Approximate Čech complexes in low and high dimensions

Michael Kerber[*] R. Sharathkumar[†]

Abstract

Čech complexes reveal valuable topological information about point sets at a certain scale in arbitrary dimensions, but the sheer size of these complexes limits their practical impact. While recent work introduced approximation techniques for filtrations of (Vietoris-)Rips complexes, a coarser version of Čech complexes, we propose the approximation of Čech filtrations directly.

For fixed dimensional point set S, we present an approximation of the Čech filtration of S by a sequence of complexes of size linear in the number of points. We generalize well-separated pair decompositions (WSPD) to well-separated simplicial decomposition (WSSD) in which every simplex defined on S is covered by some element of WSSD. We give an efficient algorithm to compute a linear-sized WSSD in fixed dimensional spaces. Using a WSSD, we then present a linear-sized approximation of the filteration of Čech complex of S.

We also present a generalization of the known fact that the Rips complex approximates the Čech complex by a factor of $\sqrt{2}$. We define a class of complexes that interpolate between Čech and Rips complexes and that, given any parameter $\varepsilon > 0$, approximate the Čech complex by a factor $(1+\varepsilon)$. Our complex can be represented by roughly $O(n^{\lceil 1/2\varepsilon \rceil})$ simplices without any hidden dependance on the ambient dimension of the point set. Our results are based on an interesting link between Čech complex and coresets for minimum enclosing ball of high-dimensional point sets. As a consequence of our analysis, we show improved bounds on coresets that approximate the radius of the minimum enclosing ball.

1 Introduction

Motivation A common theme in topological data analysis is the analysis of point cloud data representing an unknown manifold. Although the ambient space can be high-dimensional, the manifold itself is usually of relatively low dimension. Manifold learning techniques try to infer properties of the manifold, like its dimension or its homological properties, from the point sample.

An early step in this pipeline is to construct a cell complex from the point sample which shares similarities with the hidden manifold. The *Čech complex at scale* α (with $\alpha \geq 0$) captures the intersection structure of balls of radius α centered at the input points. More precisely, it is the

[*] Stanford University, Stanford, USA and Max Planck Center for Visual Computing and Communication, Saarbrücken, Germany. mkerber@mpi-inf.mpg.de

[†] Stanford University, Stanford, USA. sharathk@stanford.edu

nerve of these balls, and is therefore homotopically equivalent to their union. Increasing α from 0 to ∞ yields a *filtration*, a sequence of nested Cech complexes, which can serve as the basis of multi-scale approaches for topological data analysis.

A notorius problem with Čech complexes is their representation: Its k-skeleton can consist of up to $O(n^k)$ simplices, where n is the number of input points. Moreover, its construction requires the computation of minimum enclosing balls of point sets; we will explain this relation explicit in Section 2. A common workaround is to replace the Čech complex by the *(Vietoris-)Rips complex* at the same scale α. Its definition only depends on the diameter of point sets and can therefore be computed by only looking at the pairwise distances. Although Rips complexes permit a sparser representation, they do not resolve the issue that the final complex can consist of a large number of simplices; Sheehy [22] and Dey et al. [10] have recently adressed this problem by defining an approximate Rips filtration whose size is only linear in the input size. On the other hand, efficient methods for approximating minimum enclosing balls have been established, even for high-dimensional problems, whereas the diameter of point sets appears to be a significantly harder problem in an approximate context. This suggests that Čech complexes might be more suitable objects than Rips complexes in an approximate context.

Contribution We give two different approaches to approximate filtrations of Čech complexes, both connecting the problem to well-known concepts in discrete geometry: The first approach yields, for a fixed constant dimension, a sequence of complexes, each of linear size in the number of inut points, that approximate the Čech filtration. By approximate, we mean that the *persistence diagrams* of exact and approximate Čech filtration differ by a arbitrarily small multiplicative factor. To achieve this result, we generalize the famous *well-separated pair decomposition (WSPD)* to a higher-dimensional analogon, that we call the *well-separated simplical decomposition (WSSD)*. Intuitively, a WSSD decomposes a point set S into $O(n/\varepsilon^d)$ tuples. A k-tuple in the WSSD can be viewed as k clusters of points of S with the property that whenever a ball contains at least one point of each cluster, a small expansion of the ball contains all points in all clusters. Furthermore, these tuples cover every simplex with vertices in S, i.e., given any k-simplex σ, there is a $k+1$-tuple of clusters such that each cluster contains on vertex of σ. We consider the introduction of WSSDs to be of independent interest: given the numerous applications of WSPD, we hope that its generalization will find further applications in approximate computational topology. We finally remark that, similar to related work on the Rips filtration [22, 10], the constant in the size of our filtration depends exponentially on the dimension of the ambient space, which restricts the applicability to low- and medium-dimensional spaces.

As our second contribution, we prove a generalized version of the well-known Vietoris-Rips lemma [12, p.62] which states that the Čech complex at scale α is contained in the Rips complex at scale $\sqrt{2}\alpha$. We define a family of complexes, called *completion complexes* such that for any ε, the Čech complex at scale α is contained in a completion complex at scale $(1+\varepsilon)\alpha$. These completions complexes are parameterized by an integer k; the k-completion is completely determined by its k-skeleton, consisting of up to $O(n^k)$ complexes. To achieve $(1+\varepsilon)$-closeness to the Čech complex, we need to set $k \approx 1/(2\varepsilon)$ (see Theorem 27 for the precise statement); in particular, there is no dependence on ambient dimension to approximate the Čech complex arbitrarily closely.

For proving this result, we use *coresets* for minimum enclosing ball (meb) [3]: the meb of a set of points can be approximated by selecting only a small subset of the input which is called a *coreset*; here approximation means that an ε-expansion of the meb of the coreset contains all input points. The size of the smallest coreset is at most $\lceil 1/\varepsilon \rceil$, independent of the number of points and the ambient dimension, and this bound is tight [3]. To obtain our result, we relax the definition of coreset for minimum enclosing balls. We only require the *radius* of the meb to be approximated, not the meb itself. We prove that even smaller coresets of size roughly $\lceil 1/(2\varepsilon) \rceil$ always exist for approximating the radius of the meb. Again, we consider this coreset result to be of independent interest.

Related work Sparse representation of complexes based on point cloud data are a popular subject in current research. Standard techniques are the *alpha complex* [13, 14] which contains all Delaunay simplices up to a certain circumradius (and their faces), *simplex collapses* which remove a pair of simplices from the complex without changing the homotopy type (see [1, 19, 23] for modern references), and *witness approaches* which construct the complex only on a small subset of landmark points and use the other points as witnesses [9, 2, 11]. A more extensive treatment of some of these techniques can be found in [12, Ch.III]. Another very recent approach [21] constructs Rips complexes at several scales and connects them using *zigzag persistence* [5], an extension to standard persistence which allows insertions and deletions in the filtration. The aforementioned work by Sheehy [22] combines this theory with *net-trees* [16], a variant of hierachical metric spanners, to get an approximate linear-size zigzag-filtration of the Rips complex in a first step and finally shows that the deletions in the zigzag can be ignored. Dey et al. [10] arrive at the same result more directly by constructing an hierachical ε-net, defining a filtration from it where the elements are connected by simplicial maps instead of inclusions, and finally showing that this filtration is *interleaved* with the Rips-filtration in the sense of [6].

Outline We will introduce basic topological concepts in Section 2. Then we introduce WSSDs, our generalization of WSPDs and give an algorithm to compute them in Section 3. We show how to use WSSDs to approximates the persistence diagram of the Čech complex in Section 4. The existence of small coresets for approximating the radius of meb is the subject of Section 5. *k*-completions and the generalized Vietoris-Rips Lemma are presented in Section 6. We conclude in Section 7.

2 Preliminaries

Simplicial complexes Let S denote a finite set of universal elements, called *vertices*[1] A *(simplicial) complex* C is a collection of subsets of S, called *simplices*, with the property that whenever a simplex σ is in C, all its (non-empty) subsets are in C as well. These non-empty subsets are called the *faces* of σ; a *proper face* is a face that is not equal to σ. Setting $k := |\sigma| - 1$, where $|\cdot|$ stands for the number of elements considered as a subset, we call σ a *k-simplex*. Observe a k-simplex σ corresponds to a $(k+1)$-subset $(v_0, \ldots, v_k)$ of S; these $(k+1)$ vertices are called the *boundary vertices* of k-simplex, and we will frequently identify the simplex and its set of boundary vertices. A *subcomplex* of C is a simplicial complex that is contained in S. One example of a subcomplex

[1] The finiteness of S is not necessary for all defined concepts; however, since we will only deal with finite complexes in later sections, we decided to discuss this simpler setup.

is the *k-skeleton* of a complex C, which is the set of all ℓ-simplices in C with $\ell \leq k$. Let K and K' be two simplicial complexes with vertex sets V and V' and consider a map $f : V \to V'$. If for any simplex $(v_0, \ldots, v_k)$ of K, $(f(v_0), \ldots f(v_k))$ yields a simplex in K', then f extends to a map from K to K' which we will also denote by f; in this case, f is called a *simplicial map*.

