# HOMOLOGY INFERENCE

## Contents

Although the distance-based approach introduced in the previous lesson provides a powerful mathematical framework for shape reconstruction, it is not always possible, nor desirable, to fully reconstruct the approximated shapes from data. This chapter focuses on weaker topological invariants, homology, Betti numbers and persistent homology, that turn out to be easier to infer and that are widely used in applied topology and topological data analysis. The introduction of homology is restricted to the minimum that is necessary to understand the basic ideas of homology inference and persistent homology and its usage in topological data analysis.

## 1. Simplicial Complexes

Geometric shapes like curves, surfaces or their generalization in higher dimensions are "continuous" mathematical objects that cannot be directly encoded as a finite discrete structure usable by computers or computing devices. It is thus necessary to find representations of these shapes that are rich enough to capture their geometric structure and to comply with the constraints inherent to the discrete and finite nature of implementable data structures. On another side, when the only available data are point clouds sampled around unknown shapes, it is necessary to be able to build some continuous space on top of the data that faithfully encode the topology and the geometry of the underlying shape. Simplicial complexes offer a classical and flexible solution to overcome these difficulties.

### 1.1. Geometric Simplicial Complexes.

The points of a finite set $\mathcal{P} = \{p_0, p_1, \ldots, p_k\}$ in $\mathbb{R}^d$ are said to be *affinely independent* if they are not contained in any affine subspace of dimension less than k.

**Definition 1.1** (Simplex)**.** Given a set $\mathcal{P} = \{p_0, p_1, \ldots, p_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the *k-dimensional simplex* $\sigma$, or *k-simplex* for short, spanned by $\mathcal{P}$ is the set of convex combinations

$$\sum_{i=0}^k \lambda_i p_i, \text{ with } \sum_{i=0}^k \lambda_i = 1 \text{ and } \lambda_i \geqslant 0.$$

The points $p_0, p_1, \ldots, p_k$ are called the *vertices* of $\sigma$.

**Remark 1.2.** (i) Notice that $\sigma$ is the convex hull of the points $\mathcal{P}$ , i.e. the smallest convex subset of $\mathbb{R}^d$ containing $p_0, p_1, \ldots, p_k$.

(ii) A 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron.

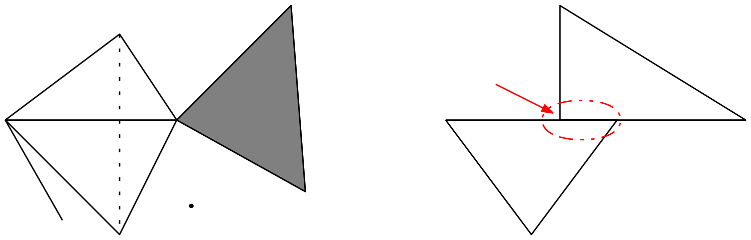
(iii) The *faces* of the simplex $\sigma$ whose vertex set is $\mathcal{P}$ are the simplices spanned by the subsets of $\mathcal{P}$.

**Definition 1.3** (Simplicial complex)**.** A (finite) simplicial complex $\mathcal{K}$ in $\mathbb{R}^d$ is a (finite) collection of simplices such that:

(i) any face of a simplex of $\mathcal{K}$ is a simplex of $\mathcal{K}$,

(ii) the intersection of any two simplices of $\mathcal{K}$ is either empty or a common face of both.

All the simplicial complexes considered here are finite. The simplices of $\mathcal{K}$ are called the *faces* of $\mathcal{K}$. The *dimension of* $\mathcal{K}$ is the highest dimension of its simplices. A complex of dimension $k$ is also called a *k-complex*. A subset of the simplices of $\mathcal{K}$ which is itself a simplicial complex is called a *subcomplex* of $\mathcal{K}$.



(A) An example of a simplicial complex.

(B) A union of simplices which is not a simplicial complex.

FIGURE 1. To be, or not to be a simplicial complex.

**Remark 1.4.** For a simplicial complex $\mathcal{K}$ in $\mathbb{R}^d$ , its *geometric realization* $|\mathcal{K}| \subset \mathbb{R}^d$ is the union of the simplices of $\mathcal{K}$. The topology of $\mathcal{K}$ is the topology induced on $|\mathcal{K}|$ by the standard topology in $\mathbb{R}^d$. When there is no risk of confusion, we do not clearly make the distinction between a complex in $\mathbb{R}^d$ and its geometric realization.

1.2. **Abstract Simplicial Complexes.** Notice that when its vertex set is known, a simplicial complex in $\mathbb{R}^d$ is fully and combinatorialy characterized by the list of its simplices. This leads to the following notion of abstract simplicial complex.

**Definition 1.5.** Let $V = \{v_1, \ldots, v_n\}$ be a finite set. An abstract simplicial complex $\tilde{\mathcal{K}}$ with vertex set $V$ is a set of finite subsets of $V$ satisfying the two conditions :

(i) The elements of $V$ belong to $\tilde{\mathcal{K}}$.
(ii) If $\tau \in \tilde{\mathcal{K}}$ and $\sigma \subset \tau$, then $\sigma \in \tilde{\mathcal{K}}$.

The elements of $\tilde{\mathcal{K}}$ are called the *simplices* or the *faces* of $\tilde{\mathcal{K}}$. If $\sigma \in \tilde{\mathcal{K}}$ has precisely $k + 1$ elements, the *dimension* of $\sigma$ is $k$ and we say that $\sigma$ is a *$k$-simplex.* The *dimension* of $\tilde{\mathcal{K}}$ is the maximal dimension of its simplices.

Any simplicial complex $\mathcal{K}$ in $\mathbb{R}^d$ naturally determines an abstract simplicial complex $\tilde{\mathcal{K}}$, called the *vertex scheme* of $\mathcal{K}$: $\mathcal{K}$ and $\tilde{\mathcal{K}}$ have the same set of vertices and the simplices of $\tilde{\mathcal{K}}$ are the sets of vertices of the simplices of $\mathcal{K}$. Conversely, if an abstract complex $\tilde{\mathcal{K}}$ is the vertex scheme of a complex $\mathcal{K}$ in $\mathbb{R}^d$ , then $\mathcal{K}$ is called a *geometric realization* of $\tilde{\mathcal{K}}$. Notice that any finite abstract simplicial complex $\tilde{\mathcal{K}}$ has a geometric realization in an Euclidean space in the following way. Let $\{v_1, v_2, \ldots, v_n\}$ be the vertex set of $\tilde{\mathcal{K}}$ where $n$ is the number of vertices of $\tilde{\mathcal{K}}$, and let $\sigma \subset \mathbb{R}^n$ be the simplex spanned by $\{e_1, e_2, \ldots, e_n\}$, where for any $i \in \{1, \ldots, n\}$, $e_i$ is the vector whose coordinates are all 0 except the $i$th one which is equal to 1. Then $\mathcal{K}$ is the subcomplex of $\sigma$ defined by $[e_{i_0}, \ldots, e_{i_k}]$ is a $k$-simplex of $\mathcal{K}$ if and only if $[v_1, v_2, \ldots, v_n]$ is a simplex of $\tilde{\mathcal{K}}$. It can also be proven that any finite abstract simplicial complex of dimension $d$ can be realized as a simplicial complex in $\mathbb{R}^{2d+1}$

**Definition 1.6** (Isomorphism of abstract simplicial complexes)**.** Two abstract simplicial complexes $\tilde{\mathcal{K}}, \tilde{\mathcal{K}}'$ with vertex sets $V$ and $V'$ are *isomorphic* if there exists a bijection $\varphi : V \to V'$ such that $\{v_0, \ldots, v_k\} \in \tilde{\mathcal{K}}$ if and only if $\{\varphi(v_0), \ldots, \varphi(v_k)\} \in \tilde{\mathcal{K}}'$

The relation of isomorphism between two abstract simplicial complexes induces homeomorphism between their geometric realizations.

Proposition 1.7. *If two simplicial complexes $\mathcal{K}, \mathcal{K}'$ are the geometric realizations of two isomorphic abstract simplicial complexes $\tilde{\mathcal{K}}, \tilde{\mathcal{K}}'$ , then $|\mathcal{K}|$ and $|\mathcal{K}'|$ are homeomorphic topological spaces. In particular, the underlying spaces of any two geometric realizations of an abstract simplicial complex are homeomorphic.*

**Remark 1.8** (About terminology)**.** As the underlying spaces of all geometric realizations of an abstract simplicial complex are homeomorphic to each other, it is usual to relate the topological properties of these underlying spaces to the complex itself. For example, when one claims that an abstract simplicial complex $\mathcal{K}$ is homeomorphic or homotopy equivalent to a topological space $X$, it is meant that the underlying space of any geometric realization of $\mathcal{K}$ is homeomorphic or homotopy equivalent to $X$.

