Persistent Homology

Chirantan Mukherjee c.mukherjee@student.uw.edu.pl

This presentation is about “Probabilistic tools for high dimensional geometric inference, topological data analysis and large-scale networks”.

1. Introduction

The amount of data is dramatically increasing with time. This necessitates the development of innovative and efficient data-processing methods. But this is difficult due to the sheer size of the data in one hand and the complexity (noisy, high dimensional, incomplete) of the data in another hand.

Techniques from the relatively new subject of ’topological data analysis’ (TDA) have provided a wealth of new insights in the study of data in an increasingly diverse set of applications — including sensor-network coverage, proteins, 3-dimensional structure of DNA, development of cells, stability of fullerene molecules, robotics, signals in images, periodicity in time series, cancer, phylogenetics, natural images, the spread of contagions, self-similarity in geometry, materials science, financial networks, diverse applications in neuroscience, classification of weighted networks, collaboration networks, analysis of mobile phone data, collective behavior in biology, time-series output of dynamical systems, natural-language analysis, and more.

TDA uses ideas and results from computational geometry and topology to develop tools for studying qualitative features of data.
Persistent Homology is a an application in TDA which uses method from Algebraic Topology to study qualitative features of data with complex structure and is computable via linear algebra.

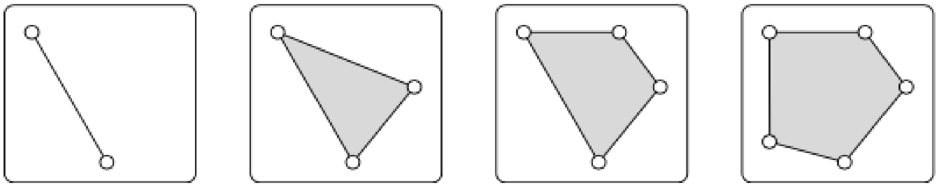
Types of data sets that can be studied with Persistent Homology include finite metric spaces (also known as **point cloud data** in the language of TDA), digital images, level sets of real-valued functions, and networks.

2. Homology

The original idea of Homology was the observation if two shapes can be distinguished by their holes? For instance, a circle is not a disk because the circle has a hole through it while the disk is solid.
It can be very difficult to compute the homology of arbitrary topological spaces. We thus approximate our spaces by combinatorial structures called ’simplicial complexes’.

2.1 Simplical Complexes and Homology Functors

Let $u_0, u_1, ..., u_k \in \mathbb{R}^d$. A point $x = \sum_{i=0}^{i=k} \lambda_i u_i$ where $\lambda_i \in \mathbb{R}$ is an **affine combination** of the u_i if $\sum_{i=0}^{i=k} \lambda_i = 1$. An **affine hull** is the set of affine combination.
A **convex combination** is an affine combination where the $\lambda_i \geq 0 \ \forall i$. Similarly, a **convex hull** is the collection of convex combination.



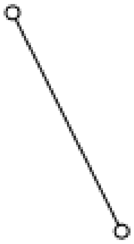
(i) Construction of a convex hull of 5 points by adding 1 point in each step

Definition 1 *The $(p+1)$ points in R^d are* **affinely independent** *iff the p-vectors $\{u_1 - u_0, u_2 - u_0, ..., u_p - u_0\}$ are linearly independent.*

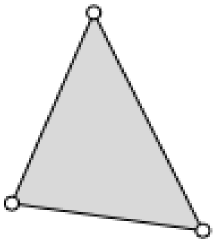
Definition 2 *A p-***simplex** *K_p is the convex hull of $(p+1)$-affinely independent points.*

The elements of K_0 (0-simplex) are called **vertices**, K_1 (1-simplex) are called **edges**, K_2 (2-simplex) are called **triangles with interior**, K_3 (3-simplex) are called **solid tetrahedron** and so on and so forth.

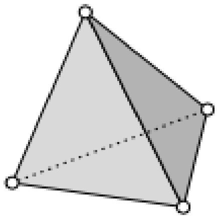
(ii) Vertex (0−simplex)



(iii) Edge (1−simplex)



(iv) Triangle with Interior (2−simplex)



(v) Solid Tetrahedron (3−simplex)

If τ and σ are simplices such that $\tau \subset \sigma$ then we say τ is a **face** of σ and σ is a **coface** of τ. In particular, the convex hull of a subset of size $(m+1)$ (of the $(p+1)$ defining points) is an m-simplex, called an m-face of the p-simplex.

