq. (1) induces a sequence of homomorphisms between

homology groups, such that all homomorphisms between homology groups of $X^{\alpha-\varepsilon}, X^{\alpha+\varepsilon}, X^{\alpha'+\varepsilon}$ are isomorphisms. It follows then from the Sandwich Lemma 3.2 that $i_* : H_k(L^\alpha) \to H_k(L^{\alpha'})$ has same rank as these isomorphisms. Now, this rank is equal to the dimension of $H_k(X^\lambda)$, since the X^β are homotopy equivalent to X^λ for all $0 < \beta < \mathrm{wfs}(X)$. It follows that $\mathrm{im}\, i_* \cong \dim H_k(X^\lambda)$, since our ring of coefficients is a field. The case of homotopy groups is a little trickier, since replacing homology groups by homotopy groups does not allow us to use the above rank argument. However, we can use the same proof as in Theorem 4.7 of [12] to conclude. □

Observe that Lemma 3.3 does not guarantee the retrieval of the homology of X. Instead, it deals with sufficiently small offsets of X, which are homotopy equivalent to one another but possibly not to X itself. In the special case where X is a smooth submanifold of $\mathbb{R}^d$ however, X^λ and X are homotopy equivalent, and therefore the theorem guarantees the retrieval of the homology of X. From an algorithmic point of view, the main drawback of Lemma 3.3 is that computing the homology of a union of balls or the image of the homomorphism i_* is usually awkward. As mentionned in [12, 15] this can be done by computing the persistence of the α-shape or λ-medial axis filtrations associated to L but there do not exist efficient algorithms to compute these filtrations in dimension more than 3. In the following we show that we can still reliably obtain the homology of X from easier to compute filtrations, namely the Rips and Witness complexes filtrations.

Consider now the Čech complex $\mathcal{C}^\alpha(L)$, for any value $\alpha > 0$. By definition, $\mathcal{C}^\alpha(L)$ is the nerve of the open cover $\{L^\alpha\}$ of L^α. Since the elements of $\{L^\alpha\}$ are open Euclidean balls, they are convex, and therefore their intersections are either empty or convex. It follows that $\{L^\alpha\}$ satisfies the hypotheses of the *nerve theorem*, which implies that $\mathcal{C}^\alpha(L)$ and L^α are homotopy equivalent – see *e.g.* [31, Corollary 4G.3]. We thus get the following diagram, where horizontal arrows are canonical inclusions, and vertical arrows are homotopy equivalences provided by the nerve theorem:

$$\begin{array}{ccc} L^\alpha & \hookrightarrow & L^{\alpha'} \\ \uparrow & & \uparrow \\ \mathcal{C}^\alpha(L) & \hookrightarrow & \mathcal{C}^{\alpha'}(L) \end{array} \tag{2}$$

Determining whether this diagram commutes is not straightforward. The following result, based on standard arguments of algebraic topology, shows that there exist homotopy equivalences between the union of balls and the Čech complex that make the above diagram commutative at homology and homotopy levels:

Lemma 3.4 *Let L be a finite set of points in $\mathbb{R}^d$ and let $0 < \alpha < \alpha'$. Then, there exist homotopy equivalences $\mathcal{C}^\alpha(L) \to L^\alpha$ and $\mathcal{C}^{\alpha'}(L) \to L^{\alpha'}$ such that, for all $k \in \mathbb{N}$, the diagram of Eq. (2) induces the following commutative diagrams:*

$$\begin{array}{ccc} H_k(L^\alpha) & \to & H_k(L^{\alpha'}) \\ \uparrow & & \uparrow \\ H_k(\mathcal{C}^\alpha(L)) & \to & H_k(\mathcal{C}^{\alpha'}(L)) \end{array} \quad \text{and} \quad \begin{array}{ccc} \pi_k(L^\alpha) & \to & \pi_k(L^{\alpha'}) \\ \uparrow & & \uparrow \\ \pi_k(\mathcal{C}^\alpha(L)) & \to & \pi_k(\mathcal{C}^{\alpha'}(L)) \end{array}$$

where vertical arrows are isomorphisms.

Proof. Our approach consists in a quick review of the proof of the nerve theorem provided in Section 4G of [31], and in a simple extension of the main arguments to our context.

As mentioned earlier, the open cover $\{L^\alpha\}$ satisfies the conditions of the nerve theorem, namely: for all points $x_0, \cdots, x_k \in L$, $\bigcap_{l=0}^k B(x_l, \alpha)$ is either empty, or convex and therefore contractible.

From this cover we construct a topological space ΔL^α as follows: let Δ^n denote the standard n-simplex, where $n = \#L - 1$. To each non-empty subset S of L we associate the face $[S]$ of Δ^n spanned by the elements of S, as well as the space $B_S(\alpha) = \bigcap_{s\in S} B(s,\alpha) \subseteq L^\alpha$. ΔL^α is then the subspace of $L^\alpha \times \Delta^n$ defined by:

$$\Delta L^\alpha = \bigcup_{\emptyset \neq S \subseteq L} B_S(\alpha) \times [S]$$

The space $\Delta L^{\alpha'}$ is built similarly. The product structures of ΔL^α and $\Delta L^{\alpha'}$ imply the existence of canonical projections $p_\alpha : \Delta L^\alpha \to L^\alpha$ and $p_{\alpha'} : \Delta L^{\alpha'} \to L^{\alpha'}$. These projections commute with the canonical inclusions $\Delta L^\alpha \hookrightarrow \Delta L^{\alpha'}$ and $L^\alpha \hookrightarrow L^{\alpha'}$, which implies that the following diagram:

$$\begin{array}{ccc} L^\alpha & \hookrightarrow & L^{\alpha'} \\ p_\alpha \uparrow & & \uparrow p_{\alpha'} \\ \Delta L^\alpha & \hookrightarrow & \Delta L^{\alpha'} \end{array} \tag{3}$$

induces commutative diagrams at homology and homotopy levels. Moreover, since $\{L^\alpha\}$ is an open cover of L^α, which is paracompact, p_α is a homotopy equivalence [31, Prop. 4G.2]. The same holds for $p_{\alpha'}$, and therefore p_α and $p_{\alpha'}$ induce isomorphisms at homology and homotopy levels.

We now show that, similarly, there exist homotopy equivalences $\Delta L^\alpha \to \mathcal{C}^\alpha(L)$ and $\Delta L^{\alpha'} \to \mathcal{C}^{\alpha'}(L)$ that commute with the canonical inclusions $\Delta L^\alpha \hookrightarrow \Delta L^{\alpha'}$ and $\mathcal{C}^\alpha(L) \hookrightarrow \mathcal{C}^{\alpha'}(L)$. This follows in fact from the proof of Corollary 4G.3 of [31]. Indeed, using the notion of *complex of spaces* introduced in [31, Section 4G], it can be shown that ΔL^α is the realization of the complex of spaces associated with the cover $\{L^\alpha\}$ — see the proof of [31, Prop. 4G.2]. Its base is the barycentric subdivision Γ^α of $\mathcal{C}^\alpha(L)$, where each vertex corresponds to a non-empty finite intersection $B_S(\alpha)$ for some $S \subseteq L$, and where each edge connecting two vertices $S \subset S'$ corresponds to the canonical inclusion $B_{S'}(\alpha) \hookrightarrow B_S(\alpha)$. In the same way, $\Delta L^{\alpha'}$ is the realization of a complex of spaces built over the barycentric subdivision $\Gamma^{\alpha'}$ of $\mathcal{C}^{\alpha'}(L)$. Now, since the non-empty finite intersections $B_S(\alpha)$ (resp. $B_S(\alpha')$) are contractible, the map $q_\alpha : \Delta L^\alpha \to \Gamma^\alpha$ (resp. $q_{\alpha'} : \Delta L^{\alpha'} \to \Gamma^{\alpha'}$) induced by sending each open set $B_S(\alpha)$ (resp. $B_S(\alpha')$) to a point is a homotopy equivalence [31, Prop. 4G.1 and Corol. 4G.3]. Furthermore, by construction, q_α is the restriction of $q_{\alpha'}$ to ΔL^α. Therefore,

$$\begin{array}{ccc} \Delta L^\alpha & \hookrightarrow & \Delta L^{\alpha'} \\ q_\alpha \downarrow & & \downarrow q_{\alpha'} \\ \Gamma^\alpha & \hookrightarrow & \Gamma^{\alpha'} \end{array} \tag{4}$$

is a commutative diagram where vertical arrows are homotopy equivalences. Now, it is well-known that Γ^α and $\Gamma^{\alpha'}$ are homeomorphic to $\mathcal{C}^\alpha(L)$ and $\mathcal{C}^{\alpha'}(L)$ respectively, and that the homeomorphisms commute with the inclusion. Combined with (3) and (4), this fact proves Lemma 3.4. □

