The Čech complex in Topological Data Analysis

Brandon Doherty

University of Western Ontario

May 24, 2018

The Vietoris-Rips complex

We've seen how the Vietoris-Rips complex is used to compute the persistent homology of a dataset, viewed as a point cloud in $\mathbb{R}^d$.

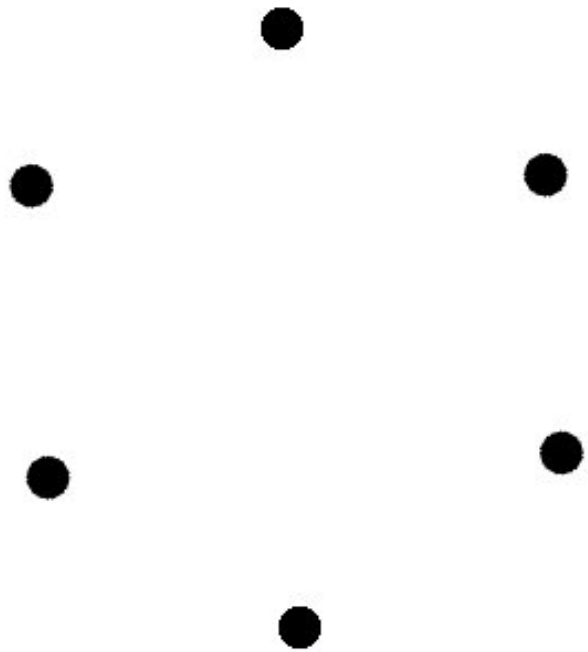
For each ϵ, we create a simplicial complex R_ϵ whose n-simplices correspond to sets of n points whose pairwise distances are all less than or equal to ϵ, and compute its simplicial homology.

The Vietoris-Rips complex

But what *is* this simplicial complex? What topological space are we describing when we compute these homology groups, and how does it relate to our point cloud?

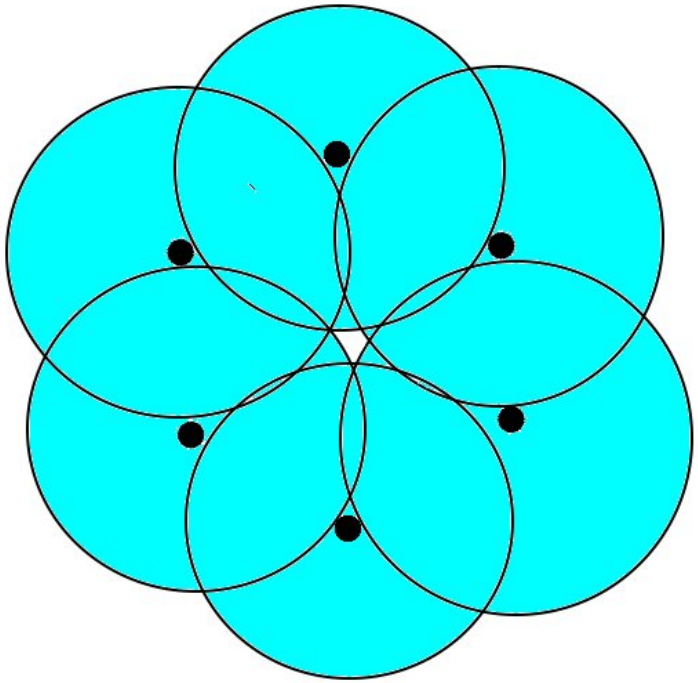
The Vietoris-Rips complex

Consider the following point cloud in $\mathbb{R}^2$, consisting of the vertices of a regular hexagon:



The Vietoris-Rips complex

Let these be the closed discs around the points of radius $\frac{\epsilon}{2}$, so each vertex is less than ϵ away from all others, except the one directly opposite itself:



The Vietoris-Rips complex

So what simplices do we get in the Rips complex R_ϵ for this point cloud?