In topology and combinatorics, it is common to "glue together" simplices (plural of simplex) to form a simplicial complex.

Definition 3 *A* **simplicial complex** K *is a set of simplices that satisfies the following conditions:*
(a) *Every face of a simplex from* K *is also in* K*.*
(b) *The intersection of any two simplices* σ_1*,* $\sigma_2 \in K$ *is either empty or a face of both.*

We say that a simplex has **dimension** p or is a p**-simplex** if it has a cardinality $p+1$.
We denote by K_p the collection of p-simplices (simplices of dimension p). While we define k**-skeleton** of K as the union of all p-simplices for $p = 0, 1, ..., k$.
The dimension of K is the maximum of the dimensions of it's simplices.
A **map of simplical complex** $f : K \to L$ is a map $f : K_0 \to L_0$ such that $f(\sigma) \in L$ $\forall \sigma \in K$.

Given a simplical complex K, let us define the Vector Space $C_p(K)$ with basis given by p-simplices of K. We call this as p**-chain**. By default we assume $C_{-1}(K) = 0$.

The **boundary** of a p-simplex is the sum of its $(p-1)$ dimensional faces. Writing $\sigma = [u_0, ..., u_p]$ for the simplex spanned by the set of listed vertices, its boundary is $d_p \cdot \sigma = \sum_{k=0}^{p} [u_0, ..., \hat{u}_k, ..., u_p]$ where the $\hat{u}_k$ indicates that u_k is omitted.
For a p-chain, $c = \sum a_i \sigma_i$, the boundary is the sum of the boundaries of its simplices
$d_p : C_p(K) \to C_{p-1}(K)$
$c \to \sum a_i d_p \sigma_i$

NOTE: Boundary of a boundary is zero, i.e. $d_{p-1} \cdot d_p = 0$
Hence, the $Im(d_{p+1})$ is contained in the $Ker(d_p)$, i.e. $Im(d_{p+1}) \subset Ker(d_p) \subset C_p$.

Definition 4 *The* p**th homology** *of a simplical complex* K *is the quotient Vector Space* $H_p(K) = \frac{Ker(d_p)}{Im(d_{p+1})}$ $\forall p = 0, 1, ...$

Definition 5 *The dimension of the pth homology is known as the the* p**th Betti Number** *of* K*.*
$\beta_p(K) = dimH_p(K) = dimKer(d_p) - dimIm(d_{p+1})$

Elements in the image of d_{p+1} $(Im(d_{p+1}))$ are called p**-boundaries**, and elements in the kernel of d_p $(Ker(d_p))$ are called p**-cycles**. Intuitively, the p-cycles that are not boundaries represent p-dimensional holes. Therefore, the pth Betti number 'counts' the number of p-holes.
Additionally, if K is a simplicial complex of dimension n, then for all $p > n$, we have that $H_p(K) = 0$, as K_p is empty and hence $C_p(K) = 0$. We therefore obtain the following sequence of vector spaces and linear maps:

$0 \xrightarrow{d_{n+1}} C_n(K) \xrightarrow{d_n} C_{n-1}K \xrightarrow{d_{n-1}} ... \xrightarrow{d_2} C_1(K) \xrightarrow{d_1} C_0(K) \xrightarrow{d_0} 0$

NOTE: A very important consequence of the functorality of the simplical complex is the following commutative diagram:
$f : K \rightarrow K'$ be a continuous map between two simplicial complex. Then, $\tilde{f_p} \cdot d_{p+1} = d'_{p+1} \cdot \tilde{f}_{p+1}$ where $\tilde{f_p} : C_p(K) \rightarrow C_p(K')$.

2.2 Building of Simplicial Complexes

Computing the homology of finite simplicial complexes boils down to linear algebra. The same is not true for the homology of an arbitrary space X. We therefore try to search for simplicial complexes whose homology approximates the homology of the space X in an appropriate sense.
An important tool for doing so is the Čech Complex ($\check{C}$).

Definition 6 *Let $\mathcal{U}$ be a finite collection of sets in $\mathbb{R}^d$. The* **nerve** *of $\mathcal{U}$ consists of all non-empty subcollections whose sets have a non-empty intersection.*
$Nrv(\mathcal{U}) = \{U_i \subseteq \mathcal{U} \mid \cap_{i=0}^{n} U_i \neq \phi\}$

If the subcollection of the sets is sufficiently 'nice,' then the Nerve Theorem (stated below) implies that the $Nrv(\mathcal{U})$ and the space X have the same homology.