Combining Lemmas 3.3 and 3.4, we obtain the following key result:

Theorem 3.5 *Let X be a compact set and L a finite set in $\mathbb{R}^d$, such that $\mathrm{d}_{\mathcal{H}}(X,L) < \varepsilon$ for some $\varepsilon < \frac{1}{4}\,\mathrm{wfs}(X)$. Then, for all $\alpha, \alpha' \in [\varepsilon, \mathrm{wfs}(X) - \varepsilon]$ such that $\alpha' - \alpha > 2\varepsilon$, and for all $\lambda \in (0, \mathrm{wfs}(X))$, we have: $\forall k \in \mathbb{N}$, $H_k(X^\lambda) \cong \mathrm{im}\ j_*$, where $j_* : H_k(\mathcal{C}^\alpha(L)) \to H_k(\mathcal{C}^{\alpha'}(L))$ is the homomorphism between homology groups induced by the canonical inclusion $j : \mathcal{C}^\alpha(L) \hookrightarrow \mathcal{C}^{\alpha'}(L)$. Given an arbitrary point $x_0 \in X$, the same result holds for homotopy groups with base-point x_0.*

Using the terminology of [38], this result means that the homology of X^λ can be deduced from the persistent homology of the filtration $\{\mathcal{C}^\alpha(L)\}_{\alpha\geq 0}$ by removing the cycles of persistence less than 2ε. Equivalently, the amplitude of the *topological noise* in the persistence barcode of $\{\mathcal{C}^\alpha(L)\}_{\alpha\geq 0}$ is bounded by 2ε, *i.e.* the intervals of length at least 2ε in the barcode give the homology of X^λ.

3.1.2 Filtrations intertwined with the Čech filtration

Using Lemma 2.1 and Theorem 3.5, we get the following guarantees on the Rips filtration:

Theorem 3.6 *Let $X \subset \mathbb{R}^d$ be a compact set, and $L \subset \mathbb{R}^d$ a finite set such that $\mathrm{d}_{\mathcal{H}}(X, L) < \varepsilon$ for some $\varepsilon < \frac{1}{9}\,\mathrm{wfs}(X)$. Then, for all $\alpha \in \left[2\varepsilon,\ \frac{1}{4}\left(\mathrm{wfs}(X) - \varepsilon\right)\right]$ and all $\lambda \in (0, \mathrm{wfs}(X))$, we have: $\forall k \in \mathbb{N}$, $H_k(X^\lambda) \cong \mathrm{im}\ j_*$, where $j_* : H_k(\mathcal{R}^\alpha(L)) \to H_k(\mathcal{R}^{4\alpha}(L))$ is the homomorphism between homology groups induced by the canonical inclusion $j : \mathcal{R}^\alpha(L) \hookrightarrow \mathcal{R}^{4\alpha}(L)$.*

Proof. From Lemma 2.1 we deduce the following sequence of inclusions:

$$\mathcal{C}^{\frac{\alpha}{2}}(L) \hookrightarrow \mathcal{R}^\alpha(L) \hookrightarrow \mathcal{C}^\alpha(L) \hookrightarrow \mathcal{C}^{2\alpha}(L) \hookrightarrow \mathcal{R}^{4\alpha}(L) \hookrightarrow \mathcal{C}^{4\alpha}(L) \tag{5}$$

Since $\alpha \geq 2\varepsilon$, Theorem 3.5 implies that Eq. (5) induces a sequence of homomorphisms between homology groups, such that $H_k(\mathcal{C}^{\frac{\alpha}{2}}(L)) \to H_k(\mathcal{C}^{4\alpha}(L))$ and $H_k(\mathcal{C}^\alpha(L)) \to H_k(\mathcal{C}^{2\alpha}(L))$ have ranks equal to $\dim H_k(X^\lambda)$. Therefore, by the Sandwich Lemma 3.2, rank j_* is also equal to $\dim H_k(X^\lambda)$. It follows that $\mathrm{im}\ j_* \cong \dim H_k(X^\lambda)$, since our ring of coefficients is a field. □

Similarly, Corollary 2.3 provides the following sequence of inclusions:

$$\mathcal{C}^{\frac{\alpha}{4}}(L) \hookrightarrow \mathcal{C}^\alpha_W(L) \hookrightarrow \mathcal{C}^{8\alpha}(L) \hookrightarrow \mathcal{C}^{9\alpha}(L) \hookrightarrow \mathcal{C}^{36\alpha}_W(L) \hookrightarrow \mathcal{C}^{288\alpha}(L),$$

from which follows a result similar to Theorem 3.6 on the witness complex, by the same proof:

Theorem 3.7 *Let X be a compact subset of $\mathbb{R}^d$, and let $L \subseteq W \subseteq \mathbb{R}^d$ be such that L is finite. Assume that $\mathrm{d}_{\mathcal{H}}(X, W) \leq \delta$ and that $\mathrm{d}_{\mathcal{H}}(W, L) \leq \varepsilon$, with $\delta \leq \varepsilon < \min\left\{\frac{1}{8}\,\mathrm{diam}_{\mathrm{CC}}(X),\ \frac{1}{1153}\,\mathrm{wfs}(X)\right\}$. Then, for all $\alpha \in \left[4\varepsilon,\ \frac{1}{288}\left(\mathrm{wfs}(X) - \varepsilon\right)\right]$ and all $\lambda \in (0, \mathrm{wfs}(X))$, we have: $\forall k \in \mathbb{N}$, $H_k(X^\lambda) \cong \mathrm{im}\ j_*$, where $j_* : H_k(\mathcal{C}^\alpha_W(L)) \to H_k(\mathcal{C}^{36\alpha}_W(L))$ is the homomorphism between homology groups induced by the canonical inclusion $j : \mathcal{C}^\alpha_W(L) \hookrightarrow \mathcal{C}^{36\alpha}_W(L)$.*

More generally, the above arguments show that the homology of X^λ can be recovered from the persistence barcode of any filtration $\{F_\alpha\}_{\alpha\geq 0}$ that is intertwined with the Čech filtration in the sense of Lemmas 2.1 and 2.2. Note however that Theorems 3.6 and 3.7 suggest a different behavior of the barcode in this case, since its topological noise might scale up with α (specifically, it might be up to linear in α), whereas it is uniformly bounded by a constant in the case of the Čech filtration. This difference of behavior is easily explained by the way $\{F_\alpha\}_{\alpha\geq 0}$ is intertwined with the Čech filtration. A trick to get a uniformly-bounded noise is to represent the barcode of $\{F_\alpha\}_{\alpha\geq 0}$ on a logarithmic scale, that is, with $\log_2 \alpha$ instead of α in abcissa.

3.2 Results on homotopy

The results on homology obtained in Section 3.1 follow from simple algebraic arguments. Using a more geometric approach, we can get similar results on homotopy. From now on, $x_0 \in X$ is a fixed point and all the homotopy groups $\pi_k(X) = \pi_k(X, x_0)$ are assumed to be with base-point x_0. Theorems 3.6 and 3.7 can be extended to homotopy in the following way:

Theorem 3.8 *Under the same hypotheses as in Theorem 3.6, we have:* $\forall k \in \mathbb{N}$, $\pi_k(X^\lambda) \cong \text{im } j_*$, *where* $j_* : \pi_k(\mathcal{R}^\alpha(L)) \to \pi_k(\mathcal{R}^{4\alpha}(L))$ *is the homomorphism between homotopy groups induced by the canonical inclusion* $j : \mathcal{R}^\alpha(L) \hookrightarrow \mathcal{R}^{4\alpha}(L)$.

Theorem 3.9 *Under the same hypotheses as in Theorem 3.7, we have:* $\forall k \in \mathbb{N}$, $\pi_k(X^\lambda) \cong \text{im } j_*$, *where* $j_* : \pi_k(\mathcal{C}^\alpha_W(L)) \to \pi_k(\mathcal{C}^{36\alpha}_W(L))$ *is the homomorphism between homotopy groups induced by the canonical inclusion* $j : \mathcal{C}^\alpha_W(L) \hookrightarrow \mathcal{C}^{36\alpha}_W(L)$.