Let the vertices S be a set of arbitrary geometric objects, embedded in an ambient space $\mathbb{R}^d$. We call $|S| := \cup_{s \in S} s \subset \mathbb{R}^d$ the *union of S*. We define a simplicial complex C as follows: A k-simplex σ is in C if the corresponding $k+1$ objects have a common intersection in $\mathbb{R}^d$. It is easy to check that C is indeed closed under face relations and thus a simplical complex, called the *nerve* of S. The famous *Nerve Theorem* [12] states that if all objects in S are convex, the union of S and its nerve are *homotopically equivalent*. This intuitively means that one can transform one into the other by bending, shrinking and expanding, but without glueing and cutting. A consequence of this theorem is that the *homology groups* of the union and the nerve are equal; see [12, 20] for thorough introductions to homology.

For a finite point set P and $\alpha > 0$, the *Cech complex* $\mathscr{C}_\alpha(P)$ is the nerve of the set of (closed) balls of radius α centered at the points in P. Note that a k-simplex of the Cech complex can be identified with $(k+1)$ points $p_0, \ldots, p_k$ in P, the centers of the intersecting balls. Let $\mathrm{meb}(p_0, \ldots, p_k)$ denote the *minimum enclosing ball of P*, that is, the ball with minimal radius that contains each p_i.

Observation 1. *A k-simplex $\{p_0, \ldots, p_k\}$ is in $\mathscr{C}_\alpha(P)$ iff the radius of* $\mathrm{meb}(p_0, \ldots, p_k)$ *is at most α.*

A widely used approximation of Cech complexes is the *(Vietoris)-Rips complex $\mathscr{R}_\alpha(P)$*. It is defined as the maximal simplicial complex whose 1-skeleton equals the 1-skeleton of the Čech complex. Described as an iterative construction, starting with the edges of the Čech complex, a triangle is added to the Rips complex when its three boundary edges are present, a tetrahedron when its four boundary triangles are present, and so forth. The Rips complex is an example of a *clique complex* (also known as *flag complex* or *Whitney complex*). That means, it is completely determined by its 1-skeleton which in turn only depends on the pairwise distance between the input points. For $k+1$ points $p_0, \ldots, p_k$ in P, let the *diameter* $\mathrm{diam}(p_0, \ldots, p_k)$ denote the maximal pairwise distance between any two points p_i and p_j with $0 \leq i \leq j \leq k$.

Observation 2. *A k-simplex $\{p_0, \ldots, p_k\}$ is $\mathscr{R}_\alpha(P)$ iff* $\mathrm{diam}(p_0, \ldots, p_k)$ *is at most α.*

For notational convenience, we will often omit the P from the notation and write $\mathscr{C}_\alpha$ and $\mathscr{R}_\alpha$ when P is clear from context.

Persistence modules For $A \subset \mathbb{R}$, a *persistent module* is a family $(F_\alpha)_{\alpha \in A}$ of vector spaces with homomorphisms $f_\alpha^{\alpha'} : F_\alpha \to F_{\alpha'}$ for any $\alpha \leq \alpha'$ such that $f_{\alpha'}^{\alpha''} \circ f_\alpha^{\alpha'} = f_\alpha^{\alpha''}$ and f_α^α is the identity function.[2] The most common class are modules induced by a *filtration*, that is, a familiy of complexes $(C_\alpha)_{\alpha \in A}$ such that $C_\alpha \subseteq C_{\alpha'}$ for $\alpha \leq \alpha'$. For some fixed dimension p, set $H_\alpha := H_p(C_\alpha)$, the p-th homology group of C_α. The inclusion map from C_α to $C_{\alpha'}$ induces an homomorphism $\hat{f}_\alpha^{\alpha'} : H_\alpha \to H_{\alpha'}$ and turns $(H_\alpha)_{\alpha \in \mathbb{R}}$ into a persistence module. Example of such filtrations and their induced modules are the *Cech filtration* $(\mathscr{C}_\alpha)_{\alpha \geq 0}$ and the Rips filtration $(\mathscr{R}_\alpha)_{\alpha \geq 0}$. However, we will also consider persistence modules which are not induced by filtrations. Generalizing the case of filtrations, given a sequence of simplicial complexes $(\mathscr{A}_\alpha)_{\alpha \in A}$ connected by simplicial

[2] This is not the most general definition of a persistent module; see [6].

maps $g_{\alpha}^{\alpha'} : \mathscr{A}_{\alpha} \to \mathscr{A}_{\alpha'}$ which satisfy $g_{\alpha'}^{\alpha''} \circ g_{\alpha}^{\alpha'} = g_{\alpha}^{\alpha''}$ and $g_{\alpha}^{\alpha} = \mathrm{id}$, the induced homology groups $H_{\alpha} := H_p(\mathscr{A}_{\alpha})$ and induced homomorphisms $\hat{g}_{\alpha}^{\alpha} : H_{\alpha} \to H_{\alpha'}$ also yield a persistence module. A persistence module $(F_{\alpha})_{\alpha \in A}$ is *tame* if the rank of F_{α} is finite for all $\alpha \in A$. As our modules in this work will consist only of homology groups over finite simplicial complexes, all modules constructed in this paper will be tame, and we will ignore this technicality from now on when referring to previous results. We will frequently denote filtrations and modules by F_* instead of $(F_{\alpha})_{\alpha \in A}$ for brevity if there is no confusion about A.

For a persistence module F_* with homomorphisms $f_{\alpha}^{\alpha'}$, we say that a generator (basis element) $\gamma \in F_{\alpha}$ is *born* at α if $\gamma \notin \mathrm{Im} f_{\alpha-\varepsilon}^{\alpha}$ for any $\varepsilon > 0$, where Im is the image of a map. If γ is born at α, we say that it *dies* at α' if α' is the smallest value such that $f_{\alpha}^{\alpha'}(\gamma) \in \mathrm{Im} f_{\alpha-\varepsilon}^{\alpha'}$ for some $\varepsilon > 0$. In other words, every generator can be represented by a point in the plane, determining its birth- and death-coordinate. F_* is completely characterized by this multiset of points, which is called the *persistence diagram* of the module and denote it as $\mathrm{Dgm}F_*$. Note that all points of the diagram lie on or above the diagonal in the birth-death-plane.

For the benfit of readers inexperienced with the concept of persistence, we explain the wealth of geometric-topological information contained in the persistence diagram, examplified on a Čech filtration of a point set S in $\mathbb{R}^3$. As discussed, we can visualize the filtration as a sequence of growing balls centered at the points in S, and the union of these balls forms a sequence of growing shapes. During this process, the shape might create *voids*, that is, pockets of air completely enclosed by the shape. The rank of the second homology group $H_2(\mathscr{C}_{\alpha})$ yields the number of voids present at scale α (this rank is also called the 2nd *Betti number*). The persistence diagram for $H_2(\mathscr{C}_*)$ provides multi-scale information about the voids in the process; every point (b,d) of the diagram represents a void that was formed for $\alpha = b$ and filled up for $\alpha = d$. The same information as for voids can be obtained for *connected components* and for *tunnels*, choosing the 0-th and 1-st homology groups, respectively.

Approximating persistence diagrams An important property of persistence diagrams is their stability under "small" perturbations of the underlying filtrations and modules; see Cohen-Steiner et al. [8] for the precise first statement of this type. We will use the more recent results by Chazal et al. [6] for this work, following Sheehy's notations and definitions [22]. For two modules F_*, G_*, we say that $\mathrm{Dgm}F_*$ is a *c-approximation* of $\mathrm{Dgm}\,G_*$ with $c \geq 1$ if there is a bijection $\pi : \mathrm{Dgm}\,F_* \to \mathrm{Dgm}\,G_*$ such that for any point (x,y) of $\mathrm{Dgm}\,F_*$, $\pi(x,y)$ lies in the axis-aligned box defined by $\frac{1}{c}(x,y)$ and $c(x,y)$. An equivalent statement is that the two diagrams have a bounded bottleneck distance on the log-scale.

