MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Diplomová práce

BRNO 2025 TOMÁŠ PETIT

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Topological data analysis

Diplomová práce

Tomáš Petit

Vedoucí práce: prof. RNDr. Jan Slovák, DrSc. **Brno 2025**

Bibliografický záznam

Autor:	Bc. Tomáš Petit Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Topological data analysis
Studijní program:	Matematika
Studijní obor:	Matematika
Vedoucí práce:	prof. RNDr. Jan Slovák, DrSc.
Akademický rok:	2024/2025
Počet stran:	?? + ??
Klíčová slova:	Topologie; Algebraická Topologie; Homologie; Persistentní Homologie; Topologická analýza dat; TDA; Topologické Metody

Bibliographic Entry

Author:	Bc. Tomáš Petit Faculty of Science, Masaryk University Department of mathematics and statistics
Title of Thesis:	Topological data analysis
Degree Programme:	Mathematics
Field of Study:	Mathematics
Supervisor:	prof. RNDr. Jan Slovák, DrSc.
Academic Year:	2024/2025
Number of Pages:	?? + ??
Keywords:	Topology; Algebraic Topology; Homology; Persistent Homology; Topological data analysis; TDA; Topological Methods

Abstrakt

V této bakalářské/diplomové/rigorózní práci se věnujeme ...

Abstract

In this thesis we study ...

MUNI
SCI

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2024/2025

Ústav:	Přírodovědecká fakulta
Student:	Bc. Tomáš Petit
Program:	Matematika
Specializace:	Matematika

Ředitel *ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

Název práce:	Topological data analysis
Název práce anglicky:	Topological data analysis
Jazyk závěrečné práce:	angličtina

Oficiální zadání:
Goal: The goal is to understand the concepts and tools of Topological Data Analysis, and to be ready to use them in practical tasks. Aim: Depending on the results of the initial period, the student will either focus on theoretical understanding and original research in Mathematics and Statistics, or the focus will be on smart use of advanced tools in solving practical problems, including the implementation issues. One of the resources for real data requiring sophisticated analysis will come form the project Machine Learning in Nanomaterial Biocompatibility Assessment (MUNI/G/1125/2022).

Literatura: RAÚL RABADÁN, ANDREW J. BLUMBERG, Topological Data Analysis for Genomics and Evolution, Cambridge University Press, 2020, DOI: 10.1017/9781316671665

Vedoucí práce:	prof. RNDr. Jan Slovák, DrSc.
Datum zadání práce:	20. 9. 2023
V Brně dne:	25. 7. 2024

Zadání bylo schváleno prostřednictvím IS MU.

Bc. Tomáš Petit, 16. 10. 2023
prof. RNDr. Jan Slovák, DrSc., 17. 10. 2023
RNDr. Jan Vondra, Ph.D., 18. 10. 2023

Poděkování

Na tomto místě bych chtěl(-a) poděkovat ...

Prohlášení

Prohlašuji, že jsem svoji bakalářskou/diplomovou práci vypracoval(-a) samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Prohlašuji, že jsem svoji rigorózní práci vypracoval(-a) samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno xx. měsíce 20xx

.....

Tomáš Petit

Contents

List of used notation

Pro snažší orientaci v textu zde čtenáři předkládáme přehled základního značení, které se v celé práci vyskytuje.

$\mathbb{C}$ množina všech komplexních čísel
$\mathbb{R}$ množina všech reálných čísel
$\mathbb{Z}$ množina všech celých čísel
$\mathbb{N}$ množina všech přirozených čísel
$\mathbb{C}$ množina všech komplexních čísel
$\mathbb{R}$ množina všech reálných čísel
$\mathbb{Z}$ množina všech celých čísel
$\mathbb{N}$ množina všech přirozených čísel
$\mathbb{C}$ množina všech komplexních čísel
$\mathbb{R}$ množina všech reálných čísel
$\mathbb{Z}$ množina všech celých čísel
$\mathbb{N}$ množina všech přirozených čísel
$\mathbb{C}$ množina všech komplexních čísel
$\mathbb{R}$ množina všech reálných čísel
$\mathbb{Z}$ množina všech celých čísel
$\mathbb{N}$ množina všech přirozených čísel
$\mathbb{C}$ množina všech komplexních čísel
$\mathbb{R}$ množina všech reálných čísel
$\mathbb{Z}$ množina všech celých čísel

Introduction

To add later.

