A Geometric Perspective on Sparse Filtrations*

Nicholas J. Cavanna† Mahmoodreza Jahanseir‡ Donald R. Sheehy§

Abstract

We present a geometric perspective on sparse filtrations used in topological data analysis. This new perspective leads to much simpler proofs, while also being more general, applying equally to Rips filtrations and Čech filtrations for any convex metric. We also give an algorithm for finding the simplices in such a filtration and prove that the vertex removal can be implemented as a sequence of elementary edge collapses.

A video illustrating this approach is available [7].

1 Introduction

Given a finite data set in a Euclidean space, it is natural to consider the balls around the data points as a way to fill in the space around the data and give an estimate of the missing data. The union of balls is often called the *offsets* of the point set. Persistent homology was originally invented as a way to study the changes in topology of the offsets of a point set as the radius increases from 0 to ∞. The input to persistent homology is usually a filtered simplicial complex, that is, an ordered collection of simplices (vertices, edges, triangles, etc.) such that each simplex appears only after its boundary simplices of one dimension lower. The Nerve Theorem and its persistent variant allow one to compute the persistent homology of the offsets by instead looking at a discrete object, a filtered simplicial complex called the nerve (see Fig. 1). The simplest version of this complex is called the Čech complex and it may be viewed as the set of all subsets of the input, ordered by the radius of their smallest enclosing ball. Naturally, the Čech complex gets very big very fast, even when restricting to subsets of constant size. A common alternative is the Rips complex but it suffers similar difficulties. Over the last few years, there have been several approaches to building sparser complexes that still give good approximations to the persistent homology [21, 17, 11, 3, 2].

Our main contributions are the following.

1. A much simpler explanation for the construction and proof of correctness of sparse filtrations. Our new geometric construction shows that the sparse complex is just a nerve in one dimension higher.

2. The approach easily generalizes to Rips, Čech and related complexes (the offsets for any convex metric). This is another advantage of the geometric view as the main result follows from convexity rather than explicit construction of simplicial map homotopy equivalences.

3. A simple geometric proof that the explicit removal of vertices from the sparse filtration can be done with simple edge contractions. This can be done without resorting to the full-fledged zigzag persistence algorithm [5, 4, 18, 19] or even the full simplicial map persistence algorithm [11, 1].

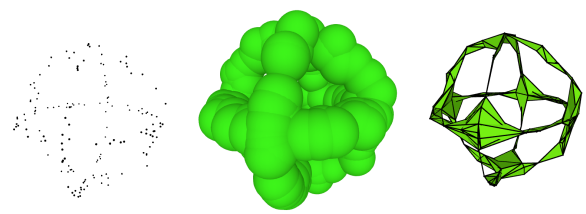

Figure 1: A point set sampled on a sphere, its offsets, and its (sparsified) nerve complex.

The most striking thing about this paper is perhaps more in what is absent than what is present. Despite giving a complete treatment of the construction, correctness, and approximation guarantees of sparse filtrations that applies to both Čech and Rips complexes, there is no elaborate construction of simplicial maps or proofs that they induce homotopy equivalences. In fact, we prove the results directly on the geometric objects, the covers, rather than the combinatorial objects, the complexes, and the result is much more direct. In a way, this reverses a common approach in computational geometry problems in which the geometry is as quickly as possible replaced with combinatorial structure; instead, we delay the transition from the offsets to a discrete representation until the very end of the analysis.

Related Work. Soon after the introduction of persistent homology by Edelsbrunner et al. [13], there was interest in building more elaborate complexes for larger and larger data sets. Following the full algebraic characterization of persistent homology by Zomorodian and

*Partially supported by the National Science Foundation under grant number CCF-1464379

†University of Connecticut nicholas.j.cavanna@uconn.edu

‡University of Connecticut reza@engr.uconn.edu

§University of Connecticut don.r.sheehy@gmail.com

Carlsson [23], a more general theory of zigzag persistence was developed [5, 4, 18, 19] using a more complicated algorithm. Zigzags gave a way to analyze spaces that did not grow monotonically; they could alternately grow and shrink such as by growing the scale and then removing points [22]. A variant of this techniques was first applied for specific scales by Chazal and Oudot in work on manifold reconstruction [9] and was implemented as a full zigzag by Morozov in his Dionysus library [12]. Later, Sheehy gave a zigzag for Rips filtrations that came with guaranteed approximation to the persistent homology of the unsparsified filtration [21]. Other later works gave various improvements and generalizations of sparse zigzags [20, 17, 11, 2].

