

CLC \_\_\_\_\_

Number \_\_\_\_\_

UDC \_\_\_\_\_

Available for reference  Yes  No



# Undergraduate Thesis

**Thesis Title: The Method of Persistent Homology  
and its Applications**

**Student Name:** Xiabing Ruan

**Student ID:** 11710936

**Department:** Mathematics

**Program:** Math and Applied Math

**Thesis Advisor:** Yifei Zhu

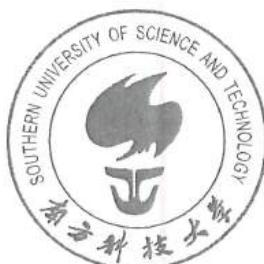
Date: 2021 年 5 月 26 日

分类号\_\_\_\_\_

编 号\_\_\_\_\_

U D C \_\_\_\_\_

密 级  Yes  No



南方科技大学  
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# 本科生毕业设计（论文）

题 目：持续同调方法及其应用

姓 名： 阮夏冰

学 号： 11710936

系 别： 数学系

专 业： 数学与应用数学

指导老师： 朱一飞

2021 年 5 月 26 日

## Contents

Contents .....	I
[摘要] .....	III
[ABSTRACT] .....	V
Notations .....	VII
Chapter 1 Introduction .....	1
Chapter 2 Preliminaries on Homology .....	3
2.1 Singular Homology .....	3
2.2 Simplicial Homology .....	7
Chapter 3 Persistence Homology .....	9
3.1 Complex Constructions .....	10
3.2 Persistence .....	11
Chapter 4 Sliding Window Persistence and its Application .....	17
4.1 Sliding Window Persistence .....	17
4.2 Application on Wheeze Detection .....	20
Chapter 5 Extended Persistence and its Application .....	25
5.1 Relative Homology and Morse Function .....	25
5.2 Extended Persistence .....	29
5.3 Application of the Extended Persistence in Protein Docking .....	34
Conclusions .....	41
References .....	43
Acknowledgements .....	45

## [摘要]

本文介绍一种数据分析的拓扑学方法——持续同调。拓扑空间的同调群包含其几何信息，粗略地说， $n$  维同调群的秩可以说明空间中  $n$  维“洞”的个数。持续同调的主要思想是计算同调群，得到数据点云所在空间的拓扑性质。具体来说，该方法从离散的数据点云出发，选取不同半径，连接距离小于半径长度数据点，构建一系列单纯复形，对每个单纯复形计算其同调群。在得到的一系列同调群中，用持续时间最长的一些同调类来估计数据点云所在空间的同调群。这种方法可以解决不同尺度的数据测量标准难以统一的问题。持续同调被广泛地应用于生物、医学、图像处理等领域中，它对数据的扰动具有稳定性。本文将深入讨论这一方法，包括其理论和计算，以及在喘息探测、蛋白质对接两个问题上的应用。

**[关键词]:** 持续同调，时间延迟嵌入，喘息探测，蛋白质对接

## [ABSTRACT]

This thesis introduces a topological method of data analysis, called persistent homology. Homology groups of a topological space contain geometric information: roughly speaking, the rank of the  $n$ -dimensional homology group indicates the number of  $n$ -dimensional “holes” in the space. The idea of persistent homology is to obtain topological features of the underlying space of a point cloud sample, by calculating homology groups. More specifically, starting with discrete point cloud data, one chooses different radii, connects points with distance less than each given radius to get a sequence of simplicial complexes, and calculates their simplicial homology groups. Then one chooses the homology classes that persist the longest to estimate the homology groups of the underlying space. This method resolves the difficulty with choosing relative measures for data in different scales. Persistent homology is widely used in areas such as biology, medical science, and image processing. It is stable under permutation of data in a precise sense. We discuss this method in detail, including theoretical and computational aspects. We give two applications, to the problems of wheeze detection in medical science and of protein docking in biology.

**[Keywords]:** persistent homology, time-delay embedding, wheeze detection, protein docking

## Notations

$\mathbb{C}$	complex number field
$\mathbb{R}$	real number field
$\mathbb{Q}$	rational number field
$\mathbb{Z}$	rational integer ring
$\mathbb{Z}_n$	the quotient ring $\mathbb{Z}/n\mathbb{Z}$
$S^n$	the $n$ -dimensional sphere $\{(x_0, \dots, x_n) \in \mathbb{R}^{n+1} \mid x_0^2 + \dots + x_n^2 = 1\}$
$\mathbb{K}\mathbb{P}^n$	the $n$ -dimensional projective space of the field $K$ , $\mathbb{K}\mathbb{P}^n = K^{n+1} - \{0\} /$ where $x \sim \lambda x$ for all $x \in K^{n+1} - \{0\}$ and $\lambda \neq 0 \in K$
$\mathbb{F}^n$	The Cartesian product $\mathbb{F} \times \dots \times \mathbb{F}$ of $n$ copies of the field $\mathbb{F}$

## Chapter 1 Introduction

There are many cases when we need to detect the “shape” of data points. For example, taking 100 points on the unit circle randomly, one can recognize the shape of circle by eyes. In dynamical systems, circular attractors imply periodicity. What if the circle is embedded in  $R^4$  or higher dimensional Euclidean space? Or even worse, what if the data is taken from a non-Euclidean topological space? This is when topological method being used. Recovering the underlying space of discrete data points is what persistent homology do.

Chapter 2 provides basic knowledge of homology. Starting with singular homology, which can be defined for all spaces, we introduce basic concepts and properties of it. We then come to simplicial homology, which is defined only for spaces with  $\Delta$ -complex structures but easier to calculate. The conclusions are that these two homology theories are isomorphic, and are both homotopic invariants. Homology groups are important in characterizing topological spaces.

Chapter 3 discusses technical details of persistent homology. A sequence of nested simplicial complexes indexed by integers, called a filtration, is constructed from a set of discrete data points. As the index grows, homology classes appear and become trivial. We track the birth and death for each homology classes and record them in the so-called persistent diagram. The persistent diagram gives us information of the homology classes who persist the longest, which are all we need to estimate the homology of the underlying space. The persistent diagram is calculated by reducing the boundary matrix of the filtration.

Chapter 4 introduces a technique called sliding window embedding. Combined with persistent homology, it can find the shape of times series data such as attractors of dynamical systems, which contain important information. This is further applied to the problem of wheeze detection, a medical problem related to lung diseases. Compared to other methods, the topological one performs better.

Chapter 5 gives a little generalization of persistent homology, that is, the extended persistence. The theory and computation of extended persistence is much like the ordinary persistence. It also start with a filtration and obtain the persistent diagram by matrix reduction, but requiring knowledge about the relative homology and Morse function, which we provide at the beginning of the chapter. Then we illustrate how it be used to the protein docking problem.

The final chapter concludes the article.

## Chapter 2 Preliminaries on Homology

To introduce the method of persistent homology, we need some basic knowledge of homology. Readers are assumed to have basic knowledge of point-set topology, smooth manifold, abstract algebra; otherwise they may refer to<sup>[1]</sup>,<sup>[2]</sup> and<sup>[3]</sup>. The main references for this chapter are<sup>[4]</sup> and<sup>[5]</sup>. The readers who are familiar with the notion of homology may skip this chapter.

Homology is used to study invariant properties of geometric objects under continuous deformations. Heuristically, it counts the holes of a space in all dimensions. For example, a circle has a one-dimensional hole, a torus has two one-dimensional holes and a two-dimensional hole and a 2-sphere  $S^2$  contains a two-dimensional hole. Different behavior of the holes yields to different homology groups.

### 2.1 Singular Homology

**Definition 2.1.1** (Standard Simplex). A **standard  $n$ -simplex** is defined to be  $\Delta_n := \{(x_0, \dots, x_n) \in \mathbb{R}^{n+1} \mid x_0, \dots, x_n \geq 0, \sum x_i = 1\}$ . Denote a standard  $n$ -simplex by  $[\nu_0, \dots, \nu_n]$ , where  $\nu_i$  refers to the corner  $(0, \dots, 0, 1, \dots, 0)$  of  $\Delta_n$  with the  $i$ -th coordinate being 1 and others being 0.

For example, a standard 0-simplex is a point; a standard 1 simplex is a line segment; a standard 2-simplex is a triangle (with interior), and so on.

**Definition 2.1.2** (Singular Simplex). Let  $X$  be a topological space, an  **$n$ -dimensional singular simplex** in  $X$  is a continuous map  $\sigma : \Delta_n \rightarrow X$ .

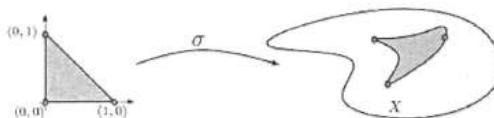


Figure 2.1: A singular simplex

**Definition 2.1.3** (Singular Chain). Let  $A$  be any abelian group. The  **$n$ -dimensional singular chain group with coefficients in  $A$**  of  $X$  is defined to be  $C_n(X; A) := \{\sum_{\text{finite}} a_\sigma \sigma \mid \sigma \text{ is an } n\text{-dimensional singualr complex in } X \text{ and } a_\sigma \in A\}$ . An element in  $C_n(X; A)$  is called an  $n$ -chain.

For example, if  $A = \mathbb{Z}$ ,  $C_n(X; A)$  is the free abelian group generated by all singular  $n$ -simplexes.  $C_0(X, \mathbb{Z})$  has a basis consists of all points in  $X$  since a map  $\Delta_0 \rightarrow X$  can be identified as a point in  $X$  and  $C_1(X, \mathbb{Z})$  has a basis consists of all continuous paths in  $X$ . When no ambiguity is caused, we omit the space  $X$ , the group  $A$  or both and simply write  $C_n(X; A)$  as  $C_n$ .

**Definition 2.1.4** (Boundary Map). Let  $\Delta_n = [v_0, v_1, \dots, v_n]$  be a standard  $n$ -simplex. Define the boundary of  $\Delta_n$  to be the formal sum  $\partial\Delta_n := \sum_{i=0}^n (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_n]$  where the hat indicates that  $v_i$  is deleted from the sequence  $v_0 \cdots v_n$  and each  $[v_0, \dots, \hat{v}_i, \dots, v_n]$  is a standard  $n-1$  simplex.

Let  $\sigma : \Delta_n \rightarrow X$  be a singular simplex, define the boundary of  $\sigma$  to be  $\partial\sigma := \sum_{i=0}^n (-1)^i \sigma|_{[v_0, \dots, \hat{v}_i, \dots, v_n]} \in C_{n-1}$ .

Extend linearly, we get a **boundary map**

$$\partial_n : C_n \rightarrow C_{n-1}, \partial_n(\sum a_\sigma \sigma) = \sum a_\sigma \partial\sigma$$

for each dimension.

Each  $\partial_n$  is of course a group homomorphism. Intuitively, the boundary of a triangle consists of three edges without the interior, which all appear in the expression of the boundary map. That is,  $\partial[v_0, v_1, v_2] = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$ . The signs indicate orientation, so the three edges are consistently oriented, forming a cycle. Similar geometric explanations hold for simplices in higher dimension.

We state without proving the following assertion, which can be checked by direct computation:

**Proposition 2.1.5.** For all  $n$ ,  $\partial_n \circ \partial_{n+1} = 0$ .

The proposition can be understood as “the boundary has no boundary”.

If we let  $C_n(X) = 0$  for  $n < 0$  and  $\partial_n = 0$  for  $n \leq 0$ , we get a sequence of abelian groups parametrized by  $\mathbb{Z}$

$$\cdots \rightarrow C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots$$

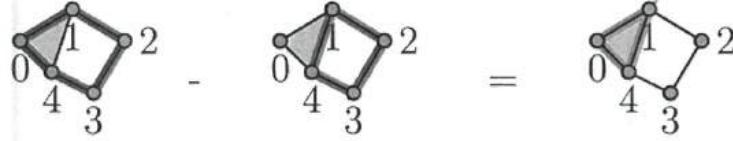
and the connecting boundary homomorphisms  $\partial_n$  with  $\partial_n \circ \partial_{n+1} = 0$ . Such a structure is called a **chain complex**. We usually omit the subscript  $n$  and write as  $\partial^2 = 0$ .

Since  $\partial_n \circ \partial_{n+1} = 0$ ,  $\text{im}\partial_{n+1}$  (the image of  $\partial_{n+1}$ )  $\subseteq \ker\partial_n$  (the kernel of  $\partial_n$ ). Both of them are subgroups of  $C_n$ , so we are able to do quotient.

**Definition 2.1.6.** Keep all the notions above. Let  $Z_n(X; A) := \ker\partial_n$ ,  $B_n(X; A) := \text{im}\partial_{n+1}$ , then  $Z_n \subseteq B_n$ . An element in  $Z_n$  is called an ( $n$ -dimensional) **closed chain**, or an ( $n$ -dimensional) **cycle**, and an element in  $B_n$  is called an ( $n$ -dimensional) **boundary chain**.

The  $n$ -dimensional singular homology group of  $X$  (with coefficients in  $A$ ) is  $H_n(X; A) := Z_n(X; A)/B_n(X; A)$ . The **homology class** of a cycle  $z \in Z_n(X; A)$  is the equivalence class  $[z]$  of  $z$  in  $H_n(X; A)$ . Two cycles  $z_1, z_2 \in Z_n(X; A)$  are called **homologous** if  $[z_1] = [z_2]$ , that is, they differ by a boundary chain  $c \in C_{n+1}(X; A)$  such that  $z_1 - z_2 = \partial c$ .

Homology groups are what we need to distinguish spaces. A cycle is a chain with no boundary, hence is “closed”, representing a “hole” of the space; Two homologous cycles represent the same “hole” of the space (see fig 2.2). Therefore, the homology group is the group of all cycles modulo the homologous relation. Before moving to the useful properties, we calculate some easy examples.



**Figure 2.2:** The two 1-cycles (01234) and (1234) are homologous, which differ by the boundary of the 2-simplex (014). They represent the same 1-dimensional hole of the space.

**Example 2.1.7.** Let  $X = \{*\}$  = a singleton. Then from every standard simplex  $\Delta_n$  there is exactly one map  $\sigma_n$  onto  $X$ , thus  $C_n(X; A) = \sigma_n \cdot A \cong A$  for all  $n \geq 0$ . By the definition of the boundary map,  $\partial_n \sigma_n = 0$  if  $n$  is odd, and  $\partial_n \sigma_n = \sigma_{n-1}$  if  $n$  is even.

If  $n$  is odd,  $Z_n(X; A) = B_n(X; A) = C_n(X; A)$ , so  $H_n(X; A) = 0$ .

If  $n \neq 0$  is even,  $Z_n(X; A) = B_n(X; A) = 0$ , so  $H_n(X; A) = 0$ .

If  $n = 0$ ,  $\partial_0 = 0$ ,  $Z_0(X; A) = C_0(X; A)$  and  $B_0(X; A) = 0$ , so  $H_0(X; A) = A$ .

Therefore, the one-point space  $X$  has trivial homology group in all dimensions except for 0, on which  $H_0(X)$  is isomorphic to the coefficient group  $A$ . That all positive-dimensional homology groups vanish indicates that there is no “hole” in a one-point set. The 0-dimensional homology group, which we will see soon, gives us information of the path components of a space.

**Proposition 2.1.8.** Let  $X$  be a path-connected topological space, then  $H_0(X; A) = A$ .

*Proof.* Since  $\partial_0 = 0$ ,  $Z_0(X; A) = C_0$ . Define  $\epsilon : C_0(X; A) \rightarrow A$ ,  $\sum a_i x_i \mapsto \sum a_i$ . Clearly  $\epsilon$  is a surjective group homomorphism. We claim that  $B_0(X; A) = \ker \epsilon = \{c = \sum a_x \cdot x \in C_0(X; A) \mid \sum a_x = 0\}$ . That is, the 0-dimensional boundary chain group consists of chains whose coefficients add up to 0.

To see this, let  $\gamma \in C_1(X; A)$  be a one-dimensional singular simplex, which can be identified with a path  $\gamma : [0, 1] \rightarrow X$ . Then  $\partial \gamma = \gamma(1) - \gamma(0)$ . Thus  $\partial \gamma \in B'$  and hence  $B_0(X; A) \subseteq \ker \epsilon$ . Conversely, let  $c = \sum_{i=0}^k a_i \cdot x_i \in \ker \epsilon$  be an element in  $C_0$  such that  $\sum_{i=0}^k a_i = 0$ . Fix a point  $y \in X$ . Then  $\sum_{i=0}^k a_i x_i = \sum_{i=0}^k a_i x_i - (\sum_{i=0}^k a_i)y = \sum_{i=0}^k a_i(x_i - y)$ . By the path-connectedness of  $X$ , there is a path (a 1-singular simplex)  $\gamma_i : [0, 1] \rightarrow X$  such that  $\gamma_i(0) = x_i$ ,  $\gamma_i(1) = y$  for  $i = 0, 1, \dots, k$ . Then easy to see that  $\partial_1(\sum_{i=0}^k a_i \gamma_i) = \sum_{i=0}^k a_i(x_i - y) = c$ . Thus  $c \in B_0(X; A)$  and  $\ker \epsilon \subseteq B_0(X; A)$ . We have the claim.

By the claim,  $\ker \epsilon = B_0(X; A)$ . By the fundamental theorem of group homomorphism,  $C_0(X; A)/\ker \epsilon = Z_0(X; A)/B_0(X; A) = H_0(X; A) = A$ .  $\square$

**Corollary 2.1.9.** If  $X = \bigsqcup_{\lambda \in \Lambda} X_\lambda$ , where the “ $\bigsqcup$ ” symbol denotes disjoint union and each  $X_\lambda$  is a path component of  $X$ , then  $H_0(X; A) = \bigoplus_{\lambda \in \Lambda} A$ .

*Proof.* Easy to see that  $C_n(X; A) = \bigsqcup_{\lambda \in \Lambda} C_n(X_\lambda; A)$ . By definition of  $H_n$ ,  $H_n(X; A) = \bigoplus_{\lambda \in \Lambda} H_n(X_\lambda; A)$ . By 2.1.8, we have the result.  $\square$

By 2.1.9, the rank of  $H_0(X)$  gives us the cardinality of path components of  $X$ .