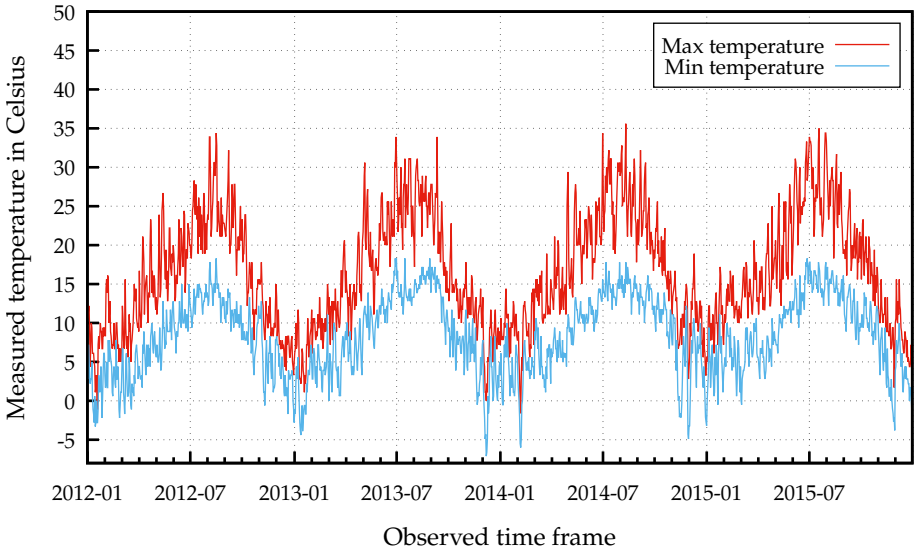
Chapter 1

Why Topology?

Data has shape. This is hardly a new or revolutionary idea in the realm of data analysis and statistics. It is an assumption that we make all the time, even if we do not say it out loud. Whenever one tries to construct a linear regression model, we all have the mental image of a straight line in our minds, which should roughly approximate the data. This is then generalized via hyperplanes in higher dimensions.

Another example would be periodic time series or signals – we all expect to see a "loop" of some sort, given a long enough time interval between the measurements, see for example 1.1. It isn't much of a stretch to imagine that we could use this loop to try to approximate the period of the time series (something that we will actually do in the following chapters, after introducing the necessary mechanisms).

Figure 1.1: Example plot of seasonal temperature changes in Seattle throughout the years.

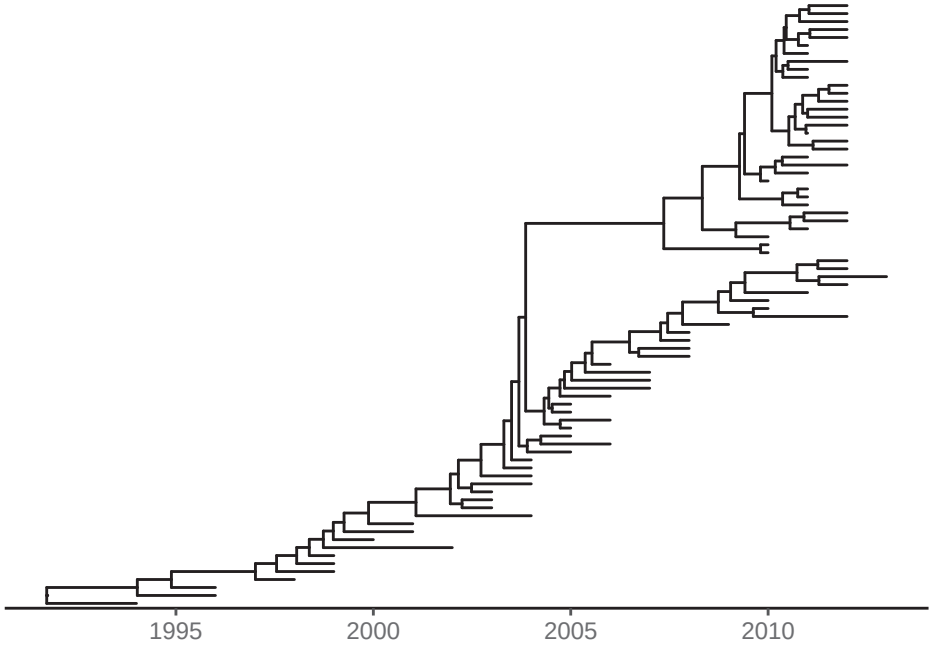


Clustering algorithms – be it k-means, hierarchical clustering and so on – all

work with the shape and geometry of our data explicitly by partitioning the available search space into the distinct clusters, once a measure of distance (which doesn't need to be a metric, per se) is chosen. The list could go on, but I believe the point was already made. Historically and traditionally, an appropriate analytic model was derived and constructed for each of the above-mentioned methods with a rich theoretical background to justify the results.