**Exercise 1.9.** Give examples of simplicial complexes in $\mathbb{R}^d$ that are homeomorphic to a ball, a sphere, and a torus.

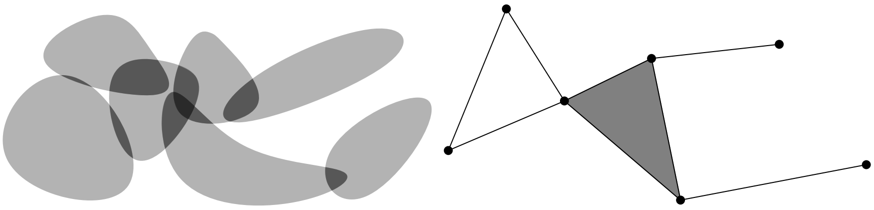
1.3. **Nerve.** As noticed in the previous section, simplicial complexes can be seen at the same time as topological spaces and as purely combinatorial objects.

**Definition 1.10** (Cover)**.** An *open cover* of a topological space $X$ is a collection $\mathcal{U} = (U_i)_{i\in I}$ of open subsets $U_i \subset X$, $i \in I$ where $I$ is a set, such that $X = \cup_{i\in I} U_i$. Similarly, a *closed cover* of $X$ is a collection of closed sets whose union is $X$.

**Definition 1.11** (Nerve of a cover)**.** Given a cover $\mathcal{U} = (U_i)_{i\in I}$ of a topological space $X$, we associate an abstract simplicial complex $\mathrm{Nerve}(\mathcal{U})$ whose vertex set is $\mathcal{U}$ and such that

$$\sigma = [U_{i_0}, \dots, U_{i_k}] \in \mathrm{Nerve}(\mathcal{U}) \text{ if and only if } \cap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

Such a simplicial complex is called the *nerve* of the cover $\mathcal{U}$.



(A) An open cover in $\mathbb{R}^2$. (B) A nerve is a simplicial complex.

Figure 2. A cover and its associated nerve.

When all the sets $U_i$ are open and all their finite intersections are *contractible*, i.e. are homotopy equivalent to a point, the Nerve Theorem relates the topology of $X$ and $\mathrm{Nerve}(\mathcal{U})$.

Theorem 1.12 (Nerve Theorem). *Let $\mathcal{U} = (U_i)_{i\in I}$ be a finite open cover of a subset $X$ of $\mathbb{R}^d$ such that any intersection of the $U_i$'s is either empty or contractible (in particular, each $U_i$ must be contractible). Then $X$ and* $\mathrm{Nerve}(\mathcal{U})$ *are homotopy equivalent.*

*Proof.* See [Hat02, Section 4.G]. □

**Remark 1.13** (About the assumptions of Theorem 1.12)**.** – Contractibility of the intersections of the $U_i$'s is necessary. For instance, take $\mathcal{U} = \{U_1\}$, where $U_1$ is a circle. Then $\mathrm{Nerve}(\mathcal{U})$ is a point, while $X = U_1$ is not contractible.

- Openness (or similar condition) is also necessary, as shown for the example of $\mathcal{U} = \{(-1, 0], (0, 1)\}$, which yields a nerve $\mathrm{Nerve}(\mathcal{U})$ that consists in two disconnected points, while $X = (-1, 1)$ is connected.
- The nerve theorem also holds for closed covers under a slightly more restrictive assumption on $X$. The following version is general enough for our purpose.

Theorem 1.14 (Nerve Theorem for Convex Covers). *Let $X \subset \mathbb{R}^d$ be a finite union of closed convex sets $\mathcal{F} = (F_i)_{i\in I}$ in $\mathbb{R}^d$. Then $X$ and* $\mathrm{Nerve}(\mathcal{F})$ *are homotopy equivalent.*

A cover satisfying the assumptions of the Nerve Theorem is sometimes called a *good cover*. The Nerve Theorem is of fundamental importance in computational topology and geometric inference: it provides a way to encode the homotopy type of continuous topological space $X$ by a simplicial complex describing the intersection pattern of a good cover. In particular, when $X$ is a (finite) union of (closed or open) balls in $\mathbb{R}^d$, it is homotopy equivalent to the nerve of this union of balls.

1.4. **Filtrations of Simplicial Complexes.** Simplicial complexes often come with a specific ordering of their simplices that plays a fundamental role in geometry inference.

**Definition 1.15** (Filtration)**.** A *filtration* of a finite simplicial complex $\mathcal{K}$ is a nested sequence of sub-complexes $\emptyset = \mathcal{K}^0 \subset \mathcal{K}^1 \subset \ldots \subset \mathcal{K}^m = \mathcal{K}$ such that $\mathcal{K}^{i+1} = \mathcal{K}^i \cup \sigma^{i+1}$ where $\sigma^{i+1}$ is a simplex of $\mathcal{K}$.

Equivalently, a filtration of $\mathcal{K}$ can be seen as an ordering of the simplices such that for any $i \geqslant 0$, the collection of the first $i$ simplices is a simplicial complex. To ensure this later condition, it is sufficient to know that every simplex $\sigma^i$ appears in the filtration after all its faces.

As a filtration of $\mathcal{K}$ is just an ordering of the simplices, in some cases, it might be more natural to index the simplices by an increasing sequence $(\alpha_i)_{1\leqslant i\leqslant m}$ of real numbers: $\emptyset = \mathcal{K}^{\alpha_0} \subset \mathcal{K}^{\alpha_1} \subset \ldots \subset \mathcal{K}^{\alpha_m} = \mathcal{K}$ In this case, it is often convenient to extend the filtration to the whole set of real numbers by defining $\mathcal{K}^\alpha = \mathcal{K}^{\alpha_i}$ for $\alpha \in [\alpha_i, \alpha_{i+1})$, $\mathcal{K}^\alpha = \emptyset$ for $\alpha < \alpha_0$ and $\mathcal{K}^\alpha = \mathcal{K}$ for $\alpha \geqslant \alpha_m$.

For example, when a function is defined on the vertices of $\mathcal{K}$, on can define a sublevel set filtration in the following way.

**Definition 1.16** (Filtration induced by a function)**.** Let $\mathcal{K}$ be a simplicial complex and let $f$ be a real valued function defined on the vertices of $\mathcal{K}$. For any simplex $\sigma = \{v_0, \ldots, v_k\}$ one defines $f(\sigma)$ by

$$f(\sigma) = \max_{0\leqslant i\leqslant k} f(v_i).$$

Ordering the simplices of $\mathcal{K}$ according to the values of each simplex defines a filtration of $\mathcal{K}$. Note that different simplices can have the same value. In this case, they are ordered according to increasing dimension and simplices of the same dimension with same value can be ordered arbitrarily.

The *filtration induced by* $f$ is the filtration by the sublevel sets $f^{-1}((-\infty, t])$ of $f$.

1.5. **Vietoris-Rips and Čech Filtrations.** As we will see later in the course, filtrations are often built on top of finite sets of points to reveal the underlying topological structure of data. We let $\mathcal{P} \subset \mathbb{R}^d$ be a (finite) set of points.