2 Background

Distances and Metrics. Throughout, we will assume the input is a finite point set P in $\mathbb{R}^d$ endowed with some convex metric $\mathbf{d}$. A closed ball with center c and radius r will be written as $\mathrm{ball}(c, r) = \{x \in \mathbb{R}^d | \mathbf{d}(x, c) \leq r\}$. For illustrative purposes, we will often draw balls as Euclidean (ℓ_2) balls.

For a non-negative $\alpha \in \mathbb{R}$, the *α-offsets* of P are defined as

$$P^\alpha := \bigcup_{p \in P} \mathrm{ball}(p, \alpha).$$

The sequence of offsets as α ranges from 0 to ∞ is called the *offsets filtration* $\{P^\alpha\}$.

The *doubling dimension* of a metric space is $\log_2 \gamma$, where γ is the maximum over all balls B, of the minimum number of balls of half the radius of B required to cover B. Metric spaces with a small constant doubling dimension are called *doubling metrics*. Such metrics allow for packing arguments similar to those used in Euclidean geometry.

Simplicial Complexes. A *simplicial complex* K is a family of subsets of a vertex set that is closed under taking subsets. The sets $\sigma \in K$ are called *simplices* and $|\sigma| - 1$ is called the *dimension* of σ. A nested family of simplicial complexes is called a *simplicial filtration*. Often the family of complexes will be parameterized by a nonnegative real number as in $\{K^\alpha\}_{\alpha \geq 0}$. Here, the filtration property guarantees that $\alpha \leq \beta$ implies that $K^\alpha \subseteq K^\beta$. In this case, the value of α for which a simplex first appears is called its birth time, and so, if there is a largest complex K^α in the filtration, the whole filtration can be represented by K^α and the birth time of each simplex. For this reason, simplicial filtrations are often called *filtered simplicial complex*.

Persistent Homology. Homology is an algebraic tool for characterizing the connectivity of a space. It captures information about the connected components, holes, and voids. For this paper, we will only consider homology with field coefficients and the computations will all be on simplicial complexes. In this setting, computing homology is done by reducing a matrix D called the boundary matrix of the simplicial complex. The boundary matrix has one row and column for each simplex. If the matrix reduction respects the order of a filtration, i.e. columns are only combined with columns to their left, then the reduced matrix also represents the so-called *persistent homology* of the filtration. Persistent homology describes the changes in the homology as the filtration parameter changes and this information is often expressed in a *barcode* (See Fig. 2). Barcodes give topological signatures of a shape [14].

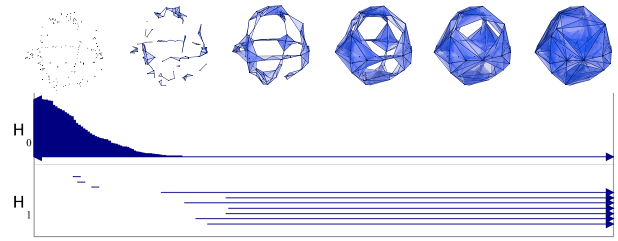

Figure 2: A filtration and its barcode.

Each bar of a barcode is an interval encoding the lifespan of a topological feature in the filtration. We say that a barcode B_1 is a (multiplicative) c-approximation to another barcode B_2 if there is a partial matching between B_1 and B_2 such that every bar $[b, d]$ with $d/b > c$ is matched and every matched pair of bars $[b, d], [b', d']$ satisfies $\max\{b/b', b'/b, d/d', d'/d\} \leq c$. A standard result on the stability of barcodes [8] implies that if two filtrations $\{F^\alpha\}$ and $\{G^\alpha\}$ are c-interleaved in the sense that $F^{\alpha/c} \subseteq G^\alpha \subseteq F^{c\alpha}$, then the barcode of $\{F^\alpha\}$ is a c-approximation to $\{G^\alpha\}$.

