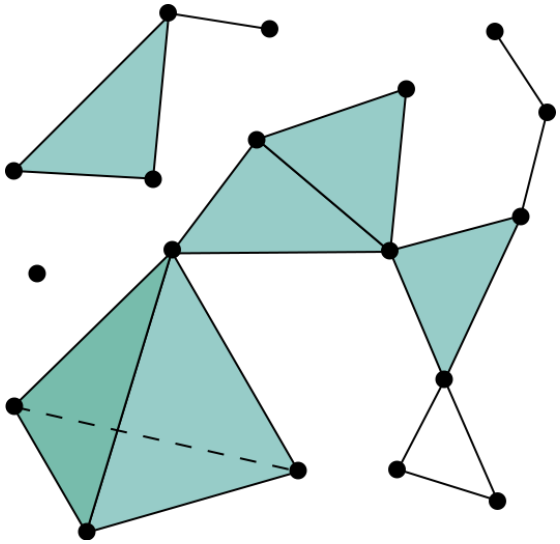
University of Utrecht

Random Geometric Complexes

Master Thesis, Mathematics

Author:
Remie Janssen
3587037

Supervisor:
Dr. Tobias Müller
Second Examiner:
Dr. Gil Cavalcanti



May 10, 2017

Preface

Let me start by saying that I rather enjoyed working on this thesis. The project was presented to me as a interdisciplinary (within mathematics) and relatively new topic. While I could still go quite in depth, it felt like I had to use a broader spectrum of my mathematical abilities than for other topics.

That said, this interdisciplinary/broader look that was needed for this thesis did not only make it more fun. It also made it more challenging to communicate. It seldom took long to understand the idea of a theorem or a proof, but to clearly write down how I understood it, I had to use language from different parts of mathematics. For example, it was quite new for me to mix notation for homology and for probability.

This not only made it harder to write things down, I also had to think about my audience a lot more. Because my supervisor is well versed in random geometric graphs (and hence complexes) and my Master's track focused more on topology and geometry, I decided the thesis had to be understandable for mathematicians with either of these backgrounds. This resulted in a long thesis, which seems to consist of two thirds introduction, and one third literature review. I apologise (again) to my supervisor and to my second examiner fro writing such a long text. I hope, nevertheless, that this thesis might serve as a useful starting point for people wanting to study random geometric complexes.

I now include a short list of some things I have learnt while writing this thesis. Besides the obvious mathematical knowledge that I have gained, there are some more general things that are notable.

- It is hard to write for more than one clear restricted audience. Next time I will either try to limit my audience more, or I will realise earlier that the writing process will be a significant part of the task.
- The website arXiv might only contain old versions of a paper, even when a paper is already published. Do not just assume all mistakes have been corrected yet. I have learnt this while reading a paper that was rather hard for me: comparing an old version of the paper and the errata of the most recent version of the paper, a lot of information seemed to be missing.
- Related to this, I learnt that not-yet published papers sometimes con-

tain quite a few mistakes. I did not realise this before writing this thesis. Somehow, I always assumed research mathematicians would not publish anything they weren't one hundred percent sure of. This is unlike biological papers, which I have read for my Biology Bachelor's: I never assumed those papers to be completely true. This unconsciously was the difference between mathematics and experimental sciences to me.

- This brings me to a part of this project that is not represented in this thesis. I have done 'experiments': I have run simulations of random complexes and computed the shape. At some point during the project, I just needed to do something 'practical'. I realised that I sometimes need to make things more tangible than just words and pictures on paper, the addition of simulations made it feel more real to me.

- This reality that I looked for in the simulations is also important to me in another way. I enjoy keeping applications in mind, and I enjoy telling people about the possible applications. It is important to me that I can communicate the relevance of my research (read: study) to other people in a non banal way: it needs to be in reasonable depth, but any mathematician in any field should be able to understand it.

I conclude this preface with a short reading guide for people who might not know where to start in this booklet of a thesis. I focus on my main chosen audiences: people from probability theory with a basic background in topology, and topologists with a basic knowledge of probability theory.

- Chapter 1 is written as a short reminder of definitions from probability theory.

- Chapter 2 contains a lot of important definitions and notation used in the main part of this thesis. Mainly focus on understanding the definitions of the structures.

- Chapter 3 is a reminder of basic simplicial homology. Topologists can likely skip this chapter. As this thesis studies the homology of random complexes, this is an essential chapter to understand.

- Chapter 4 contains the two tools we use: the nerve theorem, and the main theorem of discrete Morse theory. Understanding the statements of these theorems is most important, the rest of this chapter consists of background to understand the proofs.

- Chapter 5 is a chapter added to give the context of the well-studied random geometric graphs. It is mainly important to understand the relation between random geometric graphs and random geometric complexes.

- Chapter 6 introduces the structure that we study in this thesis. It is therefore vital to understand the definitions in this chapter. People who have studied random geometric graphs will probably recognise most of the

material: for them it probably suffices to study the differences between the two constructions for the complexes: Čech and Vietoris-Rips.

- Chapter 7-9 contain the main material of this thesis.

If one wants to go to the main part as quickly as possible, I suggest they try to understand the idea behind homology, and the two main theorems we use as tools: the nerve theorem, and the main theorem of discrete Morse theory. They can then read the definition of a random geometric complex in Chapter 6, and immediately after start studying the main material presented in the third part of this thesis.

Let me end by expressing my wish that this thesis will be of help to people who start studying random geometric graphs.

Introduction

This thesis is about the 'form' of random geometric complexes. A geometric complex can be visualised as a shape built up from elementary parts. For low dimensional (< 4) complexes, this means we take a bunch of points, line segments, triangles, and tetrahedrons, and then glue them together in a nice way. This nice way just makes sure that endpoints of lines are always glued to endpoints of lines, and similarly, that line segments are glued to line segments, and triangles to triangles, such that nothing sticks out. A random geometric complex is then a randomly chosen set of points, fixed at some position in space. Based on where these points are, we decide whether we draw lines, triangles, tetrahedrons, etc. between them.

We study the homology, i.e. the holes and twists, of these random spaces. If one realises that point clouds of experiments are basically random samplings of points in a space, and that random geometric complexes are built up of these kind of samplings, it is not hard to imagine there must be some application to data analysis. This 'topological data analysis' could be an application of the work in this thesis. In particular the work done here could provide null-hypotheses.

The relevance of such applications is clear from the following example. It is very well possible that data is not normally distributed, but is restricted to some part of the space. A very obvious example is the position in space of an object rotating around another object. Even though the average position is its point of rotation, this will never be its position. Hence it does not make sense to analyse these data by looking at a mean and variance. We want to know how it moves: in a circle, which is a topological object.

Another, possibly less clear, example of use for topological data-analysis is the following. A well known fact is that eutrophication (excess influx of nutrients) of lakes makes lakes turbid, and that we have to remove a lot of nutrients before the lakes clear up again. This is an example of hysteresis, a behaviour of a system characterised by dependence on the direction of change of a parameter. In this case the 'direction' is determined by whether nutrients are added, or removed from the system. This direction dependence is clearly visible in theoretical representations (Figure 1) of data-plots of these variables: turbidity and nutrient content. In real data, this circle is less well visible (Figure 2 [15]).

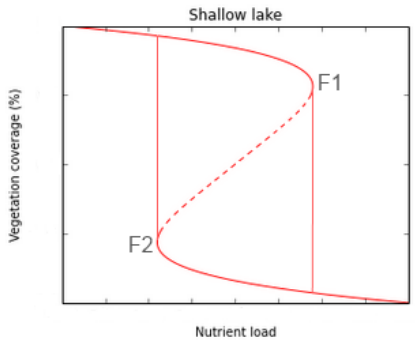


Figure 1: Data of eutrophication in lakes. Hysteresis is clear, as we expect data only on the solid lines. The dashed line is an unstable equilibrium of the system.

Note that such hysteresis is quite common in complex systems, and it would therefore be useful to have the possibility of analysing the topology of data, too.

We will however not focus on topological data analysis at all. We will purely study the shape (homology) of random geometric complexes arising from different kinds of random point processes. The focus will lie at different point distributions in Euclidean space, where we study two constructions of random complexes: Čech and Vietoris-Rips complexes. We are especially interested in limiting behaviour, when we collect increasingly many points.

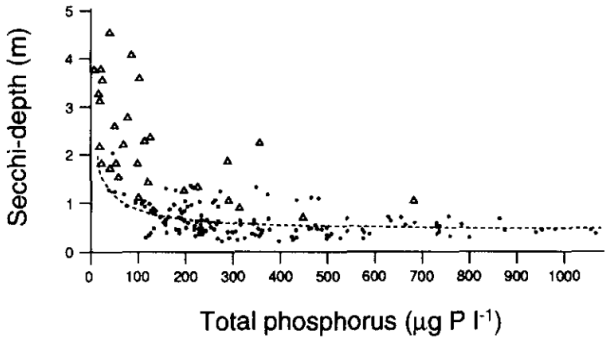


Figure 2: Data of eutrophication in lakes, the hysteresis is not directly clear, but might be indentified using topological data analysis if there is enough data. The typical turbid lakes are represented by dots, and the typical clear lakes are represented by triangles.

Chapter 1

Notation and symbols

1.0.1 Landau notation

We occasionally use Landau notation for asymptotic relationships between functions of a certain variable.

In particular, for functions $f, g : X \to \mathbb{R}$ with $X \subset \mathbb{R}$, we use the following notation. We denote $f(x) = o(g(x))$ for when f is dominated by g asymptotically; $f(x) = O(g(x))$ if $|f|$ is bounded by a constant times g in the limit $x \to \infty$; $f(x) = \Theta(g(x))$ if f and g grow approximately equally fast (up to a linear factor); and $f(x) \sim g(x)$ if f and g are asymptotically equal.

Strictly speaking we define the symbols in the following way:

$$
\begin{aligned}
f(x) = o(g(x)) &\iff \forall k > 0\ \exists N > 0\ \forall n > N &&: && |f(x)| < kg(x) \\
f(x) = O(g(x)) &\iff \exists k > 0\ \exists N > 0\ \forall n > N &&: && |f(x)| < kg(x) \\
f(x) = \Theta(g(x)) &\iff \exists k_1, k_2 > 0\ \exists N > 0\ \forall n > N &&: && k_1 g(x) < |f(x)| < k_2 g(x) \\
f(x) \sim g(x) &\iff \lim_{x \to \infty} \left| \frac{f(x)}{g(x)} \right| = 1.
\end{aligned}
$$

We might also use the 'reverse' symbols ω and Ω:

$$
\begin{aligned}
f(x) = \omega(g(x)) &\iff g(x) = o(f(x)) \\
f(x) = \Omega(g(x)) &\iff g(x) = O(f(x)).
\end{aligned}
$$

1.0.2 List of symbols

Sets

$\mathbb{R}$	the real numbers	
$\mathbb{Q}$	the rational numbers	
$\mathbb{Z}$	the integers	
$\mathbb{N}$	the natural numbers $\{0, 1, 2, \dots\}$	
$[n]$	the set of n numbers $\{1, 2, \dots, n\}$	
$A \subset B$	the set A is a (not necessarily proper) subset of B	

Spaces and categories

$\cong$	isomorphism of spaces	
$\simeq$	homotopy equivalence	
$\mathcal{N}(\mathcal{U})$	is the nerve of a cover $\mathcal{U}$	21
S^k	topological k-sphere	
D^k	topological k-dimensional ball	
$\{pt\}$	the spaceconsisting of one point	
colim	the topological colimit	39
hocolim	the topological homotopy colimit	46
$\coprod$	the topological coproduct, or disjoint sum	40
$\sqcup$	idem	
$X \sqcup_f Z$	spaces X and Y atached along a space Z	40
Set	the category of sets	
Top	the category of opological spaces	
$\mathrm{Ob}(\cdot)$	the objects of a category	
$i \downarrow I$	the over category of an object i in a category I	44

Complexes

$\triangleq$	isomorphism of simplicial complexes	14
$\mathrm{Aut}(\mathcal{K})$	the group of automorphisms of a complex $\mathcal{K}$	14
$\alpha \prec \beta$	'α is a codimension 1 face of β'	52
Δ^k	for $k \in \mathbb{N}_{\geq 0}$ is the k-simplex (as complex)	10
$\partial_i \Delta^k$	Th i-th face of the k-simplex	10
Ω^k	for $k \in \mathbb{N}_{\geq 0}$ the k^{th} cross-polytope (as complex)	20
K_n	the complete graph on n nodes	
$K_{n,m}$	the complete bipartite graph on n and m nodes	
P_n	the (linear) path-graph on n nodes	
$C_k(J)$	the circulant graph on vertices $[k]$ and edges defined by the et $J \subset [k-1]$	20
$\|\| \bullet \|\|$	the realisation operator	11
$V(\Delta)$	is the vertex support of complex Δ	10
$\mathrm{Cl}\,\Delta$	Closure of a simplicial complex Δ	15
$\mathrm{St}\,\Delta$	Star of a simplicial complex Δ	15
$\mathrm{oSt}\,\Delta$	open star of a simplicial complex Δ	15
$\mathrm{Lk}\,\Delta$	Link of a simplicial complex Δ	17
$\Delta \star \Delta'$	Join of simplicial complexes Δ and Δ'	18
$\Delta[K]$	induced subcomplex of $K \subset V(\Delta)$	12
$X(G)$	the clique complex of a graph G	19
$B(x,r)$	is the ball of radius r around point x	
$G(X,r)$	is the geometric graph for point set X and radius r	62
$G_n(\Delta)$	the number of occurrences of Δ as induced subcomplex in a random complex on n points; the type of random complex is assumed to be clear from context.	76
$J_n(\Delta)$	the number of occurrences of Δ as a component in a random complex on n points; the type of random complex is assumed to be clear from context.	76
$K_n(\Delta)$	the number of (unlabelled) occurrences of Δ as subcomplex of a random geometric complex	81
$\check{\mathcal{C}}(X,r)$	is the Čech complex for point set X and radius r	72
$\mathcal{R}(X,r)$	is the Rips complex for point set X and radius r	73
$\mathrm{Sk}_k(\mathcal{K})$	is the k-skeleton of a complex $\mathcal{K}$	13

Homology

δ	boundary operator, used sometimes for homology, but mostly for spaces and for simplicial complexes	28
$H_k(X)$	the k-th homology group of a space X	29
$\beta_k(X)$	the k^{th} Betti number of a space X	27
$\mathrm{vsupp}(\gamma)$	is the vertex support of a (simplicial homological) chain γ	

Probability

$\mathbb{P}(A)$	the probability of an event A	
$\mathbb{E}(X)$	the expected value of a stochastic variable X	
$\mathcal{N}(\mu, \sigma)$	the normal distribution with mean μ and standard deviation σ	3
$\text{Poi}(\lambda)$	the poisson distribution with parameter λ	3
$\mathbb{1}$	the identity map	
$\xrightarrow{D}$	converges in distribution	4
$\mathcal{X}_n$	Binomial point process with n points i.i.d. distributed with density f, which is assumed to be understood from context	65
$\mathcal{P}_n$	Poisson point process with $\text{Poi}(n)$ points i.i.d. distributed with density f, which is assumed to be understood from context	66
Φ_n	Binomial or Poisson point process as above	67
d_{TV}	Total variation distance	5

Contents

Part I

Prerequisites

Chapter 1

Probability

We start with a short chapter about probability, which will mainly introduce notation. Most of these things are incredibly basic, except maybe for the total variation distance.

1.1 Variables

1.1.1 Poisson

A Poisson random variable is an $\mathbb{N}$-valued variable with a parameter $\lambda > 0$. We denote such a variable with $\mathrm{Poi}(\lambda)$, and its probability mass function is given by

$$\mathbb{P}(\mathrm{Poi}(\lambda) = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

It is well known that $\mathbb{E}(\mathrm{Poi}(\lambda)) = \mathrm{Var}(\mathrm{Poi}(\lambda)) = \lambda$, and that all higher moments

$$\mathbb{E}\Big(\mathrm{Poi}(\lambda)^p\Big) = \sum_{i=0}^{\infty} i^p \mathbb{P}(\mathrm{Poi}(\lambda) = i)$$

are also bounded. The exact values are known, but are not relevant here.

1.1.2 Normal

A normal distributed variable is an $\mathbb{R}$-valued variable with two parameters, the mean μ and the standard deviation σ. Its probability density function is given by the familiar bell-shaped curve. We denote such a variable by $\mathcal{N}(\mu, \sigma)$. It suffices to know that the normal distribution is quite ubiquitous, and is hence of great importance.

1.1.3 Some terminology for multiple variables

Sometimes, we have more than one random variables. If we talk about sequences of random variables, we will always assume that they have the same probability space.

In particular we often have a sequence of multiple independent copies of one random variable. If we have such a set of independent and identically distributed variables X_n, we will often abbreviate "independent and identically distributed" to "i.i.d.".

If a set of random variables is not independent, it is important to keep track of the dependencies. The following definition helps to achieve this.

Definition 1.1.1. *Let $\{X_i\}_{i\in I}$ be a set of random variables indexed by I. A graph with vertex set I is called a dependency graph of $\{X_i\}_{i\in I}$ if it has the following property regarding the edges: Let $I_1, I_2 \subset I$ be disjoint, then the disjoint sets of random variables $\{X_i\}_{i\in I_1}$ and $\{X_i\}_{i\in I_2}$ are independent if there are no edges between I_1 and I_2 in the dependency graph.*

1.2 Limits

Whereas it is easy to envision limits in $\mathbb{R}$, it is harder to do the same for limits of random variables, even when they are $\mathbb{R}$-valued. We will look at a few definitions that we will use throughout this thesis.