Definition 7 *Given two topological spaces X and Y, a* **homotopy equivalence** *between X and Y is a pair of continuous maps $f : X \rightarrow Y$ and $g : Y \rightarrow X$, such that gf is homotopic to the identity map id_X and fg is homotopic to id_Y. If such a pair exists, then X and Y are said to be* **homotopy equivalent**, *or of the same* **homotopy type**.

Intuitively, two spaces X and Y are homotopy equivalent if they can be transformed into one another by bending, shrinking and expanding operations.

Theorem 8 (NERVE THEOREM) *Let $\mathcal{U}$ be a finite collection of closed, convex sets in Euclidean space. Then the nerve of $\mathcal{U}$ and the union of the sets in $\mathcal{U}$ have the same homotopy type.*

The Nerve Theorem implies that the nerve of the cover and the space X have the same homology. Suppose, we have a finite set of points S in a metric space X. We can define, for every $\epsilon > 0$, the space $S_\epsilon = \cup_{x \in S} B(x, \epsilon)$, where $B(x, \epsilon)$ denotes the closed ball with radius ϵ centered at x. It follows that the set $\{B(x, \epsilon) \mid x \in S\}$ is a cover of S_ϵ, and the nerve of this cover is the **Čech complex** on S at scale ϵ. We denote this complex by $\check{C}_\epsilon(S)$.
If the space X is Euclidean space, then the Nerve Theorem guarantees that the simplicial complex $\check{C}_\epsilon(S)$ recovers the homology of S_ϵ.

3. Persistent Homology

The central concept of Persistent Homology is motivated by the practical need to cope with noise in data. This includes defining, recognizing, and possibly eliminating noise.

3.1 Filtered Complexes and Homology

Definition 9 *Let K be a finite simplicial complex, and let $\phi = K_0 \subset K_1 \subset K_2 \subset ... \subset K_l = K$ be a finite sequence of nested subcomplexes of K. The simplicial complex K with such a sequence of subcomplexes is called a* **filtered simplicial complex***.*

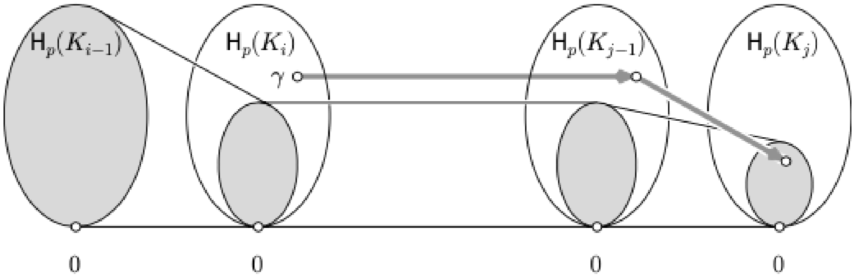
We can apply homology to each of the subcomplexes. For all p, the inclusion maps $K_i \to K_j$ induce $\mathbb{F}_2$-linear maps $f^{i,j} : H_p(K_i) \to H_p(K_j)\ \forall i, j \in \{1, ..., l\}$ with $i \leq j$. By functorality, $f_p^{k,j} f_p^{i,k} = f_p^{i,j}\ \forall i \leq k \leq j$

Definition 10 *Let $\phi = K_0 \subset K_1 \subset K_2 \subset ... \subset K_l = K$ be a filtered simplicial complex. The* **pth persistent homology** *of K is $H_p^{i,j} = Imf_p^{i,j}$, where $\forall i, j \in \{0, ..., l\}$ with $i \leq j$, the linear maps $f^{i,j} : H_p(K_i) \to H_p(K_j)$ are the maps induced by the inclusion maps $K_i \to K_j$.*

The corresponding, **p-th persistent Betti numbers** are the ranks of these groups, $\beta_p^{i,j} =$ rank $H_p^{i,j}$.

The pth persistent homology of a filtered simplicial complex gives more refined information than just the homology of the individual subcomplexes. The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,
$0 = H_p(K_0) \to H_p(K_1) \to ... \to H_p(K_n) = H_p(K)$
again one for each dimension p. As we go from K_{i-1} to K_i, we might gain new homology classes and we might lose some when they become trivial or merge with each other. We collect the classes that are born at or before a given threshold and die after another threshold in groups.