Nerve Complexes and Filtrations. Let $\mathcal{U} = \{U_1, \ldots, U_n\}$ be a collection of closed, convex sets. Let $\bigcup \mathcal{U}$ denote the union of the sets in $\mathcal{U}$, i.e. $\bigcup \mathcal{U} := \bigcup_{i=1}^n U_i$. We say that the set $\mathcal{U}$ is a *cover* of the space $\bigcup \mathcal{U}$. The *nerve* of $\mathcal{U}$, denoted $\mathrm{Nrv}(\mathcal{U})$ is the abstract simplicial complex defined as

$$\mathrm{Nrv}(\mathcal{U}) := \left\{ I \subseteq [n] \mid \bigcap_{i \in I} U_i \neq \emptyset \right\}.$$

This construction is illustrated in Fig.3. The Nerve Theorem [16, Cor. 4G.3] implies that $\mathrm{Nrv}(\mathcal{U})$ is homotopy equivalent to $\bigcup \mathcal{U}$.

Similarly, one can construct a nerve filtration from a cover of a filtration by filtrations. Let $\mathcal{U} = \{\{U_1^\alpha\}, \ldots \{U_n^\alpha\}\}$ be a collection of filtrations parameterized by real numbers such that for each $i \in [n]$ and each $\alpha \geq 0$, the set U_i^α is closed and convex. Let $\mathcal{U}^\alpha$

Figure 3: The nerve has an edge for each pairwise intersection, a triangle for each 3-way intersection (right), etc.

to denote the set $\{U_1^\alpha, \ldots, U_n^\alpha\}$. As before, the Nerve Theorem implies that $\bigcup \mathcal{U}^\alpha$ is homotopy equivalent to $\mathrm{Nrv}(\mathcal{U}^\alpha)$. The Persistent Nerve Lemma [9] implies that the filtrations $\{\bigcup \mathcal{U}^\alpha\}_{\alpha \geq 0}$ and $\{\mathrm{Nrv}(\mathcal{U}^\alpha)\}_{\alpha \geq 0}$ have identical persistent homology.

Čech and Rips Filtrations. A common filtered nerve is the *Čech filtration*. It is defined as $\{\mathcal{C}_\alpha(P)\}$, where

$$\mathcal{C}_\alpha(P) := \mathrm{Nrv}\{\mathrm{ball}(p_i, \alpha) \mid i \in [n]\}.$$

Notice that this is just the nerve of the cover of the α-offsets by the α-radius balls. Thus, the Persistent Nerve Lemma implies that $\{P^\alpha\}$ and $\{\mathcal{C}_\alpha(P)\}$ have identical persistence barcodes.

A similar filtration that is defined for any metric is called the *(Vietoris-)Rips filtration* and is defined as $\{\mathcal{R}_\alpha(P)\}$, where

$$\mathcal{R}_\alpha(P) := \{J \subseteq [n] \mid \max_{i,j \in J} \mathbf{d}(p_i, p_j) \leq 2\alpha\}.$$

Note that if $\mathbf{d}$ is the max-norm, ℓ_∞, then $\mathcal{R}_\alpha(P) = \mathcal{C}_\alpha(P)$. Moreover, because every finite metric can be isometrically embedded into ℓ_∞, every Rips filtration is isomorphic to a nerve filtration.

Greedy Permutations. Let P be a set of points in some metric space with distance $\mathbf{d}$. A *greedy permutation* of P goes by many names, including landmark sets, farthest point sampling, and discrete center sets. We say that $P = \{p_1, \ldots, p_n\}$ is ordered according to a greedy permutation if each p_i is the farthest point from the first $i-1$ points. We let p_1 be any point. Formally, let $P_i = \{p_1, \ldots, p_i\}$ be the ith *prefix*. Then, the ordering is greedy if and only if for all $i \in \{2, \ldots, n\}$,

$$\mathbf{d}(p_i, P_{i-1}) = \max_{p \in P} \mathbf{d}(p, P_{i-1}).$$

For each point p_i, the value $\lambda_i := \mathbf{d}(p_i, P_{i-1})$ is known as the *insertion radius*. By convention, we set $\lambda_1 = \infty$. It is well-known (and easy to check) that P_i is a λ_i-net in the sense that it satisfies the conditions: for all distinct $p, q \in P_i$, $\mathbf{d}(p,q) \geq \lambda_i$ (**packing**) and $P \subseteq P_i^{\lambda_i}$ (**covering**).

