Data Science Methods for Disordered Materials: TDA Lecture Notes

Yossi Bokor Bleile

Department of Mathemathematical Sciences
Aalborg University
Aalborg Øst 9220
Denmark

August 2023

Contents

List of Theorems

Introduction

These are the notes for the two lectures on Topological Data Analysis, as part of the 'Data Science Methods for Disordered Materials' PhD course at Aalborg University in August 2023.

A big question most people have at the start of a series of lectures on a topic is 'What is this about and why should I care?', and to be perfectly honest, there is no easy answer to the 'why should I care' part.

Topological data analysis (TDA) concerns itself with extracting, analysing, inferring and exploiting topological and geometric information from data. These methods are well-founded in mathematics and statistics, with lots of results relating to stability and reconstruction. In these two lectures, we will introduce one of the most common tools in TDA, persistent homology, and present some ways in which it can be used to summarise and compare the shape of data.

TDA has been used in a variety of settings, including:

1. understanding neural firing patterns,
2. differentiating between different types/groups/lineages of biological objects (bones/leaves for example),
3. identifying sub-populations,
4. learning structures underlying spatial data,
5. exploring the relationship between the structure of materials and their physical properties.

Assumed knowledge includes basic linear algebra and some things about functions.

Lecture 1

Ground work

A lot of the work in TDA is done by examining the topological and geometric properties of complexes, representing the data you wish to examine. Roughly speaking, a *complex* is an object built out of pieces that fit together in very specific ways, think of Lego. There are two main types of complexes you are likely to encounter:

- simplicial, and
- cellular, also called CW.

In these lectures, we will restrict to simplicial complexes, which can be thought of as higher dimensional generalisations of graphs. The basic building blocks for simplicial complexes are *simplices*.

Definition 1.1 (Simplex)**.** Let $V_\sigma \subset \mathbb{R}^n$ be a finite set of points, say $V = \{v_0, \ldots, v_d\}$, $d \leq n$ which are in general position. The *geometric simplex* σ with vertices V_σ is the convex hull of V. We say it has dimension d.

We can also consider V_σ as an abstract set, and then we can define a *abstract simplex* σ, of dimension d.

In both cases, we call a simplex τ with vertex set $V_\tau \subset V_\sigma$ a *face* of σ, and write $\tau \subset \sigma$.

Sometimes, we will write $[v_0, \ldots, v_d]$ instead of σ, in particular for dimension 0 simplices.

Now that we have defined the building blocks we use to construct simplicial complexes, we can formally define them.

Definition 1.2 (Simplicial complex)**.** Let $V \subset \mathbb{R}^n$ be a finite set of points. A *geometric simplicial complex* X with vertex set V is a finite set of simplices σ with $V_\sigma \subseteq V$ such that

1. $\forall v \in V$ we have $[v] \in X$,
2. $\forall \sigma \in X$ and $\tau \subset \sigma$ we have $\tau in X$.

Example 1.3. For low dimensional simplices, we have names for them:

1. a 0-simplex is a point,
2. a 1-simplex is a line,
3. a 2- simplex is a triangle, and
4. a 3-simplex is a tetrahedron.

Example 1.4. Any graph is a simplicial complex.

Now that we have defined simplicial complexes, we can ask what structure does it have in a given dimension p. In particular, we are interested in the *non-deformable* structure. To do this, we will examine *groups of chains* in dimensions $p+1, p, p-1$ and the relationship between $p+1$ and p chains, as well as between p and $p-1$ chains. And so, we define p-chain groups.

Definition 1.5 (Group of p-chains)**.** Given a simplicial complex X, we can form the *group of p-chains* under formal sums with coefficiants in $\mathbb{Z}$, $(X, C_p, +, \mathbb{Z}) = (C_p, +, \mathbb{Z})$, where $+$ is a binary operation

$$+ : C_p \times C_p \to C_p$$

satisfying:

1. $\forall \sigma, \tau \in C_p$, $\sigma + \tau \in C_p$,
2. $\forall \sigma, \tau, \gamma \in C_p$, $(\sigma + \tau) + \gamma = \sigma + (\tau + \gamma)$,
3. $\exists 0 \in C_p$ such that $\forall \sigma \in C_p$, $0 + \sigma = \sigma = \sigma + 0$,
4. $\forall \sigma \in C_p$, $\exists \sigma^{-1}$ such that $\sigma + \sigma^{-1} = 0 = \sigma^{-1} + \sigma$, often write $-\sigma$ instead of σ^{-1}.

