(Multi)Filtering Noise in Geometric Persistent Homology*

Donald R. Sheehy†

Abstract

The beauty of persistent homology for topological data analysis is that it obviates the need to choose an explicit scale at which to view the data. Not only does one skip the problem of tuning parameters, but also the output shows explicitly which features are robust to perturbations of scale. Unfortunately, de-noising the data as a preprocess often leads to new parameters to choose. We show how to replace the Euclidean distance with a family of distance functions to de-noise the data as part of the persistence computation. The result is an instance of multidimensional persistence where we can tell not only what topological features are present but also how robust they are to changes in the de-noising parameters.

1 Introduction

In practice, topological inference has three phases: one statistical, one geometric, and the third, topological. First, the data is filtered for noise. Second, the geometry of the points drives the construction of a filtered simplicial complex. Third, the persistent homology of the filtered complex is computed. Usually, the emphasis is placed on the latter two phases with the first treated as a necessary evil. And it *is* necessary; even a small number of outliers can generate spurious persistent features that foil existing methods.

The process of de-noising the data introduces a new set of parameters, one for the scale at which to define density and one to threshold between signal and noise. So, although no explicit scale is chosen to compute the persistent homology, one *is* chosen to de-noise the data. The problem is both aesthetic and practical. It is both more elegant to do all three phases without tuning any parameters, and it is also be more useful, as de-noising parameters can be difficult to choose for topological applications (see [7]).

In our approach, we replace the usual distance to the input set P, with the kth nearest neighbor distance, d_k. This replaces the α-*offsets*, $P^\alpha = \bigcup_{p\in P} \text{ball}(p,\alpha)$ with the (k,α)-*offsets*, $P_k^\alpha = d_k^{-1}(-\infty,\alpha]$ as our estimate of the shape at scale α.

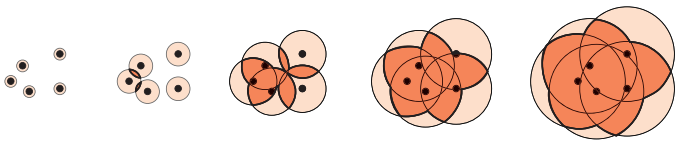

Figure 1: The α-offsets overlaid with the $(2,\alpha)$-offsets.

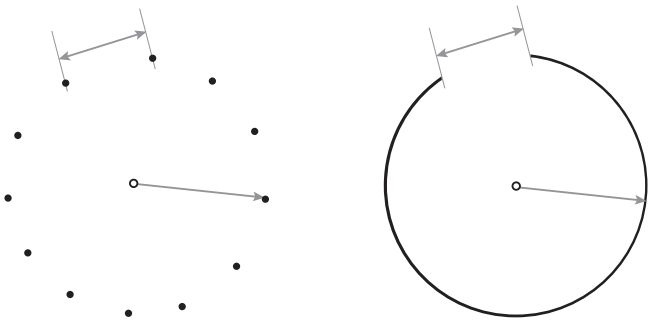

Figure 2: Two inputs yielding persistent cycles with the same birth and death times. The cycle on the left uses relatively few points while the cycle on the right uses unboundedly many. Our approach allows us to tell the difference between these two cases.

Our contributions:

- We present *barycentric multifiltrations*, a natural way to filter a filtered complex by a second parameter. We then show that the barycentric mutifiltration of the Čech filtration captures exactly the persistent homology of the (k,α)-offsets.
- We prove approximation guarantees for mesh-based filtrations of more general smooth functions. This allows us to construct a multifiltered complex that approximates the (k,α)-offsets that is the same size as the corresponding mesh filtration for the α-offsets. Perhaps surprisingly, for $k>1$, both the filtration and its analysis are simpler than for the usual distance function, d_1. Thus these results are both simpler to understand and implement while also being more generally applicable.

*This work was partially supported by the National Science Foundation under grant number CCF-0635257.