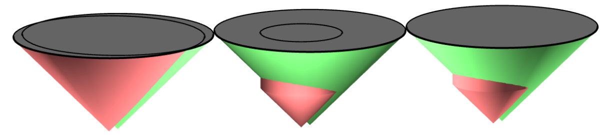

Figure 4: Left: two growing balls trace out cones in one dimension higher. Center: One of the cones has a maximum radius. Right: Limiting the height of one cone guarantees that the top is covered.

3 Perturbed Distances

A convenient first step in making a sparse version of the Čech filtration is to "perturb" the distance. Given a greedy permutation, we perturb the distance function so that as the radius increases, only a sparse subset of points continues to contribute to the offsets. This can most easily be viewed as changing the radius of the balls slightly so that some balls will be completely covered by their neighbors and thus will not contribute to the union. Fix a constant $\varepsilon < 1$ that will control the sparsity. As we will show in Lemma 1, at scale α, there is an $\varepsilon\alpha$-net of P whose perturbed offsets cover the perturbed offsets of P. Assuming the points $P = \{p_1, \ldots, p_n\}$ are ordered by a greedy permutation with insertion radii $\lambda_1, \ldots, \lambda_n$, we define the radius of p_i at scale α as

$$r_i(\alpha) := \begin{cases} \alpha & \text{if } \alpha \leq \lambda_i(1+\varepsilon)/\varepsilon \\ \lambda_i(1+\varepsilon)/\varepsilon & \text{otherwise.} \end{cases}$$

The *perturbed α-offsets* are defined as

$$\widetilde{P}^\alpha := \bigcup_{i \in [n]} \mathrm{ball}(p_i, r_i(\alpha)).$$

To realize the sparsification as described, we want to remove balls associated with some of the points as the scale increases. This is realized by defining the α-ball for a point $p_i \in P$ to be

$$b_i(\alpha) := \begin{cases} \mathrm{ball}(p_i, r_i(\alpha)) & \text{if } \alpha \leq \lambda_i(1+\varepsilon)^2/\varepsilon \\ \emptyset & \text{otherwise.} \end{cases}$$

The usefulness of this perturbation is captured by the following covering lemma, which is depicted in the tops of the cones in Fig. 4.

Lemma 1 (Covering Lemma) *Let $P = \{p_1, \ldots, p_n\}$ be a set of points ordered by a greedy permutation with insertion radii $\lambda_1, \ldots, \lambda_n$. For any $\alpha, \beta \geq 0$, and any $p_j \in P$, there exists a point $p_i \in P$ such that*

1. *if $\beta \geq \alpha$ then $b_j(\alpha) \subseteq b_i(\beta)$, and*
2. *if $\beta \geq (1+\varepsilon)\alpha$, then $\mathrm{ball}(p_j, \alpha) \subseteq b_i(\beta)$.*

Proof. Fix any $p_j \in P$. We may assume that $\beta \geq \lambda_j(1+\varepsilon)^2/\varepsilon$, for otherwise, choosing $p_i = p_j$ suffices to satisfy both clauses, the first because $b_j(\alpha) \subseteq b_j(\beta)$ and the second because $\mathrm{ball}(p_j, \alpha) = b_j(\alpha) \subseteq b_j(\beta)$. This assumption is equivalent to the assumption that $b_j(\beta) = \emptyset$.

By the covering property of the greedy permutation, there is a point $p_i \in P$ such that $\mathbf{d}(p_i, p_j) \leq \varepsilon\beta/(1+\varepsilon)$ and $\lambda_i \geq \varepsilon\beta/(1+\varepsilon)$. It follows that $r_i(\beta) = \beta$ and $b_i(\beta) = \mathrm{ball}(p_i, \beta)$. Recall that $\lambda_1 = \infty$ by convention, so $b_1(\beta) \neq \emptyset$, and for large values of β, choosing $p_i = p_1$ suffices.

To prove the first clause, fix any point $x \in b_j(\alpha)$. By the triangle inequality,

$$\begin{aligned}\mathbf{d}(x, p_i) &\leq \mathbf{d}(x, p_j) + \mathbf{d}(p_i, p_j) \leq r_j(\alpha) + \varepsilon\beta/(1+\varepsilon)\\ &\leq \lambda_j(1+\varepsilon)/\varepsilon + \varepsilon\beta/(1+\varepsilon) \leq \beta = r_i(\beta).\end{aligned}$$

So, $x \in b_i(\beta)$ and thus, $b_j(\alpha) \subseteq b_i(\beta)$ as desired.