We often just write $(C_p, +)$, where it is clear from context, or when the underlying space and coefficients are not explicitly required.

Exercise 1.6.
1. Show that 0 is unique.
2. Show that σ^{-1} is unique.

Remark. The group of p-chains $(C_p, +)$ really is a group, and it is generated by the p-simplices of X.

Okay, now that we have defined p-chain groups (and hence $p+1$- and $p-1$-chain groups), we introduce the tools needed for us to explore some relationships between them. We begin with maps between chain groups.

Definition 1.7 (Maps between chain groups)**.** Given $(C, +_C), (D, +_D)$ two groups of chains (could be from different simplicial complexes and different p's) a *homeomorphism*

$$h : (C, +_C) \to (D, +_D)$$

is a map from C to D which respects $+_C$ and $+_D$. That is,

1. $\forall \sigma, \tau \in C$, $h(\sigma +_C \tau) = h(\sigma) +_D h(\tau)$,
2. $h(0_C) = 0_D$,
3. $\forall \sigma \in C$, $h(\sigma^{-1}) = h(\sigma)^{-1}$.

For any such map, we can look at two sets related to it:

I the kernel: $\ker(h) := \{\sigma \in C \mid h(\sigma) = 0_D\}$,

II the image: $\text{im}(h) := \{h(\sigma) \mid \sigma \in C\} = \{\tau \in D \mid \exists \sigma \in C \text{ s.t. } h(\sigma) = \tau\}$.

The map (or homeomorphism) we use to understand the relationship between $p+1$- and p-chains and p- and $p-1$-chains is the *boundary operator*, which takes a p-chain σ and gives the $p-1$-chain that forms its boundary.

Definition 1.8 (Boundary operator)**.** Let X be a simplicial complex, let $(C_p, +), (C_{p-1}, +)$ be the p- and $p-1$- chain groups respectively. Then the *dimension p boundary operator* $\partial_p : (C_p, +) \to (C_{p-1}, +)$ is the map defined on a p-simplex $\sigma = [v_0, \ldots, v_p]$ by

$$\partial_p(\sigma) := \sum_{i=0}^{p} (-1)^i [v_0, \ldots, \widehat{v_i}, \ldots, v_p],$$

where $\widehat{v_i}$ means that the i^{th} vertex has been removed. We then extend this definition linearly to C_p.

Example 1.9.

 ©2023

We now have everything we need to explore the relationship between chains via the boundary operators ∂_p. In particular, we consider kernels $\ker(\partial_p)$ and images $\text{im}(\partial_p)$, which allow us to define two classes of p-chains.

Definition 1.10 (p-cycles and p-boundaries)**.** Consider a p-chain $c \in C_p$. We say c is a

1. p-cycle if $\partial_p(c) = 0_{C_{p-1}}$ or $c \in \ker(\partial_p)$,
2. p-boundary if $\exists c' \in C_{p+1}$ with $\partial_{p+1}(c') = c$ or $c \in \text{im}(\partial_{p+1})$.

Topology can be thought of as the study of shape up to some nice, smooth changes, and so we need to say what kind of changes we ignore. We will use *homologous* or up to *homology* as our notion.

Definition 1.11 (Homologous cycles)**.** Given p-chains γ, γ', we say they are *homologous*, $\gamma \sim \gamma'$ if $\gamma - \gamma' \in \ker(\partial_{p+1})$. We say $[\gamma] := \{\gamma' \text{ s.t } \gamma \sim \gamma'\}$ the coset/equivalence class/homology class of γ.

We can now formally define the p-homology group of a simplicial complex X.

Definition 1.12 (Homology groups)**.** Let X be a simplicial complex, the simplicial p-homology group $H_p(X)$ is the group of equivalence classes of p-chains $\gamma on C_p(X)$. That is, $H_p(X)$ is the quotient of $\ker(\partial_p)$ by $\text{im}(\partial_{p+1})$:

$$H_p(X) = {}^{\ker(\partial_p)}\!/_{\text{im}(\partial_{p+1})}$$

Often, we write $H_\bullet(X)$ to refer to the collection of homology groups of X, and the maps between them induced by the boundary operator.

Great, we are now ready to proceed and think about homology groups of simplicial complexes.

