Sparse Higher Order Čech Filtrations

Mickaël Buchet
Institute of Geometry, TU Graz, Austria

Bianca B. Dornelas
Institute of Geometry, TU Graz, Austria

Michael Kerber
Institute of Geometry, TU Graz, Austria

Abstract

For a finite set of balls of radius r, the k-fold cover is the space covered by at least k balls. Fixing the ball centers and varying the radius, we obtain a nested sequence of spaces that is called the k-fold filtration of the centers. For $k = 1$, the construction is the union-of-balls filtration that is popular in topological data analysis. For larger k, it yields a cleaner shape reconstruction in the presence of outliers. We contribute a sparsification algorithm to approximate the topology of the k-fold filtration. Our method is a combination and adaptation of several techniques from the well-studied case $k = 1$, resulting in a sparsification of linear size that can be computed in expected near-linear time with respect to the number of input points.

2012 ACM Subject Classification Theory of computation → Computational geometry; Theory of computation → Computational complexity and cryptography

Keywords and phrases Sparsification, k-fold cover, Higher order Čech complexes

Digital Object Identifier 10.4230/LIPIcs.SoCG.2023.20

Related Version *Full Version*: `http://arxiv.org/abs/2303.06666` [9]

Funding This research has been supported by the Austrian Science Fund (FWF), grant numbers P 33765-N and W1230.

Acknowledgements We thank Alexander Rolle for his valuable input and encouragement of this work.

1 Introduction

Persistent homology [10,22,23] is a major branch of topological data analysis with applications, for instance, in shape recognition [5], material science [34] and biology [29,35]. It studies the homological properties of sequences of topological spaces. A standard construction is to take the homogeneous union of balls, with increasing radius, centered at finitely many points of $\mathbb{R}^d$. We call these points *sites* and refer to that filtration as the *union-of-balls filtration*. For computational purposes, one considers the homologically equivalent *Čech filtration*, which is a sequence of simplicial complexes that captures the intersection patterns of the balls in the union-of-balls filtration [22, Chap.3; 30].

The drawback of the Čech filtration (as well as of the closely-related Vietoris-Rips filtration) is that for n sites, it consists of up to $\binom{n}{m+1}$ m-simplices because every $(m+1)$-subset of balls intersects at a sufficiently large radius. A technique to overcome this large size is to *approximate* the Čech (or Vietoris-Rips) filtration with another, much smaller simplicial filtration with similar topological properties. Technically, that means that the persistence modules induced by the homology of the Čech filtration and its approximation are ϵ-interleaved for an arbitrary $\epsilon > 0$ [12]. Several strategies have been devised to construct such approximations with total size linear in n for any fixed ϵ (see related work). Many of these approaches work by selecting only a subset of the simplices of the Čech filtration, in which case we refer to the approximation as a *sparsification*.

The union-of-balls filtration is a special case of the *k-fold filtration* built upon the *k-fold cover*. For n sites in $\mathbb{R}^d$ and $k \geq 1$ fixed, the k-fold cover is the subset of $\mathbb{R}^d$ consisting of points contained in at least k balls of radius r centered at the sites. Besides being a natural extension, k-fold filtrations are tightly related with the kth neighbor distance that arises in the context of outlier removal and processing of non-homogeneous data densities [13, 25, 42, 44]. For that reason, they have received increased attention recently, both with regards to computational [17, 24] and structural aspects [4].

For fixed k, the k-fold filtration can be equivalently expressed by its nerve, which is a simplicial filtration called the *kth order Čech filtration*. It captures the intersection patterns of all k-wise intersections of balls, which we call *lenses*. The aforementioned size issue for Čech filtrations is even more important in the kth order case: the filtration is defined over $\binom{n}{k}$ vertices (one for each k-subset of sites) and consequently consists of $\binom{\binom{n}{k}}{m+1}$ m-simplices, making it unrealistic to compute even for small values of n. Therefore we need to reduce its size considerably while maintaining a good approximation quality.

Contributions. We propose the first sparsification of the k-fold filtration for a fixed k. It is a simplicial filtration that, for n sites in $\mathbb{R}^d$ (with constant d) and a given parameter $\epsilon > 0$, is (multiplicatively) $(1+\epsilon)$-interleaved with the k-fold filtration. Moreover, the number of m-simplices in our sparsification is

$$\mathcal{O}\left(nk^{k(m+1)}\left(\frac{96}{\epsilon}\right)^{\delta k(m+1)}\right), \tag{1}$$

where δ is the doubling dimension of $\mathbb{R}^d$. We point out that for constant k and ϵ, the size of the filtration is linear in the number of sites. This is remarkable because the kth order Čech filtration, which captures the k-fold filtration exactly, already contains $\binom{n}{k}$ vertices. Hence our construction avoids including the vast majority of lenses into the sparsification.

We give an output-sensitive algorithm to compute our sparsification up to dimension m_{max} in

$$\mathcal{O}\left(nk\log n\log\Phi + Xk^{k+1}\left(\frac{96}{\epsilon}\right)^{k\delta}\cdot m_{max}\right)$$

expected time. Here Φ is the spread of the point set (i.e., the ratio of diameter and smallest distance of two distinct points) and X is the size of the output complex, upper bounded by (1) with m replaced by m_{max}. Again considering everything but n as constant, we get a running time of $\mathcal{O}(n\log n)$.

Techniques and related work. The seminal work by Sheehy [45] was the first one to introduce a sparsification technique for Vietoris-Rips filtrations yielding linear size and $\mathcal{O}(n\log n)$ running time (assuming all other parameters as constant). His technique extends to Čech complexes as well with minor adaptations. Subsequent work [6, 8, 11, 19, 46] introduces several extensions, variations, and simplifications of Sheehy's original sparsification; all these works share essentially the same size and complexity bounds.

Our results are achieved by combining several of these techniques used for approximating in the case $k = 1$, which required non-trivial adaptation for larger values of k. The main idea is that for every site p, we define a *removal radius* such that, for radii larger than this removal radius, all lenses involving p are ignored. That means, for larger and larger radii, we construct simplicial complexes with fewer and fewer sites to keep the size small. To determine the removal radii of sites, we introduce the *k-distance permutation* which is an ordering of the sites based on the distance to the kth closest neighbor. The k-distance permutation is a generalization of the farthest point sampling [31] used in some sparsification schemes [19, 46] and induces covering and packing properties analogous to those of nets.

Although we opted to extend sparsification techniques, there is an alternative line of research by Choudhary et al. [14–16] that defines approximations of Čech complexes which are not sparsifications. They arrive at slightly improved bounds than the sparsification for $k = 1$. Approximate filtrations are also actively researched in practice [3, 7, 20, 37, 41].

The k-fold cover and the higher order Čech complexes are also studied with relation to *multiparameter persistence*: considering the order k as a second varying parameter, we obtain the *multicover bifiltration*. Blumberg and Lesnick [4] survey different multiparameter persistence approaches and show a particularly strong stability result for multicovers. Sheehy [44] introduces the barycentric bifiltrations, which is equivalent to the multicover but whose size is prohibitively large. The question of computing the multicover bifiltration exactly has been studied by Edelsbrunner and Osang [25], whose results have been refined by Corbet et al. [17]. The latter authors obtain an equivalent bifiltration to the multicover one but has total size (over all choices of k) $\mathcal{O}(n^{d+1})$ for n points in $\mathbb{R}^d$ [17, Prop. 5]. Their construction rely on using *higher order Voronoi diagrams* and *Delaunay complexes* [26]. That reduces the size of Čech complexes, but cannot lead to linear size without further improvements: the Delaunay filtration's d-skeleton is of size $\mathcal{O}(n^{\lceil d/2 \rceil})$ [43], which is a substantial improvement over the $\mathcal{O}(n^d)$ size of the Čech d-skeleton, but still super-linear for $d \geq 3$. Our approximation foregoes those constructions to reduce the size dependency on n further, with the trade-off that we get an exponential dependency on k.

The size reduction in our construction is a consequence of ignoring lenses after their removal radius. The idea of removing a lens beyond a certain radius is justified geometrically by the fact that the remaining lenses cover its entire area after a certain radius. This is only true, however, if we *freeze* a lens before removing it, that is, keep it unchanged for a short time while the surrounding lenses keep growing. This concept was already introduced in [11], from where we also adapt the elegant technique of lifting the lenses to convex *cones* in $\mathbb{R}^d \times \mathbb{R}$. The additional dimension, which is the radius r, is needed because removing simplices is not possible in filtrations.