To prove the second clause of the lemma, fix any $x \in \mathrm{ball}(p_j, \alpha)$. By the triangle inequality,

$$\begin{aligned}\mathbf{d}(x, p_i) &\leq \mathbf{d}(x, p_j) + \mathbf{d}(p_i, p_j) \leq \alpha + \varepsilon\beta/(1+\varepsilon)\\ &\leq \beta/(1+\varepsilon) + \varepsilon\beta/(1+\varepsilon) = r_i(\beta).\end{aligned}$$

So, as before, $x \in b_i(\beta)$ and thus, $\mathrm{ball}(p_j, \alpha) \subseteq b_i(\beta)$ as desired. □

Corollary 2 *Let $P = \{p_1, \ldots, p_n\}$ be a set of points ordered by a greedy permutation with insertion radii $\lambda_1, \ldots, \lambda_n$. For all $\alpha \geq 0$, $\widetilde{P}^\alpha = \bigcup_i b_i(\alpha)$ and $\widetilde{P}^\alpha \subseteq P^\alpha \subseteq \widetilde{P}^{(1+\varepsilon)\alpha}$.*

A proof may be found in the full paper [6]. Corollary 2 implies the following proposition using standard results on the stability of persistence barcodes [8].

Proposition 3 *The persistence barcode of the perturbed offsets $\{\widetilde{P}^\alpha\}_{\alpha\geq 0}$ is a $(1+\varepsilon)$-approximation to the persistence barcode of the offsets $\{P^\alpha\}_{\alpha\geq 0}$.*

4 Sparse Filtrations

The *sparse Čech complex* is defined as $Q^\alpha := \mathrm{Nrv}\{b_i(\alpha) \mid i \in [n]\}$. Notice that because $b_i(\alpha) = \emptyset$ unless λ_i is sufficiently large compared to α, there are fewer vertices as the scale increases. This is the desired sparsification. Unfortunately, it means that the set of complexes $\{Q^\alpha\}$ is not a filtration, but this is easily remedied by the following definition. The *sparse Čech filtration* is defined as $\{S^\alpha\}$, where

$$S^\alpha := \bigcup_{\delta\leq\alpha} Q^\delta = \bigcup_{\delta\leq\alpha} \mathrm{Nrv}\{b_i(\delta) \mid i \in [n]\}.$$

This definition makes it clear that the sparse complex is a union of nerves, but it not obvious that it has the same persistent homology as the filtration defined by the perturbed offsets $\widetilde{P}^\alpha := \bigcup_i b_i(\alpha)$. For such a statement, it would be much more convenient if $\{S^\alpha\}$ was itself a nerve filtration rather than a union of nerves, in which case the Persistent Nerve Lemma could be applied directly. In fact, this can be done by adding an extra dimension corresponding to the filtration parameter extending the balls $b_i(\alpha)$ into the perturbed cone shapes

$$U_i^\alpha := \bigcup_{\delta\leq\alpha} (b_i(\delta) \times \{\delta\}).$$

These sets, depicted in Figs. 4 and 5, allow the following equivalent definition of the complexes in the sparse Čech filtration.

$$S^\alpha := \mathrm{Nrv}\{U_i^\alpha \mid i \in [n]\}.$$

Theorem 4 *The persistence barcode of the sparse nerve filtration $\{S^\alpha\}_{\alpha\geq 0}$ is a $(1+\varepsilon)$-approximation to the persistence barcode of the offsets $\{P^\alpha\}_{\alpha\geq 0}$.*

Proof. For all i, the set U_i^α is convex because r_i is concave (see the full paper [6] for a proof). It follows that the sets U_i^α satisfy the conditions of the Persistent Nerve Lemma. So, $\{S^\alpha\}$ has the same persistence barcode as the filtration $\{B^\alpha\}$, where $B^\alpha := \bigcup_i U_i^\alpha$.

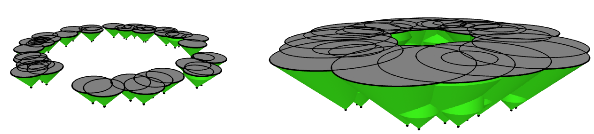

Figure 5: The collection of cones B^α at two different scales. The top of B^α is the union of (perturbed) balls.