**Definition 1.17** (Čech)**.** Given $\alpha > 0$, the *Čech complex* with vertex set $\mathcal{P}$ and parameter $\alpha$ is the nerve $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha)$ of the unions of balls centered on $\mathcal{P}$ with radius $\alpha$. The simplices of $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha)$ are characterized by

$$[x_0, \ldots, x_k] \in \check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha) \Leftrightarrow \cap_{i=0}^k \mathrm{B}(x_i, \alpha) \neq \emptyset.$$

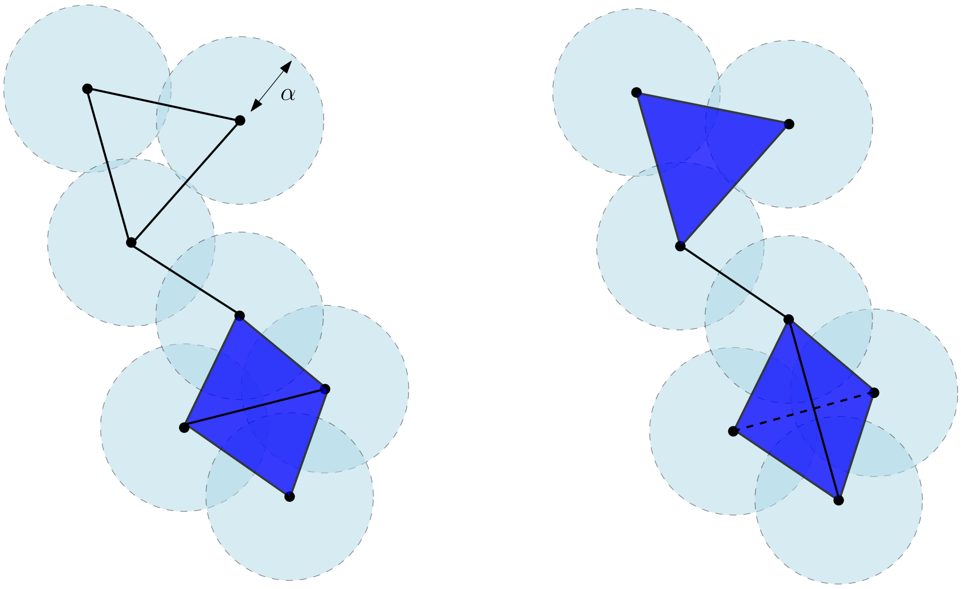


Figure 3. The Čech (left) and Vietoris-Rips (right) complexes built on top of a finite set of points in $\mathbb{R}^2$. Note that they both contains a 3-simplex and are thus not embedded in $\mathbb{R}^2$.

As $\alpha$ goes from 0 to $\infty$, the nested sequence of complexes $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha)$ defines the *Čech complex filtration.*

Given a $k$-dimensional face $\sigma$ of the simplex of dimension $|\mathcal{P}|-1$, the smallest $\alpha$ such that $\sigma \in \check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha)$ is the radius of the smallest ball enclosing the vertices of $\alpha$.

As a consequence, the $k$-dimensional skeleton of the Čech filtration can be computed by computing the $\mathcal{O}(|\mathcal{P}|^k)$ minimum enclosing balls of all the subsets of at most $k$ points of $\mathcal{P}$. Although the computation of the minimum ball enclosing a set of $k$ points can be done in time $\mathcal{O}(k)$, the computation of the whole Čech filtration quickly becomes intractable in practice. Given $\alpha > 0$, the computation of the $k$-skeleton of $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha)$ can be done by first computing all the cliques of at most $(k+1)$ vertices of the 1-skeleton of $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha)$ which is a graph, and second by selecting the cliques whose minimum enclosing ball has its radius upper bounded by $\alpha$.

A simplicial complex which is closely related to the Čech filtration is the Vietoris-Rips filtration, $\mathrm{Rips}(\mathcal{P})$.

**Definition 1.18** (Vietoris-Rips)**.** Given $\alpha > 0$, the *Vietoris-Rips complex* with vertex set $\mathcal{P}$ and parameter $\alpha$ is characterized by

$$[x_0,\ldots,x_k] \in \mathrm{Rips}(\mathcal{P},\alpha) \Leftrightarrow \|x_i - x_j\| \leqslant \alpha \text{ for all } i,j \in \{0,\ldots,k\}.$$

As $\alpha$ goes from 0 to $\infty$, the nested sequence of complexes $\mathrm{Rips}(\mathcal{P},\alpha)$ defines the *Vietoris-Rips complex filtration.*

The Vietoris-Rips complex is much simpler to compute than the Čech filtration as it just involves distance comparisons. The Vietoris-Rips complex is the largest simplicial complex that has the same 1-skeleton as the Čech complex. It is thus completely characterized by its 1-skeleton. The whole $k$-dimensional skeleton of the Vietoris-Rips filtration can be computed by computing the diameter of all the subsets of at most $k$ points of $\mathcal{P}$.

The Čech and the Vietoris-Rips filtrations are related by the following interleaving property that plays a fundamental role in homology inference.

LEMMA 1.19. *Let $\mathcal{P}$ be a finite set of points in $\mathbb{R}^d$. for any $\alpha \geqslant 0$*

$$\mathrm{Rips}(\mathcal{P},\alpha) \subset \check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha) \subset \mathrm{Rips}(\mathcal{P},2\alpha).$$

*Proof.* If $\sigma = [x_0,\dots,x_k] \in \mathrm{Rips}(\mathcal{P},\alpha)$, then $x_0 \in \cap_{i=0}^{k} \mathrm{B}(x_i,\alpha)$. So, $\sigma \in \check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha)$, which proves the first inclusion.

Now, if $\sigma = [x_0,\dots,x_k] \in \check{\mathrm{C}}\mathrm{ech}(\mathcal{P},\alpha)$, there exists $y \in \mathbb{R}^d$ such that $y \in \cap_{i=0}^{k} \mathrm{B}(x_i,\alpha)$. As a consequence, for all $i,j \in \{0,\dots,k\}$, $\|x_i - x_j\| \leqslant \|x_i - y\| + \|y - x_j\| \leqslant 2\alpha$ and $\sigma \in \mathrm{Rips}(\mathcal{P},2\alpha)$, which proves the second inclusion. □

## 2. Simplicial Homology

In this section we introduce the basic notions of simplicial homology. To avoid minor technical discussions about the orientation of simplices, we restrict to the homology with coefficients in the finite field $\mathbb{Z}/2\mathbb{Z} = \{0,1\}$. As above, $\mathcal{K}$ denotes a finite $d$-dimensional simplicial complex.

### 2.1. **The Space of $k$-chains.**

For any non negative integer k, the space of *$k$-chains* is the vector space of all the formal sums (with coefficient in $\mathbb{Z}/2\mathbb{Z}$) of $k$-dimensional simplices of $\mathcal{K}$. More precisely, if $\{\sigma_1,\dots,\sigma_p\}$ is the set of $k$-simplices of $\mathcal{K}$ any $k$-chain $c$ can be uniquely written

$$c = \textstyle\sum_{i=1}^{p} \varepsilon_i \sigma_i, \text{ with } \varepsilon_i \in \mathbb{Z}/2\mathbb{Z}.$$

If $c' = \sum_{i=1}^{p} \varepsilon_i' \sigma_i$, is another $k$-chain, the sum of two $k$-chains and the product of a chain by a scalar are defined by

$$c + c' = \textstyle\sum_{i=1}^{p} (\varepsilon_i + \varepsilon_i')\sigma_i, \text{ and } \lambda.c = \sum_{i=1}^{p} (\lambda\varepsilon_i)\sigma_i,$$

where the sums $\varepsilon_i + \varepsilon_i'$ and $\lambda\varepsilon_i$ are modulo 2.

**Definition 2.1** (Space of $k$-chains)**.** The *space of $k$-chains* of $\mathcal{K}$ is the set $C_k(\mathcal{K})$ of the simplicial $k$-chains of $\mathcal{K}$ with the two operations defined above. This is a $\mathbb{Z}/2\mathbb{Z}$-vector space whose zero element is the empty chain $0 = \sum_{i=1}^{p} 0.\sigma_i$.

Notice that the set of $k$-simplices of $\mathcal{K}$ is a basis of $C_k(\mathcal{K})$.

**Example 2.2.** or the simplicial complex $\mathcal{K}$ of Figure 4, $C_1(\mathcal{K})$ is the $\mathbb{Z}/2\mathbb{Z}$-vector space generated by the edges $e_1 = [a,b]$, $e_2 = [b,c]$, $e_3 = [c,a]$ and $e_4 = [c,d]$; i.e.

$$\begin{aligned} C_1(\mathcal{K}) &= \mathrm{span}_{\mathbb{Z}/2\mathbb{Z}}(e_1,e_2,e_3,e_4) \\ &= \{0, e_1, e_2, e_3, e_4, e_1+e_2, e_1+e_3, e_1+e_4, e_2+e_3, e_2+e_4, e_3+e_4, \\ &\qquad e_1+e_2+e_3, e_1+e_2+e_4, e_1+e_3+e_4, e_2+e_3+e_4\}. \end{aligned}$$

Summing $e_1+e_2$ with $e_2+e_3+e_4$ gives $e_1+e_3+e_4$. One may describe $C_0(\mathcal{K})$ and $C_2(\mathcal{K})$ in a similar way.

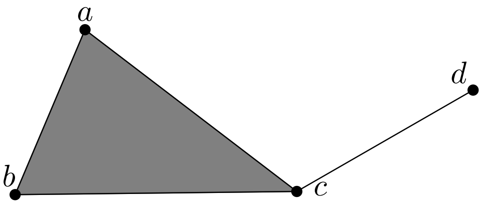


Figure 4. The simplicial complex of Example 2.2, made of four vertices, four edges and one triangle.