It is quite common to say that something is true almost surely (a.s.) when we mean that the probability of it happening is 1. Similarly, we say that an event A_n occurs asymptotically almost surely (a.a.s.) if

$$\lim_{n\to\infty} \mathbb{P}(A_n) = 1.$$

1.2.1 Weak convergence

Now what do we do for sequences of random variables? We want to know something about the limit of such a sequence. We can then, of course, define several kinds of convergence. Some will be stronger than others, but they might all have some probabilistic meaning. We will here just shortly introduce two convergences, and note some properties.

Definition 1.2.1. *Let X_n be a sequence of random variables, we say that they converge in distribution to a variable X, or*

$$X_n \xrightarrow{D} X,$$

if

$$\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$$

for any 'nice' event set A. Nice meaning that A is a continuity set of X: a Borel set with zero-measure boundary. We also say that X_n converges weakly to X.

This is somewhat like point-wise convergence if we consider these event sets A as points. The next kind of convergence we will see is stronger, because it is to uniform convergence what weak convergence is to point-wise convergence. To define this convergence, we first define a distance for random variables.

Definition 1.2.2. *Let X and Y be random variables, then the total variation distance between X and Y is*

$$d_{TV}(X,Y) = \sup_{A \in F} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|,$$

where F is the sigma algebra of the sample space, or the 'set of events'.

The convergence we talked about earlier is defined by convergence in this distance. As mentioned earlier, it is stronger than weak convergence and it is named after the distance it stems from: convergence in total variation(al distance). We are mostly interested in the fact it implies weak convergence. This fact is quite trivial when one sees the parallels with point-wise and uniform convergence.

Lemma 1.2.3. *Let $X, X_0, X_1, X_2, \ldots$ be a sequence of random variables. If we have the limit*

$$\lim_{n \to \infty} d_{TV}(X_n, X) = 0,$$

then also $X_n \xrightarrow{D} X$.

Proof. Let A be a Borel set of X, and say X_n converges to X in total variation. Then, per definition, for every $\epsilon > 0$ there exists $N > 0$ such that for all $n > N$

$$d_{TV}(X, X_n) = \sup_{B \in F} |\mathbb{P}(X \in B) - \mathbb{P}(X_n \in B)| < \epsilon$$

. In particular we also have

$$|\mathbb{P}(X_n \in A) - \mathbb{P}(X \in A)| < \epsilon$$

for all $n > N$. Hence we conclude that

$$\lim_{n \to \infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A),$$

as we wanted to prove.

The correspondence hinted at before is easily realised by considering Borel sets as points of the maps defining the random variables.

Moments

Note that for both these kinds of convergence, there is no guarantee that any of the moments also converge to the corresponding moment of the limiting variable.

With this we mean that the following two quantities are not guaranteed to be equal

$$\lim_{n\to\infty} \mathbb{E}(X_n^p), \quad \mathbb{E}(X^p).$$

A good example is the sequence X_n of $\mathbb{Z}$ valued variables, where the probability mass function is

$$\mathbb{P}(X_n = k) = \begin{cases} 1-(1/n) & \text{if } k = 0 \\ (1/n) & \text{if } k = 2^n \\ 0 & \text{otherwise} \end{cases}$$

This sequence converges in distribution, as well as in total variation, to the random variable X which is 0 with probability 1. It is easy to check that $\mathbb{E}(X_n^p) = (2^n/n)^p$, and $\mathbb{E}(X^p) = 0$. Hence we see that convergence in distribution does not imply convergence of moments to the corresponding moment of the limit variable.

1.2.2 Convergence in probability

The second kind of convergence is quite a lot stronger than convergence in distribution. This one is called convergence in probability. We give the definition for vector valued random variables.

Definition 1.2.4. *Let X_n be a sequence of vector valued random variables, where the vector space has metric d, then X_n converges in probability to a random variable X if*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

for all $\epsilon > 0$.

1.2.3 Some more facts

Oftentimes, we are interested in the sum $S_n = X_i + \cdots + X_n$ of such i.i.d. variables $\{X_i\}_{i\in\mathbb{N}}$. In certain cases, the average of this sum has a nice property: The average S_n/n converges almost surely and in probability to the expected value $\mu = \mathbb{E}(X_i)$:

$$\mathbb{P}(\lim_{n\to\infty} S_n/n = \mu) = 1,$$

$$\lim_{n\to\infty} \mathbb{P}(|S_n/n - \mu| \geq \epsilon) = 0.$$

There also is the following convergence in distribution

$$\frac{S_n - \mu}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

When this happens, we say that a central limit theorem (CLT) holds, and it is also the reason the normal distribution is ubiquitous.

In this thesis we will look at some related situations, where there are other reasons for the distribution converging to a normal distribution. We will also call these central limit theorems.

One last remark about the relation between the Poisson and the normal distribution, which seems related. If he parameter of a Poisson variable goes to infinity, its distribution will look like a normal distribution with mean going to infinity. Strictly speaking, this is nonsense, but we can make sense of it by normalising the variables, giving the following limit in distributions:

$$\frac{\mathrm{Poi}(\lambda) - \lambda}{\sqrt{\lambda}} \xrightarrow{D} \mathcal{N}(0,1).$$

Normalisation is a concept we will encounter quite often later in this thesis.

Chapter 2

Simplicial complexes

In this thesis, we are interested in the topology of random complexes. These random complexes are simplicial complexes. Simplicial complexes are widely used in algebraic topology, and in particular, in combinatorial topology. This is the case because simplicial complexes are actually combinatorial constructions, somewhat like hypergraphs. On the other hand they are spaces, so that we can study their topology. In combinatorial topology, we use the combinatorial structure to study the topological structure. In particular, we look at simplicial homology, which is by its construction quite easily computable. One can easily imagine why algebraic topology started with the study of simplicial homology. In this chapter we focus on the structure of simplicial complexes; in next chapter we define (simplicial) homology.

2.1 Abstract and topological simplicial complexes

Simplicial complexes can be thought of as a combinatorial way to describe certain topological spaces. This means we have two ways of looking at these complexes: as topological spaces, and as combinatorial constructions. It will often be convenient to use the combinatorial description for calculations, whereas the study of topological properties is easier when we consider the topological description. As long as we are provided with a description of the simplices, we can go back and forth between the two descriptions. In this section we give definitions for both descriptions.

Definition 2.1.1. *An abstract simplicial complex is a collection $\mathcal{K}$ of finite sets, such that for all $\sigma \subset \tau \in \mathcal{K}$, also $\sigma \in \mathcal{K}$. An element $\sigma \in \mathcal{K}$ of cardinality $|\sigma| = k + 1$ is called a k-simplex, k-cell, or simplex/cell of dimension k. The dimension of $\mathcal{K}$ is the maximal dimension of a cell in $\mathcal{K}$, if this does not exist we say $\mathcal{K}$ is of dimension ∞.*

Most of the time, we will work with finite simplicial complexes, i.e. $|\mathcal{K}|$ is finite. This automatically means that there are $n < \infty$ 0-cells, and that $\mathcal{K}$ is a

subset of the powerset of $[n]$. Consequently, the dimension of $\mathcal{K}$ can be at most $n-1$, this happens precisely when $\mathcal{K}$ is the n-simplex. For this reason, it is convenient to have notation for the set of 0-cells of a complex.

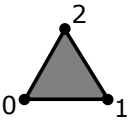
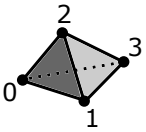
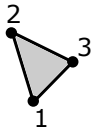
Definition 2.1.2. *The set of* 0*-cells, also called vertices or nodes, of a complex* $\mathcal{K}$ *is denoted* $V(\mathcal{K})$*.*

The topological description of a simplicial complex comes in the form of the realisation of an abstract complex. As any topological simplicial complex can be described as the realisation of a combinatorial one, we will only talk about such realisations. To describe the realisation, we must first know what the realisation of one simplex is. The realisation of the complex will be built with these topological simplices.

Definition 2.1.3. *The standard topological* k*-simplex* Δ^k *is the subspace of* $\mathbb{R}^{k+1}$ *given by the convex hull of the basis vectors* $\{e_i\}_{i=0}^k$*. The* i*-th face* $\partial_i\Delta^k$ *of the simplex is the convex hull of all basis vectors except for the* i*-th. A topological simplex is any space homeomorphic to a standard topological simplex.*

In table 2.1 the first few simplices are depicted, together with the combinatorial description and their 0-th face.

Table 2.1: Standard k-simplices, their topological realisation and their 0-th face, for $0 \leq k \leq 3$.

k	$\| \Delta^k \|$	Δ^k	δ_0	$\| \delta_0\Delta^k \|$
0		$\Big\{\{0\}, \emptyset\Big\}$	$\Big\{\emptyset\Big\}$	$\emptyset$
1		$\Big\{\{0\}, \{1\}, \{0,1\}, \emptyset\Big\}$	$\Big\{\{1\}, \emptyset\Big\}$	
2		$\Big\{\{0\}, \{1\}, \{2\}, \{0,1\}, \{0,2\}, \{1,2\}, \{0,1,2\}, \emptyset\Big\}$	$\Big\{\{1\}, \{2\}, \{1,2\}, \emptyset\Big\}$	
3		$\Big\{\{0\}, \{1\}, \{2\}, \{3\}, \{0,1\}, \{0,2\}, \{0,3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{0,1,2\}, \{0,1,3\}, \{0,2,3\}, \{1,2,3\}, \{0,1,2,3\}, \emptyset\Big\}$	$\Big\{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \emptyset\Big\}$	

There are natural inclusions of faces of topological simplices, as the faces are simply subspaces of the simplex. Note that these faces are simplices of one dimension lower, just like in the abstract case.

The inclusions of faces can thus be used to give a topological interpretation to abstract simplicial complexes. We can 'glue' simplices of the complex to other simplices on their boundaries. Doing this as the combinatorial description prescribes, we get the topological complex. The following definition makes this precise by substituting the rigorous concept of quotients for the intuitive concept of gluing. An example of a simplicial complex in its topological form can be found in Figure 2.1.

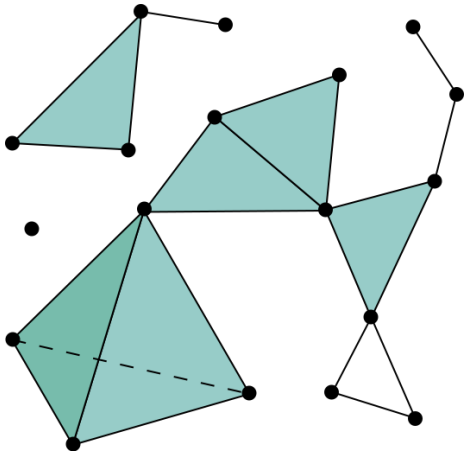


Figure 2.1: A topological simplicial complex, note that it is built with a collection of simplices.

Definition 2.1.4. *The realisation $||\mathcal{K}||$ of an abstract simplicial complex $\mathcal{K}$ is the following quotient space:*

$$\coprod_{\sigma\in\mathcal{K}} \Delta_\sigma^{|\sigma|}/\sim,$$

where $\coprod$ denotes the disjoint union of spaces, and $\Delta_\sigma^{|\sigma|}$ denotes the standard topological $|\sigma|$-simplex indexed by σ. The relation $\sim$ is such that a simplex gets identified with its faces in the expected way, i.e. it is generated by $x \sim i_{\tau\subset\sigma}(x)$ for $x \in \Delta_\tau^{|\tau|}$, τ a face of σ, and $i_{\tau\subset\sigma}$ the inclusion of the topological face as described above.

There is a shorter but more abstract definition of the realisation in terms of a colimit. This description might be useful to keep in mind when we talk about homotopy colimits later. Readers who are not familiar with colimits can skip the following definition, and return to it after reading the short reminder of colimits in Section 4.1.

Definition 2.1.5. *The realisation of an abstract simplicial complex $\mathcal{K}$ is given by the colimit*

$$||\mathcal{K}|| = \text{colim}_F,$$

where F is the functor sending the poset-category $(\mathcal{K}, \subset)$ to **Top***:*

$$F(\{x_0, x_1, \ldots, x_k\}) = \Delta^{|k+1|}$$
$$F(\{x_0, \ldots, \hat{x}_i, \ldots, x_k\} \subset \{x_0, \ldots, x_k\}) = i_{\delta_i \Delta^{k+1} \subset \Delta^{k+1}},$$

where $i_{\sigma \subset \tau}$ is the topological inclusion of the face $\sigma \subset \tau$.

In Section 2.2 we continue our overview of the link between the topological and the abstract description of a simplicial complex. In particular we will see how basic definitions of topology translate to the case of abstract simplicial complexes. First it is useful to have some terminology ready to study parts of complexes.

2.1.1 Subcomplexes and maps

In topology, it is often useful to study just part of a space. As we will view complexes as spaces, we want to be able to study parts of complexes. It is hence convenient to introduce notation and terms for subcomplexes. Similarly, topology is only interesting if we consider (continuous) maps of spaces. Therefore we also introduce maps of complexes here.

Subcomplexes

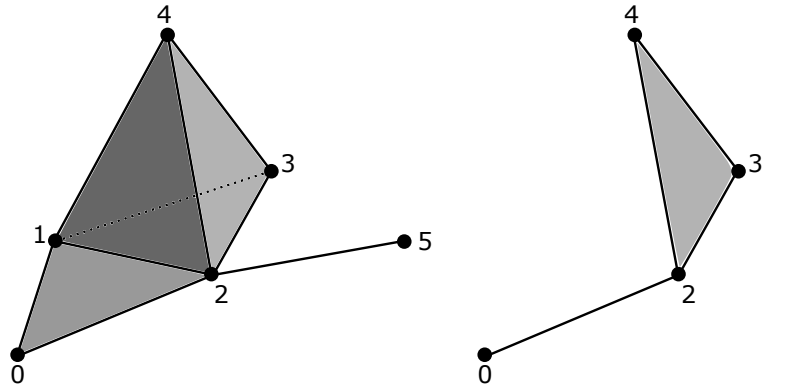
Definition 2.1.6. *Let $\mathcal{K}$ be a simplicial complex, then $\mathcal{K}'$ is a subcomplex of $\mathcal{K}$ if $\mathcal{K}'$ is a complex and $\mathcal{K}' \subset \mathcal{K}$.*

A specific kind of subcomplex that we will encounter is the induced subcomplex. Which is basically the subcomplex we get by throwing away all vertices not in some set A and also removing the corresponding higher dimensional simplices. The following definition is illustrated in Figure 2.2.

Definition 2.1.7. *Let $\mathcal{K}$ be a simplicial complex and $A \subset V(\mathcal{K})$ a subset of the vertices of $\mathcal{K}$, then we denote with $\mathcal{K}(A)$ the induced subcomplex of $\mathcal{K}$ on A. This complex is the subcomplex consisting of all simplexes with only vertices in the set A.*

Another subcomplex of great importance is the k-skeleton, which reduces a complex to a k-dimensional one. These skeleta can be quite convenient if we want to study properties of the complexes.

For example, if we want to count the components of a complex, we only need to know the 1-skeleton (the underlying graph). In this way, also, a lot of properties of a complex can be inferred from some skeleton. An important example is the computation of homology. Informally this means that the number of 'k-dimensional holes' is only dependent on the $k+1$-skeleton of the complex.



(a) Complex $\mathcal{K}$ (b) Induced subcomplex $\mathcal{K}(V)$

Figure 2.2: A simplicial complex with vertices $0, \ldots, 5$ with, on the right, the induced subcomplex $\mathcal{K}(V)$ corresponding to vertex set $V = \{0, 2, 3, 4\}$.

Definition 2.1.8. *Let $\mathcal{K}$ be a simplicial complex, then the k-skeleton of K is the subcomplex*

$$\mathrm{Sk}_k(\mathcal{K}) = \{\sigma \in \mathcal{K} | \dim(\sigma) \leq k\}.$$

Because the skeleta form a sequence of inclusions, they can also be used in an inductive way. This is a ubiquitous proof technique, because going from one skeleton to the next, we only add simplices of a certain dimension. The implications of adding a simplex are often easy to understand.

Maps

As for all structures, we want maps of complexes to respect the structure of the complexes. The following definition makes this precise. It should be noted that we will often use the term 'map' without stating what kind of map it is. For complexes we will generally mean a map of complexes as we will define next, for spaces we will always mean a continuous map unless stated otherwise, because maps of the underlying sets are seldom interesting.

Definition 2.1.9. *Let $\mathcal{K}, \mathcal{K}'$ be simplicial complexes, a map $f : V(\mathcal{K}) \to V(\mathcal{K}')$ is a map of complexes if it induces a map of simplices. With this we mean that the map*

$$\begin{aligned} f :& \mathcal{K} \to \mathcal{K}' \\ f :& \{x_0, \ldots, x_k\} \mapsto \{f(x_0), \ldots, f(x_k)\}, \end{aligned}$$

is well defined.

Just like for other structures, we have special kinds of maps. Examples of these are inclusions of spaces and isomorphisms of groups. The next definition is about these special maps for complexes.