The major geometric predicate for our computation is whether a set of balls is intersecting, which can be dualized to computing the radius of the minimal enclosing ball of the ball's centers [21, 27, 28, 38]. However, the aforementioned freezing of lenses makes this problem technically more challenging. This question seems to be unaddressed in previous work, and we give an efficient solution in the Euclidean setting.

Outline. Section 2 provides background definitions and results. Section 3 defines a k-distance and uses it to construct the k-distance permutation of a point set P. In Section 4 the permutation is used to define a sparse lens filtration that approximates the k-fold cover. That results in a nerve filtration that approximates the kth order Čech complex, as shown in Section 5. The size bound of that filtration is given in Section 6. Section 7 provides an algorithm for computing the discrete sparse Čech filtration. We conclude with Section 8.

2 Background

Lenses and k-fold covers. Given a point set $P \subseteq \mathbb{R}^d$ and a fixed $k \in \mathbb{N}$, an element $p \in P$ is called a *site* and a *k-subset* of P is a subset with k sites. Let $\binom{P}{k}$ be the collection of all k-subsets of P and $A \in \binom{P}{k}$. Let also $B_r(a)$ denote the closed ball centered at a of radius r.

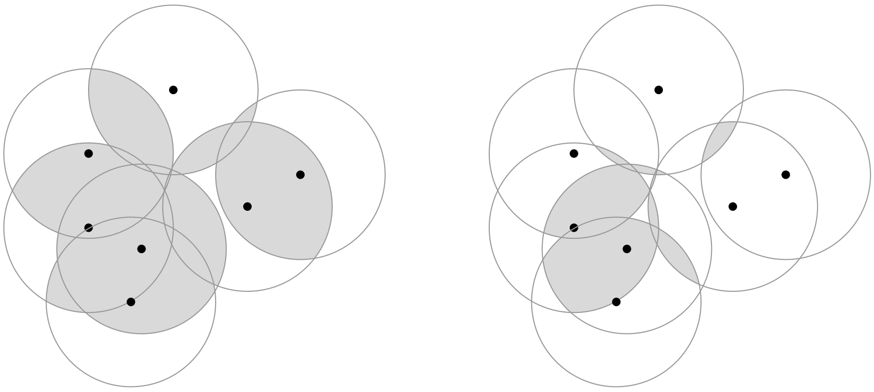

Figure 1 Example of 2- (left) and 3-fold (right) covers for a fixed radius.

The *lens* corresponding to the k-subset A at scale r is

$$L_r(A) := \bigcap_{a \in A} B_r(a).$$

The *k-fold cover* of P at scale r is the union of lenses at scale r over all k-subsets:

$$\mathrm{L}(r, P) := \bigcup_{A \in \binom{P}{k}} L_r(A).$$

See Figure 1 for an example. Note the omission of k, which we consider fixed, in the notation. When P is clear from context, we drop it from the notation as well and write L_r instead.

Nerves. We assume that the reader is familiar with (abstract) simplicial complexes [22, Chap. 3]. For a finite collection C of subsets of $\mathbb{R}^d$, we can define a simplicial complex with vertex set C, called the *nerve* of C, as the set of all subsets of C that have a non-empty mutual intersection. Note that the nerve can contain simplices of larger dimension than d. The nerve of the set of all lenses of k-subsets of P at scale r is called the *kth order Čech complex with radius r* over P, denoted by $\check{\mathrm{C}}\mathrm{ech}_r\,(P, k)$.

Filtrations and equivalence. A collection of topological spaces (e.g., subsets of $\mathbb{R}^d$) $\mathcal{C} = \{C_r\}_{r \geq 0}$ is called a *filtration* if for all $r \leq r'$, it holds that $C_r \subseteq C_{r'}$. The letter r denotes the *scale parameter* of the filtration. For P and k fixed, the previous concepts yield two different ways of obtaining filtrations. On the one hand, since $L_r \subseteq L_{r'}$ for $r \leq r'$, we get the *k-fold filtration* $\mathcal{L} := \{L_r\}_{r \geq 0}$. On the other hand, we observe that $\check{\mathrm{C}}\mathrm{ech}_r\,(P, k)$ is a subcomplex of $\check{\mathrm{C}}\mathrm{ech}_{r'}\,(P, k)$ for $r \leq r'$ and hence we get the *kth order Čech filtration* $\{\check{\mathrm{C}}\mathrm{ech}_r\,(P, k)\}_{r \geq 0}$.

Let $\mathcal{C}$ and $\mathcal{D}$ be two filtrations. We say that $\mathcal{C}$ is *(homotopy) equivalent* [40, Chap. 9] to $\mathcal{D}$ if there exists a family of maps $\{f_r \colon C_r \to D_r\}_{r \geq 0}$ that are homotopy equivalences of spaces and additionally commute with the inclusion maps of $\mathcal{C}$ and $\mathcal{D}$.

Interleaving and approximations. Let $\epsilon \geq 0$. Two filtrations $\mathcal{C}$ and $\mathcal{D}$ are *(multiplicatively) $(1+\epsilon)$-interleaved* if there exist families of linear maps $f_\cdot, g_\cdot$ such that the diagram

$$\begin{tikzcd}
C_r \arrow[rr, hook] \arrow[dr, "f_r"] & & C_{r(1+\epsilon)^2} \arrow[rr, hook] \arrow[dr, "f_{r(1+\epsilon)^2}"] & & C_{r(1+\epsilon)^4} \\
& D_{r(1+\epsilon)} \arrow[ur, "g_{r(1+\epsilon)}"'] \arrow[rr, hook] & & D_{r(1+\epsilon)^3} \arrow[ur, "g_{r(1+\epsilon)^3}"'] &
\end{tikzcd}$$

commutes for all r. Informally, interleaved filtrations with small ϵ are good approximations of each other because every D_r sits in between two instances of $\mathcal{C}$ with close-by scale parameters.

If $\{C_r\}_{r\geq 0}$, $\{D_r\}_{r\geq 0}$ are $(1+\epsilon_1)$-interleaved and $\{D_r\}_{r\geq 0}$, $\{E_r\}_{r\geq 0}$ are $(1+\epsilon_2)$-interleaved, then $\{C_r\}_{r\geq 0}$ and $\{E_r\}_{r\geq 0}$ are $(1+\epsilon_1)(1+\epsilon_2)$-interleaved. Moreover, if $C_r \subseteq D_r \subseteq C_{r(1+\epsilon)}$ for all $r \geq 0$, then $\mathcal{C}$ and $\mathcal{D}$ are $(1+\epsilon)$-interleaved.

A filtration $\mathcal{C}$ is a $(1+\epsilon)$-*approximation* of another filtration $\mathcal{D}$ if there exist filtrations $\mathcal{C}'$ and $\mathcal{D}'$ such that $\mathcal{C}'$ is equivalent to $\mathcal{C}$, $\mathcal{D}'$ is equivalent to $\mathcal{D}$ and $\mathcal{C}'$ and $\mathcal{D}'$ are $(1+\epsilon)$-interleaved. This is a symmetric relationship, so we can say that $\mathcal{C}$ and $\mathcal{D}$ are $(1+\epsilon)$-approximate. If additionally $C_r \subseteq D_r$ for all $r \geq 0$ we call $\mathcal{C}$ a $(1+\epsilon)$-*sparsification* of $\mathcal{D}$. We point out that an approximation between two filtrations implies interleaved *persistence modules* (see [22, Chap. 7]) in the sense of [12].

The Persistent Nerve Theorem. Consider a finite index set I and a family of filtrations $\{U_r^{(i)}\}_{r\geq 0}$ over $\mathbb{R}^d$, one for each $i \in I$. The union filtration is $\{U_r\}_{r\geq 0}$, where $U_r := \bigcup_{i\in I} U_r^{(i)}$, and the nerve filtration is $\{N_r\}_{r\geq 0}$, where N_r is the nerve of U_r. The Persistent Nerve Theorem [2, Thm. 3.9] states that if every $U_r^{(i)}$ is closed and convex, then $\{U_r\}_{r\geq 0}$ and $\{N_r\}_{r\geq 0}$ are equivalent. As a consequence, the Persistent Nerve Theorem implies that the k-fold and the kth order Čech filtrations are equivalent: choose I as the set of all k-subsets of P and $U_r^{(i)}$ as the lens indexed by i at radius r, which is a closed and convex set.