†Department of Computer Science, Carnegie Mellon University, dsheehy@cs.cmu.edu

2 Background

Filtrations and Persistent Homology. The 0th, 1st, and 2nd homology groups describe the connected components, holes, and bubbles in a space respectively. Similarly, the higher order homology groups describe the non-bounding cycles of higher dimension.

A *filtration* is a nested family of topological spaces, parameterized by single variable. In *persistent homology*, the changes in homology over the course of a filtration are computed. Instead of a static snapshot of the topology of F^α, we get a movie of the topological changes in $\{F^\alpha\}_{\alpha\geq 0}$ as α grows. Persistent homology can be computed for several different types of filtrations in time polynomial in the complexity of the filtration (see [9] for more a primer on persistence).

If a space is filtered by more than one parameter, we have a *multifiltration*. There are polynomial-time algorithms for multidimensional persistence as well as for the one-dimensional case [1].

Persistence Diagrams. The output of the persistence algorithm is a *persistence diagram* that marks each homology class with a point in the plane, using the birth and death times as the x and y coordinates. Features that persist for a long time, those with a large gap between birth and death times, appear far from the diagonal $y = x$, whereas short-lived features (topological noise) concentrate around this diagonal.

The theory of approximate persistence diagrams is derived from the stability of persistence diagrams [4]. For stability, the goal is to show that two similar inputs yield similar outputs. For approximation, we replace one of these inputs with the true filtration that we want to approximate, P_k^α in our case, and argue that our approximate filtration will produce a persistence diagram that is provably close to the persistence diagram of the true of filtration. The main result we use for this is a special case of the Strong Stability Theorem of Chazal et al. [2], rephrased into the language of multiplicative approximations (see also [6]).

Theorem 1 (Chazal et al. [2])**.** *Let $\{F^\alpha\}$ and $\{G^\alpha\}$ be two tame filtrations. If $F^{\alpha/c} \subseteq G^\alpha \subseteq F^{c\alpha}$ for all $\alpha \geq 0$, then the persistence diagram of $\{F^\alpha\}$ is a c-approximation to the persistence diagram of $\{G^\alpha\}$.*

Distance Functions and Offsets. Let $d_P(x)$ be the distance from x to the nearest points of P. The sublevel $d_P^{-1}(-\infty, \alpha]$ is called the α-*offsets*, and is denoted P^α. Equivalently, the α-offsets are the union of closed α-balls centered at points of P.

A simple modification gives a distance function that is more robust to outliers. Define the *kth nearest neighbor distance*, d_k, to be the distance to k points of P. The sublevels of d_k are the (k, α)-offsets, denoted P_k^α:

$$P_k^\alpha = d_k^{-1}(-\infty, \alpha].$$

Equivalently, the (k, α)-*offsets* are the points contained in at least k balls of radius α centered at points in P (see Figure 1). The family of sets $\{P_k^\alpha\}_{k,\alpha}$ is a filtration in both α and $|P| - k$. We are interested in computing the persistence diagram of the (k, α)-offsets.

Simplicial Complexes. Simplicial complexes discretize the topological spaces in a filtration. An *abstract simplicial complex* is a family of subsets of a vertex set V that is closed under taking subsets. A set $\sigma \subset V$ of a simplicial complex is called a *simplex* and the *dimension* of σ is defined to be $|\sigma - 1|$, where $|\cdot|$ denotes cardinality. The subsets of σ are its *faces*.

A simplicial complex, $\mathcal{K}$, is *filtered* if there is an assignment of nonnegative real numbers to simplices such that every simplex is assigned a value greater than or equal to that of its faces, i.e. $t : \mathcal{K} \to \mathbb{R}$ such that $t(\sigma') \leq t(\sigma)$ for all $\sigma' \subset \sigma$. Thus, we get a filtration $\{\mathcal{K}^\alpha\}_{\alpha\geq 0}$, where $\mathcal{K}^\alpha = \{\sigma \in K : t(\sigma) \leq \alpha\}$.