Now we want to know more about the relations of maps between spaces and their homology groups.

Let  $f : X \rightarrow Y$  be a continuous map, then  $f$  induces a map  $C_n(X) \rightarrow C_n(Y)$  by mapping  $\sigma$  to  $f \circ \sigma$ . Denote this map still by  $f$ . One can easily check that  $f(Z_n(X)) \subseteq Z_n(Y)$  and  $f(B_n(X)) \subseteq B_n(Y)$ . Therefore,  $f$  induces a well-defined homomorphism  $f_* : H_n(X) \rightarrow H_n(Y)$ . It is also not hard to check that  $(f \circ g)_* = f_* \circ g_*$  for every  $g : Y \rightarrow Z$ .

For example, if  $f$  is a homeomorphism, then  $f_*$  is an isomorphism, since  $(f_*)^{-1} = (f^{-1})_*$ .

Intuitively, the  $\mathbb{R}^2$  plane and a point  $\{\ast\}$  both contain one path component and no non-trivial higher dimension “holes”. They are of course not homeomorphic, since there is no bijection between them; but we expect they have the same homology groups. To describe this phenomenon in general, we need an equivalence relation among topological spaces weaker than homeomorphic, called homotopic.

**Definition 2.1.10** (Homotopy). (1) Let  $X, Y$  be topological spaces,  $f, g : X \rightarrow Y$  be continuous maps between them. A **homotopy between  $f$  and  $g$**  is a continuous map  $F : [0, 1] \times X \rightarrow Y$  such that  $F(0, x) = f(x), F(1, x) = g(x)$ .

If there is a homotopy between  $f, g$ , then  $f$  and  $g$  are called **homotopic**, denoted  $f \simeq g : X \rightarrow Y$ .

(2) Let  $f : X \rightarrow Y, h : Y \rightarrow X$ .  $h$  is called a **homotopic inverse** of  $f$  if  $f \circ h \simeq id_Y$  and  $h \circ f \simeq id_X$ .

If there is a pair of homotopic inverses between  $X$  and  $Y$ , then  $X$  and  $Y$  are called homotopic (or homotopic equivalent).

A homotopy can be thought of as “a continuous transformation between maps”. Two spaces being homotopic means that they can be continuously deformed to each other. We give some concrete examples.

**Example 2.1.11.** (1) If  $X$  is homeomorphic to  $Y$ , then  $X$  and  $Y$  are homotopic, since there are maps  $f : X \rightarrow Y$  and  $g : Y \rightarrow X$  such that  $f \circ g = id_Y$  and  $g \circ f = id_X$ .

(2)  $\mathbb{R}^n$  is homotopic to a point. To see this, let  $i : \{0\} \hookrightarrow \mathbb{R}^n$  be the inclusion from the origin to  $\mathbb{R}^n$ . Let  $r : \mathbb{R}^n \rightarrow \{0\}$  be the only map from  $\mathbb{R}^n$  to  $\{0\}$ . Then  $r \circ i = id$  is the identity map on  $\{0\}$ . Let  $F : [0, 1] \times \mathbb{R}^n \rightarrow \{0\}, F(t, x) = (1 - t)x$ . Then  $F$  is continuous and  $F(0, x) = x, F(1, x) = 0$ . Thus  $F$  is a homotopy between  $id_{\mathbb{R}^n}$  and  $i \circ r$ , which maps everything to  $\{0\}$ .  $i$  and  $r$  are homotopic inverses to each other.

Similarly,  $\mathbb{R}^n$  is homotopic to the disc  $D_n := \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid \sum x_i^2 \leq 1\}$ . Note that none of these spaces are homeomorphic.

We admit without proving the following theorem, which tells us that homotopic spaces do have same homology groups.

**Theorem 2.1.12** (Homotopy Invariance of Homology). *Let  $f \simeq g : X \rightarrow Y$  be homotopic maps from  $X$  to  $Y$ , then  $f_* = g_* : H_n(X) \rightarrow H_n(Y)$  for all  $n$ .*

**Corollary 2.1.13.** *If  $X$  and  $Y$  are homotopic, then  $H_n(X) = H_n(Y)$  for all  $n$ .*

*Proof.* Let  $f : X \rightarrow Y, g : Y \rightarrow X$  be homotopic inverses to each other. Then by 2.1.12,  $(f \circ g)_* = (id)_* = id = f_* \circ g_*$ . Thus  $f_*$  and  $g_*$  are group isomorphisms.  $\square$

## 2.2 Simplicial Homology

The singular homology is defined for all topological spaces, so we can compute the singular homology group for all kind of spaces, no matter how strange they are. However, the computation is difficult even for simple geometric objects such as  $S^n$  and  $\mathbb{RP}^n$ . This is not surprising given its tedious and abstract set-up. In this section, we introduce another kind of homology theory, called simplicial homology, whose computation is much easier but can only be defined for fewer spaces. This is actually the kind of homology that we will use in the next chapter. We shall see that the two homology theories are in fact identical.

**Definition 2.2.1.** Let  $\Delta_n$  be the standard  $n$ -simplex. The *i-th face* of  $\Delta_n$  is an  $n - 1$  simplex  $[v_0, \dots, \hat{v}_i, \dots, v_n]$  spanned by  $n - 1$  corners in  $\Delta_n$  with the original order. There is a canonical homeomorphism  $d_n^i$  between  $\Delta_{n-1}$  and  $[v_0, \dots, \hat{v}_i, \dots, v_n]$  that preserves the order of vertices, called the *i-th face map*.

**Remark.** As in 2.1.4, the boundary chain of  $\Delta_n$  consists of all faces of  $\Delta_n$  with alternative signs.

The simplicial homology can only be computed for spaces that are “good enough”, that is, spaces with a nice structure on it, called  $\Delta$ -complex.

**Definition 2.2.2 ( $\Delta$ -complex).** Let  $X$  be a topological space. A  $\Delta$ -complex structure on  $X$  is a collection of continuous maps  $f_\alpha^n : \Delta_n \rightarrow X$ , where  $n$  depending on  $\alpha$ , called an *n-cell*, such that the following conditions hold:

(1) For every  $\alpha, n$ ,  $f_\alpha^n|_{\overset{\circ}{\Delta}_n} : \overset{\circ}{\Delta}_n \rightarrow X$  is injective, and  $X = \bigsqcup_{\alpha, n} f_\alpha^n(\overset{\circ}{\Delta}_n)$ . Here  $\overset{\circ}{\Delta}_n := \Delta_n - \partial\Delta_n$  denotes the interior of  $\Delta_n$ .

(2) For every  $\alpha, n, i$ ,  $f_\alpha^n \circ d_n^i = f_\beta^{n-1}$  for some  $\beta$ . In other words, the restriction of each  $f_\alpha^n$  to a face is one of the maps  $f_\beta^{n-1}$  in the  $\Delta$ -complex structure.

(3) A set  $A \subseteq X$  is open if and only if  $(f_\alpha^n)^{-1}(A)$  is open in  $\Delta_n$  for each  $f_\alpha^n$ .

A space that admits a  $\Delta$ -complex structure is called **triangulable**. A space with a  $\Delta$ -complex structure is called a  **$\Delta$ -complex**.

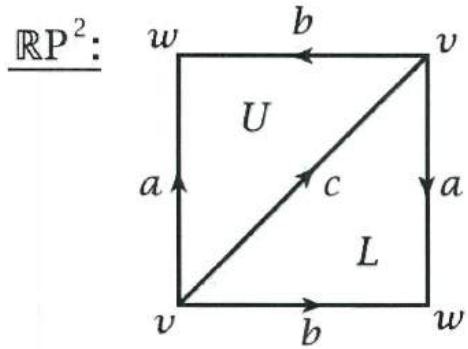
By (1) (2), one can construct a  $\Delta$ -complex step by step. Start with a discrete set  $\{f_\alpha^0(\Delta_0)\} := X_0$ , then attach  $\Delta_1$  along its faces via the map  $f_\alpha^1 \circ d_1^i (= f_\beta^0 \text{ for some } \beta)$  and get the space  $X_1$ , and similarly attach  $\Delta_2$  and so on. Inductively,  $X_i = (X_{i-1} \sqcup \{f_\alpha^i(\Delta_i)\}) / \sim$  is the quotient space where the equivalence relation identifies points mapped to the same point in  $X$ . We have  $X = \bigcup X_i$ . By (3), the two spaces have the same topology.  $X_i$  is called the *i-skeleton* of  $X$ .

We also use the name “*n-cell*” to regard the image  $f_\alpha^n(\Delta_n)$  of an *n-cell*  $f_\alpha^n$ . Then the *i-skeleton*  $X_i$  is the union of all cells of  $X$  with dimension less than or equal to  $i$ .

**Definition 2.2.3.** Let  $X$  be a  $\Delta$ -complex. A **subcomplex**  $Y$  of  $X$  is a closed subspace  $Y$  of  $X$  that is a union of cells of  $X$ . Easy to verify that a subcomplex is itself a  $\Delta$ -complex.

**Example 2.2.4.** (1)  $S^n$  can be thought of as  $(\Delta_n \sqcup \Delta_n) / \sim$ , where the two standard *n*-simplexes are attached along their boundaries via the identity map.

(2) The projective plane  $\mathbb{RP}^2$  can also be constructed by attaching two 2-simplices along their boundaries, as in figure 2.3.



**Figure 2.3:** The  $\Delta$ -complex structure of the projective plane. Arrows with same letters are identified in the way indicated by the arrows. Cf.<sup>[4]</sup>.

**Definition 2.2.5** (Simplicial Homology). Let  $X$  be a  $\Delta$ -complex and  $A$  an abelian group. Define  $C_n^\Delta(X; A) := \{ \sum_{\text{finite}} a_f f \mid f \text{ equals to some } f_\alpha^n : \Delta_n \rightarrow X \text{ in the } \Delta\text{-complex structure of } X, a_f \in A \}$ . In particular, if  $A = \mathbb{Z}$ , then  $C_n^\Delta(X; A)$  is the free abelian group generated by all  $n$ -cells in the  $\Delta$ -complex structure of  $X$ .

There is a boundary map  $\partial_n : C_n^\Delta(X; A) \rightarrow C_{n-1}^\Delta(X; A)$  defined by  $\partial_n(f_\alpha^n) := f_\alpha^n|_{\partial\Delta_n}$  and extend linearly. By 2.2.2 (2), the boundary map is well-defined. One can check that  $\partial_n \circ \partial_{n-1} = 0$ , so

$$\cdots \rightarrow C_{n+1}^\Delta \xrightarrow{\partial_{n+1}} C_n^\Delta \xrightarrow{\partial_n} C_{n-1}^\Delta \xrightarrow{\partial_{n-1}} \cdots$$

is a chain complex (meaning that  $\partial^2 = 0$ ).

We can define the notion of **cycle**, **boundary chain** and **simplicial homology group** as in singular homology. I.e.,  $Z_n^\Delta(X; A) := \ker \partial_n$ ,  $B_n^\Delta(X; A) := \text{im } \partial_{n+1}$  and  $H_n^\Delta(X; A) := Z_n^\Delta(X; A)/B_n^\Delta(X; A)$ .

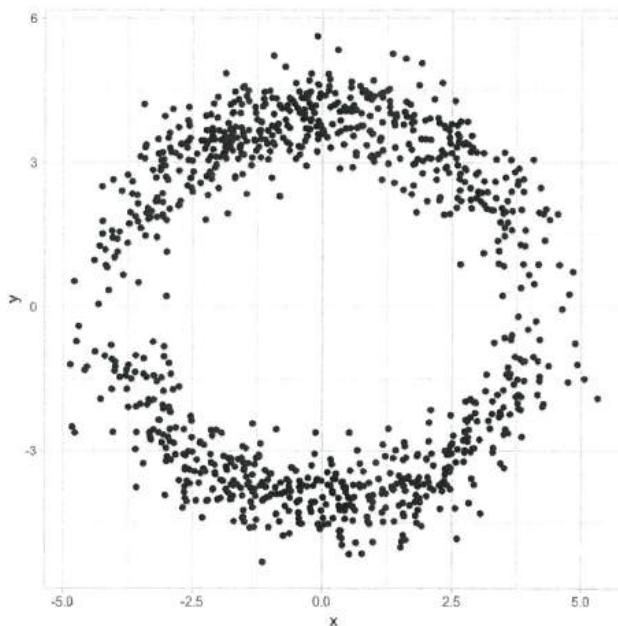
We omit the proof of the following important theorem, saying that computing the singular and simplicial homology will yield the same result.

**Theorem 2.2.6.** *Let  $X$  be a  $\Delta$ -complex, then  $H_n(X; A) \cong H_n^\Delta(X; A)$  for all  $n \geq 0$ .*

We have seen the definition, examples and basic properties of homology. Homology groups give us the information about the homotopy class of a space. In the next chapter, we will see how to analyze the shape of data using homology.

## Chapter 3 Persistence Homology

If we take 100 arbitrary points from a circle, we can recognize the shape of the circle by our eyes, as in figure 3.1. If the points are taken from a torus in  $\mathbb{R}^3$ , we might still be able to recover the shape of the torus in our mind. What if the points are taken from a Klein bottle or a projective plane embedded in  $\mathbb{R}^4$ , or  $S^n$  that lies in  $\mathbb{R}^{n+1}$ ? In many cases we have similar problems, having to recover the original space (for example, an attractor of a dynamical system) from which a discrete point cloud is taken out. One may think of calculating the homology to find its homotopy type, but the topology on a finite set is discrete, giving almost trivial homology. A strategy is to connect close points, fill in the blank spaces near each point to get a non-discrete space and then calculate the homology of that space. This might work, but the new question is the definition of “close”. If the points come from a large circle, we may have to connect points whenever their distance is less than 100. But if they come from a small torus with diameter 1, the parameter 100 will connect every two points and simply yield a contractible space. Similarly, if we let the threshold of closeness to be 1 for the large circle, we would get nothing new but the discrete set again. Different parameters change the homotopy type and homology groups, while it turns out that there is no systematic method to choose the parameter. The solution of this problem is to try all parameters from small to large and to calculate the homology groups for every parameter. After that, compare all the homology groups and pick the homology classes that persist longer than others, which are more possible to be the homology class of the original space, and regard them as the homology classes for the real space. This is what persistent homology does. In this section, we will talk about persistent homology in detail. The main reference of this chapter is<sup>[6]</sup>.



**Figure 3.1:** A statistical circle

### 3.1 Complex Constructions

The method to get a non-discrete space from a point cloud is to construct a complex by joining certain points together. Here we introduce two kind of complexes: Vietoris-Rips complex and Čech complex.

**Definition 3.1.1** (Čech Complex). Let  $X = \{x_1, \dots, x_n\}$  be a point cloud in a metric space  $(M, d)$ . The **Čech complex at scale  $\alpha$**  is the set  $\check{C}_\alpha(X) := \{\sigma \subseteq X \mid \bigcap_{x \in \sigma} B(x, \alpha) \neq \emptyset\}$ , where  $B(x, \alpha)$  is the ball centered at  $x$  with radius  $\alpha$ .

$\check{C}_\alpha(X)$  can be thought of as a  $\Delta$ -complex as follows: let  $|\sigma|$  be the convex hull spanned by all points of  $\sigma = \{x_{k_1}, \dots, x_{k_m}\}$  with  $k_1 < \dots < k_m$ , that is,  $|\sigma| := \{\sum_{i=1}^m a_i x_{k_i} \mid a_i \geq 0, \sum a_i = 1\}$ . Then identify  $\sigma$  with the only linear map  $\Delta_m \rightarrow |\sigma|$  that preserves the order of the vertices.

$|\check{C}_\alpha(X)| := \bigcup_{\sigma \in \check{C}_\alpha} |\sigma| \subseteq \mathbb{R}^d$  is called the **geometric realization** of  $\check{C}_\alpha(X)$ , whose  $\Delta$ -complex structure is given by maps in  $\check{C}_\alpha(X)$ . In this case, we call  $\check{C}_\alpha(X)$  an **abstract  $\Delta$ -complex**.

When no ambiguity is caused, we identify an abstract  $\Delta$ -complex and its geometric realization.

**Remark.** Once regarding  $\check{C}_\alpha(X)$  as a  $\Delta$ -complex, we can talk about its  $n$ -cycles,  $n$ -boundary chains and homology groups. It is clear that  $\check{C}_\alpha(X) \subseteq \check{C}_{\alpha'}(X)$  when  $\alpha < \alpha'$ . So there is a natural inclusion  $C_n^\Delta(\check{C}_\alpha(X)) \hookrightarrow C_n^\Delta(\check{C}_{\alpha'}(X))$  for every  $\alpha < \alpha'$  and  $n \geq 0$ .

**Definition 3.1.2** (Vietoris-Rips Complex). Let  $X$  be the same as in 3.1.1. The **Vietoris-Rips complex at scale  $\alpha$**  (or simply Rips complex) is the set  $\mathcal{V}_\alpha(X) := \{\sigma \subseteq X \mid B(x_i, \alpha) \cap B(x_j, \alpha) \neq \emptyset, \forall x_i, x_j \in \sigma\}$ .

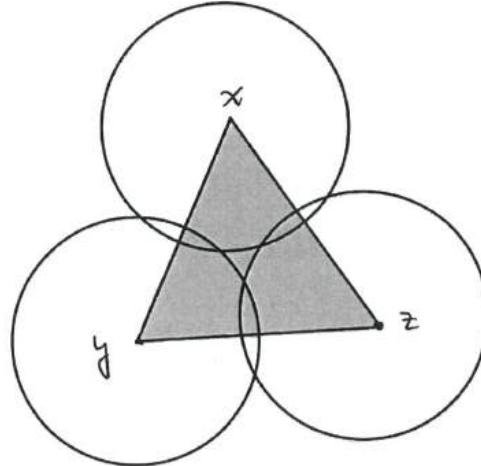
$\mathcal{V}_\alpha(X)$  can be thought of as a  $\Delta$ -complex in the same way as  $\check{C}_\alpha(X)$ . The homology groups can still be computed and the natural inclusion for every  $\alpha < \alpha'$  still exists.