Doubling dimension. The *doubling constant* Δ of $\mathbb{R}^d$ is such that any ball of radius r can be covered with at most Δ balls of radius $r/2$, for all $r \geq 0$. The *doubling dimension* of $\mathbb{R}^d$ is $\delta := \log_2 \Delta$, which is of order $\Theta(d)$ and hence constant for this paper. Note that for finite point sets in $\mathbb{R}^d$ the doubling dimension can be significantly smaller than d, for instance if the points all lie close to a low-dimensional subspace.

To cover a ball B of radius r with balls of radius $r/4$, one needs at most Δ^2 balls; with balls of radius $r/8$ one needs Δ^3 balls and so on. Thus, to cover B with balls of radius r', we have to find the smallest t such that $r/2^t \leq r'$. That is $t = \lceil \log_2 r/r' \rceil$. Then, $\Delta^t \leq \Delta^{\log_2 r/r'+1} = 2^\delta \left(r/r'\right)^\delta$ and $(2r/r')^\delta$ balls of radius r' are sufficient to cover B.

Quadtreaps. A *quadtreap* [39] is a dynamic data structure for spherical range search. We summarize its properties in a simplified form suitable for us: for a set X of n points in $\mathbb{R}^d$ (with d constant), it can be built in $\mathcal{O}(n \log n)$ expected time. It supports deletions of points in X in expected $\mathcal{O}(\log n)$ time. Moreover, given a query point q and a radius r, it returns a list $S \subseteq X$ which is guaranteed to contain all points in X of distance $\leq r$ from q, and is guaranteed not to contain any point in X of distance $\geq 2r$ from q. The running time for such a query is $\mathcal{O}(\log n + |S|)$.

3 k-distance permutation

Given some integer $k \geq 1$ and a finite data set $P \subseteq \mathbb{R}^d$ of $n \geq k$ sites, we define an order on the points in P in which the sites are denoted by $p_1, \ldots, p_n$. Writing $P_i := \{p_1, \ldots, p_i\}$, our order ensures that the k-fold cover over P_i approximates the k-fold cover over P, with increasing approximation quality when i increases.

The k-*distance* of $x \in \mathbb{R}^d$ to P, denoted by $\mathrm{d}^k(x, P)$, is the distance from x to its kth closest neighbor in P. We define the k-*distance permutation* incrementally as follows: we choose $p_1, \ldots, p_k$ as arbitrary, pairwise distinct sites from P. If $p_1, \ldots, p_{i-1}$ are chosen for

$k < i \leq n$, we set

$$p_i := \operatorname*{argmax}_{q \in P \setminus P_{i-1}} \mathrm{d}^k(q, P_{i-1}).$$

Note that for $k = 1$, we obtain the well-known farthest point sampling. We also define

$$\lambda_i := \mathrm{d}^k(p_i, P_{i-1})$$

for $k+1 \leq i \leq n$ and set $\lambda_1, \ldots, \lambda_k$ to ∞, so that the sequence $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ is non-increasing. The next two properties of the k-distance permutation are reminiscent of the packing and covering properties of ϵ-nets [47, Chap. 14].

▶ **Lemma 1** (Covering). *For all $k \leq i \leq n-1$, we have $\mathrm{L}(r, P_i) \subseteq L(r, P) \subseteq \mathrm{L}(r + \lambda_{i+1}, P_i)$.*

Proof. Recall the notation $L_r = L(r, P)$. $P_i \subseteq P$ immediately implies $\mathrm{L}(r, P_i) \subseteq L_r$. Consider $x \in L_r$. Then, $x \in L_r(A)$ for some $A = \{a_1, a_2, \ldots, a_k\} \subseteq P$. If $A \subseteq P_i$, the result follows. Otherwise, without loss of generality let $a_1 \notin P_i$. By definition of λ_{i+1}, $\mathrm{d}^k(a_1, P_i) \leq \lambda_{i+1}$ and hence there are sites $b_1, b_2, \ldots, b_k \in P_i$ with $\mathrm{d}(a_1, b_j) \leq \lambda_{i+1}$ for all $1 \leq j \leq k$. Consequently, $\mathrm{d}(x, b_j) \leq \mathrm{d}(x, a_1) + \mathrm{d}(a_1, b_j) \leq r + \lambda_{i+1}$ and the k closest sites to x in P_i are within distance $r + \lambda_{i+1}$ of x, implying $x \in \mathrm{L}(r + \lambda_{i+1}, P_i)$. ◀

▶ **Lemma 2** (Packing). *For all $k+1 \leq i \leq n$, each $p \in P_i$ has $\mathrm{d}^k(p, P_i \setminus \{p\}) \geq \lambda_i/2$.*

Proof. We do induction on i. For $i = k+1$, let q be the kth closest neighbor of p_{k+1} in P_k. We have $\mathrm{d}^k(p_{k+1}, P_{k+1} \setminus \{p_{k+1}\}) = \lambda_{k+1} \geq \lambda_{k+1}/2$ and, for any $p \in P_{k+1} \setminus \{p_{k+1}\}$,

$$\mathrm{d}^k(p, P_{k+1} \setminus \{p\}) = \max_{p' \in P_{k+1} \setminus \{p\}} \mathrm{d}(p, p') \geq \frac{\mathrm{d}(p, q) + \mathrm{d}(p, p_{k+1})}{2} \geq \frac{\mathrm{d}(q, p_{k+1})}{2} = \frac{\lambda_{k+1}}{2}.$$

Hence the statement is true for $i = k+1$. Next we assume, for some $i \geq k+1$, that for every $p \in P_i$, $\mathrm{d}^k(p, P_i \setminus \{p\}) \geq \lambda_i/2$, and show the statement for $i+1$.

For p_{i+1}, we have $\mathrm{d}^k(p_{i+1}, P_{i+1} \setminus \{p_{i+1}\}) = \lambda_{i+1} \geq \lambda_{i+1}/2$ and the statement follows. Consider $p \in P_{i+1} \setminus \{p_{i+1}\}$. If p_{i+1} is not among the k nearest neighbors of p in P_{i+1}, then

$$\mathrm{d}^k(p, P_{i+1} \setminus \{p\}) = \mathrm{d}^k(p, P_i \setminus \{p\}) \geq \frac{\lambda_i}{2} \geq \frac{\lambda_{i+1}}{2}$$

by the induction hypothesis and because the λ-values are non-increasing. Otherwise, p_{i+1} is among the k nearest neighbors of p in $P_{i+1} \setminus \{p\}$ and $\mathrm{d}^k(p, P_{i+1} \setminus \{p\}) \geq \mathrm{d}(p, p_{i+1})$.

If $\mathrm{d}(p, p_{i+1}) \geq \lambda_{i+1}/2$, the claim follows. Otherwise, every site at distance smaller than $\lambda_{i+1}/2$ of p is at distance smaller than λ_{i+1} of p_{i+1}. Since $\lambda_{i+1} = \mathrm{d}^k(p_{i+1}, P_i)$, there can be at most $k-2$ sites of $P_i \setminus \{p\}$ at distance smaller than λ_{i+1} of p_{i+1}. Thus, counting p_{i+1} as well, there can be at most $k-1$ sites of $P_{i+1} \setminus \{p\}$ at distance smaller than $\lambda_{i+1}/2$ of p and it follows that $\mathrm{d}^k(p, P_{i+1} \setminus \{p\}) \geq \lambda_{i+1}/2$. ◀

Computation. We give a simple algorithm for computing the k-distance permutation that has quadratic running time in the number of input points and discuss an approach for improving it. We call a site *ordered* if it has already been assigned its index in the k-distance permutation and *unordered* otherwise.

The simple approach is the following. Pick k sites $p_1, \ldots, p_k$ and compute, for each $y \in P \setminus P_k$, the distances from y to p_i, $1 \leq i \leq k$. Store them in a max-heap T_y that also has a fixed entry identifying y. Up until this point we need $\mathcal{O}(nk)$ time. The next steps are repeated iteratively. For all unordered y, group the T_y in a list L. When $p_1, \ldots, p_{i-1}$

are chosen, the algorithm picks p_i by scanning over L and choosing the point with largest distance to its kth nearest ordered neighbor, which takes $\mathcal{O}(n)$ time. When p_i is picked and becomes ordered, remove its entry from L. Then, by traversing all remaining elements in L, identify each unordered y with p_i among y's k nearest neighbors in P_i and insert $\mathrm{d}(p_i, y)$ to T_y. The $(k+1)$-distance from y to the ordered sites, which was a previous entry in T_y, is removed. This takes $\mathcal{O}(\log k)$ time per element of L and hence $\mathcal{O}(n \log k)$ per iteration. Since there are n iterations, the total running time is of $\mathcal{O}(n^2 \log k)$.

