Approximating Persistent Homology in Euclidean Space Through Collapses

Magnus Botnan  
botnan@math.ntnu.no

Gard Spreemann  
spreeman@math.ntnu.no

March 4, 2014

Abstract

The Čech complex is one of the most widely used tools in applied algebraic topology. Unfortunately, due to the inclusive nature of the Čech filtration, the number of simplices grows exponentially in the number of input points. A practical consequence is that computations may have to terminate at smaller scales than what the application calls for.

In this paper we propose two methods to approximate the Čech persistence module. Both constructions are built on the level of spaces, i.e. as sequences of simplicial complexes induced by nerves. We also show how the bottleneck distance between such persistence modules can be understood by how tightly they are sandwiched on the level of spaces. In turn, this implies the correctness of our approximation methods.

Finally, we implement our methods and apply them to some example point clouds in Euclidean space.

1 Introduction

Topological data analysis in general, and persistent homology in particular, have shown great promise as tools for analyzing real-world data arising in the sciences. Examples of successful applications range from image analysis [6,25], to cancer research [1], virology [7] and sensor networks [13].

Central to persistent homology are standard constructions for recovering the homology of an underlying topological space from a finite sample set, chiefly the Čech and Vietoris–Rips complexes. Unfortunately, due to the inclusive nature of their filtrations, the number of simplices grows exponentially in the number of sample points. This may be unfortunate, as simplices added at small scales may contribute little to homology at larger, possibly more interesting, scales.

An extreme example may be a constant region in a measurement signal (perhaps from faulty equipment or downtime) under time-delay embedding [24]. In such a case, a large proportion of the point cloud may lie in, say, a dense lump of \(N\) points that contributes nothing to the cloud’s overall homology, yet introduces \(\binom{N}{k+1}\) \(k\)-simplices in the complex from an early scale.

Preprocessing of the point cloud, such as density considerations, may sometimes rectify the situation, but such schemes are often decidedly “off-line” in the sense that they require a one-off decision about which sparsifications to effectuate ahead of persistence computations. We propose more “on-line” methods wherein a decision to attempt a simplification of the simplicial complex may be made at any time during computations when it is deemed necessary. The simplification operation itself requires only that the
point cloud comes supplied with its complete linkage hierarchical clustering, which may be computed ahead of time once and for all, or the computation of nets.

1.1 Contributions

The well-known Nerve lemma [20] allows us to capture the topology of a continuous space using discrete structures. However, central to this lemma is the notion of a good cover, i.e. a cover wherein every finite intersection of covering sets are contractible. This means that whenever we have a parametrized sequence of covers, connected by maps of covers, the persistence diagram captured by the nerves equals the persistence diagram computed by singular homology at the level of spaces.

A central result in this paper is a way to bound the bottleneck distance between these two persistence diagrams when the covers are not necessarily good. Using this theorem we provide an approximation to the Čech persistence module built on a finite sample from Euclidean space. This method enjoys several favorable properties: it approximates the Čech persistence module well and allows for size reduction at a heuristic basis, i.e. only when the complex becomes to large too store. Unfortunately, computing the weights of the simplices turns out to be computationally expensive, making it inapplicable in most settings. To mend this we propose an easy to compute approximation which performs surprisingly well on real data sets. Using our aforementioned result we also show that the net-tree construction as introduced by Sheehy [26] and Dey et al. [16] works well for the Čech complex in Euclidean space. This approach enjoys very powerful theoretical bounds, e.g. a linear growth in the number of simplices as a function of sampled points. In practice, however, it is difficult to prevent the complex from growing too large. Having implemented an algorithm to compute persistence diagrams of simplicial complexes connected by simplicial maps, we conclude the paper by applying our approximations to a variety of point samples in Euclidean space.

To the best of our knowledge, this is the first paper where persistence computations are performed on simplicial complexes connected by more general simplicial maps than inclusion.

1.2 Outline

In Section 2 we review background material and Dey et al.’s algorithm [16] for computing persistent homology of simplicial complexes connected by simplicial maps. In particular, we introduce the concept of sequences of covers, and in Section 3 we give a homotopy colimit argument which relates the persistence module associated to a sequence of covers to that formed by the covering sets at the level of spaces. This relation is used in Section 3.1 to prove a sandwich type of theorem for sequences of covers. Applying this theorem we give two approaches to approximating the Čech persistence module in Section 4. The paper concludes with Section 5 where we compute the persistence diagrams of point clouds in Euclidean space using the aforementioned approximations.

1.3 Related work

In low-dimensional Euclidean space the alpha complex [17] offers a memory efficient way to compute the persistence diagram of a point cloud. Unfortunately, the number of simplices grows exponentially in the ambient dimension, making it inefficient in high-dimensional space. The witness complex [12] is a simplicial complex on a subset of the sample, called landmarks. The persistence diagrams of the associated filtration may, however, depending heavily on the choice of landmarks. Sheehy [26] and later Dey et al. [14] approximate the Vietoris–Rips complex using net-trees, and Kerber and Sharathkumar [21] arrive at similar results for the Čech complex in Euclidean space.
Using quadtrees. Our constructions in Section 4 is an adaption of the work of Dey et al. [14] to the Čech complex in Euclidean space. Chazal and Oudot [10] prove the results in Section 5 for the case where all the simplicial maps are inclusions.

2 Background material

In this section, we survey prerequisite background material and fix notation. We assume familiarity with basic concepts from algebraic topology, and basic knowledge of persistent homology. For introductions see [20] and [18], respectively.

Throughout the paper, all simplicial complexes are assumed to be finite and unoriented, and we write a $k$-simplex with vertices $i_0, \ldots, i_k$ as $[i_0, \ldots, i_k]$ or any permutation thereof. For a simplicial complex $K$, we will denote its geometric realization by $|K|$. Moreover, if $f : K \to L$ is a simplicial map between simplicial complexes, then $|f| : |K| \to |L|$ denotes the continuous map between their geometric realizations defined by $f$ on the vertices and extended linearly using barycentric coordinates. The $p$-th singular homology of a topological space $X$ with coefficients in a field $\mathbb{Z}_2$ will be denoted by $H_p(X)$, and for a continuous map $f : X \to Y$ we denote its induced map on homology by $f_* : H_p(X) \to H_p(Y)$. When $X = |K|$ is the geometric realization of a simplicial complex, we will make no distinction between the $p$-th simplicial homology module of $K$ and the $p$-th singular homology group of $|K|$. Cohomology vector spaces over $\mathbb{Z}_2$ are similarly denoted by $H^p(X)$.

A collection of open sets $U = \{U_i \subseteq U \mid i \in I\}$ indexed by a finite set $I$ is said to be a finite cover of $U$ if $U = \bigcup_{i \in I} U_i$. The nerve $NU$ of the cover $U$ is the simplicial complex with vertex set $I$ and a $k$-simplex $[i_0, \ldots, i_k] \in NU$ if $U_{i_0} \cap \cdots \cap U_{i_k} \neq \emptyset$. Let $U = \{U_i \mid i \in I\}$ and $V = \{V_j \mid j \in J\}$ be covers of topological spaces $U \subseteq V$. A map of sets $F : I \to J$ is said to be a map of covers if $U_i \subseteq V_{F(i)}$ for all $i \in I$. It is easy to check that $F$ extends to a simplicial map $F : NU \to NV$ between the nerves of the covers. By a sequence of covers we will mean a collection of covers $\{U(\alpha) \mid \alpha \in A \subset [0, \infty)\}$, each indexed respectively by $I(\alpha)$, together with maps of covers $F_\alpha' : I(\alpha) \to I(\alpha')$ such that $F_{\alpha''} = id$ and $F_{\alpha}'' = F_{\alpha'}'' \circ F_{\alpha}'$ for all $\alpha'' \geq \alpha' \geq \alpha$. Such a sequence will be denoted by a pair $(U, F)$. Similarly, for any sequence of covers we have an induced sequence of nerves which will be denoted by $(NU, F)$.

2.1 Persistence modules

A persistence module $\mathcal{V}$ over $A \subseteq \mathbb{R}$ is a collection of $k$-vector spaces $\{V(\alpha) \mid \alpha \in A\}$ and linear maps $v_{\alpha} : V(\alpha) \to V(\alpha')$ for all $\alpha \leq \alpha'$ such that $v_{\alpha} = id$ and $v_{\alpha} = v_{\alpha}'' \circ v_{\alpha}'$. The direct sum of two persistence modules $\mathcal{U}$ and $\mathcal{W}$, both indexed over the same set, is the persistence module $\mathcal{V} = \mathcal{U} \oplus \mathcal{W}$ where $V(\alpha) = U(\alpha) \oplus W(\alpha)$ and $v_{\alpha} = u_{\alpha} \oplus w_{\alpha}$. We say that $\mathcal{V}$ is indecomposable if the only decompositions of $\mathcal{V}$ are the trivial decompositions $0 \oplus \mathcal{V}$ and $\mathcal{V} \oplus 0$.