The proofs of these two results being mostly identical, we focus exclusively on the Rips complex. We will use the following lemma, which is an immediate generalization of Proposition 4.1 of [12]:

Lemma 3.10 *Let* X *be a compact set and* L *a finite set in* $\mathbb{R}^d$, *such that* $\mathrm{d}_\mathcal{H}(X, L) < \varepsilon$ *for some* $\varepsilon < \frac{1}{4}\,\text{wfs}(X)$. *Let* $\alpha, \alpha' \in [\varepsilon, \text{wfs}(X) - \varepsilon]$ *be such that* $\alpha' - \alpha \geq 2\varepsilon$. *Given* $k \in \mathbb{N}$, *two* k-loops $\sigma_1, \sigma_2 : \mathbb{S}^k \to (L^\alpha, x_0)$ *in* L^α *are homotopic in* $X^{\alpha'+\varepsilon}$ *if and only if they are homotopic in* $L^{\alpha'}$.

Proof of Theorem 3.8. As mentionned at the begining of the proof of Lemma 3.3, we can assume without loss of generality that $2\varepsilon < \alpha < \frac{1}{4}(\text{wfs}(X) - \varepsilon)$. Consider the following sequence of inclusions:

$$\mathcal{C}^{\frac{\alpha}{2}}(L) \subset \mathcal{R}^\alpha(L) \subset \mathcal{C}^\alpha(L) \subset \mathcal{C}^{2\alpha}(L) \subset \mathcal{R}^{4\alpha}(L) \subset \mathcal{C}^{4\alpha}(L)$$

We use the homotopy equivalences $h_\beta : L^\beta \to \mathcal{C}^\beta(L)$ provided by Lemma 3.4 for all values $\beta > 0$, which commute with inclusions at homotopy level. Note that, for any element σ of $\pi_k(\mathcal{C}^\beta(L))$, there exists a k-loop in L^β that is mapped through h_β to a k-loop representing the homotopy class σ. In the following, we denote by σ_g such a k-loop. Let E, F and G be the images of $\pi_k(\mathcal{C}^{\frac{\alpha}{2}}(L))$ in $\pi_k(\mathcal{C}^\alpha(L))$, $\pi_k(\mathcal{C}^{2\alpha}(L))$ and $\pi_k(\mathcal{C}^{4\alpha}(L))$ respectively, through the homomorphisms induced by inclusion. We thus have a sequence of surjective homomorphisms:

$$\pi_k(\mathcal{C}^{\frac{\alpha}{2}}(L)) \to E \to F \to G$$

Note that, by Theorem 3.5, F and G are isomorphic to $\pi_k(X^\lambda)$. Let $\sigma \in F$ be a homotopy class. Since F is the image of $\pi_k(\mathcal{C}^{\frac{\alpha}{2}}(L))$, we can assume without loss of generality that $\sigma_g \subset L^{\frac{\alpha}{2}}$. Assume that the image of σ in G is equal to 0. This means that σ_g is null-homotopic in $L^{4\alpha}$ and, since $L^{4\alpha} \subset X^{4\alpha+\varepsilon}$, σ_g is also null-homotopic in $X^{4\alpha+\varepsilon}$. But $\sigma_g \subset L^{\frac{\alpha}{2}} \subset X^{\frac{\alpha}{2}+\varepsilon}$, and $X^{2\alpha+\varepsilon}$ deformation retracts onto $X^{\frac{\alpha}{2}+\varepsilon}$, by the Isotopy Lemma 3.1. As a consequence, σ_g is null-homotopic in $X^{\frac{\alpha}{2}+\varepsilon}$, which is contained in $L^{2\alpha}$ since $\frac{\alpha}{2} + 2\varepsilon < 2\alpha$. Hence, σ_g is null-homotopic in $L^{2\alpha}$, namely: $\sigma = 0$ in F. So, the homomorphism $F \to G$ is injective, and thus it is an isomorphism. As a consequence, $F \to \pi_k(\mathcal{R}^{4\alpha}(L))$ is injective, and it is now sufficient to prove that the image of $\phi_* : \pi_k(\mathcal{R}^\alpha(L)) \to \pi_k(\mathcal{C}^{2\alpha}(L))$ induced by the inclusion is equal to F.

Obviously, F is contained in the image of ϕ_*. Now, let $\sigma \in \pi_k(\mathcal{R}^\alpha(L))$ and let $\phi_*(\sigma)_g$ be a k-loop in $L^{2\alpha}$ that is mapped through $h_{2\alpha}$ to a k-loop representing the homotopy class $\phi_*(\sigma)$. Since $\phi_*(\sigma)$ is in the image of ϕ_*, and since $\mathcal{R}^\alpha(L) \subset \mathcal{C}^\alpha(L)$, we can assume that $\phi_*(\sigma)_g$ is contained in L^α. Let $\tilde{\sigma}_g$ be the image of $\phi_*(\sigma)_g$ through a deformation retraction of $X^{2\alpha+\varepsilon}$ onto X^{α_0}, where $0 < \alpha_0 < \frac{\alpha}{2}$ is such that $\frac{\alpha}{2} - \alpha_0 > \varepsilon$. Obviously, $\tilde{\sigma}_g$ and $\phi_*(\sigma)_g$ are homotopic in $X^{2\alpha+\varepsilon}$. It follows then from Lemma 3.10 that $\tilde{\sigma}_g$ and $\phi_*(\sigma)_g$ are homotopic in $L^{2\alpha}$. And since $\tilde{\sigma}_g$ is contained in $X^{\alpha_0} \subset L^{\frac{\alpha}{2}}$, the equivalence class of $h_{\frac{\alpha}{2}}(\tilde{\sigma}_g)$ in $\pi_k(\mathcal{C}^{\frac{\alpha}{2}}(L))$ is mapped to $\phi_*(\sigma) \in \pi_k(\mathcal{C}^{2\alpha}(L))$ through the homomorphism induced by $\mathcal{C}^{\frac{\alpha}{2}}(L) \hookrightarrow \mathcal{C}^{2\alpha}(L)$, which commutes with the homotopy equivalences. As a result, $\phi_*(\sigma)$ belongs to F, which is therefore equal to im ϕ_*. □

4 The case of smooth submanifolds of $\mathbb{R}^d$

In this section, we consider the case of submanifolds X of $\mathbb{R}^d$ that have positive *reach*. Recall that the reach of X, or $\mathrm{rch}(X)$ for short, is the minimum distance between the points of X and the points of its medial axis [1]. A point cloud $L \subset X$ is an *ε-sample* of X if every point of X lies within distance ε of L. In addition, L is *ε-sparse* if its points lie at least ε away from one another.

Our main result is a first attempt at quantifying a conjecture of Carlsson and de Silva [18], according to which the witness complex filtration should have *cleaner* persistence barcodes than the Čech and Rips filtrations, at least on smooth submanifolds of $\mathbb{R}^d$. By *cleaner* is meant that the amplitude of the topological noise in the barcodes should be smaller, and also that the long intervals should appear earlier. We prove this latter statement correct, at least to some extent:

Theorem 4.1 *There exist a constant $\varrho > 0$ and a continuous, non-decreasing map $\bar{\omega} : [0, \varrho) \to [0, \frac{1}{2})$, such that, for any submanifold X of $\mathbb{R}^d$, for all ε, δ satisfying $0 < \delta \leq \varepsilon < \varrho\, \mathrm{rch}(X)$, for any δ-sample W of X and any ε-sparse ε-sample L of W, $\mathcal{C}^{\alpha}_{W}(L)$ contains a subcomplex $\mathcal{D}$ homeomorphic to X and such that the canonical inclusion $\mathcal{D} \hookrightarrow \mathcal{C}^{\alpha}_{W}(L)$ induces an injective homomorhism between homology groups, provided that α satisfies: $\frac{8}{3}(\delta + \bar{\omega}(\frac{\varepsilon}{\mathrm{rch}(X)})^2\varepsilon) \leq \alpha < \frac{1}{2}\,\mathrm{rch}(X) - (3 + \frac{\sqrt{2}}{2})(\varepsilon + \delta)$.*

This theorem guarantees that, for values of α ranging from $O(\delta + \bar{\omega}(\frac{\varepsilon}{\mathrm{rch}(X)})^2\varepsilon)$ to $\Omega(\mathrm{rch}(X))$, the topology of X is captured by a subcomplex $\mathcal{D}$ that injects itself suitably in $\mathcal{C}^{\alpha}_{W}(L)$. As a result, long intervals showing the homology of X appear around $\alpha = O(\delta + \bar{\omega}(\frac{\varepsilon}{\mathrm{rch}(X)})^2\varepsilon)$ in the persistence barcode of the witness complex filtration. This can be much sooner than the time $\alpha = 2\varepsilon$ prescribed by Theorem 3.7, since $\bar{\omega}(\frac{\varepsilon}{\mathrm{rch}(X)})$ can be arbitrarily small. Specifically, the denser the landmark set L, the smaller the ratio $\frac{\varepsilon}{\mathrm{rch}(X)}$, and therefore the smaller $\delta + \bar{\omega}(\frac{\varepsilon}{\mathrm{rch}(X)})^2\varepsilon$ compared to 2ε. We have reasons to believe that this upper bound on the appearance time of long bars is tight. In particular, the bound cannot depend solely on δ, since otherwise, in the limit case where $\delta = 0$, we would get that the homology groups of X can be injected into the ones of the standard witness complex $\mathcal{C}_W(L)$, which is known to be false [30, 35]. The same argument implies that the amplitude of the topological noise in the barcode cannot depend solely on δ either. However, whether the upper bound $O(\varepsilon)$ on the amplitude of the noise can be improved or not is still an open question.