Chains with coefficient in $\mathbb{Z}/2\mathbb{Z}$ have an obvious geometric interpretation: since any $k$-chain can be uniquely written as $c = \sigma_{i_1} + \ldots + \sigma_{i_m}$ where the $\sigma_{i_j}$'s are $k$-simplices, $c$ can be considered as the union of the simplices $\sigma_{i_j}$. The sum of two $k$-chains is equal to their symmetric difference.

### 2.2. **The Boundary Operator and Homology Groups.**

**Definition 2.3** (Boundary of a Simplex)**.** The *boundary* $\partial(\sigma)$ of a $k$-simplex $\sigma$ is the sum of its $(k-1)$-faces. This is a $(k-1)$-chain.

If $\sigma = [v_0, \ldots, v_k]$ is a $k$-simplex, then

$$\partial(\sigma) = \sum_{i=0}^{k} [v_0, \ldots, \hat{v}_i, \ldots, v_k],$$

where $[v_0, \ldots, \hat{v}_i, \ldots, v_k]$ is the $(k-1)$-simplex spanned by the set of all the vertices of $\sigma$ except $v_i$.

**Remark 2.4.** In the general case where the coefficient of the chains are taken in another field than $\mathbb{Z}/2\mathbb{Z}$ it is important to take into account the ordering of the vertices in $\sigma$ and the boundary of $\sigma$ has to be defined as $\partial(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$.

The boundary operator, defined on the simplices of $\mathcal{K}$, extends linearly to $C_k(\mathcal{K})$.

**Definition 2.5** (Boundary Operator)**.** The *boundary operator* is the linear map defined by

$$\begin{aligned} \partial \colon C_k(\mathcal{K}) &\to C_{k-1}(\mathcal{K}) \\ c &\mapsto \partial(c) = \sum_{\sigma \in c} \partial(\sigma). \end{aligned}$$

Notice that one should denote $\partial_k$ the above defined operator but to avoid heavy notations one usually omits the index in the notations.

Proposition 2.6. *The boundary of the boundary of a chain is always zero:*

$$\partial\partial = \partial \circ \partial = 0.$$

*Proof.* Since the boundary operator is linear, it is sufficient to check the property for a simplex. Let $\sigma = [v_0, \ldots, v_k]$ be a $k$-simplex.

$$\begin{aligned}\partial\partial(\sigma) &= \partial\left(\sum_{i=0}^{k}[v_0,\ldots,\hat{v}_i,\ldots,v_k]\right)\\ &= \sum_{i=0}^{k}\partial([v_0,\ldots,\hat{v}_i,\ldots,v_k])\\ &= \sum_{j<i}[v_0,\ldots,\hat{v}_j,\ldots,\hat{v}_i,\ldots,v_k] + \sum_{j>i}[v_0,\ldots,\hat{v}_i,\ldots,\hat{v}_j,\ldots,v_k]\\ &= 0. \qquad \square\end{aligned}$$

The boundary operator defines a sequence of linear maps between the spaces of chains.

**Definition 2.7** (Chain Complex)**.** The chain complex associated to a complex $\mathcal{K}$ of dimension $d$ is the following sequence of linear operators

$$\{0\} \xrightarrow{\partial} C_d(\mathcal{K}) \xrightarrow{\partial} \ldots \xrightarrow{\partial} C_{k+1}(\mathcal{K}) \xrightarrow{\partial} C_k(\mathcal{K}) \xrightarrow{\partial} C_{k-1}(\mathcal{K}) \xrightarrow{\partial} \ldots \xrightarrow{\partial} C_0(\mathcal{K}) \xrightarrow{\partial} \{0\}.$$

For $k \in \{0,\ldots,d\}$, the set $Z_k(\mathcal{K})$ of $k$-*cycles* of $\mathcal{K}$ is the kernel of $\partial : C_k \to C_{k-1}$:

$$Z_k(\mathcal{K}) := \ker(\partial : C_k \to C_{k-1}) = \{c \in C_k(\mathcal{K}) | \partial c = 0\}.$$

The image $B_k(\mathcal{K})$ of $\partial : C_{k+1} \to C_k$ is the set $k$-*boundaries*, i.e. the $k$-chains bounding a $(k+1) - chain$:

$$B_k(\mathcal{K}) := \mathrm{Im}(\partial : C_{k+1} \to C_k) = \{c \in C_k(\mathcal{K}) | \exists c' \in C_{k+1}, c = \partial c'\}.$$

Examples of chains, cycles and boundaries are given in Figure 5.

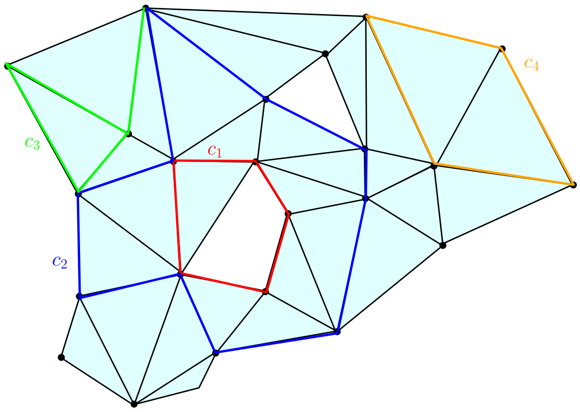


FIGURE 5. Some examples of chains, cycles and boundaries on a 2-dimensional complex $\mathcal{K}$: $c_1, c_2$ and $c_4$ are 1-cycles; $c_3$ is a 1-chain but not a 1-cycle; $c_4$ is a 1-boundary, namely the boundary of the 2-chain obtained as the sum of the two triangles surrounded by $c_4$; The cycles $c_1$ and $c_2$ span the same element in $H_1(\mathcal{K})$ as their difference is the 2-chain represented by the union of the triangles surrounded by the union of $c_1$ and $c_2$.

The linear spaces $B_k$ and $Z_k$ are subspaces of $C_k$, and according to Proposition 2.6, one has

$$B_k(\mathcal{K}) \subset Z_k(\mathcal{K}) \subset C_k(\mathcal{K}).$$

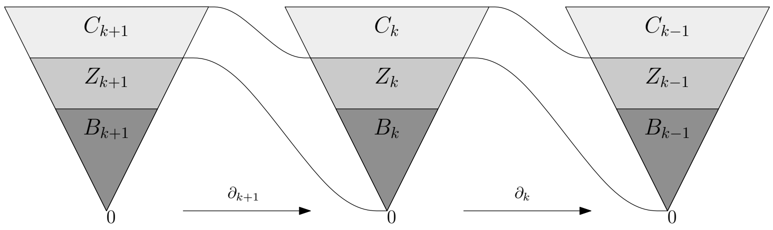


FIGURE 6. A chain complex, with the property $B_k(\mathcal{K}) \subset Z_k(\mathcal{K}) \subset C_k(\mathcal{K})$ dislayed. The composition of any two consecutive boundary maps is the zero map.

That is, said with words, $k$-boundaries are $k$-cycles. However, not all cycles are boundaries, with motivates the following definition.

**Definition 2.8** (Homology Groups, Betti Numbers)**.** The *kth homology group* of $\mathcal{K}$ is the quotient linear space

$$H_k(\mathcal{K}) = Z_k(\mathcal{K})/B_k(\mathcal{K}).$$

$H_k(\mathcal{K})$ is a vector space and its elements are the *homology classes* of $\mathcal{K}$. The dimension $\beta_k(\mathcal{K}) = \dim H_k(\mathcal{K})$ is called the *kth Betti number* of $\mathcal{K}$.

The homology class of a cycle $c \in Z_k(\mathcal{K})$ is the set $c + B_k(\mathcal{K}) = \{c + b | b \in B_k(\mathcal{K})\}$. Two cycles $c, c'$ that are in the same homology class are said to be *homologous.*

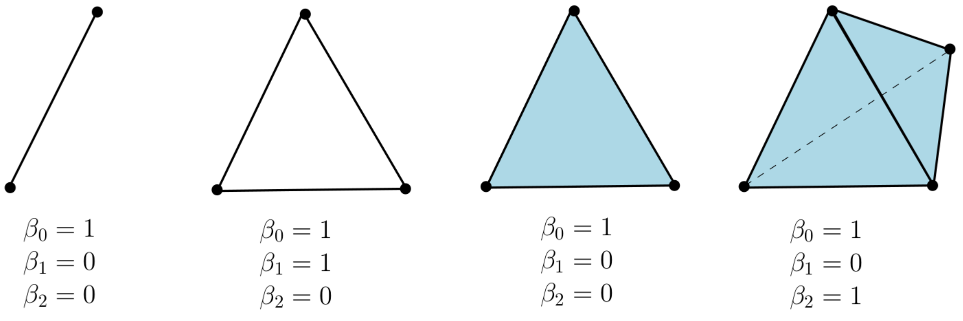


FIGURE 7. Examples of Betti numbers for simple simplicial complexes: from left to right, an edge, the boundary of a triangle, a triangle and the boundary of a tetrahedron.