The Covering Lemma implies that the linear projection of B^α to $\widetilde{P}^\alpha$ that maps (x, δ) to x is a homotopy equivalence as each fiber is simply connected. Moreover, the projection clearly commutes with the inclusions $B^\alpha \hookrightarrow B^\beta$ and $\widetilde{P}^\alpha \hookrightarrow \widetilde{P}^\beta$, from which, it follows that $\mathrm{Pers}\{\widetilde{P}^\alpha\} = \mathrm{Pers}\{B^\alpha\} = \mathrm{Pers}\{S^\alpha\}$. So, the claim now follows from Proposition 3. □

5 Algorithms

In this section, we present an algorithm to construct the sparse filtration. In previous work, it was shown how to use metric data structures [15] to compute the sparse Rips filtration in $O(n \log n)$ time [21] when the doubling dimension is constant. The same approach also works for the sparse nerve filtrations described here. However, in our implementation, we found it to be efficient to construct the edge set in $O(n^2)$ time and then find the remaining simplices in linear time.

In order to find the edges in the sparse filtration, we consider every two points and determine whether their corresponding balls have a common intersection. If the balls intersect, it returns the birth time of the corresponding edge and ∞ otherwise. We store every edge as a directed edge, which is used to find other k-simplices. Edges are directed from smaller to larger insertion radius. This will allow us to charge the simplices we find to their vertex of minimum insertion radius.

Let $E(v)$ be the vertices adjacent to a vertex v with larger insertion radius. To find a k-simplex for $k > 1$ containing a vertex v, we consider all subsets $\{u_1, \ldots, u_k\}$ of k vertices in $E(v)$. If $\{v, u_1, \ldots, u_k\}$ forms a $(k+1)$-clique, we check the clique to see whether it creates a k-simplex and compute its birth time. The birth time of a k-simplex σ in a nerve filtration is the minimum α such that $U_j^\alpha \neq \emptyset$. If no such α exists, then we define the birth time to be ∞. We assume the user provides a method, SimplexBirthTime, to compute birth times for their metric that runs in time polynomial in k. This function takes a $(k+1)$-clique as input. If at some scale α, the corresponding balls have a common intersection, it returns the minimum such α, otherwise, it returns ∞ indicating the $(k+1)$-clique is not a k-simplex in the sparse filtration.

For the case of Rips filtrations (i.e. ℓ_∞), SimplexBirthTime(σ) needs to compute the maximum birth time of the edges and compare it to $\min_{p_i \in \sigma} \lambda_i(1+\varepsilon)^2/\varepsilon$ (the first time t after which some $p_i \in \sigma$ has $b_i(t) = \emptyset$). For ℓ_2, the corresponding computation is a variation of the minimum enclosing ball problem.

Algorithm 1 finds the k-simplices and birth times in a sparse filtration. Here, $G = (V, E)$ is a directed graph, and the output S is the set of pairs (σ, t), where σ is a k-simplex with birth time t.

Algorithm 1 Find all k-simplices and birth times

procedure FindSimplices$(G = (V, E), k)$
  $S \leftarrow \emptyset$
  for all $v \in V$ **do**
    for all $\{u_1, \ldots, u_k\} \subseteq E(v)$ **do**
      if $\{v, u_1, \ldots, u_k\}$ is a $(k+1)$-clique **then**
        $\sigma \leftarrow \{v, u_1, \ldots, u_k\}$
        $t \leftarrow$SimplexBirthTime(σ)
        if $t < \infty$ **then**
          $S \leftarrow S \cup (\sigma, t)$
  return S

Theorem 5 *Given the edges of a sparse filtration, Algorithm 1 finds the k-simplices of $\{S^\alpha\}$ in $((1+\varepsilon)^2/\varepsilon)^{O(k\rho)} n$ time, where ρ is the doubling dimension of the input metric.*

Proof. In Algorithm 1, for every vertex v in the directed graph G, there are $\binom{|E(v)|}{k}$ subsets with size k. Therefore, if we find an upper bound Δ on the number of adjacent vertices for all $v \in V$, the total running time of the algorithm will be $O(\Delta^k n)$.