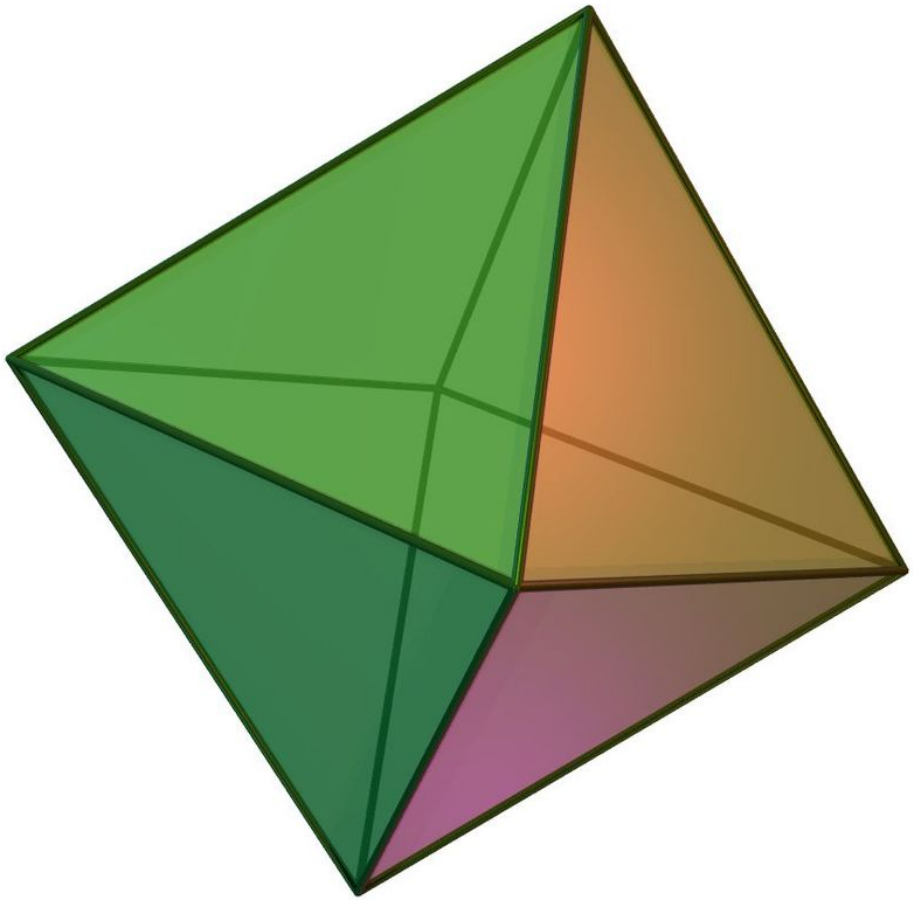
- For $n \geq 3$, any set of $n+1$ vertices must include at least one pair of opposite vertices, hence they will not all have pairwise distance less than or equal to ϵ. So R_ϵ has no n-simplices.
- For $n = 2$, we see that any set of three points, none of them opposite, spans a 2-simplex.

The Vietoris-Rips complex

- For each vertex v, there are four ways to choose a pair of non-opposite points, excluding the vertex opposite v. Thus there are four triangles meeting at each vertex, for a total of eight triangles.
- Similarly, given an edge uv, there are two ways to choose a third vertex, excluding those opposite u and v. Thus two triangles meet at each edge.

The Vietoris-Rips complex

So the realization of this simplicial complex is the boundary of an octahedron.



(Image source: Wikipedia, "Octahedron". Originally uploaded by user "Cyp".)

The Vietoris-Rips complex

So in this example, the complex we got had the homotopy type of S^2. But S^2 can't be embedded in $\mathbb{R}^2$. When we talk about "the shape of the data", we'd like to think of a cloud of points in $\mathbb{R}^d$ as living in some kind of d-dimensional space.

The Čech complex

The Čech complex is a simplicial complex which we can associate to a point cloud $\{p_i\}_{1 \leq i \leq N}$ in $\mathbb{R}^d$, which has some nice topological properties.

The Čech complex

For a fixed ϵ, consider the closed balls of radius $\frac{\epsilon}{2}$ about each point in the dataset, $\bar{B}_\epsilon(p_i)$.

For each n, we define the n-simplices of the simplicial complex C_ϵ to be the sets of $n+1$ points $\{p_{i_k}\}_{0\leq k\leq n}$ such that $\bigcap_{k=0}^{n} \bar{B}_{\frac{\epsilon}{2}}(p_{i_k}) \neq \emptyset$.