Definition 11 *A pth homology class $[\gamma]$ is* **born** *at K_i if $[\gamma] \in H_p(K_i)$, but $[\gamma] \notin H_p(K_{i-1})$. It* **dies** *entering K_j if it merges with a class that was born earlier. Formally stated, $f^{i,j-1}([\gamma]) \notin H_p^{i-1,j-1}$ but $f^{i,j}([\gamma]) \in H_p^{i-1,j}$.*



(vi) Birth and Death

The definition of birth is almost straightforward. But, the definition of death is a little subtle. When a class is merged with another class, we choose to kill the class that is born the latest. We call the difference in function value the **persistence**, $pers(\gamma) = a_j - a_i$. Sometimes, emphasizing on the index we take persistence as $j - i$.

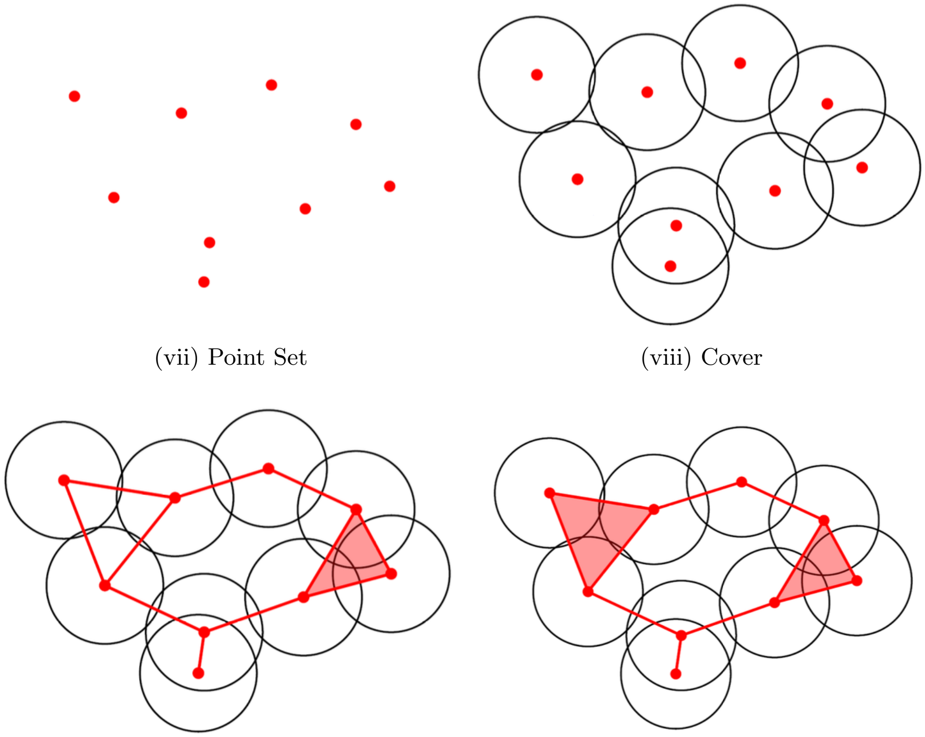
3.2 Filtered Simplical Complex

We introduced the Čech Complex ($\check{C}$), a classical simplicial complex from algebraic topology. However, there are many other simplicial complexes that are better suited for studying data from applications. We discuss them in this section.

3.2.1 Vietoris–Rips Complex

One of the disadvantages of the Čech Complex is that one has to check for a large number of intersections, which is difficult. But this can be overcome with the use of Vietoris–Rips (VR) Complex, which approximates the Čech Complex. Let S be a subset of a metric space (X, d). For a non-negative real number ϵ, the **Vietoris–Rips Complex** $VR_\epsilon(S)$ at scale ϵ is defined as:
$VR_\epsilon(S) = \{\sigma \subset S \mid d(x,y) \leq 2\epsilon \; \forall x, y \in \sigma\}$

Lemma 12 (VIETORIS-RIPS LEMMA) *Letting S be a finite set of points in some Euclidean space and letting $\epsilon \geq 0$, we have $\check{C}_\epsilon(S) \subseteq VR_\epsilon(S) \subseteq \check{C}_{\sqrt{2}\epsilon}(S)$.*



(vii) Point Set

(viii) Cover

(ix) Čech Complex

(x) Vietoris–Rips Complex

3.2.2 Delaunay Complex

For the Delaunay Complex, one usually considers $X = \mathbb{R}^d$. We subdivide the space $\mathbb{R}^d$ into regions of points that are closest to any of the points in S. Let S be a subset of a metric space (X, d).