**Remark.** The Rips Complex is a **flag complex**, or a **full complex**: That is, it is the maximal element in all  $\Delta$ -complexes with the given 1-skeleton. This is easy to check by the definition. Therefore, the 0-skeleton (vertices) and 1-skeleton completely determine a Rips complex, making the Rips complex less expensive than other complexes in computation.

For both Čech and Rips complex, if the parameter  $\alpha$  is small enough,  $\check{C}_\alpha(X)$  and  $\mathcal{V}_\alpha(X)$  contain only discrete points; if  $\alpha$  is larger than the diameter of  $X$ , then both  $\check{C}_\alpha(X)$  and  $\mathcal{V}_\alpha(X)$  contain the full complex spanned by points as vertices in  $X$  (that is, all subsets of  $X$  are simplices in  $\check{C}_\alpha(X)$  and  $\mathcal{V}_\alpha(X)$ ). For certain medium  $\alpha$ ,  $\check{C}_\alpha(X)$  and  $\mathcal{V}_\alpha(X)$  may very similar to the the space where  $X$  is taken from, topologically.

It is clear that for the same scale  $\alpha$ ,  $\check{C}_\alpha(X) \subseteq \mathcal{V}_\alpha(X)$ . Fig 3.2 illustrates a situation when  $\check{C}_\alpha(X)$  contains strictly less simplices than  $\mathcal{V}_\alpha(X)$ . Conversely, the Čech complex also contains the Rips complex with smaller radius, as illustrated in proposition 3.1.3.

**Proposition 3.1.3.** For any point cloud  $X$  and  $\alpha > 0$ ,  $\mathcal{V}_{\alpha/2}(X) \subseteq \check{C}_\alpha(X) \subseteq \mathcal{V}_\alpha(X)$ .



**Figure 3.2:** Complex construction for the point cloud  $X = \{x, y, z\}$ . Each circle has radius 1. Then  $\check{C}_1(X) = \{\{x\}, \{y\}, \{z\}, \{xy\}, \{xz\}, \{yz\}\}$  and  $\mathcal{V}_1(X) = \{\{x\}, \{y\}, \{z\}, \{xy\}, \{xz\}, \{yz\}, \{xyz\}\}$ . The shaded triangle is not contained in  $\check{C}_1(X)$  but is in  $\mathcal{V}_1(X)$ .

*Proof.* The second inclusion is obvious. For the first inclusion, suppose  $\sigma = \{x_1, \dots, x_m\} \in \mathcal{V}_{\alpha/2}(X)$ , then  $B(x_i, \alpha) \cap B(x_j, \alpha) \neq \emptyset, \forall x_i, x_j \in \sigma$ . That is,  $d(x_i, x_j) < \alpha, \forall x_i, x_j \in \sigma$ . Then the diameter of the set  $\sigma$  is less than  $\alpha$ , so we can find a ball  $B(y, \alpha)$  with radius  $\alpha$ , centered at some  $y \in \mathbb{R}^d$ , such that  $\sigma \subseteq B(y, \alpha)$ . Then  $y$  is in the intersection of all  $B(x_i, \alpha), x_i \in \sigma$ . Thus  $\sigma \in \mathcal{V}_\alpha(X)$ .  $\square$

This proposition allows us to compare the homology computed using the two kind of complexes.

### 3.2 Persistence

In this section we will deal with the varying scale, and extract important homology classes from each  $H_n^\Delta(\check{C}_\alpha(X))$  and  $H_n^\Delta(\mathcal{V}_\alpha(X))$ .

**Definition 3.2.1.** Let  $K$  be an abstract  $\Delta$ -complex with geometric realization  $|K|$ . A **filtration** of  $K$  is a sequence of subcomplexes  $= K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$ .

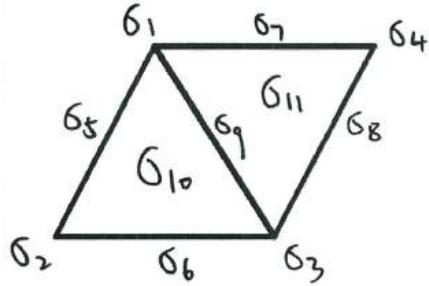
A function  $f : K \rightarrow \mathbb{R}$  is called **monotonic** if  $f(\sigma) \leq f(\tau)$  whenever  $\sigma$  is a face of  $\tau$ .

**Example 3.2.2.** (1) A filtration can be generated using a monotonic function. Let  $a_1 < \dots < a_n$  be the distinct function values of  $f$  and  $K_i := f^{-1}(-\infty, a_i]$ . By monotonicity of  $f$ , each  $K_i$  is a subcomplex of  $K$  and  $K_1 \subseteq K_2 \subseteq \dots \subseteq K_n = K$ .

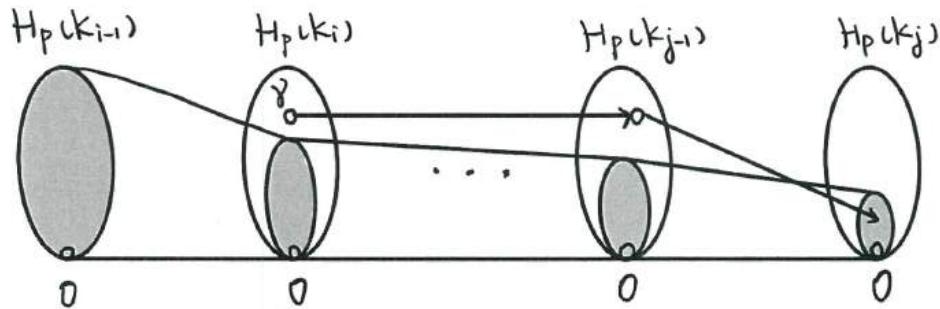
In particular, one may have  $K_i - K_{i-1} = \{\sigma_i\}$  for some  $p$ -simplex  $\sigma_i$ . In this case, the filtration illustrates the construction of  $K$  by adding one simplex at a time. See fig 3.3.

(2) If we take values  $0 < \alpha_1 < \dots < \alpha_n$ , then the Vietoris-Rips complex or Čech complex give a filtration  $\mathcal{V}_{\alpha_1}(X) \subseteq \dots \subseteq \mathcal{V}_{\alpha_n}$  or  $\check{C}_{\alpha_1}(X) \subseteq \dots \subseteq \check{C}_{\alpha_n}(X)$ .

Let  $= K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  be a filtration. For every  $i \leq j$ , we have  $K_i \hookrightarrow K_j$  the inclusion. This induces a map between homology groups, which we denoted by



**Figure 3.3:** A filtration with  $K_i = \{\sigma_1, \dots, \sigma_i\}$ . Here  $\sigma_1, \dots, \sigma_4$  are vertices,  $\sigma_5, \dots, \sigma_9$  are edges and  $\sigma_{10}, \sigma_{11}$  are faces.



**Figure 3.4:** A homology class  $\gamma$  born at  $K_i$  and dies entering  $K_j$ .

$f_p^{i,j} : H_p(K_i) \rightarrow H_p(K_j)$  for all dimension  $p$ . So we obtain a sequence of homology groups connected by homomorphisms:

$$\cdots \rightarrow H_p(K_{i-1}) \rightarrow H_p(K_i) \rightarrow H_p(K_{i+1}) \rightarrow \cdots$$

where  $f_p^{j,k} \circ f_p^{i,j} = f_p^{i,k}$  for every  $i \leq j \leq k$ . We can actually define this structure more generally, although this formal definition is not so important in this article:

**Definition 3.2.3.** A **persistent vector space** is a set of vector spaces  $\{V_\alpha\}_{\alpha \in P}$  indexed by a total ordered set  $P$  together with linear maps  $f_{\alpha,\alpha'} : V_\alpha \rightarrow V_{\alpha'}$  for every  $\alpha < \alpha'$  such that  $f_{\alpha,\alpha} = \text{id}_{V_\alpha}$ ,  $\forall \alpha$  and  $f_{\alpha',\alpha''} \circ f_{\alpha,\alpha'} = f_{\alpha,\alpha''}$  for every  $\alpha \leq \alpha' \leq \alpha''$ .

Easy to see that a filtration induces a persistent vector space indexed by a finite set  $\{0, 1, \dots, n\}$ .

**Definition 3.2.4.** The  $p$ -th persistent homology group of the filtration  $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$  (with respect to  $i, j$  is  $H_p^{i,j} := \text{im}(f_p^{i,j}) \subseteq H_p(K_j)$ ). The corresponding  $p$ -th persistence betti number is  $\beta_p^{i,j} := \text{rank}(H_p^{i,j})$ .

We say a homology class  $\gamma \in H_p(K_i)$  **born at  $K_i$**  if  $\gamma \notin H_p^{i-1,i}$ ; say  $\gamma$  **dies entering  $K_j$**  if  $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$  but  $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$ . Define the **persistence** of  $\gamma$  to be  $\text{pers}(\gamma) := j - i$ . If  $\gamma$  is born at  $K_i$  and never dies, set  $\text{pers}(\gamma) = \infty$ . See figure 3.4

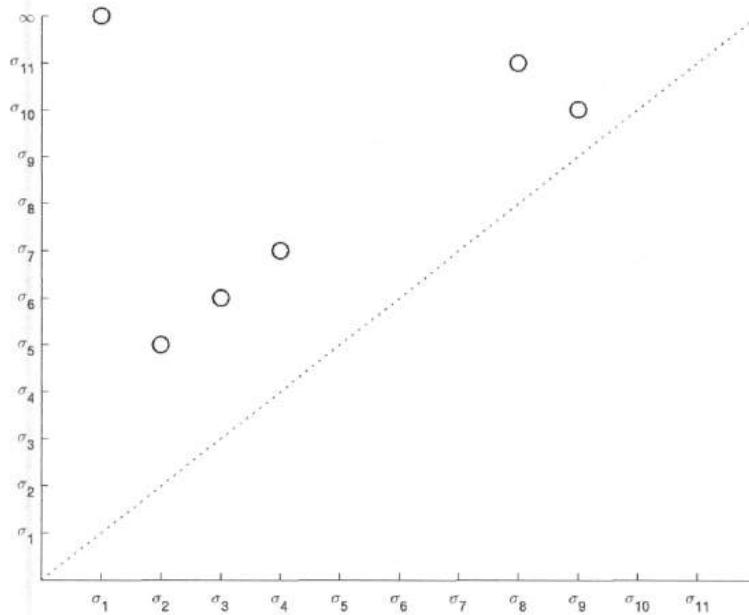
From  $K_i$  to  $K_j$ , new cycles may be born (for example, when the simplex  $\sigma_8$  is added in figure 3.3, creating new 1-cycle). Existing cycles may vanish, or become homologous

to older cycles (for example,  $\sigma_9, \sigma_{11}$  kill 1-cycles in figure 3.3). The persistent homology group  $H_p^{i,j}$  contains the homology classes in  $K_i$  that are still alive independently in  $K_j$ . The betti number  $\beta_p^{i,j}$  indicates the number of such classes. The persistent of a homology class  $\gamma$  illustrates how long does this class exist. We should of course take the classes who persistent longer into consideration. To visualize the persistence, we use the following diagram:

**Definition 3.2.5.** The  $p$ -th **persistent diagram** of the filtration  $\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  is the multiset of points in the extended real plane  $\bar{\mathbb{R}}^2 := (\mathbb{R} \cup \{\pm\infty\})^2$  such that  $(i, j)$  appears  $\mu_p^{i,j}$  times, where  $\mu_p^{i,j}$  is the number of  $p$ -dimensional homology classes born at  $K_i$  and dying entering  $K_j$ .

If the filtration is given by a monotonic function  $f : K \rightarrow \mathbb{R}$ , then we denote its persistent diagram by  $\text{Dgm}_p(f)$ .

**Example 3.2.6.** In figure 3.3, additions of  $\sigma_1, \sigma_2, \sigma_3$  and  $\sigma_4$  create new 0-cycles, that is, new connected components.  $\sigma_5, \sigma_6$  and  $\sigma_8$  kill the components created by  $\sigma_2, \sigma_3$  and  $\sigma_4$  respectively, making them homologous to the 0-class  $[\sigma_1]$ .  $\sigma_7$  creates a new 1-cycle  $(\sigma_5 \sigma_6 \sigma_8 \sigma_7)$ .  $\sigma_9$  further creates another 1-cycle, making the second hole.  $\sigma_{10}$  kills the class generated by  $\sigma_9$ , making it homologous to the 1-cycle from  $\sigma_8$ . Finally,  $\sigma_{11}$  kills the remaining 1-cycle from  $\sigma_8$ . The persistent diagram of this filtration is shown in fig



**Figure 3.5:** The persistence diagram in example 3.2.6.

The next lemma tells us that the persistent diagram contains all the information about the persistence homology groups, and vice versa:

**Theorem 3.2.7** (Fundamental Lemma of Persistent Homology). *Let  $\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  be a filtration. Then for every  $p \geq 0$ ,*

- (1)  $\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$
- (2)  $\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$ .

*Proof.* (1) Indeed, the difference in the first bracket is the number of independent classes born at or before  $K_i$  and dying entering  $K_j$ ; the difference in the second bracket is the number of independent classes born at or before  $K_{i-1}$  and dying entering  $K_j$ . By definition, their difference is  $\mu_p^{i,j}$ , which is the number of independent classes that are born at  $K_i$  and dies entering  $K_j$ .

(2)  $\beta_p^{k,l}$  is the number of independent classes in  $K_i$  that are still alive in  $K_j$ , which is the number of independent classes that are born before  $K_i$  and dies after  $K_j$ . So  $\beta_p^{k,l}$  is the number of points (counted with multiplicity) in the left upper quadrant with corner point  $(k, l)$ . This is the summation in the right hand side.  $\square$

Now we proceed to the calculation of the persistent diagram. During the specific calculation, we usually consider the homology with coefficient  $\mathbb{Z}_2$ , which is indeed a field. The homology groups are therefore vector spaces over  $\mathbb{Z}_2$ . By choosing this coefficient group, we can ignore the sign (orientation) of the chains and spaces. Moreover, since  $\mathbb{Z}_2$  is the simplest nontrivial group, the calculation is also easier.

Let  $\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  be a filtration. We assume  $K_i - K_{i-1} = \{\sigma_i\}$ . Since each  $K_i$  is a simplicial complex,  $i < j$  whenever  $\sigma_i$  is a face of  $\sigma_j$ . Then we store all the simplices and its boundary in a matrix  $\partial$ , called the **boundary matrix** of  $K$  where  $\partial(i, j) = 1$  if  $\sigma_i$  is a codimension-1 face of  $\sigma_j$ , and  $\partial(i, j) = 0$  otherwise. So  $\partial$  is a  $n \times n$  strictly upper-triangular 0 – 1 matrix, whose columns store the boundary chains of the simplices in  $K$ .

The information of the persistent diagram is hidden in the boudary matrix. To extract it, we need to reduce this matrix.

**Definition 3.2.8.** Let  $R$  be an  $n \times n$  0 – 1 matrix. For  $1 \leq j \leq n$ , define  $low(j)$  to be the row index of the lowest 1 in column  $j$ . If the column  $j$  is zero, set  $low(j)$  undefined. We say  $R$  is **reduced** if  $low(j) \neq low(j')$  whenever  $j \neq j'$  and  $low(j), low(j')$  are defined.

Here is the pseudocode of an algorithm reducing  $\partial$ :

---

#### Algorithm 1 An algorithm reducing $\partial$

---

```

 $R = \partial$ 
for  $j=1$  to  $n$  do do
    while  $\exists j_0 < j$  with  $low(j_0) = low(j)$  do
        add column  $j_0$  to column  $j$ 
    end while
end for
return  $R$ 

```

---

The output  $R$  is a reduced matrix.

**Example 3.2.9.** Let  $\emptyset = K_0 \subseteq \dots \subseteq K_n = K$  be the same filtration as in example 3.2.6 and figure 3.3. Then the boundary matrix in coefficient  $\mathbb{Z}_2$  and the reduced  $R$  by the algorithm

are

$$\partial = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 \\ & & 1 & 1 & 1 & 1 & 1 \\ & & & 1 & 1 & 1 & 1 \\ & & & & 1 & 1 & 1 \\ & & & & & 1 & 1 \\ & & & & & & 1 \\ & & & & & & & 1 \\ & & & & & & & & 1 \\ & & & & & & & & & 1 \\ & & & & & & & & & & 1 \\ & & & & & & & & & & & 1 \\ & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & & & & 0 \\ & & & & & & & & & & & & & & & & & & 0 \\ & & & & & & & & & & & & & & & & & & 0 \\ & & & & & & & & & & & & & & & & & & 0 \end{bmatrix} R = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ & 1 & 1 & 1 & 1 & 1 & 0 \\ & & 1 & 1 & 1 & 1 & 1 \\ & & & 1 & 1 & 1 & 1 \\ & & & & 1 & 1 & 1 \\ & & & & & 1 & 1 \\ & & & & & & 1 \\ & & & & & & & 1 \\ & & & & & & & & 1 \\ & & & & & & & & & 1 \\ & & & & & & & & & & 1 \\ & & & & & & & & & & & 1 \\ & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & & 1 \\ & & & & & & & & & & & & & & & & & 0 \\ & & & & & & & & & & & & & & & & & 0 \\ & & & & & & & & & & & & & & & & & 0 \end{bmatrix}$$

Note that since  $\partial$  is upper-triangular and we are only allowing additions from left to right,  $R$  is also upper-triangular. By example 3.2.9 and 3.2.6, the pairs  $(j, \text{low}(j))$  in the reduced matrix  $R$  are precisely the finite points in the persistent diagram. The following theorem states that this is not a coincidence.