!!!¡¡¡ WARNING ¡¡¡!!!

From here forward, we assume we are working over ${}^{\mathbb{Z}}\!/_{2\mathbb{Z}}$. This means we do not have to worry about torison, and our algorithms are more efficient.

Now that we have defined homology groups for a simplicial complex X, we seek a method for computing (at least the dimension of) the homology groups. One way is to find the Smith Normal form of the boundary matrix,

and then use rank-nullity to obtain the dimension of the homology groups. Recall that for any $m \times n$ matrix A (over a principal ideal domain), there are $m \times m$ and $n \times n$ matrices S and T such that SAT is a diagonal matrix of the form:

$$\begin{pmatrix} \alpha_1 & & & & & & & \\ & \alpha_2 & & & & & & \\ & & \alpha_3 & & & & & \\ & & & \ddots & & & & \\ & & & & \alpha_r & & & \\ & & & & & 0 & & \\ & & & & & & \ddots & \\ & & & & & & & 0 \end{pmatrix}$$

with $\alpha_i \mid \alpha_{i+1}$. This can be obtained using elementary column operations.

So, to find the dimension of the homology groups, we can find the Smith Normal form for each ∂_p, and then read off the information about the rank of $\ker(\partial_p)$ and $\mathrm{im}(\partial_p)$, which will then give us the dimension of the homology groups:

$$\dim(H_p) = \dim(\ker(\partial_p)) - \dim(\mathrm{im}(\partial_{p+1})).$$

We end this lecture by introducing what we are all here for:

persistent homology.

We will only do it in terms of simplicial complexes and with coefficients in $\mathbb{Z}/2\mathbb{Z}$, but it can be done in more generality.

First, we define *filtered* simplicial complexes.

Definition 1.13 (Filtered simplicial complex)**.** A *filtered simplicial complex* (X, h) consists of a simplicial complex X and a function $h : X \to \mathbb{R}$ (you can replace R by any poset) such that for each $\alpha \in \mathbb{R}$,

$$X_\alpha := h^{-1}((-\infty, \alpha]) = \{\sigma \in X \mid h(\sigma) \in (-\infty, \alpha]\}$$

is a proper subcomplex. We often write $\{X_\alpha\}$, leaving the h implied. Note that for all $\alpha \leq \beta \in \mathbb{R}$, we have an inclusion map

$$i_\alpha^\beta : X_\alpha \to X_\beta.$$

 ©2023

Persistent homology tracks how the homology of X_α changes as α varies.

Definition 1.14 (Persistent homology of a filtered simplicial complex)**.** Given a filtered simplicial complex $(X, h) = \{X_\alpha\}$, the p^{th} persistent homology of $\{X_\alpha\}$ is the sequence of homology groups

$$\to H_p(X_\alpha) \xrightarrow{j_\alpha^\beta} H_p(X_\beta) \to$$

with the maps j_α^β induced by i_α^β, that is

$$j_\alpha^\beta\left([\gamma]\right) = \left[i_\alpha^\beta(\gamma)\right].$$

Lecture 2

Summaries and methods

We ended Lecture 1 with the definition of persistent homology for a filtered simplicial complex, but we are still lacking a way of calculating it. To do this, we will introduce a structured boundary matrix for $\{X_\alpha\}$, and then reduce it to read off the persistence.

Definition 2.1 (Boundary matrix for $\{X_\alpha\}$)**.** Let $X = (X, h) = \{X_\alpha\}$ be a filtered simplicial complex with n simplices in it. Order the simplices such that for all p, the p-simplices come before $p+1$-simplices, and for all $\alpha \leq \beta$ the simplices of X_α come before the simplices of X_β. We now refer to the simplices by their index in this ordering. Then, the *boundary matrix* $B_X = B$ is the $n \times n$ matrix with

$$B_{i,j} = \begin{cases} 1 & \text{if } \sigma_i \subset \sigma_j \text{ and } \dim(\sigma_i) = \dim(\sigma_j) - 1 \\ 0 & \text{otherwise} \end{cases}$$

Now that we have a framework we can use to explore the shape/structure of simplicial complexes, we are going to look at ways we can build simplicial complexes from data, and how we can visualise and summarise this information.

Example 2.2.