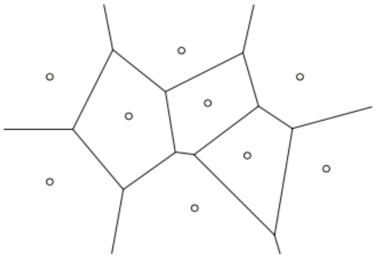
Definition 13 *For any $s \in S$, we define the set $V_S = \{x \in \mathbb{R}^d \mid d(x,s) \leq d(x,s') \ \forall s' \in S\}$. The collection of sets V_s is a cover of $\mathbb{R}_d$ that is called the* **Voronoi decomposition** *of $\mathbb{R}^d$ with respect to S.*

The nerve of this cover is called the **Delaunay Complex** of S and is denoted by $Del(S;\mathbb{R}^d)$.

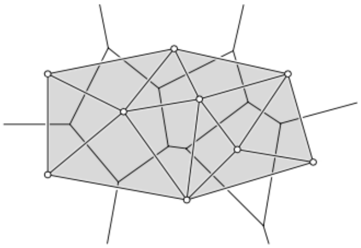
3.2.3 Alpha Complex

We assume that S is a finite set of points in $\mathbb{R}^d$. The Alpha complex is a subcomplex of the Delaunay complex.

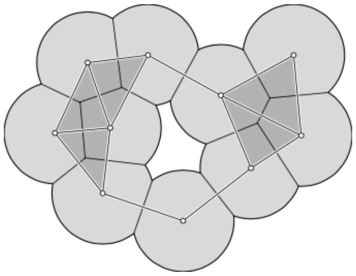
Let $\epsilon > 0$, and let $S_\epsilon = \cup_{s \in S} B(s, \epsilon)$. For every $s \in S$, consider the intersection $V_s \cap B(s, \epsilon)$. The collection of these sets forms a cover of S_ϵ, and the nerve complex of this cover is called the **Alpha Complex** of S at scale ϵ and is denoted by $A_\epsilon(S)$.
The Nerve Theorem implies that $A_\epsilon(S)$ has the same homology as S_ϵ.



(xi) Voronoi decomposition



(xii) Delaunay Complex



(xiii) Alpha Complex

3.3 Persistent Diagrams

We visualize the collection of persistent Betti numbers by drawing points in two dimensions. A visual representation of the the persistent homology can be created by drawing a collection of points in the plane. We consider the extended plane $(\mathbb{R}, +\infty)$ on which we represent a birth paired with the death as a point with two coordinates. Some of the classes may never die and thus be represented as points at infinity. Some others may have same coordinates because they may be born and die at the same time. This happens only when we allow multiple homology classes being created or destroyed at the same function value or filtration point.

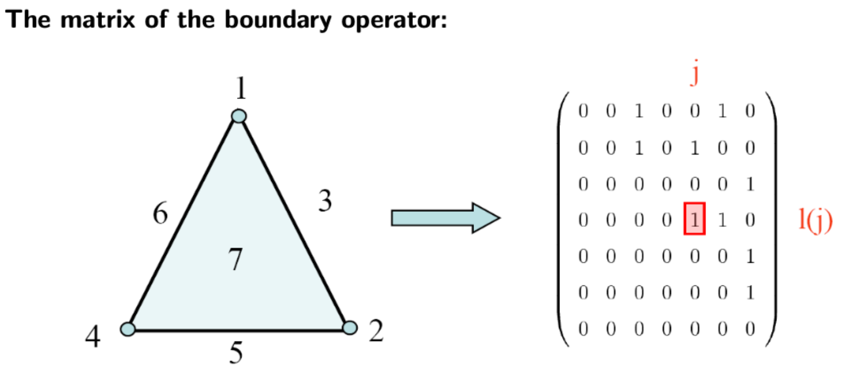
Let $\mu_p^{i,j}$ represent the number of independent p-dimensional classes that are born at K_i and die entering K_j, we have:
$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})\ \forall i < j$ and $\forall p$

Definition 14 *The* **persistence diagram** $Dgm_p(f)$ *of a filtration is induced by a function f and is obtained by drawing a point (a_i, a_j) with multiplicity $\mu_p^{i,j}$ on the extended plane where the diagonal D is added with infinite multiplicity.*