**Theorem 3.2.10.** *Let  $\emptyset = K_0 \subseteq \dots \subseteq K_n = K$  be a filtration such that  $K_i - K_{i-1} = \{\sigma_i\}$  for some simplex  $\sigma_i$ . Let  $\partial$  be its boundary matrix and  $R$  be the reduced matrix. Denote its persistent diagram by  $D$ , then*

- (1) *If the  $j$ -th column of  $R$  is 0, then  $\beta_p(K_j) = \beta_p(K_{j-1})$ . The addition of the  $p$ -simplex  $\sigma_j$  creates a new cycle. In this case, we call  $\sigma_j$  **positive**;*
- (2) *If the  $j$ -th column of  $R$  is nonzero, let  $i = \text{low}(j)$ . Then the addition of the  $p$ -simplex  $\sigma_j$  kills a  $(p-1)$ -dimensional homology class  $\gamma$  created by the addition of  $\sigma_i$ . In particular,  $\gamma$  has persistence  $j-i$  and  $(i, j) \in D$ . In this case, we call  $\sigma_j$  **negative**.*
- (3) *The point  $(i, \infty) \in D$  if and only if the  $i$ -th column of  $R$  is 0 and  $\nexists j$  such that  $\text{low}(j) = i$ .*

*Proof.* (1) If the  $j$ -th column of  $R$  is 0, by the reduction algorithm, the  $j$ -th column of  $\partial$  is a linear combination of the previous  $j-1$  columns. Since the columns represent boundary chains, that is,  $\partial\sigma_j = \sum \partial\sigma_k$  where each  $k < j$ . Then  $\partial(\sigma_j - \sum \sigma_k) = 0$ ,  $\sigma_j - \sum \sigma_k$  is a cycle. This cycle is not homologous to any existing cycles because it contains  $\sigma_j$ . Thus the addition of  $\sigma_j$  creates the new cycle  $\sigma_j - \sum \sigma_k$ .

(2) If  $i = \text{low}(j)$  is defined, we claim that  $\partial\sigma_j$  is a nontrivial cycle killed by the addition of  $\sigma_j$ .  $\partial\sigma_j$  is a cycle since  $\partial(\partial\sigma_j) = 0$ . It is also obviously killed by  $\sigma_j$  since it is the boundary. To see the cycle  $\partial\sigma_j$  is created by the addition of  $\sigma_i$ , let  $c$  be the chain in the  $j$ -th column of  $R$ . By the algorithm,  $c$  is a linear combination of the chains in the first  $j$  columns of  $\partial$ , that is,  $c = \partial\sigma_j + \sum \partial\sigma_k$  where each  $k < j$ . Then  $c - \partial\sigma_j = \partial \sum \sigma_k \in B_p(K)$ . Therefore,  $c$  and  $\partial\sigma_j$  differ by a boundary, hence homologous. We denote this class by  $\gamma = [c] = [\partial\sigma] \in H_p(K)$ . Now  $c$  is clearly created by the addition of  $\sigma_i$ , since  $\sigma_i$  has the largest index among all simplices in the chain  $c$ . Thus the whole homology class  $\gamma = [\partial\sigma_j]$  is created by the addition of  $\sigma_i$ . This also implies that  $(i, j) \in D$ .

(3)  $(i, \infty) \in D$  if and only if  $\sigma_i$  creates a homology class that never dies. By (1) and (2), the statement is obvious.  $\square$

There is an alternative representation of the persistent diagram, called the **barcode**. Specifically, the barcode  $bcd$  of a persistence diagram  $D$  is a multiset of intervals  $\{[a_i, a_j] \mid (a_i, a_j) \in D\}$ . The two things contain equivalent information, namely the birth and death of

the homology classes. However, the barcode is more intuitive, since the length of the intervals indicate the persistence of homology classes. We can visualize the persistence homology by drawing the barcodes. We will see examples of barcodes in the next chapters.

Therefore, given a filtration, we can systematically reduce its boundary matrix and get the persistent diagram or the barcode. We are done with the technical details of persistence homology. In the next chapters, we will see how it combine with other techniques and apply to real-world problems.

## Chapter 4 Sliding Window Persistence and its Application

In this chapter, we will introduce how to use persistent homology to detect the shape of an attractor in a dynamical system. The tool translating an attractor to a point cloud in an Euclidean space is sliding window embedding. Then we see how it being applied to wheeze detection, which is an important medical problem. The main references of this chapter are<sup>[7]</sup> and<sup>[8]</sup>.

### 4.1 Sliding Window Persistence

We first introduce some basic notions of dynamical systems.

**Definition 4.1.1** (Dynamical system). A (global continuous time) **dynamical system** is a pair  $(M, \Phi)$  where  $M$  is a topological space and  $\Phi : \mathbb{R} \times M \rightarrow M$  is a continuous map, or a **flow**, such that  $\Phi(0, p) = p$  and  $\Phi(t + s, p) = \Phi(s, \Phi(t, p))$  for all  $p \in M, t, s \in \mathbb{R}$ .

**Remark.** Alternatively, a flow can be thought of as an action of the additive group  $\mathbb{R}$  to  $M$ . The two requirements in the definition hold by the group law of  $(\mathbb{R}, +)$ .

Dynamical systems are mathematical abstraction for time-dependent physical processes. Points in the space  $M$  (usually a manifold) flow with time, forming different orbits. For any fixed  $p$ , we have a map  $\Phi_p : \mathbb{R} \rightarrow M$ ,  $\Phi_p(t) = \Phi(t, p)$ . This map illustrates the motion of  $p$  with time. For any fixed  $t$  we also have map  $\Phi_t : M \rightarrow M$ ,  $\Phi_t(p) = \Phi(t, p)$ . The image of  $p \in M$  is the position of  $p$  after  $t$  time. Each  $\Phi_t$  is a homeomorphism with inverse  $\Phi_{-t}$ , and is a diffeomorphism in the case  $M$  being a smooth manifold and  $\Phi$  being a smooth map.

The changing of weather in Shenzhen with time, the trajectory of a ball with initial speed, the periodic moving of a pendulum... all form a dynamical system. Typical examples are the solutions of differential equations.

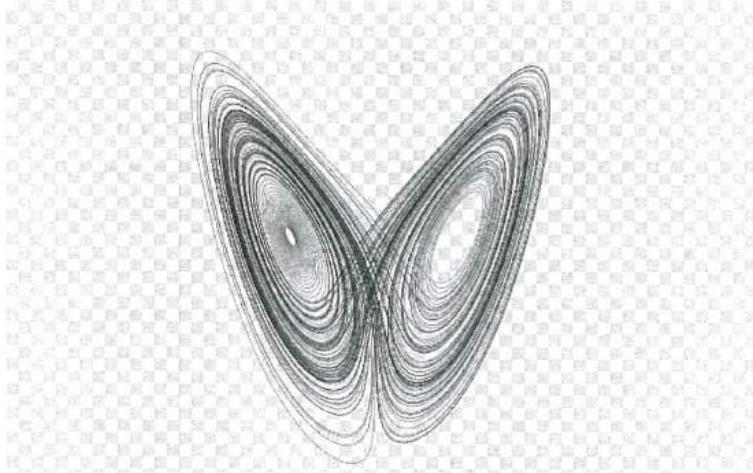
**Example 4.1.2** (Lorentz System). Let  $x, y, z$  be real-valued differentiable functions with respect to  $t$  and  $\sigma, \rho, \beta \in \mathbb{R}$  be constants. Solving the system of equations

$$\begin{aligned} x'(t) &= \sigma \cdot (y - x) \\ y'(t) &= x \cdot (\rho - z) - y \\ z'(t) &= xy - \beta z \end{aligned}$$

yields a dynamical system  $(\mathbb{R}^3, \Phi)$  where  $\Phi : \mathbb{R} \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$  is given by  $\Phi(t, (x, y, z)) = (x(t), y(t), z(t))$ .

This is indeed a chaotic system, that is, a system whose evolution is extremely sensitive to initial conditions. The rigorous definition of chaotic system is complicated and not universally consistent, so we will not state here. To describe it more precisely, one can imagine the “distance” between any two closed points grows exponentially with time. The orbits in a chaotic system is wild and not bounded by initial states.

Some points of  $M$  tends to gather to certain subsets of  $M$  as time passes. These subsets are like magnets attracting nearby points. We call them attractors.



**Figure 4.1:** the butterfly attractor in the Lorentz system

**Definition 4.1.3** (Attractor). An **attractor** of a dynamical system  $(M, \Phi)$  is a compact subset  $A$  of  $M$  such that

- (1)  $A$  is invariant under  $\Phi$ . That is, if  $a \in A$ , then  $\Phi(t, a) \in A$  for all  $t \geq 0$ .
- (2)  $A$  has an open **basin of attraction**. That is, there is an open neighborhood  $B(A)$  of  $A$  such that  $\bigcap_{t \geq 0} \{\Phi(t, p) \mid p \in B(A)\} = A$ .

Condition (1) means that all points entering an attractor  $A$  is trapped forever. By condition (2), points in the basin of attraction approach to  $A$  infinitely close as time goes by.

Attractors are of vital importance in the study of dynamical systems. In particular, the shape of attractors give us much information about the dynamical system. For example, attractors homeomorphic to  $S^1$  imply periodicity, while a tori  $\mathbb{T}^n := S^1 \times \cdots \times S^1$  implies quasiperiodicity; non-integral Hausdorff dimension (whose formal definition is not given here) of an attractor implies chaotic behavior. Figure 4.1 shows a strange attractor with non-integral Hausdorff dimension of the Lorentz system in example 4.1.2. Therefore, detecting the shape of attractors is crucial. To do so, we would like to use the method of persistence homology, with the help of another tool, called sliding window embedding.

In practice, the explicit information of each state  $p \in M$  is usually unknown. Instead, what we have is a measurement, that is, a continuous function  $F : M \rightarrow \mathbb{R}$ , called the *observation function*. Given a state  $p \in M$ , let  $\varphi_p : \mathbb{R} \rightarrow \mathbb{R}$  be the composite  $\varphi_p(t) := F \circ \Phi(t, p)$ .  $\varphi_p$  characterizes the change of the measurement of the point with time. Such a one dimensional time series data seems to loss much information of the original space (usually a higher dimensional manifold). Thanks to the *Takens's embedding theorem*, from almost all observation functions and almost all dynamical systems, we could recover our orbit in the original space.

Let  $M, N$  be two smooth manifolds and  $C^k(M, N)$  be the set of all maps from  $M$  to  $N$  with continuous  $k$ -th partial derivatives. We can endow  $C^k(M, N)$  with a topology, called the **Whitney topology**, by which two functions are “close” if all their derivatives up to order

$k$  are close on compact subsets of  $M$ . The definition of Whitney topology is not stated here, since all we need is the fact that this makes  $C^k(M, N)$  a topological space.

**Theorem 4.1.4** (Taken's Embedding Theorem). *Let  $M$  be a smooth, compact, Riemannian manifold (a manifold with a metric) and  $\tau > 0$  a real number. Let  $d \geq 2 \dim M$  be an integer. Then the delay map  $\varphi : M \rightarrow \mathbb{R}^{d+1}$ ,  $\varphi(p) = (\varphi_p(0), \varphi_p(\tau), \dots, \varphi_p(d\tau))$ , where  $\varphi_p(t) = F \circ \Phi(t, p)$  is the composition, is an embedding for generic  $\Phi \in C^2(\mathbb{R} \times M, M)$  and generic  $F \in C^2(M, \mathbb{R})$ . That is,  $M$  is diffeomorphic to its image under  $\varphi$ .*

Here “generic” implies that the set of all  $\Phi$  and  $F$  making  $\varphi$  an embedding is open and dense with respect to the Whitney topology. Although a proof of theorem 4.1.4 is not given here, it is not so surprising considering all manifolds can be embedded into high dimensional Euclidean spaces. Taken's embedding theorem gives us a particular way to do this. In particular, every orbit is also embedded into  $\mathbb{R}^{d+1}$  with  $M$ . Motivated by this, we have the following definition:

**Definition 4.1.5.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a real-valued function,  $\tau > 0$  a real number and  $d > 0$  an integer. The sliding window embedding with parameters  $\tau$  and  $d$  is the vector valued function

$$\begin{aligned} SW_{d,\tau} f : \mathbb{R} &\rightarrow \mathbb{R}^{d+1} \\ t &\mapsto (f(t), f(t + \tau), f(t + 2\tau), \dots, f(t + d\tau)) \end{aligned}$$

Here  $d+1$  is the dimension of the Euclidean space,  $\tau$  is called the **delay** and  $d\tau$  is called the **window size**.

Let  $T \subseteq \mathbb{R}$  be a subset of the real numbers, the set

$$\mathbb{SW}_{d,\tau} f := \{SW_{d,\tau} f(t) \mid t \in T\} \subseteq \mathbb{R}^{d+1}$$

is called the **sliding point cloud** associated to the sample set  $T$ .

If  $f$  is one of the composite measurement maps  $\varphi_p$  in theorem 4.1.4, then  $\mathbb{SW}_{d,\tau} f$  is embedded into  $\mathbb{R}^{d+1}$  as a subset of  $\varphi(M)$ . Given a sample set  $T\mathbb{R}$ ,  $\mathbb{SW}_{d,\tau} \varphi_p$  recovers part of the orbit of  $p \in M$  in  $\mathbb{R}^{d+1}$ . In other words,  $\mathbb{SW}_{d,\tau} \varphi_p$  gives a topological copy of  $\{\Phi(t, p) \mid t \in T\}$ .

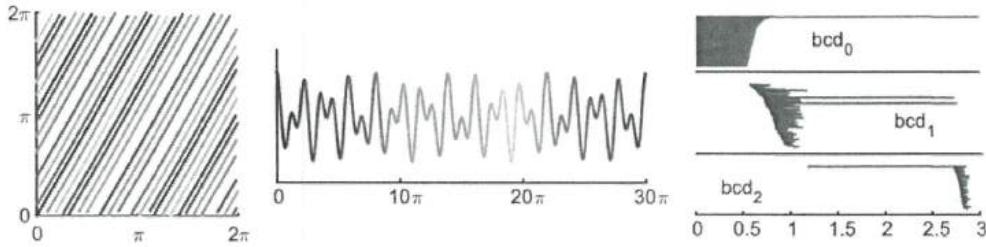
In practice, we usually use such sets to approximate attractors. Therefore, what we should do is choosing an appropriate observation function and two parameters  $d, \tau$ , using Taken's theorem to recover the orbit of a point and then using persistent homology introduced in 3 to analyze its shape. Information of dynamical systems hide in the homotopy type of the attractors. We finish this section with an example of sliding window persistence.

**Example 4.1.6.** Let  $M = \mathbb{T}^2 = S^1 \times S^1 \subseteq \mathbb{C}^2$  and  $\omega \in \mathbb{R}$ . Consider the following flow  $\Phi$  and observation function  $F$ :

$$\begin{aligned} \Phi : \mathbb{R} \times \mathbb{T}^2 &\rightarrow \mathbb{T}^2 \\ (t, (z_1, z_2)) &\mapsto (e^{it} z_1, e^{i\omega t} z_2) \\ F : \mathbb{T}^2 &\rightarrow \mathbb{R} \\ (z_1, z_2) &\mapsto \operatorname{Re}(z_1 + z_2) \end{aligned}$$

If  $\omega \in \mathbb{Q}$ , then every orbit  $\{\Phi(t, p) \mid t \in \mathbb{R}\}$  is periodic, that is,  $\Phi(t + \tau, p) = \Phi(t, p)$  for all  $t \in \mathbb{R}$  for some constant  $\tau$ . However, if  $\omega$  is irrational, then the orbit  $\{\Phi(t, p) \mid t \in \mathbb{R}\}$  of  $p$  never intersect itself and turns out to be dense in  $M$ . By results in dynamical system, the whole space  $M = \mathbb{T}^2$  is the only attractor.

Here we choose  $p = (1, 1)$  and  $\omega = \sqrt{3}$ . Then the time series data  $f(t) = F \circ \Phi(t, p) = \operatorname{Re}(e^{it} + e^{i\omega t}) = \cos t + \cos(\sqrt{3}t)$ . Figure 4.2 illustrates the orbits, the corresponding time series data and the barcode of the embedded point cloud. One can see that there are one significant 0-dimensional cycles, two significant 1-dimensional cycles and one significant 2-dimensional cycle. This coincides with the homology of a torus.



**Figure 4.2:** Left: The orbit of a point on a torus  $S^1 \times S^1$ , colored from blue to red. Center: The corresponding time series data. Right: The barcodes of the sliding window embedding. Cf<sup>[7]</sup>.

## 4.2 Application on Wheeze Detection

Wheezes are abnormal lung sounds, usually implying obstructive airway diseases. Therefore, detecting wheeze signals is essential in diagnosing lung diseases. People have worked on this problem in different ways, see reference . In this section, we mainly follow reference , introducing a topological method to detect the wheeze signal. The idea is to use the sliding window time delay embedding to embed the wheeze signal to  $\mathbb{R}^2$  and compute its barcodes, then the first persistent Betti number distinguishes between wheeze and non-wheeze signals.

To get started, we model the wheeze signal. The frequency of a monophonic wheeze signal is approximately piecewisely constant with respect to time. Therefore, we give a model for wheeze signals as a continuous piecewise sinusoidal function with different periods and phases and time-varying amplitude. That is,

$$w(t) = \sum_{i=1}^n g_i(t)$$

where

$$g_i(t) = \begin{cases} w_i(t) & t_{i-1} \leq t < t_i, \\ 0 & \text{otherwise.} \end{cases}$$

and  $w_i$  for  $i = 1, 2, \dots, n$ , with  $t_0 = a$  and  $t_n = b$ , are defined as

$$w_i(t) = A(t) \sin\left(\frac{2\pi}{T_i}t + \phi_i\right)$$

where  $A(t)$  is a nonzero continuous amplitude function. To make  $w(t)$  continuous, the phases should meet the requirement  $\phi_i = \phi_{i-1} + 2\pi t_{i-1}(\frac{1}{T_{i-1}} - \frac{1}{T_i})$ . The fact is that this model approximate wheeze signals very well, by a measurement called *Hausdorff distance*. However, for non-wheeze signals, this model not necessarily perform well.

**Topological features of delay embeddings.** Now we want to embed this time series data to Euclidean spaces, using the sliding window embedding. Let  $\tau > 0$  be the time delay. We will consider the following sets

$$W(t) := \{(w(t), w(t + \tau)) \mid t \in [a, b]\},$$

$$W_i(t) := \{(w_i(t), w_i(t + \tau)) \mid t_{i-1} \leq t, t + \tau \leq t_i\}.$$

$W(t)$  is the embedding of the whole data and  $W_i(t) \subseteq W(t)$  is the  $i$ -th piece of it. In the following discussion, we first focus on the constant amplitude case, and then generalize it to the varying amplitude  $A(t)$  case. We will elaborate that how the first Betti number of the delay embedding  $W(t)$  being at least 1. The following theorem indicates that the delay embedding sets of two sinusoidal functions that differ by a phase are equal after a reparametrization.