Nerves and Barycentric Decomposition. Given a collection of closed sets U, the *nerve* of U is a simplicial complex with vertex set U and simplices for subsets of U with a common intersection. The Nerve Theorem states that the nerve of collection of sets is homotopy equivalent to their union if all intersections of finitely many sets are either empty or contractible. Such a family U is called a *good closed cover* of $\bigcup_{u\in U} u$. The Persistent Nerve Lemma of Chazal and Oudot [3] allows us to move easily between filtrations on geometric spaces and filtered simplicial complexes.

Given a simplicial complex, $\mathcal{K}$, the *barycentric decomposition*, $\tilde{\mathcal{K}}$, is a new simplicial complex with vertex set $\mathcal{K}$ and simplices $\{\sigma_0, \ldots, \sigma_j\} \subset \mathcal{K}$ whenever $\sigma_0 \subset \cdots \subset \sigma_j$. Every simplicial complex is topologically equivalent to its barycentric decomposition.

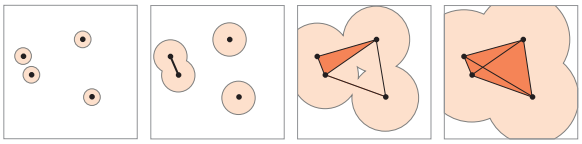

Figure 3: The Čech filtration and the α-mesh filtration.

The Čech filtration. The *Čech complex* at scale α, $\mathcal{C}^\alpha$, is the nerve of the set $\{\text{ball}(p, \alpha)\}_{p\in P}$. If α is at least half the diameter of P, the Čech complex contains every possible simplex, i.e. $\mathcal{C}^{\text{diam}(P)/2} = 2^P$. To avoid this combinatorial blowup in the size of the complex, it is common to truncate the filtration at some maximum scale, $\alpha_{\max}$. It is also common to only consider simplices up to the dimension of the ambient space.

3 Barycentric Multifiltration

The vertices of the barycentric decomposition $\tilde{\mathcal{K}}$ of a complex $\mathcal{K}$ are simplices of $\mathcal{K}$ and thus have a dimension associated with them. To avoid confusion we refer to this as the *number* of the vertex. This leads to a natural filtration on $\tilde{\mathcal{K}}$ defined to be $\{\tilde{\mathcal{K}}_k\}_k$, where $\tilde{\mathcal{K}}_k$ is the subcomplex induced on the vertices numbered at least $k-1$ (the filter parameter here goes down rather than up but this is not a problem). If we have a filtered complex $\{\mathcal{K}^\alpha\}_\alpha$, then we can apply this method to form the *barycentric multifiltration*, $\{\tilde{\mathcal{K}}_k^\alpha\}_{k,\alpha}$.

The barycentric Čech complex Let $\tilde{\mathcal{C}}^\alpha$ be the barycentric decomposition of the Čech complex at scale α, and let $\{\tilde{\mathcal{C}}_k^\alpha\}_{k,\alpha}$ be its barycentric multifiltration. The following theorem establishes a topological equivalence between $\tilde{\mathcal{C}}_k^\alpha$ and P_k^α.

Theorem 2. *The barycentric Čech complex, $\tilde{\mathcal{C}}_k^\alpha$, is homotopy equivalent to the (k,α)-offsets, P_k^α, for all $\alpha \geq 0$ and all $k \in \mathbb{N}$.*

Proof. Let $\mathcal{N}_k^\alpha$ be the nerve of all k-wise intersections of α-balls centered at points of P and let $\tilde{\mathcal{N}}_k^\alpha$ denote its barycentric decomposition. By the Nerve Theorem, $\mathcal{N}_k^\alpha$ is homotopy equivalent P_k^α and therefore $\tilde{\mathcal{N}}_k^\alpha$ is also. It will suffice to demonstrate a homotopy equivalence from $\tilde{\mathcal{N}}_k^\alpha$ to $\tilde{\mathcal{C}}_k^\alpha$. We will first show that $\tilde{\mathcal{C}}_k^\alpha$ can be identified with a subcomplex of $\tilde{\mathcal{N}}_k^\alpha$. Next, we will show that the desired homotopy is a deformation retraction onto this subcomplex induced by a projection of the vertex set.