While this approach works for most simple cases the average data analyst will encounter on an Excel spreadsheet, thanks to the advancements made in computing power and software engineering, we're collecting massive amounts of complicated, high-dimensional data living faster than we did before, especially in the fields of biology and medical sciences.

Figure 1.2: Time-scaled phylogenetic tree of H3 influenza viruses inferred by BEAST using molecular clock model.



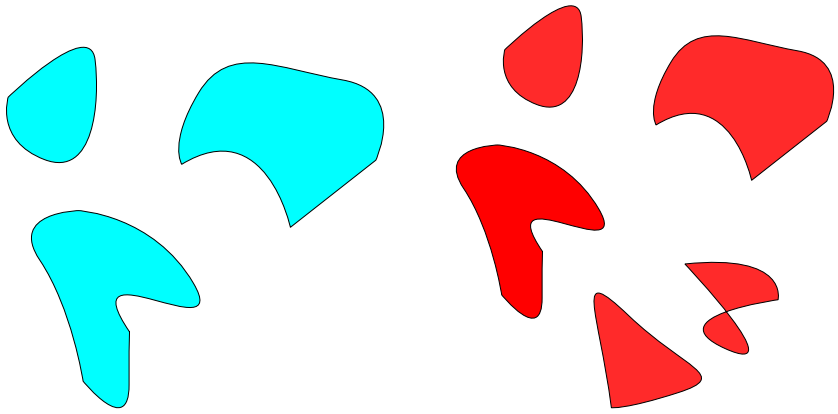
A good example of that would be phylogenetic and evolutionary trees, like the one in 1.2. We might be interested to know whether any mutation occurred, where did they happen and how far were they transmitted down the tree. We might look for re-combinations of the genomic material or both horizontal and vertical transfer of it. All those questions pertain to the shape of the phylogenetic tree and its branching. [1]

The situation only gets more complex with each passing day and batch of collected data. One *could* try to develop analytical models for each case, provide the rigorous theory and obtain the conclusions that they seek. But a far more reasonable and general method would be to instead study the shape itself and approximate the datasets by selecting the shape that describes it the "best". This is

[1] Data downloaded from the following link, visited on the 01.08.2024

where topology comes into play, as this gives us the tools and theory to do exactly that, with the help of algebraic topology.

Algebraic topology will help us to qualitatively distinguish the two situations in 1.3, where even on an intuitive level we can see that the difference lies in the number of "blobs" in each figure; or more rigorously in the number of connected components.

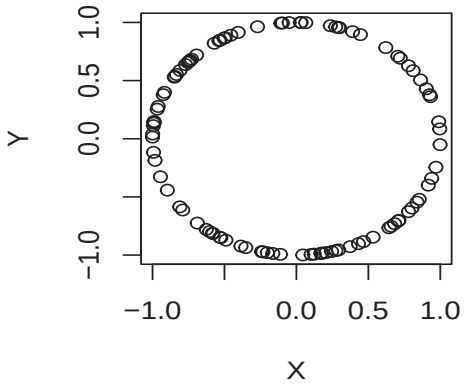


(a) 3 blobs (b) 5 blobs

Figure 1.3: Counting the number of blobs in each figure.

Likewise, using the already established machinery, we will be able to describe and quantify the fact that in 1.4 there is a hole in the middle of it.

Figure 1.4: Data sampled from a unit circle characterized by the distinct hole in the middle.



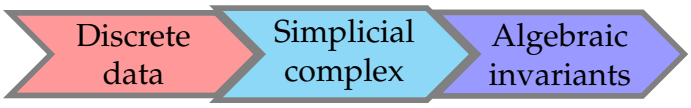
Both of these can be thought of as counting $n-$dimensional holes: connected components being of dimension 0, while the hole in 1.4 is a $1-$dimensional one. We will see how this approach and its extension will be the foundation of what we call TDA – Topological Data Analysis.