**Exercise 2.9.** What are the Betti numbers of a $n$-simplex? Of a torus? Of the skeleton of a tetrahedron? Of the skeleton of a cube?

2.3. **Singular Homology and Topological Invariance.** The homology groups and Betti numbers are topological invariants.

THEOREM 2.10. *If $\mathcal{K}$ and $\mathcal{K}'$ are two simplicial complexes with homotopy equivalent geometric realizations then their homology groups are isomorphic and their Betti numbers are equal.*

*Proof.* Beyond the scope of this course. See [Hat02, Section 2.1] for a complete proof. □

Singular homology is another notion of homology that allows to consider general spaces that are not necessarily homeomorphic to simplicial complexes. The definition of singular homology is similar to the one of simplicial homology except that it relies on the notion of singular simplex.

Let $\Delta_k$ be the standard $k$-dimensional simplex in $\mathbb{R}^k$ , i.e.. the geometric simplex spanned by the origin and the vertices $x_i$'s, $i \in \{1, \dots, k\}$, whose coordinates are all 0 except the $i$th one which is equal to 1. Given a topological space $X$, a singular $k$-simplex is a continuous map $\sigma : \Delta_k \to X$. As in the case of simplicial homology, the space of singular $k$-chains is the vector space of formal linear combinations of singular $k$-simplices. The boundary $\partial\sigma$ of a singular $k$-simplex is the sum of the restriction of $\sigma$ to each of the $(k-1)$-faces of $\Delta_k$. Proposition 2.6 still holds for the (singular) boundary operator and the $k$th singular homology group of $X$ is defined similarly as the quotient of the space of cycles by the space of boundaries.

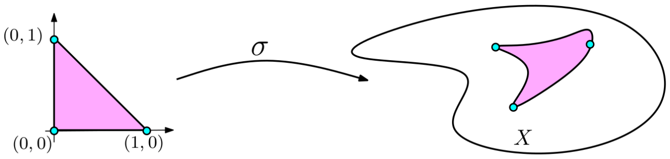


FIGURE 8. $\Delta_k$ is the standard simplex in $\mathbb{R}^k$. A singular $k$-simplex in a topological space $X$ is a continuous map $\sigma : \Delta_k \to X$.

Some of its important properties are:

- Singular homology is defined for any topological space $X$.
- If $X$ is homotopy equivalent to the geometric realization of a simplicial complex, then the singular and simplicial homology coincide.

Another important property of singular (and thus simplicial) homology is that continuous maps between topological spaces canonically induce homomorphisms between their homology groups.

PROPOSITION 2.11 (Homology and Continuous Maps). *if $f : X \to Y$ is a continuous map and $\sigma : \Delta_k \to X$ a simplex in $X$, then $f \circ \sigma : \Delta_k \to Y$ is a simplex in $Y$.*

*As a consequence, $f$ canonically induces linear maps between homology groups:*

$$f_* : H_k(X) \to H_k(Y).$$

*Furthermore, if $f : X \to Y$ is an homeomorphism or an homotopy equivalence then $f_*$ is an isomorphism.*

*Proof.* See [Hat02, Theorem 2.10]. □

As a consequence, two spaces that are homotopy equivalent have the same Betti numbers. Notice that, when $X$ is not homotopy equivalent to a finite simplicial complex, its Betti numbers might not be finite.

## 3. Deterministic Betti Numbers Inference

Singular homology allows to consider Betti numbers of compact sets in $\mathbb{R}^d$ and of their offsets. Using its connection to simplicial homology and the distance functions framework of the previous chapter on *reconstruction of compact sets*, we derive explicit methods to infer the Betti numbers of compact subsets with positive weak feature size.

### 3.1. **A First Method.**

Let $K \subset \mathbb{R}^d$ be a compact set with $\mathrm{wfs}(K) > 0$ and let $\mathcal{P} \in \mathbb{R}^d$ be a finite set of points such that $\mathrm{d_H}(K, \mathcal{P}) < \varepsilon$ for some given $\varepsilon > 0$. Recall that, from Grove's isotopy lemma, all the $r$-offsets $K^r$ of $K$, for $0 < r < \mathrm{wfs}(K)$, are homeomorphic and thus have isomorphic homology groups. The goal of this section is to provide an effective method to compute the Betti numbers $\beta_k(K^r)$, $0 < r < \mathrm{wfs}(K)$, from $\mathcal{P}$.

Theorem 3.1 (Chazal, Lieutier). *Let $K \subset \mathbb{R}^d$ be a compact set with $\mathrm{wfs}(K) > 0$ and let $\mathcal{P} \subset \mathbb{R}^d$ be a finite set of points such that $\mathrm{d_H}(K, \mathcal{P}) < \varepsilon$ for $\varepsilon > 0$ such that $\mathrm{wfs}(K) > 4\varepsilon$. For $\alpha > 0$ such that $4\varepsilon + \alpha < \mathrm{wfs}(K)$, let $i : \mathcal{P}^{\alpha+\varepsilon} \hookrightarrow \mathcal{P}^{\alpha+3\varepsilon}$ be the canonical inclusion.*

*Then for all integer $k \geqslant 0$ and $0 < r < \mathrm{wfs}(K)$,*

$$H_k(K^r) \cong \mathrm{Im}(i_* : H_k(\mathcal{P}^{\alpha+\varepsilon}) \to H_k(\mathcal{P}^{\alpha+3\varepsilon})),$$

*where* Im *denotes the image of the homomorphism and $\cong$ means two groups are isomorphic.*

*Proof.* Since $\mathrm{d_H}(K, \mathcal{P}) < \varepsilon$, we have the following sequence of inclusion maps

$$K^\alpha \subset \mathcal{P}^{\alpha+\varepsilon} \subset K^{\alpha+2\varepsilon} \subset \mathcal{P}^{\alpha+3\varepsilon} \subset K^{\alpha+4\varepsilon}$$

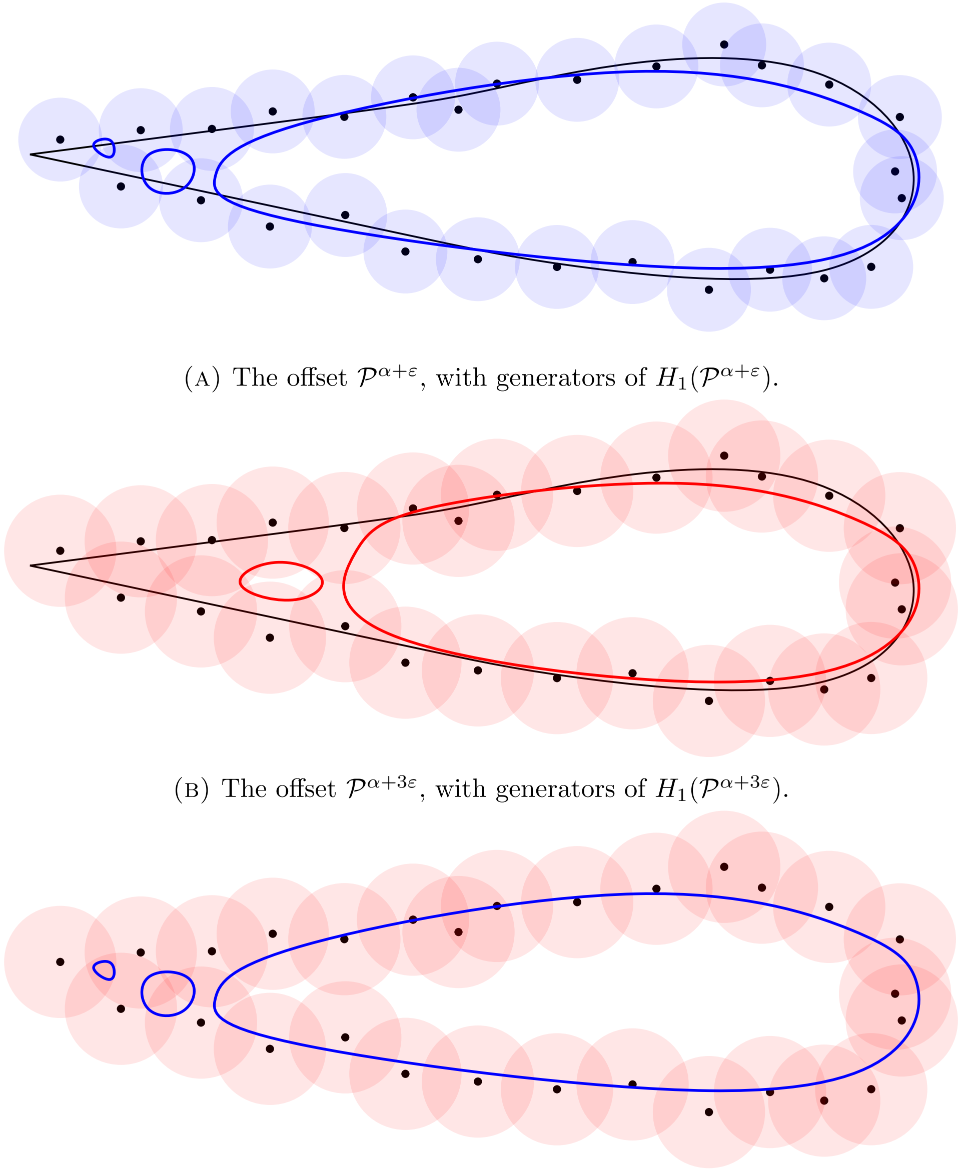
that induces he following sequence of homomorphisms (the one induced by the canonical inclusion maps) at the homology level