We start by defining an injective map from $\tilde{\mathcal{C}}_k^\alpha$ to $\tilde{\mathcal{N}}_k^\alpha$. A vertex of $\tilde{\mathcal{C}}_k^\alpha$ corresponds to a collection S of at least k points whose α-balls intersect. It follows that $\binom{S}{k}$ corresponds to a vertex in $\tilde{\mathcal{N}}_k$. This map from vertices S in $\tilde{\mathcal{C}}_k^\alpha$ to vertices $\binom{S}{k} \in \tilde{\mathcal{N}}_k^\alpha$ extends naturally to the higher order simplices.

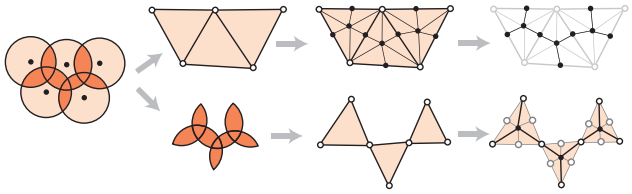

Figure 4: We transform the collection of balls in two different ways to get homotopy equivalent complexes, $\tilde{\mathcal{C}}_k^\alpha$ (top) and $\tilde{\mathcal{N}}_k^\alpha$ (bottom) for $k = 2$.

Now, we will give a surjective map from $\tilde{\mathcal{N}}_k^\alpha$ to $\tilde{\mathcal{C}}_k^\alpha$, again defined by mapping the vertices. Let $V = \{V_0, \ldots, V_j\} \subset \binom{P}{k}$ be a vertex of $\tilde{\mathcal{N}}_k^\alpha$. Since it is a vertex, this means that $\bigcap_{i=0}^{j} \bigcap_{p \in V_i} \text{ball}(p,\alpha) \neq \emptyset$. This is equivalent to the statement that $\bigcap_{p \in Q_V} \text{ball}(p,\alpha) \neq \emptyset$, where $Q_V = \bigcup_{V_i \in V} V_i$. This fact along with the observation that $|Q_V| \geq k$ imply that Q_V is a vertex of $\tilde{\mathcal{C}}_k^\alpha$. We map all such vertices V in $\tilde{\mathcal{N}}_k^\alpha$ to their corresponding vertices Q_V in $\tilde{\mathcal{C}}_k^\alpha$. This map extends easily to the higher order simplices.

The composition of the two maps takes vertices $V \in \tilde{\mathcal{N}}_k^\alpha$ to vertices $V' = \binom{Q_V}{k} \in \tilde{\mathcal{N}}_k^\alpha$. For each such vertex V, there is a corresponding subset of $\mathbb{R}^d$, $S_V = \bigcap_{u \in V} \bigcap_{p \in u} \text{ball}(p,\alpha)$. We can observe that S_V and $S_{V'}$ are identical:

$$S_V = \bigcap_{u \in V} \bigcap_{p \in u} \text{ball}(p,\alpha) = \bigcap_{p \in Q_V} \text{ball}(p,\alpha)$$
$$= \bigcap_{u \in \binom{Q_V}{k}} \bigcap_{p \in u} \text{ball}(p,\alpha) = S_{V'}. \tag{3.1}$$

The simplices $\{V_0, \ldots, V_j\}$ of $\tilde{\mathcal{N}}_k^\alpha$ are those families of k-element sets of P such that $S_{V_0} \subseteq \cdots \subseteq S_{V_j}$. So, Equation (3.1) implies that if σ is a simplex of $\tilde{\mathcal{N}}_k^\alpha$ containing V then $\sigma \cup \{V'\}$ is also a simplex of $\tilde{\mathcal{N}}_k^\alpha$. Thus, the projection that takes V to V' for all vertices in $\tilde{\mathcal{N}}_k^\alpha$ induces a homotopy equivalence, because it merely projects simplices to a faces on their boundary. □

This homotopy equivalence can be combined with standard topological methods to yield the following theoretical guarantee.