Definition 2.1.10. *Let $\mathcal{K}, \mathcal{K}'$ be simplicial complexes.*
An inclusion of complexes is a map (of complexes) $i : \mathcal{K} \to \mathcal{K}'$ which is an injection on the vertices. Note that there is such an injection iff $\mathcal{K}$ is a subcomplex of $\mathcal{K}$.
An isomorphism of complexes is a bijective map $f : \mathcal{K} \to \mathcal{K}'$. If there exist such an isomorphism between $\mathcal{K}$ and $\mathcal{K}'$, we say that they are isomorphic, which we symbolically represent by writing $\mathcal{K} \triangleq \mathcal{K}'$.
An automorphism of a complex $\mathcal{K}$ is a bijective map $f : \mathcal{K} \to \mathcal{K}'$. The group of automorphisms of $\mathcal{K}$ is denoted $\mathrm{Aut}(\mathcal{K})$.

The complexes we study are mostly finite, and this finiteness can be used in the maps, too. Where for spaces there may be a infinite number of maps of a space to another space; for complexes, this number is obviously bounded by the number of maps from the vertices of the first complex, to the vertices of the second. As both sets of vertices are finite, there is only a finite number of possible maps.

Another nice property that follows, is that the group of automorphisms (isomorphisms from a complex to itself) is finite. We will use this later, when we want to count the number of copies of some kind of complex. The copies we want to count are 'unlabelled' subcomplexes.

Definition 2.1.11. *Let $\mathcal{K}$ be a complex. A labelled (occurrence of a) subcomplex $\mathcal{K}' \subset_i \mathcal{K}$ is a particular injection of $\mathcal{K}'$ into $\mathcal{K}$. An unlabelled (occurrence of a) subcomplex is the class of labelled subcomplexes that have the same image (as a subset of $\mathcal{K}$).*

If we are only able to count occurrences of labelled complexes, we might find the complex too often.

Example 2.1.12. *If we want to count the number of triangles in Δ^2: there are 6 different isomorphisms from the triangle complex to Δ^2, but there is only one actual triangle in Δ^2. Dividing the number of labelled occurrences by the number of automorphisms of the complex we want to find, gives us the number of unlabelled occurrences.*

We will return to this definition when we actually count subcomplexes in Section 6.3. That section also includes another example, where the importance of having the distinction between labelled and unlabelled becomes obvious.

2.2 Topological parallels: Closure, star, link and join

To work with simplicial complexes, we often have to look locally. The locality we know from topological spaces must then be transferred to abstract simplicial complexes. This is done with the introduction of the 'star'. A related concept is the 'link', which then brings convenient notation for working with simplicial subcomplexes. Lastly we have the closure, which is in essence just the topological closure, but phrased in a combinatorial way.

2.2.1 Closure

We will first look at the closure, which corresponds to the topological closure, as said earlier.

Definition 2.2.1. *Let $A \subset \mathcal{K}$ be a set of simplices, the closure of A denoted* $\mathrm{Cl}\, A$ *is defined to be the minimal simplicial subcomplex of $\mathcal{K}$ containing A.*

The definition of closure talks about the minimal subcomplex containing a set of simplices. This is because a set of simplicies is not automatically a simplicial complex; the definition of a complex needs the set of simplices to be closed under taking faces. This means that to get this minimal complex, we only need to add all the faces of the simplices in our set.

The correspondence to the topological closure is clear when we note that a subcomplex will always be a closed subset. Because the closure of A is the smallest (closed)complex containing A, this corresponds directly to the topological definition of the closure of a set being the smallest closed set containing it.

2.2.2 Star

The star construction corresponds to another concept from topology: neighbourhoods. As with neighbourhoods, we have open and closed versions, here represented by the open and closed star. It is convenient to first describe the star of one simplex, as this is easier to envision. Note that neighbourhoods here are relatively 'big' as we must take complete simplices.

Definition 2.2.2. *Let $\sigma \in \mathcal{K}$ be a simplex. The open star of σ denoted* $\mathrm{oSt}\, \sigma$ *is:*

$$\mathrm{oSt}\, \sigma = \{\tau \in \mathcal{K} | \sigma \subset \tau\}$$

The (closed) star of σ denoted $\mathrm{St}\, \sigma$ *is:*

$$\mathrm{St}\, \sigma = \{\tau \in \mathcal{K} | \sigma \cup \tau \in \mathcal{K}\}$$

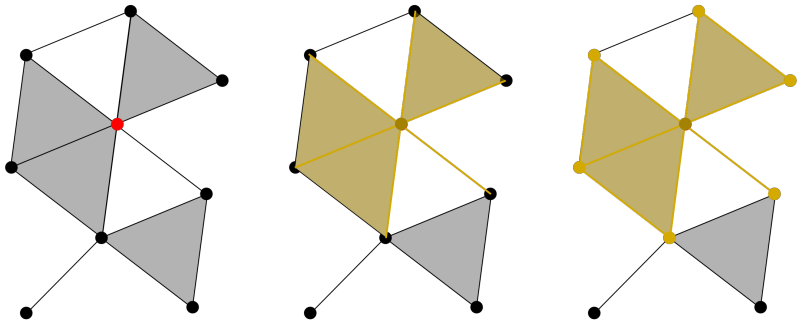
Note that the open star is seldom a simplicial subcomplex, the exception being when σ is an isolated simplex. The closed star, however, is per definition always a simplicial complex.

Instead of looking at the star of one simplex, we can also take the star of a set of simplices. We define this such that it still corresponds with topological neighbourhoods.

Definition 2.2.3. *The (closed) star* $\mathrm{St}(A)$ *of A, where $A \subset \mathcal{K}$ is a subset of simplices, is*

$$\cup_{\sigma \in A} \mathrm{St}\, \sigma,$$

and similarly for the open star.



(a) Complex and node v (b) Open star (c) Closed star

Figure 2.3: A simplicial complex with node v indicated in red. In the middle figure we see the open star $\mathrm{oSt}\, v$ of v, and the right figure depicts the closed star $\mathrm{St}\, v$ of v.

Of course, a closed and open neighbourhood can be related to eachother. The closure of an open neighbourhood should be a closed neighbourhood. This is reflected in the following proposition, which treats this for the case of simplicial complexes.

Proposition 2.2.4. *As functions, applying the open star and then the closure, is equivalent to applying the closed star, i.e.*

$$\mathrm{Cl} \circ \mathrm{oSt} = \mathrm{St}$$

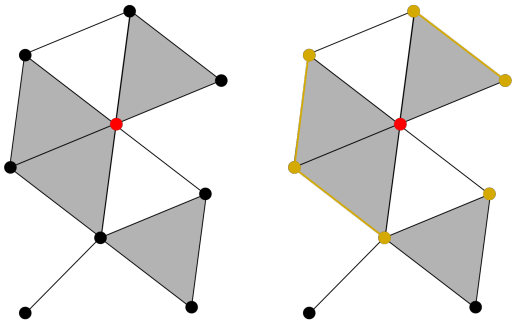
Proof. Let $A \subset \mathcal{K}$ be a subset of the simplices of complex $\mathcal{K}$. We prove both inclusions as sets, so that we can conclude that $\mathrm{Cl} \circ \mathrm{oSt} = \mathrm{St}$.

Let σ be a simplex in $\mathrm{Cl}(\mathrm{oSt}(A))$, then $\sigma \subset \tau$ for some simplex $\tau \in \mathrm{oSt}\, A$. This means there is some simplex $v \in A$ such that $v \subset \tau$. Now look at $\sigma \cup v \subset \tau$: because $\tau \in \mathcal{K}$ we know that $\sigma \cup v \in \mathcal{K}$, hence $\sigma \in \mathrm{St}\, v$ and consequently $\sigma \in \mathrm{St}\, A$.

Now suppose $\sigma \in \mathrm{St}\, A$, then per definition there is some $v \in A$ such that $\sigma \in \mathrm{St}\, v$. Using the definition of the star, we see that $\tau := \sigma \cup v \in \mathcal{K}$. Because $v \subset \tau$, we see that $\tau \in \mathrm{oSt}\, v$, and because $\sigma \subset \tau$, we see that $\sigma \in \mathrm{Cl}\, \tau$. Hence we conclude that $\sigma \in \mathrm{Cl}\, \mathrm{oSt}\, v$. Seeing that the closure of a set of simplices is the union of the closures, we conclude that $\sigma \in \mathrm{Cl}\, \mathrm{oSt}\, A$. ◆

2.2.3 Link

In our comparison of topological and complex-definitions, the link seems to correspond to the boundary of a neighbourhood within the topological complex. In this section we will see that this correspondence can be made precise. This is



(a) Complex and node v (b) Link of v

Figure 2.4: A simplicial complex with node v indicated in red. The right figure depicts the link $\operatorname{Lk} v$ of v. Compare with Figure 2.3 to see the relation $\operatorname{St} v = \operatorname{oSt} v \cup \operatorname{Lk} v$.

intuitively already clear when we see how the link, the star, and the open star are related.

Definition 2.2.5. *Let v be a vertex in a simplicial complex $\mathcal{K}$. The link of v denoted* $\operatorname{Lk} v$ *is:*

$$\operatorname{Lk} v = \{\sigma \in \mathcal{K} | v \notin \sigma, \sigma \cup \{v\} \in \mathcal{K}\}$$

Just like the closed star, the link is a simplicial subcomplex. The link is related to the star and the open star by $\operatorname{Lk}(v) = \operatorname{St}(v) \setminus \operatorname{oSt}(v)$ for v a vertex. This gives us an easy way to go back and forth between these related sets of simplices. We prove this in the following proposition.

Proposition 2.2.6. *For a vertex v of a simplicial complex, we have* $\operatorname{Lk}(\{v\}) = \operatorname{St}(\{v\}) \setminus \operatorname{oSt}(\{v\})$.

Proof. The proof is a straightforward unrolling of the definitions. Say $\sigma \in \operatorname{St}(\{v\}) \setminus \operatorname{oSt}(\{v\})$, this happens if and only if $\sigma \notin \operatorname{oSt}(\{v\})$ and $\sigma \in \operatorname{St}(\{v\})$. Using the definitions of the star and the open star, we can see that this situation occurs iff $\sigma \in \mathcal{K}$, $v \notin \sigma$ and $\{v\} \cup \sigma \in \mathcal{K}$. These are precisely the conditions for σ being an element of $\operatorname{Lk}(v)$.

Noting that we can reverse the argument because all steps consist of an 'if and only if', we conclude that $\sigma \in \operatorname{Lk}(\{v\})$ iff $\sigma \in= \operatorname{St}(\{v\}) \setminus \operatorname{oSt}(\{v\})$. ◆

For arbitrary subsets of simplices, the relation as above does not hold precisely the same. In these cases we have the following relation, which clearly makes our intuition in the topological sense work again.

Definition 2.2.7. *Let $A \subset \mathcal{K}$ be a subset of simplices, then the link of this set is defined as*

$$\operatorname{Lk} A = \operatorname{St} A \setminus \operatorname{oSt} \circ \operatorname{Cl} A.$$

Note that by the proposition above, this definition coincides in the case of vertices because the closure operator acts as the identity on vertices.

2.2.4 Join

The last construction we will look at is the join. It is the abstract version of the topological join.

Definition 2.2.8. *Let $\mathcal{K}$ and $\mathcal{K}'$ be simplicial complexes with $V(\mathcal{K}) \cap V(\mathcal{K}') = \emptyset$. The join $\mathcal{K} \star \mathcal{K}'$ is:*

$$\{\sigma \subset V(\mathcal{K}) \cup V(\mathcal{K}') | \sigma \cap V(\mathcal{K}) \in \mathcal{K} \text{ and } \sigma \cap V(\mathcal{K}') \in \mathcal{K}'\}$$

We will sometimes denote $\mathcal{K} \star \sigma := \mathcal{K} \star \mathrm{Cl}(\sigma)$ for the join with a simplex. The closure here is strictly speaking needed, because σ on itself is not a simplicial complex.

Example 2.2.9. *The cone of a complex $\mathcal{K}$ is the join $\mathcal{K} \star \{v\}$ with a complex $\{v\}$ with just one vertex not contained in the vertex support of $\mathcal{K}$. Joining $\mathcal{K}$ with an independent set of n vertices is called n-coning. Another term for 2-coning is 'taking the suspension'.*

Proposition 2.2.10. *Let v be a vertex of the simplicial complex $\mathcal{K}$, then*

$$\mathrm{Lk}(v) \star v = \mathrm{St}(v).$$

Proof. Let $\sigma \in \mathrm{Lk}(v) \star v$. Then we have, either $v \notin \sigma$ which implies $\sigma \in \mathrm{Lk}(v) \subset \mathrm{St}(v)$, or $v \in \sigma$ and $\sigma \cap V(\mathrm{Lk}\, v) = \tau \in \mathrm{Lk}\, v$. With $\tau \in \mathrm{Lk}\, v$, we get $\tau \in \mathrm{St}\, v \setminus \mathrm{oSt}\, v \subset \mathrm{St}\, v$. Therefore $\tau \cup \{v\} \in \mathrm{St}\, v$, and since $\sigma = \tau \cup \{v\}$, we see that $\sigma \in \mathrm{St}\, v$.

Now suppose $\sigma \in \mathrm{St}(v)$, and note firstly that the vertex set of σ denoted $V(\sigma)$ is contained in $V(\mathrm{Lk}\, v) \cup \{v\}$. It is clear that $\sigma \cap \{v\}$ is a simplex in the complex $\{v\}$. Now we look at $\sigma \cap V(\mathrm{Lk}\, v)$, this is either σ if $v \notin \sigma$, or it is $\sigma \setminus \{v\}$ if $v \in \sigma$. In both cases, we see that $v \notin \sigma \cap V(\mathrm{Lk}\, v)$ and $(\sigma \cap \mathrm{Lk}\, v) \cup \{v\} \in \mathcal{K}$ because $\sigma \in \mathrm{St}\, v$. Hence $\sigma \in \mathrm{Lk}(v) \star v$. ◆

Note that a similar proposition is false if we substitute for v a subcomplex $\mathcal{K}'$.

2.3 Clique complexes

An easy way to construct a simplicial complex is by taking the maximal complex that is supported by some graph. This kind of complex is called a clique complex. It will turn up as one of the ways in which we get our random geometric complexes, which makes the comparison with the field of random geometric graphs quite obvious. It is therefore worthwhile to look at some properties of these complexes.

Definition 2.3.1. *Let G be an undirected graph with vertices $V(G)$ and edges $E(G)$, then the clique complex on G, denoted $X(G)$, is the simplicial complex defined by*

$$X(G) = \{\sigma \subset V(G) | (v, w) \in E(H) \text{ for all } v \neq w \in \sigma\}$$

These kind of complexes are rather restricted, as for example, a triangle is already not a clique complex, because it will be 'filled' with a 2-simplex. In this section we will look at special properties of clique complexes compared to general simplicial complexes.

2.3.1 Examples and properties

Obviously, not every subcomplex of a clique complex is also a clique complex. The following proposition states that the link of a vertex in a clique complex is also a clique complex. This fact will be used later when we look at the homology of clique complexes.

Proposition 2.3.2. *Let $X(H)$ be the clique complex on the graph H, and let v be a vertex of H. The link* $\mathrm{Lk}(v)$ *in this complex is also a clique complex.*

Proof. Suppose the vertices $v_0, \ldots, v_k$ form a clique in $\mathrm{Lk}(v)$. To show that $\{v_0, \ldots, v_k\} \in \mathrm{Lk}(v)$, it is sufficient to prove that the simplex $\{v_0, \ldots, v_k\} \cup \{v\}$ is a simplex in $X(H)$, as then $\{v_0, \ldots, v_k\} \in \mathrm{Lk}(v)$. To prove this, we only have to prove that all the faces of $\{v_0, \ldots, v_k\} \cup \{v\}$ are in $X(H)$. We proceed by using induction on the dimension k of the simplex.

Basis Suppose $k = 1$ and we find $\{v_0, v_1\}$ forming a clique in the link $\mathrm{Lk}(v)$. Then trivially $\{v_0, v_1\}$ forms a clique in $\mathrm{Lk}(v)$.

Hypothesis Let $k < n$ and suppose $\{v_0, \ldots, v_k\}$ forms a clique in $\mathrm{Lk}(v)$, then $\{v_0, \ldots, v_k\}$ is a simplex in $\mathrm{Lk}(v)$.

Step Now suppose we find $\{v_0, \ldots, v_n\}$ forming a clique in $\mathrm{Lk}(v)$. By the induction hypothesis, the faces $\{v_0, \ldots, \hat{v_i}, \ldots, v_n\}$ (for all $0 \leq i \leq n$) are simplices in the link. By the definition of the link, we can conclude that the simplices $\{v_0, \ldots, \hat{v_i}, \ldots, v_n\} \cup \{v\}$ are elements of the clique complex. Also, because $\{v_0, \ldots, v_n\}$ is a clique in the link, it also is in the clique complex $X(H)$, and we conclude that $\{v_0, \ldots, v_n\}$ is a simplex of the clique complex. Hence all faces of $\{v_0, \ldots, v_n\} \cup \{v\}$ are in the clique complex, and consequently, so is the the simplex $\{v_0, \ldots, v_n\} \cup \{v\} \in X(H)$. We conclude that $\{v_0, \ldots, v_n\} \in \mathrm{Lk}(v)$.

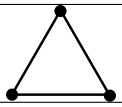
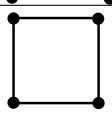
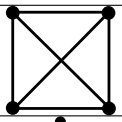
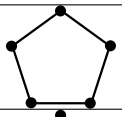
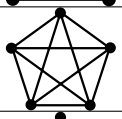
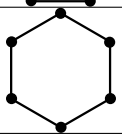
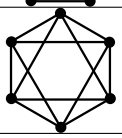
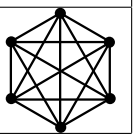
Circulant graphs

A special type of clique complexes, are the clique complexes arising from circulant graphs.