Chapter 2

Simplicial Complexes and Homology

The goal of this and the following chapters is to establish and set up the pipeline for extracting the algebraic invariants of our data. Usually, we can only work with sampled and discrete data coming from some set of measurements. As such, we can't directly use methods of algebraic topology, since we won't typically be working with discrete topological spaces and to properly use these methods, we would need an uncountable amount of data; something that isn't feasible from a computational point of view.

This forces us to use different methods to somehow approximate and recover the topology of the ambient space given only a finite set of points. Secondly, we also need to consider the *scale* of the data – some interesting properties may be more apparent only after we "zoom" in closely on them, some may not become apparent at all. All in all, we will construct the following pipeline:



and repeat this step for all scales at once, effectively measuring the evolution of the algebraic invariants through the changes in the feature scale.

2.1 Simplicial complexes

Definition 2.1.1 (Simplex)**.** For $k \geq 0$, a k-simplex σ of dimension k in a Euclidean space $\mathbb{R}^n$ is the convex hull of a set P of $(k+1)$ affinely independent points in $\mathbb{R}^n$. For $0 \leq m \leq k$, an m-face of σ is an m-simplex that is the convex hull of a nonempty subset of P. A *proper face* of σ is a simplex that is the convex hull of a proper subset of P (any face except σ). $(k-1)$ faces of σ are called *facets* of σ.

Typically, we refer to a 0-simplex as a *vertex*, a 1-simplex as an *edge*, a 2-simplex as a *triangle* and so on. An illustration of those can be seen in 2.1.

Definition 2.1.2 (Geometric simplicial complex)**.** A *geometric simplicial complex* K is a set with finitely many simplices that satisfy the following:

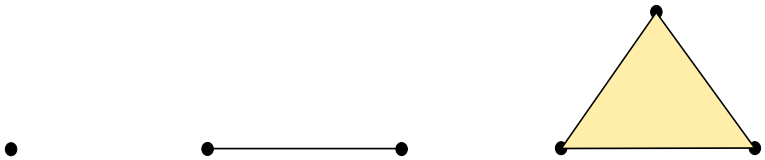


Figure 2.1: From the left: a 0-simplex, a 1-simplex and a 2-simplex

- K contains every face of each simplex in K.
- For any two simplices $\sigma, \tau \in K$, their intersection $\sigma \cap \tau$ is either empty or a face or both σ and τ.

This is also known as a *triangulation*, where the *dimension* k of K is the maximum dimension of any simplex in K. The two definitions above are highly geometric and easy to visualize and imagine. The next definition is more technical and abstract but nonetheless important.

Definition 2.1.3 (Abstract simplex)**.** A collection K of non-empty subsets of a given set $V(K)$ is an *abstract simplicial complex*, if every element $\sigma \in K$ has all of its non-empty subsets $\sigma' \subseteq \sigma$ also in K. Each element σ with a cardinality $|\sigma| = k + 1$ is called a k-simplex and each of its subsets $\sigma' \subseteq \sigma$ with $|\sigma'| = k' + 1$ is called a k'-face. Finally, a $(k-1)$-face of a k-simplex is called its *facet*.

Remark. One could also dually define a k-coface, cofacet and its codimension but it's not terribly important.

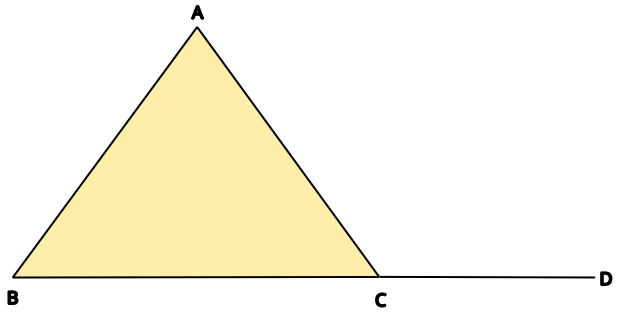


Figure 2.2: A simplicial complex with 4 vertices, 4 edges and 1 triangle.

A geometric simplicial complex K in $\mathbb{R}^n$ is called a *geometric realization* of an abstract simplicial complex K', if and only if there is an embedding $e : V(K') \to \mathbb{R}^n$, that takes every k-simplex $\{v_0, \ldots, v_k\}$ in K' to a k-simplex in K that is the convex hull of $e(v_0), \ldots, e(v_k)$. An example is shown in 2.2 as this is the geometric realization of the abstract complex with vertices A, B, C, D, edges $\{A, B\}, \{A, C\}$, $\{B, C\}, \{C, D\}$ and 1 triangle $\{A, B, C\}$.