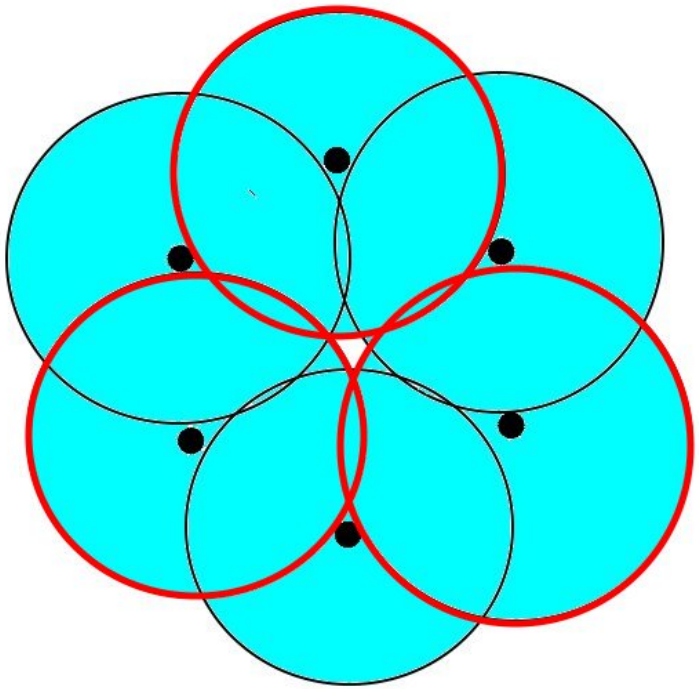
As with the Rips complex, face maps simply omit points.

The Čech complex

To return to our example, what would the Čech complex of those six points in $\mathbb{R}^2$ look like?

The Čech complex

We lose the two opposite faces of the octahedron corresponding to triples of non-adjacent points, but keep all the rest. Thus we get something homtopy-equivalent to S^1.



The Čech complex

So what makes this complex a better topological model for the point cloud?

The nerve theorem provides the answer.

The Nerve Theorem

Given a cover $\mathcal{U}$ (not necessarily open) of a space X, the nerve of $\mathcal{U}$ is a simplicial complex $\mathcal{NU}$ whose n-simplices consist of sets of $n+1$ elements of $\mathcal{U}$ with non-empty intersection.

The Čech complex C_ϵ is thus the nerve of the collection of closed discs of radius $\frac{\epsilon}{2}$ around the points of the cloud, thought of as a cover of its union.

The Nerve Theorem

There are many distinct theorems which go by this name. The most commonly cited one is:

The Nerve Theorem

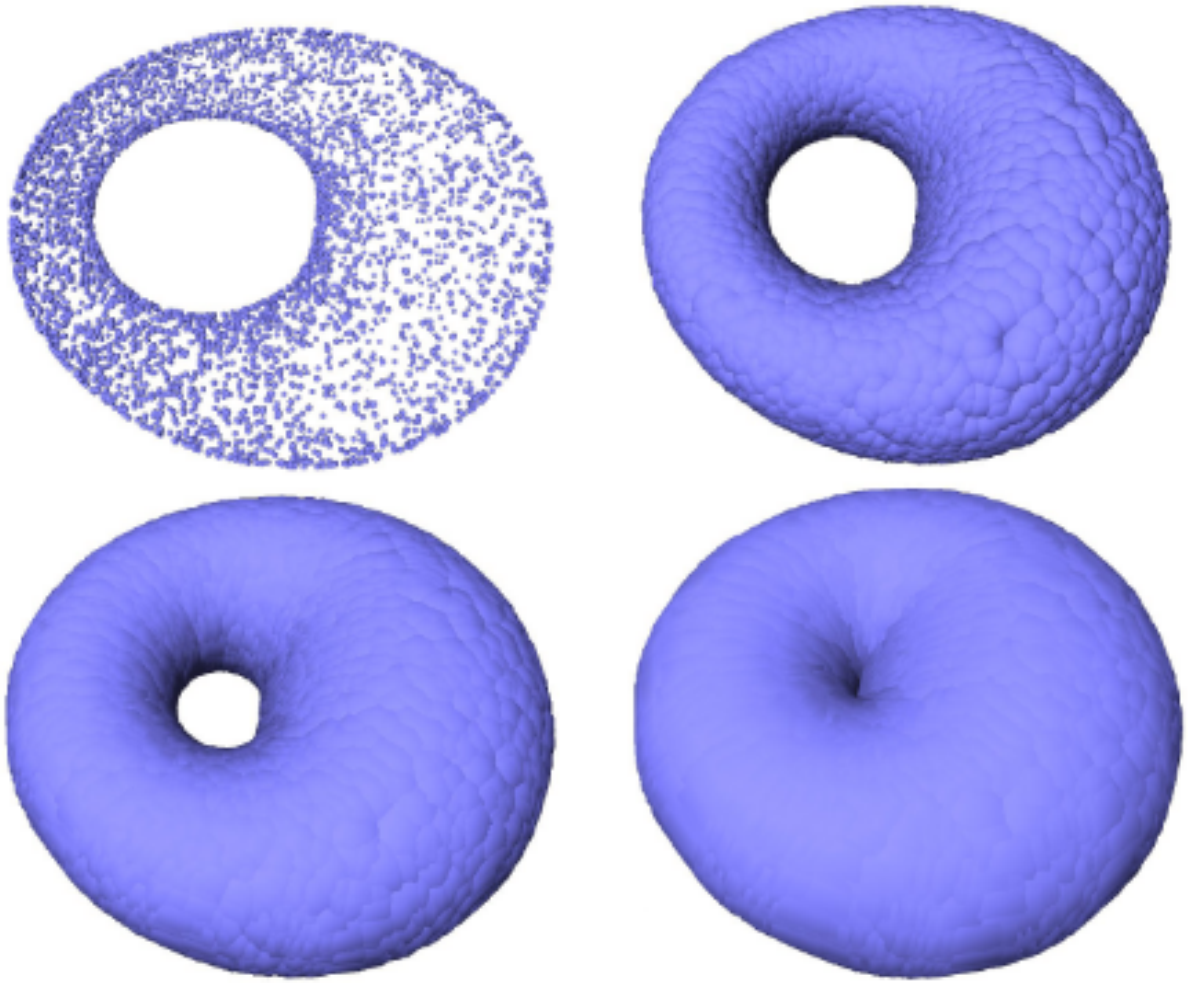
If X is a paracompact space, and $\mathcal{U}$ is an open cover of X such that the intersection of any finite subfamily of $\mathcal{U}$ is either empty or contractible, then the realization of the nerve of $\mathcal{U}$ is homotopy equivalent to X. (Hatcher, section 4G.)

More relevantly for our purposes:

The Nerve Theorem

Let $\mathcal{U}$ be a collection of closed balls in $\mathbb{R}^d$; then the realization of the nerve of $\mathcal{U}$ is homotopy equivalent to $\bigcup_{V \in \mathcal{U}} V$. (Oudot, section 4.3; see also Borsuk).

The Nerve Theorem



(Image source: Chazal and Michal, "An Introduction to Topological Data Analysis".)

The Nerve Theorem

Note that the dimension of the Čech complex can still be larger than d – for sufficiently large ϵ, we get an $(N-1)$-simplex, where N is the number of points in the cloud.

But the Čech complex still "behaves" like a subspace of $\mathbb{R}^d$, in that it is homotopy equivalent to such a subspace, and thus has all the same homology groups.

Computations

So why not just compute persistent homology using the Čech complex? Why bother with the Rips complex at all?

Computing with the Čech complex is difficult – we must remember either the entire complex and its boundary operator, or the precise distances between vertices.

The Rips complex is a flag complex, meaning that a set of $n+1$ vertices spans an n-simplex if and only if any two of them span a 1-simplex. Thus it's completely determined by its 1-skeleton.

Computations

So rather than storing the entire Rips complex in memory, we can just store the 1-skeleton and reconstitute higher simplices as necessary when computing homology.

Example: When computing the n^{th} homology, any unfilled $n+1$ simplex in R_ϵ^n is indeed the boundary of an $(n+1)$-simplex in R_ϵ^{n+1}, while identifying boundaries in C_ϵ requires precise knowledge of distances, which cannot be determined from the 1-skeleton.