So, we now (finally) have a way to read off the persistence cycles of a filtered simplicial complex, but we don't have a nice way of visualising them. One convenient way to present information about the persistent homology of a filtered simplicial complex is a persistence diagram. However, we need some terminology to describe the changes in homology. Given a filtered simplicial complex $\{X_\alpha\}$, let $\mathcal{H}_p(X) := \bigcup_\alpha H_p(X_\alpha)$ be the set of all homology

classes of $\{X_\alpha\}$. For any $[\gamma] \in \mathcal{H}_p(X)$, we say it is *born* at time b if it first appears in $H_p(X_b)$, and we say it *dies* at time d if it becomes the boundary of some $p+1$-chain in X_d. We call (b, d) the birth-death pair of $[\gamma]$. Note, $d = \infty$ is possible, and we call such homology classes *essential classes*.

We can now define *persistence diagrams*.

Definition 2.3 (Persistence Diagrams)**.** Let $X = \{X_\alpha\}$ be a filtered simplicial complex. The p^{th} *persistence diagram* $D_p(X)$ of X is the multiset of points in $\mathbb{R}^2$ consisting of the birth-death pairs $(b_{[\gamma]}, d_{[\gamma]})$ for all $[\gamma] \in \mathcal{H}_p(X)$.

Example 2.4.

Okay, great, given a filtered simplicial complex X, we have a way of calculating and visualising its persistent homology. Now lets talk about some things we can use to compare persistence diagrams:

1. Wasserstein distances
2. rank functions
3. accumulated persistence functions
4. persistence images
5. persistence landscapes

We will define the Wasserstein distance, rank functions and accumulated persistence diagrams. Persistence images are intended to use with machine learning/computer vision methods.

Definition 2.5 (Wasserstein distance)**.** Take two persistence diagrams D_1, D_2, and let Δ be the multiset of points containing infinitely many copies of $(x, x) \forall x \in \mathbb{R}$. A *matching* between D_1 and D_2 is a bijective map

$$m : D_1 \cup \Delta \to D_2 \cup \Delta.$$

Then, the *q-Wasserstein distance* $W_q(D_1, D_2)$ between D_1 and D_2 is

$$W_q(D_1, D_2) := \inf_{\text{over all matchings } m} \left(\sum_{x \in D_1 \cup \Delta} \|x - m(x)\|^q \right)^{1/q}$$

!!!¡¡¡ WARNING ¡¡¡!!!

Two very different diagrams can be close in the Wasserstein distance.

Fact. Persistence diagrams are stable with respect to the Wasserstein distance: consider two functions $f, g : X \to \mathbb{R}$ on a simplicial complex X, and let D_f, D_g bethe p-persistence diagrams of $(X, f), (X, g)$ respectively. Then if

$$\|f - g\|_\infty \leq \varepsilon$$

we have

$$W_q(D_f, D_g) \leq \varepsilon.$$

Unfortunately, the Wasserstein distance is not always good for statistical analysis. So, we will define two other options: rank functions and accumulated persistence functions.

Definition 2.6 (Rank function)**.** Let D be a persistence diagram. Its *rank function* $RK_D : \mathbb{R}^2 \to \mathbb{N}$ is the function defined by

$$RK_D(x, y) = \begin{cases} \sum_{(b,d)\in D} \mathbb{1}(b < x, d > y) & x \leq y \\ 0 & x > y \end{cases}.$$

Definition 2.7 (Accumualted persistence function)**.** Let D be a persistence diagram. Its *accumulated persistence function* $APF_D : \mathbb{R}^2 \to \mathbb{N}$ is the function defined by

$$rk_D(x) = \sum_{(b,d)\in D} (d - b)\mathbb{1}\left(\frac{b + d}{2} \leq x\right).$$

Using either of these summaries, we can construct (dis)similarity matrices and do principal component analysis.

There is one big thing we have not really discussed yet: how to obtain a filtered simplicial complex from data! There are a few different ways, including:

1. Vietoris-Rips complex - d-simplex $[v_0, \ldots, v_d]$ has filtration value δ if $\forall i, j\ B_\delta(v_i) \cap B_\delta(v_j) \neq \emptyset$,
2. Čech complex: d-simplex $[v_0, \ldots, v_d]$ has filtration value δ if $\bigcap_i B_\delta(v_i) \neq \emptyset$,

 ©2023

3. α-complex: d-simplex $[v_0, \ldots, v_d]$ filtration value δ if there is a spehere of radius δ which has $\{v_0, \ldots, v_d\}$ on its boundary but contains no other points in the data set.