Theorem 3. *For any fixed k, the persistence diagram of the barycentric Čech filtration, $\{\tilde{\mathcal{C}}_k^\alpha\}$, is identical to the persistence diagram of the (k,α)-offsets, $\{P_k^\alpha\}$.*

4 Filtrations on Meshes

In this section we will construct a filtered simplicial complex whose persistence diagram is provably close to that of (k,α)-offsets. The advantage of this approximation is that it has only linear size, whereas the barycentric Čech filtration could have size doubly exponential in n.

Let M be a superset of the input set P and let $\text{Vor}(M)$ denote its Voronoi diagram. Let $\text{Vor}(v)$ be the Voronoi cell of a vertex $v \in M$ restricted to some compact bounding box that contains M. The *in-radius*, r_v, is the radius of the largest ball centered at v contained in $\text{Vor}(v)$. The *out-radius*, R_v, is the radius of the smallest ball centered at v containing $\text{Vor}(v)$. The *aspect ratio* of $\text{Vor}(v)$ is $\frac{R_v}{r_v}$. We say M is ρ-well-spaced if every Voronoi cell has aspect ratio at most ρ. A set M is *ε-refined* if for every vertex v in M, the outer radius R_v is at most $\varepsilon d_2(v)$, where d_2 is measured with respect to P. Given n points P, standard Voronoi refinement meshing algorithms can produce a superset

M of P that is both ρ-well-spaced and ε-refined [5, 6]. Moreover, for reasonable inputs, $|M| = O(n)$[1].

For any function $f : \mathbb{R}^d \to \mathbb{R}$, we can construct the following filtrations based on f, $\mathrm{Vor}(M)$, and $\mathrm{Del}(M)$:

1. The *sublevel filtration*: $F^\alpha = f^{-1}(-\infty, \alpha]$.
2. The *Voronoi filtration*: $\mathrm{V}^\alpha = \bigcup_{\substack{v \in M \\ f(v) \le \alpha}} \mathrm{Vor}(v)$.
3. The *Delaunay filtration*: $\mathcal{D}^\alpha = \{\sigma \in \mathrm{Del}(M) : \forall v \in \sigma, f(v) \le \alpha\}$.

The Persistent Nerve Lemma implies that the Voronoi and Delaunay filtrations are homotopy equivalent. We will follow the pattern that algorithms operate on the Delaunay filtration and proofs work with the Voronoi filtration.

Given a function f that is sufficiently large and sufficiently smooth, the Delaunay filtration on a ρ-well-spaced, ε-refined set M gives a constant factor approximation to the persistence diagram of the sublevel filtration of f. The following two technical lemmas show how the Voronoi filtration is interleaved with the sublevel filtration.

Lemma 4. *If for all $v \in M$ and all $x \in \mathrm{Vor}(v)$, $\frac{1}{c} f(x) \le f(v) \le c f(x)$ for some constant $c \ge 1$, then $\mathrm{V}^{\alpha/c} \subseteq F^\alpha \subseteq \mathrm{V}^{c\alpha}$, for all $\alpha \ge 0$.*

Proof. First we prove that $\mathrm{V}^{\alpha/c} \subseteq F^\alpha$. If x is a point in $\mathrm{V}^{\alpha/c}$ then $f(v) \le \alpha/c$. It follows that $f(x) \le \alpha$, and so $x \in F^\alpha$. Next, we prove that $F^\alpha \subseteq \mathrm{V}^{c\alpha}$. If x is in F^α then $f(x) \le \alpha$ and thus $f(v) \le c\alpha$. It follows that $\mathrm{Vor}(v) \subset \mathrm{V}^{c\alpha}$, and so $x \in \mathrm{V}^{c\alpha}$. □