Definition 1. Let $J \subseteq A$ be an interval, i.e. if $s, t \in J$ and $s < r < t$ then $r \in J$. The interval module over $J$ is the persistence module $I^J$ defined by

$$I^J(\alpha) = \begin{cases} k & \text{if } \alpha \in J \\ 0 & \text{otherwise} \end{cases}$$

and $i^J_\alpha = id : I^J(\alpha) \to I^J(\alpha')$ whenever $\alpha, \alpha' \in J$ and $0$ otherwise.

It is not difficult to show that $I^J$ is indecomposable, and the Krull–Remak–Schmidt–
Azumaya theorem \[3\] tells us that if

$$\mathcal{V} \simeq \bigoplus_{l \in L} I^l$$

then there is a bijection $\sigma : L \to M$ such that $J_l = K_{\sigma(l)}$ for all $l \in L$. So whenever $\mathcal{V}$ admits such a decomposition we can characterize it by the multiset $\{J_l \mid l \in L\}$ of intervals called the persistence diagram $\mathbf{D}(\mathcal{V})$ of $\mathcal{V}$, which is usually drawn as a collection of points in $(\mathbb{R} \cup \{-\infty\})^2$. An interval $(b, d) \in \mathbf{D}(\mathcal{V})$ represents a feature of $\mathcal{V}$ with birth and death time $b$ and $d$, respectively. A recent theorem by Crawley-Boevey \[11\] asserts that $\mathcal{V}$ admits a decomposition into interval modules if $V_\alpha$ is finite-dimensional for all $\alpha \in \mathbb{R}$. For an example of a persistence module which does not admit an interval decomposition, see \[8\).

For a sequence of covers $(\mathcal{U}, F)$ we get a persistence module $(H_p(\mathcal{U} \cup \{x\}), F_\alpha)$ with vector spaces $\{H_p(\mathcal{U}(\alpha)) \mid \alpha \in A \subseteq [0, \infty]\}$ and maps $(F_\alpha^p)_*$. As the covers are finite, all the homology groups will have finite dimension, and thus persistence diagrams are well-defined. In particular, if $P \subseteq M$ is a finite set of points in a metric space $M$ and $B(p; \alpha)$ is the open ball of radius $\alpha$ centered at $p$ we get a sequence of covers by defining $\mathcal{U}(\alpha) = \{B(p; \alpha) \mid p \in P\}$ and $F = \text{id}$. The induced sequence of nerves is known as the Čech filtration and the associated persistence module is the Čech persistence module. In the remainder of this paper $C(P; \alpha)$ denotes the nerve of the Čech filtration of $P$ at scale $\alpha$.

Another popular construction is the Vietoris–Rips complex $\mathcal{R}(P; \alpha)$ which is defined as the largest simplicial complex with the same 1-skeleton as $C(P; \alpha)$. By definition, it follows that $C(P; \alpha) \subseteq \mathcal{R}(P; \alpha)$, and for $P \subseteq \mathbb{R}^n$, it is also true that $\mathcal{R}(P; \alpha) \subseteq C(P; \sqrt{2}\alpha)$ \[13\].

### 2.2 Metrics and approximations

In the following, $\Delta$ denotes the multiset of all pairs $(x, x) \in (\mathbb{R} \cup \{-\infty\})^2$, each with countably infinite multiplicity. A partial matching between two persistence diagrams $D$ and $D'$ is a bijection $\gamma : B \cup \Delta \to B' \cup \Delta$, and we denote all such by $\Gamma(D, D')$.

The following defines a metric on persistence diagrams:

**Definition 2.** The bottleneck distance between two persistence diagrams $B$ and $B'$ is

$$d_B(B, B') = \inf_{\gamma \in \Gamma(D, D')} \sup_{(b, d) \in B} \|((b, d) - \gamma((b, d)))\|_\infty$$

where

$$\|((b_1, d_1) - (b_2, d_2))\|_\infty = \max(|b_1 - b_2|, |d_1 - d_2|).$$

The theory of interleavings \[9\] offers a generalization of the bottleneck distance to persistence modules that do not admit a decomposition into indecomposables. Importantly, if there exists an $\epsilon$-interleaving between two persistence modules, then their bottleneck distance is at most $\epsilon$. In this paper we adopt the conventions of \[21\] and use a slight reformulation of the ordinary theory of interleavings.

**Definition 3.** Two persistence modules $U$ and $V$ indexed over $[0, \infty)$ are said to be $c$-approximate if there exist a constant $c \geq 1$ and two families of homomorphisms $\{\phi_\alpha : U(\alpha) \to V(c\alpha)\}_{\alpha \geq 0}$ and $\{\psi_\alpha : V(\alpha) \to U(c\alpha)\}_{\alpha \geq 0}$ such that the following four
diagrams commute for all $\alpha \leq \alpha'$:

With this the following theorem [9] is immediate:

**Theorem 4.** If $U$ and $V$ are $c$-approximate, then their bottleneck distance is bounded by $\log c$ on the log-scale.

The above result can be seen as a general version of the relationship between the Čech and Vietoris–Rips filtrations. Indeed, while the bottleneck distance between their persistence diagrams may be arbitrarily large, the inclusions

$$C(P; \alpha) \subseteq R(P; \alpha) \subseteq C(P; \sqrt{2}\alpha)$$

ensure that a feature $(b,d)$ in the Vietoris–Rips persistence module is also a feature in the Čech persistence module if $d - b \geq \sqrt{2}b$, and vice versa.

### 2.3 Computing persistent homology using annotations

Many widely implemented and used algorithms for computing persistent homology assume that the maps in the persistence module are induced by inclusions of simplicial complexes, i.e. that the underlying sequence is a filtration. As shall become clear, we will need to compute in the setting of general simplicial maps.

**Definition 5.** A simplicial map $f : K \to K'$ with the property that

$$f(\sigma) = \begin{cases} \sigma \setminus \{b\} & \text{if } a, b \in \sigma \\ \{a\} \cup \sigma \setminus \{b\} & \text{if } a \notin \sigma, b \in \sigma \\ \sigma & \text{otherwise.} \end{cases}$$

is called an **edge contraction** of $[a,b]$ to $[a]$. Simplices $\sigma, \sigma' \in K$ are called **mirror simplices** (for $f$) if $f(\sigma) = f(\sigma')$.

We will often refer to an edge contraction like that above by $[a,b] \mapsto [a]$. Since any simplicial map $K \to K'$ decomposes into a finite sequence of inclusions and edge contractions, we only need to deal with those two types and adjust the persistence module indices accordingly to reflect the addition of extra maps. Likewise, as is normal, we decompose inclusions into ones of the form $K \to K \cup \{\sigma\}$ and refer to these as “adding a simplex $\sigma$”.

We will use Dey et al.’s method of **persistence annotations** [16] to compute (the persistence diagrams of) persistence modules with simplicial maps, and now quickly review their algorithm and our implementation details.

The method of annotation tracks homology with $\mathbb{Z}_2$ coefficients across a persistence module by storing the value of all cohomology generators at each simplex and updating...
these “annotations” to reflect the inclusion of a simplex or the contraction of an edge. Care should be taken to notice a slight difference in terminology: our definition of annotations reflects Dey’s valid annotations.

**Definition 6.** An annotation for a simplicial complex $K$ is a linear map $\Phi_p : C_p(K) \to \mathbb{Z}^2_2$ with the property that

$$\varphi_1 = [c \mapsto \Phi_p(c)_1], \ldots, \varphi_n = [c \mapsto \Phi_p(c)_n]$$

is a basis for $H^p(K)$. Here $\Phi_p(c)_i$ denotes the $i$'th component of $\Phi_p(c) \in \mathbb{Z}^2_2$.

A key observation is the following: the persistent homology of a sequence of simplicial complexes can be obtained by dualizing on the level of chains and taking cohomology. This is true since when working over field coefficients, the map $H^p(K) \to \text{Hom}(H_p(K), \mathbb{Z}_2)$ defined by $\alpha([f])([c]) = f(c)$ is an isomorphism. Thus, intervals in persistent cohomology are dual to intervals in persistent homology. Therefore, we shall interchangeably speak of a homology class born at persistence index $i$ as a cohomology class in the opposite direction dying at persistence index $i$.

By storing the value of $\Phi_p$ at each $p$-simplex, that simplex’ contribution to the (co)homology vector space is known and so allows us to only make changes to homology near the site of a contraction. This “locality” of the changes introduced by an edge contraction is summarized in the following definition [15] and proposition [2].

**Definition 7.** The link of a simplex $\sigma$ in a simplicial complex $K$ is the set

$$\text{lk}\sigma = \{\tau \setminus \sigma \mid \sigma \subseteq \tau \in K\}.$$

An edge $[a, b]$ satisfies the **link condition** if $\text{lk}[a] \cap \text{lk}[b] = \text{lk}[a, b]$.