We will use the following result which is a reformulation of [6, Def.4.2+Thm.44]:

Theorem 3. *Let $(F_{\alpha})_{\alpha \geq 0}$ and $(G_{\alpha})_{\alpha \geq 0}$ be two persistence module with two families of homomorphisms $\{\phi : F_{\alpha} \to G_{c\alpha}\}_{\alpha \geq 0}$ and $\{\psi : G_{\alpha} \to F_{c\alpha}\}_{\alpha \geq 0}$ such that all the following diagrams* *commute:*

(2.1)

$$\begin{tikzcd}
F_{\frac{\alpha}{c}} \arrow[rrr] \arrow[dr] & & & F_{c\alpha'} \\
& G_{\alpha} \arrow[r] & G_{\alpha'} \arrow[ur] &
\end{tikzcd}
\qquad
\begin{tikzcd}
& F_{c\alpha} \arrow[r] & F_{c\alpha'} \\
G_{\alpha} \arrow[ur] \arrow[r] & G_{\alpha'} \arrow[ur] &
\end{tikzcd}$$

$$\begin{tikzcd}
& F_{\alpha} \arrow[r] & F_{\alpha'} \arrow[dr] & \\
G_{\frac{\alpha}{c}} \arrow[ur] \arrow[rrr] & & & G_{c\alpha'}
\end{tikzcd}
\qquad
\begin{tikzcd}
F_{\alpha} \arrow[r] \arrow[dr] & F_{\alpha'} \arrow[dr] & \\
& G_{c\alpha} \arrow[r] & G_{c\alpha'}
\end{tikzcd}$$

Then, the persistence diagrams of F_α and G_α are c-approximations of each other.

In the case of modules induced by filtrations, there is a simple corollary, called the “Persistence Approximation Lemma” in [22]:

Lemma 4. *If two filtrations $(A_\alpha)_{\alpha\geq 0}$ and $(B_\alpha)_{\alpha\geq 0}$ satisfy $A_{\frac{\alpha}{c}} \subset B_\alpha \subset A_{c\alpha}$ for all $\alpha \geq 0$, then the persistence diagrams are c-approximations of each other.*

3 Well-separated simplicial decompositions

In this section, we introduce the notion of Well-separated simplicial decomposition (WSSD) of point sets. WSSD can be seen as a generalization of well-separated pair decomposition of a point set. We first revisit the definition of WSPD and then generalize it to WSSD.

Notations. Let $S \subset \mathbb{R}^d$ be a fixed point set with minimal distance $1/\sqrt{d}$ between two points and such that all points are contained in a axis-parallel hypercube q with side length 2^L. We consider a *quadtree* Q of q where each node represents a hypercube; the root represents q, and when an internal node represents a hypercube q', its children represent the hypercubes obtained by splitting q' into 2^d congruent hypercubes. From now on, we will usually identify the quadtree node and the hypercube that it represents. We call a node of Q *empty* if it does not contain any point of S. For any internal node q', the *height* of q' in Q is i if the side length of q' is 2^i; the construction ends at height 0; by construction, each leaf contains at most one point of S.[3] The nodes of Q at height i induce a grid G_i where the side length of every cell of G_i is 2^i. For $e > 0$ and a ball $\mathbb{B}$ with center c and radius r, we let $e\mathbb{B}$ denote the ball with center c and radius $e \cdot r$. We state the following property, which follows directly by triangle inequality, but is used several times in our arguments:

Observation 5. *Let $\mathbb{B}$ be a ball with radius r that intersects a convex object M whose diameter is at most λr. Then, $M \subseteq \lambda\mathbb{B}$.*

Finally, whenever we make statements that depend on a parameter ε, it is implicitely assumed that $\varepsilon \in (0,1)$ from now on.

[3] This “construction” is only conceptual; in an actual implementation, only non-empty would be stored. Moreover, the quadtree should be represented in *compressed* form to avoid dependance on the *spread* of the point set; see [15, §2] for details.

Well-Separated Pair Decomposition. Let Q be a quadtree for S. A pair of quadtree cells (q,q') is called *ε-well separated* if $\max(\mathrm{diam}(q),\mathrm{diam}(q')) \leq \varepsilon d(q,q')$; here $\mathrm{diam}(q)$ is the diameter of a quadtree cell (which equals $2^h\sqrt{d}$ if h is the height of q) and $d(q,q')$ is the closest distance between cells q and q'. We state a simple consequence which appears somewhat indirect, but allows a generalization to multivariate tuples:

Lemma 6. *If (q,q') is ε-well separated, any ball $\mathbb{B}$ that contains at least one point of q and one point of q', the ball $(1+2\varepsilon)\mathbb{B}$ contains all of q and all of q'.*

Proof. Let $\mathbb{B}$ be a ball with radius r intersecting both q and q', which means that $r \geq d(q,q')/2$. Because (q,q') is well-separated,

$$\mathrm{diam}(q) \leq \varepsilon d(q,q') \leq 2\varepsilon r,$$

implying that $(1+2\varepsilon)\mathbb{B}$ contains all of q by Observation 5. The same argument applies for q'. □

For a pair $(p,p') \in S \times S$ we say that a pair of quadtree cells (q,q') *covers* (p,p') if $p \in q$ and $p' \in q'$, or $p \in q'$ and $p' \in q$. An *ε-well separated pair decomposition* (ε-WSPD) of S is a set of pairs $\Gamma = ((q_1,q_1'),(q_2,q_2'),\ldots,(q_m,q_m'))$ such that all pairs are ε-well separated and every edge in $S \times S$ is covered by some pair in Γ. We rely on the following properties of WSPDs, proved first in [4]; see also [15, §3] for a modern treatement:

Theorem 7. *A ε-WSPD of size $O(n/\varepsilon^d)$ can be computed in $O(n\log n + n/\varepsilon^d)$ time.*

Well-Separated Simplicial decomposition. We generalize the construction of WSPD to higher dimensions: Let S and Q be as above. We call a $(k+1)$-tuple $(q_0,\ldots,q_k)$ of quadtree cells an *ε-well separated tuple* (ε-WST), if for any ball $\mathbb{B}$ that contains at least one point of each q_ℓ, we have that

$$q_0 \cup q_1 \cup \ldots q_k \subseteq (1+\varepsilon)\mathbb{B}. \tag{3.1}$$

Moreover, we say that $(q_0,\ldots,q_k)$ *covers* a k-simplex $\sigma = (p_0,\ldots,p_k)$, $p_0,\ldots,p_k \in S$ if there is a permutation π of $(0,\ldots,k)$ such that $p_{\pi(\ell)} \in q_\ell$ for all $0 \leq \ell \leq k$.

Definition 8. *A set of $(k+1)$-tuples $\Gamma = \{\gamma_1,\ldots,\gamma_m\}$ is a (ε,k)-*well separated simplicial decomposition *((ε,k)-WSSD), if each γ_ℓ is a ε-well separated tuple and each k-simplex of S is covered by some γ_ℓ. An ε-*WSSD *is the union of (ε,k)-WSSDs over all $1 \leq k \leq d$.*

It is easy to see with Lemma 6 that an $\frac{\varepsilon}{2}$-WSPD is an $(\varepsilon,1)$-WSSD.

Our algorithm. We present an iterative algorithm for computing an (ε,k)-WSSD. In the first iteration of our algorithm, we compute an $\frac{\varepsilon}{2}$-WSPD which is an $(\varepsilon,1)$-WSSD using the algorithm from [15, Fig. 3.3]. We now describe iteration $k>1$ of our algorithm, where we compute a (ε,k)-WSSD Γ_k using the $(\varepsilon,k-1)$-WSSD Γ_{k-1} from the previous iteration:

We initialize Γ_k as the empty set and iterate over the elements in Γ_{k-1}. For an ε-WST $\gamma = (q_0,q_1,\ldots q_{k-1}) \in \Gamma_{k-1}$, let $\mathbb{B}_\gamma = \mathrm{meb}(q_0 \cup q_1 \cup \ldots q_{k-1})$, and let r denote its radius. Consider the grid G_h formed by all quadtree cells of height h such that $2^h \leq \frac{\varepsilon r}{2\sqrt{d}} \leq 2^{h+1}$. We compute the set of non-empty quadtree cells G_h that intersect the ball $2 \cdot \mathbb{B}_\gamma$. For each such cell q', we add the $(k+1)$-tuple $(q_0,\ldots,q_{k-1},q')$ to Γ_k. See Figure 3.1 for an illustration.

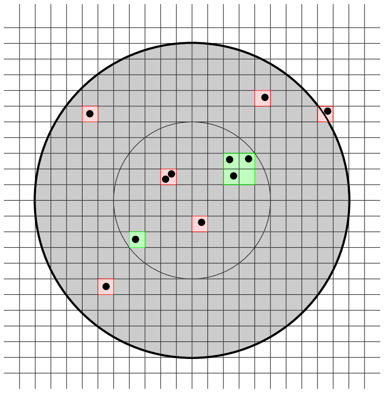

Figure 3.1: Example for the construction of Γ_2 from Γ_1: Let the pair of green boxes be a WST γ of Γ_1 (that is, a well-separated pair). Now, the algorithm creates a triple consisting of the two green boxes and any grid cell at height h that intersects $2\mathbb{B}_\gamma$ (shaded area). In this example, this would be 10 triples - 6 with the red boxes, and 4 additional ones coming from the non-empty boxes in the green areas.