Our proof of Theorem 4.1 generalizes and argument used in [26] for the planar case, which stresses the close relationship that exists between the α-witness complex and the so-called *weighted restricted Delaunay triangulation* $\mathcal{D}^X_\omega(L)$. Given a submanifold X of $\mathbb{R}^d$, a finite landmark set $L \subset \mathbb{R}^d$, and an assignment of non-negative weights to the landmarks, specified through a map $\omega : L \to [0, \infty)$, $\mathcal{D}^X_\omega(L)$ is the nerve of the restriction to X of the *power diagram*[1] of the weighted set L. Under the hypotheses of the theorem, we show that $\mathcal{C}^{\alpha}_{W}(L)$ contains $\mathcal{D}^X_\omega(L)$, which, by a result of Cheng *et al.* [14] (see Theorem 4.2 below), is homeomorphic to X. The main point of the proof is then to show that $\mathcal{D}^X_\omega(L)$ injects itself *nicely* into $\mathcal{C}^{\alpha}_{W}(L)$.

The rest of the section is devoted to the proof of Theorem 4.1. After introducing the weighted restricted Delaunay triangulation in Section 4.1 and stressing its relationship with the α-witness complex in Section 4.2, we detail the proof of Theorem 4.1 in Section 4.3.

[1] More on power diagrams and on restricted triangulations can be found in [3] and [23] respectively.

4.1 The weighted restricted Delaunay triangulation

Given a finite point set $L \subset \mathbb{R}^d$, an *assignment of weights over* L is a non-negative real-valued function $\omega : L \to [0, \infty)$. The quantity $\max_{u \in L, v \in L \setminus \{u\}} \frac{\omega(u)}{\|u-v\|}$ is called the *relative amplitude* of ω. Given $p \in \mathbb{R}^d$, the *weighted distance* from p to some weighted point $v \in L$ is $\|p - v\|^2 - \omega(v)^2$. This is actually not a metric, since it is not symmetric. Given a finite point set L and an assignment of weights ω over L, we denote by $\mathcal{V}_\omega(L)$ the power diagram of the weighted set L, and by $\mathcal{D}_\omega(L)$ its nerve, also known as the weighted Delaunay triangulation. If the relative amplitude of ω is at most $\frac{1}{2}$, then the points of L have non-empty cells in $\mathcal{V}_\omega(L)$, and in fact each point of L belongs to its own cell [13]. For any simplex σ of $\mathcal{D}_\omega(L)$, $\mathrm{V}_\omega(\sigma)$ denotes the face of $\mathcal{V}_\omega(L)$ dual to σ.

Given a subset X of $\mathbb{R}^d$, we call $\mathcal{V}_\omega^X(L)$ the restriction of $\mathcal{V}_\omega(L)$ to X, and we denote by $\mathcal{D}_\omega^X(L)$ its nerve, also known as the weighted Delaunay triangulation of L restricted to X. Observe that $\mathcal{D}_\omega^X(L)$ is a subcomplex of $\mathcal{D}_\omega(L)$. In the special case where all the weights are equal, $\mathcal{V}_\omega(L)$ and $\mathcal{D}_\omega(L)$ coincide with their standard Euclidean versions, $\mathcal{V}(L)$ and $\mathcal{D}(L)$. Similarly, $\mathrm{V}_\omega(\sigma)$ becomes $\mathrm{V}(\sigma)$, and $\mathcal{V}_\omega^X(L)$ and $\mathcal{D}_\omega^X(L)$ become respectively $\mathcal{V}^X(L)$ and $\mathcal{D}^X(L)$.

Theorem 4.2 (Lemmas 13, 14, 18 of [14], see also Theorem 2.5 of [4]) *There exist*[2] *a constant* $\varrho > 0$ *and a non-decreasing continuous map* $\bar{\omega} : [0, \varrho) \to [0, \frac{1}{2})$*, such that, for any manifold* X *and any* ε*-sparse* 2ε*-sample* L *of* X*, with* $\varepsilon < \varrho \operatorname{rch}(X)$*, there is an assignment of weights* ω *of relative amplitude at most* $\bar{\omega}\left(\frac{\varepsilon}{\operatorname{rch}(X)}\right)$ *such that* $\mathcal{D}_\omega^X(L)$ *is homeomorphic to* X*.*

This theorem guarantees that the topology of X is captured by $\mathcal{D}_\omega^X(L)$ provided that the landmarks are sufficiently densely sampled on X, and that they are assigned suitable weights. Observe that the denser the landmark set, the smaller the weights are required to be, as specified by the map $\bar{\omega}$. In the particular case where X is a curve or a surface, $\bar{\omega}$ can be taken to be the constant zero map, since $\mathcal{D}^X(L)$ is homeomorphic to X [1, 2]. On higher-dimensional manifolds though, positive weights are required, since $\mathcal{D}^X(L)$ may fail to capture the topological invariants of X [35].

The proof of the theorem given in [14] shows that $\mathcal{V}_\omega^X(L)$ satisfies the so-called *closed ball property*, which states that every face of the weighted Voronoi diagram $\mathcal{V}_\omega(L)$ intersects the manifold X along a topological ball of proper dimension, if at all. Under this condition, there exists a homeomorphism h_0 between the nerve $\mathcal{D}_\omega^X(L)$ and X, as proved by Edelsbrunner and Shah [23]. Furthermore, h_0 sends every simplex of $\mathcal{D}_\omega^X(L)$ to a subset of the union of the restricted Voronoi cells of its vertices, that is: $\forall \sigma \in \mathcal{D}_\omega^X(L)$, $h_0(\sigma) \subseteq \bigcup_{v \text{ vertex of } \sigma} \mathrm{V}_\omega(v) \cap X$. This fact will be instrumental in the proof of Theorem 4.1.

4.2 Relationship between $\mathcal{D}_\omega^X(L)$ and $\mathcal{C}_W^\alpha(L)$

As mentioned in introduction, the use of the witness complex filtration for topological data analysis is motivated by its close relationship with the weighted restricted Delaunay triangulation:

Lemma 4.3 *Let* X *be a compact subset of* $\mathbb{R}^d$*,* $W \subseteq X$ *a* δ*-sample of* X*, and* $L \subseteq W$ *an* ε*-sparse* ε*-sample of* W*. Then, for all assignment of weights* ω *of relative amplitude* $\bar{\omega} \leq \frac{1}{2}$*,* $\mathcal{D}_\omega^X(L)$ *is included in* $\mathcal{C}_W^\alpha(L)$ *whenever* $\alpha \geq \frac{2}{1-\bar{\omega}^2}\left(\delta + \bar{\omega}^2 \varepsilon\right)$*.*

This result implies in particular that $\mathcal{D}^X(L)$ is included in $\mathcal{C}_W^\alpha(L)$ whenever $\alpha \geq 2\delta$, since $\mathcal{D}^X(L)$ is nothing but $\mathcal{D}_\omega^X(L)$ for an assignment of weights of relative amplitude zero.

[2] Note that ϱ and $\bar{\omega}$ are the same as in Theorem 4.1. In fact, these quantities come from Theorem 4.2.