Lemma 5. *If M is ε-refined and f is a t-Lipschitz function with $f \ge d_2$, then*

$$\frac{1}{1+\varepsilon_0} f(x) \le f(v) \le (1+\varepsilon_0) f(x).$$

for all $v \in M$ and $x \in \mathrm{Vor}(v)$, where $\varepsilon_0 = \frac{t\varepsilon}{1-t\varepsilon}$.

Proof. Let $v \in M$ and $x \in \mathrm{Vor}(v)$ be chosen arbitrarily. Then we may bound $f(x)$ as follows.

$$\begin{aligned} f(x) &\le f(v) + t|v - x| && [\text{f is } t\text{-Lipschitz}] \\ &\le f(v) + t\varepsilon d_2(v) && [R_v \le \varepsilon d_2(v)] \\ &\le (1 + t\varepsilon) f(v) && [d_2 \le f] \\ &< (1 + \varepsilon_0) f(v). && [\varepsilon_0 > t\varepsilon] \end{aligned}$$

[1]For unreasonable inputs, other tricks can guarantee linear size (see [8])

Similarly, we can bound $f(v)$:

$$\begin{aligned} f(v) &\le f(x) + t|v - x| && [f \text{ is } t\text{-Lipschitz}] \\ &\le f(x) + t\varepsilon d_2(v) && [R_v \le \varepsilon d_2(v)] \\ &\le f(x) + t\varepsilon f(v) && [d_2 \le f] \\ &\le \frac{1}{1 - t\varepsilon} f(x) && [\text{Collect terms}] \\ &= (1 + \varepsilon_0) f(x) && \left[1 + \varepsilon_0 = \frac{1}{1 - t\varepsilon}\right] \end{aligned}$$

□

The preceding lemmas and Theorem 1 imply the following theorem.

Theorem 6. *If M is an ε-refined, and $f \ge d_2$ is t-Lipschitz, then the persistence diagram of the Delaunay (or equivalently, the Voronoi) filtration on f and M is a $\frac{1}{1-t\varepsilon}$-approximation to the persistence diagram of the sublevels filtration of f.*

The result for general functions applies easily to the class of kth nearest neighbor distances to yield the following corollary.

Corollary 7. *If M is ρ-well-spaced and $k \ge 2$, then the persistence diagram of the Delaunay (or equivalently, the Voronoi) filtration on d_k and M is a $\frac{1}{1-\varepsilon}$-approximation to the persistence diagram of the sublevels filtration of f.*

References

[1] G. Carlsson, G. Singh, and A. Zomorodian. Computing multidimensional persistence. In *ISAAC*, 2009.

[2] 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*, 2009.

[3] F. Chazal and S. Y. Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the 24th ACM Symposium on Computational Geometry*, 2008.

[4] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proceedings of the 21st ACM Symposium on Computational Geometry*, 2005.

[5] B. Hudson, G. Miller, and T. Phillips. Sparse Voronoi Refinement. In *Proceedings of the 15th International Meshing Roundtable*, pages 339–356, Birmingham, Alabama, 2006. Long version available as Carnegie Mellon University Technical Report CMU-CS-06-132.

[6] B. Hudson, G. L. Miller, S. Y. Oudot, and D. R. Sheehy. Topological inference via meshing. In *Symposium on Computational Geometry*, 2010.

[7] J. Kloke and G. Carlsson. Topological de-noising: Strengthening the topological signal. *arXiv:0910.5947v2*, 2010.

[8] G. L. Miller, T. Phillips, and D. R. Sheehy. Linear-size meshes. In *CCCG: Canadian Conference in Computational Geometry*, 2008.

[9] A. Zomorodian. *Topology for Computing*. Cambridge Univ. Press, 2005.