Correctness. In order to prove the correctness of our construction procedure, we need to show that the generated tuples indeed form a (ε, k)-WSSD.

Lemma 9. *Every tuple added by our procedure is an ε-WST.*

Proof. We do induction on k, noting that for $k = 1$, the statement is true because an $\frac{\varepsilon}{2}$-WSPD is an $(\varepsilon, 1)$-WSSD. For $k \geq 2$, assume that our algorithm creates a k-tuple $(q_0, \ldots, q_{k-1}, q')$ by adding the cell q' while considering the ε-WST $(q_0, \ldots, q_{k-1})$. Let $\mathbb{B}$ be a ball that contains at least one point from each of the cells $(q_0, \ldots, q_{k-1}, q')$. We have to argue that $(1+\varepsilon)\mathbb{B}$ contains the cells $q_0, \ldots, q_{k-1}, q'$; by induction hypothesis, it is clear that $q_0 \cup \ldots \cup q_{k-1} \subseteq (1+\varepsilon)\mathbb{B}$ and moreover,

$$r = \operatorname{rad}(q_0, \ldots, q_{k-1}) \leq (1+\varepsilon)\operatorname{rad}(\mathbb{B}). \tag{3.2}$$

Finally, by construction,

$$\operatorname{diam}(q') \leq \frac{\sqrt{d}\varepsilon r}{2\sqrt{d}} \leq \frac{\varepsilon(1+\varepsilon)\operatorname{rad}(\mathbb{B})}{2} \leq \varepsilon \cdot \operatorname{rad}(\mathbb{B}),$$

so $q' \subseteq (1+\varepsilon)\mathbb{B}$ by Observation 5. □

For showing that all k-simplices are covered, we use the following result which is taken from [3] – we note that the required bound also follows as a simple corollary of the main result of [3], but we decided to give a more low-level argument for clarity.

Lemma 10. *Let P be a point set with $|P| \geq 3$. Then, there exists a point $p \in P$ such that*

$$p \in \frac{1+1/d}{\sqrt{1-1/d^2}}\operatorname{meb}(P \setminus \{p\}).$$

In particular, $p \in 2\operatorname{meb}(P - \setminus\{p\})$ for $d \geq 2$.

Proof. Note that the statement is trivial if there exists a point $p \in P$ whose removal does not change the minimum enclosing ball. Therefore, assume wlog that $|P| \leq d+1$, and all points of P are at the boundary of $\mathrm{meb}(P)$. Let c be the center and r be the radius of $\mathrm{meb}(P)$. The points in P span a polytope T; take the smallest ball $\mathbb{B}$ centered at c that is contained in T. By [3, Lem. 3.2], its radius is at most r/d. Moreover, $\mathbb{B}$ touches at least one facet of T. Let p be the point opposite of this facet, set $P' := P \setminus \{p\}$ and let c' and r' denote the center and radius of the meb of P'. Following the argumentation of [3, Lem. 3.3], it holds that

$$r' \geq r\sqrt{1-(1/d^2)}$$

and moreover, c' is the point where $\mathbb{B}$ touches the facet, so that $\|c-c'\| \leq r/d$. Now, by triangle inequality

$$\begin{aligned} \|p-c'\| &\leq \|p-c\| + \|c-c'\| \\ &\leq r + r/d \\ &\leq (1+1/d)\frac{r'}{\sqrt{1-1/d^2}} \end{aligned}$$

which implies the first claim. The second part follows easily by noting that

$$\frac{1+1/d}{\sqrt{1-1/d^2}} \leq 2$$

for all $d \geq 5/3$. □

Lemma 11. *The set of $(k+1)$-tuples Γ_k generated by our procedure covers all k-simplices over S.*

Proof. We do induction on k. For the base case $k=1$, by definition, all pairs of points in $S \times S$ are covered by some pair (q,q') in an $\frac{\varepsilon}{2}$-WSPD. Assume that the computed $(\varepsilon,k-1)$-WSSD covers all $(k-1)$-simplices and consider any k-simplex $\sigma = (p_0,\ldots,p_k)$. By Lemma 10, there exists a point among the p_i, say p_0, such that $p_0 \in 2\mathrm{meb}(\sigma')$, where $\sigma' = (p_1,\ldots,p_k)$. By induction hypothesis, there exists a ε-WST $t = (q_1,\ldots,q_k)$ that covers σ'. Clearly, $p_0 \in 2\mathrm{meb}(t)$ as well. Let q be the cell of G_h that contains p_0. By construction, our algorithm adds $(q_1,\ldots,q_k,q)$ to Γ_k, and this tuple covers σ. □

With Lemma 9 and Lemma 11, it follows that the constructed set Γ_k is an (ε,k)-WSSD.

Analysis. We bound the size of the (ε,k)-WSSD generated by our algorithm and the total time taken to compute it.

Lemma 12. *Let Γ_k be the (ε,k)-WSSD generated by our algorithm. Then, $|\Gamma_k| = n(d/\varepsilon)^{O(dk)}$.*

Proof. By Theorem 7, the size of the $(\varepsilon,1)$-WSSD (or $\frac{\varepsilon}{2}$-WSPD) is $O(n(d/\varepsilon)^{O(d)})$. Let us assume that the size of Γ_{k-1} is $O(n(d/\varepsilon)^{O(d(k-1))})$. It suffices to show that for every $\gamma \in \Gamma_{k-1}$, we add at most $O((d/\varepsilon)^d)$ ε-WSTs to Γ_k.

As in the algorithm, set $\mathbb{B}_\gamma := \mathrm{meb}(\gamma)$ and $r := \mathrm{rad}(\gamma)$. By construction, the side length of a cell in G_h is at least $\frac{\varepsilon r}{4\sqrt{d}}$. By a simple packing argument, the total number of cells of G_h that intersect $2\mathbb{B}_\gamma$ is $O((d/\varepsilon)^d)$. We add (at most) one ε-WST to Γ_k for each of these $O((d/\varepsilon)^d)$ cells. □

By Theorem 7, an $\frac{\varepsilon}{2}$-WSPD can be constructed in $O(n\log n+n(d/\varepsilon)^d)$ time. To construct Γ_k from Γ_{k-1}, for every $\gamma\in\Gamma_{k-1}$, our algorithm has to compute the meb $\mathbb{B}_\gamma$ of the involved cells and find all cells in G_h that intersect $2\mathbb{B}_\gamma$. This can be done, for instance, by finding the cell q that contains the center of $\mathbb{B}_\gamma$ and traverse the cells in increasing distance from q. All these operations can be done in time proportional to the number of cells visitied, and a constant that only depends on d. Since the total number of visited cells is at most $O((d/\varepsilon)^d)$, the running time of computing Γ_k from Γ_{k-1} is $O(|\Gamma_{k-1}|(d/\varepsilon)^d)=O(n(d/\varepsilon)^{O(dk)})$. It follows that the total running time for computing $\Gamma_1,\ldots,\Gamma_k$ is bounded by $O(n\log n+n(d/\varepsilon)^{O(dk)})$.

We end the section with a property of our computed WSTs which will be useful in Section 4.

Lemma 13. *For any ε-WST $t=(q_0,\ldots,q_k)$ generated by our algorithm, let $\rho=\mathrm{rad}(t)$. Then, the height λ of each q_i satisfies:*

$$2^\lambda\leq\frac{\varepsilon\rho}{\sqrt{d}}.$$

Proof. We do induction on k. For $k=1$, every pair $(q,q')\in\Gamma_1$ is an $\frac{\varepsilon}{2}$-well separated pair. With $\ell:=d(q,q')$ the minimum distance between q and q', it is clear that $\rho\geq\ell/2$. From the well-separated property, we know that $\max(\mathrm{diam}(q),\mathrm{diam}(q'))\leq\frac{\varepsilon\ell}{2}$ and therefore, the maximum height λ of q and q' is such that $2^\lambda\leq\frac{\varepsilon\ell}{2\sqrt{d}}\leq\frac{\varepsilon\rho}{\sqrt{d}}$ as required.