$$H_k(K^\alpha) \to H_k(\mathcal{P}^{\alpha+\varepsilon}) \to H_k(K^{\alpha+2\varepsilon}) \to H_k(\mathcal{P}^{\alpha+3\varepsilon}) \to H_k(K^{\alpha+4\varepsilon}).$$

Since $\mathrm{wfs}(K) > \alpha + 4\varepsilon$, it follows from Grove's isotopy lemma that the homomorphisms $u : H_k(K^\alpha) \to H_k(K^{\alpha+2\varepsilon})$ and $v : H_k(K^{\alpha+2\varepsilon}) \to H_k(K^{\alpha+3\varepsilon})$ induced by the inclusions maps are indeed isomorphisms, so that we display the following diagram

$$\begin{tikzcd}
H_k(K^\alpha) \arrow[rr, "u"] \arrow[dr, "a"] & & H_k(K^{\alpha+2\varepsilon}) \arrow[rr, "v"] \arrow[dr, "c"] & & H_k(K^{\alpha+4\varepsilon}) \\
& H_k(\mathcal{P}^{\alpha+\varepsilon}) \arrow[ur, "b"] \arrow[rr, "i_*"] & & H_k(\mathcal{P}^{\alpha+3\varepsilon}) \arrow[ur, "d"] &
\end{tikzcd}$$

It follows from elementary linear algebra (see Lemma 3.2) that the rank of $i_* : H_k(\mathcal{P}^{\alpha+\varepsilon}) \to H_k(\mathcal{P}^{\alpha+3\varepsilon})$ is equal to the rank of these isomorphisms which is equal to $\beta_k(K^\alpha)$.



(A) The offset $\mathcal{P}^{\alpha+\varepsilon}$, with generators of $H_1(\mathcal{P}^{\alpha+\varepsilon})$.

(B) The offset $\mathcal{P}^{\alpha+3\varepsilon}$, with generators of $H_1(\mathcal{P}^{\alpha+3\varepsilon})$.

(C) The generators of $H_1(\mathcal{P}^{\alpha+\varepsilon})$ when seen in $H_1(\mathcal{P}^{\alpha+3\varepsilon})$: the two small 1-cycles of $Z_1(\mathcal{P}^{\alpha+\varepsilon})$ on the left are sent to zero in $H_1(\mathcal{P}^{\alpha+3\varepsilon})$, since they belong to $B_1(\mathcal{P}^{\alpha+3\varepsilon})$. Hence, $\mathrm{Im}(i_* : H_1(\mathcal{P}^{\alpha+\varepsilon}) \to H_1(\mathcal{P}^{\alpha+3\varepsilon}))$, has dimension equal to 1.

Figure 9. The idea behind Theorem 3.1 is the following: both $\mathcal{P}^{\alpha+\varepsilon}$ and $\mathcal{P}^{\alpha+3\varepsilon}$, taken individually, may contain topological artifacts due to sampling. However, in c we observe that most of these homological features do not survive the transition. All the extra connected components in a are merged into the large component in b. Similarly, all the extra cycles in a are filled up in b.

Indeed, as $u = b \circ a$ and $v = d \circ c$ are isomorphisms, we get that $a$ and $c$ are injective, and that $b$ and $d$ are surjective. As a result, noticing that $i_* \circ a = c \circ u$, we obtain

$$\mathrm{rank}(i_*) = \mathrm{rank}(i_* \circ a) = \mathrm{rank}(c \circ u) = \mathrm{rank}(u) = \beta_k(K^\alpha),$$

which concludes the proof. □

Lemma 3.2 (Sandwich Lemma). *Consider the following sequence of homomorphisms between finite-dimensional vector spaces over a same field:*

$$A \to B \to C \to D \to E \to F.$$

- *If* $\mathrm{rank}(A \to F) = \mathrm{rank}(C \to D)$*, then this quantity also equals the rank of* $B \to E$*.*
- *In the same way, if* $\mathrm{rank}(A \to F) = \dim C$*, then* $\mathrm{rank}(B \to E) = \dim C$*.*

*Proof.* Observe that, for any sequence of homomorphisms $F \xrightarrow{f} G \xrightarrow{g} H$, we have $\mathrm{rank}(g \circ f) \leqslant \min\{\mathrm{rank} f, \mathrm{rank} g\}$. Applying this fact to maps $A \to F$, $B \to E$, and $C \to D$, which are nested in the sequence of the lemma, we get: $\mathrm{rank}(A \to F) \leqslant \mathrm{rank}(B \to E) \leqslant \mathrm{rank}(C \to D)$, which proves the first statement of the lemma. As for the second statement, it is obtained from the first one by letting $D = C$ and taking $C \to D$ to be the identity map. □

3.2. **Using Simplicial Complexes.** Theorem 3.1 shows that the Betti numbers of the offsets of $K$ can be deduced from the offsets of $\mathcal{P}$. However, the direct computation of the homology groups of a union of balls, which is a continuous object and not a finite simplicial complex, is not obvious. To overcome this issue, recall that the Nerve Theorem 1.14 implies that for any $r \geqslant 0$, $\mathcal{P}^r$ is homotopy equivalent to $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, r)$. As a consequence $H_k(\mathcal{P}^r)$ and $H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, r))$ are isomorphic. Moreover, one can show that the isomorphisms can be chosen to commute with the ones induced by inclusions maps, making the following diagram commutative

$$\begin{tikzcd}
H_k(\mathcal{P}^r) \arrow[r] & H_k(\mathcal{P}^{r'}) \\
H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, r)) \arrow[r] \arrow[u, "\wr|"'] & H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, r')) \arrow[u, "\wr|"']
\end{tikzcd}$$

We immediately obtain the following result.

Theorem 3.3 (Chazal, Oudot). *Assume that* $\mathrm{d_H}(K, \mathcal{P}) < \varepsilon$ *and* $\mathrm{wfs}(K) > 4\varepsilon$*. For* $\alpha > 0$ *such that* $4\varepsilon + \alpha < \mathrm{wfs}(K)$*, let* $i : \check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha + \varepsilon) \hookrightarrow \check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha + 3\varepsilon)$ *be the canonical inclusion.*

*Then for all integer* $k \geqslant 0$ *and* $0 < r < \mathrm{wfs}(K)$*,*

$$H_k(K^r) \cong \mathrm{Im}(i_* : H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha + \varepsilon)) \to H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha + 3\varepsilon)).$$

Thanks to the previous proposition, inferring the Betti numbers of $K^r$ now boils down to homology computation on finite Čech complexes. However, as already noticed in Section 1.5, computing Čech complexes require to determine if finite sets of balls intersect, which quickly becomes prohibitive as $d$ and the cardinality of $\mathcal{P}$ increase. Using the interleaving property between Čech and Vietoris-Rips filtrations established in Lemma 1.19, we obtain the following theorem.