This simple algorithm can be improved with the main insight that when p_i is determined, the kth nearest ordered neighbor of all remaining unordered sites is at most λ_i away. Hence, unordered sites further than λ_i away from p_i do not have to be updated. Whenever a site p_i is ordered, we can employ a quadtreap (Section 2) to only update the unordered sites within distance λ_i. This last step can also be done in general metric spaces with elementary but rather tedious techniques; see [33, Sec. 3.1]. Using the packing property, the total number of updates reduces to $\mathcal{O}(nk \log \Phi)$ (with a constant that depends exponentially on the doubling dimension of the point set). Finally, we replace the list L by a max-heap to avoid the linear scan to search for the next ordered points. Appendix A of [9] provides further details on how to achieve this improvement, which results in the next theorem.

▶ **Theorem 3.** *The k-distance permutation can be computed in expected time $\mathcal{O}(nk \log n \log \Phi)$ with Φ the spread of the point set.*

4 A sparse union of lenses

We define several spaces in this section and the following. Figure 2 has an overview.

Recall that the k-fold cover is defined as the union of all lenses at radius r, where every lens is given by k sites. For large values of r, most of these lenses intersect, yielding a size explosion in its nerve, the kth order Čech complex. At the same time, many lenses are eventually covered by the union of other lenses and so may be removed from consideration.

To define the precise threshold for removal of a lens, recall that in Section 3 we ordered the sites as $p_1, \ldots, p_n$ and obtained values $\lambda_1, \lambda_2, \ldots, \lambda_k = \infty$, $\lambda_{k+1} \geq \ldots \geq \lambda_n$. Fix $\epsilon \in (0, 1]$. Since it is fixed, we drop ϵ from the upcoming notation. The *freezing radius* of a site p_i is

$$\mathrm{frz}\,(p_i) := \frac{(1+\epsilon)\lambda_i}{\epsilon}.$$

We extend the definition to lenses by setting $\mathrm{frz}\,(A) = \min_{p \in A} \mathrm{frz}\,(p)$. Then at radius r we only consider lenses whose freezing radius is at least r and set

$$U_r := \bigcup_{\mathrm{frz}(A) \geq r} L_r(A).$$

Notice that $\mathcal{U} := \{U_r\}_{r \geq 0}$ is *not* a filtration: Figure 3 illustrates that U_r might not be a subset of $U_{r'}$ for $r < r'$. Even so, some useful inclusions hold as we see on the next lemma.

▶ **Lemma 4.** $U_r \subseteq L_r \subseteq U_{(1+\epsilon)r}$.

Proof. The first inclusion is clear. For the second inclusion, consider $x \in L_r$ and let i be the maximal index such that

$$r \leq \frac{\lambda_i}{\epsilon}. \qquad (*)$$

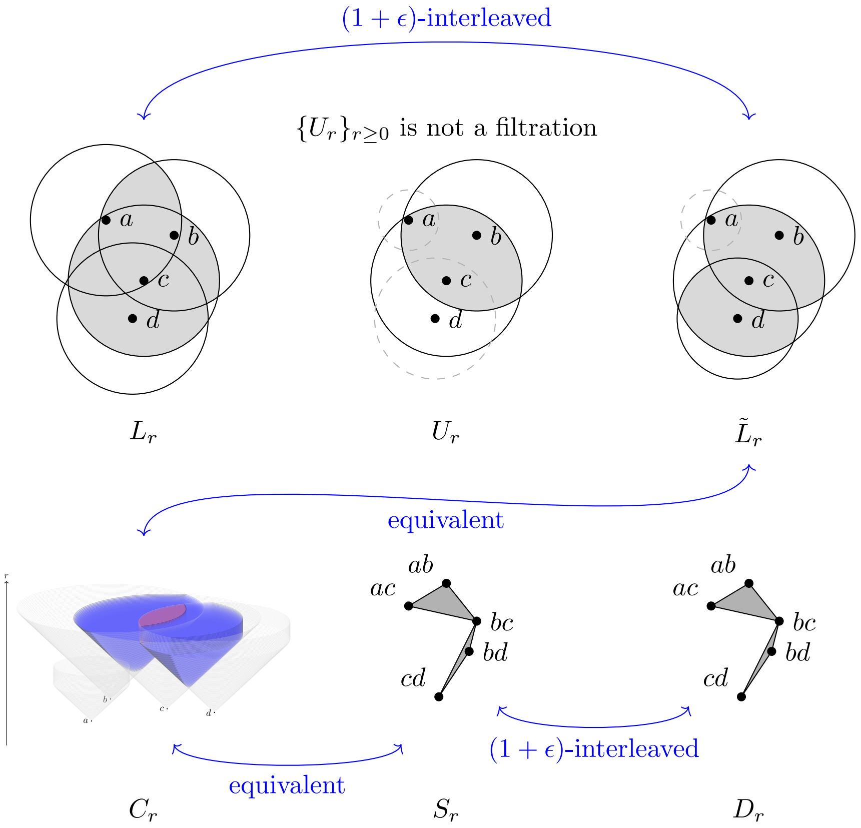
$(1+\epsilon)$-interleaved
$\{U_r\}_{r\geq 0}$ is not a filtration
a
b
c
d
L_r
U_r
$\tilde{L}_r$
equivalent
ab
ac
bc
bd
cd
$(1+\epsilon)$-interleaved
equivalent
C_r
S_r
D_r

Figure 2 Schematical view of the different filtrations introduced in Sections 4 and 5. We consider $k = 2$, $r > (1+\epsilon)\,\mathrm{frz}\,(a)$ and $r \in (\mathrm{frz}\,(d)\,, (1+\epsilon)\,\mathrm{frz}\,(d)]$.

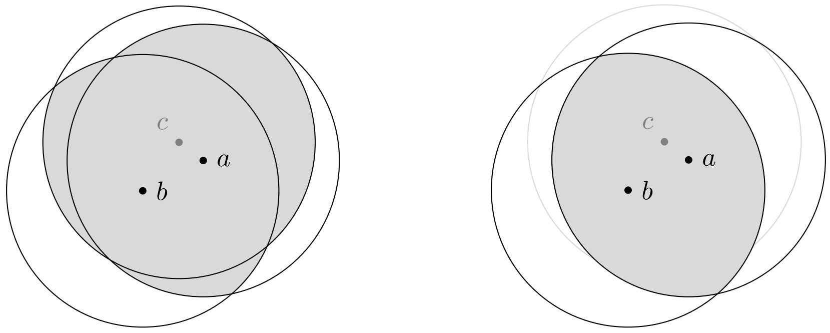

Figure 3 Example in $\mathbb{R}^2$ with $k = 2$. *Left:* U_r at radius $r = \mathrm{frz}\,(c)$. *Right:* $U_{r'}$ at radius immediately after $\mathrm{frz}\,(c)$. Even though $r < r'$, $U_r \nsubseteq U_{r'}$.

If $i = n$, then there is $A \subseteq P_n$ with $x \in L_r(A)$ because $x \in L_r$ and $P = P_n$. By definition of the freezing radius and inequality $(*)$, $\mathrm{frz}\,(A) \geq \mathrm{frz}\,(p_i) = (1+\epsilon)\lambda_i/\epsilon \geq r(1+\epsilon)$ and thus $L_{r(1+\epsilon)}(A) \subseteq U_{(1+\epsilon)r}$. Since $L_r(A) \subseteq L_{(1+\epsilon)r}(A)$, the result follows.

For $i < n$, notice that the Covering Property (Lemma 1) guarantees that x is contained in a lens $L_{r+\lambda_{i+1}}(A)$ for some $A \subseteq P_i$. Since i is maximal, $\lambda_{i+1}/\epsilon < r$ and so $L_{r+\lambda_{i+1}}(A) \subseteq L_{(1+\epsilon)r}(A)$. Moreover, $A \subseteq P_i$ and inequality $(*)$ imply $\mathrm{frz}\,(A) \geq \mathrm{frz}\,(p_i) \geq (1+\epsilon)r$. Hence the lens of A contributes to $U_{(1+\epsilon)r}$ and as it contains x, the statement follows. ◀

This lemma suggests that a $(1+\epsilon)$-interleaving with the filtration $\mathcal{L} = \{L_r\}_{r\geq 0}$ consisting of all lenses should be possible if we adjust $\mathcal{U}$ to obtain an actual filtration. To do that we slightly delay the removal of lenses. More precisely, we define

$$\tilde{L}_r(A) := \begin{cases} L_r(A) & r < \operatorname{frz}(A) \\ L_{\operatorname{frz}(A)}(A) & \operatorname{frz}(A) \leq r \leq (1+\epsilon)\operatorname{frz}(A) \\ \emptyset & (1+\epsilon)\operatorname{frz}(A) < r. \end{cases}$$

One can visualize the evolution of a lens as a continuous process for increasing r: the lens $\tilde{L}_r$ grows until it reaches its freezing radius and remains unchanged (it is "frozen") for the interval $[\operatorname{frz}(A), (1+\epsilon)\operatorname{frz}(A)]$. Afterwards it completely disappears. We call $(1+\epsilon)\operatorname{frz}(A)$ the *removal radius* of A. The construction is an adaptation of a similar one by Sheehy [45].