**Theorem 4.2.1.** Suppose  $v_1(t) = A \sin(\frac{2\pi}{T}t + \phi_1)$  and  $v_2(t) = A \sin(\frac{2\pi}{T}t + \phi_2)$ ,  $t \in [0, \infty)$  and  $\phi_2 > \phi_1$ . For  $\tau > 0$ , consider  $V_i(t) = \{(v_i(t), v_i(t + \tau)) \mid t \in [0, \infty)\}$  for  $i = 1, 2$ . Let  $t' = t + \frac{\phi_2 - \phi_1}{2\pi}T$  and  $V_i = \{(v_i(t'), v_i(t' + \tau)) \mid t \in [\frac{\phi_2 - \phi_1}{2\pi}T, \infty)\}$ , then  $V_1 = V_2(t)$ .

Therefore, we can ignore the phase during our analysis. The following theorem shows that for suitable parameter  $\tau$ , the time delay embedding of a sinusoidal function is an ellipse.

**Theorem 4.2.2.** Suppose  $u(t) = A \sin(\frac{2\pi}{T}t + \phi)$ ,  $t \in [0, \infty)$  and  $\tau \neq k\frac{T}{2}$  for all  $k \in \mathbb{Z}$ . Then the time delay embedding  $U(t)$  for  $u(t)$  yields an ellipse with radii  $\alpha, \beta = A\sqrt{1 \pm \cos(\frac{2\pi}{T})}\tau$  centered at the origin with angle of rotation  $\pm 45^\circ$ . The circumscribed square around the ellipse has length  $2A$ .

In the theorem above, the period  $T$  is fixed while the time delay  $\tau$  can vary. The following theorem gives the relation of the variation of  $\tau$  and  $T$ .

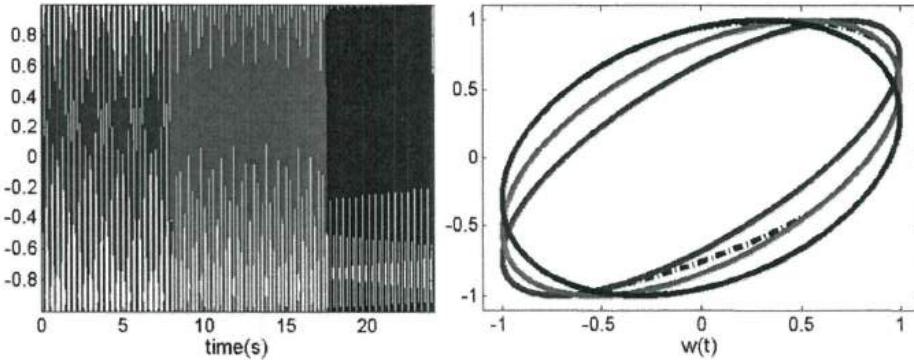
**Theorem 4.2.3.** Suppose  $u_i(t) = A \sin(\frac{2\pi}{T_i}t)$ . Let  $U_i(t) = \{(A \sin(\frac{2\pi}{T_i}t), A \sin(\frac{2\pi}{T_i}(t + \tau_i)))\}$  with time delay  $\tau_i$  for  $i = 1, 2$ . Then the time delay embedding sets  $U_1(t)$  and  $U_2(t)$  differ by a reparametrization if and only if  $\frac{\tau_1}{\tau_2} = \frac{T_1}{T_2}$ .

Now we discretize the above data. Let  $T_s$  be a small sampling time and  $I := \{T_s, 2T_s, \dots, nT_s\}$  be a finite set of time. Let  $W = \{W(t_j) \mid t_j \in I\}$  and  $W_i = \{W_i(t_j) \mid t_j \in I, t_j, t_j + \tau \in [t_{i-1}, t_i]\}$ . Set  $K := \{(w_i(t_j), w_{i+1}(t_j + \tau)) \mid t_j \in I, t_j \in [t_{i-1}, t_i], t_j + \tau \in [t_i, t_{i+1}]\}$  to be the set of transition points from different  $w_i$  segments, then we have

$$W = \bigcup_{i=1}^n W_i \bigcup K.$$

We note that  $K$  only occupies a small fraction of  $W$ , for example, in a real-world experiment,  $K$  only contains 1.6% of the whole data points. Theorem 4.2.2 tells us that the radii of the ellipse depend only on the delay when the period is fixed. Theorem 4.2.3 says that changing delay can yield to same effects as changing the period. Therefore, when the period of  $w(t)$  changes piecewisely constantly with time, the radii of the ellipse varies accordingly. As a result,  $\bigcup_{i=1}^n W_i$  with constant amplitude  $A$  is a set of concentric ellipses with angle of rotation  $\pm 45^\circ$ , all contained in a square with side length  $2A$ . If  $A$  varies with time, the only thing changed is that different ellipse would have circumscribed squares with different side lengths.

**Theorem 4.2.4.** Suppose  $\bigcup_{i=1}^n W_i \subseteq \mathbb{R}^2$  is a union of concentric ellipses, with radii  $\alpha_i, \beta_i = A_i \sqrt{1 \pm \cos(\frac{2\pi}{T_i} \cdot \tau)}$ , all centered at zero with angle of rotation  $\pm 45^\circ$ . Then the persistence diagram of  $\bigcup_{i=1}^n W_i$  has at least one 1-dimensional persistent barcode.



**Figure 4.3:** The frequency of  $w(t)$  changed for three times, in different colors (left) and the corresponding time delay embedding set as ellipses including the set  $K$  in dashed curve (right). Cf<sup>[8]</sup>.

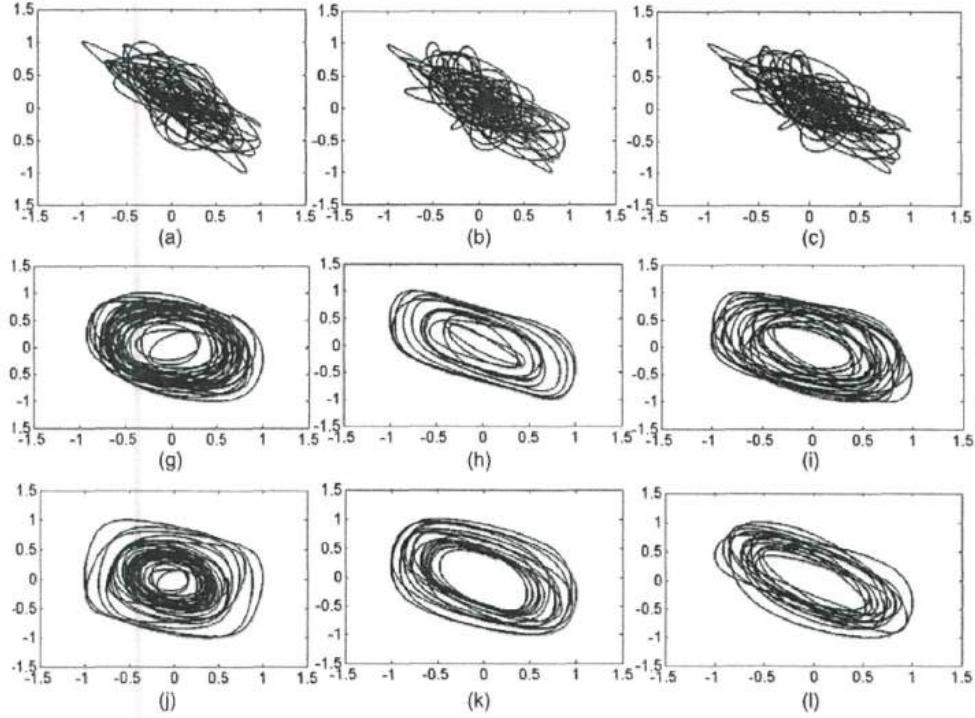
An example of  $\bigcup_{i=1}^n W_i$  can be seen in figure 4.3. From the above analysis, we have the following conclusion:

**Corollary 4.2.5.** Let  $w$  be the model of Wheeze signals. The corresponding time delay embedding  $W$  with suitable parameter  $\tau$  is a set of concentric ellipses  $\bigcup_{i=1}^n W_i$  with angles of rotation  $\pm 45^\circ$ , with different radii and different-sized circumscribed squares. Therefore, the Rips complex associated to the point cloud  $W$  always has at least one 1-dimensional persistent barcode.

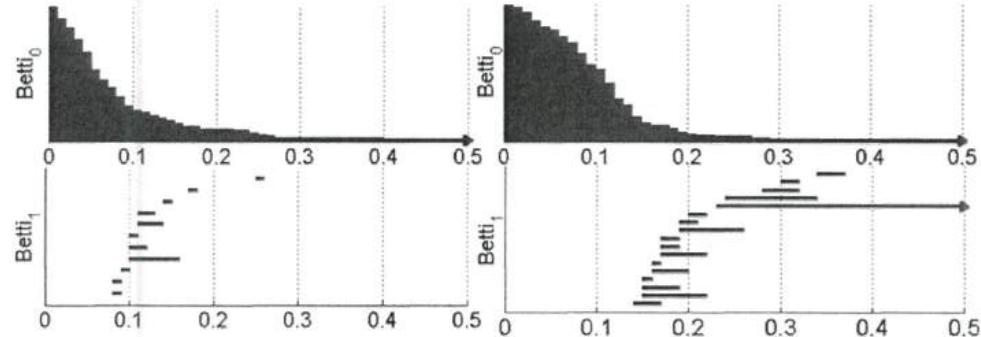
The resulting 1-dimensional persistent barcodes of Wheeze signals always exist, while the experiments have shown that this is not the case for non-wheeze signals. Therefore, the first persistent barcodes can effectively distinguish between these two kinds of signals.

**Experimental Results** Figure 4.4 shows nine lung signals consisting of three normal sounds and six types of wheezes. One can see that compared to non-wheeze signals, the wheeze

signals behave apparent periodic features. One-dimensional holes appear on them. Figure 4.5 prints the corresponding barcodes, with the wheeze data having a significant long 1-dimensional barcode.



**Figure 4.4:** The time delay embedding set of non-wheeze signals (a,b,c) and wheeze signals (g,h,i,j,k,l). Cf<sup>[8]</sup>.



**Figure 4.5:** The barcodes for non-wheeze sound signals (left) and wheeze sound signals (right). The significant one-dimensional hole is indicated by the red bar, which is the key to distinguish between these two. Cf<sup>[8]</sup>.

Compared to other methods for wheeze detection, this one achieves very high accuracy, which is 98.39%. This is perhaps because the persistent homology extract the topological features of the data, which is more important than geometric ones in the problem of wheeze detection.

## Chapter 5 Extended Persistence and its Application

In this chapter, we introduce another kind of persistent homology which extends the notion of persistent homology introduced in chapter 3. This new method, called *extended persistence*, can be used to define a function characterizing shape on surfaces in  $\mathbb{R}^3$ . Using this “shape function”, we can detect the place of “cavities” and “protrusions” of a surface by calculating the local maxima of the shape function. Detecting shape can further be used to problems such as protein docking, as we shall see in the last section of this chapter. The main references for this chapter are<sup>[2], [6]</sup> and<sup>[9]</sup>.

To officially get started, we need more knowledge about algebraic topology.

### 5.1 Relative Homology and Morse Function

Basic notions of relative homology and Morse function appear in the extended persistence, so we introduce here.

**Definition 5.1.1.** A **pair of spaces**, or simply a **pair**, is an ordered pair  $(X, Y)$  where  $X$  is a topological space and  $Y \subseteq X$  is a subspace.

A space  $X$  can be identified with the pair  $(X, \emptyset)$ . Therefore, space pairs are generalization of topological spaces.

**Definition 5.1.2** (Relative Chain). Let  $A$  be an abelian group. The  **$n$ -dimensional relative chain with coefficients in  $A$**  of the pair  $(X, Y)$  is defined to be  $C_n(X, Y; A) := C_n(X)/C_n(Y)$ . An element in  $C_n(X, Y; A)$  is called a relative  $n$ -chain.

By the definition of quotient spaces, one can regard a relative chain as a regular chain in  $C_n(X)$ , ignoring the coefficients on singular simplices that completely lie in  $Y$ . The boundary map  $\partial_n : C_n(X) \rightarrow C_{n-1}(X)$  induces naturally the boundary map  $\partial_n : C_n(X, Y; A) \rightarrow C_{n-1}(X, Y; A)$  (we abuse the same notation) by taking boundary of a representative of an element in  $C_n(X, Y; A)$ , and then ignoring the coefficients on singular simplices in  $Y$ .  $\partial_n \circ \partial_{n+1} = 0$  follows directly from the property of boundary maps in the ordinary chain complex, see proposition 2.1.5.

**Definition 5.1.3** (Relative Homology). Keep all the notions above. Let  $Z_n(X, Y; A) := \ker \partial_n$ ,  $B_n(X, Y; A) := \text{im } \partial_{n+1}$ , then  $Z_n \subseteq B_n$ . An element in  $Z_n$  is called an ( $n$ -dimensional) **relative closed chain**, or an ( $n$ -dimensional) **relative cycle**. An element in  $B_n$  is called an ( $n$ -dimensional) **relative boundary chain**. The notion of homologous is the same as in definition 2.1.6

The quotient group  $H_n(X, Y; A) := Z_n(X, Y; A)/B_n(X, Y; A)$  is called the  **$n$ -th relative homology group of  $(X, Y)$  with coefficients in  $A$** .

In the case  $Y = \emptyset$ , the relative homology coincides with the regular homology. If  $Y = X$ , then each relative chain groups  $C_n(X)/C_n(Y)$  will be trivial and so do the homology groups  $H_n(X, Y)$ . Geometrically, a relative cycle is a chain in  $X$  with nonempty boundary, but whose boundary is contained in  $Y$ . Similarly, a relative boundary chain is the boundary of some higher dimensional chain after ignoring the coefficients on simplices in  $Y$ .

Maps  $f : (X_1, Y_1) \rightarrow (X_2, Y_2)$  between space pairs (which by definition is a map  $f : X_1 \rightarrow X_2$  such that  $f(Y_1) \subseteq Y_2$ ) also induce maps  $f_* : H_*(X_1, Y_1; A) \rightarrow H_*(X_2, Y_2; A)$  by the same way as in the regular homology and taking quotients. Homotopic maps  $f, g$  also have the same induced maps  $f_* = g_*$ , as before.

For a  $\Delta$ -complex  $K$  and a subcomplex  $L$  of  $K$ , we also have **relative simplicial homology**, that is the homology groups of the chain complex  $\{C_n^\Delta(K, L; A) := C_n^\Delta(K; A)/C_n^\Delta(L; A)\}$ . Again, the relative singular homology coincides with the relative simplicial homology for  $\Delta$ -complexes.

We now talk about the Morse function. We have encountered differentiable functions on manifolds and their derivatives several times and took it for granted. In previous cases, readers can simply regard them as continuous functions on topological spaces without losing much information. However, when discussing the Morse function, the derivatives are important. Thus we officially introduce the notion of smooth manifolds here.

**Definition 5.1.4.** A topological space  $M$  is a **(topological) manifold** of dimension  $n$  if it is Hausdorff, second-countable and satisfies the following: for any  $x \in M$ , there is an open neighborhood  $U$  of  $x$  that is homeomorphic to some open subset of  $\mathbb{R}^n$ . In other words, a manifold is locally homeomorphic to an open set in  $\mathbb{R}^n$ , but unknown globally.

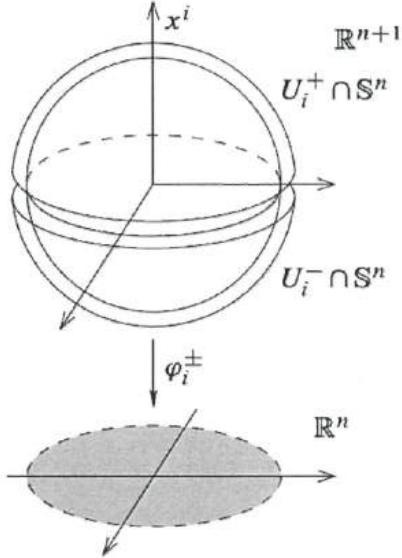
A pair  $(U, \varphi)$  where  $U \subseteq M$  is an open subset and  $\varphi : U \rightarrow V$  is a homeomorphism from  $U$  to some open  $V \subseteq \mathbb{R}^n$ , is called a (local) **chart** of  $M$ . If  $U = M$ , then  $(U, \varphi)$  is called a **global chart**. By definition, every point of a manifold is contained in the domain  $U$  of some chart of  $M$ .

An **atlas** on  $M$  is a collection of charts  $\{(U_\alpha, \varphi_\alpha)\}$  such that the union of all  $U_\alpha$ 's covers  $M$ . By definition, every manifold admits a chart.

**Example 5.1.5.** (1) Any open subspace  $U$  of  $\mathbb{R}^n$  is a manifold with a global chart  $(U, id)$ .

(2) The  $n$ -sphere  $S^n := \{(x_0, \dots, x_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^n x_i^2 = 1\} \subseteq \mathbb{R}^{n+1}$  is an  $n$ -manifold, covered by  $2(n+1)$  charts: let  $U_i^+ := \{(x_0, \dots, x_n) \in S^n \mid x_i > 0\}$ , then  $\varphi_i : U_i^+ \rightarrow \overset{\circ}{D}_n$ ,  $(x_0, \dots, x_n) \mapsto (x_0, \dots, \hat{x}_i, \dots, x_n)$  is a homeomorphism. Here  $D_n = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 \leq 1\} \subseteq \mathbb{R}^n$  is the  $n$ -disk and  $\overset{\circ}{D}_n$  is its interior. The hat symbol  $\hat{x}_i$  indicates deleting  $x_i$  from the  $(n+1)$ -tuple. Thus we obtain charts  $(U_i^+, \varphi_i)$ . Similarly, let  $U_i^- := \{(x_0, \dots, x_n) \in S^n \mid x_i < 0\}$  and  $\Psi_i : U_i^- \rightarrow \overset{\circ}{D}_n$ ,  $(x_0, \dots, x_n) \mapsto (x_0, \dots, \hat{x}_i, \dots, x_n)$ , then  $(U_i^-, \Psi_i)$  is also a chart of  $S^n$ . Since the union of all  $U_i$  and  $V_i$ 's cover  $S^n$ ,  $S^n$  is covered by local charts and hence an  $n$ -manifold. See figure 5.1.