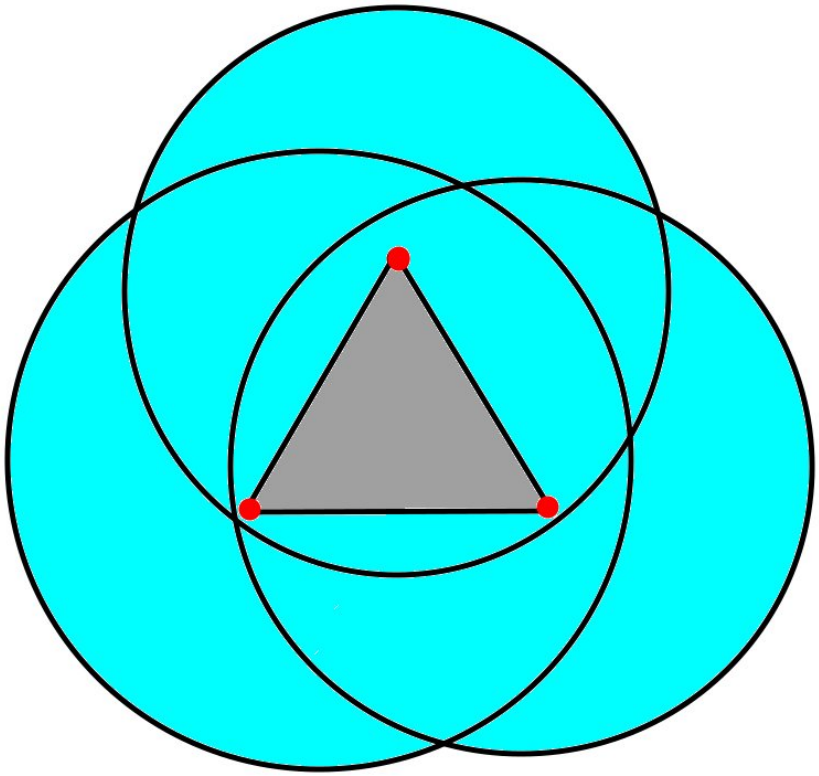
Relating the Čech complex to the Rips complex

Fortunately, these nice properties of the Čech complex can tell us something about the Rips complex as well.

If the closed balls of radius $\frac{\epsilon}{2}$ about a set of n points intersect, then the points must have pairwise distance less than or equal to ϵ by the triangle inequality. So we have an embedding $C_\epsilon \hookrightarrow R_\epsilon$.

Relating the Čech complex to the Rips complex

Furthermore, if $n+1$ points are within distance $\frac{\epsilon}{2}$ of one another, then each of them is contained in the intersection of the closed balls of radius $\frac{\epsilon}{2}$. Thus we also have an embedding $R_{\frac{\epsilon}{2}} \hookrightarrow C_\epsilon$.



Relating the Čech complex to the Rips complex

So for any ϵ, we have a chain of embeddings:

$$C_{\frac{\epsilon}{2}} \hookrightarrow R_{\frac{\epsilon}{2}} \hookrightarrow C_{\epsilon} \hookrightarrow R_{\epsilon}$$

So if a particular topological feature (e.g., a homology generator) persists from ϵ to ϵ' in the Čech complex, then it also appears as a persistent feature of the Rips complex, and vice-versa, provided that $\frac{\epsilon'}{\epsilon} \geq 2$.
To see this, observe:

$$C_{\epsilon} \hookrightarrow R_{\epsilon} \hookrightarrow R_{\frac{\epsilon'}{2}} \hookrightarrow C_{\epsilon'}$$

$$R_{\epsilon} \hookrightarrow C_{2\epsilon} \hookrightarrow C_{\epsilon'} \hookrightarrow R_{\epsilon'}$$

Relating the Čech complex to the Rips complex

In fact, we can acheive an even lower bound on $\frac{\epsilon'}{\epsilon}$, using a result of de Silva and Ghrist. They show that we have inclusions:

$$C_\epsilon \hookrightarrow R_\epsilon \hookrightarrow C_{\gamma_d \epsilon}$$

where $\gamma_d = \sqrt{\frac{2d}{d+1}}$. In particular, we thus have inclusions $C_\epsilon \hookrightarrow R_\epsilon \hookrightarrow C_{\sqrt{2}\epsilon}$ regardless of the dimension d.