In the directed graph G, a vertex p_j is adjacent to vertex p_i if the insertion radius of p_i is less than insertion radius of p_j and their corresponding balls intersect at some scale α, i.e. $b_i(\alpha) \cap b_j(\alpha) \neq \emptyset$. We know that $\lambda_i \leq \lambda_j$, and also they intersect before p_i disappears, so

$$b_i(\lambda_i(1+\varepsilon)^2/\varepsilon) \cap b_j(\lambda_i(1+\varepsilon)^2/\varepsilon) \neq \emptyset.$$

The distance between p_i and p_j is bounded as follows.

$$\begin{aligned}\mathbf{d}(p_i, p_j) &\leq r_i(\lambda_i(1+\varepsilon)^2/\varepsilon) + r_j(\lambda_i(1+\varepsilon)^2/\varepsilon)\\ &\leq \lambda_i(1+\varepsilon)/\varepsilon + \lambda_i(1+\varepsilon)^2/\varepsilon < 2\lambda_i(1+\varepsilon)^2/\varepsilon\end{aligned}$$

Thus, all adjacent vertices to p_i lie in a ball with center p_i and radius $2\lambda_i(1+\varepsilon)^2/\varepsilon$. Moreover, the insertion radii of the neighbors are all at least λ_i, so by a standard packing argument for doubling metrics, $|E(p_i)| = ((1+\varepsilon)^2/\varepsilon)^{O(\rho)}$. Consequently, the running time of this algorithm will be $((1+\varepsilon)^2/\varepsilon)^{O(k\rho)} n$. □

6 Removing Vertices

Because the sparse filtration is a true filtration, no vertices are removed. When the cone is truncated, no new simplices will be added using that vertex, but it is still technically part of the filtration. The linear-size guarantee is a bound on the total number of simplices in the complex. Thus, by using methods such as zigzag persistence or simplicial map persistence to fully remove these vertices when they are no longer needed cannot improve the asymptotic performance. Still, it may be practical to remove them (see [2]). A full theoretical or experimental analysis of the cost tradeoff of using a heavier algorithm to do vertex removal is beyond the scope of this paper.

In this section, we show that the geometric construction leads to a natural choice of elementary simplicial maps (edge collapses) which all satisfy the so-called link condition. In the persistence by simplicial maps work of Dey et al. [11] and Boissonat et al. [1], a key step in updating the data structures to contract an edge is to first add simplices so that the so-called Link Condition is satisfied. The *link* of a simplex σ in a complex K is defined as $\mathrm{Lk}\,\sigma = \{\tau \setminus \sigma \mid \tau \in K \text{ and } \sigma \subseteq \tau\}$. That is, the link σ is formed by removing the vertices of σ from each of its cofaces. An edge $\{u, v\} \in K$ satisfies the *Link Condition* if and only if

$$\mathrm{Lk}\,\{u, v\} = \mathrm{Lk}\,\{u\} \cap \mathrm{Lk}\,\{v\}.$$

Dey et al. [10] proved that edge contractions induce homotopy equivalences when the link condition is satisfied. Thus, it gives a minimal local condition to guarantee that the contraction preserves the topology. More recently, it was shown that such a contraction does not change the persistent homology [11].

Proposition 6 *If* $(P, \mathbf{d})$ *is a finite subset of a convex metric space and* $\{S^\alpha\}$ *is its corresponding sparse filtration, then the last vertex* p_n *has a neighbor* p_i *such that the edge* $\{p_n, p_i\} \in S^\alpha$ *satisfies the link condition, where* $\alpha = \lambda_n(1+\varepsilon)^2/\varepsilon$ *and* λ_n *is the insertion radius of* p_n*.*

Proof. It follows directly from the definition of a link that $\mathrm{Lk}\,\{u,v\} \subseteq \mathrm{Lk}\,\{u\} \cap \mathrm{Lk}\,\{v\}$ for all edges $\{u,v\}$. By the Covering Lemma (Lemma 1), we know that there exists a $p_i \in P$ such that $b_n(\alpha) \subseteq b_i(\alpha)$. Thus, it suffices to check that $\mathrm{Lk}\,\{i\} \cap \mathrm{Lk}\,\{n\} \subseteq \mathrm{Lk}\,\{i,n\}$. Because the vertices are ordered according to a greedy permutation, $\lambda_n \geq \lambda_j$ for all $p_j \in P$. It follows that a simplex $J \in S^\alpha$ if and only if $\bigcap_{i\in J} b_j(\alpha) \neq \emptyset$.