Definition 2.1.4 (Underlying space)**.** The *underlying space* of an abstract simplicial complex K, denoted by $|K|$, is the pointwise union of its simplices in its geometrical realization, i.e., $|K| = \bigcup_{\sigma \in K} |\sigma|$, where $|\sigma|$ is the restriction of this realization on σ. If K is geometric, then its geometric realization can be taken as itself.

Unless it is considered necessary, we won't be making the distinction between the two due to this equivalence between geometric and abstract simplicial complexes.

Definition 2.1.5 (k-skeleton)**.** For any $k \geq 0$, the k-skeleton of a simplicial K complex, denoted by K^k, is the subcomplex formed by all simplices of dimension at most k.

Given this, in 2.2, the 1-skeleton consists of the vertices A, B, C, D and the edges joining those.

2.2 Nerves, Čech and Rips complexes

Given any open cover of a topological space, we are able to construct a simplicial complex on top of it. As we'll see, there isn't only one kind of complex we can build, depending on the properties we're looking for and its size, which has to be considered whenever we talk about any software implementation of the algorithms.

Definition 2.2.1 (Nerve)**.** Given a finite collection of sets $\mathfrak{U} = \{U_\alpha\}_{\alpha \in A}$, we define the *nerve* of the set $\mathfrak{U}$ to be the simplicial complex $N(\mathfrak{U})$, whose vertex set is the index set A, and where a subset $\{\alpha_0, \ldots, \alpha_k\} \subseteq A$ spans a k-simplex in $N(\mathfrak{U})$ if and only if $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$.

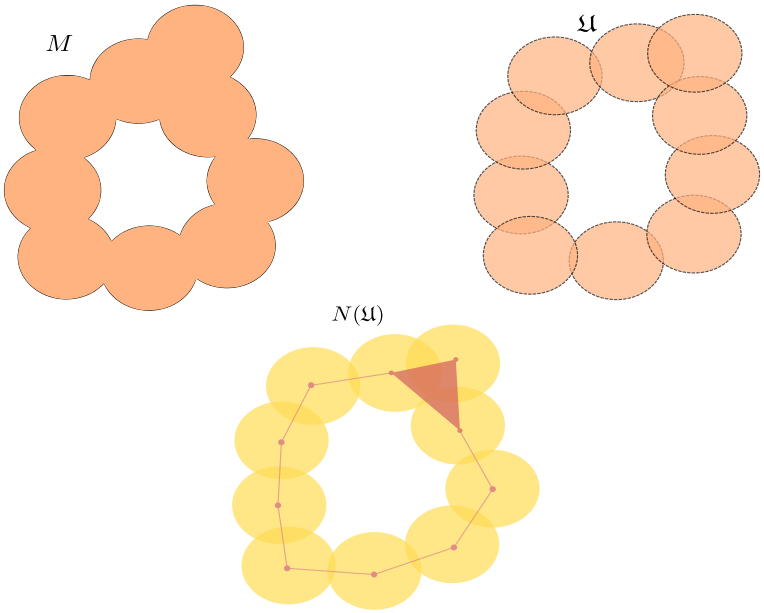


Figure 2.3: An example of a space M, its open cover $\mathfrak{U}$ and its nerve $N(\mathfrak{U})$.

The following important theorem about nerves tells us when are the nerves "equivalent" to the original space. There are various formulation of this statement but since we are working primarily with finite metric spaces, we'll adopt the appropriate version for it.

Theorem 2.2.1 (Nerve theorem)**.** Given a finite cover $\mathfrak{U}$ (open or closed) of a metric space M, the underlying space $|N(\mathfrak{U})|$ is homotopy equivalent to M, if every non-empty intersection $\cap_{i=0}^{k} U_{\alpha_i}$ of cover elements is homotopy equivalent to a point, i.e., contractible.

For those interested in a proof of this statement, see [Bor48] for example. From this we can see, that the nerve is homotopy equivalent to M in 2.3. Now we can finally present the first construction of an abstract simplicial complex using the concept of a nerve, given a finite subset P of a metric space (M, d).