We write $\tilde{L}_r$ for the union of $\tilde{L}_r(A)$ over all $A \in \binom{P}{k}$ and $\tilde{\mathcal{L}} := \{\tilde{L}_r\}_{r\geq 0}$. We show next that $\tilde{\mathcal{L}}$ is a filtration.

▶ **Lemma 5.** *$\tilde{\mathcal{L}}$ is a filtration, i.e., for any $r \leq r'$, $\tilde{L}_r \subseteq \tilde{L}_{r'}$.*

Proof. If the interval $(r, r']$ does not contain any removal radius, $\tilde{L}_r \subseteq \tilde{L}_{r'}$ because the inclusions hold lens-wise. Since the number of different removal radii is bounded by the number of sites and hence finite, it suffices to show that at a removal radius s, any lens that is removed is already covered by lenses that are not being removed at s. In fact, we show that such a lens is covered by lenses that are not yet frozen at s.

Let A be the k-subset associated with a lens being removed at s and $x \in \tilde{L}_s(A)$. By definition, $s = (1+\epsilon)t$ with $t = \operatorname{frz}(A)$. This implies that $x \in \tilde{L}_t(A) = L_t(A) \subseteq L_t$ because the lens is frozen from radius t on. By Lemma 4, it follows that $x \in U_{(1+\epsilon)t} = U_s$, and therefore x is contained in a lens $L_s(B)$ with $\operatorname{frz}(B) \geq s$. Thus $\tilde{L}_s(A)$ is covered by lenses $L_s(B)$ with $\operatorname{frz}(B) \geq s$ and $\tilde{\mathcal{L}}$ is a filtration. ◀

▶ **Lemma 6.** *$\tilde{\mathcal{L}}$ and $\mathcal{L}$ are $(1+\epsilon)$-interleaved.*

Proof. We show that $\tilde{L}_r \subseteq L_r \subseteq \tilde{L}_{(1+\epsilon)r}$. Note that by definition, $\tilde{L}_r \subseteq L_r$. For the second inclusion, observe that $U_r \subseteq \tilde{L}_r$ follows directly from their definition. Then, Lemma 4 yields $L_r \subseteq U_{(1+\epsilon)r} \subseteq \tilde{L}_{(1+\epsilon)r}$. ◀

5 A sparse simplicial filtration

Since $\tilde{L}_r(A)$ is closed and convex for every r, the nerve of all (non-empty) $\tilde{L}_r(A)$ yields a simplicial complex with the same homotopy type as the k-fold cover at radius r. However, the collection of simplicial complexes obtained when varying r does not form a filtration because simplices disappear from the nerve when passing a removal radius. To overcome this problem, we adapt a construction of Cavanna et al. [11] that is similar to a function's graph.

Cones. The idea is to "stack-up" the lenses $\tilde{L}_r(A)$ for all radii: the *cone* of A at radius r is

$$C_r(A) := \bigcup_{\alpha\in[0,r]} \left(\tilde{L}_\alpha(A) \times \{\alpha\}\right) \subseteq \mathbb{R}^d \times \mathbb{R}.$$

We write C_r for the union of $C_r(A)$ over all $A \in \binom{P}{k}$. Figure 4 shows one cone.

▶ **Lemma 7.** *The filtrations $\mathcal{C} = \{C_r\}_{r\geq 0}$ and $\tilde{\mathcal{L}}$ are equivalent.*

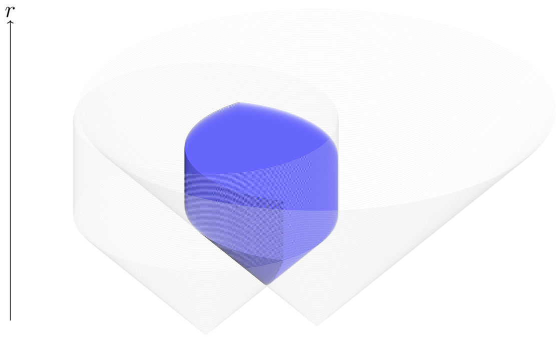

Figure 4 A cone representing the evolution of a lens for $k = 2$. At first the lens grows, until it is frozen. Then, the lens has static size and afterwards it disappears. At the radius where the lens disappears, it is completely covered by other lenses (which are not displayed in the figure).

Proof. In the same sense as above C_r is a stacked-up version of $\tilde{L}_\alpha$ for all $\alpha \leq r$, and we can consider $\tilde{L}_r$ as a subspace of C_r via the map $x \mapsto (x, r)$ for $x \in \tilde{L}_r$. Since $\tilde{\mathcal{L}}$ is a filtration, there is a strong deformation retraction R from C_r to $\tilde{L}_r$, given by $R((x, \alpha), t) = (x, (1-t)\alpha + tr)$, which naturally commutes with the canonical inclusions. The result follows. ◀

We define the nerve of the cones as the *sparse kth order Čech complex*,

$$S_r := \operatorname{Nrv}\left\{C_r(A) \mid A \in \binom{P}{k}\right\}.$$

S_r is a subcomplex of the kth order Čech complex for every r because $\tilde{L}_r \subseteq L_r$. Moreover, if $r \leq r'$, $C_r(A) \subseteq C_{r'}(A)$ for all A. Hence $\mathcal{S} = \{S_r\}_{r\geq 0}$ is a filtration. By the Persistent Nerve Theorem [2, Thm. 3.9] and Lemmas 6 and 7 we obtain:

▶ **Lemma 8.** *The filtrations $\mathcal{S}$ and $\mathcal{C}$ are equivalent. As a consequence, $\mathcal{S}$ is a $(1+\epsilon)$-approximation of the k-fold filtration.*

Discretization of the radius. The filtration $\mathcal{S}$ is challenging to compute, due to the freezing of lenses. We elaborate on these issues in Section 7. We now define a variant of $\mathcal{S}$ which is easier to compute and also is $(1+\epsilon)$-interleaved with the k-fold filtration.

Recall that the filtrations $\tilde{\mathcal{L}}$, $\mathcal{C}$ and $\mathcal{S}$ are defined based on the freezing radii of sites, which depend on a parameter $\epsilon > 0$. To obtain a $(1+\epsilon)$-approximation for $\epsilon \in (0, 1]$ in the end, we consider the above construction of $\mathcal{S}$ with parameter $\epsilon' = \frac{\epsilon}{3}$, obtaining a $(1 + \frac{\epsilon}{3})$-approximation of the k-fold filtration.

Next, for every $r \geq 0$, let $z \in \mathbb{Z}$ be such that $(1+\frac{\epsilon}{3})^z \leq r < (1+\frac{\epsilon}{3})^{z+1}$ and define

$$D_r := S_{(1+\epsilon/3)^z}.$$

We call $\mathcal{D} := \{D_r\}_{r\geq 0}$ the *discrete sparse kth order Čech filtration*. It is formed by a discrete set of snapshots of $\mathcal{S}$ and kept unchanged except when passing over a snapshot radius (this is also referred to as the *Left Kan extension* of a discrete filtration [36, Chap. 10]).