**Proposition 8.** The contraction $f : K \to K'$ of an edge that satisfies the link condition induces a homotopy equivalence $\vert f\vert : \vert K\vert \to \vert K'\vert$, and hence an isomorphism $f_* : H_*(K) \to H_*(K')$.

Suppose

$$K = (K_0 \xrightarrow{f_0} K_1 \xrightarrow{f_1} \cdots \xrightarrow{f_{m-1}} K_m)$$

is a sequence of simplicial complexes (with the $f_i$’s simplicial maps) whose persistence module

$$H_*(K) = (H_*(K_0) \xrightarrow{(f_0)_*} H_*(K_1) \xrightarrow{(f_1)_*} \cdots \xrightarrow{(f_{m-1})_*} H_*(K_m))$$

has been computed, and write $\Phi_p^i$ for the annotation of $H^p(K_i)$ and $n$ for its dimension. To compute the persistence module of

$$K' = (K_0 \xrightarrow{f_0} \cdots \xrightarrow{f_{m-1}} K_m \xrightarrow{f_m} K_{m+1}),$$

there are four cases to handle:

1. $f_m$ adds a single simplex $p$-simplex $\sigma$, and...

   (a) $\Phi_{p-1}^m(\partial \sigma) = 0$. This corresponds to a generator of $H_p(K')$ being born at persistence index $m + 1$, or equivalently to a generator of $H^p(K')$ dying at $m$ going left (see Proposition 5.2 in [16]). Define $\Phi_{p+1}^m$ on the $p$-simplices of $K_{m+1}$ by

   $$\Phi_{p+1}^m(\tau) = \begin{cases} (\Phi_p^m(\tau)_0, \ldots, \Phi_p^m(\tau)_n, 0) & \text{if } \tau \neq \sigma \\ (\Phi_p^m(\tau)_0, \ldots, \Phi_p^m(\tau)_n, 1) & \text{if } \tau = \sigma \end{cases}$$

   and extend linearly. In other dimensions $q \neq p$, we set $\Phi_{q+1}^m = \Phi_q^m$.
As suggested in [4, 5], the simplicial complex $\Phi^m_\partial(K) = \cdots = \Phi^m_\partial(\partial \sigma)_{ij} = 1$ for some $l \geq 1$. In this case $\sigma$ kills a class in $H_{p-1}(K')$ at $m+1$, or equivalently gives birth to one of the generators $\varphi_i$, $\cdots$, $\varphi_i$ of $H^{p-1}(K')$ in the reverse direction (see Proposition 5.2 in [16]). We kill the youngest homology class, say the one numbered $u$ (so $\varphi_u$ is born in the reverse direction). Define $\Phi^m_{p-1}$ on the $(p-1)$-simplices of $K_{m+1}$ by

$$\Phi^m_{p-1}(\tau) = \begin{cases} 
\Phi^m_{p-1}(\tau) + \Phi^m_\partial(\partial \sigma) & \text{if } \Phi^m_{p-1}(\tau)_u = 1 \\
\Phi^m_{p-1}(\tau) & \text{otherwise},
\end{cases}$$

wherein $\sigma$ is the . In other dimensions $q \neq p$, we set $\Phi^m_q = \Phi^m_q$. 

2. $f_m$ contracts $[a, b]$ to $[a]$, and...

(a) $[a, b]$ satisfies the link condition. Let

$$M_{p-1} = \{ \sigma \in K_m \mid \dim \sigma = p-1, a \in \sigma \text{ and } \sigma \text{ has a mirror under } f_m \},$$

and note that to any $\tau \in M_{p-1}$, there is a unique $g_\tau \in K_m$ with $\tau \subseteq g_\tau$, $\dim g_\tau = p$ and $[a, b] \subseteq g_\tau$. Define $\Phi^m_{p-1}$ on the $p$-simplices of $K_{m+1}$ by

$$\Phi^m_{p-1}(\sigma) = \Phi^m_p(\sigma) + \sum_{\sigma \supseteq \tau \in M_{p-1}} \Phi^m_p(g_\tau),$$

noting that the sum may be empty. This corresponds to Dey’s “annotation transfers” – see Proposition 4.4 and 4.5 of [16] for a more detailed explanation.

(b) $[a, b]$ does not satisfy the link condition. As discussed earlier, we add the simplices in $(\text{lk}[a] \cap \text{lk}[b]) \setminus \text{lk}[a, b]$, which are cofaces of $[a, b]$, repeatedly hitting the cases [1a] and [1b] until the link condition becomes fulfilled and we finally reach case [25]. Some bookkeeping is of course required if one wants to consider the potentially many homology changes from the inclusions as occurring at persistence index $m + 1$.

Dey et al. show in [16] (Proposition 5.1) that $\Phi^m_{m+1}$ as constructed above is an annotation for $H^*(K_{m+1})$. With $K_0 = \emptyset$ and the associated empty annotation, then, the above is a correct algorithm for computing persistent homology.

### 2.3.1 Some implementation details

As suggested in [4,5], the simplex tree is a data structure that is well-suited for storing the simplicial complex in the above algorithm.

A simplex tree is a trie (also called a prefix tree), which is a tree $T$ that stores a simplicial complex $K$ whose vertices $V$ have a total ordering $\leq$ by the following rules:

- $T$ contains a distinguished root.
- Every non-root node $n \in T$ carries the data of a label $L(n) \in V$. The root is labelled by a distinguished symbol, say $*$, and we extend the ordering to $* \leq v$ for all $v \in V$ to ease notation.
- Nodes have zero or more children.
- If $n$ is a child of $p$, then $L(n) > L(p)$.
- If $n$ and $m$ are both children of $p$, then $L(n) \neq L(m)$. 

7
The simplicial complex $K$ to be encoded corresponds to all paths to the root of $T$, and we write $S(n) \in K$ for the simplex corresponding to the path from $n \in T$. We will also refer to the root having depth 0, and in general a node as having depth $k + 1$ if its parent has depth $k$. Thus $\text{depth}(n) = \dim S(n) + 1$.

In terms of implementation, every node holds a pointer to its parent and a dictionary of pointers to its children, keyed on their labels. Furthermore, we augment the tree by adding to each node a “cousin pointer”: We call $m$ a cousin of $n$ if $\text{depth}(m) = \text{depth}(n)$ and $L(m) = L(n)$. Every node holds a pointer to one of its cousins in such a way that they form a cyclic linked list that visits every cousin at the same depth precisely once (per cycle). In addition, an arbitrary representative of each such cyclic linked list is maintained in a dictionary keyed on labels and depths.

Figure 1 shows an example of the basic part of a simplex tree, along with an example of annotations (intermediate data structures are dropped from the figure, and annotations are attached directly to the simplices for ease of visualization).

![Figure 1: A somewhat simplified simplex tree representation of a simplicial complex. Annotation values on the 1-simplices are included for a persistence module in which the simplices are added in the order ..., [1, 5], [4, 5], [2, 5], [3, 5], [1, 4], [2, 3], [1, 4, 5], [2, 3, 5], [3, 4], [1, 2], leading up to the top row situation. To contract the edge [1, 2], the link condition must be fulfilled, requiring the inclusion of [1, 2, 5] (middle row). The situation after contraction is shown in the bottom row.

Boissonnat and Maria show that this data structure allows us to efficiently insert and remove simplices, and compute their faces and cofaces. For details, see [5].

To tie the simplex tree to the annotations discussed earlier, we want to associate to each node (i.e. each simplex) its annotation value. Since multiple simplices are likely to share the same annotation value, we go by way of a union find structure. Each node thus contains a pointer to a node in a forest, wherein each tree represents an annotation.

\footnote{A dictionary is here any data structure with logarithmic lookup time complexity for keys.}
value shared by multiple cohomologous simplices. The root of each tree in the forest points to the actual annotation value of the simplices pointing to nodes in that tree.

The annotation values themselves are also kept referenced in a dictionary (keyed on the annotation values) for easy access and updating as used in the algorithm outlined earlier.

3 Persistent homology of sequences of covers

In the following we assume that all covering sets are subsets of some metric space and that every cover is finite. In particular, this means that all our spaces are paracompact. Moreover, the constructions in this section can be seen as special cases of the much more general construction of a homotopy colimit of a diagram of topological spaces.

To any open covering \( \mathcal{U} = \{ U_i \mid i \in I \} \) of an open set \( U \) we assign a topological space \( \Delta U \) defined as the disjoint union

\[
\bigcup_{S \subseteq N U} |S| \times \bigcap_{i \in S} U_i
\]

under the equivalence relation \((s, x) \sim (t, x)\) if \( s \in |S|, t \in |T|, S \subseteq T \) and \( s = t \). This construction comes equipped with continuous projection maps \( \pi_1: \Delta U \to |N U| \) and \( \pi_2: \Delta U \to U \) given by projecting onto the first and second factor, respectively.