Definition 2.3.3. *For $k > 0$ and $J \subset [k-1]$, the circulant graph $C_k(J)$ is the graph on $[k]$ where there is an edge (x, y) iff there exists $j \in J$ such that $x + j \equiv y \mod k$ or $x - j \equiv y \mod k$.*

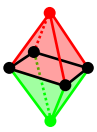
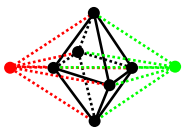
For $J = [n]$ we denote $C_k([n]) = C_k(n)$.

Table 2.2: Circulant graphs $C_k(n)$ for several values of k and n.

k\n	1	2	3
3			
4			
5			
6			

We introduce these kind of graphs, because a subtype is quite interesting in the study of homology of clique complexes. Cross polytopes are a specific class of complexes on circulant graphs, which have a special role in homology of clique complexes. We will later see that these are the smallest 'hollow' subcomplexes of clique complexes (Theorem 3.4.4).

Table 2.3: The first 4 cross polytopes, where the 'new' vertices and edges (compared to the one 1 dimension lower), are indicated in red and green.

k	0	1	2	3
Ω^k				

Definition 2.3.4. *The k-dimensional cross-polytope Ω^k is the clique complex of $C_{2k+2}(k)$. In other words $\Omega^k = X(C_{2k+2}(k))$.*

It is easy to see that the k-dimensional cross-polytope is homeomorphic to a k-sphere, for example by inductively taking the suspension, as we see done in Figure 2.3. We will use this fact later when we are interested in homology of these complexes. It is a fun exercise to look at the other circulant graphs and deduce their topological structure.

2.4 The nerve

A way to use some topological space to create a simplicial complex is by taking a cover of the space (a collection of subsets, the union of which is X), and using that to build a simplicial complex. This construction, that we define in this section, is called the nerve construction. If we take a sufficiently nice space and cover, we can retrieve the space from the resulting complex. This will be made precise in the very useful nerve theorem which we state and prove in chapter . The main reason to treat this construction, is because it is used for the definition of a type of random geometric complex, called the Čech complex.

Definition 2.4.1. *Let $\mathcal{U} = \{U_i\}_{i\in I}$ be a cover, then we have a simplicial complex $\mathcal{N}(\mathcal{U})$. Its set of 0-simplices is I and a finite set $S \subset I$ is a simplex of $\mathcal{N}(\mathcal{U})$ iff $\cap_{i\in S} U_i \neq \emptyset$.*

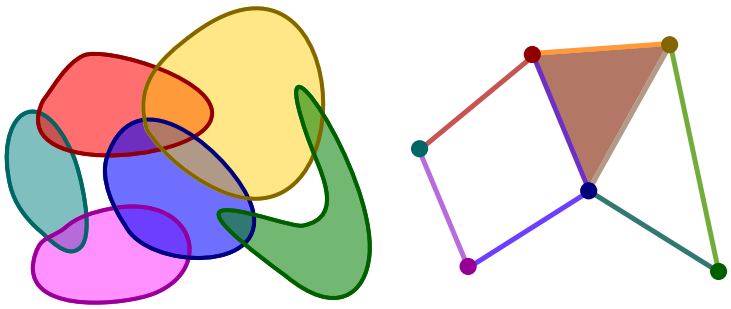


Figure 2.5: The nerve of a cover, coloured sets correspond to the vertices of the same colour.

An example of the nerve construction can be found in Figure 2.5. As alluded to in the text above the definition, the nerve construction can give really nice results, if the cover is nice enough. However, when we take a cover such as in Figure 2.6, then the complex and the covered space are hardly related at all. In Chapter 4 we will see when a cover is nice enough to get a nerve complex actually related to the covered space.

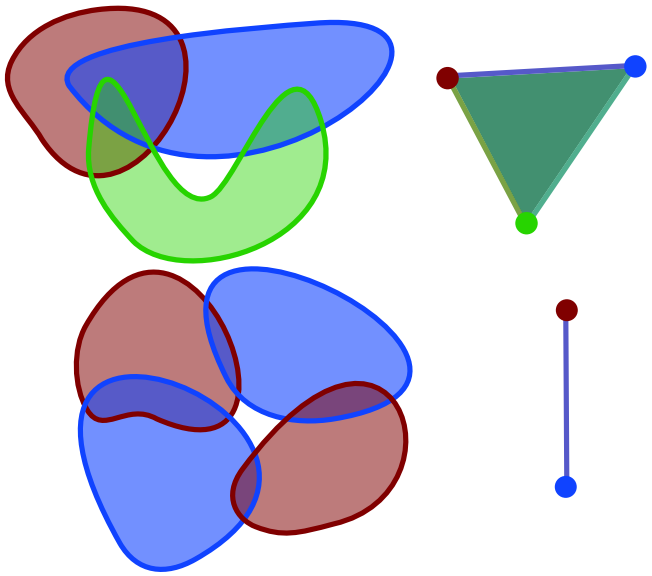


Figure 2.6: The nerve of a cover for two covers. coloured sets correspond to the vertices of the same colour. Note that the complex and the covered space are topologically completely different in both cases.

Chapter 3

Homology

Homology is an abstract algebraic method of studying of spaces. It gives a lot of information about the shape of a space in the sense that we get a global picture. In particular, we get information about the holes and twists in a space. For example, a circle has one (one-dimensional) hole, which can be characterised with the terminology of homology.

Although homology gives a lot of global information, it can be computed with local information together with information about how the parts are patched together. This means that using homology works especially well if there is a lot of information about subspaces. In particular, a filtration (sequence of subspaces) is easy to work with. A good example is the skeletal filtration of a simplicial complex, i.e. the sequence of d-dimensional skeleta. In this chapter we give an overview of well known properties of (simplicial) homology.

3.1 Intuition

It is easy to get confused with all the abstract definitions in homology. To prevent this, we first try to get some intuition for the things we want to study with homology. Informally, these things are holes and twists in topological spaces.

3.1.1 Holes

There are many kinds of holes in spaces. We can classify them by their dimension. Zero dimensional holes, for example, are the separations of components; One dimensional holes are things you can poke your finger through, like the hole in a ring; Two dimensional holes are cavities, such as the inside of a balloon.

It is tempting to immediately think of these holes as the inside of a sphere, but there are also other kinds. Think for example of the cavity in a torus: a torus evidently has only one cavity. One can see this easily if we think of a

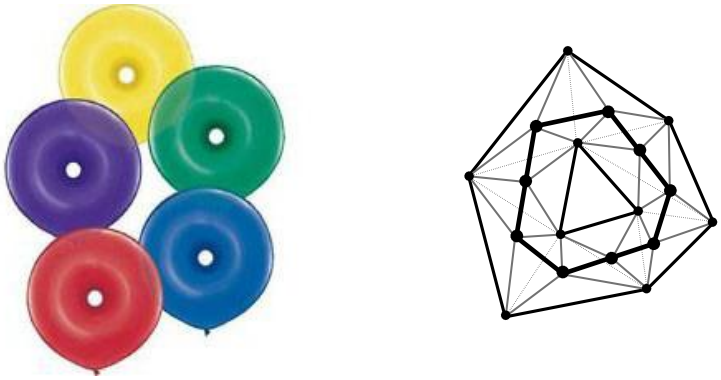


Figure 3.1: Left: A torus-shaped balloon can be blown up through one hole. Source: https://mathsgear.co.uk/products/torus-balloons; Right: A simplicial complex homeomorphic to a torus.

torus-shaped balloon (Figure 3.1), the balloon can be blown up through the one hole. The cavity however, is not like the cavity in a 2-sphere.

Homology solves this problem by defining cycles, which can be seen as boundary-less subspaces of a certain dimension. If such a subspace cannot be filled up in the space, then it represents a hole.

3.1.2 Twists

The other thing we can measure with homology is the twists of the space. These are harder to imagine than holes, as spaces with twists cannot be embedded in three-dimensional euclidean space. Two well known examples of spaces with twists are the real projective plane, and the Klein bottle. Every twists has an order, which is 'the number of times you have to go round to get back where you started'.

It is hard to make the twists more precise without going into details of homology. Therefore we just treat an example where we gloss over all the details by assuming we can continuously deform a cycle without changing the equivalence class of the cycle.

Example 3.1.1. *This example is illustrated in Figure 3.2. It is meant to give a feeling for what a twist in a space constitutes, and how it manifests itself in homology. Let us start with the representation of* $\mathbb{R}P^2$ *as a disc with its boundary glued to itself by identifying opposite points on the boundary. It is clear that the red and the green line in the first part of the figure both represent the same cycle, because they are equal.*

We want to see that such a cycle is actually 'equivalent' to the cycle traversed in the other direction, because then the space should have some kind of twist in it. To see this, we for now assume that we know that cycles can be transformed

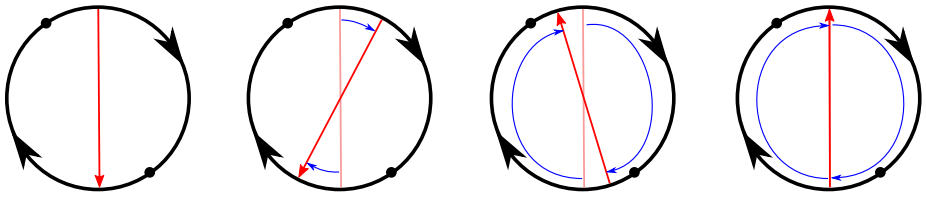


Figure 3.2: The reason $\mathbb{R}P^2$ has a twist. The red path in each part represents the same cycle, it is just rotated as the blue arrows indicate. Note that the red cycle in the first part is the inverse of the red cycle in the last part, i.e. the red cycle is its own inverse.

in a continuous way, such that the actual cycle does not change. In the picture, we just rotate the green cycle continuously, until it is back to where it started, but in the other direction. If we now 'add' the green and the red cycle, the sum is 0.

We see that by starting with two copies of the cycle (going through the cycle twice), we actually have a trivial cycle, and there is a twist in the space. The difference with a hole, is that in that case no multiple copies of the cycle will be trivial.

3.1.3 Coefficients

In the example of last section, we talked about two cycles and said we had one (1) of both, so we started with two times one particular cycle. In that case we were counting the cycles with the integers ($\mathbb{Z}$), as we assumed that the number two (2) had meaning. We can, however, choose to 'count' the cycles using a different group. This group is then called the group of coefficients, and it is crucial to know which coefficients we use.

The coefficients are always a group, so we can add cycles. This means that we can sometimes split them, too. The following example shows why this is very important.

Example 3.1.2. *In the example above, we could have started with the same cycle, but only one time, so with coefficient 1. If we had coefficients in $\mathbb{Q}$ instead of in $\mathbb{Z}$, we could split the cycle in twice a half times the cycle. This does not mean we take only part of the 'path', we just count the one cycle as 'twice a cycle with coefficient $1/2$'. We can then use the same rotation as in the previous example to see that one $1/2$ cycle is the reverse of the other $1/2$ cycle. This means that with coefficients in $\mathbb{Q}$, the twist is not discernible, because the half cycles cancel each other out.*

In the same way, no twist can ever be seen in homology with coefficients in $\mathbb{Q}$ or $\mathbb{R}$. The only thing we register there is the number of holes. This means that the number of holes is easily defined by taking homology with coefficients in $\mathbb{Q}$. The number of holes will be called the Betti number. This is the quantity that we study mainly in this thesis.

3.2 Chain complexes and their homology

Homology is easily defined by using chain complexes, which are defined in an abstract algebraic way. Examples of these chain complexes in algebraic topology are often quite geometric, which makes interpreting the homology in a geometric way easier. The fact that chain complexes are defined abstractly, makes them a convenient framework for the study of spaces.

We begin by just defining chain complexes and maps between chain complexes. This will be followed by a definition of homology which arises naturally from the definition of chain complexes.

3.2.1 Chain complexes

A chain complex can be viewed as a sequence of objects with nice maps. This niceness is exactly defined so that we can produce homology from the objects, which must also have certain properties.

Definition 3.2.1. *A chain complex is a collection* $\mathcal{C} = \{C_i\}_{i\in\mathbb{Z}}$ *of groups or modules together with maps* $d_i : C_i \to C_{i-1}$ *such that* $\mathrm{Im}(d_i) \subset \mathrm{Ker}(d_{i-1})$*, or equivalently such that* $d \circ d = 0$*.*

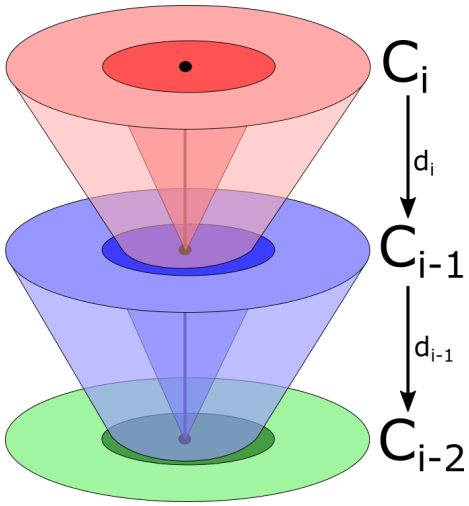


Figure 3.3: A schematic showing the structure of a chain complex. Every horizontal level shows a group or module C_i (large circle), with its kernel (smaller circle) in corresponding colours. Note that d maps the whole C_i into the kernel of C_{i-1}, which then gets mapped to the identity represented by the black line and dots.

In this thesis, we will mostly work with chain complexes where the C_i are vector spaces, i.e. modules over a field. This will considerably simplify calculations because the C_i are then characterized by their dimension as a vector

space. The maps are also easy to work with as we only have to know the image of a basis.

Note that the term chain complex is not named after the sequence, which some might call a chain, but after the elements of the C_i. These elements are called chains for their appearance in geometric context, which we will see in next section. The next definition introduces more terms like chain.

Definition 3.2.2. *Let $\mathcal{C}$ be the following chain complex*

$$\cdots \to C_{i+1} \xrightarrow{d} C_i \xrightarrow{d} C_{i-1} \to \cdots,$$

and let $c \in C_i$. The element c is called a chain, if $c = d(c')$ for any $c' \in C_{i+1}$, then c is called a boundary. The set of boundaries in C_i is denoted B_i. If $d(c) = 0$, then c is said to be a cycle. The set of cycles in C_i is denoted Z_i (for the German word Zyklus).

All of these terms will be explained geometrically when we look at simplicial homology.

3.2.2 Homology

Recall that the holes we wanted to study where cycles (boundaryless objects) that could not be filled up. A cycle can be filled up iff it is a boundary. So that means homology must be some object that counts cycles, but disregards them when they are a boundary. This is exactly how we define homology, as can be seen in the following definition.

Definition 3.2.3. *The homology $H(\mathcal{C})$ of a chain complex $\mathcal{C}$ is defined as follows. It consists of groups or modules*

$$H_k := \frac{Z_k}{B_k}.$$

An element of H_k is called a homology class, and two cycles are said to be homologous if they are in the same homology class. Cycles which are elements of the trivial homology class are called trivial cycles.

Homology in this sense is quite easily defined, but also very abstract. In practice we will take chain complexes corresponding to topological spaces, and the construction of the chain complex will be quite geometrical. Still, the homology groups of a space might not directly give you an idea of what the space looks like, even though they encode information like the number 'twists' and 'holes' in the space. In this thesis, we mainly study this number of holes.

Definition 3.2.4. *The k-th Betti number $\beta_k(X) := \beta_k(\mathcal{C})$ of a topological space X with chain complex $\mathcal{C}$, is defined to be the rank of $H_k(\mathcal{C})$ when working over $\mathbb{Z}$, or if we work over a field of characteristic 0, as the vector space dimension of $H_k(\mathcal{C})$.*

In next section we will look at simplicial homology, a concrete geometrical and almost combinatorial kind of homology. All concepts of this section will be clarified with examples, and the nomenclature should become obvious in the geometrical context.

3.3 Simplicial homology

Simplicial homology is the kind of homology we will use most in this thesis, as it is defined for simplicial complexes. The big advantage of this framework is that homology of this kind is easily computable. Additionally, this simplicial homology is equal to singular homology and de Rham homology. Hence, we do not actually make a limiting choice here.

3.3.1 The simplicial chain complex

In simplicial homology, the chains are simply given in terms of the simplices. The boundary operation is then easily defined by using the face maps. Of course we have to keep in mind some kind of orientation, this will be incorporated in the boundary map. With this the homology is defined as for any chain complex. To define simplicial homology, we will first have to pick the chain complex.

Definition 3.3.1. *A simplicial k-chain c in a simplicial complex Δ is a formal R-linear combination of simplices:*

$$c = \sum_{\sigma \in S} r_\sigma \sigma,$$

where $S \subset \Delta$ is a finite set of k-simplices in Δ (i.e. $\sigma \in S$ then $|\sigma| = k$), R is a ring, and $r_\sigma \in R$. The set (in fact module) of k-chains is denoted $C_k(\Delta, R)$.