Proof. Let σ be a simplex of $\mathcal{D}_\omega^X(L)$. If σ is a vertex, then it clearly belongs to $\mathcal{C}_W^\alpha(L)$ for all $\alpha \geq 0$, since $L \subseteq W$. Assume now that σ has positive dimension, and consider a point $c \in \mathrm{V}_\omega(\sigma) \cap X$. For any vertex v of σ and any point p of L (possibly equal to v), we have: $\|v-c\|^2 - \omega(v)^2 \leq \|p-c\|^2 - \omega(p)^2$, which yields: $\|v-c\|^2 \leq \|p-c\|^2 + \omega(v)^2 - \omega(p)^2$. Now, $\omega(p)^2$ is non-negative, while $\omega(v)^2$ is at most $\bar{\omega}^2\|v-p\|^2$, which gives: $\|v-c\|^2 \leq \|p-c\|^2 + \bar{\omega}^2\|v-p\|^2$. Replacing $\|v-p\|$ by $\|v-c\|+\|p-c\|$, we get a semi-algebraic expression of degree 2 in $\|v-c\|$, namely: $(1-\bar{\omega}^2)\|v-c\|^2 - 2\bar{\omega}^2\|p-c\|\|v-c\| - (1+\bar{\omega}^2)\|p-c\|^2 \leq 0$. It follows that $\|v-c\| \leq \frac{1+\bar{\omega}^2}{1-\bar{\omega}^2}\,\|p-c\|$. Let now w be a point of W closest to c in the Euclidean metric. Using the triangle inequality and the fact that $\|w-c\| \leq \delta$, we get: $\|v-w\| \leq \|v-c\| + \|w-c\| \leq \frac{1+\bar{\omega}^2}{1-\bar{\omega}^2}\,\|p-c\| + \delta$. This holds for any point $p \in L$, and in particular for the nearest neighbor p_w of w in L. Therefore, we have $\|v-w\| \leq \frac{1+\bar{\omega}^2}{1-\bar{\omega}^2}\,\|p_w-c\| + \delta$, which is at most $\frac{1+\bar{\omega}^2}{1-\bar{\omega}^2}\,(\|p_w-w\|+\delta) + \delta \leq \|p_w-w\| + \frac{2}{1-\bar{\omega}^2}\left(\delta + \bar{\omega}^2\varepsilon\right)$ because $\|w-c\| \leq \delta$ and $\|w-p_w\| \leq \varepsilon$. Since this inequality holds for any vertex v of σ, and since the Euclidean distances from w to all the landmarks are at least $\|p_w-w\|$, w is an α-witness of σ and of all its faces as soon as $\alpha \geq \frac{2}{1-\bar{\omega}^2}\left(\delta + \bar{\omega}^2\varepsilon\right)$. Since this holds for every simplex σ of $\mathcal{D}_\omega^X(L)$, the lemma follows. □

4.3 Proof of Theorem 4.1

The proof is mostly algebraic, but it relies on two technical results. The first one is Dugundji's extension theorem [20], which states that, given an abstract simplex σ and a continuous map $f : \partial\sigma \to \mathbb{R}^d$, f can be extended to a continuous map $f : \sigma \to \mathbb{R}^d$ such that $f(\sigma)$ is included in the Euclidean convex hull of $f(\partial\sigma)$, noted $\mathrm{CH}(f(\partial\sigma))$. This convexity property of f is used in the proof of the second technical result, stated as Lemma 4.5 and proved at the end of the section.

Proof of Theorem 4.1. Since $\delta \leq \varepsilon$, L is an ε-sparse 2ε-sample of X, with $\varepsilon < \varrho\ \mathrm{rch}(X)$. Therefore, by Theorem 4.2, there exists an assignment of weights ω over L, of relative amplitude at most $\bar{\omega}\left(\frac{\varepsilon}{\mathrm{rch}(X)}\right)$, such that $\mathcal{D}_\omega^X(L)$ is homeomorphic to X. Taking $\mathcal{D} = \mathcal{D}_\omega^X(L)$, we then have: $\forall k \in \mathbb{N},\ H_k(X) \cong H_k(\mathcal{D})$. Moreover, by Lemma 4.3, we know that $\mathcal{D} = \mathcal{D}_\omega^X(L)$ is included in $\mathcal{C}_W^\alpha(L)$, since $\alpha \geq \frac{8}{3}\left(\bar{\omega}\left(\frac{\varepsilon}{\mathrm{rch}(X)}\right)^2\varepsilon + \delta\right) \geq \frac{2}{1-\bar{\omega}\left(\frac{\varepsilon}{\mathrm{rch}(X)}\right)^2}\left(\bar{\omega}\left(\frac{\varepsilon}{\mathrm{rch}(X)}\right)^2\varepsilon + \delta\right)$. There remains to show that the inclusion map $j : \mathcal{D}_\omega^X(L) \hookrightarrow \mathcal{C}_W^\alpha(L)$ induces injective homomorphisms j_* between the homology groups of $\mathcal{D}_\omega^X(L)$ and $\mathcal{C}_W^\alpha(L)$, which will conclude the proof of the theorem.

Our approach to showing the injectivity of j_* consists in building a continuous map[3] $h : \mathcal{C}_W^\alpha(L) \to \mathcal{D}_\omega^X(L)$ such that $h \circ j$ is homotopic to the identity in $\mathcal{D}_\omega^X(L)$. This implies that $h_* \circ j_* : H_k(\mathcal{D}_\omega^X(L)) \to H_k(\mathcal{D}_\omega^X(L))$ is an isomorphism (in fact, it is the identity map), and thus that j_* is injective.

We begin our construction with the homeomorphism $h_0 : \mathcal{D}_\omega^X(L) \to X$ provided by the theorem of Edelsbrunner and Shah [23]. Taking h_0 as a map $\mathcal{D}_\omega^X(L) \to \mathbb{R}^d$, we extend it to a continuous map $\tilde{h}_0 : \mathcal{C}_W^\alpha(L) \to \mathbb{R}^d$ by the following iterative process: while there exists a simplex $\sigma \in \mathcal{C}_W^\alpha(L)$ such that $\tilde{h}_0$ is defined over the boundary of σ but not over its interior, apply Dugundji's extension theorem, which extends $\tilde{h}_0$ to the entire simplex σ.

Lemma 4.4 *The above iterative process extends h_0 to a map $\tilde{h}_0 : \mathcal{C}_W^\alpha(L) \to \mathbb{R}^d$.*

[3]Note that this map does not need to be simplicial, since we are using singular homology.

Proof. We only need to prove that the process visits every simplex of $\mathcal{C}_W^\alpha(L)$. Assume for a contradiction that the process terminates while there still remain some unvisited simplices of $\mathcal{C}_W^\alpha(L)$. Consider one such simplex σ of minimal dimension. Either σ is a vertex, or there is at least one proper face of σ that has not yet been visited – since otherwise the process could visit σ. In the former case, σ is a point of L, and as such it is a vertex[4] of $\mathcal{D}_\omega^X(L)$, which means that h_0 is already defined over σ (contradiction). In the latter case, we get a contradiction with the fact that σ is of minimal dimension. □

Now that we have built a map $\tilde{h}_0 : \mathcal{C}_W^\alpha(L) \to \mathbb{R}^d$, our next step is to turn it into a map $\mathcal{C}_W^\alpha(L) \to X$. To do so, we compose it with the projection p_X that maps every point of $\mathbb{R}^d$ to its nearest neighbor on X, if the latter is unique. This projection is known to be well-defined and continuous over $\mathbb{R}^d \setminus \mathrm{M}$, where M denotes the medial axis of X [24].

Lemma 4.5 *Let $X, W, L, \delta, \varepsilon$ satisfy the hypotheses of Theorem 4.1. Then, $\tilde{h}_0(\mathcal{C}_W^\alpha(L)) \cap \mathrm{M} = \emptyset$ as long as $\alpha < \frac{1}{2}\,\mathrm{rch}(X) - \left(3 + \frac{\sqrt{2}}{2}\right)(\varepsilon + \delta)$.*

Since by Lemma 4.5 we have $\tilde{h}_0(\mathcal{C}_W^\alpha(L)) \cap \mathrm{M} = \emptyset$, the map $p_X \circ \tilde{h}_0 : \mathcal{C}_W^\alpha(L) \to X$ is well-defined and continuous. Our final step is to compose it with h_0^{-1}, to get a continuous map $h = h_0^{-1} \circ p_X \circ \tilde{h}_0 : \mathcal{C}_W^\alpha(L) \to \mathcal{D}_\omega^X(L)$. The restriction of h to $\mathcal{D}_\omega^X(L)$ is simply $h_0^{-1} \circ p_X \circ h_0$, which coincides with $h_0^{-1} \circ h_0 = \mathrm{id}$ since $h_0(\mathcal{D}_\omega^X(L)) = X$. It follows that $h \circ j$ is homotopic to the identity in $\mathcal{D}_\omega^X(L)$ (in fact, it is the identity), and therefore that the induced map $h_* \circ j_*$ is the identity. This implies that $j_* : H_k(\mathcal{D}_\omega^X(L)) \to H_k(\mathcal{C}_W^\alpha(L))$ is injective, which concludes the proof of Theorem 4.1. □

We end the section by providing the proof of Lemma 4.5:

Proof of Lemma 4.5. First, we claim that the image through $\tilde{h}_0$ of any simplex of $\mathcal{C}_W^\alpha(L)$ is included in the Euclidean convex hull of the restricted Voronoi cells of its simplices, that is: $\forall \sigma \in \mathcal{C}_W^\alpha(L)$, $\tilde{h}_0(\sigma) \subseteq \mathrm{CH}\left(\bigcup_{v \text{ vertex of } \sigma} \mathrm{V}_\omega(v) \cap X\right)$. This is clearly true if σ belongs to $\mathcal{D}_\omega^X(L)$, since in this case we have $\tilde{h}_0(\sigma) = h_0(\sigma) \subseteq \bigcup_{v \text{ vertex of } \sigma} \mathrm{V}_\omega(v) \cap X$, as mentioned after Theorem 4.2. Now, if the property holds for all the proper faces of a simplex $\sigma \in \mathcal{C}_W^\alpha(L)$, then by induction it also holds for the simplex itself. Indeed, for each proper face $\tau \subset \sigma$, we have $\tilde{h}_0(\tau) \subseteq \mathrm{CH}\left(\bigcup_{v \text{ vertex of } \tau} \mathrm{V}_\omega(v) \cap X\right) \subseteq \mathrm{CH}\left(\bigcup_{v \text{ vertex of } \sigma} \mathrm{V}_\omega(v) \cap X\right)$. Therefore, $\mathrm{CH}\left(\bigcup_{v \text{ vertex of } \sigma} \mathrm{V}_\omega(v) \cap X\right)$ contains $\mathrm{CH}\left(\tilde{h}_0(\partial\sigma)\right)$, which, by Dugundji's extension theorem, contains $\tilde{h}_0(\sigma)$. Therefore, the property holds for every simplex of $\mathcal{C}_W^\alpha(L)$.

We can now prove that the image through $\tilde{h}_0$ of any arbitrary simplex σ of $\mathcal{C}_W^\alpha(L)$ does not intersect the medial axis of X. This is clearly true if σ is a simplex of $\mathcal{D}_\omega^X(L)$, since in this case $\tilde{h}_0(\sigma) = h_0(\sigma)$ is included in X. Assume now that $\sigma \notin \mathcal{D}_\omega^X(L)$. In particular, σ is not a vertex. Let v be an arbitrary vertex of σ. Consider any other vertex u of σ. Edge $[u, v]$ is α-witnessed by some point $w_{uv} \in W$. We then have $\|v-u\| \leq \|v-w_{uv}\| + \|w_{uv}-u\| \leq 2\mathrm{d}_2(w_{uv}) + 2\alpha$, where $\mathrm{d}_2(w_{uv})$ stands for the Euclidean distance from w_{uv} to its second nearest landmark. According to Lemma 3.4 of [4], we have $\mathrm{d}_2(w) \leq 3(\varepsilon+\delta)$, since L is an $(\varepsilon+\delta)$-sample of X. Thus, all the vertices of σ are included in the Euclidean ball $B(v, 2\alpha + 6(\varepsilon+\delta))$. Moreover, for any vertex u of σ and any point $p \in \mathrm{V}_\omega(u) \cap X$, we have $\|p-u'\| \leq \varepsilon+\delta$, where u' is a landmark closest to p in the Euclidean metric. Combined with the fact that $\|p-u\|^2 - \omega(u)^2 \leq \|p-u'\|^2 - \omega(u')^2$, we get: $\|p-u\|^2 \leq \|p-u'\|^2 + \omega(u)^2 \leq 2(\varepsilon+\delta)^2$, since by Lemma 3.3 of [4] we have $\omega(u) \leq 2\ \bar{\omega}\left(\frac{\varepsilon}{\mathrm{rch}(X)}\right)(\varepsilon + \delta) \leq \varepsilon + \delta$. Hence, $\mathrm{V}_\omega(u) \cap X$ is

[4]Indeed, every point $p \in L$ lies on X and belongs to its own cell, since ω has relative amplitude less than $\frac{1}{2}$. Therefore, $\mathrm{V}_\omega(p) \cap X \neq \emptyset$, which means that p is a vertex of $\mathcal{D}_\omega^X(L)$.

included in $B(u, \sqrt{2}(\varepsilon+\delta)) \subset B(v, 2\alpha + (6+\sqrt{2})(\varepsilon+\delta))$. Since this is true for every vertex u of σ, we get: $\tilde{h}_0(\sigma) \subseteq \mathrm{CH}\left(\bigcup_{u \text{ vertex of } \sigma} \mathrm{V}_\omega(u) \cap X\right) \subseteq B(v, 2\alpha + (6+\sqrt{2})(\varepsilon+\delta))$. Now, v belongs to $L \subseteq W \subseteq X$, and by assumption we have $2\alpha + (6+\sqrt{2})(\varepsilon+\delta) < \mathrm{rch}(X)$, therefore $\tilde{h}_0(\sigma)$ does not intersect the medial axis of X. □

5 Application to reconstruction

Taking advantage of the structural results of Section 3, we devise a very simple yet provably-good algorithm for constructing nested pairs of complexes that can capture the homology of a large class of compact subsets of $\mathbb{R}^d$. This algorithm is a variant of the greedy refinement technique of [30], which builds a set L of landmarks iteratively, and in the meantime maintains a suitable data structure. In our case, the data structure is composed of a nested pair of simplicial complexes, which can be either $\mathcal{R}^\alpha(L) \hookrightarrow \mathcal{R}^{\alpha'}(L)$ or $\mathcal{C}^\alpha_W(L) \hookrightarrow \mathcal{C}^{\alpha'}_W(L)$, for specific values $\alpha < \alpha'$. Both variants of the algorithm can be used in arbitrary metric spaces, with similar theoretical guarantees, although the variant using witness complexes is likely to be more effective in practice. In the sequel we focus on the variant using Rips complexes because its analysis is somewhat simpler.

5.1 The algorithm

The input is a finite point set W drawn from an arbitrary metric space, together with the pairwise distances $l(w, w')$ between the points of W. In the sequel, W is identified as the set of witnesses.

Initially, $L = \emptyset$ and $\varepsilon = +\infty$. At each iteration, the point of W lying furthest away[5] from L in the metric l is inserted in L, and ε is set to $\max_{w \in W} \min_{v \in L} l(w, v)$. Then, $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$ are updated, and the persistent homology of $\mathcal{R}^{4\varepsilon}(L) \hookrightarrow \mathcal{R}^{16\varepsilon}(L)$ is computed using the persistence algorithm [38]. The algorithm terminates when $L = W$. The output is the diagram showing the evolution of the persistent Betti numbers versus ε, which have been maintained throughout the process. As we will see in Section 5.2 below, with the help of this diagram the user can determine a relevant scale at which to process the data: it is then easy to generate the corresponding subset L of landmarks (the points of W have been sorted according to their order of insertion in L during the process), and to rebuild $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$. The pseudo-code of the algorithm is given in Figure 2.

5.2 Guarantees on the output

For any $i > 0$, let $L(i)$ and $\varepsilon(i)$ denote respectively L and ε at the end of the ith iteration of the main loop of the algorithm. Since $L(i)$ keeps growing with i, $\varepsilon(i)$ is a decreasing function of i. In addition, $L(i)$ is an $\varepsilon(i)$-sample of W, by definition of $\varepsilon(i)$. Hence, if W is a δ-sample of some compact set $X \subset \mathbb{R}^d$, then $L(i)$ is a $(\delta + \varepsilon(i))$-sample of X. This quantity is less than $2\varepsilon(i)$ whenever $\varepsilon(i) > \delta$. Therefore, Theorem 3.6 provides us with the following theoretical guarantee:

Theorem 5.1 *Assume that the input point set W is a δ-sample of some compact set $X \subset \mathbb{R}^d$, with $\delta < \frac{1}{18}\mathrm{wfs}(X)$. Then, at each iteration i such that $\delta < \varepsilon(i) < \frac{1}{18}\mathrm{wfs}(X)$, the persistent homology groups of $\mathcal{R}^{4\varepsilon(i)}(L(i)) \hookrightarrow \mathcal{R}^{16\varepsilon(i)}(L(i))$ are isomorphic to the homology groups of X^λ, for all $\lambda \in (0, \mathrm{wfs}(X))$.*

[5] At the first iteration, since L is empty, an arbitrary point of W is chosen.