Definition 2.2.2 (Čech complex)**.** Let (M, d) be a metric space and P a finite subset of it. Given a real $r > 0$, the Čech complex $\mathbb{C}^r(P)$ is defined to be the nerve of the set $\{B(p_i, r)\}$, where

$$B(p_i, r) = \{x \in M \mid d(p_i, x) \leq r\}.$$

One can easily deduce that if M happens to be a Euclidean metric space, according to Theorem 2.2.1, the Čech complex will be homotopy equivalent to the space of union of the balls. The Čech complex has nice theoretical properties, but the one predominantly implemented in most software packages is the Vietoris-Rips complex bellow.

Definition 2.2.3 (Vietoris-Rips complex)**.** Let (P, d) be a finite metric space. Given a real $r > 0$, the Vietoris-Rips (VR for short) complex is the abstract simplicial complex $\mathbb{VR}^r(P)$, where a simplex $\sigma \in \mathbb{VR}^r(P)$, if and only if $d(p, q) \leq 2r$ for every pair of vertices of σ.

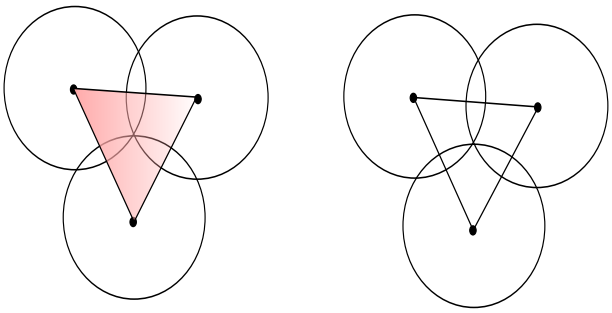


Figure 2.4: On the left, the VR complex of three pairwise intersecting triangles. On the right, the corresponding Čech complex.

A simple example, where we can see the difference between the Čech and VR complex can be seen on 2.4. For an interactive visualization of the two, the reader may want to access the following sites - [Sus24] and [Nat24]. With the movement of the sliders there, the reader is introduced to another important concept that we will make use of - changing the radius r and constructing a growing sequence of either Čech or VR complexes.

Nevertheless, we don't have to worry too much about the differences between the two, given the following result.

Theorem 2.2.2. Let P be a finite subset of a metric space (M, d). Then

$$\mathbb{C}^r(P) \subseteq \mathbb{V}^r(P) \subseteq \mathbb{C}^{2r}(P). \tag{2.1}$$

2.3 Sparse complexes

While the VR complex is the one you will usually encounter in most talks about TDA, it has a size problem. More often than not, both the Čech and VR complexes grow too large, even in low dimensions. A VR complex constructed out of a few thousand points can easily have millions of triangles. A short example of this behaviour can be seen in 2.5. As such, sometimes it is computationally less expensive to use more sparse alternatives instead. [1]

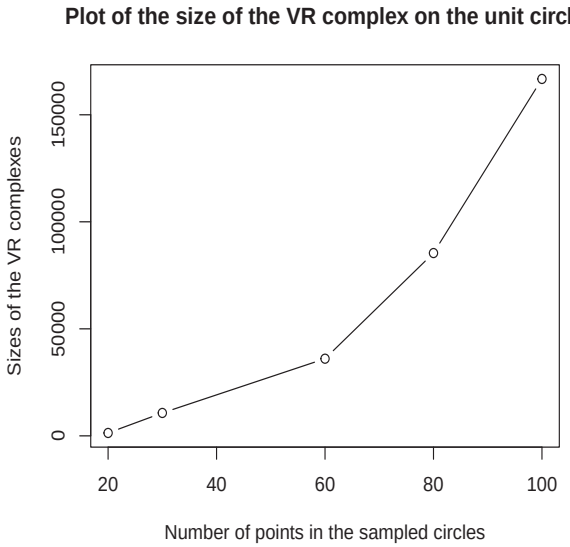


Figure 2.5: An illustration of the growth of the VR complex when we increase the number of sampled points.

2.3.1 Delaunay complex

This complex is usually used in various applications, such as mesh generation or 3D triangulation (for example, see the following). However, it remains computationally expensive in dimensions greater than 3, and other complexes are preferred instead.

Definition 2.3.1 (Delaunay complex)**.** Let P be a finite point set in R^n. A k-simplex σ is called *Delaunay*, if its vertices are in P and there is an open n-ball whose boundary contains the boundary of this ball. A *Delaunay complex* of P, denoted by Del P, is the simplicial complex with vertices in P, in which every simplex is Delaunay and $|$ Del $P|$ coincides with the convex hull of P.