Lemma 9. The fiber projecting map \( \pi_2: \Delta U \to U \) is a homotopy equivalence.

Sketch. As \( U \) is assumed to be paracompact we can choose a partition of unity \( \{ \phi_i \}_{i \in I} \) subordinate to \( U \) and define \( g: U \to \Delta U \) by

\[
g(x) = \sum_{i \in I} (\phi_i(x)v_i, x),
\]

where \( v_i \) is the vertex corresponding to \( U_i \). Then \( \pi_2 \circ g = id_U \) and it is not difficult to show that \( g \circ \pi_2 \simeq id_{\Delta(U)} \). For a complete proof see [20].

Now let \( \mathcal{V} = \{ V_j \mid j \in J \} \) be a finite covering of \( V \supseteq U \) and \( F: I \to J \) a map of coverings. Recall that \( |F|: |NU| \to |NV| \) denotes the continuous map defined on the vertices by the induced simplicial map between the nerves. If we let \( \text{inc}_U^V: U \hookrightarrow V \) denote the inclusion of \( U \) into \( V \) we get the commutative diagram

\[
\begin{array}{ccc}
U & \xrightarrow{\text{inc}_U^V} & V \\
\downarrow{\pi_2} & \simeq & \downarrow{\pi_2} \\
\Delta U & \xrightarrow{|F| \times \text{inc}_U^V} & \Delta V \\
\downarrow{\pi_1} & & \downarrow{\pi_1} \\
|NU| & \xrightarrow{|F|} & |NV|
\end{array}
\]

By passing to (singular) homology and using that \( \pi_2 \) is a homotopy equivalence we can reverse arrows to find the following commutative diagram:

\[
\begin{array}{ccc}
H_*(U) & \xrightarrow{(\text{inc}_U^V)_*} & H_*(V) \\
\downarrow{\text{inc}_U^V_*} & & \downarrow{\text{inc}_U^V_*} \\
H_*(|NU|) & \xrightarrow{|F|_*} & H_*(|NV|)
\end{array}
\]
Example 10. Note that Diagram 1 does not commute on the level of spaces: let $U = \{U\}$ and $V = \{U, V\}$ where $U \cap V \neq \emptyset$. If $x \in U \cap V$ then $(|F| \circ \pi_1 \circ g)(x)$ is a point in $|NU|$ whereas $(\pi_1 \circ g \circ \text{inc})(x)$ can be any point along the edge $|[U, V]|$, depending on the choice of partition of unity. See Figure 2.

![Figure 2](image)

Figure 2: This diagram is an example of the diagram from Example 10 not commuting on the level of spaces.

Definition 11. A cover $\mathcal{U} = \{U_i \mid i \in I\}$ is said to be good if every finite intersection of open sets in $\mathcal{U}$ is contractible.

The following theorem is one of the great pillars of computational algebraic topology. It allows us to use discrete information to capture the topology of a continuous space. For a proof see Section 4.G. of [20].

**Theorem 12.** If $\mathcal{U}$ is a good cover, then the base projection map $\pi_1 : \Delta \mathcal{U} \to |\mathcal{N}|\mathcal{U}$ is a homotopy equivalence.

**Corollary 13.** If $\mathcal{U}$ is a good cover, then the composition $(\pi_1 \circ \pi_2^{-1})_*$ is an isomorphism.

### 3.1 A sandwich theorem for sequences of covers

We will use the results from the previous section to prove a sandwich type of theorem for sequences of covers. The idea is that if a sequence of covers can be sandwiched between two sequences of good covers, then the middle persistence module associated to the middle sequence of covers approximates the persistence modules associated to the good covers.

Let $\left(\mathcal{U}, F_{\mathcal{U}(\alpha)}^{\mathcal{U}(-)}\right), \left(\mathcal{V}, F_{\mathcal{V}(\alpha)}^{\mathcal{V}(-)}\right)$ and $\left(\mathcal{W}, F_{\mathcal{W}(\alpha)}^{\mathcal{W}(-)}\right)$ be sequences of covers satisfying

$$U(\alpha) \subseteq V(\alpha) \subseteq W(\alpha) \subseteq U(\alpha)$$

(2)

together with maps of covers

$$F_{\mathcal{V}(\alpha')}^{\mathcal{W}(\alpha')} : V(\alpha) \to W(\alpha')$$

(3)

$$F_{\mathcal{W}(\alpha)}^{\mathcal{V}(\alpha')} : W(\alpha) \to V(\alpha')$$

for all $\alpha' \geq \alpha$ and a fixed constant $c \geq 1$. We will assume that the maps of covers satisfy the coherence relations

$$F_{\mathcal{W}(\alpha')}^{\mathcal{W}(\alpha')} \circ F_{\mathcal{V}(\alpha')}^{\mathcal{W}(\alpha')} = F_{\mathcal{V}(\alpha')}^{\mathcal{V}(\alpha')} \circ F_{\mathcal{V}(\alpha)}^{\mathcal{V}(\alpha')}$$

$$F_{\mathcal{W}(\alpha/c)}^{\mathcal{W}(\alpha/c)} \circ F_{\mathcal{V}(\alpha/c)}^{\mathcal{W}(\alpha/c)} = F_{\mathcal{V}(\alpha/c)}^{\mathcal{V}(\alpha/c)}$$

(4)

for all $\alpha'' \geq \alpha' \geq \alpha$. 


If Theorem 14. Also, note that the map \( \eta \) making the following diagrams commute:

\[
\begin{align*}
H_p(U(\alpha)) \xrightarrow{\text{inc}^{V(\alpha')}} H_p(V(\alpha')) \\
H_p([NU(\alpha)]) \xrightarrow{\eta\text{inc}^{V(\alpha')}} H_p([NV(\alpha')]) \\
H_p(W(\alpha)) \xrightarrow{\text{inc}^{W(\alpha')}} H_p(U(\alpha')) \\
H_p([NW(\alpha)]) \xrightarrow{\eta\text{inc}^{W(\alpha')}} H_p([NW(\alpha')])
\end{align*}
\]

Hence, there are well-defined homomorphisms

\[
\begin{align*}
\phi_\alpha &= \eta_{\text{inc}^{V(\alpha')}}^V : H_p([NU(\alpha)]) \to H_p([NV(\alpha)]) \\
\psi_\alpha &= \eta_{\text{inc}^{W(\alpha')}}^W : H_p([NV(\alpha)]) \to H_p([NU(\alpha)])
\end{align*}
\]

Also, note that the map \( \eta_{\text{inc}^{W(\alpha')}}^W \) is the unique map that makes Diagram 1 commute.

**Theorem 14.** If \( (\mathcal{U}, F_{\mathcal{U}(-)}(-)) \) and \( (\mathcal{W}, F_{\mathcal{W}(-)}(-)) \) are sequences of good covers, then the families of homomorphisms \( \{\phi_\alpha\}_{\alpha \in [0, \infty)} \) and \( \{\psi_\alpha\}_{\alpha \in [0, \infty)} \) defined in Equation 5 satisfy the diagrams of Definition 3. In particular, the persistence modules

\[
(H_p([NU]), \eta_{\text{inc}^{U(-)}}^U) \quad \text{and} \quad (H_p([NV]), \eta_{\text{inc}^{V(-)}}^V)
\]

are \( c \)-approximate.

**Proof.** We need to show that the following four relations in Definition 3 are satisfied:

\[
\begin{align*}
\psi_{\alpha'} \circ \eta_{\text{inc}^{V(\alpha')}}^V \circ \phi_{\alpha'/c} &= \eta_{\text{inc}^{(\alpha')}}^U \\
\eta_{\text{inc}^{W(\alpha')}}^W \circ \psi_{\alpha'/c} &= \eta_{\text{inc}^{(\alpha')}}^W \\
\phi_{\alpha'} \circ \eta_{\text{inc}^{(\alpha')}}^U \circ \psi_{\alpha'/c} &= \eta_{\text{inc}^{W(\alpha')}}^W \\
\phi_{\alpha'} \circ \eta_{\text{inc}^{W(\alpha')}}^W \circ \psi_{\alpha'/c} &= \eta_{\text{inc}^{V(\alpha')}}^V
\end{align*}
\]

The proof follows from the coherence relations in Equation 4 and the uniqueness of \( \eta_{\text{inc}^{U(\alpha)}}^U, \eta_{\text{inc}^{V(\alpha)}}^V, \eta_{\text{inc}^{W(\alpha)}}^W \) and \( \eta_{\text{inc}^{V(\alpha')}}^V, \eta_{\text{inc}^{W(\alpha')}}^W \).

1. The composition \( \psi_{\alpha'} \circ \eta_{\text{inc}^{V(\alpha')}}^V \circ \phi_{\alpha'/c} : \mathcal{U}(\alpha/c) \to \mathcal{U}(\alpha') \) has the same domain and co-domain as \( \eta_{\text{inc}^{(\alpha')}}^U \) and therefore they are equal by uniqueness.