Since manifolds are locally Euclidean, it is a natural desire to do calculus on it. It is also not surprising that the calculus is done on charts. To define the concept of smooth function on  $M$ , we do locally: we want to say  $f : M \rightarrow \mathbb{R}$  is smooth if  $f \circ \varphi^{-1}$  is smooth for every chart

Fig. 1.3 Charts for  $\mathbb{S}^n$ 

are graph coordinates for  $\mathbb{S}^n$ . Since each point of  $\mathbb{S}^n$  is in the domain of at least one of these  $2n + 2$  charts,  $\mathbb{S}^n$  is a topological  $n$ -manifold. //

Figure 5.1: Charts on the sphere  $S^2$ 

$(U, \varphi)$ . However, different charts may lead to different kind of smoothness on  $M$ . Therefore, we need the compatibility of charts.

**Definition 5.1.6.** Let  $M$  be a smooth manifold. Two charts  $(U, \varphi)$  and  $(V, \Psi)$  on  $M$  are said to be **(smoothly) compatible** if the **transition function**  $\Psi \circ \varphi^{-1} : \varphi(U \cap V) \rightarrow \Psi(U \cap V)$  (which is a homeomorphism) is smooth.

An atlas  $\{(U_\alpha, \varphi_\alpha)\}$  is said to be smooth if each  $(U_\alpha, \varphi_\alpha)$  is compatible with each other. A smooth atlas of  $M$  is also called a **smooth structure** on  $M$ . A manifold with a smooth structure is called a **smooth manifold**.

Now we can formally define the smoothness of a function.

**Definition 5.1.7.** Let  $M$  be a manifold and  $\{(U_\alpha, \varphi_\alpha)\}$  be a smooth structure on it. A function  $f : M \rightarrow \mathbb{R}$  is said to be smooth if  $f \circ \varphi_\alpha^{-1} : \varphi_\alpha(U_\alpha) \rightarrow \mathbb{R}$  is smooth for every chart  $(U_\alpha, \varphi_\alpha)$  in the smooth structure.

Each  $f \circ \varphi_\alpha^{-1}$  is said to be a **coordinate representation** of  $f$ .

**Example 5.1.8.** The height function  $h : S^n \rightarrow \mathbb{R}$ ,  $h(x_0, \dots, x_n) = x_n$  is smooth, as one can check.

Once a function is smooth, we can differentiate it infinitely many times. In this article, we still only do the differentiation locally. Fortunately, this is enough to give the definition of the most interesting smooth function: the Morse function.

We differentiate  $f$  by differentiating the coordinate representations  $f \circ \varphi^{-1}$  of  $f$ . Denote the derivatives by  $d(f \circ \varphi^{-1})$ , which is vector-valued (think about the gradient vector in

classical calculus). It is not hard to check that if  $p \in U \cap V$ , where  $(U, \varphi)$  and  $(V, \Psi)$  are two smooth charts, then  $d(f \circ \varphi^{-1}) = 0$  if and only if  $d(f \circ \Psi^{-1}) = 0$ . We may simply write as  $df(p) = 0$ . Hence the following definition makes sense.

**Definition 5.1.9.** Let  $f : M \rightarrow \mathbb{R}$  be a smooth function. A point  $p \in M$  is said to be a **critical point** of  $f$  if  $df(p) = 0$ . A critical point  $p$  is said to be **degenerate** if the Hessian matrix  $Hess_f(p)$  is singular, and **nondegenerate** otherwise.

A smooth function  $f : M \rightarrow \mathbb{R}$  is said to be **Morse** if

- (1) All critical points of  $f$  are nondegenerate;
- (2) All critical points of  $f$  have distinct values.

There are interesting properties for Morse functions. We will not give proof for most of them, but focus more on their geometric intuition and applications.

**Theorem 5.1.10** (Morse lemma). *Let  $f : M \rightarrow \mathbb{R}$  be a Morse function and  $p \in M$  be a critical point of  $f$ . Then there is a smooth chart  $(U, \varphi)$  containing  $p$  such that  $f$  has coordinate representation  $f(x_1, \dots, x_n) = f(p) - x_1^2 - \dots - x_\lambda^2 + x_{\lambda+1}^2 + \dots + x_n^2$ , where  $\lambda$  is an integer between 1 and  $n$ . Moreover,  $\lambda$  does not depend on the choice of chart and is called the **index** of  $f$  at  $p$ .*

**Corollary 5.1.11.** *If  $f : M \rightarrow \mathbb{R}$  is a Morse function, then the set of all critical points of  $f$  is discrete.*

*Proof.* By theorem 5.1.10, for a critical point  $p$  of  $f$ , it has a neighborhood  $U$  where the derivative of  $f$  is non-vanish except for at  $p$ . In other words, every critical point is not a limit point, hence the set is discrete.  $\square$

Before digging in more interesting properties, we give an example.

**Example 5.1.12.** As in figure 5.2, we have a torus  $\mathbb{T}$  embedded in  $\mathbb{R}^3$ .  $f : \mathbb{T} \rightarrow \mathbb{R}$  given by  $f(x, y, z) = z$  is the height function on the torus. Then  $f$  is Morse, having critical value  $a_0, a_1, a_2, a_3$  with index 0, 1, 1, 2 respectively.

Let  $M_a := f^{-1}(-\infty, a]$ , then  $M_a = \emptyset$  when  $a < a_0$ ;  $M_a$  is contractible when  $a_0 \leq a < a_1$ ;  $M_a$  is homotopic to a disk with two points on the boundary joined by a curve, which is the side face of a cylinder; by further connecting two points on the boundary of the cylinder, we obtain  $M_a$  for  $a_1 \leq a < a_2$ ; Finally, the torus is obtained by filling the hole using a 2-disk.

As we can see, the homotopy type of  $M_a$  and  $M_b$  are the same if there is no critical value between  $a$  and  $b$ . Moreover, passing through a critical value upward effect the homotopy type by adding a cell of dimension the index of the corresponding critical point. This is not a coincidence, as we shall see in the next theorem.

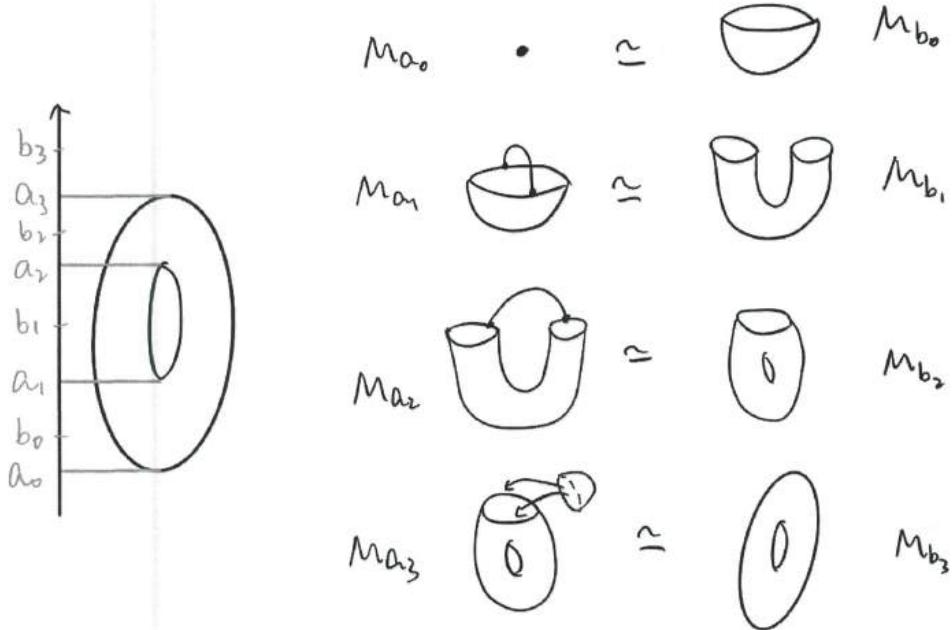
**Theorem 5.1.13** (Fundamental Theorem for Morse Function). *Let  $M$  be a smooth manifold and  $f : M \rightarrow \mathbb{R}$  be a Morse function. Let  $M_a = f^{-1}(-\infty, a]$ .*

(1) *If  $a < b \in \mathbb{R}$  are real numbers such that there is no critical value between  $a$  and  $b$ , then  $M_a$  is homeomorphic to  $M_b$  and  $M_a$  is a deformation retract to  $M_b$ .*

(2) *Let  $p$  be a critical point of index  $\lambda$  of  $f$  and set  $c = f(p)$ . Then for  $\epsilon > 0$  sufficiently small,  $M_{p+\epsilon}$  is homotopic to  $M_{p-\epsilon}$  with a  $\lambda$ -cell attached.*

Therefore, the cellular decomposition of a manifold can be given by a Morse function. As in example 5.1.12, the four critical points correspond to a 0-cell, two 1-cells and a 2-cell in the *CW*-complex structure of the torus.

With the basic knowledge in hand, we are ready to play with the extended persistence.



**Figure 5.2:** The height function gives the cellular decomposition of the torus

## 5.2 Extended Persistence

In the ordinary persistence, we started with a filtration  $\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  of a simplicial complex  $K$ . We can generalize them to the category of space pairs:

$$(\emptyset, \emptyset) = (K_0, \emptyset) \subseteq (K_1, \emptyset) \subseteq \dots \subseteq (K_n = K, \emptyset)$$

Moreover, let  $K^i := K - K_i$ , we get another sequence of inclusions of space pairs:

$$(K, \emptyset) = (K, K^n) \subseteq (K, K^{n-1}) \subseteq \dots \subseteq (K, K^0) = (K, K)$$

Put them together, we obtain an extended filtration of  $K$ :

$$(\emptyset, \emptyset) = (K_0, \emptyset) \subseteq \dots \subseteq (K_n, \emptyset) = (K, K^n) \subseteq (K, K^{n-1}) \subseteq \dots \subseteq (K, K^0) = (K, K)$$

The filtration and the inclusion maps induce maps between the homology groups for each  $p$ :

$$\dots \rightarrow H_p(K_i, K^j) \rightarrow \dots$$

which is also a persistent vector space as in definition 3.2.3. Therefore, the persistent diagram characterizing how homology classes created and dying of  $\{H_p(K_i, K^j)\}$  can be drawn for each dimension  $p$ .

An important observation is that both the first and the last term,  $(\emptyset, \emptyset)$  and  $(K, K)$  in the extended filtration have trivial relative homology (all  $H_p = 0$ ). The persistence vector space start and end with no nontrivial homology class, so all classes being created at some time are killed within finite time. In other words, there is no infinite point in the (relative)

persistence diagram of the extended filtration. Up to the term  $(K_n, \emptyset)$ , the relative homology groups coincide with the ordinary ones, so the finite points diagram is the same as that in the ordinary persistent diagram. The infinite points, however, are killed by the following relative terms  $(K, K^i)$  in the extended filtration. There is a theorem characterizing how this be done in details.

From no on, we consider filtration obtained by Morse functions. Let  $f : M \rightarrow \mathbb{R}$  be a Morse function,  $M_a = f^{-1}(-\infty, a]$  and  $M^a = f^{-1}[a, \infty)$ . Suppose  $f$  has critical values  $a_1 < \dots < a_n$ . Let  $b_0, b_1, \dots, b_n \in \mathbb{R}$  be real numbers such that  $b_0 < a_1 < b_1 < \dots < a_n < b_n$ . Then we have two filtration,  $\emptyset = M^{b_n} \subseteq M^{b_{n-1}} \subseteq \dots \subseteq M^{b_0} = M$  and  $(M, \emptyset) = (M, M^{b_n}) \subseteq (M, M^{b_{n-1}}) \subseteq \dots \subseteq (M, M^{b_0}) = (M, M)$ .

**Definition 5.2.1.** In the filtration  $\emptyset = M^{b_n} \subseteq M^{b_{n-1}} \subseteq \dots \subseteq M^{b_0} = M$ , call a class  $\gamma \in H_p(M^{b_i})$  **essential** if  $\text{pers}(\gamma) = \infty$ , and **inessential** otherwise.

**Theorem 5.2.2.** *With the notions above,*

- (1) *An inessential dimension  $p$  homology class of  $M^b$  gets born at the same time that a dimension  $p+1$  relative homology class of  $(M, M^b)$  gets born.*
- (2) *A dimension  $p$  homology class of  $M^b$  dies at the same time that a  $p+1$  relative homology class of  $(M, M^b)$  dies.*
- (2) *An essential dimension  $p$  homology class of  $M^b$  gets born at the same time that a dimension  $p$  relative homology class dies.*

By the theorem, we can analyze the relative homology classes of  $(M, M^b)$  by taking complement and going backward, looking at the ordinary homology classes in  $M^b$ . Instead of giving a proof, we look at an example and understand the geometric explanation.

**Example 5.2.3.** Figure 5.3 illustrates a genus 2 surface  $M$  embedded in  $\mathbb{R}^3$ .  $f$  is a height function on it, with critical values  $a_1 < \dots < a_{10}$ .

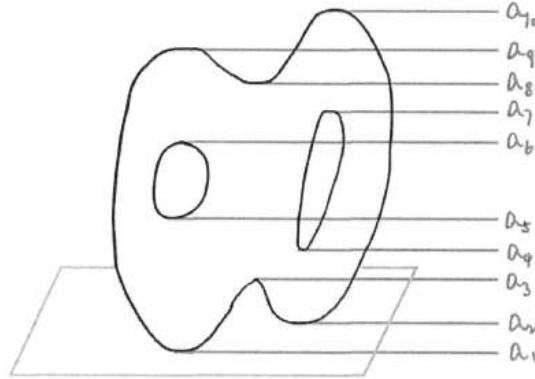
Going up, the critical point  $a_1$  gives birth to a 0-cycle, namely a connected component, making  $M_{b_1}$  homeomorphic to a disk.  $a_2$  gives birth to another 0-cycle, which is killed by  $a_3$ . Each of  $a_4, a_5, a_6, a_7, a_8$  creates a new 1-cycle in  $H_1$  and the class created by  $a_8$  is killed by  $a_9$ .  $a_{10}$  finally creates a 2-cycle. All of the classes created by  $a_1, a_4, a_5, a_6, a_7$  and  $a_{10}$  are essential, living forever in the ordinary persistence.

Coming down,  $a_{10}$  gives birth to an essential 0-cycle in  $M^{a_{10}}$ . By theorem 5.2.2 (3), a 0-relative cycle dies in  $(M, M^{a_{10}})$ . This is the connected component created by  $a_0$  upward. If one know more about homology, notice that  $H_*(M, M^{10}) = \tilde{H}_*(M \cup CM^{10})$ , where  $\tilde{H}$  denotes the reduced homology and  $CM^{10} := M^{10} \times [0, 1]/M^{10} \times \{1\}$  is the cone over  $M^{10}$ . When passing to reduced homology, the only connected component vanishes.

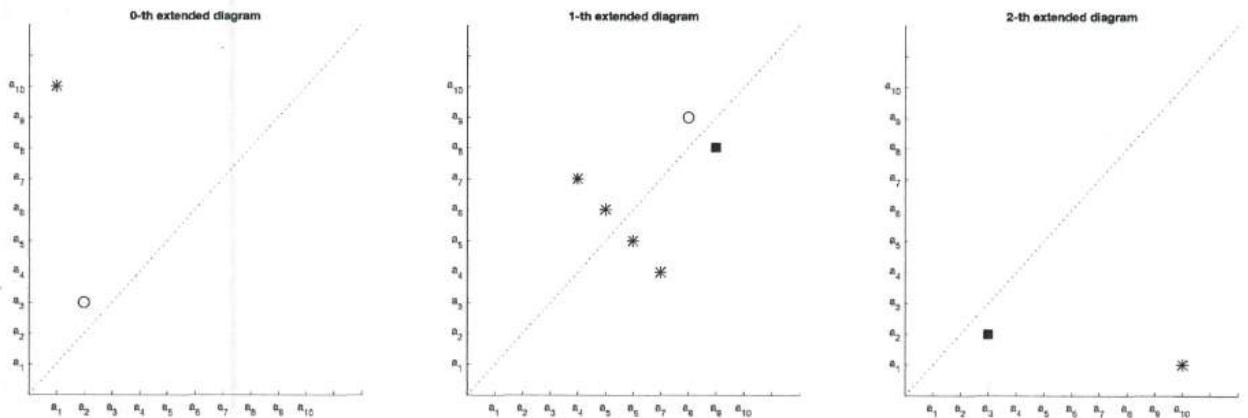
$a_9$  gives birth to an inessential 0-cycle in  $M^{a_9}$ , which is killed by  $a_8$ . By theorem 5.2.2 (1) (2),  $a_9$  gives birth to a 1-relative class in  $(M, M^{a_9})$  killed by  $a_8$ . Actually, an inessential 0-cycle is the boundary of some 1-chain, which is contained in  $M^{a_8}$  in this case. This 1-chain is a relative 1-cycle in  $(M, M^{a_9})$  and become trivial in  $(M, M^{a_8})$ .

Each of  $a_7, a_6, a_5, a_4$  gives birth to an essential 1-dimensional homology class in  $M^{a_7}, M^{a_6}, M^{a_5}, M^{a_4}$  respectively. By theorem 5.2.2 (3) again,  $a_7, a_6, a_5, a_4$  kill the upward essential 1-cycles created by  $a_4, a_5, a_6, a_7$ , respectively. As one can check, the 1-cycles created by  $a_4, a_5, a_6, a_7$  are contained in  $M^{a_7}, M^{a_6}, M^{a_5}, M^{a_4}$  respectively, hence being trivial in the corresponding space pairs.