An example of a Delaunay complex can be seen on 2.6. The Delaunay complex is dual to another construction that you may or may not encounter in the wild - the Voronoi diagram.

[1] Technically, the graph is based on the Rips filtration, a term we'll see in the future although it doesn't change the point presented here.

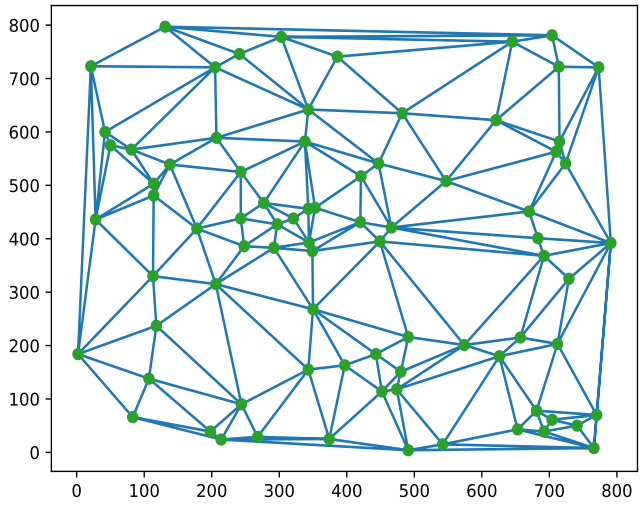


Figure 2.6: Delaunay complex on a sample of randomly generated points.

Definition 2.3.2 (Voronoi diagram)**.** Given a finite point set $P \subset \mathbb{R}^n$ in generic position, the Voronoi diagram Vor(P) of P is the tessellation of the embedding space $\mathbb{R}^n$ into convex cells V_p for every $p \in P$ where

$$V_p = \{x \in \mathbb{R}^n \mid d(x,p) \leq d(x,q), \forall q \in P\}.$$

Additionally, a k-face of Vor(P) is the intersection of $(d-k+1)$ Voronoi cells.

The duality between a Delaunay complex and Voronoi diagram is better expressed through this little theorem.

Theorem 2.3.1. For $P \in\subset \mathbb{R}^n$, Del($P$) is the nerve of the set of Voronoi cells $\{V_p\}_{p\in P}$, which is a closed cover of $\mathbb{R}^n$.

More specifically, a Delaunay k-simplex in Del(P) is dual to a Voronoi $(d-k)$-face in Vor(P). That duality can be seen on 2.7. [2] The reasons why Delaunay complexes are popular in dimensions < 3 are the following

Theorem 2.3.2. A triangulation of a point set $P \subset \mathbb{R}^n$ is a geometric simplicial complex whose vertex set is P and whose simplices tessellate the convex hull of P. Among all triangulations of a point set $P \subset \mathbb{R}^n$, Del(P) achieves the following:

1. In $\mathbb{R}^2$, Del(P) maximizes the minimum angle of triangles in the complex.
2. In $\mathbb{R}^2$, Del(P) minimizes the largest circumcircle for triangles in the complex.
3. For a simplex in Del(P), let its min-ball be the smallest ball that contains the simplex in it. In all dimensions, Del(P) minimizes the largest min-ball.

Unfortunately, the size of a Delaunay complex is $O(n^{\lceil d/2 \rceil})$, with the same cost in computation (see [Cha93]).

[2] Both 2.6 and 2.7 were generated using [VGO+20].

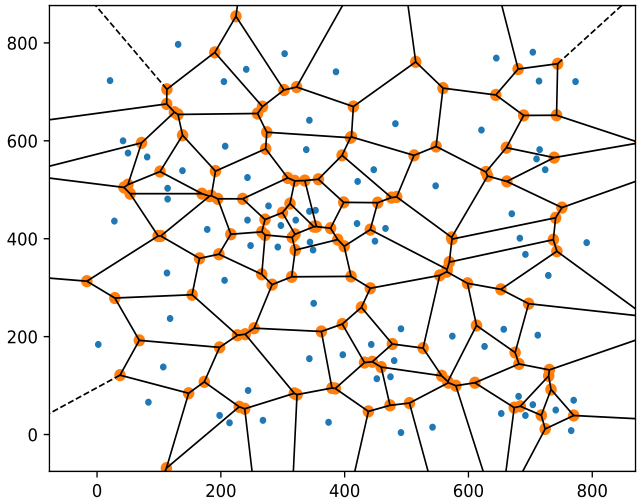


Figure 2.7: Voronoi diagram on a sample of randomly generated points dual to 2.6.