2. By definition \( \phi_{\alpha'} \circ \eta_{\text{inc}^{U(\alpha')}}^U \circ \psi_{\alpha'/c} = \eta_{\text{inc}^{W(\alpha')}}^W \circ \psi_{\alpha'/c} = \eta_{\text{inc}^{W(\alpha')}}^W \circ \psi_{\alpha'/c} \). Using that the composition of the three leftmost maps has same domain and co-domain as \( \psi_{\alpha'/c} \), we are left with \( \eta_{\text{inc}^{W(\alpha')}}^W \circ \psi_{\alpha'/c} = \psi_{\alpha'/c} \). Here the last equality follows from Equation 4.
This section is an adaption of the work in [14] to the case of Čech complexes in Euclidean space. Throughout this section, we choose parameters $\varepsilon_0, \epsilon \geq 0$ and define a sequence of point sets $P_k$ for $k = 0, 1, \ldots, m$ such that $P_0 = P$ and $P_{k+1}$ is an $\varepsilon_0 \alpha^2 (1 + \epsilon)^{k-1}$-net of $P_k$. We refer to such a collection $P_0, \ldots, P_m$ as a net-tree. Furthermore, let $\mathcal{C}(P_k; \alpha)$ be the Čech complex at scale $\alpha$ built upon the vertex set $P_k$, and $U(P_k; \alpha)$ the union of open balls of radius $\alpha$ centered at each point in $P_k$. We clearly have maps $\pi_k : P_k \to P_{k+1}$ which send a vertex $p \in P_k$ to its most nearby vertex in $P_{k+1}$.

3. Once again, using the coherence relations in Equation (4) we find $\psi_{(\alpha')} \circ \eta_{V(\alpha')} = \eta_{U(\alpha')} \circ \eta_{V(\alpha')} \circ \eta_{V(\alpha)} = \eta_{U(\alpha')} \circ \eta_{V(\alpha)}$. Both $\eta_{U(\alpha')} \circ \eta_{V(\alpha)}$ and $\eta_{U(\alpha)} \circ \eta_{V(\alpha)}$ equal $\eta_{V(\alpha)}$ by uniqueness.

4. Both compositions in the last equation equal $\eta_{V(\alpha)}$ by uniqueness.

The following is a corollary of the proof.

**Corollary 15.** Any pair of the persistence modules

$$
\left( H_p(|NU|), \eta_{U(-)} \right), \left( H_p(|NV|), \eta_{V(-)} \right) \text{ and } \left( H_p(|NW|), \eta_{W(-)} \right)
$$

are $c$-approximate.

Note that we do not require the covers in the sequence $(V, I_{V(-)})$ to be good. An application of that is the following. Let us say that we filter a space by a sequence of covers which is not necessarily good. Now, for every scale parameter we get a good cover by taking the convex hull of each covering set. The previous theorem now tells us that the error in the bottleneck distance between the persistence module formed by singular homology at the level of spaces, and that associated to the sequence of covers, can be understood from studying how much bigger the covering set becomes by taking the convex hull. That is, the error is bounded by $\log c$ on the log scale, where $c$ is a constant such that the covering set at scale $\alpha$ covers the covering formed by taking convex hulls at scale $\alpha$. In the following section we will utilize this idea to construct an approximation to the Čech persistence module.

## 4 Approximating the Čech complex in Euclidean space

In this section we construct two different approximation schemes for the Čech persistence module built on a finite set of points $P$ in Euclidean space.

It is clear that for any $c$-approximation of the Čech persistence module, a $\sqrt{2c}$-approximation can be had via the Vietoris–Rips complex built on its 1-skeleton. For a treatment of approximate Vietoris–Rips complexes in general metric spaces see [14, 26].

### 4.1 Linear-size approximation of the Čech persistence module

This section is an adaption of the work in [14] to the case of Čech complexes in Euclidean space. Throughout this section, $P \subseteq \mathbb{R}^n$.

**Definition 16.** For a set of points $P$, we say that $P' \subseteq P$ is a $\delta$-net of $P$ if

1. for every $p \in P$ there exists a $p' \in P'$ such that $\|p - p'\| \leq \delta$

2. for any $p, q \in P'$, $\|p - q\| > \delta$.

Choose parameters $\varepsilon_0, \epsilon \geq 0$ and define a sequence of point sets $P_k$ for $k = 0, 1, \ldots, m$ such that $P_0 = P$ and $P_{k+1}$ is an $\varepsilon_0 \alpha^2 (1 + \epsilon)^{k-1}$-net of $P_k$. We refer to such a collection $P_0, \ldots, P_m$ as a net-tree. Furthermore, let $\mathcal{C}(P_k; \alpha)$ be the Čech complex at scale $\alpha$ built upon the vertex set $P_k$, and $U(P_k; \alpha)$ the union of open balls of radius $\alpha$ centered at each point in $P_k$. We clearly have maps $\pi_k : P_k \to P_{k+1}$ which send a vertex $p \in P_k$ to its most nearby vertex in $P_{k+1}$.
Lemma 17. For every $k = 0, \ldots, m - 1$ we have inclusions
\[ U(P; \alpha_0(1 + \epsilon)^k) \subseteq U(P_{k+1}; \alpha_0(1 + \epsilon)^{k+1}). \]

Proof. Let $p \in P = P_0$ and $x \in \mathbb{R}^n$ be any point such that $||p - x|| < \alpha_0(1 + \epsilon)^k$. Since $P_1$ is an $\alpha_0 \epsilon^2 (1 + \epsilon)^{-1}$-net of $P$ we can find $\pi_0(p) \in P_1$ such that $||\pi_1(p) - p|| \leq \alpha_0 \epsilon^2 (1 + \epsilon)^{-1}$. Similarly, we can find $p' = (\pi_k \circ \cdots \circ \pi_0)(p) \in P_{k+1}$ such that
\[ ||p' - x|| \leq ||\pi_k \circ \cdots \circ \pi_0(p) - x|| + \alpha_0 \epsilon^2 (1 + \epsilon)^{k-1} \]
\[ \leq ||p - x|| + \sum_{i=0}^{k} \alpha_0 \epsilon^2 (1 + \epsilon)^{i-1} \]
\[ \leq ||p - x|| + \frac{\alpha_0 \epsilon^2}{1 + \epsilon} \cdot \frac{(1 + \epsilon)^{k+1} - 1}{\epsilon} \]
\[ < \alpha_0 (1 + \epsilon)^k + \alpha_0 \epsilon (1 + \epsilon)^k = \alpha_0 (1 + \epsilon)^{k+1}. \]

An immediate consequence is that $\pi_k : P_k \to P_{k+1}$ is a map of covers
\[ \pi_k : \mathcal{U}(P_k; \alpha_0(1 + \epsilon)^k) \to \mathcal{U}(P_{k+1}; \alpha_0(1 + \epsilon)^{k+1}). \]

Using this we define a sequence of covers associated to the net tree by defining $U^\text{net}(P_k; \alpha) = U(P_k; \alpha_0(1 + \epsilon)^k)$ where $k$ is the greatest integer such that $\alpha_0(1 + \epsilon)^k \leq \alpha$. The maps between the covers are given by compositions of $\pi_k$’s. We will denote the induced sequence of nerves by $C^\text{net}(P)$ and the persistence module by $(H_p(C^\text{net}(P)), \pi_*)$.

Proposition 18. The persistence modules $(H_p(C^\text{net}(P)), \pi_*)$ and $(H_p(C(P)), \id_*)$ are $(1 + \epsilon)^2$-approximate.

Proof. Using that $U^\text{net}(P_k; \alpha) = U(P_k; \alpha_0(1 + \epsilon)^k)$ together with Lemma 17 we have the chain of inclusions
\[ U^\text{net}(\alpha) \subseteq U(P; \alpha) \subseteq U(P; \alpha_0(1 + \epsilon)^{k+1}) \subseteq U(P_{k+2}; \alpha_0(1 + \epsilon)^{k+2}) = U^\text{net}(\alpha(1 + \epsilon)^2). \]

The rest of the proof follows by applying Theorem 14 with $\mathcal{U} = U^\text{net}(P)$ and $\mathcal{V} = \mathcal{W} = \mathcal{U}(P)$.

Proposition 19. Let $P \subseteq \mathbb{R}^n$ be a set of $m$ points. Then the number of $p$-simplices in $C^\text{net}(P; (1 + \epsilon)^k)$ is $O((\frac{1}{\epsilon})^{O(np)} m)$.

Proof. This is Theorem 6.3 in [16] together with the fact that the doubling dimension of $\mathbb{R}^n$ is $O(n)$.