Finally, by the theorem again,  $a_3$  gives birth to a 2-relative cycle killed by  $a_2$ , and  $a_1$  kills the 2-cycle created by  $a_{10}$  going up. The extended persistence diagram is shown in figure 5.4. Note that there is no infinite points.



**Figure 5.3:** Height function on a genus 2 surface



**Figure 5.4:** The 0-th, 1-th and 2-th extended diagram  $D$  of the height function on a genus 2 surface. For each  $(a_i, a_j) \in D$ , there is an extended homology class born at  $a_i$  and dying at  $a_j$ .

One may have observed the symmetry in the above extended persistent diagram in complementary dimension. This is not a coincidence.

**Definition 5.2.4.** Let  $f : M \rightarrow \mathbb{R}$  be a Morse function and  $Dgm_p(f)$  be the extended persistent diagram of  $f$ . That is, the extended filtration comes from the subspaces  $M_b$  and  $M^b$ , where each  $b$  is a critical point of  $f$ . We call homology classes born and dying going up **ordinary**; call those born going up and dying coming down **extended**; call those born and dying coming down **relative**. We obtained the ordinary, extended and relative subdiagrams of  $Dgm(f)$  consisting of ordinary, extended and relative homology classes respectively. Denote these subdiagrams by  $Ord(f)$ ,  $Ext(f)$  and  $Rel(f)$ .

In figure 5.4, we used symbols  $\circ$ ,  $*$ ,  $\blacksquare$  to denote the ordinary, extended and relative classes respectively. One can easily notice the symmetric pattern on it, which is not a coincidence.

**Theorem 5.2.5** (Persistence Duality Theorem). *If  $f : M \rightarrow \mathbb{R}$  is a Morse function, where  $M$  is an  $n$ -dimensional manifold, then*

$$\text{Ord}_p(f) = \text{Rel}_{d-p}^T(f)$$

$$\text{Ext}_p(f) = \text{Ext}_{d-p}^T(f)$$

$$\text{Rel}_p(f) = \text{Ord}_{d-p}^T(f)$$

where the superscript ‘ $T$ ’ indicates the reflection along the main diagonal  $(a, b) \mapsto (b, a)$ . Equivalently,  $\text{Dgm}_p(f) = \text{Dgm}_{d-p}^T(f)$ .

**Theorem 5.2.6** (Persistence Symmetry Theorem). *With the same notations in theorem 5.2.5, we have*

$$\text{Ord}_p(f) = \text{Ord}_{d-p-1}^R(-f)$$

$$\text{Ext}_p(f) = \text{Ext}_{d-p}^o(-f)$$

$$\text{Rel}_p(f) = \text{Rel}_{d-p+1}^R(-f)$$

where the superscript ‘ $R$ ’ indicates the reflection along the minor diagonal  $(a, b) \mapsto (-b, -a)$  and ‘ $o$ ’ indicates the reflection with respect to the origin  $(a, b) \mapsto (-a, -b)$ .

**Computation.** We now proceed to the computation of the extended persistence. In the ordinary persistence, we obtained the ordinary persistent diagram by reducing the boundary matrix  $\partial$ . To compute the extended diagram, we do the same reduction but on an augmented matrix.

To begin with, we need to discretize the smooth manifold to a simplicial complex  $K$ , as well as the Morse function, to an increasing function on  $K$ . In fact, there always exist a  $\Delta$ -complex structure on a 2-manifold  $M$ . Let  $K$  be a triangulation of  $M$ , that is, a  $\Delta$ -complex structure on  $M$ . We obtain a function, still denoted by  $f : K \rightarrow \mathbb{R}$ , given by  $f(\sigma) = \max_{x \in \sigma} f(x)$ . Here  $\sigma$  is a simplex in  $K$  and we abuse the notation  $\sigma$  to also denote its image in  $M$ . By definition,  $f(\sigma) \leq f(\tau)$  whenever  $\sigma$  is a face of  $\tau$ . Hence  $f$  is an increasing function.

**Definition 5.2.7.** Let  $K$  be a simplicial complex and  $\sigma \in K$  be a simplex. A **face** of  $\sigma$  is a simplex in  $K$  spanned by a subset of vertices of  $\sigma$ . The **star of  $\sigma$**  is the set of all cofaces of  $\sigma$ , denoted  $\text{St}(\sigma) := \{\tau \in K \mid \sigma \leq \tau\}$ .

Given  $f : K \rightarrow \mathbb{R}$  an increasing function, the **lower star** of  $\sigma$  is  $\text{St}_-(\sigma) := \{\tau \in \text{St}(\sigma) \mid f(\tau) \leq f(\sigma)\}$ . The **upper star** of  $\sigma$  is  $\text{St}^+(\sigma) := \{\tau \in \text{St}(\sigma) \mid f(\tau) \geq f(\sigma)\}$ .

Now given  $f : K \rightarrow \mathbb{R}$  increasing, let  $v_1, \dots, v_n$  be the vertices, that is, 0-simplices of  $K$ , ordered by function values  $f(v_1) < f(v_2) < \dots < f(v_n)$  (we may assume the genericity of  $f$  that all vertices have distinct function value). Let  $K_i$  be the full subcomplex of  $K$  spanned by  $v_1, \dots, v_i$  and  $K^i$  be the full subcomplex spanned by  $v_{i+1}, \dots, v_n$ . Then

$$\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K = (K, \emptyset) \subseteq (K, K^1) \subseteq \dots \subseteq (K, K^n) = (K, K) \quad (5.1)$$

is an extended filtration. Easy to verify that  $K_i = K_{i-1} \cup \text{St}_-v_i = \bigsqcup_{j=1}^i \text{St}_-v_j$  and  $K^i = K^{i+1} \cup \text{St}^+v_{i+1} = \bigsqcup_{j=i+1}^n \text{St}^+v_j$ . Hence this is also called the **lower and upper star filtration**.

To write the boundary matrix, we have to order the simplices in  $K$ . Among each  $K_i - K_{i-1}$ , there may be more than one simplices. Order them arbitrarily, as long as in non-decreasing dimension. Then we have an ordering  $\sigma_1, \sigma_2, \dots, \sigma_m$  on all simplices of  $K$  and the corresponding boundary matrix  $A_{m \times m}$  (with coefficient field  $\mathbb{Z}_2$ ) where  $A[i, j] = 1$  if  $\sigma_j$  is a face of  $\sigma_i$  and  $A[i, j] = 0$  otherwise. Going backward, each  $K^{i-1} - K^i$  also contains several simplices. Again, we order them arbitrarily, just making sure that the simplices are in non-decreasing dimension, and get another total ordering  $\sigma'_1, \sigma'_2, \dots, \sigma'_{m'}$  of the simplices of  $K$ . By this ordering we obtain a boundary matrix  $B$ . Now let  $P$  be the permutation matrix between  $(1, 2, \dots, m)$  and  $(1', 2', \dots, m')$  and  $\partial = \begin{pmatrix} A & P \\ 0 & B \end{pmatrix}$  be the  $(2m) \times (2m)$  block upper-triangular matrix.

**Theorem 5.2.8.** *With the notations above, let  $R$  be the reduced matrix obtained by reducing the augmented matrix  $\partial$  by the algorithm ???. Then there is a one-to-one correspondence between the lowest 1's in  $R$  and points in the extended persistence diagram  $D$  of the filtration 5.1.*

*This theorem is an analogue of theorem 3.2.10. More precisely, the lowest 1 of the  $j$ -th column is at the  $i$ -th row if and only if the point  $(i, j) \in D$ .*

Therefore, we get the extended persistent diagram by doing reduction on the augmented boundary matrix  $\partial$ , just as in the ordinary persistence when we reducing the boundary matrix  $\partial$ . We will give a proof of this theorem, which requires more knowledge of homology. Readers without this knowledge can skip the proof with no influence on the understanding of the following sections.

*Proof of theorem 5.2.8.* There is an isomorphism  $H_p(X, Y) \cong \tilde{H}_p(X \cup CY)$ , where  $\tilde{H}$  denotes the reduced homology and  $CY := Y \times [0, 1]/Y \times \{1\}$  is the cone over  $Y$ . Thus the extended persistence homology groups also read

$$0 \rightarrow H_p(K_1) \rightarrow \dots \rightarrow \tilde{H}_p(K \cup CK^i) \rightarrow \dots \rightarrow \tilde{H}_p(K \cup CK^{n-1}) \rightarrow \tilde{H}_p(CK) = 0$$

Note that every the cone  $CK$  is contractible for every space  $X$ , hence the reduced homology end up with 0. To represent the cone, we add a dummy vertex  $v_0$ . For a simplicial complex  $K$ ,  $CK = K \cup \{(v_0\sigma) \mid \sigma \in K\}$ , where  $(v_0\sigma)$  is the  $p+1$  simplex with vertex  $v_0$  and bottom  $\sigma$  for a  $p$ -simplex  $\sigma$ .

Since both of  $\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  and  $(K, \emptyset) \subseteq (K, K^1) \subseteq \dots \subseteq (K, K)$  are filtrations and the simplices in each  $K_i - K_{i-1}$  are ordered in non-decreasing dimension, both the boundary matrices  $A$  and  $B$  are upper-triangular and so does  $\partial$  and  $R$ . The part  $\begin{bmatrix} A \\ 0 \end{bmatrix}$  of  $\partial$  represents the boundary of the bottom  $K$  of  $CK$ , as we can easily see.

Now we will show that the part  $\begin{bmatrix} P \\ B \end{bmatrix}$  represents the boundary of  $\{(v_0\sigma)\} \subseteq K$ . Actually, each column in  $\begin{bmatrix} P \\ B \end{bmatrix}$  stands for a simplex of the form  $(v_0\sigma_{i'})$  and we have  $\partial(v_0\sigma_{i'}) = \sigma_k + (v_0\partial\sigma_{i'})$ , where  $k$  and  $i'$  are the order of  $\sigma$  in the ordering of  $A$  and  $B$  respectively. In the matrix  $\partial$ , the first  $m$  rows and columns stand for the first  $m$  simplices in  $K$ , while the last  $m$  rows and columns stand for those simplices  $(v_0\sigma)$  in  $CK - K$ , with the order  $(1', 2', \dots, m')$ .

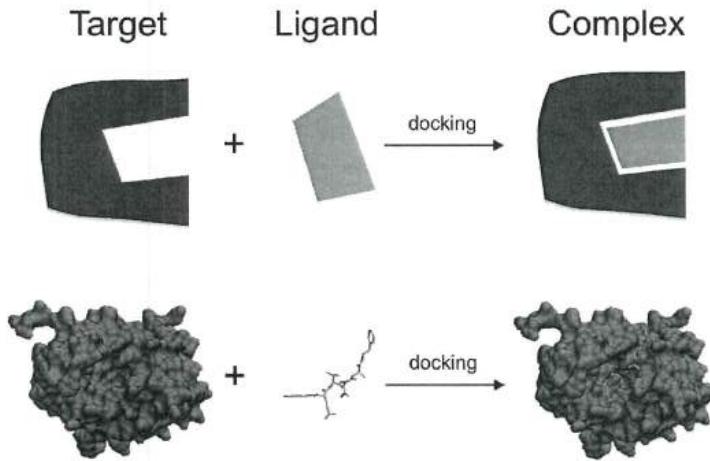
The  $(v_0 \partial \sigma_{i'})$  part in  $\partial(v_0 \sigma_{i'})$  appear in  $B$ , just  $v_0$  and thinking of the definition of  $B$ . The left term is  $\sigma_k$  itself. By the definition of the permutation matrix  $P$ , we have  $P[i, k] = 1$ . This is just saying that  $P$  gives the  $\sigma_k$  part of  $\partial(v_0 \sigma_{i'})$ . See figure  $\square$

We have talked about the idea and computation of the extended persistence. In the next section, we will give an application of it in the protein docking problem.

### 5.3 Application of the Extended Persistence in Protein Docking

Protein molecules interact with each other via cavities and protrusions on their protein surfaces, see figure 5.5. Suitable shapes of different molecules tend to bond and lead to biochemical interactions. Therefore, we would like to know the places of cavities and protrusions on a protein surface and use this information to predict whether two proteins will interact with each other. This is the protein docking problem. The main reference of this section is reference .

In this section, we use the extended persistence to define a function characterizing the shape of a surface, called the elevation function. The local maximums of the elevation function give the places of cavity and protrusion. It is an analogue of the height function on earth: we have the notion of “high” and “low” on earth by defining the height of a point as the “mean sea level”. This is reasonable because of the relatively uniform geometric shape of the earth, whose mass distribute uniformly around its center of mass in all directions. The local maximums and minimums of the height function on the earth give the cavity and protrusion on the earth. For general surfaces such as torus, this method might not work well and the elevation function is a good analogue. We note that the cavity and protrusion information is much geometric: it is not preserved under homeomorphism.



**Figure 5.5:** Two proteins interact with each other

We begin to define the elevation function. During this section,  $M$  denotes a 2-manifold embedded in  $\mathbb{R}^3$ , such as a torus.

In figure 5.3, the height function  $f$  on a 2-torus  $M$  is a Morse function. It actually pairs the critical points two by two, see figure 5.4. At each critical point  $a_i$ , we define its **elevation** to be  $|f(a_i) - f(a_j)|$ , where  $a_j$  is the critical point paired with  $a_i$  by  $f$ . For a general point  $x$  on  $M$ , the idea is that there is a “height function” in another direction for which  $x$  is a critical point.

**Definition 5.3.1.** Let  $u \in S^2$ , representing a direction. The height function at direction  $u$  is defined to be  $f_u : M \rightarrow \mathbb{R}$ ,  $f_u(x) := \langle x, u \rangle$ .

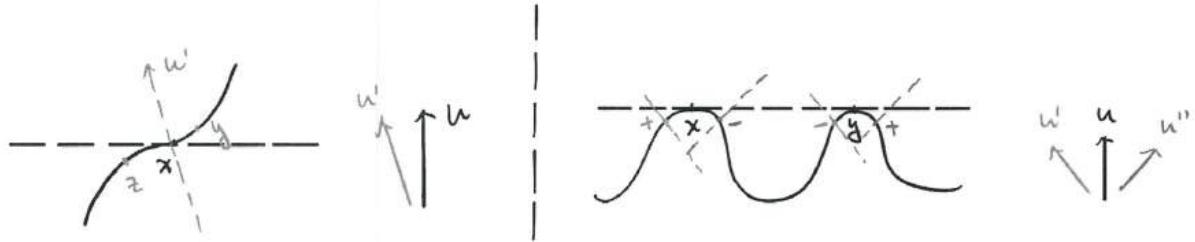
Clearly, for a smooth surface  $M \subseteq \mathbb{R}^3$  and  $x \in M$ ,  $x$  is a critical point of  $f_u$  if and only if  $u$  is perpendicular to the tangent plane  $T_x$  at  $x$ . If  $f_u$  is a Morse function, then we can pair the critical points of  $f_u$  and define their elevation. A point  $x \in M$  is critical in two opposite directions  $\pm u$ , perpendicular to  $T_x$ . We clearly have  $f_{-u}(x) = -f_u(x)$ . By the persistence symmetry theorem (theorem 5.2.6),  $(x, y)$  is get paired by  $f_u$  if and only if it is get paired by  $f_{-u} = -f_u$ . Therefore, the elevation of critical points of  $f_u$  is well-defined whenever  $f_u$  is Morse.

What if  $f_u$  is not Morse? Recalling definition 5.1.9 for the definition of Morse functions, we call a point  $x \in M$  **singular** if  $x$  is a degenerate critical point for  $f_u$ , or  $x$  is a critical point of  $f_u$  having the same function value with another critical point. Here  $u$  is the normal vector of  $T_x$ . There are two cases.

(1) Flat point. In this case,  $f_u$  has a degenerate critical point. There is a geodesic through  $x$ , restricted to which  $f_u$  has an inflection point, see figure 5.6 left. The point  $x$  is a flat point for  $f_u$ . If we disturb the direction  $u$  slightly to  $u'$ , two points  $y, z$  near  $x$  are get paired. As  $u'$  approaching  $u$ ,  $y, z$  in the pair  $(y, z)$  get closer and finally disappear at  $x$ . There is no more pairs near  $x$  when we disturb  $u$  to the other direction. For this reason,  $x$  is also called a birth-death point.

(2) Shared tangent plane. In this case,  $f_u$  has two degenerate critical points with the same function value, see figure 5.6 right.  $x, y$  are two different critical points of  $f_u$  with  $f_u(x) = f_u(y)$ . This happens when  $x, y$  has the same tangent plane. If we disturb  $u$  slightly to  $u'$  to the left, we get two paired points near  $x$  and  $y$ , with the point near  $x$  higher in the direction of  $u'$ . As  $u'$  approaches to  $u$ , the two points converge to  $x, y$  respectively. When disturbing  $u$  to  $u''$  to the right, we get another pair of points near  $x, y$  getting paired, with the one near  $y$  higher in the direction of  $u''$ . The order of height near  $x, y$  changed when the direction passing through  $u$ , so  $x, y$  are also called interchange points.

In each cases, the analysis is based on moving  $u$  along one degree of freedom. If we move  $u$  through another degree of freedom on  $S^2$ , we get a curve of singularities, on which all points are flat or interchange points. Curves of singularities may intersect, causing higher order singularities. We wish that  $M$  is not too bad, so we make the **genericity assumption** on  $M$ : *The only singular points of  $M$  are flat points, interchange points or intersections of two curves of singularities.*

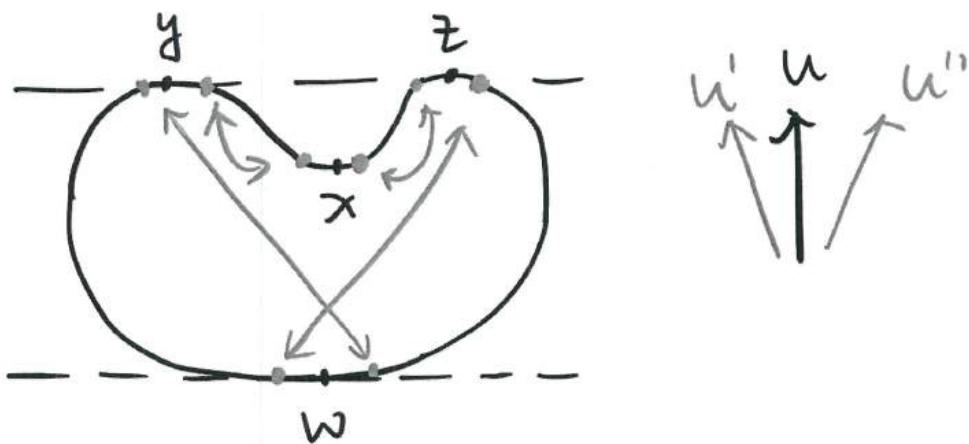


**Figure 5.6:** Two cases when  $f_u$  is not Morse. Left: flat point/birth-death point. Right: interchange point.

With the genericity assumption, we can define the elevation of  $x \in M$  for almost every  $x$ , by  $\text{elevation}(x) := |f_u(x) - f_u(y)|$ , where  $u$  is the normal vector of  $T_x$  and  $y$  is the point paired with  $x$  by  $f_u$ , assuming  $f_u$  being Morse. As discussed above, when  $f_u$  is not Morse, we want to define the elevation at  $x$  by analyzing the elevation on the neighborhood of  $x$  and taking limit. This is when we have to talk about the continuity.