Theorem 3.4 (Chazal, Oudot). *Assume that* $\mathrm{d_H}(K, \mathcal{P}) < \varepsilon$ *and* $\mathrm{wfs}(K) > 9\varepsilon$*. For all* $2\varepsilon < \alpha < \frac{1}{4}(\mathrm{wfs}(K) - \varepsilon)$ *and all* $0 < r < \mathrm{wfs}(K)$*, we have*

$$\beta_k(K^r) = \mathrm{rank}\,(i_* : H_k(\mathrm{Rips}(\mathcal{P}, \alpha)) \to H_k(\mathrm{Rips}(\mathcal{P}, 4\alpha))\,,$$

*where* $i : \mathrm{Rips}(\mathcal{P}, \alpha) \hookrightarrow \mathrm{Rips}(\mathcal{P}, 4\alpha)$ *denotes the canonical inclusion.*

*Proof.* From Lemma 1.19 we have the sequence of inclusions:

$$\begin{tikzcd}
\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha/2) \arrow[r, hook] & \mathrm{Rips}(\mathcal{P}, \alpha) \arrow[r, hook] & \check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha) \arrow[d, hook] \\
\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, 4\alpha) & \mathrm{Rips}(\mathcal{P}, 4\alpha) \arrow[l, hook'] & \check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, 2\alpha) \arrow[l, hook']
\end{tikzcd}$$

Since $\alpha \geqslant 2\varepsilon$, Theorem 3.3 implies that in the sequence of induced homomorphisms at the homology level, $H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha/2) \to H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, 4\alpha)$ and $H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha) \to H_k(\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, 2\alpha)$ have ranks equal to $\beta_k(K^r)$. Hence, $\mathrm{rank}\,(i_*)$ is also equal to $\beta_k(K^r)$ (see Lemma 3.2), which concludes the proof. □

## 4. Rates of Convergence for Random Point Clouds

4.1. **Minimax Upper Bound.** In the smooth case where $K$ has positive reach ($\mu = 0$), one can actually prove that $K$ is homotopy equivalent to its offsets, for small enough radii, as stated in the following lemma.

Lemma 4.1. *Let* $M \subset \mathbb{R}^d$ *be a* $k$*-dimensional compact submanifold with positive reach* $\mathrm{reach}(M) \geqslant \tau > 0$*.*
*Then for all* $r \in (0, \tau)$*,* $M$ *is homotopy equivalent to* $M^r$*.*

*Proof.* Consider $f = \pi_M : M^r \to M$ the projection map onto $M$. $f$ is well defined, since $M^r \subset \mathrm{Med}(M)^c$. Write $g = \mathrm{id}_M$, so that $f \circ g = \mathrm{id}_M$. Let us show that $g \circ f = f$ is homotopy equivalent to $\mathrm{id}_{M^r}$ by considering $f(x,t) = t\pi_M(x) + (1-t)x$. $f$ is continuous, and $f : M^r \to M^r$, since for all $x \in M^r$,

$$\mathrm{d}_M(f(x,t)) \leqslant \|f(x,t) - \pi_M(x)\| = (1-t)\,\|x - \pi_M(x)\| \leqslant r.$$

□

Corollary 4.2 (Homology Inference under Reach Condition). *Let* $M \subset \mathbb{R}^d$ *be a* $k$*-dimensional compact submanifold with* $\mathrm{reach}(M) \geqslant \tau > 0$*. Assume that* $\tau > 9\varepsilon$*. Then for all integer* $\ell \geqslant 0$ *and* $2\varepsilon < \alpha < \frac{1}{4}(\tau - \varepsilon)$*, we have*

$$\beta_\ell(M) = \mathrm{rank}\,(i_* : H_\ell(\mathrm{Rips}(\mathcal{P}, \alpha)) \to H_\ell(\mathrm{Rips}(\mathcal{P}, 4\alpha))\,,$$

*where* $i : \mathrm{Rips}(\mathcal{P}, \alpha) \hookrightarrow \mathrm{Rips}(\mathcal{P}, 4\alpha)$ *denotes the canonical inclusion.*

*Proof.* Combine Theorem 3.4 with Lemma 4.1 and use $\mathrm{wfs}(M) \geqslant \mathrm{reach}(M)$. □

**Definition 4.3** (Statistical Model)**.** We let $\mathcal{Q}_{k,\tau,a}$ denote the set of Borel probability distributions $Q$ on $\mathbb{R}^d$ such that
– $M = \mathrm{supp}\, Q$ is a $k$-dimensional submanifold with $\mathrm{reach}(M) \geqslant \tau > 0$.
– $Q$ is $(a,k)$-standard at scale $r_0 \geqslant \tau/40$.

The following result provides a convergence rate for inference of homology groups from random point clouds.

Proposition 4.4 (Homology Inference under Reach Condition). *Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be a i.i.d. $n$-sample of some $Q \in \mathcal{Q}_{k,\tau,a}$. For all integer $\ell \geqslant 0$ write*

$$\hat{\beta}_\ell = \operatorname{rank}\left(i_* : H_\ell(\mathrm{Rips}(\mathbb{X}_n, \alpha_0)) \to H_\ell(\mathrm{Rips}(\mathbb{X}_n, 4\alpha_0)\right),$$

*where $\alpha_0 = \tau/5$ and $i : \mathrm{Rips}(\mathcal{P}, \alpha_0) \hookrightarrow \mathrm{Rips}(\mathcal{P}, 4\alpha_0)$ denotes the canonical inclusion.*
*Then for $n$ large enough,*

$$\mathbb{P}_Q\left(\exists \ell \geqslant 0 \text{ such that } \beta_\ell(M) \neq \hat{\beta}_\ell\right) \leqslant \frac{4^k}{a'\tau^k} \exp\left(-na'\tau^k\right),$$

*where $a' = a/20^k$.*

**Remark 4.5.** This bound only depends on $a, k$ and $\tau$.

*Proof.* From Proposition 6.4 of previous lesson and Corollary 4.2, denoting $\varepsilon = \mathrm{d_H}(M, \mathbb{X}_n)$, we have

$$\begin{aligned}
&\mathbb{P}_Q\left(\exists \ell \geqslant 0 \text{ such that } \beta_\ell(M) \neq \hat{\beta}_\ell\right) \\
&\leqslant \mathbb{P}_Q\left(9\varepsilon \geqslant \tau \text{ or } 2\varepsilon \geqslant \alpha_0 \text{ or } \alpha_0 \geqslant \frac{1}{4}(\tau - \varepsilon)\right) \\
&\leqslant \mathbb{P}_Q\left(\mathrm{d_H}(M, \mathbb{X}_n) > \tau/20\right) \\
&\leqslant \frac{4^k}{a'\tau^k} \exp\left(-na'\tau^k\right),
\end{aligned}$$

where $a' = a/20^k$. □

**Definition 4.6** (Minimax Risk for Homology Inference)**.** We let

$$R_n\big(\mathcal{Q}_{k,\tau,a}\big) = \inf_{\hat{\beta}} \sup_{Q \in \mathcal{Q}_{k,\tau,a}} \mathbb{P}_Q\left(\exists \ell \geqslant 0 \text{ such that } \beta_\ell(M) \neq \hat{\beta}_\ell\right),$$

where $\hat{\beta} : (\mathbb{R}^d)^n \to \mathbb{N}^{\mathbb{N}}$ ranges among all the estimators $\hat{\beta} = \hat{\beta}(X_1, \ldots, X_n)$ based on $n$-samples.

Corollary 4.7. *For all $n \geqslant 1$,*

$$R_n\big(\mathcal{Q}_{k,\tau,a}\big) \leqslant \frac{4^k}{a'\tau^k} \exp\left(-na'\tau^k\right),$$

*where $a' = a/20^k$.*

*Proof.* Follows straightforwardly from Proposition 4.4. □

4.2. **Minimax Lower Bound.**

**Definition 4.8** (Total Variation)**.** Given two probability distributions $Q, Q'$ on a measurable space $(\mathcal{X}, \mathcal{A})$, the *total variation* between them is

$$\mathrm{TV}(Q, Q') = \sup_{A \in \mathcal{A}} \left|Q(A) - Q'(A)\right|.$$

TV is a distance on the space of probability measures on $(\mathcal{X}, \mathcal{A})$, and

$$\mathrm{TV}(Q, Q') = \frac{1}{2}\int_{\mathcal{X}} |q - q'| \mathrm{d}\nu = 1 - \int_{\mathcal{X}} q \wedge q' \mathrm{d}\nu,$$

where $\nu$ is any measure dominating $Q$ and $Q'$, and $q = \mathrm{d}Q/\mathrm{d}\nu$, $q' = \mathrm{d}Q'/\mathrm{d}\nu$.