The net-tree construction exhibits great theoretical results both on the level of approximating the Čech persistence module and in terms of size complexity. In practice however, as we shall see in Section 5 the complex often grows too large to be stored. Not doing a single collapse between scale $\alpha_0(1 + \epsilon)^k$ and scale $\alpha_0(1 + \epsilon)^{k+1}$ will in many situations introduce too many new simplices. To mend this we introduce a complex which allows for more numerous collapses, at the expense of computation time and poorer error bounds.
4.2 Approximations through non-good covers

We propose a general framework to approximate persistence modules associated to sequences of good covers. Using this framework we give an explicit approximation of the Čech persistence module in Euclidean space.

Let \((\mathcal{U}, F)\) be a sequence of covers with index sets \(\{I(\alpha)\}_{\alpha \geq 0}\) and \(J(I(\alpha))\) a partition of \(I(\alpha)\). We make the following assumption on the partitions: if \(J \in J(I(\alpha))\) then for all \(\alpha' \geq \alpha\) there exists \(J' \in J(I(\alpha'))\) such that \(J \subseteq J'\). In other words, if two elements are partitioned together at some scale \(\alpha\), they will be partitioned together at all scales \(\alpha' \geq \alpha\). Moreover, if \(J = \{j_0, \ldots, j_k\} \in J(I(\alpha))\) then \(F_\alpha(J)\) denotes the set \(J' \in J(I(\alpha'))\) such that \(J \subseteq J'\).

**Lemma 20.** For each \(\alpha \geq 0\), let \(J(I(\alpha))\) be a partition of \(I(\alpha)\) as described above. Then the pair

\[
(\tilde{\mathcal{U}}, F), \quad \tilde{\mathcal{U}}(\alpha) = \left\{ \tilde{U}_J(\alpha) = \bigcup_{j \in J} U_j(\alpha) \mid J \in J(I(\alpha)) \right\}
\]

is a sequence of covers.

**Proof.** This follows from that \(J \subseteq F_\alpha'(J)\) for all \(J \in J(I(\alpha))\). \(\square\)

For such a choice of partitions we say that \((\tilde{\mathcal{U}}, F)\) is a coarsening of \((\mathcal{U}, F)\).

Let \((\tilde{\mathcal{U}}(P), \text{id})\) be any coarsening of the Čech sequence of covers \(\mathcal{U}(P)\) on a finite point set \(P \subseteq \mathbb{R}^n\). Furthermore, define an associated sequence of good covers \((\text{Conv}\tilde{\mathcal{U}}, \text{id})\) where

\[
\text{Conv}\tilde{\mathcal{U}}(\alpha) = \left\{ \text{Conv}\left(\tilde{U}_k(\alpha)\right) \mid \tilde{U}_k(\alpha) \in \tilde{\mathcal{U}}(\alpha) \right\},
\]

and \(\text{Conv}(-)\) denotes the convex hull. In the following proposition \((\tilde{\mathcal{C}}(P), \text{id})\) denotes the induced sequence of nerves of \((\tilde{\mathcal{U}}(P), \text{id})\).

**Proposition 21.** If there exists a constant \(c \geq 1\) such that \(\text{Conv}\left(\tilde{U}_J(\alpha)\right) \subseteq \bigcup_{j \in J} U_j(c\alpha)\) for all \(\alpha \geq 0\) and \(J \in J(\alpha)\), then the persistence modules \((H_p(\mathcal{C}(P)), \text{id}_*)\) and \((H_p(\tilde{\mathcal{C}}(P)), \text{id}_*)\) are \(c\)-approximate.

**Proof.** Using Theorem 14 we see that the inclusion condition is satisfied by assumption:

\[
U(\alpha) \subseteq U(\alpha) \subseteq \text{Conv}U(\alpha) \subseteq U(c\alpha).
\]

Moreover, \(\tilde{U}(P; \alpha)\) and \(\text{Conv}\tilde{U}(P; \alpha)\) have the same indexing set, so the coherence relations of Equation 4 are trivially satisfied. \(\square\)

We see that every time we make our cover coarser, the number of 0-simplices in the nerve is reduced, and hence so is the size of the simplicial complex.

4.2.1 An explicit approximation

In the previous section we provided a general framework for constructing \(c\)-approximations to the Čech persistence module. We now give an explicit construction using Proposition 12.

**Lemma 22.** Let \(P = \{p_0, p_1, \ldots, p_k\} \subset \mathbb{R}^n\) where \(p_0 = 0\) and \(\|p_i\| \leq \alpha\) for all \(i\). Then for any point \(x \in \text{Conv}(P)\) there exists \(p_i \in P\) such that \(\|x - p_i\| \leq \alpha/\sqrt{2}\).
As an example, if implying that for some computationally expensive. The next section details one method for doing so.

\[ \alpha \]

the smallest memory efficient construction comes at the expense of computing weights of simplices.

\[ \text{such that } \|x - p_i\| > 1/\sqrt{2} \]

Using the law of cosines:

\[ 1 = \|p_i\|^2 = \|(p_i - x) + x\|^2 = \|p_i - x\|^2 + \|x\|^2 - 2\|p_i - x\| \cdot \|x\| \cos(\angle p_0xp_i) \]

\[ > \frac{1}{2} + \frac{1}{2} - 2\|p_i - x\| \cdot \|x\| \cos(\angle p_0xp_i) \]

\[ = 1 - 2\|p_i - x\| \cdot \|x\| \cos(\angle p_0xp_i) \]

implying that \( \cos(\angle p_0xp_i) > 0 \). But then

\[ (p_i - x) \cdot (x - x) = -p_i \cdot x_1 + x : 1^2 = \|p_i - x\| \cdot \|x\| \cos(\angle p_0xp_i) > 0 \]

and therefore \( p_i < x_1 \), contradicting that \( x \) was enclosed in the convex hull of \( P \).

**Proposition 23.** Let \( P = \{p_0, p_1, \ldots, p_k\} \subset \mathbb{R}^n \) be a set of points such that \( \|p_i - p_j\| \leq \epsilon \alpha \) for some \( \epsilon \geq 0 \). Then for all \( \alpha \geq 0 \) we have the following relation

\[ \text{Conv} \left( \bigcup_{0 \leq i < k} B(p_i; \alpha) \right) \subseteq \bigcup_{0 \leq i < k} B(p_i; \alpha \sqrt{1 + \frac{\epsilon^2}{2}}) . \]

**Proof.** First, observe that we have the following equality

\[ \text{Conv} \left( \bigcup_{0 \leq i < k} B(p_i; \alpha) \right) = \{ x \in \mathbb{R}^n \mid \exists y \in \text{Conv} P, \|x - y\| < \alpha \} . \]

Now let \( x \in \mathbb{R}^n \) be a point such that \( \|x - p\| < \alpha \) for some \( p \in \text{Conv} \{p_0, \ldots, p_k\} \) and \( k \leq n - 1 \). We will prove the proposition by induction on \( k \). If \( k = 0 \) the result follows trivially, so assume that it holds true for \( k < n - 1 \). Let \( x' \) be the orthogonal projection of \( x \) down on the affine space spanned by \( \{p_0, \ldots, p_k\} \). If \( p' \in \text{Conv} \{p_0, \ldots, p_k\} \) it follows from Lemma 22 that there exists a \( p_{i_j} \) such that

\[ \|p_{i_j} - x\|^2 = \|p_{i_j} - x'\|^2 + \|x' - x\|^2 \leq \frac{\epsilon^2 \alpha^2}{2} + \frac{\alpha^2}{2} \]

\[ = \alpha^2 \left( 1 + \frac{\epsilon^2}{2} \right) . \]

If \( x' \not\in \text{Conv} \{p_0, \ldots, p_k\} \) it implies the existence of a point \( p' \) on the boundary of \( \text{Conv} \{p_0, \ldots, p_k\} \) such that \( \|x - p'\| \leq \|x - p\| < 1 \) and we are done by induction.

By combining Propositions 21 and 23 we have shown the following.

**Proposition 24.** If \( \tilde{U} \) is a coarsening of \( U \) such that for all \( i, i' \in J \in J(I(\alpha)) \) we have that \( \|p_i - p_{i'}\| \leq \alpha \cdot \epsilon \) for some \( \epsilon \geq 0 \), then \( H_p(\tilde{C}(P), \text{id}_*) \) is a \( \sqrt{1 + \frac{\epsilon^2}{2}} \)-approximation of the \( \check{C} \)ech persistence module built on \( P \).