▶ **Theorem 9.** *$\mathcal{D}$ is a $(1+\epsilon)$-approximation of the k-fold filtration.*

Proof. From the definition, $D_r \subseteq S_r \subseteq D_{(1+\epsilon/3)r}$. This interleaving implies that $\mathcal{D}$ is a $(1+\frac{\epsilon}{3})$-approximation of $\mathcal{S}$. Since $\mathcal{S}$ is a $(1+\frac{\epsilon}{3})$-approximation of the k-fold filtration, by transitivity, we get that D_r is a $(1+\frac{\epsilon}{3})^2$-approximation of the k-fold filtration. The result follows by noting that $(1+\frac{\epsilon}{3})^2 \leq 1+\epsilon$ for all $\epsilon \in (0, 1]$. ◀

6 Size analysis

We bound the size of $\mathcal{D}$, i.e., the number of simplices it contains. Since $D_r \subseteq S_r$ for all $r \geq 0$, it is enough to bound the size of $\mathcal{S}$ with parameter $\epsilon' = \epsilon/3$. Let

$$C_\infty(A) := \bigcup_{r\geq 0} C_r(A)$$

be the *cone of* A (without dependence on a radius). Then, the size of $\mathcal{S}$ equals the size of the nerve of the cones $C_\infty(A)$, where A ranges over all k-subsets of sites. However, the number of vertices is not necessarily $\binom{n}{k}$ because many cones are empty: this happens in particular when the smallest radius for which the balls around the sites of A intersect is larger than the removal radius of the lens. In fact, our argument shows that this is the case for the vast majority of cones. The proof for vertices extends readily to the case of m-wise intersections of cones, i.e. for $(m-1)$-simplices, without change and thus we treat the general case directly.

For fixed $m \geq 1$, we derive an upper bound for the number of sets $\{A_1, \ldots, A_m\}$ such that the cones $C_\infty(A_1), \ldots, C_\infty(A_m)$ intersect. Such sets are in one-to-one correspondence to the $(m-1)$-simplices of the sparse kth order Čech filtration, hence we refer to these sets as $(m-1)$-simplices. Let $\sigma = \{C_\infty(A_1), \ldots, C_\infty(A_m)\}$ be a $(m-1)$-simplex and the set of sites P be ordered according to the k-distance permutation. We call a site p_i *involved* in σ if p_i belongs to one of the sets $A_1, \ldots, A_m$. Note that there are at most km sites involved in σ. We say that σ is *associated* to a site p_i if p_i is involved in σ and all other involved sites have index smaller than i. Our strategy is to upper bound the number of $(m-1)$-simplices associated to an arbitrary p_i. We only need to consider simplices associated to p_i that appear in the filtration, i.e., simplices whose defining cones intersect.

Fix p_i and $\omega_i := (1+\epsilon')\,\mathrm{frz}\,(p_i)$. Let B denote the ball of radius $2\omega_i$ centered at p_i.

▶ **Lemma 10.** *If $\sigma := \{C_\infty(A_1), \ldots, C_\infty(A_m)\}$ is an $(m-1)$-simplex associated to p_i whose cones intersect, then all sites involved in σ are contained in B.*

Proof. Let α denote the minimal radius such that all the balls around sites involved in σ intersect. This is the radius of the minimum enclosing ball of the involved sites. Any common intersection of the cones must happen at scale $r \geq \alpha$.

On the other hand, assume wlog that $p_i \in A_1$. Since p_i has maximal index in A_1, we have that $\mathrm{frz}\,(A_1) = \mathrm{frz}\,(p_i)$. Hence the removal radius of A_1 is equal to ω_i and it follows that the cone of A_1 is empty for all radii greater than ω_i. Therefore any common intersection of the cones of σ must happen at scale $r \leq \omega_i$. Hence, as we assume that the cones do intersect, we must have that $\alpha \leq \omega_i$.

Then, since $\mathrm{d}(q, p_i) \leq 2\alpha$, any involved site q lies within distance $2\omega_i$ from p_i. ◀

Hence the involved sites of σ are close to p_i in the sense of the lemma. We can furthermore guarantee that the points of P_i are not too densely packed in B using the packing property of the k-distance permutation.

▶ **Lemma 11.** *The ball B contains at most $\Gamma := k\left(\frac{96}{\epsilon}\right)^\delta$ sites of $P_i = \{p_1, \ldots, p_i\}$, where δ is the doubling dimension of $\mathbb{R}^d$.*

Proof. We cover B by balls of radius $\lambda_i/4$. That can be done with at most $\zeta = \left(\frac{16(1+\epsilon')^2}{\epsilon'}\right)^\delta$ balls (see *Doubling Dimension* in Section 2). By the Packing Lemma 2, each open ball of radius $\lambda_i/4$ contains at most k sites of P_i, thus the total number of sites in B is at most $k\zeta$. The bound follows because $\epsilon' = \frac{\epsilon}{3}$ and $\epsilon \leq 1$, hence $(1+\epsilon')^2 \leq \frac{16}{9} < 2$. ◀

Bounding the number of non-empty m-intersections of cones is now a matter of simple combinatorics. Recall the Γ notation from Lemma 11.

▶ **Theorem 12.** *The number of $(m-1)$-simplices of the sparse kth order Čech complex with non-empty cone intersection is at most*

$$n \cdot \Gamma^{km} = nk^{km}\left(\frac{96}{\epsilon}\right)^{\delta km}.$$

Proof. Fix p_i. Every $(m-1)$-simplex associated to p_i with non-empty cone intersection has up to km involved sites, which all lie in B by Lemma 10. Moreover, all involved sites are in P_i and, by Lemma 11, there are at most Γ of those sites in B to choose from. It follows that there are at most Γ^{km} different choices possible. This upper bound holds for every p_i, so multiplying by the number of sites n yields the result. ◀

We remark that the bounds on this section are not tight and slightly better ones could be easily achieved, by keeping binomials in place or avoiding some approximations. However the improvements would be minor.

7 Computation

We now present an algorithm to construct the discrete sparse Čech filtration. As in the previous sections, let us fix a finite set $P \subseteq \mathbb{R}^d$, an integer $k > 0$ and $\epsilon \in (0, 1]$. Assume that $P = \{p_1, \ldots, p_n\}$ has the indices ordered with respect to the k-distance permutation, and that we have computed the corresponding values $\lambda_1, \ldots, \lambda_n$ as discussed in Section 3. The algorithm outputs the discrete sparse kth order Čech filtration $\{D_r\}_{r\geq 0}$ as a list of simplices with their corresponding *critical value*, i.e., the smallest parameter value r for which the simplex is part of the filtration. Note that by definition of the discrete sparse Čech filtration, every critical value is of the form $(1+\epsilon/3)^z$ for some integer z.

Friends. Our algorithm follows the approach and notation of Section 6. For every p_i, we compute all simplices associated to p_i in the filtration together with their critical value. To do so, we first find, among $p_1, \ldots, p_{i-1}$, all sites of distance at most $2\omega_i$ from p_i, where $\omega_i = (1+\epsilon')\,\mathrm{frz}\,(p_i)$ (compare Lemma 10). We call these points *friends* of p_i. We compute friends using a quadtreap data structure, as introduced in Section 3, which we query for every p_i at $2\omega_i$. p_i is added to the quadtreap after the ith iteration (adding an element costs $\mathcal{O}(\log n)$ in expectation as well). Hence the expected running time for this loop is $\mathcal{O}(n \log n + \Sigma)$, where Σ is the number of reported points. These reported points have distance at most $4\omega_i$ from the respective p_i (since the queries are approximate), and by the same argument as in Lemma 11, the number of sites reported for p_i is at most $k\left(\frac{192}{\epsilon}\right)^{\delta} = \mathcal{O}(k(1/\epsilon)^{\delta})$. We traverse the list and remove all “false friends” of distance more than $2\omega_i$. Thus we get the friends of p_i for all sites p_i in expected time

$$\mathcal{O}(n \log n + nk(1/\epsilon)^{\delta}). \tag{2}$$

Note that the number of friends is bounded by Γ as defined in Lemma 11.

Vertices. Next, we compute the vertices of the filtration associated to p_i, for each $i \geq k$. We proceed by brute-force, just enumerating all k-tuples formed by p_i and $k-1$ of its friends and checking for every k-tuple whether their cone is non-empty. The last condition is simple to check, as the cone is non-empty if and only if the radius α of the minimum enclosing ball of the k sites is at most $\operatorname{frz}(p_i)$. In this case, the critical radius of the vertex is set to $(1+\epsilon/3)^z$, where z is the smallest integer such that $(1+\epsilon/3)^z \geq \alpha$. Computing α and z per vertex requires $\mathcal{O}(k)$ in expectation [18]. Hence we calculate all vertices in $\mathcal{O}((n-k)k\Gamma^{k-1})$ time.