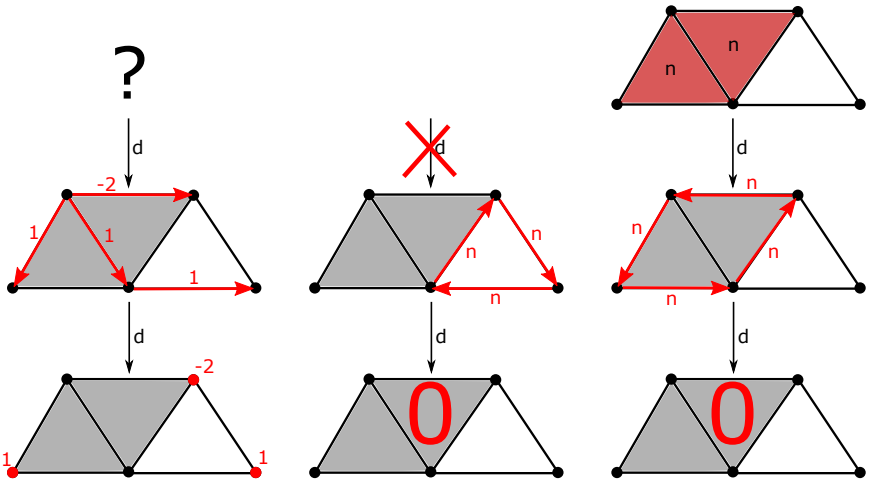
Typically the ring R is equal to one of $\mathbb{Z}$, $\mathbb{Z}/p\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$. the last two of these do not give any different information, as intuitively explained in Section 3.1.

To get homology, we now only need to have a boundary map. To define this boundary map, we assume there is an orientation on the complex, which manifests itself as a (partial) order on the simplices such that we can compare any two vertices in a simplex.

It hence also gives an orientation on each simplex, which can be thought of as an ordering of the vertices of the simplex. The reason this is an orientation is very clear in the case of a one-dimensional simplex. There the ordering defines a starting point and an end point of the line segment. For higher dimensional simplices something similar holds.

Definition 3.3.2. *The boundary map $d_k : C_k(\Delta, R) \to C_{k-1}(\Delta, R)$ is defined by*

$$d_k \sigma := \sum_{i=0}^{k} (-1)^i \partial_i \sigma,$$



(a) Chain (b) Non-trivial cycle (c) Boundary

Figure 3.4: A few examples of elements from C_1 for the complex at the top. The first one is a chain with non-trivial boundary. The second one is a boundary; a pre-image is given. The last one is a non-trivial cycle.

where $\partial_i\sigma = (v_0, \dots, \hat{v_i}, \dots, v_k)$. *For a* k*-chain* $\gamma = \sum r_\sigma \sigma$ *we have*

$$d\gamma := \sum r_\sigma d\sigma.$$

We will often drop the index k *from* d_k*, as most of the time, it is known which degree chain we work with.*

Proposition 3.3.3. *For* d *as defined above we have* $d \circ d = 0$.

Proof. We prove this fact for $\gamma = 1\sigma$, and $\sigma = (v_0, \dots, v_k)$ an arbitrary simplex. The result follows easily because d is a module homomorphism and the simplices form a basis of he module. We now just compute the coefficient of $(v_0, \dots, \hat{v_i}, \dots, \hat{v_j}, \dots, v_k)$ in $d \circ d(\sigma)$ for any $0 \leq i < j \leq k$. We can first remove the ith vertex and then jth, or vice versa (note that in the first case the jth vertex becomes the $(j-1)$th). This corresponds to the fact that the coefficient will be

$$\begin{aligned}(-1)^i\partial_i((-1)^j\partial_j\sigma) + (-1)^{j-1}\partial_{j-1}((-1)^i\partial_i\sigma) &= ((-1)^{i+j} + (-1)^{i+j-1})\partial_j\partial_i\sigma \\ &= 0.\end{aligned}$$

As we get 0 for each pair $i < j$, we know that $d \circ d(\sigma) = 0$. ◆

With our chains and boundary maps defined, we also have homology.

Definition 3.3.4. *Simplicial homology groups* $H_k(\mathcal{K})$ *of a complex* $\mathcal{K}$ *with coefficients in* R *is the homology corresponding to the chain complex* $(C_k(\mathcal{K}, R), d_k)$ *with* $C_k(\mathcal{K}, R)$ *and* d_k *as above.*

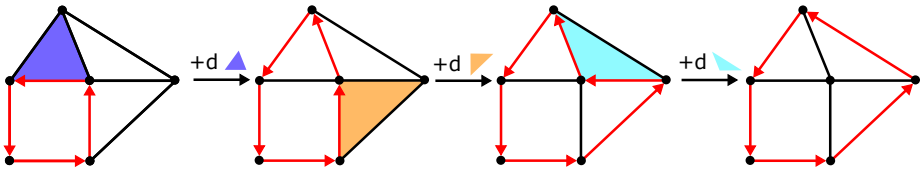


Figure 3.5: Stepwise explanation of why two chains are in the same class. In each step we add a boundary, which makes us stay in the same class. Of course the equivalence could be seen in one step by adding the sum of the boundaries.

3.3.2 Some useful facts about simplicial homology

We now look at a few general results about simplicial homology which we will use throughout this thesis. These results are quite basic facts which can be found in standard works like Hatcher's book [6]. Two of the following results regard long exact sequences, algebraic objects which are useful in the computation of homology.

First we treat a very basic but important fact about homology. In many treatments of homology it is seen as an axiom, because it can be used to define homology. The way we treat homology, it should be proven, but we will omit the proof as it serves no purpose here.

Lemma 3.3.5 (Corollary 2.11 and Theorem 2.27 in Hatcher [6])**.** *Let X and Y be two homotopy equivalent spaces, then $H_k(X) = H_k(Y)$ for all k.*

We will use this lemma mainly in the application of the nerve theorem, which gives us a homotopy equivalence between a complex and some space. There we skip over the fact that simplicial and singular homology are equal, which is also not so trivial to prove.

Long exact sequences of homology

A long exact sequence is a convenient computational tool. When part of the sequence is known, exactness gives a lot of information about the part that is not yet known.

Definition 3.3.6. *A long exact sequence is a sequence of (Abelian) groups or modules*

$$\cdots \to C_{i+1} \xrightarrow{f_{i+1}} C_i \xrightarrow{f_i} C_{i-1} \to \cdots$$

such that $\operatorname{Ker} f_i = \operatorname{Im} f_{i-1}$.

We can compute things with these sequences using the exactness. For example if there is a part of the sequence that looks like $\cdots \to 0 \to A \to B \to 0 \to \cdots$, then we can conclude that A is isomorphic to B.

The first of the two LES's that we discuss, is the long exact sequence of a pair. A pair in topology is a pair of spaces, one contained in the other. These pairs need to have some nice properties, but for subcomplexes these are already

satisfied. We suggest the reader to think about the maps on the level of cycles, as this works quite well, even though the maps are strictly speaking maps of equivalence classes of cycles.

Proposition 3.3.7 (Long exact sequence of a pair)**.** *Let $\mathcal{K}' \subset \mathcal{K}$ be complexes, then there exists a long exact sequence*

$$\cdots \xrightarrow{j_*} H_{k+1}(\mathcal{K}/\mathcal{K}') \xrightarrow{\delta} H_k(\mathcal{K}') \xrightarrow{i_*} H_k(\mathcal{K}) \xrightarrow{j_*} H_k(\mathcal{K}/\mathcal{K}') \xrightarrow{\delta} H_{k-1}(\mathcal{K}') \xrightarrow{i_*} \cdots$$

of homology, where the map i_* is induced by the injection $i : \mathcal{K}' \to \mathcal{K}$ and the map j_* by the quotient map $j : \mathcal{K} \to \mathcal{K}/\mathcal{K}'$ as topological spaces.

The map δ in the definition above is harder to explain than i_* and j_*. This is especially the case because we directly go to the homology of the quotient space $\mathcal{K}/\mathcal{K}'$, instead of rigorously defining the homology of a pair of spaces. If the reader wants to understand the map δ better, we refer them to the section about relative homology in Chapter 2 of Hatcher's book [6].

The second LES is the Mayer-Vietoris sequence. For this one, we divide the space in two subspaces. The homology of the subspaces together with the homology of the intersection give a lot of information about the homology of the total space. For these sequences we again invite the reader to understand them better by studying the maps. The boundary-map δ_* merits most study, because it is often defined very abstractly, but can be understood intuitively.

Proposition 3.3.8 (Mayer-Vietoris)**.** *Let $A, B \subset X$ be complexes, then there is a long exact sequence*

$$\cdots \xrightarrow{\delta_*} H_k(A \cap B) \xrightarrow{(i_*, j_*)} H_k(A) \oplus H_k(B) \xrightarrow{k_* - l_*} H_k(X) \xrightarrow{\delta_*} H_{k-1}(A \cap B) \to \cdots$$

of homology, where the maps are induced by the inclusion maps

$$\begin{aligned} i : A \cap B &\to A, \quad j : A \cap B \to B, \\ k : A &\to X, \quad l : B \to X. \end{aligned}$$

A nice application of the Mayer-Vietoris sequence is the computation of the homology of the spheres. We treat this application in terms of spaces instead of simplicial complexes. We do this because notation for complexes isomorphic to the spheres is unnecessarily complicated, especially when we can just talk about the spheres directly. It is however possible to rephrase the following lemma in terms of cross polytopes.

Lemma 3.3.9. *Let S^n denote the n-dimensional sphere with $n > 0$, then*

$$H_k(S^n, \mathbb{Z}) = \begin{cases} \mathbb{Z} & \text{if } k = n, 0 \\ 0 & \text{otherwise} \end{cases},$$

and

$$H_k(S^n, \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } k = n, 0 \\ 0 & \text{otherwise} \end{cases}.$$

Proof. Let D_1^n and D_2^n be two discs covering S^n with intersection a small open neighbourhood of the equator. Note that this intersection is homotopy equivalent to a $(n-1)$-sphere. This means the homology of the intersection is the same as the homology of S^{n-1}. We use the Mayer Vietoris sequence with this cover:

$$\cdots \to H_k(S^{n-1}) \to H_k(D^n) \oplus H_k(D^n) \to H_k(S^n) \to H_{k-1}(S^{n-1}) \to \cdots .$$

Knowing that $H_0(S^0) = \mathbb{Z} \oplus \mathbb{Z}$, we can inductively use the Mayer-Vietoris sequence of $D_1^n, D_2^n \subset S^n$ to arrive at the result. To clearly see this, we explicitly give the relevant parts of the sequences. The cases $n = 1$ and $n > 1$ are treated separately as the induction basis, and the step.

Basis First we look at the case $n = 1$. Only the end of the sequence is relevant, as for $k > 1$ all the homology groups in the sequence are 0. The end of the sequence is

$$\cdots \to H_1(D^1) \oplus H_1(D^1) \to H_1(S^1) \to H_0(S^0) \to H_0(D^1) \oplus H_0(D^1) \to H_0(S^1) \to 0.$$

We note that $H_k(D^n) = 0$ for all $k > 0$, because D^n is homotopy equivalent to a point; and we note that S^1 has one component, so $H_0(S^1) = \mathbb{Z}$. We fill this in to get the exact sequence

$$\cdots \to 0 \oplus 0 \to H_1(S^1) \to \mathbb{Z} \oplus \mathbb{Z} \to \mathbb{Z} \oplus \mathbb{Z} \to \mathbb{Z} \to 0.$$

Using exactness (in this case rank-nullity), it is easy to see that $H_1(S^1) = \mathbb{Z}$. We conclude that the theorem is correct for $n = 1$.

Hypothesis The theorem is correct for $n = i - 1$.

Step We again use the Mayer-Vietoris sequence for $D_1^i, D_2^i \subset S^i$. Using our knowledge of the homology of the discs we fill in the sequence for $k > 1$

$$\cdots \to H_k(D^i) \oplus H_k(D^i) \to H_k(S^i) \to H_{k-1}(S^{i-1}) \to H_{k-1}(D^i) \oplus H_{k-1}(D^i) \to \cdots$$

to get

$$\cdots \to 0 \oplus 0 \to H_k(S^i) \to H_{k-1}(S^{i-1}) \to 0 \oplus 0 \to \cdots .$$

This sequence implies that $H_k(S^i)$ and $H_{k-1}(S^{i-1})$ are isomorphic if $k > 1$. The induction hypothesis then implies

$$H_k(S^i, \mathbb{Z}) = \begin{cases} \mathbb{Z} & \text{if } k = i, 0 \\ 0 & \text{otherwise} \end{cases}.$$

Replacing $\mathbb{Z}$ by $\mathbb{Q}$ everywhere proves the second statement of the lemma. 

Minimal non-trivial cycle

When we look at homology of random complexes later we will want to 'estimate' homology. In certain cases, the so called 'sparse regime', most components will be very small. Hence the homology can be determined by looking at very small components. To use this, we want to know which small components can contribute to homology. The following lemma is about the smallest of these components.

Proposition 3.3.10 (Minimal non-trivial cycle)**.** *Let $k \geq 1$, then in a simplicial complex, a non-trivial k-cycle γ is supported on a complex with at least $k+2$ vertices. Additionally, if the complex has exactly $k+2$ nodes, then it is the hollow $(k+1)$-simplex $\delta\Delta^{k+1}$.*

Proof. Suppose we have a complex with $k+1$ vertices, then there is at most one k-simplex. This simplex has a non-zero boundary, so there are no non-trivial k-cycles.

We now assume the complex has $k+2$ vertices. It should be clear that the only boundary-less chains are the ones picking each possible k-simplex with the same coefficient, as each face is shared by exactly two of these simplices. If the $(k+1)$ simplex on all vertices is present, such a cycle can be filled up, so there can only be a non-trivial cycle if the complex is the boundary of the $(k+1)$-simplex, which we call the hollow/empty simplex of dimension k, or the hollow/empty k-simplex.

Note that Clique complexes cannot take this form, and hence their minimal non-trivial cycle is supported on a different complex. We will see later which form this complex takes.

Betti numbers

In this thesis, we will often try to find the Betti number of a random complex, as we want to know the number of holes in a random complex.

The way we try to find the Betti number, is by counting certain components. For that purpose we introduce some counting variables here. Note that some of these variables just count the number of unlabelled occurrences of a certain complex, and others also need this complex to be a component of the larger complex.

Definition 3.3.11. *Let $\mathcal{K}$ be a simplicial complex. The following are variables*

which relate to the number of occurrences of a certain type of subcomplex:

$$\begin{aligned}
s_k(\Delta) &:= \#\{A \subset V(\mathcal{K}) | \mathcal{K}[A] \triangleq \delta\Delta^{k+1}\} \\
\tilde{s_k}(\mathcal{K}) &:= \#\{A \subset V(\mathcal{K}) | \mathcal{K}[A] \triangleq \delta\Delta^{k+1}, \mathcal{K}[A] \text{ a component}\} \\
o_k(\mathcal{K}) &:= \#\{A \subset V(\mathcal{K}) | \mathcal{K}[A] \triangleq \Omega^k\} \\
\tilde{o_k}(\mathcal{K}) &:= \#\{A \subset V(\mathcal{K}) | \mathcal{K}[A] \triangleq \Omega^k, \mathcal{K}[A] \text{ a component}\} \\
f_k^{=i}(\mathcal{K}) &:= \#\{\mathcal{K}' \subset \mathcal{K} | \mathcal{K}' \triangleq \Delta^k, |C(\mathcal{K}')| = i\} \\
f_k^{\geq i}(\mathcal{K}) &:= \#\{\mathcal{K}' \subset \mathcal{K} | \mathcal{K}' \triangleq \Delta^k, |C(\mathcal{K}')| \geq i\}.
\end{aligned}$$

where $C(\mathcal{K}')$ is the component of subcomplex $\mathcal{K}' \subset \mathcal{K}$.

The counting variables count special kinds of subcomplexes: s_k counts the number of hollow simplices of dimension k; o_k counts the number of unlabelled occurrences of the k-th cross-polytope; $f_k^{=}i$ counts the number of k-simplices in complexes of size exactly i, et cetera. All these are useful in computing homology as we will see in the following propositions.

Proposition 3.3.12. *Let $\mathcal{K}' \subset \mathcal{K}$ be simplicial complexes and $k \geq 1$, then*

$$\beta_k(\mathcal{K}) \leq \#\{k\text{-simplices in } \mathcal{K}\} \leq \binom{|V(\mathcal{K})|}{k+1},$$

and

$$|\beta_k(\mathcal{K}) - \beta_k(\mathcal{K}')| \leq \sum_{j=k}^{k+1} \#\{j\text{-simplices in } \mathcal{K} \setminus \mathcal{K}'\}.$$

Proof. We prove this proposition by induction, using long exact sequences of homology with coefficients in $\mathbb{Q}$ (because we are interested in the Betti number). Each step of the induction adds one simplex. In particular, we prove that if $\mathcal{K} = \mathcal{K}' \cup \{\sigma\}$, where σ is of dimension d, then

$$\beta_n(\mathcal{K}) \in \begin{cases} \{\beta_n(K'), \beta_n(K') - 1\} & \text{if } n = d-1 \\ \{\beta_n(K'), \beta_n(K') + 1\} & \text{if } n = d \\ \{\beta_n(K')\} & \text{otherwise} \end{cases}.$$

The proof of this is easy when we use the long exact sequence of the pair $(\mathcal{K}, \mathcal{K}')$, and note that $\mathcal{K}/\mathcal{K}'$ as a topological space is homeomorphic to S^d, of which we know the homology by Lemma 3.3.2. The long exact sequence now gives us

$$\cdots \to H_{n+1}(S^d) \to H_n(\mathcal{K}') \to H_n(\mathcal{K}) \to H_n(S^d) \to H_{n-1}(\mathcal{K}') \to \cdots.$$

For $n, n+1 \neq d$, we can easily see that this part of the sequence becomes

$$\cdots \to 0 \to H_n(\mathcal{K}') \to H_n(\mathcal{K}) \to 0 \to \cdots,$$

and hence in those cases $H_n(\mathcal{K}') = H_n(\mathcal{K})$ and hence $\beta_n(\mathcal{K}') = \beta_n(\mathcal{K})$. If $n = d$ then we get

$$\cdots \to 0 \to H_n(\mathcal{K}') \to H_n(\mathcal{K}) \to \mathbb{Q} \to \cdots.$$

Now if $\mathbb{Q}$ is the kernel of the last depicted map, then exactness gives $H_n(\mathcal{K}) = H_n(\mathcal{K}') \oplus \mathbb{Q}$. If the kernel of that same map is 0, then $H_n(\mathcal{K}') = H_n(\mathcal{K})$.