The previous proposition allows us to build good approximations to the \( \check{C} \)ech persistence module with far fewer simplices. A problem with this approach is that such a memory efficient construction comes at the expense of computing weights of simplices. As an example, if \( J(I(\alpha)) \) consists of \( k \) partitions, each with \( m \) elements, then computing the smallest \( \alpha \) at which they have a \( k \)-intersection has time complexity \( O(m^k) \). To mend this we seek methods to approximate this persistence module by ones that are less computationally expensive. The next section details one method for doing so.
4.2.2 Choosing a representative

In the following $\tilde{U}$ refers to the construction in Section 4.2.1. For every $\alpha \geq 0$ and every $J \in J(\alpha)$ choose a representative $j \in J$. Then the set map $\text{id}^\alpha_{\alpha'}: J(I(\alpha)) \to J(I(\alpha'))$ sends a representative at scale $\alpha$ to its representative at scale $\alpha'$. Denote the set of representatives at scale $\alpha$ by $P_\alpha$ and let $C^{\text{rep}}(P_\alpha; \alpha)$ be the minimal simplicial complex containing $C(P_\alpha; \alpha)$ such that that $\text{id}: C(P_\alpha; \alpha') \to C^{\text{rep}}(P_\alpha; \alpha)$ is a simplicial map for all $\alpha \leq \alpha'$. Equivalently, $C^{\text{rep}}(P_\alpha; \alpha)$ is the union $\cup_{\alpha \leq \alpha'} \text{id}^{\alpha'}_{\alpha}(C(P_\alpha, \alpha))$ where $\text{id}^{\alpha'}_{\alpha}(C(P_\alpha; \alpha))$ is the image of the map $\text{id}^{\alpha'}_{\alpha}: \tilde{C}(P; \alpha) \supseteq C(P_\alpha; \alpha) \to \tilde{C}(P; \alpha')$.

**Proposition 25.** The persistence modules $(H_p(\tilde{C}), \text{id}_\epsilon)$ and $(H_p(C^{\text{rep}}), \text{id}_\epsilon)$ are $\frac{1}{1+\epsilon}$-approximate.

**Proof.** The simplicial complexes $C^{\text{rep}}(P_\alpha; \alpha)$ and $\tilde{C}(P; \alpha)$ are defined over the same indexing set $J(I(\alpha))$ for every $\alpha \geq 0$. Clearly, $C^{\text{rep}}(P_\alpha; \alpha) \subseteq \tilde{C}(P; \alpha)$, so if we can find a $c$ such that $\text{id}^{\alpha'}_{\alpha}: \tilde{C}(P; \alpha) \to C^{\text{rep}}(P; \alpha)$ is well-defined, then the result follows from Definition 3 and Theorem 3. Let $x \in U_j \in \hat{U}(\alpha)$, where $\|x - p_j\| < \alpha$ for some $j \in J$ and let $p$ be the representative of $\text{id}^{\alpha/(1-\epsilon)}(J) \in J(I(\alpha/(1-\epsilon)))$. Then

$$\|p - x\| \leq \|p - p_j\| + \|p_j - x\| < \frac{\alpha\epsilon}{1-\epsilon} + \alpha = \frac{\alpha}{1-\epsilon}.$$ 

Hence, the composition $\tilde{C}(P; \alpha) \subseteq C(P; \frac{1}{1+\epsilon}) \to C^{\text{rep}}(P_\alpha/(1-\epsilon); \frac{\alpha}{1+\epsilon})$ is well-defined. \hfill \Box

5 Computational experiments

This section details our implementation of the approximation schemes described above, as well as some computational examples examining its efficacy and practical applicability.

5.1 Implementation

We realize an implementation of the approximation scheme detailed in Section 4.2.1 as a C++ program in the following way.

The program takes as parameters $\epsilon \geq 0$ (as in Section 4.2.1), a maximal scale $\alpha_{\text{max}} > 0$ (as usual when computing persistence), a maximal simplex dimension $D > 0$ (as usual) and $L \in \mathbb{N}$ (to be explained later). Given an input point cloud $P = \{p_1, \ldots, p_N\} \subseteq \mathbb{R}^d$, we first use Müllner’s fastcluster 22 to compute its hierarchical clustering $HC(P)$ with the complete linkage criterion. This is considered a preprocessing step.

A cluster is a pair $(p, X)$ with $p \in P$ and $X \subseteq P$ nonempty, wherein $p$ will be called the cluster’s representative and $X$ its members. At initialization time, we begin with $N$ clusters

$$c_0^i = (p_1, \{p_1\}), c_0^2 = (p_2, \{p_2\}), \ldots, c_0^N = (p_N, \{p_N\}).$$

and denote their enumeration by $C^0 = \{1, \ldots, N\}$.

We shall regard $HC(P)$ as the data of a series of linkage events of the form $(s, i, j) \in \mathbb{R} \times N \times N$ ordered by the first component, and (arbitrarily) with the convention that $i < j$. An event like this signifies the linking of clusters $c_i^l = (p_i^l, X_i^l)$ and $c_j^l = (p_j^l, X_j^l)$ at scale $s$, from which we form a new cluster $c_i^{l+1} = (p_i^{l+1}, X_i^{l+1})$ where $p_i^{l+1} \in X_i^{l+1}$; in principle the new representative $p_i^{l+1}$ can be chosen arbitrarily from $X_i^l \cup X_j^l$, but for heuristic reasons we pick the point in the member set $X_i^l \cup X_j^l$ closest to that set’s centroid.
We maintain a priority queue $Q$ of simplices prioritized by their persistence time. At initialization, the queue contains the 0-simplices $[1], \ldots, [N]$ all at persistence time 0. A simplex tree, along with associated annotations and other data structures as described in Section 2.3 are also initialized empty. These data structures that track homology will jointly be referred to as PH below, and we shall abuse language and speak of a simplex as “belonging to PH” when the simplex is present in the simplicial complex. We also initialize $\alpha' = 0$ and $l = 0$ to begin with.

The implementation code then proceeds in the following steps:

1. If $Q$ is empty, we are done and go to step 6. If not, pop a simplex $\sigma$ and its persistence scale $\alpha$ from the front, and continue.

2. If $\alpha > \alpha_{\text{max}}$, we are done and go to step 6. Otherwise continue.

3. If $\sigma$ is not already in PH, add it according to Section 2.3. In both cases, continue.

4. If $\dim \sigma > D$, go to step 5. Otherwise, for each simplex $\tau \in \{\sigma \cup \{i\} \mid i \in C_l\}$: compute the radius $r_\tau$ of the smallest enclosing ball of the set $\{p_i \mid i \in \tau\} \subseteq P$, and add $\tau$ to $Q$ at persistence scale $r_\tau$. Go to step 5.

5. If at least $L$ simplices have been added to PH since the last time this step was reached, we (possibly) perform a simplification by going to step 5a. Otherwise go to step 1.

   (a) For each linkage event $(s, i, j) \in HC(P)$ for which $s \in [\alpha', \epsilon\alpha)$, perform the edge contraction $[i, j] \mapsto [i]$ according to Section 2.3, taking care to adjust persistence times to reflect a (possible) series of inclusions to satisfy the link condition. If there were no linkage events in the given interval, go to step 1. Otherwise, denote the clusters present after handling linkage event, as explained earlier in this section, by

$$\{C^{l+1}_{i_1}, \ldots, C^{l+1}_{i_{N_l+1}}\} \subseteq \{C^d_{j_1}, \ldots, C^d_{j_{N_l}}\}$$

and go to step 5b.

   (b) Clear $Q$ and reset it to contain the 0-simplices $[i_1], \ldots, [i_{N_l}]$, all at persistence scale 0. Update $l$ to $l + 1$ and $\alpha'$ to $\epsilon\alpha$, and go to step 1.

6. We are done. Any persistent homology generators not yet killed off are recorded as on the form $(b, \infty)$.

The algorithm above may be summarized as follows: Compute Čech persistence until the underlying simplicial complex has at least $L$ simplices. When that is the case, walk up the complete linkage dendrogram of the point cloud until scale $\epsilon\alpha$ is reached, where $\alpha$ is the persistence scale. Any linkage event encountered corresponds to an edge contraction, which is performed. After that, computation of Čech persistence resumes as before, albeit on a reduced and changed point cloud, and collapses may happen again when $L$ more simplices have been added. We terminate upon reaching $\alpha_{\text{max}}$, and ignore simplices of dimension above $D$ (thus computing homology in dimensions $0, \ldots, D - 1$).

Note that $L$ is merely a parameter to reduce computational overhead involved in the collapses, as a higher value postpones contractions until the simplicial complex is denser. In principle, $L$ can be thought of as zero. Also observe that $\epsilon = 0$ corresponds to computing ordinary Čech persistence.

\footnote{Our implementation uses Gärtner’s Miniball \cite{19} for this computation.}
5.2 Experiments

This section describes three experiments designed to test the feasibility of our implementation.

A calculation ranging from scale 0 to scale $\alpha_{\text{max}}$ will have its resulting persistence diagram drawn as the region above the diagonal in $[0, \alpha_{\text{max}}]^2$. Generators still alive at $\alpha_{\text{max}}$ will be referred to as on the form $(b, \infty)$ and plotted as triangles, while generators of the form $(b, d)$ with $b < \alpha_{\text{max}}$ will be plotted as dots. See Figure 4 for an example of drawing conventions.