Simplices. For higher-dimensional simplices associated to p_i, we proceed inductively by dimension, up to a maximal dimension m_{max}. Fix a $(m-1)$-simplex $\sigma = \{A_1, \ldots, A_m\}$ associated to p_i. We compute all cofacets of σ in the filtration, that is, all m-simplices that contain σ and one further vertex A_{m+1}. Notice that one could order the k-subsets and avoid computing all cofacets, computing instead only cofacets with larger index in the ordering. This would remove a k factor from the computation expected time locally, but does not change the final bound in Theorem 13, which has a k^k factor. We compute all cofacets for simplicity. Since each element of A_{m+1} is either a friend of p_i or p_i itself, we enumerate all k-tuples consisting of these sites and check whether they form a vertex of the filtration. This takes $\mathcal{O}(k)$ time per vertex, just by re-doing the check from the previous step, except that k-tuples associated to p_j must be checked at radius $\min\{\operatorname{frz}(p_j), \omega_i\}$. That is because if the k-tuple cone becomes non empty only after ω_i, then it cannot contribute to a coface of σ.

For a vertex A_{m+1} of the filtration, check whether the cones of $A_1, \ldots, A_{m+1}$ intersect is technically challenging because the cones might intersect for a radius where one or several cones are frozen. One cannot resolve this question by a simple minimal enclosing ball computation. In fact, we are not aware of an efficient way to compute the smallest intersection radius of such cones in general. However, as demonstrated in [9, App. B], given a collection of cones $A_1, \ldots, A_{m+1}$ and a fixed radius r, we can decide whether the cones (or rather, the corresponding lenses $\tilde{L}_r$) intersect at radius r by a reduction to a minimum enclosing ball of balls instance [28] in expected $\mathcal{O}(k(m+1))$ time. We use this predicate and query whether the cones intersect at the smallest removal radius of $A_1, \ldots, A_{m+1}$. That decides whether the m-simplex is in the filtration, and its critical radius can be computed by two more minimal enclosing ball computations. This is only possible because we have discretized the filtration, as discussed in [9, App. B]. It follows that the expected running time spent per $(m-1)$-simplex of the filtration is $\mathcal{O}(mk\Gamma^k)$, and doing this over all simplices of the filtration up to dimension m_{max} yields a total expected complexity of $\mathcal{O}(X \cdot m_{max} \cdot k \cdot \Gamma^k)$ for this step, where X is the total number of simplices in the filtration.

This concludes the description of the algorithm. Recalling (2) we obtain a complexity of

$$\mathcal{O}\left(n\log n + \frac{nk}{\epsilon^\delta} + (n-k)k\Gamma^{k-1} + X \cdot m_{max} \cdot k \cdot \Gamma^k\right),$$

where the second and third terms are dominated by the last one because $X \geq n-k$ (see [9, App. C]). Together with the algorithm from Section 3, we arrive at the result

▶ **Theorem 13.** *Given a set P of n points, $k \geq 0$ and $\epsilon \in (0,1]$, the discrete sparse kth order Čech filtration up to dimension m_{max} can be computed in time*

$$\mathcal{O}\left(nk\log n\log\Phi + Xk^{k+1}m_{max}\left(\frac{96}{\epsilon}\right)^{k\delta}\right),$$

where Φ is the spread of P and X is the total number of simplices in the filtration.

8 Conclusions

We introduced the first $(1+\epsilon)$-approximate filtration of the higher order k-fold filtration and provided an algorithm for computing it. If k and ϵ are considered as constants and the input point set has constant spread, the algorithm runs in time $\mathcal{O}(n \log n)$ and yields a filtration of size $\mathcal{O}(n)$, which are the same favorable properties of the well-studied case $k = 1$.

There are various avenues to strengthen and generalize our results. First of all, our method has concentrated on the Euclidean case, but our approach mostly generalizes to point sets in arbitrary metric spaces – the algorithm cannot use the quadtreap data structure anymore, but there is no need for it, since the algorithm by Har-Peled and Mendel [33, Sec. 3.1] can be adapted to the k-distance case with little effort. Also, the friends of p_i (Section 7) can be computed with a slight adaptation of their techniques; we used quadtreaps mostly for the ease of presentation. However, the computation of critical values of simplices described in [9, App. B] is for the Euclidean case only, and the complexity of this step remains unspecified for a general metric space. This is common in related work; see, for instance [11, Sec. 5].

Another natural goal is to remove the dependence on the spread. This dependence is caused by the computation of the k-distance permutation which is inspired by the algorithm of [33, Sec. 3.1]. In the same paper [33, Sec. 3.2–3.3], they describe an approach to remove the spread from the bound (for $k = 1$) using an approximate version of the greedy permutation. While our construction of the sparsified filtration can be easily adapted to work with an approximate version of the k-distance permutation, it seems less straight-forward to generalize the computation to the k-distance, even in the Euclidean case. We leave this for future work.

The k-distance permutation relates to the *Distance to Measure* (DTM) [13], which is the square average of the distances to the k nearest neighbors. The DTM has the advantage of being robust, in terms of the Wasserstein distance, to perturbations on the sample [13, Sec. 3]. However, most of the existing methods for sparsifying filtrations obtained via the DTM [1,8,32] require a preliminary approximation by weighted distances. Our approach might be adaptable to directly sparsify DTM filtrations.

While we concentrate on the case of a single value of k, we pose the question whether our methods can be used to approximate the multi-cover bifiltration, as studied in [17]. The extension is not straight-forward because there is no direct relation between the approximate filtrations on level k and $k+1$. We speculate that the technique of double-nerve constructions of [17] could be useful in this context. The presence of an exponential factor on k in our size bounds suggests a restriction of our approach's usability to small portions of the bifiltration, for small k. The exponential factor on k also carries over to the expected computation time. Reducing that dependency on k is another possible line of future work. Note that [17, Prop. 5] gives a size bound of $\mathcal{O}\left(n^{d+1}\right)$ for the exact version, but we ask whether a polynomial bound on k could be achieved without such a blow-up in the dependency on n.

Finally, a natural question is the practicality of our algorithm. We remark that even for $k = 1$, while some work has been devoted to practical aspects of computing sparsifications [3, 7, 20, 37, 41], the actual practical computation is still an unresolved problem. We think that the natural order for a practically efficient solution would be to first identify best practices in the simpler $k = 1$ case and subsequently try to adapt them to larger values of k. So, while we would be curious about the performance of our algorithm, such an evaluation seems to be premature at the moment.

References

1 Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. Dtm-based filtrations. In Gill Barequet and Yusu Wang, editors, *35th International Symposium on Computational Geometry, SoCG 2019, June 18-21, 2019, Portland, Oregon, USA*, volume 129 of *LIPIcs*, pages 58:1–58:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.SoCG.2019.58.

2 Ulrich Bauer, Michael Kerber, Fabian Roll, and Alexander Rolle. A unified view on the functorial nerve theorem and its variations, 2022. arXiv:2203.03571.

3 Nello Blaser and Morten Brun. Sparse nerves in practice. In Andreas Holzinger, Peter Kieseberg, A M. Tjoa, and Edgar R. Weippl, editors, *Machine Learning and Knowledge Extraction - CD-MAKE 2019, Canterbury, UK, August 26-29, 2019, Proceedings*, volume 11713 of *Lecture Notes in Computer Science*, pages 272–284. Springer, 2019. doi:10.1007/978-3-030-29726-8_17.

4 Andrew J. Blumberg and Michael Lesnick. Stability of 2-parameter persistent homology. *CoRR*, 2020. arXiv:2010.09628.

5 Thomas Bonis, Maks Ovsjanikov, Steve Oudot, and Frédéric Chazal. Persistence-based pooling for shape pose recognition. In Alexandra Bac and Jean-Luc Mari, editors, *Computational Topology in Image Context - 6th International Workshop, CTIC 2016, Marseille, France, June 15-17, 2016, Proceedings*, volume 9667 of *Lecture Notes in Computer Science*, pages 19–29. Springer, 2016. doi:10.1007/978-3-319-39441-1_3.

6 Magnus B. Botnan and Gard Spreemann. Approximating persistent homology in euclidean space through collapses. *Applicable Algebra in Engineering, Communication and Computing*, 26(1-2):73–101, January 2015. doi:10.1007/s00200-014-0247-y.

7 Bernhard Brehm and Hanne Hardering. Sparips. *CoRR*, 2018. arXiv:1807.09982.

8 Mickaël Buchet, Frédéric Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust persistent homology for measures. *Comput. Geom.*, 58:70–96, 2016. doi:10.1016/j.comgeo.2016.07.001.

9 Mickaël Buchet, Bianca B. Dornelas, and Michael Kerber. Sparse higher order Čech filtrations, 2023. arXiv:2303.06666.

10 Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009. doi:10.1090/S0273-0979-09-01249-X.