For $k>1$, assume that for every $(\varepsilon,k-1)$-WST, the lemma holds. Let $\gamma'=q_0,\ldots,q_{k-1}\in\Gamma_{k-1}$ be any $(\varepsilon,k-1)$-WST and $\rho'=\mathrm{rad}(\gamma')$. Assume that our algorithm generates $\gamma=(q_0,\ldots,q_{k-1},q')$; then q' is a cell of level h with $2^h\leq\frac{\varepsilon\rho'}{2\sqrt{d}}$. Because $\rho=\mathrm{rad}(\gamma)\geq\rho'$, this implies that the statement is true for q', and also holds for $q_0,\ldots,q_{k-1}$ by induction hypothesis. □

4 Cech approximations of linear size

In this section, we will define a persistence module which is a $(1+\varepsilon)$-approximation of the Čech module in the sense of Section 2. We start with a summary of our construction: we first define a sequence of (non-nested) simplicial complexes $(\mathscr{A}_\alpha)_{\alpha\geq0}$, which we define using a WSSD from Section 3. Then, we construct simplicial maps $g_\alpha^{\alpha'}:\mathscr{A}_\alpha\to\mathscr{A}_{\alpha'}$ such that $g_{\alpha'}^{\alpha''}\circ g_\alpha^{\alpha'}=g_\alpha^{\alpha''}$ and $g_\alpha^\alpha=\mathrm{id}$. As discussed in Section 2, applying the homology functor to that sequence yields a persistent module. To show that the constructed module approximates the Čech module, we define simplicial *cross-maps* $\phi:\mathscr{C}_{\frac{\alpha}{1+\varepsilon}}\to\mathscr{A}_\alpha$ and $\psi:\mathscr{A}_\alpha\to\mathscr{C}_\alpha$ that connect the two sequences on a simplicial level. We then show that the induced maps on homology groups all commute and finally apply Theorem 3 to show that the constructed module $(1+\varepsilon)$-approximates the Čech module. We remark that this strategy follows the approach by Dey et al. [10] who get a similar result for the Rips module, simplifying the previous work of Sheehy [22].

More notations. Throughout the section, we assume a finite point set $S\subset\mathbb{R}^d$ and a quadtree Q, and we reuse the notation on quadtrees from the previous section. Moreover, we will use assume the existence of an $\frac{\varepsilon}{12}$-WSSD defined over cells of Q, computed with the algorithm from Section 3. We will mostly omit the "$\frac{\varepsilon}{12}$" and just talk about the WSSD and its WSTs from now on. Having a WST $t=(q_0,\ldots,q_k)$, we write $\mathrm{rad}(t)$ for the radius of the minimum enclosing ball of $q_0\cup\ldots\cup q_k$. For a non-empty quadtree cell q, we choose a *representative* $\mathrm{rep}(q)$ in S with the property that if

q is internal, its representative is chosen among the representatives of its children. Moreover, for any quadtree cell q of height i or less, we define $\mathrm{qcell}(q,i)$ for its (unique) ancestor at level i.

We fix the following additonal parameters: Set $\theta_\ell := (1+\frac{\varepsilon}{2})^\ell$ for any integer ℓ. Let Δ_α denote the integer such that

$$\theta_{\Delta_\alpha} \leq \alpha < \theta_{\Delta_\alpha+1}.$$

Furthermore, we define h_α as the integer such that

$$2^{h_\alpha} \leq \frac{\varepsilon \theta_{\Delta_\alpha}}{3\sqrt{d}} \leq 2^{h_\alpha+1}.$$

When there is no ambiguity about α, we will skip the suffixes and write $\Delta := \Delta_\alpha$ and $h := h_\alpha$.

To give a rough intuition about the chosen terms, the approximate complex will be only changing at discrete values; more precisely, all $\alpha \in [\theta_\ell, \theta_{\ell+1})$ will result in the same approximation. This motivates the definition of Δ_α which determines the range in which α falls in. The second parameter h_α determines the grid size on which the approximation is constructed. Note that h_α rather depends on Δ_α than on α itself. Consequently, for any $\alpha \in [\theta_k, \theta_{k+1})$, the same h_α is chosen. Before we formally describe our construction, we prove the following useful lemma:

Lemma 14. *Let $\alpha > 0$, $\Delta := \Delta_\alpha$ and $h := h_\alpha$ as defined above. If an $\frac{\varepsilon}{12}$-WST $t = (q_0, \ldots, q_k)$ satisfies* $\mathrm{rad}(t) \leq \theta_{\Delta+1}$, *the height of each q_i is h or smaller.*

Proof. Since $\mathrm{rad}(t) \leq \theta_{\Delta+1}$, Lemma 13 implies that the height h' of each q_i satisfies $2^{h'} \leq \frac{\varepsilon\theta_{\Delta+1}}{12\sqrt{d}}$. Note that $\theta_{\Delta+1} = (1+\varepsilon/2)\theta_\Delta \leq 2\theta_\Delta$, and therefore,

$$2^{h'} \leq \frac{\varepsilon\theta_\Delta}{6\sqrt{d}} < \frac{2^{h+1}}{2} \leq 2^h. \qquad \square$$

The approximation complex Recall that G_ℓ denotes the set of all quadtree cell at height ℓ. We construct a simplicial complex $\mathscr{A}_\alpha$ with vertex set G_h in the following way: For any WST $t' = (q_0, \ldots, q_k)$ with all q_i at height h or less, let $t = (\mathrm{qcell}(q_0,h), \ldots, \mathrm{qcell}(q_k,h))$. If $\mathrm{rad}(t) \leq \theta_\Delta$, we add the simplex t to $\mathscr{A}_\alpha$. Note that some of the $\mathrm{qcell}(q_\ell,h)$ can be the same, so that the resulting simplex might be of dimension less than k. It is clear by construction and Lemma 12 that $\mathscr{A}_\alpha$ consists of at most $n(d/\varepsilon)^{O(d^2)}$ simplices, but it requires a proof to show that it is well-defined:

Lemma 15. *$\mathscr{A}_\alpha$ is a simplicial complex.*

Proof. Let $(q_0, \ldots, q_k) \in \mathscr{A}_\alpha$. We need to show that its faces are in $\mathscr{A}_\alpha$ as well. Wlog consider $(q_0, \ldots, q_\ell)$ with $\ell < k$. Since each q_i is non-empty, we can choose some $v_i \in q_i$ and consider the simplex $\tau = (v_0, \ldots, v_\ell)$. By the covering property of WSSD, there exists a WST $t' = (q'_0, \ldots, q'_\ell)$ that covers τ. Note that

$$\mathrm{rad}(\tau) \leq \mathrm{rad}(q_0, \ldots, q_\ell) \leq \mathrm{rad}(q_0, \ldots, q_k) \leq \theta_\Delta.$$

Now, because t' is $\frac{\varepsilon}{12}$-well-separated and the meb of τ intersects all q'_i,

$$\mathrm{rad}(t') \leq (1+\frac{\varepsilon}{12})\mathrm{rad}(\tau) < \theta_{\Delta+1}$$

It follows by Lemma 14 that all q'_i are at most on level h. In particular, for all i, $\mathrm{qcell}(q'_i,h) = q_i$ because both cells contain v_i, and q_i is on height h by construction. Because $\mathrm{rad}(q_0, \ldots, q_\ell) \leq \theta_\Delta$, it follows that $(q_0, \ldots, q_\ell)$ belongs to $\mathscr{A}_\alpha$ because of the WST t'. $\square$

We define maps between the $\mathscr{A}_\alpha$ next: Consider two scales $\alpha_1 < \alpha_2$. We set $h_1 := h_{\alpha_1}$ and define h_2, Δ_1, and Δ_2 accordingly. Since $h_1 \le h_2$, there is a natural map $g_{\alpha_1}^{\alpha_2} : G_{h_1} \to G_{h_2}$, mapping a quadtree cell at height h_1 to its ancestor at height h_2. This naturally extends to a map

$$g_{\alpha_1}^{\alpha_2} : \mathscr{A}_{\alpha_1} \to \mathscr{A}_{\alpha_2},$$

by mapping a simplex $\sigma = (v_0,\ldots,v_k)$ to $g_{\alpha_1}^{\alpha_2}(\sigma) := (g_{\alpha_1}^{\alpha_2}(v_0),\ldots,g_{\alpha_1}^{\alpha_2}(v_k))$. It is easy to verify that $g_{\alpha'}^{\alpha''} \circ g_{\alpha}^{\alpha'} = g_{\alpha}^{\alpha''}$ and $g_\alpha^\alpha = \mathrm{id}$.

Lemma 16. *$g := g_{\alpha_1}^{\alpha_2} : \mathscr{A}_{\alpha_1} \to \mathscr{A}_{\alpha_2}$ is a simplicial map.*

Proof. Let $t = (q_0,\ldots,q_k)$ be a k-simplex of $\mathscr{A}_{\alpha_1}$. In particular, $\mathrm{rad}(t) \le \theta_{\Delta_1}$ and all cells are at level h_1. Let $q'_\ell = g(q_\ell)$ denote the ancestor of q_ℓ at level h_2. We need to show that $t' = (q'_0,\ldots,q'_k) \in \mathscr{A}_{\alpha_2}$. For that, it suffices to show that $\mathrm{rad}(t') \le \theta_{\Delta_2}$. Note that $\Delta_1 = \Delta_2$ implies $h_1 = h_2$, so $t = t'$ and the statement is trivial. So, assume that $\Delta_1 < \Delta_2$.