Proof of the bound

The statement of the theorem is equivalent to saying: if the pairwise distances of a collection of points $\{x_0, ..., x_{d'}\}$ in $\mathbb{R}^d$ are all less than or equal to ϵ, then there is some point within distance $\frac{\epsilon'}{2}$ of all the x_i, provided that $\frac{\epsilon'}{\epsilon} \geq \sqrt{\frac{2d}{d+1}}$.

First, consider the case $d' \leq d$. Define $f : \mathbb{R}^d \to \mathbb{R}$ by $f(y) = \max_{0 \leq i \leq d'} ||x_i - y||$. This function is continuous, and tends to $+\infty$ as $y \to \infty$. Thus f has some global minimum, say $f(y_0)$.

Define the *critical vertices* to be the x_i such that $||x_i - y_0|| = f(y_0)$.

Proof of the bound

Now we will show, by contradiction, that y_0 lies within the convex hull of the critical vertices.

If it did not, then they would be separated by some hyperplane H, and there would be a vector v, normal to H, such that $v \cdot (x_i - y_0) > 0$ for all critical x_i.

Proof of the bound

Suppose there were such a v; then for any critical x_i and any $\lambda > 0$, we could calculate:

$$||x_i - y_0||^2 = ||x_i - (y_0 + \lambda v)||^2 + 2\lambda v \cdot (x_i - y_0) + ||\lambda v||^2$$
$$> ||x_i - (y_0 + \lambda v)||^2$$

So for a sufficiently small λ we would have $f(y_0 - \lambda v) < f(y_0)$, contradicting the minimality of $f(y_0)$.

Proof of the bound

So y_0 lies within the convex hull of the critical vertices x_i. For each i, let $\hat{x}_i = x_i - y_0$; then 0 is within the convex hull of the critical $\hat{x}_i$.

Relabeling if necessary, we can find a convex combination $\sum_{i=0}^{d''} a_i \hat{x}_i = 0$, where each x_i is critical, all a_i are positive, and $a_0 \geq a_i$ for all i. Rearranging, we get:

$$-\hat{x}_0 = \sum_{i=1}^{d''} (\frac{a_i}{a_0}) \hat{x}_i$$

We can take the dot product with $\hat{x}_0$ on each side to get:

$$-f(y_0)^2 = -||\hat{x}_0||^2 = \sum_{i=1}^{d''} (\frac{a_i}{a_0}) \hat{x}_i \cdot \hat{x}_0$$

Proof of the bound

$$-f(y_0)^2 = -||\hat{x}_0||^2 = \sum_{i=1}^{d''}(\frac{a_i}{a_0})\hat{x}_i \cdot \hat{x}_0$$

For at least one term on the right-hand side, we must have $(\frac{a_i}{a_0})\hat{x}_i \cdot \hat{x}_0 \leq -\frac{f(y_0)^2}{d''}$. Rearranging, we get:

$$\frac{f(y_0)^2}{d''} \leq -(\frac{a_i}{a_0})\hat{x}_i \cdot \hat{x}_0$$

$$\therefore \frac{f(y_0)^2}{d} \leq \frac{f(y_0)^2}{d''} \leq -(\frac{a_i}{a_0})\hat{x}_i \cdot \hat{x}_0 \leq -\hat{x}_i \cdot \hat{x}_0$$

Proof of the bound

Noting that $||\hat{x}_0||^2 = ||\hat{x}_i||^2 = f(y_0)^2$, we can use this inequality to calculate:

$$f(y_0)^2(1 + \frac{2}{d} + 1) \leq ||\hat{x}_0||^2 - 2\hat{x}_0 \cdot \hat{x}_i + ||\hat{x}_i||^2$$
$$= ||\hat{x}_0 - \hat{x}_i||^2 = ||x_0 - x_i||^2 \leq \epsilon^2$$

Proof of the bound

So $2f(y_0)^2(\frac{d+1}{d}) \leq \epsilon^2$. Thus:

$$f(y_0) \leq \frac{\epsilon}{2}\sqrt{\frac{2d}{d+1}} \leq \frac{\epsilon'}{2}$$

So $\max_{0\leq i\leq d'}||x_i - y_0|| \leq \frac{\epsilon'}{2}$; thus y_0 is in the intersection of the closed balls $\bar{B}_{\frac{\epsilon'}{2}}(x_i)$.