11 Nicholas J. Cavanna, Mahmoodreza Jahanseir, and Donald R. Sheehy. A geometric perspective on sparse filtrations. In *Proceedings of the 27th Canadian Conference on Computational Geometry, CCCG 2015, Kingston, Ontario, Canada, August 10-12, 2015*. Queen's University, Ontario, Canada, 2015. URL: http://research.cs.queensu.ca/cccg2015/CCCG15-papers/01.pdf.

12 Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Oudot. Proximity of persistence modules and their diagrams. In John Hershberger and Efi Fogel, editors, *Proceedings of the 25th ACM Symposium on Computational Geometry, Aarhus, Denmark, June 8-10, 2009*, pages 237–246. ACM, 2009. doi:10.1145/1542362.1542407.

13 Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Found. Comput. Math.*, 11(6):733–751, 2011. doi:10.1007/s10208-011-9098-0.

14 Aruni Choudhary, Michael Kerber, and Sharath Raghvendra. Improved topological approximations by digitization. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2675–2688. SIAM, 2019. doi:10.1137/1.9781611975482.166.

15 Aruni Choudhary, Michael Kerber, and Sharath Raghvendra. Polynomial-sized topological approximations using the permutahedron. *Discret. Comput. Geom.*, 61(1):42–80, 2019. doi:10.1007/s00454-017-9951-2.

16 Aruni Choudhary, Michael Kerber, and Sharath Raghvendra. Improved approximate rips filtrations with shifted integer lattices and cubical complexes. *J. Appl. Comput. Topol.*, 5(3):425–458, 2021. doi:10.1007/s41468-021-00072-4.

17 René Corbet, Michael Kerber, Michael Lesnick, and Georg Osang. Computing the multicover bifiltration. *Discret. Comput. Geom.*, 2023. doi:10.1007/s00454-022-00476-8.

18 Mark de Berg, Otfried Cheong, Marc J. van Kreveld, and Mark H. Overmars. *Computational geometry: algorithms and applications, 3rd Edition.* Springer, 2008. URL: https://www.worldcat.org/oclc/227584184.

19 Tamal K. Dey, Fengtao Fan, and Yusu Wang. Computing topological persistence for simplicial maps. In Siu-Wing Cheng and Olivier Devillers, editors, *30th International Symposium on Computational Geometry, SoCG 2014, Kyoto, Japan, June 08 - 11, 2014*, page 345. ACM, 2014. doi:10.1145/2582112.2582165.

20 Tamal K. Dey, Dayu Shi, and Yusu Wang. Simba: An efficient tool for approximating rips-filtration persistence via simplicial batch collapse. *ACM J. Exp. Algorithmics*, 24(1):1.5:1–1.5:16, 2019. doi:10.1145/3284360.

21 Martin E. Dyer. A class of convex programs with applications to computational geometry. In David Avis, editor, *Proceedings of the Eighth Annual Symposium on Computational Geometry, Berlin, Germany, June 10-12, 1992*, pages 9–15. ACM, 1992. doi:10.1145/142675.142681.

22 Herbert Edelsbrunner and John L. Harer. *Computational topology.* American Mathematical Society, Providence, RI, 2010. An introduction. doi:10.1090/mbk/069.

23 Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discret. Comput. Geom.*, 28(4):511–533, 2002. doi:10.1007/s00454-002-2885-2.

24 Herbert Edelsbrunner and Georg Osang. A simple algorithm for higher-order delaunay mosaics and alpha shapes. *CoRR*, 2020. arXiv:2011.03617.

25 Herbert Edelsbrunner and Georg Osang. The multi-cover persistence of euclidean balls. *Discret. Comput. Geom.*, 65(4):1296–1313, 2021. doi:10.1007/s00454-021-00281-9.

26 Herbert Edelsbrunner and Raimund Seidel. Voronoi diagrams and arrangements. *Discret. Comput. Geom.*, 1:25–44, 1986. doi:10.1007/BF02187681.

27 Kaspar Fischer. *Smallest enclosing balls of balls: Combinatorial structure & algorithms.* PhD thesis, Swiss Federal Institute of Technology, ETH Zürich, 2005. URL: https://people.inf.ethz.ch/emo/DoctThesisFiles/fischer05.pdf.

28 Kaspar Fischer and Bernd Gärtner. The smallest enclosing ball of balls: combinatorial structure and algorithms. *Int. J. Comput. Geom. Appl.*, 14(4-5):341–378, 2004. doi:10.1142/S0218195904001500.

29 Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, and Vidit Nanda. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32(1):1–17, 2015. doi:10.1007/s13160-014-0153-5.

30 Robert W. Ghrist. *Elementary applied topology*, volume 1. Createspace Seattle, WA, 2014.

31 Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi:10.1016/0304-3975(85)90224-5.

32 Leonidas J Guibas, Quentin Mérigot, and Dmitriy Morozov. Witnessed k-distance. *Discret. Comput. Geom.*, 49(1):22–45, 2013. doi:10.1007/s00454-012-9465-x.

33 Sariel Har-Peled and Manor Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput.*, 35(5):1148–1184, 2006. doi:10.1137/S0097539704446281.

34 Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G. Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26):7035–7040, 2016. doi:10.1073/pnas.1520877113.

35 Lida Kanari, Pawel Dlotko, Martina Scolamiero, Ran Levi, Julian C. Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018. doi:10.1007/s12021-017-9341-1.

36 Saunders Mac Lane. *Categories for the working mathematician.*, volume 5. New York, NY: Springer, 1998. doi:10.1007/978-1-4757-4721-8.

37 Clément Maria, Pawel Dlotko, Vincent Rouvreau, and Marc Glisse. Rips complex. In *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.5.0 edition, 2022. URL: `https://gudhi.inria.fr/doc/3.5.0/group__rips__complex.html`.

38 Nimrod Megiddo. On the ball spanned by balls. *Discret. Comput. Geom.*, 4:605–610, 1989. `doi:10.1007/BF02187750`.

39 David M. Mount and Eunhui Park. A dynamic data structure for approximate range searching. In David G. Kirkpatrick and Joseph S. B. Mitchell, editors, *Proceedings of the 26th ACM Symposium on Computational Geometry, Snowbird, Utah, USA, June 13-16, 2010*, pages 247–256. ACM, 2010. `doi:10.1145/1810959.1811002`.

40 James Munkres. *Topology*. Prentice Hall, 2ed edition, 2000.

41 Julian B. Pérez, Sydney Hauke, Umberto Lupo, Matteo Caorsi, and Alberto Dassatti. Giotto-ph: A python library for high-performance computation of persistent homology of vietoris-rips filtrations. *CoRR*, 2021. `arXiv:2107.05412`.

42 Jeff M. Phillips, Bei Wang, and Yan Zheng. Geometric inference on kernel density estimates. In Lars Arge and János Pach, editors, *31st International Symposium on Computational Geometry, SoCG 2015, June 22-25, 2015, Eindhoven, The Netherlands*, volume 34 of *LIPIcs*, pages 857–871. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2015. `doi:10.4230/LIPIcs.SOCG.2015.857`.

43 Raimund Seidel. On the number of faces in higher-dimensional voronoi diagrams. In D. Soule, editor, *Proceedings of the Third Annual Symposium on Computational Geometry, Waterloo, Ontario, Canada, June 8-10, 1987*, pages 181–185. ACM, 1987. `doi:10.1145/41958.41977`.

44 Donald R. Sheehy. A multicover nerve for geometric inference. In *Proceedings of the 24th Canadian Conference on Computational Geometry, CCCG 2012, Charlottetown, Prince Edward Island, Canada, August 8-10, 2012*, pages 309–314, 2012. URL: `http://2012.cccg.ca/papers/paper52.pdf`.

45 Donald R. Sheehy. Linear-size approximations to the vietoris-rips filtration. *Discret. Comput. Geom.*, 49(4):778–796, 2013. `doi:10.1007/s00454-013-9513-1`.

46 Donald R. Sheehy. A sparse delaunay filtration. In Kevin Buchin and Éric Colin de Verdière, editors, *37th International Symposium on Computational Geometry, SoCG 2021, June 7-11, 2021, Buffalo, NY, USA (Virtual Conference)*, volume 189 of *LIPIcs*, pages 58:1–58:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. `doi:10.4230/LIPIcs.SoCG.2021.58`.

47 Wilson A. Sutherland. *Introduction to metric and topological spaces*. Oxford Mathematics Series. Oxford University Press, 2 edition, 2009.