Consider the minimum enclosing ball of t. Note that this ball contains q_ℓ, and therefore also at least one point of each q'_ℓ, for $0 \le \ell \le k$. We increase the radius by (at least) the diameter of a quadtree cell on level h_2. The enlarged ball then contains q'_ℓ completely (compare Observation 5). The diameter of the cells at level h_2, however, is at most

$$\frac{\varepsilon\theta_{\Delta_2}}{3\sqrt{d}}\sqrt{d} \le \frac{\varepsilon}{3}\theta_{\Delta_2}.$$

Moreover, because Δ_1 is strictly smaller than Δ_2, $\theta_{\Delta_1} \le \frac{\theta_{\Delta_2}}{1+\frac{\varepsilon}{2}}$. It follows that

$$\mathrm{rad}(t') \le \mathrm{rad}(t) + \frac{\varepsilon}{3}\theta_{\Delta_2} \le \frac{1+\frac{\varepsilon}{3}+\frac{\varepsilon^2}{6}}{1+\frac{\varepsilon}{2}}\theta_{\Delta_2} \le \theta_{\Delta_2}$$

for all $\varepsilon \le 1$. Therefore, $t' \in \mathscr{A}_{\alpha_2}$. □

Cross maps Next, we investigate the *cross-map* $\phi : \mathscr{C}_{\frac{\alpha}{1+\varepsilon}} \to \mathscr{A}_\alpha$. To define it for a vertex $v \in \mathscr{C}_{\frac{\alpha}{1+\varepsilon}}$ (which is a point of S), set $\phi(v) = q$, where q is the quadtree cell at level h that contains v. For a simplex $(v_0,\ldots,v_k)$, define $\phi(v_0,\ldots,v_k) = (\phi(v_0),\ldots,\phi(v_k))$.

Lemma 17. *ϕ is a simplicial map.*

Proof. Fix a simplex $\sigma = (v_0,\ldots,v_k) \in \mathscr{C}_{\frac{\alpha}{1+\varepsilon}}$. Take a WST $t = (q_0,\ldots,q_k)$ that covers σ. By the properties of the $\frac{\varepsilon}{12}$-WSSD, it follows that

$$\mathrm{rad}(t) \le (1+\frac{\varepsilon}{12})\mathrm{rad}(\sigma) \le \frac{(1+\frac{\varepsilon}{12})}{1+\varepsilon}\alpha.$$

Now, since $\frac{1+\frac{\varepsilon}{12}}{1+\varepsilon}\alpha < \alpha \le \theta_{\Delta+1}$, we can apply Lemma 14 which guarantees that all q_ℓ are at level at most h. Let $t' = (q'_0,\ldots,q'_k)$ with $q'_\ell = \mathrm{qcell}(q_\ell,h)$. Note that $q'_\ell = \phi(v_\ell)$, so all we need to show is that $t' \in \mathscr{A}_\alpha$. As mentioned in the proof of Lemma 16, the diameter of a cell at level h is at most $\frac{\varepsilon}{3}\theta_\Delta$. It follows that the minimum enclosing ball of t enlarged by $\frac{\varepsilon}{3}\theta_\Delta$ covers t'. Therefore,

$$\mathrm{rad}(t') \le \mathrm{rad}(t) + \frac{\varepsilon}{3}\theta_\Delta \le \frac{(1+\frac{\varepsilon}{12})}{1+\varepsilon}\alpha + \frac{\varepsilon}{3}\theta_\Delta$$

Since $\alpha \leq (1+\frac{\varepsilon}{2})\theta_\Delta$, this implies

$$\mathrm{rad}(t') \leq \frac{1+\frac{5}{12}\varepsilon+\frac{3}{8}\varepsilon^2}{1+\varepsilon}\theta_\Delta \leq \theta_\Delta$$

for $\varepsilon \leq \frac{14}{9}$. It follows that $t' \in \mathscr{A}_\alpha$. □

In the other direction, we have a map $\psi : \mathscr{A}_\alpha \to \mathscr{C}_\alpha$ defined by mapping a quadtree cell q at level h to its representative $\mathrm{rep}(q)$. It is easy to see that this map is simplicial: For $t = (q_0,\ldots,q_k)$ in $\mathscr{A}_\alpha$, we have that $\mathrm{rad}(t) \leq \theta_\Delta \leq \alpha$. Setting $\sigma := (\mathrm{rep}(q_0),\ldots,\mathrm{rep}(q_k))$, it is clear that $\mathrm{rad}(\sigma) \leq \mathrm{rad}(t) \leq \alpha$, so $\sigma \in \mathscr{C}_\alpha$.

Interleaving sequences We fix some integer $p \geq 0$ and consider the persistence modules

$$(\hat{\mathscr{C}}_\alpha)_{\alpha\geq 0} := (H_p(\mathscr{C}_\alpha))_{\alpha\geq 0}, \quad (\hat{\mathscr{A}}_\alpha)_{\alpha\geq 0} := (H_p(\mathscr{A}_\alpha))_{\alpha\geq 0}$$

and the induced homomorphisms $\hat{f}_{\alpha_1,\alpha_2}$ (induced by inclusion) and $\hat{g}_{\alpha_1,\alpha_2}$, respectively. Moreover, since the cross maps are simplicial, the induced homomorphisms $\hat{\phi} : \hat{\mathscr{C}}_{\frac{\alpha}{1+\varepsilon}} \to \hat{\mathscr{A}}_\alpha$ and $\hat{\psi} : \hat{\mathscr{A}}_\alpha \to \hat{\mathscr{C}}_\alpha$ connect the two modules. We show that the crossmaps $\hat{\phi}$, $\hat{\psi}$ commute with the module maps $\hat{f}$, $\hat{g}$ in the next three lemmas.

Lemma 18. *The diagram*

$$\begin{tikzcd}
\hat{\mathscr{C}}_{\frac{\alpha}{1+\varepsilon}} \arrow[dr, "\hat{\phi}"] & \\
\hat{\mathscr{A}}_{\frac{\alpha}{1+\varepsilon}} \arrow[u, "\hat{\psi}"] \arrow[r, "\hat{g}"] & \hat{\mathscr{A}}_\alpha
\end{tikzcd}$$

commutes, that means, $\hat{\phi} \circ \hat{\psi} = \hat{g}$.

Proof. The maps commute already on the simplicial level, that is, $\phi \circ \psi = g$, as one can easily verify from the definition of the maps. □

For the next two lemmas, we need the following definition: Two simplicial maps $h_1, h_2 : K \to L$ are *contiguous* if for any simplex $(v_0,\ldots,v_k) \in K$, the points $(h_1(v_0),\ldots,h_1(v_k),h_2(v_0),\ldots,h_2(v_k))$ form a simplex in L. In this case, the induced homomorphisms $\hat{h_1}, \hat{h_2}$ are equal [20, p.67].

Lemma 19. *The diagram*

$$\begin{tikzcd}
\hat{\mathscr{C}}_{\frac{\alpha}{1+\varepsilon}} \arrow[r, "f"] \arrow[dr, "\hat{\phi}"'] & \hat{\mathscr{C}}_\alpha \\
& \hat{\mathscr{A}}_\alpha \arrow[u, "\hat{\psi}"]
\end{tikzcd}$$

commutes, that means, $\hat{\psi} \circ \hat{\phi} = \hat{f}$.

Proof. Note the simplicial maps do not commute here; we will show instead that they are contiguous. So, fix a simplex $\sigma = (v_0, \ldots, v_k)$ in $\mathscr{C}_{\frac{\alpha}{1+\varepsilon}}$. Consider its image $(q_0, \ldots, q_k)$ under ϕ. All q_ℓ are on level h, $v_\ell \in q_\ell$, and $\mathrm{rad}(q_0, \ldots, q_k) \leq \theta_\Delta \leq \alpha$. Let $(w_0, \ldots, w_k)$ be the image of $(u_0, \ldots, u_k)$ under ψ, that is, w_ℓ is the representative of q_ℓ. In particular, we have that $w_\ell \in q_\ell$. It follows that the set $\{v_0, \ldots, v_k, w_0, \ldots, w_k\}$ is contained in the union $q_0 \cup \ldots \cup q_k$ and therefore, $\mathrm{rad}(v_0, \ldots, v_k, w_0, \ldots, w_k) \leq \alpha$. It follows that the simplex $(v_0, \ldots, v_k, w_0, \ldots, w_k)$ is in $\mathscr{C}_\alpha$. Hence, $\psi \circ \phi$ and f are contiguous. □

Lemma 20. *For $\alpha_1 \leq \alpha_2$, the diagram*

$$\begin{tikzcd}
\hat{\mathscr{C}}_{\alpha_1} \arrow[r, "\hat{f}"] & \hat{\mathscr{C}}_{\alpha_2} \\
\hat{\mathscr{A}}_{\alpha_1} \arrow[u, "\hat{\psi}"] \arrow[r, "\hat{g}"] & \hat{\mathscr{A}}_{\alpha_2} \arrow[u, "\hat{\psi}"]
\end{tikzcd}$$

commutes, that means, $\hat{\psi} \circ \hat{g} = \hat{f} \circ \hat{\psi}$.