Proof of the bound

For $d' > d$, Helly's Theorem says that the intersection of $d' + 1$ convex sets in $\mathbb{R}^d$ is nonempty if and only if the intersection of any subfamily of size $d + 1$ is nonempty. So the same bound applies in this case.

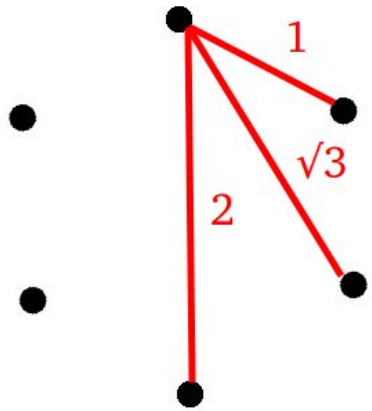
Thus, any simplex of R_ϵ is also a simplex of $C_{\epsilon'}$, provided that $\frac{\epsilon'}{\epsilon} \geq \sqrt{\frac{2d}{d+1}}$.

Proof of the bound

Returning to the example of the six points in $\mathbb{R}^2$, how long does the nontrivial 2-cycle in the Rips complex persist? Could it appear in any Čech complexes?

Proof of the bound

The 2-cycle corresponding to the unfilled boundary of the octahedron appears in the Rips complex when $\sqrt{3} \leq \epsilon < 2$.



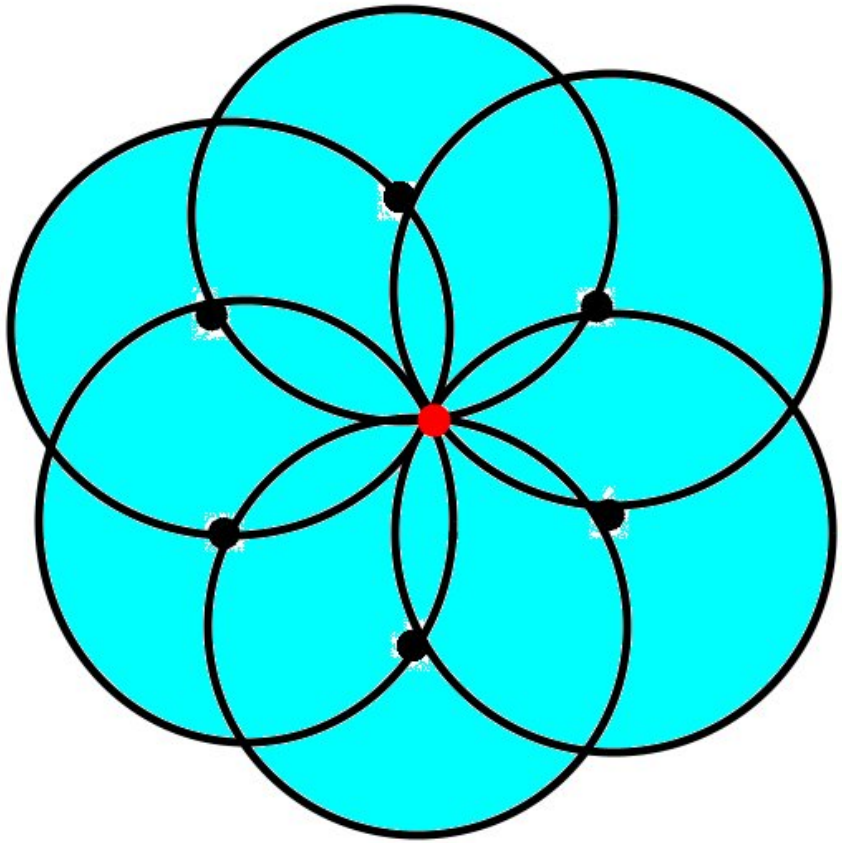
For $\epsilon < \epsilon'$ in this interval, we have:

$$\frac{\epsilon'}{\epsilon} < \frac{2}{\sqrt{3}} = \sqrt{\frac{2(2)}{2+1}}$$

This is *exactly* the bound provided by Ghrist and de Silva for $d = 2$, so their theorem does not imply that the 2-cycle appears in any Čech complex – and, indeed, it does not.