Input: W finite, together with distances $l(w, w')$ for all $w, w' \in W$.
Init: Let $L := \emptyset$, $\varepsilon := +\infty$;
While $L \subsetneq W$ do
  Let $p := \text{argmax}_{w \in W} \min_{v \in L} l(w, v)$; *// p chosen arbitrarily in W if* $L = \emptyset$
  $L := L \cup \{p\}$;
  $\varepsilon := \max_{w \in W} \min_{v \in L} l(w, v)$;
  Update $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$;
  Compute persistent homology of $\mathcal{R}^{4\varepsilon}(L) \hookrightarrow \mathcal{R}^{16\varepsilon}(L)$;
End_while
Output: diagram showing the evolution of persistent Betti numbers versus ε.

Figure 2: Pseudo-code of the algorithm.

This theorem ensures that, when the input point cloud W is sufficiently densely sampled from a compact set X, there exists a range of values of $\varepsilon(i)$ such that the persistent Betti numbers of $\mathcal{R}^{4\varepsilon(i)}(L(i)) \hookrightarrow \mathcal{R}^{16\varepsilon(i)}(L(i))$ coincide with the ones of sufficiently small offsets X^λ. This means that a plateau appears in the diagram of persistent Betti numbers, showing the Betti numbers of X^λ. In view of Theorem 5.1, the width of the plateau is at least $\frac{1}{18}\text{wfs}(X) - \delta$. The theorem also tells where the plateau is located in the diagram, but in practice this does not help since neither δ nor $\text{wfs}(X)$ are known. However, when δ is small enough compared to $\text{wfs}(X)$, the plateau is large enough to be detected (and thus the homology of small offsets of X inferred) by the user or a software agent. In cases where W samples several compact sets with different weak feature sizes, Theorem 5.1 ensures that several plateaus appear in the diagram, showing plausible reconstructions at various scales – see Figure 1 (right). These guarantees are similar to the ones provided with the low-dimensional version of the algorithm [30].

Once one or more plateaus have been detected, the user can choose a relevant scale at which to process the data: as mentioned in Section 5.1 above, it is then easy to generate the corresponding set of landmarks and to rebuild $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$. Differently from the algorithm of [30], the outcome is not a single embedded simplicial complex, but a nested pair of abstract complexes whose images in $\mathbb{R}^d$ lie at Hausdorff distance[6] $O(\varepsilon)$ of X, such that the persistent homology of the nested pair coincides with the homology of X^λ.

5.3 Update of $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$

We will now describe how to maintain $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$. In fact, we will settle for describing how to rebuild $\mathcal{R}^{16\varepsilon}(L)$ completely at each iteration, which is sufficient for achieving our complexity bounds. In practice, it would be much preferable to use more local rules to update the simplicial complexes, in order to avoid a complete rebuilding at each iteration.

Consider the one-skeleton graph G of $\mathcal{R}^{16\varepsilon}(L)$. The vertices of G are the points of L, and its edges are the sets $\{p, q\} \subseteq L$ such that $\|p - q\| \leq 16\varepsilon$. Now, by definition, a simplex that is not a vertex belongs to $\mathcal{R}^{16\varepsilon}(L)$ if and only if all its edges are in $\mathcal{R}^{16\varepsilon}(L)$. Therefore, the simplices of $\mathcal{R}^{16\varepsilon}(L)$ are precisely the cliques of G. The simplicial complex can then be built as follows:

1. build graph G,
2. find all maximal cliques in G,

[6]Indeed, every simplex of $\mathcal{R}^{16\varepsilon}(L)$ has all its vertices in $X^{\varepsilon+\delta} \subseteq X^{2\varepsilon}$, and the lengths of its edges are at most 16ε.

3. report the maximal cliques and all their subcliques.

Step 1. is performed within $O(|L|^2)$ time by checking the distances between all pairs of landmarks. Here, $|G|$ denotes the size of G and $|L|$ the size of L. To perform Step 2., we use the output-sensitive algorithm of [37], which finds all the maximal cliques of G in $O(k\ |L|^3)$ time, where k is the size of the answer. Finally, reporting all the subcliques of the maximal cliques is done in time linear in the total number of cliques, which is also the size of $\mathcal{R}^{16\varepsilon}(L)$. Therefore,

Corollary 5.2 *At each iteration of the algorithm, $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$ are rebuilt within $O(|\mathcal{R}^{16\varepsilon}(L)|\ |L|^3)$ time, where $|\mathcal{R}^{16\varepsilon}(L)|$ is the size of $\mathcal{R}^{16\varepsilon}(L)$ and $|L|$ the size of L.*

5.4 Running time of the algorithm

Let $|W|, |L|, |\mathcal{R}^{16\varepsilon}(L)|$ denote the sizes of $W, L, \mathcal{R}^{16\varepsilon}(L)$ respectively. At each iteration, point p and parameter ε are computed naively by iterating over the witnesses, and for each witness, by reviewing its distances to all the landmarks. This procedure takes $O(|W||L|)$ time. According to Corollary 5.2, $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$ are updated (in fact, rebuilt) in $O(|\mathcal{R}^{16\varepsilon}(L)||L|^3)$ time. Finally, the persistence algorithm runs in $O(|\mathcal{R}^{16\varepsilon}(L)|^3)$ time [22, 38]. Hence,

Lemma 5.3 *The running time of one iteration of the algorithm is $O(|W||L| + |\mathcal{R}^{16\varepsilon}(L)||L|^3 + |\mathcal{R}^{16\varepsilon}(L)|^3)$.*

There remains to find a reasonable bound on the size of $\mathcal{R}^{16\varepsilon}(L)$, which can be done in Euclidean space $\mathbb{R}^d$, especially when the landmarks lie on a smooth submanifold:

Lemma 5.4 *Let L be a finite ε-sparse point set in $\mathbb{R}^d$. Then, $\mathcal{R}^{16\varepsilon}(L)$ has at most $2^{33^d}|L|$ simplices. If in addition the points of L lie on a smooth m-submanifold X of $\mathbb{R}^d$ with reach $\mathrm{rch}(X) > 16\varepsilon$, then $\mathcal{R}^{16\varepsilon}(L)$ has at most $2^{35^m}|L|$ simplices.*

Proof. Given an arbitrary point $v \in L$, we will show that the number of vertices in the star of v in $\mathcal{R}^{16\varepsilon}(L)$ is at most 33^d. From this follows that the number of simplices in the star of v is bounded by 2^{33^d}, which proves the first part of the lemma. Let Λ be the set of vertices in the star of v. These vertices lie within Euclidean distance 16ε of v, and at least ε away from one another. It follows that they are centers of pairwise-disjoint Euclidean d-balls of same radius $\frac{\varepsilon}{2}$, included in the d-ball of center v and radius $(16 + \frac{1}{2})\varepsilon$. Therefore, their number is bounded by $\frac{\mathrm{vol}B(v,(16+1/2)\varepsilon)}{\mathrm{vol}B(v,\varepsilon/2)} = \left(\frac{16+1/2}{1/2}\right)^d = 33^d$.

Assume now that v and the points of Λ lie on a smooth m-submanifold X of $\mathbb{R}^d$, such that $16\varepsilon < \mathrm{rch}(X)$. It follows then from Lemma 6 of [28] that, for all $u \in \Lambda$, we have $\|u-u'\| \leq \frac{\|u-v\|^2}{2\mathrm{rch}(X)} \leq \frac{\varepsilon^2}{2\mathrm{rch}(X)} < \frac{\varepsilon}{32}$, where u' is the orthogonal projection of u onto the tangent space of X at v, $T(v)$. As a consequence, the orthogonal projections of the points of Λ onto $T(v)$ lie at least $\frac{31\varepsilon}{32}$ away from one another, and still at most 16ε away from v. As a result, they are centers of pairwise-disjoint open m-balls of same radius $\frac{31\varepsilon}{64}$, included in the open m-ball of center v and radius $\left(16 + \frac{31}{64}\right)\varepsilon$ inside $T(v)$. Therefore, their number is bounded by $\left(\frac{16+31/64}{31/64}\right)^m \leq 35^m$, which proves the second part of the lemma, by the same argument as above. □

In cases where the input point cloud W lies on a smooth m-submanifold X of $\mathbb{R}^d$, the above result[7] suggests that the course of the algorithm goes through two phases: first, a transition phase,

[7]Note that, at every iteration i of the algorithm, $L(i)$ is an $\varepsilon(i)$-sparse point set, since the algorithm always inserts in L the point of W lying furthest away from L — see *e.g.* [30, Lemma 4.1].

in which the landmark set L is too coarse for the dimensionality of X to have an influence on the shapes and sizes of the stars of the vertices of $\mathcal{R}^{16\varepsilon}(L)$; second, a stable phase, in which the landmark set is dense enough for the dimensionality of X to play a role. This fact is quite intuitive: imagine X to be a simple closed curve, embedded in $\mathbb{R}^d$ in such a way that it roughly fills in the space within the unit d-ball. Then, for large values of ε, the landmark set L is nothing but a sampling of the d-ball, and therefore the stars of its points in $\mathcal{R}^{16\varepsilon}(L)$ are d-dimensional.