Lastly, if $n = d - 1$ we have part of the sequence

$$\cdots \to \mathbb{Q} \to H_n(\mathcal{K}') \to H_n(\mathcal{K}) \to 0 \to \cdots,$$

and for similar reasons we get $H_n(\mathcal{K}') = H_n(\mathcal{K}) \oplus \mathbb{Q}$ or $H_n(\mathcal{K}) = H_n(\mathcal{K}')$. In terms of the Betti numbers, this is exactly what the equation above tells us. By induction over the simplices in $\mathcal{K} \setminus \mathcal{K}'$, we can easily see that the proposition holds.

An easy consequence of this proposition is the following one.

Proposition 3.3.13. *For a simplicial complex Δ, we have $\tilde{s}_k(\Delta) \leq \beta_k(\Delta) \leq \tilde{s}_k(\Delta) + f_k^{\geq k+3}(\Delta)$.*

Proof. The first inequality is obvious, as each isolated hollow simplex contributes 1 to the Betti number (Lemma), but there may be more components that contribute. The second inequality is a bit harder to see: Note that non-trivial chains can only occur in components on $k+2$ or more vertices, and that the only complex on $k+2$ vertices that sustains non-trivial chains, is the hollow simplex (Proposition 3.3.10). Therefore, we look at the hollow simplices counted by $\tilde{s}_k(\Delta)$ and at all components with at least $k+3$ vertices. Using the previous proposition, we see that for each of these components, the Betti number is bounded by the number of k-simplices in the component. This is exactly what is being counted by $f_k^{\geq k+3}(\mathcal{K})$.

3.4 Homology of clique complexes

Clique complexes have slightly different properties than general simplicial complexes. This means that their homology has some stronger properties. In particular the minimal non-trivial cycles cannot be supported on a hollow simplex, as these do not exist in a clique complex. We will mainly look at the implications of this fact for the bounds on the Betti number of clique complexes.

3.4.1 Minimal nontrivial cycles

In this section we look at nontrivial (homological) chains in clique complexes. In particular we will see how many nodes are needed to support such a nontrivial chain. We will follow the proofs of lemma 5.2 and lemma 5.3 from [9] and fill in the gaps in the arguments. This means the proofs are due to Mathew Kahle, but the proofs here contain more details.

We start by defining the restriction of chains to subcomplexes. In particular, we want to make sense of restricting a cycle to the link of a vertex. This link may of course be of lower dimension than the cycle, but we do not want it to automatically become trivial. So we define the following restrictions.

Definition 3.4.1. *Let γ be a chain in a simplicial complex $\mathcal{K}$ given by*

$$\gamma = \sum \lambda_\sigma \sigma,$$

and v a vertex of the complex. Then we define

$$\gamma \cap \operatorname{oSt} v := \sum_{\sigma \in \operatorname{oSt} v} \lambda_\sigma \sigma, \quad \gamma \cap \operatorname{Lk} v \quad := \sum_{\sigma \in \operatorname{oSt} v} \lambda_\sigma (\sigma \setminus \{v\}),$$

the restrictions of the cycle to $\operatorname{oSt} v$ *and to* $\operatorname{Lk} v$.

With these definitions we can prove the following lemma about non-triviality of these restricted cycles.

Lemma 3.4.2 ([9] lemma 5.2)**.** *Suppose $\gamma \in C_k(\Delta)$ represents a non-trivial k-cycle with minimal vertex support, $k \geq 1$. Then, for any $v \in \operatorname{vsupp}(\gamma)$, the $(k-1)$-chain $\gamma \cap \operatorname{Lk}(v)$ represents a nontrivial cycle in* $\operatorname{Lk}(v)$.

The proof is divided in two parts: we first prove that $\gamma \cap \operatorname{Lk}(v)$ actually is a cycle, by showing that it is the boundary of some chain in the total complex Δ; then we prove that $\gamma \cap \operatorname{Lk}(v)$ is nontrivial in $\operatorname{Lk}(v)$, by showing that it is not the boundary of any chain supported in this link.

Proof. Let it be noted that an orientation of the complex is implicit in this proof. This orientation is, however, important in the calculations of the proof.

We first see whether $\gamma \cap \operatorname{Lk} v$ is indeed a cycle by computing its boundary. To do this, we just compute that $\gamma \cap \operatorname{Lk}(v) = \partial(\gamma \cap \operatorname{oSt}(v))$. We can understand this intuitively by noting that γ is a cycle, and the restriction of γ can hence only have boundary on the boundary of the complex it is restricted to. Because $\gamma \cap \operatorname{Lk}(v)$ is a boundary, it is automatically a cycle.

More rigorously we have the following reasoning, which is quite abstract. Any σ with non-zero coefficient in $\partial(\gamma \cap \operatorname{oSt}(v))$ is a face of a simplex in $\operatorname{oSt}(v)$ and hence we just have $\sigma \in \operatorname{St} v = \operatorname{oSt}(v) \cup \operatorname{Lk} v$. Suppose $\sigma \in \operatorname{oSt} v$, then any simplex that has σ as a face must also lie in $\operatorname{oSt} v$. This means that $\mu_\sigma = \mu'_\sigma$, where μ and μ' are the coefficients of σ in $\delta\gamma$ and $\delta(\gamma \cap \operatorname{oSt} v)$. Because γ is a cycle, we conclude that $\mu'_\sigma = 0$. Now suppose $\sigma \in \operatorname{Lk} v$. Then the only simplex in $\operatorname{oSt} v$ contributing to the coefficient of σ is $\sigma \cup \{v\}$, and we can easily see that $\mu'_\sigma = \lambda_{\sigma \cup \{v\}}$. Hence we conclude that $\gamma \cap \operatorname{Lk}(v) = \partial(\gamma \cap \operatorname{oSt}(v))$.

Now we prove the cycle is non-trivial in $\operatorname{Lk}(v)$ by contradiction. Suppose $\gamma \cap \operatorname{Lk} v$ is trivial, then it is the boundary $\partial(\beta)$ for some k-chain β with support in $\operatorname{Lk}(v)$. Write $\beta = \sum_{\sigma \in \operatorname{supp}(\beta)} \mu_\sigma \sigma$ and define the $(k+1)$-chain

$$\beta \star \{v\} := \sum_{\sigma \in \operatorname{supp}(\beta)} \mu_\sigma (\sigma \cup \{v\}),$$

A simple computation reveals that the boundary of this chain is $\gamma \cap \mathrm{oSt}(v) + (-1)^{k+2}\beta$. This is because the $\star$ just expands the link to the open star, and similarly it expands $\partial(\beta) = \gamma \cap \mathrm{Lk}\, v$ to $\gamma \cap \mathrm{oSt}\, v$ by adding v to each of the simplices. The term $(-1)^{k+2}\beta$ is explained by the 'new boundary' of the open star compared to the link. The factor $(-1)^{k+2}$ is an artifact of the fact that we do not know the orientation of the complex.

To complete the argument, we just define $\gamma' := (\gamma - \gamma \cap \mathrm{St}(v)) + (-1)^{k+3}\beta$, and note that γ and γ' are in the same homology class, but v is in the vertex support of γ but not in the vertex support of γ', this contradicts the condition of minimal vertex support for γ, which was set in the lemma.

If the proof is unclear, we encourage the reader to draw an example. The correspondences of all cycles as stated in the proof become quite clear visually.

We can use this lemma to prove a more interesting one about the minimal cycles in a clique complex.

Lemma 3.4.3 ([9] lemma 5.3)**.** *If γ represents a nontrivial k-cycle for $k > 0$ in a clique complex $X(H)$, then $|\,\mathrm{vsupp}(\gamma)| \geq 2k + 2$.*

Proof. The proof of this lemma goes by induction on k. The preceding lemma plays a crucial role in the induction step, where it will bound the vertex degree for each vertex in the support of the nontrivial cycle γ.

Basis For $k = 1$, we obviously need at least 4 vertices to support a nontrivial cycle; there are no clique complexes with a hole on less than 4 vertices. Hence the case $k = 1$ is clear.

Hypothesis For all $k' < k$: if γ' is a nontrivial k'-cycle in a clique complex, then $|\,\mathrm{vsupp}(\gamma')| \geq 2k' + 2$.

Step Suppose γ is a nontrivial k-cycle, and $|\,\mathrm{vsupp}(\gamma)| \leq 2k + 1$: we will show this leads to a contradiction. Let $v \in \mathrm{vsupp}(\gamma)$ be any vertex in the support of γ. By the preceding lemma, $\gamma \cap \mathrm{Lk}(v)$ is a nontrivial $(k-1)$-cycle in $\mathrm{Lk}(v)$. By Proposition 2.3.2 the link $\mathrm{Lk}(v)$ is also a clique complex. Hence we can use the induction hypothesis to see that $|\,\mathrm{vsupp}(\gamma \cap \mathrm{Lk}(v))| = 2k$, because $\gamma \cap \mathrm{Lk}(v)$ is supported on at least $2k$ vertices in the link and it must be supported on less vertices than γ. Consequently we also have the equality $|\,\mathrm{vsupp}(\gamma)| = 2k + 1$. This means that the degree of v in H is $2k$. We can repeat this argument for each $v \in \mathrm{vsupp}(\gamma)$, hence the subcomplex of $X(H)$ supported on $\mathrm{supp}(\gamma)$ is a $(2k+1)$-clique. As this subcomplex is contractible it cannot support a nontrivial cycle, hence we have reached a contradiction, from which we conclude that $|\,\mathrm{vsupp}(\gamma)| \geq 2k + 2$.

In a later chapter we will use these facts in the following form, which is just a bit stronger than last lemma. We can think of it as a lemma about minimal spheres in the clique complex.

Theorem 3.4.4. *A nontrivial k-cycle γ in a clique complex $X(H)$ has $|\operatorname{vsupp}(\gamma)| \geq 2k+2$. If equality holds, then γ must be supported on the cross-polytope Ω^k.*

Proof. Note that by previous lemma, we only have to prove the second statement: "If $|\operatorname{vsupp}(\gamma)| = 2k+2$, then γ must be supported on a subcomplex isomorphic to the cross-polytope Ω^k."

The proof uses a counting argument similar to the one in previous lemma. Suppose γ is a non-trivial k-cycle supported on a complex with $2k+2$ vertices. As in the proof of previous lemma, we can see that for each vertex v we have $|\operatorname{vsupp}(\gamma \cap \operatorname{Lk}(v))|$ is $2k$ or $2k+1$. If there is a vertex v for which this quantity equals $2k+1$, then v has edges to all other vertices. Consequently, the clique complex is contractible to v. Hence there cannot be a non-trivial cycle. We conclude $|\operatorname{vsupp}(\gamma \cap \operatorname{Lk}(v))| = 2k$ for all vertices v. It is clear that the only clique complex where all vertices have $2k$ edges is the cross-polytope Ω^k. 

For clique complexes we can also estimate the Betti number with a sandwich between two counting variables as in Proposition 3.3.13. In fact, the same relation holds, because clique complexes are simplicial complexes. Although we can see that these bounds do not tell us much, because $\tilde{s}_k(X(G)) = 0$. With the preceding discussion, we can get similar but more useful bounds for clique complexes.

Proposition 3.4.5. *For a clique complex $X(H)$, we have $\tilde{o}_k(\Delta) \leq \beta_k(\Delta) \leq \tilde{o}_k(\Delta) + f_k^{\geq 2k+3}(\Delta)$.*

Proof. The proof is the same as for Proposition 3.3.13. The difference is that Clique complexes have minimal k-cycles supported on k-dimensional cross-polytopes, which are counted by $\tilde{o}_k$. 

Chapter 4

Topological tools

4.1 Homotopy colimits

For the proof of a quite strong theorem about the nerve, we will have to use some more advanced ways to construct spaces. In particular, we will need to know about homotopy colimits, which are closely related to the normal colimits of spaces. We assume the colimit of spaces is known here, but for the sake of comparison, we will also recall its (categorical) definition.

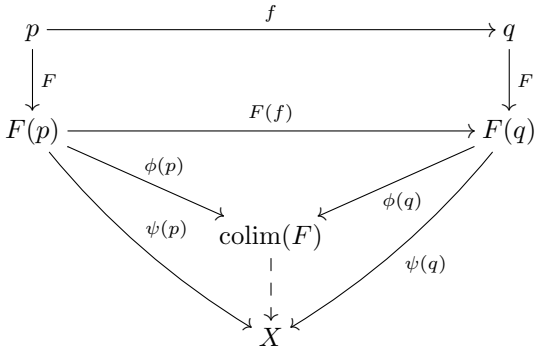
4.1.1 Colimits

In this section we recall the definition of a colimit of spaces and we will look at simple but interesting examples. In these examples we encounter a property relating to homotopy which we would like a construction to have. The colimit construction fails this property.

To define the colimit we first remind the reader that a small category is a category in which the objects and the maps are sets. In general, a category might have a class of objects that cannot be described as a set. Think for example of the category of sets (**Set**): it is well-known that it leads to contradictions to assume that there is a set of all sets (Russell's paradox).

Definition 4.1.1. *Let $F : C \to$ **Top** be a functor from a small category C to the category of topological spaces. The colimit of this functor is a space* $\operatorname{colim} F$ *together with maps $\phi(p) : F(p) \to \operatorname{colim}(F)$ that commute with all F_f in the following sense: the upper part of the diagram (without X and corresponding*

arrows)



commutes. The colimit has the following universal property: for any space X coming with maps $\psi_p : F(p) \to X$ which also commute with all the F_f, the diagram above commutes.

The mentioned space is the following

$$\operatorname{colim} F = \coprod_{p \in P} F(p)/\sim,$$

where the equivalence relation is generated by the induced maps, i.e. $x \sim F_f(x)$ for all maps $f : p \to p'$ in P and elements $x \in F(p)$. The maps are the obvious maps sending an element to its equivalence class.

In fact the above definition has some redundance: If we have a space and maps with this universal property, then the explicit description of the space and maps follow. We have however included them in the definition because we will mostly work with the explicit description, as it is easier visualized. Keep in mind that abstract proofs using the universal property can be very short and elegant.

Note that the definition uses a small category P, in the cases we encounter, this will often be a poset. In particular we will mostly use the poset of inclusions for a cover and all intersections of the cover.

To get an idea of what a colimit is, we look at a few examples:

Example 4.1.2 (Coproduct)**.** *The simplest example of a colimit is the coproduct, which is called disjoint sum in the category of spaces. The diagram consists of a collection of spaces with no maps between them; there are only the identity maps. The colimit of this diagram is then just the disjoint sum of these spaces. This should be clear from the definition, because the equivalence relation in the formula is induced by maps, which we do not have in this case.*

Example 4.1.3 (Pushout)**.** *A pushout is a colimit where diagram of spaces is of the form*

$$\begin{array}{ccc} Z & \xrightarrow{f} & X\ . \\ \downarrow{\scriptstyle g} & & \\ Y & & \end{array}$$

The pushout of this diagram is often denoted $X \sqcup_Z Y$. When both maps f and g are inclusions, i.e. $Z \subset X$ and $Z \subset Y$, then the colimit is just the union $X \cup Y$ (Figure 4.1).

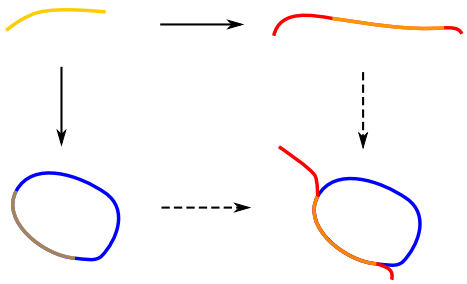


Figure 4.1: The pushout of a diagram where a line segment maps to two different spaces separately. The image of a space under each map has the same colour as the space. If we assume the solid arrows are inclusions, then we can view the colimit as the union of the line and the circle.

Another nice pushout is the attachment of a cell to a space (Figure 4.2). To get this, take the standard inclusion $g : S^k \hookrightarrow D^{k+1}$ and a map $f : S^k \to X$, to get the space which we denote $X \cup_f D^k$. This is the space X where we attach a cell D^k along its boundary to X, where the way its boundary attaches to X is given by the map f.

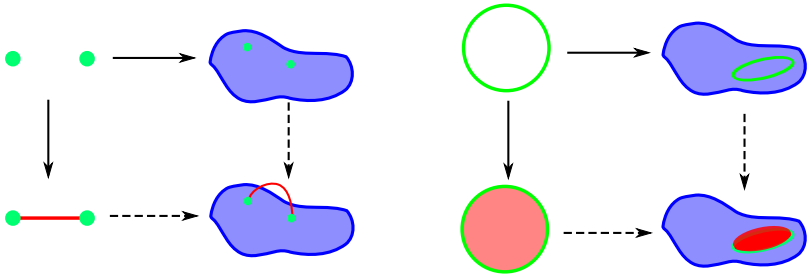


Figure 4.2: Two pushouts resulting in the attachment of a cell to a space. The attached cell on the left is of dimension one, and on the right a 2-dimensional cell is attached.