Proof. Again, the corresponding simplicial maps do not commute in general (they do only if $h_{\alpha_1} = h_{\alpha_2}$). We will show that the simplicial maps are contiguous. Fix some $t = (q_0, \ldots, q_k) \in \mathscr{A}_{\alpha_1}$ and let v_ℓ be the representative of q_ℓ; in particular $f \circ \psi(q_\ell) = v_\ell$. Now, set $q'_\ell := g(q_\ell)$. It is clear that $q_\ell \subseteq q'_\ell$. Moreover, by definition of $\mathscr{A}_{\alpha_2}$, we have that $\mathrm{rad}(q'_0, \ldots, q'_k) \leq \theta_{\Delta_2} \leq \alpha_2$, where $\Delta_2 := \Delta_{\alpha_2}$. Set $w_\ell := \psi(g(q_\ell)) = \psi(q'_\ell)$ be the representative of q'_ℓ. By construction, $v_0, \ldots, v_k, w_0, \ldots, w_k$ are all contained in the union $q'_0 \cup \ldots \cup q'_k$ and therefore, $\mathrm{rad}(v_0, \ldots, v_k, w_0, \ldots, w_k) \leq \alpha_2$. This implies that the two maps are contiguous. □

Theorem 21. *The persistence module $\hat{\mathscr{A}}_*$ is a $(1+\varepsilon)$-approximation of the persistence module $\hat{\mathscr{C}}_*$.*

Proof. Using Lemmas 18-20, one can show that all diagrams in (2.1) commute by splitting them into subdiagrams. The result follows from Theorem 3. □

5 Coresets for minimal enclosing ball radii

Recall that for a point set $P = \{p_1, \ldots, p_n\} \subset \mathbb{R}^d$, we denote by $\mathrm{meb}(P)$ the *minimum enclosing ball of P*. Let $\mathrm{center}(P) \in \mathbb{R}^d$ denote the center and $\mathrm{rad}(P) \geq 0$ the radius of $\mathrm{meb}(P)$. Fixing $\varepsilon > 0$, we call a subset $C \subseteq P$ a *meb-coreset for P* if the ball centered at $\mathrm{center}(C)$ and with radius $(1+\varepsilon)\mathrm{rad}(C)$ contains P. We call $C \subseteq P$ a *radius-coreset for P* if $\mathrm{rad}(P) \leq (1+\varepsilon)\mathrm{rad}(C)$. Informally, a radius-coreset approximates only the radius of the minimum enclosing ball, whereas the meb-coreset approximates the ball itself. A meb-coreset is also a radius-coreset by definition, but the opposite is not always the case; see Figure 5.1 for an example.

Obviously, a point set is a coreset of itself, so coresets exist for any point set. We are interested in the coresets of small sizes. For the meb-coreset, this question is answered by Bădoiu and Clarkson [3]. We summarize their result in the following statement:

Theorem 22. *For $\varepsilon > 0$, and any (finite) point set, there exists a meb-coreset of size $\lceil \frac{1}{\varepsilon} \rceil$, and there exist point sets where any meb-coreset has size at least $\lceil \frac{1}{\varepsilon} \rceil$.*

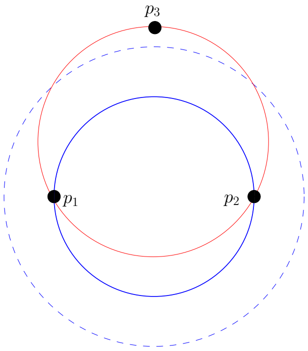

Figure 5.1: Consider the regular triangle with points $p_1 = (-1,0)$, $p_2 = (1,0)$ and $p_3 = (0,\sqrt{3})$ in the plane; let $P = \{p_1, p_2, p_3\}$ and $C = \{p_1, p_2\}$. Then, center$(P) = (0,\sqrt{1/3})$, rad$(P) = \sqrt{\frac{4}{3}}$, center$(C) = (0,0)$, and rad$(C) = 1$. For $\varepsilon = 0.5$, it is thus clear that C is a radius-coreset of P. However, C is not a meb-coreset because the ball with radius 1.5 around the origin does not contain p_3.

Note that the size of the coreset is independent of both the number of points in P and the ambient dimension. However, since radius-coresets are a relaxed version of meb-coresets, we can hope for even smaller coresets. We start by showing a lower bound:

Lemma 23. *There is a point set such that any radius-coreset has size at least*

$$\delta := \lceil \frac{1}{2\varepsilon + \varepsilon^2} + 1 \rceil.$$

Proof. Consider the standard $(d-1)$-simplex in d dimensions, that is, P is the point set given by the d unit vectors in $\mathbb{R}^d$. By elementary calculations, it can be verified that center$(P) = (\frac{1}{d}, \ldots, \frac{1}{d})$ and rad$(P) = \sqrt{\frac{d-1}{d}}$. Fixing a subset $C \subseteq P$ of size k, its points span a standard simplex in $\mathbb{R}^k$ and therefore, rad$(C) = \sqrt{\frac{k-1}{k}}$ by the same argument. Hence, C is a radius-coreset of P if and only if

$$\sqrt{\frac{d-1}{d}} \leq (1+\varepsilon)\sqrt{\frac{k-1}{k}}.$$

Isolating k yields the equivalent condition that

$$k \geq \lceil \frac{(1+\varepsilon)^2}{(1+\varepsilon)^2 - \frac{d-1}{d}} \rceil = \lceil 1 + \frac{1}{\frac{d}{d-1}(2\varepsilon + \varepsilon^2 + \frac{1}{d})} \rceil.$$

The last expression is monotonously increasing in d, and converges to δ for $d \to \infty$. It follows that, for d large enough, any radius-coreset of a standard $(d-1)$-simplex has size at least δ. □

We will show next that any point set has a radius coreset of size δ. For a point set P in $\mathbb{R}^d$ and $1 \leq k \leq d$, let $r_k(P)$ denote the maximal radius of a meb among all subsets of P of cardinality k. We can assume that P contains at least $d+1$ points; otherwise it is contained in a lower-dimensional Euclidean space. On the other hand, if P contains at least $d+1$ points, there exists a subset P' of

P containing exactly $d+1$ points such that the meb of P' equals the meb of P, which implies that $r_{d+1}(P) = \mathrm{rad}(P)$. Moreover, $r_2(P) = \mathrm{diam}(P)$ is the diameter of P. We use a result by Henk [17, Thm.1] (we adapt his notation to our context):

Theorem 24 (Generalized Jung's Theorem)**.** *Let $P \subset \mathbb{R}^d$ be a point set, and let i, j two integers with $2 \leq j \leq i \leq d+1$. Then*

$$r_i(P) \leq \sqrt{\frac{j(i-1)}{i(j-1)}} r_j(P)$$

The theorem generalizes an older result by Jung [18] which states the following relation between the circumradius and the diameter of P:

$$\mathrm{rad}(P) = r_{d+1}(P) \leq \sqrt{\frac{2d}{d+1}} r_2(P) = \sqrt{\frac{2d}{d+1}} \mathrm{diam}(P). \tag{5.1}$$

We sketch the proof of Theorem 24 for completeness. It relies on the following property: Given a point set Q of $k+1$ linearly independent points in $\mathbb{R}^k$. Then,

$$\mathrm{rad}(Q) \leq \frac{k}{\sqrt{k^2-1}} r_k(Q), \tag{5.2}$$

in other words, there is a subset of k points whose circumradius is large in some sense; see also [3, Lemma 3.3]. We assume for simplicity that the i-subset of points of P that realizes $r_i(P)$ is linearly independent; otherwise, we can switch to an independent subset and a similar argument applies. Iteratively applying (5.2) yields that

$$r_i(P) \leq \prod_{t=j}^{i-1} \frac{t}{\sqrt{t^2-1}} r_j(P).$$

However, it is a straight-forward to prove by induction that

$$\prod_{t=j}^{i-1} \frac{t}{\sqrt{t^2-1}} = \sqrt{\frac{j(i-1)}{i(j-1)}}.$$

Theorem 25. *For $\varepsilon > 0$, any finite point set P has a radius-coreset of size δ.*