**Remark 4.9.** The notation $\int_{\mathcal{X}} |\mathrm{d}Q - \mathrm{d}Q'| := \int_{\mathcal{X}} |q - q'|\mathrm{d}\nu$ is often used to emphasize that this quantity does not depend on the chosen dominating measure $\nu$.

*Proof.* TV clearly is well defined, separated, symmetric, and satisfies triangle inequality. Furthermore, taking $A_0 = \{q \geqslant q'\}$ yields

$$0 = \int_{\mathcal{X}} (q - q')\mathrm{d}\nu = \int_{A_0} (q - q')\mathrm{d}\nu - \int_{A_0^c} (q' - q)\mathrm{d}\nu,$$

so that

$$\begin{aligned}\int_{\mathcal{X}} |q - q'|\mathrm{d}\nu &= \int_{A_0} (q - q')\mathrm{d}\nu + \int_{A_0^c} (q' - q)\mathrm{d}\nu \\ &= 2\int_{A_0} (q - q')\mathrm{d}\nu.\end{aligned}$$

As a consequence,

$$\mathrm{TV}(Q, Q') \geqslant Q(A_0) - Q'(A_0) = \frac{1}{2}\int_{\mathcal{X}} |q - q'|\mathrm{d}\nu.$$

On the other hand, for all $A \in \mathcal{A}$,

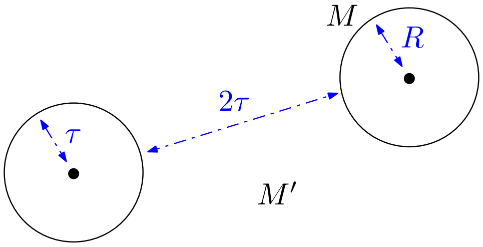
$$\begin{aligned}|Q(A) - Q'(A)| &= \left| \int_{A\cap A_0} (q - q')\mathrm{d}\nu + \int_{A\cap A_0^c} (q - q')\mathrm{d}\nu \right| \\ &\leqslant \max\left\{ \int_{A\cap A_0} (q - q')\mathrm{d}\nu, \int_{A\cap A_0^c} (q' - q)\mathrm{d}\nu \right\} \\ &= \frac{1}{2}\int_{\mathcal{X}} |q - q'|\mathrm{d}\nu.\end{aligned}$$

The last claim follows from the identity $|q - q'| = q + q' - 2q \wedge q'$. □

Lemma 4.10 (Le Cam). *Let $\mathcal{Q}$ be a set of probability distributions, and $\theta : \mathcal{Q} \to \Theta$ be a parameter of interest, where $(\Theta, \rho)$ is a metric space. Then for all $Q, Q' \in \mathcal{Q}$,*

$$\inf_{\hat{\theta}} \sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q^n} \rho\big(\theta(Q), \hat{\theta}_n\big) \geqslant \frac{1}{2}\rho\left(\theta(Q), \theta(Q')\right)\left(1 - \mathrm{TV}(Q, Q')\right)^n,$$

*where $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ ranges among all the measurable maps $\hat{\theta}_n : \mathcal{X}^n \to \Theta$ based on an i.i.d. $n$-sample.*



*Proof.* Let $\nu$ be a measure that dominates both $Q$ and $Q'$, with associated densities $q, q'$. For all measurable $\hat{\theta}_n : \mathcal{X}^n \to \Theta$,

$$\begin{aligned}\sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q^n}\rho\big(\theta(Q),\hat{\theta}_n\big) &\geqslant \frac{1}{2}\Big(\mathbb{E}_{Q^n}\rho\big(\theta(Q),\hat{\theta}_n\big) + \mathbb{E}_{Q'^n}\rho\big(\theta(Q'),\hat{\theta}_n\big)\Big)\\ &= \frac{1}{2}\int_{\mathcal{X}^n}\Big(\rho\big(\theta(Q),\hat{\theta}_n\big)q^{\otimes n} + \rho\big(\theta(Q'),\hat{\theta}_n\big)q'^{\otimes n}\Big)\,\mathrm{d}\nu^{\otimes n}\\ &\geqslant \frac{1}{2}\int_{\mathcal{X}^n}\Big(\rho\big(\theta(Q),\hat{\theta}_n\big) + \rho\big(\theta(Q'),\hat{\theta}_n\big)\Big)\, q^{\otimes n}\wedge q'^{\otimes n}\mathrm{d}\nu^{\otimes n}\\ &\geqslant \frac{1}{2}\rho\left(\theta(Q),\theta(Q')\right)\prod_{i=1}^n\int_{\mathcal{X}} q(x_i)\wedge q'(x_i)\mathrm{d}\nu(x_i)\\ &= \frac{1}{2}\rho\left(\theta(Q),\theta(Q')\right)\left(1-\mathrm{TV}(Q,Q')\right)^n.\end{aligned}$$

□

Proposition 4.11. *If $k < d$ and $\tau \leqslant 1$, then for $a > 0$ small enough,*

$$R_n(\mathcal{Q}_{k,\tau,a}) \geqslant \frac{1}{2}\exp(-n),$$

*for some $C > 0$.*

*Proof.* We apply Le Cam's lemma with $\mathcal{Q} = \mathcal{Q}_{k,\tau,a}$, $\Theta = \mathbb{N}^{\mathbb{N}}$ and $\rho(\beta,\beta') = \mathbb{1}_{\beta=\beta'}$. We let $M = \mathcal{S}^k(0,R)$ denote the centered $k$-dimensional sphere with radius $R \geqslant \tau$, and $M' = M \cup \mathcal{S}^k(\tau)$, where $\mathcal{S}^k(\tau)$ is at distance at least $2\tau$ from $M$. Write $Q$ and $Q'$ for the uniform distributions on $M$ and $M'$ respectively. $Q$ and $Q'$ are dominated by the $k$-dimensional Hausdorff measure $\nu = \mathcal{H}^k$. Denoting $\sigma_k$ for the surface area of the $k$-dimensional unit sphere, the densities of $Q$ and $Q'$ with respect to $\nu$ are $f = \frac{1}{\sigma_k R^k}\mathbb{1}_{\mathcal{S}^k(0,R)}$ and $f' = \frac{1}{\sigma_k(R^k+\tau^k)}\big(\mathbb{1}_{\mathcal{S}^k(0,R)} + \mathbb{1}_{\mathcal{S}^k(\tau)}\big)$ respectively. Hence,

$$\begin{aligned}2\,\mathrm{TV}(Q,Q') &= \int_{\mathbb{R}^d}|f-f'|\mathrm{d}\nu\\ &= \int_{\mathcal{S}(0,R)}\left|\frac{1}{\sigma_k R^k} - \frac{1}{\sigma_k(R^k+\tau^k)}\right|\mathrm{d}\mathcal{H}^k + \int_{\mathcal{S}(\tau)}\frac{1}{\sigma_k(R^k+\tau^k)}\mathrm{d}\mathcal{H}^k\\ &= \left|\frac{1}{\sigma_k R^k} - \frac{1}{\sigma_k(R^k+\tau^k)}\right|\sigma_k R^k + \frac{1}{\sigma_k(R^k+\tau^k)}\sigma_k\tau^k\\ &= \frac{2\tau^k}{R^k+\tau^k}.\end{aligned}$$

Clearly, $\beta_0(M) = 1 \neq 2 = \beta_0(M')$. Furthermore, $M$ and $M'$ are $k$-dimensional submanifolds with reach at least $\tau$, and if $a > 0$ is small enough,

$Q, Q'$ are $(a,k)$-standard. From Le Cam's Lemma 4.10,

$$R_n(\mathcal{Q}_{k,\tau,a}) \geqslant \frac{1}{2}\left(1 - \frac{\tau^k}{R^k + \tau^k}\right)^n \geqslant \frac{1}{2}\exp\left(-\frac{2n\tau^k}{R^k + \tau^k}\right),$$

where we used that $1 - t \geqslant \exp(-2t)$ whenever $0 \leqslant t \leqslant 1/2$. The result then follows by taking $R = \tau$. □

## 5. Further Sources

These notes mainly follow [BCY18] and [NSW08].

## References

[BCY18] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and Topological Inference*. Cambridge University Press, 2018. Cambridge Texts in Applied Mathematics.

[Hat02] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[NSW08] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.