5.2.1 Wedge of six circles enclosing eachother

We produced a point cloud by randomly sampling 100 points from a circle of radius 1 centered at $(0, 1)$, 200 points from a circle of radius 2 centered at $(0, 2)$ and so forth up to 600 points from a circle of radius 6 centered at $(0, 6)$. Each point was also subjected to some radial noise. The very dense region near the origin where all the circles meet (see Figure 3) contributes nothing to homology, but significantly adds to the number of simplices if no collapse is done.

Running to $\alpha_{\text{max}} = 2$, our implementation clearly limited the number of simplices — see Figure 5 and note especially the rapid increases between collapses, the regimes where the ordinary Čech filtration is formed — while producing a highly correct persistence diagram, as is shown in Figure 4.

![Figure 3: The point cloud from the example in Section 5.2.1.](image)

5.2.2 The real projective plane

We sampled $\mathbb{R}P^2$ by randomly selecting 5000 points on $S^2$ and embedding them in $\mathbb{R}^4$ under $(x, y, z) \mapsto (xy, xz, y^2 - z^2, 2yz)$ as a test of how well our scheme handles higher dimensions. Figure 6 shows that the expected persistence diagram resulted when computing to $\alpha_{\text{max}} = 0.54$ at $\epsilon = 1.0$.

Figure 7 compares our scheme (at $\epsilon = 1$) with the very beginning an ordinary Čech filtration. Our implementation keeps the number of simplices manageable, peaking at just above $3 \cdot 10^5$ simplices near the end (scale 0.54), while still recovering the correct persistence diagram. The figure also shows the simplex count for the net tree construction; notice that we were unable to correctly choose $\alpha_0$ and $\epsilon$ so as to make computations with it feasible, unlike for the example in Section 5.2.1.
Figure 4: Persistence diagrams of the (noisy) wedge of six circles in Section 5.2.1 with $\epsilon = 3/4$ and $\alpha_{\text{max}} = 2$.

Figure 5: The simplex count while computing persistence for the example in Section 5.2.1. The net tree computations were run with $\alpha_0 = 10^{-3}$ and $\epsilon = 0.7$ in the notation of Section 4.1.
Figure 6: Persistence diagrams for the 5000 point random sample of $\mathbb{R}P^2$ embedded in $\mathbb{R}^4$ as described in Section 5.2.2, with $\epsilon = 1.0$ and $\alpha_{\text{max}} = 0.54$.

Figure 7: The simplex count for the $\mathbb{R}P^2$ example from Section 5.2.2 compared to that of an ordinary Čech filtration and the net tree approach (with $\alpha_0 = 10^{-3}$ and $\epsilon = 0.7$ in the notation of Section 4.1).
5.2.3 Time-delay embedding

We solved the Lorenz system (with parameters \(\sigma = 10, \ r = 28, \ b = 8/3\) in the notation of [23]) and created a time series \(y \in \mathbb{R}^{15000}\) by adding together all three of the solution’s coordinates at each of 15000 points in time. Let \(A(i)\) denote the (discrete) correlation of \(y\) and \(y\) shifted \(i\) places to the right. The first local minimum of \(A\) occurs at 45, so that was used as delay to embed \(y\) in \(\mathbb{R}^3\) by delay-embedding. The resulting point cloud, with 15000 − 45 = 14955 points, reconstructs [27] the Lorenz attractor as seen in Figure 8.

Observe that there are regions that have a very high density of points. Our implementation computes the expected persistence diagram (Figure 9) while keeping the number of simplices low (Figure 10).

![Figure 8: Lorenz system scalar measurements (parts shown on the left) and delay-embedding reconstructed attractor (right), as detailed in Section 5.2.3.](image1)

![Figure 9: Persistence diagrams for the Lorenz attractor described in Section 5.2.3.](image2)

6 Conclusions and future work

We have presented two approximation schemes for the Čech filtration in Euclidean space. One construction uses a net-tree to build the Čech complex at fewer and fewer simplices as we increase the scale parameter. The other approach forms a coarsening of the Čech filtration by using covering sets formed by unions of open balls. Computing \(k\)-intersections of such covering sets is computationally expensive, so we approximated the
perspective module by choosing a representative at each scale. In practice we experienced far better results with this method than the net-tree approach. This contrasts with the superior theoretical guarantees enjoyed by the net-tree construction. By approximating the Čech filtration through representatives we lose much of the theoretical guarantees, but we allow for frequent collapses allowing for much greater maximum scales.

We believe that an interesting direction for future work is to find other approximations than choosing a representative for each covering set. This could be done either by choosing multiple representative points, or by using the embedding to approximate the covering sets by sets for which computing $k$-intersections is tractable.

The proofs in this paper also rely heavily on the notion of good covers. In general metric spaces a cover by a union of balls may fail to be good, and the Nerve lemma is lost. It would be interesting to see if there are similar results without this precondition. We believe it should be so, as the net-tree construction for the Vietoris–Rips filtration extends to general metric spaces.

References

[1] Javier Arsuaga, Nils A Baas, Daniel DeWoskin, Hideaki Mizuno, Aleksandr Pankov, and Catherine Park. Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer. Applicable Algebra in Engineering, Communication and Computing, 23(1-2):3–15, 2012.

[2] Dominique Attali, André Lieutier, and David Salinas. Efficient data structure for representing and simplifying simplicial complexes in high dimensions. International Journal of Computational Geometry & Applications, 22(04):279–303, 2012.

[3] Gorô Azumaya. Corrections and supplementaries to my paper concerning krrl-remak-schmidt’s theorem. Nagoya Mathematical Journal, 1:117–124, 1950.

[4] Jean-Daniel Boissonnat, Tamal K Dey, and Clément Maria. A space and time efficient implementation for computing persistent homology. INRIA Research Report, (8195), 2012.

[5] Jean-Daniel Boissonnat and Clément Maria. The simplex tree: an efficient data structure for general simplicial complexes. INRIA Research Report, (7993), 2012.
[6] Gunnar Carlsson, Tigran Ishkhanov, Vin Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

[7] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[8] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. *ArXiv e-prints*, July 2012.

[9] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, SCG ’09, pages 237–246, 2009.

[10] Frédéric Chazal and Steve Yann Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the Twenty-fourth Annual Symposium on Computational Geometry*, SCG ’08, pages 232–241, New York, NY, USA, 2008. ACM.

[11] W. Crawley-Boevey. Decomposition of pointwise finite-dimensional persistence modules. *ArXiv e-prints*, October 2012.

[12] Vin De Silva and Gunnar Carlsson. Topological estimation using witness complexes. In *Proceedings of the First Eurographics Conference on Point-Based Graphics*, SPBG’04, pages 157–166, Aire-la-Ville, Switzerland, Switzerland, 2004. Eurographics Association.

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

[14] T. K. Dey, F. Fan, and Y. Wang. Computing Topological Persistence for Simplicial Maps. *ArXiv e-prints*, 2012.

[15] Tamal K Dey, Herbert Edelsbrunner, Sumanta Guha, and Dmitry V Nekhayev. Topology preserving edge contraction. *Publ. Inst. Math. (Beograd) (NS)*, 66(80):23–45, 1999.

[16] Tamal K. Dey, Fengtao Fan, and Yusu Wang. Computing topological persistence for simplicial maps. *ArXiv e-prints*, 2012. [1208.5018v3](http://arxiv.org/abs/1208.5018v3)

[17] Herbert Edelsbrunner. The union of balls and its dual shape. In *Proceedings of the Ninth Annual Symposium on Computational Geometry*, SCG ’93, pages 218–231, New York, NY, USA, 1993. ACM.

[18] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010.

[19] Bernd Gärtner. Fast and robust smallest enclosing balls. In *Algorithms-ESA ’99*, pages 325–338. Springer, 1999. [http://www.inf.ethz.ch/personal/gaertner/miniball.html](http://www.inf.ethz.ch/personal/gaertner/miniball.html)

[20] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 1 edition, December 2001.

[21] Michael Kerber and R. Sharathkumar. Approximate cech complex in low and high dimensions. In *ISAAC*, pages 666–676, 2013.
[22] Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, (9):1–8, 2013.

[23] Encyclopedia of Matematics. Lorenz attractor. [http://www.encyclopediaofmath.org/index.php?title=Lorenz_attractor&oldid=12339](http://www.encyclopediaofmath.org/index.php?title=Lorenz_attractor&oldid=12339).

[24] Jose Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *arXiv preprint arXiv:1307.6188*, 2013.

[25] JoseA. Perea and Gunnar Carlsson. A klein-bottle-based dictionary for texture representation. *International Journal of Computer Vision*, 107(1):75–97, 2014.

[26] Donald Sheehy. Linear-size approximations to the victoris-rips filtration. *Discrete & Computational Geometry*, 49(4):778–796, 2013.

[27] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, 1981.