Proof. Applying Theorem 24 to the case that $i = d+1$ and $j = \delta$ yields

$$\mathrm{rad}(P) = r_{d+1}(P) \leq \sqrt{\frac{\delta \cdot d}{(d+1)(\delta-1)}} r_\delta(P) = \underbrace{\sqrt{\frac{d}{d+1}}}_{\leq 1} \sqrt{\frac{\delta}{\delta-1}} r_\delta(P).$$

Furthermore, since $\delta \geq \frac{1}{2\varepsilon+\varepsilon^2} + 1$, it follows that

$$\frac{\delta}{\delta-1} = 1 + \frac{1}{\delta-1} \leq (1+\varepsilon)^2.$$

So, letting C be a subset of cardinality δ with radius $r_\delta(P)$, we obtain that $\mathrm{rad}(P) \leq (1+\varepsilon)\mathrm{rad}(C)$, which means that C is a radius-coreset. □

We remark that our results immediately imply an algorithm for computing a radius-coreset of size δ: starting with the whole point set, iteratively remove points such that the remaining subset has the largest possible radius among all choices of removed points. When this process is stopped for a subset of size δ, the resulting subset is a radius-coreset. However, this algorithm is rather inefficient, because it is quadratic in n, and a natural question is how to compute radius coresets more efficiently. For meb-coresets of size $\lceil \frac{1}{\varepsilon} \rceil$, Bădoiu and Clarkson [3] prove existence algorithmically by defining an algorithm which starts with an arbitrary set of size $\lceil \frac{1}{\varepsilon} \rceil$ and alternatingly adds and removes points from the set until the set remains unchanged, and they prove that the resulting set is a meb-coreset. Their algorithm is an instance of a more general class of optimization problems as described in [7]; we were not able to find a reformulation of the radius-coreset problem in terms of this algorithmic framework.

6 A generalized Rips-Lemma

We define the following generalization of a flag-complex:

Definition 26 (i-completion)**.** *Let K denote a simplical complex. The i-*completion *of K, $\mathscr{M}_i(K)$, is maximal complex whose i-skeleton equals the i-skeleton of K.*

With that notation, we have that $\mathscr{R}_\alpha = \mathscr{M}_1(\mathscr{C}_\alpha)$. Moreover, we have that $\mathscr{C}_\alpha = \mathscr{M}_d(\mathscr{C}_\alpha)$ as a consequence of Helly's Theorem [12].

We can show the following result as an application of Theorem 25.

Theorem 27. *For $\delta = \lceil 1/(2\varepsilon + \varepsilon^2) + 1 \rceil$,*

$$\mathscr{C}_\alpha \subseteq \mathscr{M}_{\delta-1}(\mathscr{C}_\alpha) \subseteq \mathscr{C}_{(1+\varepsilon)\alpha}$$

Proof. The first inclusion is clear. Now, consider a simplex σ in $\mathscr{M}_{\delta-1}(\mathscr{C}_\alpha)$. The second inclusion is trivial if $\dim \sigma \leq \delta - 1$, so let its dimension be at least δ. By Theorem 25, the boundary vertices of σ have a coreset of size at most δ. Let τ denote the simplex spanned by such a coreset. As τ is a face of σ, it is contained in $\mathscr{M}_{\delta-1}(\mathscr{C}_\alpha)$, and because it is of dimension at most $\delta - 1$, it is in particular contained in $C(\alpha)$. By the property of coresets, the minimal enclosing ball of σ has radius at most $(1+\varepsilon)\alpha$ which implies that $\sigma \in \mathscr{C}_{(1+\varepsilon)\alpha}$. □

As a special case, consider the choice $\varepsilon = \sqrt{2} - 1$, so that $\delta = 2$. The above result yields that

$$\mathscr{R}_\alpha = \mathscr{M}_1(\mathscr{C}_\alpha) \subseteq \mathscr{C}_{\sqrt{2}\alpha},$$

which is exactly the statement of the Vietoris-Rips Lemma as stated in [12, p.62].

Theorem 27 and Lemma 4 prove the closeness of the persistence diagrams of the Čech filtration and the completion complex:

Theorem 28. *The persistence diagram of $\mathscr{M}_{\delta-1}(\mathscr{C}_*)$ with $\delta := \lceil 1/(2\varepsilon + \varepsilon^2) + 1 \rceil$ is a $(1+\varepsilon)$-approximation of the persistence diagram of $\mathscr{C}_*$.*

Note that $\mathscr{M}_k(\mathscr{C}_\alpha)$ is determined by the k-skeleton of the Čech complex, which of size $O(n^{k+1})$. In this respect, the completion complex constitutes a trade-off between simplicity (i.e., its representation size) and approximation quality of the Čech complex. We emphasize that the approximation is solely determined by k and does not depend on the ambient dimension of the point set.

7 Conclusion and Outlook

We have presented two distinct ways to approximate Cech complexes; the fixed-dimensional result on approximating the Cech filtration to linear size is a technically challenging, but conceptually straight-forward extension of recent work on the Rips filtration; however, we believe that the concept of WSSDs to be interesting and hopefully applicable in different contexts, and we plan to identify application scenarios in the future. Our high-dimensional results are a first attempt to link the areas of computational topology, where data is often high-dimensional, and geometric approximation algorithms that try to overcome the curse of dimensionality. We want to achieve algorithmic results in that context in the future; one question is whether an optimal-size radius coreset can be computed efficiently. Moreover, the introduced concept of completions is not tied to start completing simplices at a fixed dimension; in fact, one can start with any complex C (not necessarily a skeleton) and define the completion as the largest complex containing C. With such *adaptive completions*, ε-close approximations of the Cech filtration might be possible with just a slightly larger representation size than the Rips filtration. The open question is, however, whether such a representation can be computed efficiently. Finally, we pose the question whether there are other applications, besides approximating Čech complexes, where the smaller size of radius-coresets in comparison to meb-coresets could be useful.

References

[1] D. Attali, A. Lieutier, and D. Salinas. Efficient data structures for representing and simplifying simplicial complexes in high dimensions. *Int. J. of Computational Geometry & Applications*, 22:279–303, 2012.

[2] J.-D. Boissonnat, L. Guibas, and S. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Discr. Comput. Geom.*, 42:37–70, 2009.

[3] M. Bădoiu and K. Clarkson. Optimal core-sets for balls. *Computational Geometry: Theory and Applications*, 40:14–22, 2008.

[4] P. Callahan and S. Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest neighbors and n-body potential fields. *J. of the ACM*, 42(67–90), 1995.

[5] G. Carlsson and V. de Silva. Zigzag persistence. *Foundations of Computational Mathematics*, 10:367–405, 2010.

[6] F. Chazal, D. Cohen-Steiner, M. Glisse, L. Guibas, and S. Oudot. Proximity of persistence modules and their diagrams. In *Proc. 25th Symp. on Comp. Geom.*, pages 237–246, 2009.

[7] K. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. In *Proc. of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 922–931, 2008.

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

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

[10] T. Dey, F. Fan, and Y. Wang. Computing topological persistence for simplicial maps. *CoRR*, abs/1208.5018, 2012.

[11] T. Dey, F. Fan, and Y. Wang. Graph induced complex on point data. In *Proc. 29th ACM Symp. on Comp. Geom.*, 2013.

[12] H. Edelsbrunner and J. Harer. *Computational Topology, An Introduction*. American Mathematical Society, 2010.

[13] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory*, IT-29:551–559, 1983.

[14] H. Edelsbrunner and E. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.

[15] S. Har-Peled. *Geometric Approximation Algorithms*. American Mathematical Society, 2011.

[16] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics and their applications. *Siam J. on Computing*, 35:1148–1184, 2006.

[17] M. Henk. A generalization of Jung's theorem. *Geometriae Dedicata*, 42:235–240, 1992.

[18] H. Jung. Über die kleinste Kugel, die eine räumliche Figur einschliesst. *J. reine angewandte Mathematik*, 123:241–257, 1901.

[19] M. Mrozek, P. Pilarczyk, and N. Zelazna. Homology algorithm based on acyclic subspace. *Computer and Mathematics with Applications*, 55:2395–2412, 2008.

[20] J.R. Munkres. *Elements of algebraic topology*. Westview Press, 1984.

[21] S. Oudot and D. Sheehy. Zigzag zoology: Rips zigzags for homology inference. In *Proc. 29th ACM Symp. on Comp. Geom.*, 2013.

[22] D. Sheehy. Linear-size approximation to the Vietoris-Rips filtration. In *Proc. 2012 Symp. on Comp. Geom.*, pages 239–248, 2012.

[23] Afra Zomorodian. The tidy set: A minimal simplicial set for computing homology of clique complexes. In *Proc. 26th ACM Symp. on Comp. Geom.*, pages 257–266, 2010.