2.3.2 Alpha complex

Alpha complexes are parametrized subcomplexes of the Delaunay complex by some real $\alpha \geq 0$. More specifically, for a given point set P and some $\alpha \geq 0$, an alpha complex consists of all simplices in Del(P) that have a circumscribing ball of radius at most α. An alternative but equivalent definition follows like this - for each point $p \in P$, let $B(p, \alpha)$ be the closed ball of radius α centered at p. Consider the closed set D_p^α be defined as:

$$D_p^\alpha = \{x \in B(p, \alpha) \mid d(x, p) \leq d(x, q), \forall q \in P\}.$$

Then the alpha complex $\text{Del}^\alpha(P)$ is the nerve of the closed sets $\{D_p^\alpha\}_{p \in P}$. We can use the duality of the Delaunay complex to introduce a third definition - an alpha complex contains a k-simplex $\sigma = \{p_0, \ldots, p_k\}$, if and only if $\bigcup_{p \in P} B(p, \alpha)$ meets the intersection of Voronoi cells $V_{p_0} \cup \ldots \cup V_{p_k}$.

2.3.3 Graph induced complex

Lastly, we present here another sparse complex that instead uses subsampling to tackle the size problem that the VR or Čech complexes have, while capturing the topology and even the geometry of the point cloud more efficiently. This construction was introduced in [DFW13].

Definition 2.3.3 (Graph induced complex)**.** Let (P, d) be a metric space, where P is a finite set and $G(P)$ be a graph with vertices in P. Let $Q \subseteq P$ and let $\nu : P \to Q$ the mapping that sets $\nu(p)$ to be any point in argmin $d(p, Q)$. The *graph induced complex* (GIC) $\mathbb{G}(G(P), Q, d)$ is the simplicial complex containing a k-simplex $\sigma =$

$\{q_1, \ldots, q_{k+1}\}$, $q_i \in Q$, if and only if there exists a $(k+1)$-clique in $G(P)$ spanned by vertices $\{p_1, \ldots, p_{k+1}\} \subseteq P$ so that $q_i \in \nu(p_i)$ for each $i \in \{1, 2, \ldots, k+1\}$.

For the input graph $G(P)$ we may consider the neighborhood graph $G^\alpha(P) := (P, E)$, where there is an edge $\{p, q\} \in E$, if and only if $d(p, q) \leq \alpha$. If P is sufficiently dense, then $G^\alpha(P)$ should capture the local neighborhoods of the sample points.

In the following, the quality of the sampled space after subsampling with Q will be quantified with a parameter $\delta > 0$.

Definition 2.3.4. A subset $Q \subseteq P$ is called a δ-sample of a metric space (P, d), if the following condition holds:

- $\forall p \in P$, there exists a $q \in Q$, so that $d(p, q) \leq \delta$.

It is called δ-sparse, if the previous and next condition holds together:

- $\forall (q, r) \in Q \times Q$ with $q \neq r$, $d(q, r) \geq \delta$.

The metric itself is usually taken to be either the Euclidean metric or the graph metric d_G derived from the input graph $G(P)$. In those two cases we have several existence and inference results involving the GIC; see [DFW13] again for that. The paper also includes an empirical comparison of the GIC with the VR complex together with another sparse complex, the Witness complex, which we won't discuss in this thesis. The simulations show that the GIC can have a size similar to the Witness complex (which is usually smaller but also fails to capture the topology more because of it) and maintains the accuracy of the Rips complex.

Summary

Summary of the work.

Appendix A

Bibliography and sources

[Bor48] Karol Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae*, 35:217–234, 1948.

[Cha93] Bernard Chazelle. An optimal convex hull algorithm in any fixed dimension. *Discrete & Computational Geometry*, 10:377–409, 1993.

[DFW13] Tamal K. Dey, Fengtao Fan, and Yusu Wang. Graph induced complex on point data, 2013.

[Nat24] Nathaniel Saul. Čech complex playground — sauln.github.io. https://sauln.github.io/blog/nerve-playground/, 2024. [Online; accessed 16-August-2024].

[Sus24] Sushovan Majhi. Vietoris-rips complex — smajhi.com. http://www.smajhi.com/tutorials/topology/rips.html, 2024. [Online; accessed 16-August-2024].

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.