Example 4.1.4 (One map)**.** *An uninteresting, but important example here, is the colimit of the diagram $X \to Y$. The colimit of this diagram is just the space Y again. This example is important because the corresponding homotopy colimit we will see later is actually interesting.*

Now the interesting thing is that, while this construction is quite universal when we look at continuous maps, it is not very useful when we look at homotopies. By this we mean that colimits where we replace spaces by (weakly)

homotopy equivalent spaces, might not give (weakly) homotopy equivalent colimits.

Example 4.1.5 (Where the colimit fails)**.** *We can take the two pushouts*

$$\begin{array}{ccc} S^{k-1} & \longrightarrow & D^k \\ \downarrow & & \\ \{pt\} & & \end{array} \qquad\qquad \begin{array}{ccc} S^{k} & \longrightarrow & \{pt\} \\ \downarrow & & \\ \{pt\} & & \end{array},$$

where $\{pt\}$ is the space consisting of one point. Obviously D^k and $\{pt\}$ are homotopy equivalent, but the colimit of the left diagram is S^k, and the colimit of the right diagram is $\{pt\}$.

This is why we use a different but closely related construction: the homotopy colimit. In next section we look at the way homotopy colimits fix this problem.

4.1.2 Homotopy colimits

To get a homotopy colimit, we again put some spaces together in one bigger space, just like in the colimit. But now, instead of pasting them on each other directly, we put in little bridges, which lets us keep the original spaces in there. It will turn out to be the right way of putting spaces together when we want to look at homotopies of spaces.

The following definition is rather restricted in the type of categories can be used for the diagram. This is done because the definition might become unwieldy when it is given in full. To succinctly give the full definition, we would have to introduce simplicial spaces, which are far too complicated for the use we have in mind.

Definition 4.1.6. *Let $F : P \to \mathbf{Top}$ be a diagram of spaces, over a poset P. The homotopy colimit of this diagram is*

$$\operatorname{hocolim} F = \coprod_{p_0 \to \cdots \to p_n} \Delta^n \times F(p_0)/\sim,$$

where the coproduct is over all (finite) chains of composable non-identity maps in P, Δ^n is the topological n-simplex and the equivalence relation is as follows:

Let $p_0 \to \cdots \to p_n$ be a chain of composable map in P, then removing one of the p_i we get a new chain of composable maps: or $i = 0$ or $i = n$ this is clear, for $0 < i < n$, we take the map $p_{i-1} \to p_{i+1}$ to be the composition $p_{i-1} \to p_i \to p_{i+1}$. Denote this new chain by $\partial_i(p_0 \to \cdots \to p_n)$. Denote $f_i : \partial_i\Delta^i \hookrightarrow \Delta^i$ the inclusion of the ith boundary of Δ^i in the topological simplex.

Now the equivalence relation is generated by:

- *for $i > 0$ we have $(f_i(t), x) \sim (t, x)$;*

- *for $i = 0$ we have $(f_i(t), x) \sim (t, F(p_0 \to p_i)x)$.*

This definition seems hard to work with, but the idea is simple. We can imagine the homotopy colimit to be a copy of each of the spaces $F(p_i)$, where for each map $p_i \to p_j$ we attach a cylinder $F(p_i) \times I$ on the one end to $F(p_i)$ and on the other end to $F(p_j)$ by the map $F(p_i \to p_j)$. In other words, we glue in the mapping cylinder of this map. For composable chains we instead glue in 'higher dimensional mapping cylinders', which on the boundaries correspond to the mapping cylinders. Let us look at the mapping cylinder as a homotopy colimit again.

Example 4.1.7 (Mapping cylinder)**.** *Let $X \to Y$ be a diagram of spaces, then the homotopy colimit is the space*

$$X \sqcup I \times X \sqcup Y / \sim,$$

where the equivalence relation attaches one side of the cylinder by the identity map to X, and the other side by the map $X \to Y$ to Y. This is just the mapping cylinder of the map $X \to Y$.

We have to admit one thing, 'the' homotopy colimit does not exist in the category of spaces. Technically the homotopy colimit is defined in a slightly different category, where the objects are homotopy classes. This means the homotopy colimit is actually a class of spaces, of which the one in our definition is one example. If we talk about the homotopy colimit in this thesis, we talk about the choice we made in this definition.

A simple example of what the construction 'does' can be found in Figure 4.3. There we see the colimit and the homotopy colimit of a diagram of the form

$$\begin{array}{ccc} U & \hookleftarrow & V \\ & & \downarrow \\ & & V \end{array} \hookrightarrow U \ .$$

The colimit of this diagram is just the union $U \cup V$, whereas the homotopy colimit fits in a sort of bridge. Keep this example in mind because it is the main use of the homotopy colimit in the proof of the nerve theorem we will see later.

The 'bridges' in the homotopy colimit are important to keep the homotopy properties even if we choose a diagram with homotopy equivalent spaces. To illustrate this, let us return to the example of last section (example 4.1.5).

Example 4.1.8 (Where the homotopy colimit works)**.** *Take the same diagrams of spaces we had in example 4.1.5, and let us look at their homotopy colimits. Instead of using the equivalence relation directly, we use it only on the ends of the cylinders. For example, the map $f : S^{k-1} \to \{pt\}$ gives a cylinder $I \times S^{k-1}$ where the quotient $f(x) \sim x$ is used only on one end, e.g. $(0, x) \sim f(x)$. In Figure 4.4 the resulting homotopy colimits are drawn. Note that both are homotopy equivalent (even homeomorphic) to S^k.*

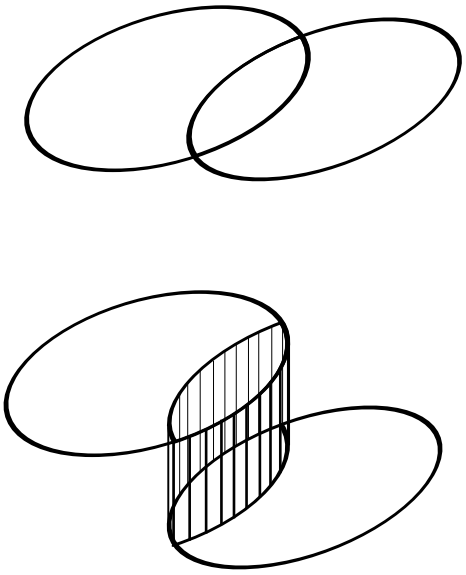


Figure 4.3: The colimit and the homotopy colimit of a diagram with two inclusions: of the intersection of two balls into the two balls.

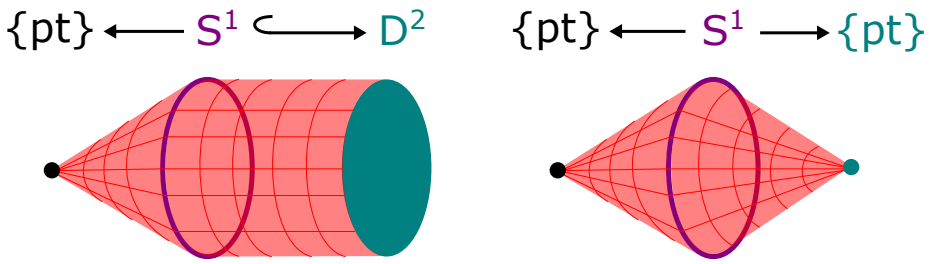


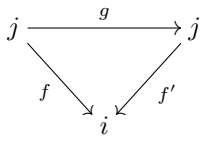
Figure 4.4: The homotopy colimits of the two diagrams of example 4.1.5, the left space corresponds to the left diagram and the right space to the right diagram. Note that the homotopy colimits of these diagrams are homotopy equivalent.

Categorical definition

Let us now look at an alternative definition, which is less constructive of nature, and more categorical. As such, it might also be harder to understand, but it is easier to work with, as it will give us a universal property to work with. To do so, we have to introduce two new constructions related to categories and simplicial complexes. The first is the over category, a way to construct a new category from an old one by looking at it from one of the objects. The second is the classifying a small category, which is a simplicial complex in the case that I is a poset.

Definition 4.1.9. *Let I be a small category and i an object of I, then the over*

category $i \downarrow I$ consists of objects: pairs $(j, f : i \to j)$ of an object j of I and a map $f : i \to j$. The arrows between (j, f) and (j', f') are maps $g : j \to j'$ such that the diagram



commutes.

If I is a small category, then the classifying space of I denoted $B(I)$ or BI is a topological space constructed from all the chains of compositions in I. Again, to define this for all small categories, we have to introduce simplicial sets. We will hence restrict to posets once more, to make the definition easier. It is helpful to keep in mind that a similar definition works for arbitrary small categories.

Definition 4.1.10. *Let P be a poset, the classifying space of P, denoted $B(P)$ is the homotopy colimit of $F : P \to$ **Top** with $F(p) = \{pt\}$ for all objects p. Equivalently, it is the simplicial complex with n-simplices $(p_{i_0}, \ldots, p_{i_n})$ for all chains $p_{i_0} \to \cdots \to p_{i_n}$.*

Even though the relation is shaky at the fundamental level, we want to note the following: the (nerve of an) over category of an object bears many similarities with the link of a vertex in a simplicial complex. There is a big difference however: in the over category we only consider outgoing arrows, where the link cannot make any distinction between edges connected to the vertex.

To practice with the definitions we prove the following lemma about the relation between different over categories of one poset.

Lemma 4.1.11. *Let P be a poset and $f : p \to q$ an arrow in P. Then there is an injection $f^* : B(q \downarrow P) \to B(p \downarrow P)$ induced by the map f.*

Proof. For any small category C and a map $f : p \to q$ in C there is the induced map (functor) $f^* : q \downarrow C \to p \downarrow C$ given by

$$f^*(r, g : q \to r) = (r, g \circ f : p \to r).$$

In general, this does not give an injection on the level of classifying spaces, we will prove that it does for $C = P$ a poset.

Because in a poset there is at most one map between two objects, there is at most one object in $i \downarrow P$ with first entry j for all i, j objects of P. Therefore, we can suppress the second entry of the objects of $i \downarrow P$ and still remain with all the information. Note that f^* is a map on pairs, which is the identity map on the first entry of the pair. This means that the map f^* is injective on the objects. In our representation with the second entry suppressed, f^* sends $r \in \mathrm{Ob}(q \downarrow P)$ to $r \in \mathrm{Ob}(p \downarrow P)$.

The simplicial complex $B(q \downarrow P)$ is then quite simply the subcomplex of $B(p \downarrow P)$ on the vertices r such that there is a map $q \to r$ (this includes q of course). In fact, both can be seen as subcomplexes of $B(P)$ in the same way.

The reason the statement about injectivity does not hold for arbitrary small categories, is that there, we may have two different maps $f, f' : p \to q$ such that $f \circ g = f' \circ g$.

With these constructions, we can easily define the homotopy colimit in a more categorical setting. For convenience, we still restrict to posets instead of small categories.

Definition 4.1.12. *Let $F : P \to \mathbf{Top}$ be a diagram of spaces over a poset P, then the homotopy colimit of F is the coequalizer of*

$$\coprod_{f:i\to j} F(i) \times B(j \downarrow I) \underset{\psi}{\overset{\phi}{\rightrightarrows}} \coprod_i F(i) \times B(i \downarrow I) \ .$$

The maps are the expected inclusions on the summands of the coproduct:

- *On summand $F(i) \times B(j \downarrow I)$ indexed by $f : i \to j$, we take $\phi = F(f) \times \mathbb{1} : F(i) \times B(j \downarrow I) \to F(j) \times B(j \downarrow I)$ composed with injection into the second coproduct;*
- *On summand $F(i) \times B(j \downarrow I)$ indexed by $f : i \to j$, we take $\psi = \mathbb{1} \times f^* : F(i) \times B(j \downarrow I) \to F(j) \times B(i \downarrow I)$ composed with injection into the second coproduct.*

This definition is (at least at first sight) quite similar to a general definition (viz. [7] Definition 8.1). Unfortunately, the general definition makes use of model categories and simplicial sets, so we cannot fully explain the similarities and differences here. We can, however, show that both our definitions are equal, which is useful because it gives us multiple ways to approach homotopy colimits.

Proposition 4.1.13. *The two definitions of the homotopy colimit given above are equivalent.*

Proof. Let us denote the homotopy colimit from the first definition with hocolim and the second one with hocolim$'$. We will show that both constructions ultimately use the same components and gluing.

Note that in the definition of hocolim$'$ the maps ψ are injective because of Lemma 4.1.11, and that we can see $B(q \downarrow P)$ as a subcomplex of $B(P)$ for all objects q. This means that in the colimit (coequalizer), all the summands of the left coproduct can be found as subspace of some summand of the right coproduct. This definition of the homotopy colimit will therefore be a quotient of the coproduct on the right.

Going through the definitions, we see that equivalence relation of the quotient is encoded in the maps $F(f : i \to j)$ in the following way:

$$(x, t) \sim (F(f)(x), t)$$

for all $t \in B(j \downarrow P)$ and $x \in F(i)$. Now we note again that $B(q \downarrow P)$ can be seen as the subcomplex of $B(P)$ on the vertices r such that there is an arrow

$q \to r$. We write the subcomplex $B(q \downarrow P)$ as union of simplices containing q as vertex, and get the following equivalent definition (to hocolim$'$) of the homotopy colimit:

$$\coprod_{p_0 \to \cdots \to p_n} F(p_0) \times \Delta^n / \sim,$$

where the equivalence relation glues the subcomplexes back together (bullet one of the equivalence relation in the definition of hocolim), and uses the equivalence relation above. Note that on the simplices this equivalence relation becomes

$$(x, \alpha_*(t)) \sim (F(f)(x), t),$$

where f is the map $f : p_0 \to p_1$, and $\alpha : (p_0 \to \cdots \to p_n) \mapsto (p_1 \to \cdots \to p_n)$ induces the inclusion α_* of the 0-th boundary of the simplex. Note that this part of the equivalence relation encodes the second bullet of the definition of hocolim. We conclude that the two definitions are equal.

Now let us turn to some important properties of homotopy colimits, which we will use in the proofs in next section about the nerve theorem.

4.1.3 Important properties

For the use of homotopy colimits in the proof of the upcoming nerve theorem, we need a few more facts about homotopy colimits. These facts come in the form of two lemmas: one which states that swapping spaces in a diagram for homotopy equivalent ones does not change the homotopy colimit; and one that says that in particular cases, the homotopy colimit projects as a homotopy equivalence to the colimit.

The proofs of the lemmas follow the course of the proofs in Kozlov's book [13], we just fill in the details. In Hatcher's book [6] equivalent statements with proofs can be found, but their notation is somewhat tedious, as everything is written in terms of mapping cones, including the homotopy colimits.

Note that the proofs only hold for our limited definition of homotopy colimits. More general results can be stated (and poved in a comparable way) if one is prepared to learn about simplicial sets and simplicial spaces. Note here that a simplicial space is not a simplicial complex, but a more general concept set in the world of categories and functors. It goes beyond the scope of this thesis to even introduce the needed terminology for treating homotopy colimits in this sense rigorously.

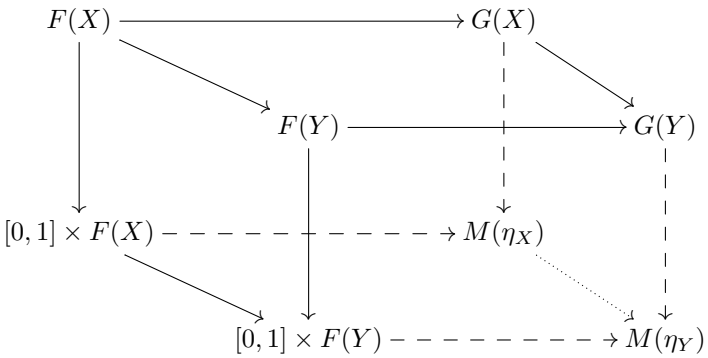
Lemma 4.1.14. *Let $F : P \to$ **Top** and $G : P \to$ **Top** be diagrams of spaces, i.e. functors from the indexing poset P to the category of topological spaces. Suppose $\eta : F \to G$ is a natural transformation which consists of only homotopy equivalences $\eta_p : F(p) \simeq G(p)$. Then the homotopy colimits*

$$\operatorname{hocolim} F \simeq_{\operatorname{hocolim}(\eta)} \operatorname{hocolim} G,$$

are also homotopy equivalent.

Proof. Note $\operatorname{hocolim}\eta$ is a homotopy equivalence if and only if $\operatorname{hocolim} F$ is a strong deformation retract of $M(\operatorname{hocolim}\eta)$. We will construct such a deformation retract inductively over the skeleton of B(P). This is possible because $M(\operatorname{hocolim}\eta) \cong \operatorname{hocolim} H$, where $H : P \to \mathbf{Top}$ is a diagram over the same poset P consisting of mapping cylinders.

Concretely we have $H(p) = M(\eta_p)$, and the maps $H(p) \to H(q)$ for $p \to q$ in P, are the maps which we get from the universal property of the mapping cylinder as a pushout:



Both the back and the front square are pushout diagrams. The solid arrows from the back to the front are induced by η_X and the dotted arrow is the map we get using the universal property of the square with Xs.