Proof of the bound

The triangles spanned by non-adjacent triples of points don't appear in the Čech complex until ϵ is large enough that their neighbourhoods contain the center of the hexagon – but then the neighbourhoods of all the other points contain it, too, so the Čech complex is just a 5-simplex.



Proof of the bound

This example shows that Ghrist and de Silva's bound is tight in the case $d = 2$.

In general, we can see that the bound is tight by considering the vertices of a regular d-simplex in $\mathbb{R}^d$.

Conclusion

The Čech complex is a theoretically reasonable model for the "shape" of a point cloud in $\mathbb{R}^d$, but performing computations with it can prove challenging.

The Rips complex is much easier to compute with, and its persistent homology accurately models that of the Čech complex.

Čech cohomology

The nerve of a cover has many uses outside of persistent homology.

Given an open cover $\mathcal{U}$ of a space X, the Čech cohomology groups of X with respect to $\mathcal{U}$, with coefficients in an abelian group G, are the simplicial cohomology groups of the nerve of $\mathcal{U}$ with coefficients in G.

Čech cohomology

If an open cover $\mathcal{V}$ of X is a refinement of $\mathcal{U}$, then we can define a simplicial map $g : \mathcal{NV} \to \mathcal{NU}$. Given $B_i \in \mathcal{V}$, choose $g(B_i)$ to be any element of $\mathcal{U}$ which contains B_i.

If $B_0, ..., B_n$ span an n-simplex of $\mathcal{NV}$ – i.e., if they have non-empty intersection – then $\bigcap_{i=0}^{n} g(B_i)$ will be non-empty as well, as it contains the intersection of the B_i. Thus g defines a simplicial map $\mathcal{NV} \to \mathcal{NU}$.

Čech cohomology

The definition of g invovled some arbitrary choices, but any two such maps will be contiguous.

If g, g' are two induced maps $\mathcal{NV} \to \mathcal{NU}$, then given an n-simplex $B = \{B_0, ..., B_n\}$ of $\mathcal{NV}$, the intersection of all sets $g(B_i), g'(B_i)$ is non-empty, as it contains the intersection of all the B_i. Thus $\{g(B_0), ..., g(B_n), g'(B_0), ..., g'(B_n)\}$ is a simplex of $\mathcal{NU}$ having $g(B)$ and $g'(B)$ as faces.

Thus, for each i, there is a unique map of cohomology groups $H^i(\mathcal{NU}; G) \to H^i(\mathcal{NV}; G)$ induced by the refinement of $\mathcal{U}$ to $\mathcal{V}$.

Čech cohomology

For each i, we have a commutative diagram of abelian groups in which the objects are the cohomology groups $H^i(\mathcal{NU}; G)$, where $\mathcal{U}$ ranges over all open covers of X, and the maps are the maps of cohomology groups induced by refinement.

The i^{th} Čech cohomology group of X with coefficients in G, denoted $\check{H}(X; G)$, is defined to be the colimit of this diagram. See Munkres for an explicit description.

Čech cohomology

This agrees with other cohomology theories for nicely behaved spaces; e.g., for simplicial complexes K we have $\check{H}^i(|K|; G) \simeq H^i(K; G)$.

For other spaces, it may not; for instance, if X is the closed topologist's sine curve, then $H^1(X; \mathbb{Z}) \simeq 0$ while $\check{H}^1(X, \mathbb{Z}) \simeq \mathbb{Z}$, where H denotes singular cohomology.

References

- K. Borsuk. On the imbedding of systems of compacta in simplicial complexes. Fund. Math. **35** (1948), 217–234.
- V. de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. Alg. Geom. Top. **7** (2007) 339-358.
- R. Ghrist. Barcodes: The persistent topology of data. Bull. Amer. Math. Soc. **45** (2008), 61-75.
- A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.
- J. Munkres. *Elements of Algebraic Topology*. Addison-Wesley, 1984.
- S. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis*. Vol. 209 of *Mathematical Surveys and Monographs*. AMS, 2015.