3.4 Matrix Reduction

To compute the Persistence Homology of a filtered simplicial complex K and obtain a **barcode**, we need to associate to it a matrix — the so-called **boundary matrix** — that stores information about the faces of every simplex. To do this, we place a total ordering on the simplices of the complex that is compatible with the filtration in the following sense:
(a) a face of a simplex precedes the simplex
(b) a simplex in the ith complex K_i precedes simplices in K_j for $j > i$, which are not in K

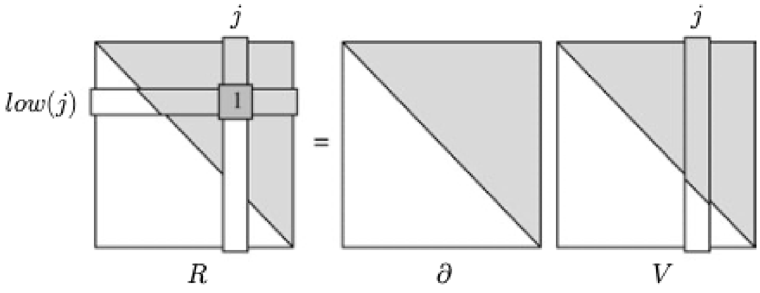
Let, n denote the total number of simplices in the complex, and let $\sigma_1, ..., \sigma_n$ denote the simplices with respect to this ordering. We construct a square matrix ∂ of dimension $n \times n$ by storing a 1 in $\partial[i, j]$ if the simplex σ_i is a face of simplex σ_j of codimension 1; otherwise, we store a 0 in $\partial[i, j]$.



$$\begin{array}{cccccccc}
 & & & & j & & & \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & \\
0 & 0 & 0 & 0 & \boxed{1} & 1 & 0 & l(j) \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 &
\end{array}$$

(xiv) Boundary Operator

Once one has constructed the boundary matrix, one has to reduce it using Gaussian elimination technique. Let $low(j)$ be the row index of the lowest 1 in column j. If the entire column is zero, then $low(j)$ is undefined. We call R reduced if $low(j) \neq low(j_0)$ whenever $j \neq j_0$, specify two non-zero columns. The algorithm reduces ∂ by adding columns from left to right. In matrix notation, this algorithm computes the reduced matrix as $R = \partial \cdot V$. Since, each simplex is preceded by its proper faces, ∂ is upper triangular. The j-th column of V encodes the columns in ∂ that add up to give the j-th column in R. Since we only add columns from left to right, V is also upper triangular and so is R.
To get the ranks of the homology groups of K, we notice that the number of zero columns of R that correspond to p-simplices is the rank of $Ker\partial_p$. Similarly, the number of non-



(xv) $R = \partial \cdot V$

zero columns gives the rank of $Im\partial_{p+1}$. The difference is the p-th Betti number $\partial_p = dimKer\partial_p - dimIm\partial_{p+1}$.

References

[CHA] Frédéric Chazal and Bertrand Michel. *Persistent homology in TDA.* Presentation at the INRIA, Barcelona (2016). Department of Computer Science and Engineering, The Ohio State University (2016).

[DEY] Tamal K. Dey. *Topological Persistence: Introduction, Stability, Algorithms.* Presentation at the Department of Computer Science and Engineering, The Ohio State University (2016).

[DUT] Kunal Dutta. *Dimensionality Reduction for Persistent Homology using k-Distance.* Presentation at the Department of Informatics, University of Warsaw, Poland (2020).

[EDE] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. *Topological Persistence and Simplification*⋆. Discrete & Computational Geometry, Springer-Verlag New York Inc. (2002).

[GRE] Marvin J Greenberg . *Lectures on algebraic topology.* W. A. Benjamin Inc (1967).

[HER] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction.* American Mathematical Society (2009).

[LOT] Martin Lotz. *Persistent homology for low-complexity models.* Proceedings, Royal Society Publications (2019).

[MOR] Dmitriy Morozov. *A Practical Guide to Persistent Homology.* Lawrence Berkeley National Lab.

[OTT] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod and Heather A Harrington. *A roadmap for the computation of persistent homology.* EPJ Data Science, a SpringerOpen Journal (2017).

[ROT] Joseph Rotman. *An Introduction to Algebraic Topology.* Graduate Texts in Mathematics, Springer-Verlag New York (1988).

[Zhu] Xiaojin Zhu. *Persistent Homology Tutorial*. Presentation at the Department of Computer Sciences, University of Wisconsin-Madison (2013).