Near a flat point, the two paired points get closer and finally emerge at  $x$ , so the elevation tends to 0 as  $u'$  tends to  $u$ . Therefore, we can simply define the elevation at a flat point to be 0.

At interchange points, however, discontinuity can appear. Figure 5.7 illustrates two interchange points  $y, z$  in the direction  $u$ . When  $u'$  tends to  $u$  from the left, points near  $y$  are paired with points near  $w$ , which are global maximum and minimum in the direction  $u'$ . Thus, going to  $u$  from the left, the elevation approaches to  $|f_u(y) - f_u(w)|$ . However, when  $u''$  tends to  $u$  from the right, points near  $y$  are paired with points near  $x$ , which give birth to and kill 0-cycles. Thus, going to  $u$  from the right, the elevation tends to  $|f_u(y) - f_u(x)|$ , which is different from the left limit. As a result, the limit of the elevation function near  $y$  does not exist. Same circumstance occurs at  $z$ , as illustrated by the pink and blue points in the figure. This discontinuity is caused by the two points  $y, z$  sharing a tangent plane. We note that not all interchange points will cause discontinuity.



**Figure 5.7:** Discontinuity at an interchange point

**Surgery.** The continuity of the elevation function is needed for well define the elevation function. Therefore, we would like to do reversible surgeries on  $M$  at interchange points to get a new manifold  $N$ , where the elevation function is continuous. Specifically, we cut  $M$  along the curves where the function is discontinuous, getting a manifold  $B$  with boundary. Then we glue the boundary of  $B$  by identifying points with the same limits. Figure 5.8 illustrates the process of a surgery on the example above. The curve on the left is a projection of the surface  $M$  on a plane. The left end at  $y$  has the same limit as the right end at  $z$  and so do the right end at  $y$  and the left end at  $z$ . We glue these two pairs together, getting a new disconnected manifold  $N$ . On  $N$ , we can safely define the elevation at singular points by the function values on their neighborhoods, whose limits exist by continuity.

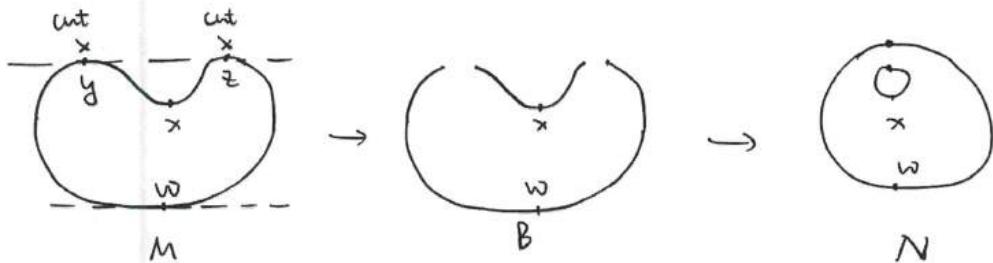


Figure 5.8: The process of a surgery

Now we get the elevation function on a surface  $M$ , defined as the height difference in certain direction between a point and another canonical point on the surface. As we promised at the beginning of this section, this function is an analogue of the height on earth whose local maximums give the cavities and protrusions on the surface. For example, points on a sphere  $S^2 \subseteq \mathbb{R}^3$  are all paired with their antipodal points, with the same elevation, namely the diameter of the sphere. Thus the elevation function on  $S^2$  has no local maximum, indicating the uniform shape of  $S^2$ . If we define the elevation function on the earth surface, it is reasonable to expect mountain peaks and basins are local maximums of the elevation function. Therefore, all we want is the local maximums of the elevation function.

**Local maximums of the elevation function** We will do the computation on the surface  $N$  after surgeries and then put the result back to analyze the shape of the original surface  $M$ .

During the surgeries, we identifies different points on  $M$ . For  $x \in N$ , let  $\mu(x)$  denote the number of its preimages on  $M$ , called the **multiplicity** of  $x$ . By our genericity assumption on  $M$ ,  $\mu(x)$  is at most 3 (when  $x$  is a triple point). Suppose by some algorithm, we have obtained a local maximum pair  $(x, y)$  for  $x, y \in N$ . We further assume that points with multiplicity 3 are only get paired with points with multiplicity 1. These leave us with four types of maximums:

The maximum  $(x, y)$  is called  $\begin{cases} \text{one-legged} & \text{if } \mu(x) = \mu(y) = 1, \\ \text{two-legged} & \text{if } \mu(x) = 1, \mu(y) = 2, \\ \text{three-legged} & \text{if } \mu(x) = 1, \mu(y) = 3, \\ \text{four-legged} & \text{if } \mu(x) = \mu(y) = 2. \end{cases}$

Figure 5.9 shows how the four types of maximums look like on  $M$ .

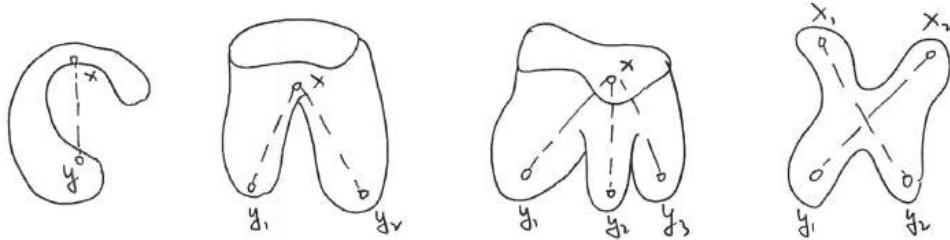


Figure 5.9: From left to right: one-legged, two-legged, three-legged and four-legged maximum on  $M$

When do a pair  $(x, y) \in N^2$  be a local maximum? By definition, the elevations of  $x, y$  are locally the longest. If  $x, y$  have more than one preimages, their tangent planes must be parallel. Geometrically, we have necessary conditions for  $(x, y)$  to be a local maximum.

**Theorem 5.3.2.** *With the notations above, let  $x_1, \dots, x_{\mu(x)} \in M$  and  $y_1, \dots, y_{\mu(y)} \in M$  be the preimages of  $x, y$  respectively. Let  $\mathbf{n}_x$  be the unit normal vector of  $T_x$ . Then the point  $x$  (or  $y$ ) is a maximum of the elevation function only if*

*One-legged case:  $\mathbf{n}_x$  is parallel or anti-parallel to  $y - x$ ;*

*Two-legged case:  $\mathbf{n}_x, y_1 - x$  and  $y_2 - x$  are linearly dependent and the orthogonal projection of  $x$  onto the line  $\overline{y_1 y_2}$  lies between  $y_1$  and  $y_2$ ;*

*Three-legged case: the orthogonal projection of  $x$  onto the plane  $\overline{y_1 y_2 y_3}$  lies inside the triangle  $\Delta y_1 y_2 y_3$ .*

*Four-legged case: the orthogonal projections of the segments  $x_1 x_2$  and  $y_1 y_2$  onto a plane parallel to both intersect.*

*In summary,  $x$  (or  $y$ ) is a local maximum only if  $\mathbf{n}_x$  is a linear combination of the vectors  $y_j - x_i$ .*

**Computation.** To compute the local maximums, as in the computation of the persistent homology, we need first to discretize the surface  $M$ . As before, we choose a triangulation  $K$  of  $M$ , which is a  $\Delta$ -complex and can be expressed combinatorially. For a smooth manifold  $M$  and  $p \in M$ , the critical directions for  $p$  is  $\pm \mathbf{n}_p$ , the normal directions of the tangent plane at  $p$ . For  $K$ , however, there can be too much critical directions. For example, if  $p$  lies on a 2-simplex of  $K$ , then it has a well-defined (up to a sign) normal direction as before; however, if  $p$  is a vertex of  $K$ , then all directions can be said to be critical. The solution is to add another restriction.

**Definition 5.3.3.** Let  $K$  be a triangulation of a surface  $M \subseteq \mathbb{R}^3$ , consisting of vertices, edges and triangles. Let  $\sigma \in K$  be a simplex and  $x \in \sigma$  be a point in  $\mathbb{R}^3$ . Let  $u \in S^2$  be a direction. We say  $x$  is **critical** for the height function  $f_u$  if

(1)  $\langle u, z - x \rangle = 0$  for all points  $z$  of  $\sigma$ , where  $\langle -, - \rangle$  denotes the standard inner product in  $\mathbb{R}^3$ ;

(2) the lower link  $Lk_-(\sigma) := \{\tau \in St_-(\sigma) \mid \tau \cap \sigma = \emptyset\}$  of  $\sigma$  is not contractible to a point (that is, homotopic equivalent to a point).

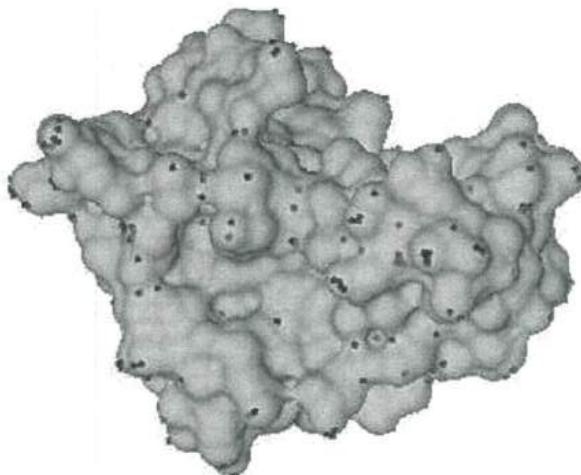
For example, the lower link of a local minimum/maximum is empty or a circle, both not contractible. For further reasons why this is necessary for  $x$  being a critical point, please refer chapter VI.3 of reference<sup>[6]</sup>. To this end, we just admit that the additional condition (2) helps us find reasonable candidates for critical directions for  $x$ .

Let  $N(x_1, x_2, \dots, x_n) \subseteq S^2$  be the set of directions that is critical on all of  $x_1, \dots, x_n$ . The necessary conditions in theorem 5.3.2 translate to:

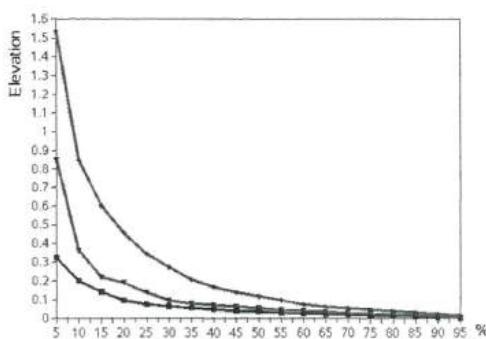
- One-legged case: The candidates for one-legged maximums on  $K$  are pairs of points  $(x, y)$  such that the direction  $(y - x)/\|y - x\|$  is contained in  $N(x, y)$ ;
- Two-legged case: The candidates for two-legged maximums on  $K$  are triplets of points  $(x, y_1, y_2)$  such that the orthogonal projection  $z$  of  $x$  onto the line  $\overline{y_1 y_2}$  lies between  $y_1$  and  $y_2$  and the direction  $(z - x)/\|z - x\|$  is contained in  $N(x, y_1, y_2)$ ;
- Three-legged case: The candidates for three-legged maximums are quadruplets of points  $(x, y_1, y_2, y_3)$  such that the orthogonal projection  $z$  of  $x$  onto the plane  $\overline{y_1 y_2 y_3}$  lies inside the triangle  $\triangle y_1 y_2 y_3$  and the direction  $(z - x)/\|z - x\|$  is contained in  $N(x, y_1, y_2, y_3)$ ;
- Four-legged case: The candidates for four-legged maximums are quadruplets of points  $(x_1, x_2, y_1, y_2)$  such that the shortest line segment  $zw$  connecting the lines  $\overline{x_1 x_2}$  and  $\overline{y_1 y_2}$  touches both the line segments  $x_1 x_2$  and  $y_1 y_2$  and the direction  $(z - w)/\|z - w\|$  is contained in  $N(x_1, x_2, y_1, y_2)$ .

The fact is that after checking for all points, we get a finite set of candidates for each kind of maximums. It remains to check whether these candidates are indeed local maximums. That is, whether the points  $x, y$  on the modified surface  $N$  is paired by the extended persistence algorithm. But this is done in section 5.2, where we reduced the augmented boundary matrix  $\partial$  to find all pairs in the extended persistence. Therefore, after judging all candidate local maximums, we get the genuine local maximums of the elevation function, as well as the information of cavities and protrusions on the surface  $M$ .

**A real-world experiment.** The article<sup>[9]</sup> provides an example on the protein *Ibrs*. The data of this protein surface is downloaded from the protein data bank. *Ibrs* contains 864 atoms excluding hydrogens, which are too small to be resolved by x-ray. The triangulation of the surface has approximately 50 thousand vertices, see figure 5.10.



**Figure 5.10:** Protein 1brs. The marked points are elevation maximums with top 100 elevation values on the surface, with heads dotted in green. Cf<sup>[9]</sup>



**Figure 5.11:** The percentage of maximums with elevation exceeding the threshold on the vertical axis. From top to bottom: the three-legged, four-legged, and two-legged maximum curves. Cf<sup>[9]</sup>.

The table below lists the number of all kinds of elevation maximums. One can see that there are significantly more two-legged type than others. This comes from the specific geometric shape of the protein, including the way that atoms bonding with each other. The one and three legged maximums are less in quantity, but when they appear the maximums tend to have higher elevation value. Figure 5.11 sorts all kinds of maximums in the order of decreasing elevation. We can observe that the fraction of higher elevation maximums of three-legged ones is higher than that of two and four legged ones. Even more, although there are only 5 one-legged maximums, but four of them have elevation more than 5 units. The statistics for other proteins are similar.

#legs	one	two	three	four
#max	5	3617	728	1103

One interesting result is that the elevation maximums are to some extend uniformly distributed on the surface. This contradicts the conjecture that the binding site of a protein may have more or higher maximums. However, this should be considered together with the fact that in many cases the protein pockets identify the location of the binding site.

## Conclusions

We have seen the method of persistent homology and its various applications. There are further topics in the persistent homology and applied topology, such as the stability and structural theorem of persistent homology, discrete Morse theory, graph theory, computation on surfaces and so on. One can refer to<sup>[6]</sup>. The knowledge we introduced about homology, differential manifolds and Morse functions are shallow and without much rigorous discussion. Interested readers may refer to<sup>[2]</sup>,<sup>[4]</sup> and<sup>[5]</sup>.

## References

- [1] MUNKRES J. Featured titles for topology: Topology[M/OL]. Prentice Hall, Incorporated, 2000. <https://books.google.com/books?id=XjoZAQAAIAAJ>.
- [2] LEE J M. Introduction to smooth manifolds[J]. 2000.
- [3] DUMMIT D S, FOOTE R M. Abstract algebra[M]. 3rd ed ed. [S.l.]: Wiley, 2004.
- [4] HATCHER A. Algebraic topology[M/OL]. Cambridge: Cambridge Univ. Press, 2000. <https://cds.cern.ch/record/478079>.
- [5] 姜伯驹. 同调论[M]. [出版地不详]: 北京大学出版社, 2015.
- [6] EDELSBRUNNER H, HARER J. Computational topology - an introduction.[M]. [S.l.]: American Mathematical Society, 2010: I-XII, 1-241.
- [7] PEREA J A. Topological time series analysis[Z]. [S.l.: s.n.], 2018.
- [8] EMRANI S, GENTIMIS T, KRIM H. Persistent homology of delay embeddings and its application to wheeze detection[J/OL]. IEEE Signal Processing Letters, 2014, 21(4):459-463. DOI: 10.1109/LSP.2014.2305700.
- [9] AGARWAL P H H J W Y, Pankaj K. Extreme elevation on a 2-manifold[J]. Discrete Computational Geometry, 2006, 36(3):553-572.

## Acknowledgements

The thesis is a summary of my study on the field of applied topology during my senior year, as well as my undergraduate study and life. It would not be finished without the help of so many kind people. I would like to give them my sincere acknowledgement.

I want to give my thanks to my academic supervisor Prof. Yifei Zhu, who introduced me this topic and gave me much instruction during the study. He also gave me precious advice in my mathematical study and personal growing. He sets an example of a good mathematician and person for all his students and inspires us imperceptibly.

Most of my study of this topic is done in the applied and computational topology seminar. I want to thank all the members of this seminar. We give presentations in turns and discuss together every week. Our website is <https://sustech-topology.github.io/>.

Many professors in the mathematical department of SUSTech gave me valuable advice in both mathematics and other aspects, including but not limited to Prof. Yong Hu, Yannan Qiu, Zhen Zhang, Jitao Wu and Jiang Yang. All professors gave wonderful lectures, as well as the helpful faculties in the mathematical department, deserve my respect and gratitude.

I also would like to thank all good friends of mine in SUSTech and other places, who had great time with me and accompanied me during difficult times. I hope we can be lifetime friends.

I am very grateful to my parents, who are open-minded and interesting people. They have always been supportive and caring to me, without whom I would not be able to pursue further in mathematics.

Finally, I want to thank mathematics for being so elegant and charming. Thank you for being my major and dream.

阮夏冰  
May, 2021