Let J be any simplex in $\mathrm{Lk}\,\{i\} \cap \mathrm{Lk}\,\{n\}$. So, $i, n \notin J$ and $\bigcap_{j\in J\cup\{n\}} b_j(\alpha) \neq \emptyset$. Because $b_n(\alpha) \cap b_i(\alpha) = b_n(\alpha)$, it follows that $\bigcap_{j\in J\cup\{i,n\}} b_j(\alpha) \neq \emptyset$. Thus, we have $J \in \mathrm{Lk}\,\{i,n\}$ as desired. □

7 Conclusion

In this paper, we gave a new geometric perspective on sparse filtrations for topological data analysis that leads to a simple proof of correctness for all convex metrics. By considering a nerve construction one dimension higher, the proofs are primarily geometric and do not require explicit construction of simplicial maps. This geometric view clarifies the non-zigzag construction, while also showing that removing vertices can be accomplished with simple edge contractions.

References

[1] J.-D. Boissonnat, T. K. Dey, and C. Maria. The compressed annotation matrix: An efficient data structure for computing persistent cohomology. In *Algorithms – ESA 2013*, pages 695–706, 2013.

[2] M. B. Botnan and G. Spreemann. Approximating persistent homology in Euclidean space through collapses. *Applicable Algebra in Engineering, Communication and Computing*, pages 1–29, 2015.

[3] M. Buchet, F. Chazal, S. Y. Oudot, and D. R. Sheehy. Efficient and robust persistent homology for measures. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 168–180, 2015.

[4] G. Carlsson and V. de Silva. Zigzag persistence. *Foundations of Computational Mathematics*, 10(4):367–405, 2010.

[5] G. Carlsson, V. de Silva, and D. Morozov. Zigzag persistent homology and real-valued functions. In *Proceedings of the 25th ACM Symposium on Computational Geometry*, pages 247–256, 2009.

[6] N. J. Cavanna, M. Jahanseir, and D. R. Sheehy. A geometric perspective on sparse filtrations. 2015, arXiv:1506.03797.

[7] N. J. Cavanna, M. Jahanseir, and D. R. Sheehy. Visualizing sparse filtrations. In *Proceedings of the 31st Symposium on Computational Geometry (Multimedia Session)*, 2015.

[8] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the 25th ACM Symposium on Computational Geometry*, pages 237–246, 2009.

[9] F. Chazal and S. Y. Oudot. Towards persistence-based reconstruction in Euclidean spaces. In *Proceedings of the 24th ACM Symposium on Computational Geometry*, pages 232–241, 2008.

[10] T. K. Dey, H. Edelsbrunner, S. Guha, and D. V. Nekhayev. Topology preserving edge contraction. *Publications de l'Institut Mathematique (Beograd)*, 60:23–45, 1999.

[11] T. K. Dey, F. Fan, and Y. Wang. Computing topological persistence for simplicial maps. In *Proceedings of the 30th Annual Symposium on Computational Geometry*, pages 345–354, 2014.

[12] Dionysus. By Dmitriy Morozov (`http://www.mrzv.org/software/dionysus/`).

[13] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 4(28):511–533, 2002.

[14] R. Ghrist. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.*, 45(1):61–75, 2008.

[15] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006.

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

[17] M. Kerber and R. Sharathkumar. Approximate Čech complexes in low and high dimensions. In *24th International Symposium on Algorithms and Computation*, volume LNCS 8283, pages 666–676, 2013.

[18] C. Maria and S. Y. Oudot. Zigzag Persistence via Reflections and Transpositions. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, January 2015.

[19] N. Milosavljevic, D. Morozov, and P. Skraba. Zigzag persistent homology in matrix multiplication time. In *Proceedings of the 27th ACM Symposium on Computational Geometry*, 2011.

[20] S. Y. Oudot and D. R. Sheehy. Zigzag zoology: Rips zigzags for homology inference. *Foundations of Computational Mathematics*, pages 1–36, 2014.

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

[22] A. Tausz and G. Carlsson. Applications of zigzag persistence to topological data analysis. arxiv:1108.3545, Aug 2011.

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