Let i_0 be the last iteration of the transition phase, *i.e.* the last iteration such that $\varepsilon(i_0) \geq \frac{1}{16}\operatorname{rch}(X)$. Then, Lemmas 5.3 and 5.4 imply that the time complexity of the transition phase is $O(|W||L(i_0)|^2 + 8^{33^d}|L(i_0)|^5)$, while the one of of the stable phase is $O(8^{35^m}\ |W|^5)$. We can get rid of the terms depending on d in at least two ways:

• The first approach has a rather theoretical flavor: it consists in amortizing the cost of the transition phase by assuming that W is sufficiently large. Specifically, since $L(i_0)$ is an $\varepsilon(i_0)$-sparse sample of X, with $\varepsilon(i_0) \geq \frac{1}{16}\operatorname{rch}(X)$, the size of $L(i_0)$ is bounded from above by some quantity $c_0(X)$ that depends solely on the (smooth) manifold X – see *e.g.* [5] for a proof in the special case of smooth surfaces. As a result, we have $8^{33^d}|L(i_0)|^k \leq 8^{35^m}|W|^k$ for all $k \geq 1$ whenever $|W| \geq 8^{33^d - 35^m}\ c_0(X)$. This condition on the size of W translates into a condition on δ, by a similar argument to the one invoked above.

• The second approach has a more algorithmic flavor, and it is based on a backtracking strategy. Specifically, we first run the algorithm without maintaining $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$, which simply sorts the points of W according to their order of insertion in L. Then, we run the algorithm backwards, starting with $L = L(|W|) = W$ and considering at each iteration j the landmark set $L(|W| - j)$. During this second phase, we do maintain $\mathcal{R}^{4\varepsilon}(L)$ and $\mathcal{R}^{16\varepsilon}(L)$ and compute their persistent Betti numbers. If W samples X densely enough, then Theorem 5.1 ensures that the relevant plateaus will be computed before the transition phase starts, and thus before the size of the data structure becomes independent of the dimension of X. It is then up to the user to stop the process when the space complexity becomes too large.

In both cases, we get the following complexity bounds:

Theorem 5.5 *If W is a point cloud in Euclidean space $\mathbb{R}^d$, then the running time of the algorithm is $O(8^{33^d}|W|^5)$, where $|W|$ denotes the size of W. If in addition W is a δ-sample of some smooth m-submanifold of $\mathbb{R}^d$, with δ small enough, then the running time becomes $O(8^{35^m}\ |W|^5)$.*

6 Conclusion

This paper makes effective the approach developped in [12, 15] by providing an efficient, provably good and easy-to-implement algorithm for topological estimation of general shapes in any dimensions. Our theoretical framework can also be used for the analysis of other persistence-based methods. Addressing a weaker version of the classical reconstruction problem, we introduce an algorithm that ultimately outputs a nested pair of complexes at a user-defined scale, from which the homology of the underlying shape X are inferred. When X is a smooth submanifold of $\mathbb{R}^d$, the complexity scales up with the intrinsic dimension of X. These results provide a new step towards reconstructing (low-dimensional) manifolds in high-dimensional spaces in reasonnable time with topological guarantees. It is now tempting to tackle the more challenging problem of constructing an embedded simplicial complex that is topologically and geometrically close to the sampled shape. As a first step, we intend to adapt our method to provide a single output complex that has the

same homology as X, using for instance the *sealing* technique of [25].

References

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

[2] N. Amenta, M. Bern, and D. Eppstein. The crust and the β-skeleton: Combinatorial curve reconstruction. *Graphical Models and Image Processing*, 60:125–135, 1998.

[3] F. Aurenhammer. Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405, September 1991.

[4] J.-D. Boissonnat, L. J. Guibas, and S. Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. In *Proc. 23rd ACM Sympos. on Comput. Geom.*, pages 194–203, 2007.

[5] J.-D. Boissonnat and S. Oudot. Provably good sampling and meshing of surfaces. *Graphical Models*, 67(5):405–451, September 2005.

[6] J.-D. Boissonnat and S. Oudot. Provably good sampling and meshing of Lipschitz surfaces. In *Proc. 22nd Annu. Sympos. Comput. Geom.*, pages 337–346, 2006.

[7] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, June 2007.

[8] F. Cazals and J. Giesen. Delaunay triangulation based surface reconstruction. In J.D. Boissonnat and M. Teillaud, editors, *Effective Computational Geometry for Curves and Surfaces*, pages 231–273. Springer, 2006.

[9] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in Euclidean space. In *Proc. 22nd Annu. ACM Sympos. Comput. Geom.*, pages 319–326, 2006.

[10] F. Chazal and A. Lieutier. The λ-medial axis. *Graphical Models*, 67(4):304–331, July 2005.

[11] F. Chazal and A. Lieutier. Topology guaranteeing manifold reconstruction using distance function to noisy data. In *Proc. 22nd Annu. Sympos. on Comput. Geom.*, pages 112–118, 2006.

[12] F. Chazal and A. Lieutier. Stability and computation of topological invariants of solids in $\mathbb{R}^n$. *Discrete Comput. Geom.*, 37(4):601–617, 2007.

[13] S.-W. Cheng, T. K. Dey, H. Edelsbrunner, M. A. Facello, and S.-H. Teng. Sliver exudation. *Journal of the ACM*, 47(5):883–904, 2000.

[14] S.-W. Cheng, T. K. Dey, and E. A. Ramos. Manifold reconstruction from point samples. In *Proc. 16th Sympos. Discrete Algorithms*, pages 1018–1027, 2005.

[15] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proc. 21st ACM Sympos. Comput. Geom.*, pages 263–271, 2005.

[16] V. de Silva. A weak definition of Delaunay triangulation. Technical report, Stanford University, October 2003.

[17] V. de Silva. A weak characterisation of the Delaunay triangulation. Submitted to *Geometriae Dedicata*, 2007.

[18] V. de Silva and G. Carlsson. Topological estimation using witness complexes. In *Proc. Sympos. Point-Based Graphics*, pages 157–166, 2004.

[19] V. de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007.

[20] J. Dugundji. An extension of Tietze's theorem. *Pacific J. Math.*, 1:353–367, 1951.

[21] H. Edelsbrunner. The union of balls and its dual shape. *Discrete Comput. Geom.*, 13:415–440, 1995.

[22] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[23] H. Edelsbrunner and N. R. Shah. Triangulating topological spaces. *Int. J. on Comp. Geom.*, 7:365–378, 1997.

[24] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.

[25] D. Freedman and C. Chen. Measuring and localizing homology classes. Technical Report, Rensselaer Polytechnic Institute, May 2007.

[26] J. Gao, L. J. Guibas, S. Y. Oudot, and Y. Wang. Geodesic Delaunay triangulations and witness complexes in the plane. In *Proc. ACM-SIAM Sympos. Discrete Algorithms*, 2008.

[27] R Ghrist. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.*, October 2007.

[28] J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds with high co-dimension. *Discrete and Computational Geometry*, 32:245–267, 2004.

[29] K. Grove. Critical point theory for distance functions. In *Proc. of Symposia in Pure Mathematics*, volume 54, 1993. Part 3.

[30] L. G. Guibas and S. Y. Oudot. Reconstruction using witness complexes. In *Proc. 18th Sympos. on Discrete Algorithms*, pages 1076–1085, 2007.

[31] A. Hatcher. *Algebraic Topology.* Cambridge University Press, 2001.

[32] T. Kaczynski, K. Mischaikow, and M. Mrozek. *Computational Homology.* Number 157 in Applied Mathematical Sciences. Springer-Verlag, 2004.

[33] P. McMullen. The maximal number of faces of a convex polytope. *Mathematika*, 17:179–184, 1970.

[34] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, to appear.

[35] S. Y. Oudot. On the topology of the restricted Delaunay triangulation and witness complex in higher dimensions. Manuscript. Preprint available at `http://geometry.stanford.edu/member/oudot/drafts/Delaunay_hd.pdf`, November 2006.

[36] V. Robins. Towards computing homology from approximations. *Topology*, 24:503–532, 1999.

[37] S. Tsukiyama, M. Ide, H. Ariyoshi, and I. Shirakawa. A new algorithm for generating all the maximal independent sets. *SIAM J. on Computing*, 6:505–517, 1977.

[38] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.