Now we use our knowledge of the maps to see that indeed

$$\operatorname{hocolim} H \cong M(\operatorname{hocolim}\eta),$$

where $\operatorname{hocolim}\eta : \operatorname{hocolim} F \to \operatorname{hocolim} G$ is the map induced by the maps

$$\mathbb{1} \times \eta_{p_0} : \Delta^n \times F(p_0) \to \Delta^n \times G(p_0),$$

for all chains of maps $p_0 \to \cdots \to p_n$ in the index category. This makes it clear that we can see $M_{\operatorname{hocolim}\eta}$ as a bunch of mapping cylinders glued together exactly as in $\operatorname{hocolim} H$.

Now that we can work with $M(\operatorname{hocolim}\eta) \cong \operatorname{hocolim} H$ as a space over $B(P)$, we want to see how it looks over parts of the skeleton of $B(P)$. For the induction, it turns out to be useful to define H^n: the part of $\operatorname{hocolim} H$ above the n-skeleton of $B(P)$ together with $\operatorname{hocolim} F$, i.e.:

$$H^n = \operatorname{hocolim} F \cup \coprod_{p_0 \to \cdots \to p_k, k \leq n} F(p_0) \times \Delta^k / \sim,$$

with the equivalence relation as for the homotopy colimit. We are now ready to use induction on H^n with $n \geq 0$.

Basis The inclusion of $F(p)$ in mapping the cylinder $M(\eta_p)$ is a deform retract iff $F(p) \simeq_h G(p)$. This holds for all objects p, as this is the condition in the lemma. Hence H^0 deform retracts onto $\operatorname{hocolim} F$ above all the vertices, which is everywhere.

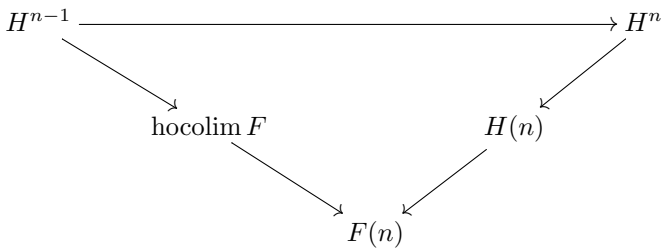
Hypothesis H^k deform retracts onto H^{k-1} and onto hocolim F for all $0 \leq k \leq n-1$

Step We now prove that H^n strong deformation retracts onto H^{n-1} and onto hocolim F.

Note that $g : H^{n-1} \to H^n$ is a closed inclusion and a neighbourhood deformation retract (NDR). Hence it is a strong deformation retract iff the inclusion is a homotopy equivalence. This is a basic fact of topology as can be found in Hatcher's.

As the deformation retract can be done separately over n-simplices, we just work over an n-simplex (with vertices $\{0, 1, \ldots, n\}$) now, and the deform retract is defined in the same way over each n-simplex.

Note by induction hypothesis: $H^{n-1} \simeq_f$ hocolim F, where f^{n-1} is the inclusion. By definition of the homotopy colimit, hocolim F deform retracts onto $F(n)$ the space above vertex n. This is clear once we realize we are working over an n-simplex. For the same reason, H^n deformation retracts onto $H(n) = M(\eta_n)$, where $\eta_n : F(n) \to G(n)$ is the relevant map above vertex n. By the condition of the lemma, this is a homotopy equivalence, hence $H(n)$ deform retracts to $F(n)$. It is clear that the diagram:



with the inclusion on the top and the deformation retracts downwards, commutes. Hence the inclusion $H^{n-1} \to H^n$ is a homotopy equivalence, and by the comment about the NDR earlier, it is a strong deformation retract. Also, because hocolim F is a strong deformation retract of H^{n-1}, it is a strong deformation retract o H^n, too.

By concatenating the (possibly infinite) number of strong deformation retracts, we get hocolim F as a deformation retract of $M(\text{hocolim}\,\eta)$, and hence that hocolim η is a homotopy equivalence.

Lastly in this section, we define the nerve diagram, which is very much related to the nerve as we saw earlier. This relation will become clear when we treat the nerve theorem.

Definition 4.1.15. *Let $\{U\}_{i\in I}$ be an (open) cover, then the nerve diagram of this cover is the poset consisting of objects $(i_0, \ldots, i_k)$ such that*

$$\cap_{j=0}^{k} U_{i_j} \neq$$

and arrows $(i_0, \ldots, i_k) \to (i_0, \ldots, \hat{i}_j, \ldots, i_k)$ for all $0 \leq j \leq k$ representing the inclusions.

Next lemma is preparation for the proof of the nerve theorem. It is, in particular, one of the homotopy equivalences that we use sequentially to prove this theorem.

Lemma 4.1.16. *Let X be a paracompact space, $\{U\}_{i\in I}$ a locally finite open cover of X, and let N be the nerve diagram of the cover. This gives a diagram of spaces $F : N \to$ **Top**, and we have the following homotopy equivalence:*

$$\operatorname{hocolim} F \simeq \operatorname{colim} F = X.$$

Proof. There is the canonical projection $p : \operatorname{hocolim} F \to \operatorname{colim} F$. This projection can most easily be imagined by noting that all elements in $\operatorname{hocolim} F$ can be represented by $((\sigma, t), x)$, where $\sigma = (i_0, \ldots, i_k)$ represents the intersection of some cover elements, t in the interior of Δ^k (drop the word interior if $k = 0$), and $x \in \cap_{0\leq j\leq k} U_{i_j}$. The projection map p is then just the map sending $((\sigma, t), x)$ to x. This means the inverse of x for x contained in exactly $k + 1$ of the U_i, say $U_{i_0}, \ldots, U_{i_k}$, consists of all simplices in the nerve corresponding to faces of σ.

The idea now is to find a nice section of the projection map. This is done using the fact that the space is paracompact, and therefore there is a partition of unity respecting the cover. Let $\phi_i : X \to \mathbb{R}$ be this partition of unity, and set

$$s(x) := \Big((\{i \in I | x \in U_i\}, \sum_{i\in I} \phi_i), x \Big).$$

This map is continuous because the partition of unity is, and for each $x \in X$ it picks an element in the pre-image $f^{-1}(x)$, so it is a section. It can also easily be checked that the contraction of Δ^k to the point $\sum_{i\in I} \phi_i$ can be done in each fiber simultaneously, which gives us the homotopy equivalence.

4.2 The nerve theorem

In this section we discuss the nerve theorem, which relates the nerve of a cover of some space to the space. If the cover is nice enough, we can recover the space from the nerve. An easy condition for this 'nice enough' is to take a finite cover, this is the version we will prove. Alternatively, we could choose to assume our space is paracompact, and take an arbitrary good open cover.

Theorem 4.2.1 (Nerve Theorem, Borsuk [3])**.** *For a finite open cover $\mathcal{U}$ of a paracompact space X we have the following: if every nonempty intersection of elements in U is contractible, then $\mathcal{N}(\mathcal{U}) \simeq X$.*

Later in this thesis, we will build simplicial complexes from random sets of points. One type of these complexes is built using these points and balls of fixed radius around them. Taking balls (of fixed radius) as the cover ensures that all intersections are contractible. This is the reason this theorem will be quite useful.

The proof of the theorem is quite abstract, unfortunately. Even when we try to envision the spaces that are used, we quickly have to go to high dimensional spaces, of which we can hardly produce insightful images. Hence illustrations might be lacking. The reader is still encouraged to think about the structures geometrically, although the pictures are missing here.

Proof. Let $F : N \to \mathbf{Top}$ to be the diagram of spaces over the nerve diagram N of $\mathcal{U}$ sending $(i_0, dots, i_k)$ to $\cap_{j=0}^{k} U_{i_j}$ and $G : N \to \mathbf{Top}$ sending all objects to the one element space $\{pt\}$. Using Lemma 4.1.14 and the fact that all intersections are contractible, we can easily see that $\mathcal{N}(\mathcal{U}) \cong \operatorname{hocolim} G \simeq \operatorname{hocolim} F$. Next, by Lemma 4.1.16 we see that $\operatorname{hocolim} F \simeq \operatorname{colim} F \cong X$. Combining the two, we get $\mathcal{N}(\mathcal{U}) \simeq X$ as we wanted.

Note that we already (in Section 2.4) gave some examples of situations where not all conditions of the theorem are met, and where therefore the implication of the theorem does not hold. Note also that when we work in $\mathbb{R}^d$, and if we take balls of certain radius as the cover the conditions of the theorem will be met. Because manifolds are paracompact, the theorem also holds for manifolds, if we choose suitable covers. It might not yet be clear, but these are the reasons this theorem is very important in the study of random complexes.

4.3 Discrete Morse theory

Discrete Morse theory, first developed by Forman [5], can be seen as a tool for understanding collapses in a simplicial complex. The theory is used to find out topological properties of a given simplicial complex. In particular the homotopy type is scrutinized by finding equivalent spaces with a limited number of cells of certain dimension.

In essence, the theory can be related to regular Morse theory, but not much of that is easily retrieved by the combinatorial description in which it is usually given. Like in regular Morse theory, we can start with Morse functions, from which we get information about the space on which it is a function. In discrete Morse theory, however, we often skip this step and immediately go to 'vector fields', with corresponding critical points.

In this section we will give a quick review of some important constructions and theorems from discrete Morse theory. We will mostly follow chapter 11 from Kozlov's book [13] on combinatorial algebraic topology, and for more details the book by Jonsson on simplicial complexes [8].

4.3.1 Collapses

The way we will use discrete Morse theory, is mainly by its main theorem. This theorem tells us something about the homotopy type of our complex. We construct a sort of vector field, which tells us how we can collapse our complex to another homotopy equivalent complex. The way we can think about these collapses is as a generalization of the simplicial collapse in the sense of Whitehead. The difference is that we also allow for internal collapses. In the proof we will only use cellular collapses, but the internal collapses provide an intuitive version of the proof.

We will only be concerned with collapses of cells to faces of codimension 1. We hence introduce notation to indicate that a cell is a codimension 1 face of another cell. With codimension one face we of course mean a face of the cell which is of one dimension lower.

Definition 4.3.1. *Let $\sigma \subset \tau$ be cells in a complex, and let σ be a codimension 1 face of τ, i.e. $|\sigma| = |\tau| - 1$. Then we write $\sigma \prec \tau$, or equivalently $\tau \succ \sigma$.*

We now turn to collapses. The simplest of those are the elementary collapses, which intuitively remove parts of the 'outside' of the complex. For these collapses, it is easy to see that the homotopy type of the complex does not change.

Definition 4.3.2. *Let σ, τ be a pair of cells in a complex $\mathcal{K}$ such that $\tau \prec \sigma$, σ a maximal cell (i.e. there is no other cell that contains σ), and σ is the only cell that contains τ. Then we call τ a free face of σ. For any such pair we have the elementary collapse of $\mathcal{K}$, which is the complex*

$$\mathcal{K}_{\tau \prec \sigma} := \mathcal{K} \setminus \{\sigma, \tau\}.$$

This elementary collapse is well defined because of the restrictions on the pair σ, τ from the definition. Removal of these two simplices does not change anything about the attachments of the other cells, because σ is maximal, and τ is no part of any attachment except that of σ.

Note also, that the homotopy type of the complex will not change because the inclusion of the horn Λ^k_i (the union of all faces of Δ^k except for the i-th) in the simplex Δ^k is a deformation retract. This induces a deformation retract for $\mathcal{K}_{\tau \prec \sigma}$ in $\mathcal{K}$.

To imagine what is happening, we also want to consider collapses of cells which are not 'on the outside.' These collapses make it possible to reduce the number of cells of the complex, even if there are no free faces. The following is a definition for these internal collapses.

Definition 4.3.3. *Let σ, τ be cells in a complex $\mathcal{K}$ with cells C, such that $\tau \prec \sigma$, and some additional conditions hold. Then the internal collapse of $\mathcal{K}$ w.r.t. these cells is the complex $\mathcal{K}^{\tau \prec \sigma}$, which is the complex with cells $C \setminus \{\sigma, \tau\}$, and attaching maps induced by the retraction of σ to $\delta\sigma \setminus \tau$.*

Note that an internal collapse may take us away from the class of simplicial complexes, to the more general class of cell (CW-) complexes. An example can be found in Figure 4.5. Also, the structure of the new complex is not as obvious as in the case of the elementary collapse, where we could just remove the two cells. For internal collapses, we have to patch things together, simply removing the two cells might not even result in a cell complex! In the definition, this is taken care of by redefining the attaching maps slightly. A collapsing cell will be considered to retract to other cells, and if a cell is attached to a collapsing cell, it will be attached to these other cells after the collapse (Figure 4.6).

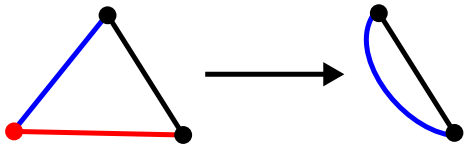


Figure 4.5: An example of an internal collapse giving a non-simplicial complex.

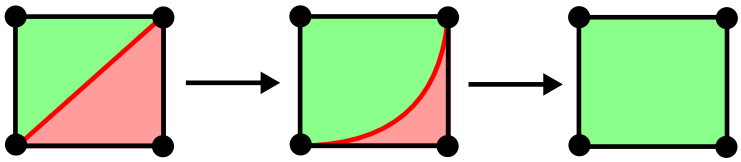


Figure 4.6: An example of the 'reattachment' for internal collapses.

It is not directly clear when internal collapses change the homotopy type of the space, hence they are not suitable for use in a proof. It turns out that the internal collapses can be done in most cases that we encounter: in particular the case of doing multiple internal collapses in a simplicial complex, where any cell is part of a collapse at most one time in a nice way. We explain this in next section.

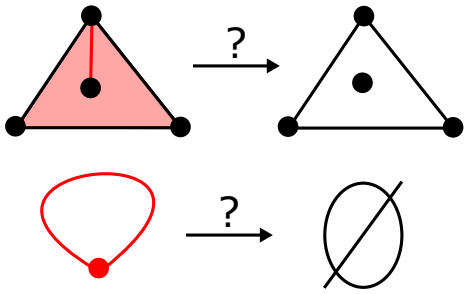


Figure 4.7: An example of how internal collapses can give spaces which are not homotopy equivalent to the original.

4.3.2 Discrete vector field

In the main theorem, we will want to do multiple collapses. For a simplicial complex, picking any pair of cells $\tau \prec \sigma$, gives a nice collapse which does not change the homotopy type. However, for general cell complexes, this is not the case. Take for example the circle, made from on 0-cell and one 1-cell. Collapsing the 1-cell to the 0-cell cannot, in any way, give a space homotopic to the circle. Because internal collapses can take us from simplicial complexes to general cell complexes, we need to be careful. Hence we will need to put some restrictions on the set of collapses that we do. These restrictions are phrased most commonly in terms of discrete vector fields.

Definition 4.3.4. *Let $\mathcal{K}$ be a simplicial complex. A discrete vector field is a collection of pairs of cells $\{\alpha \prec \beta\}$ such that each cell of $\mathcal{K}$ is in at most one of the pairs. A cell that is not paired is called critical.*

There is an alternative way of looking at these vector fields which might help visualize the next definitions. For this alternative, we need to see the simplicial complex as a directed graph.

Definition 4.3.5. *The Hasse diagram of a simplicial complex $\mathcal{K}$ is the directed graph where the vertices are the simplices of $\mathcal{K}$, and the edge set consists of all pairs of simplices $(\alpha \prec \beta)$.*

A discrete vector field is now just a choice of edges in the Hasse diagram, which induces a disconnected subgraph. All isolated nodes in this subgraph correspond to critical cells. For the purposes of Morse theory we want a more restricted kind of vector field, which is called an acyclic discrete vector field.

Definition 4.3.6. *A discrete vector field V is called acyclic if there is no sequence*

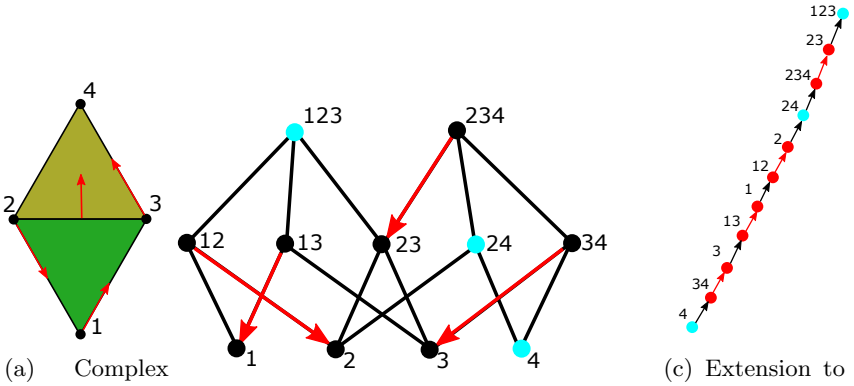
$$\alpha_0 \prec \beta_0 \succ \alpha_1 \prec \cdots \succ \alpha_i \prec \beta_i \succ \alpha_0$$

with $\alpha_j \neq \alpha_{j+1}$ and $\{\alpha_j \prec \beta_j\} \in V$.

The name acyclic vector field is visualised easiest in the Hasse diagram: Take the Hasse diagram of a complex, and invert all the arrows corresponding to a pair in the discrete vector field. The Discrete vector field is acyclic if the resulting directed graph is acyclic. It is easy to see that this gives the same definition as above.

We can now use a nice property of the directed acyclic graph that we got. We can extend the order this graph gives to a linear order, such that paired nodes 'stay together'. This linear order will give us the necessary data to prove the main theorem of discrete Morse theory: it gives the order in which we do collapses.

Lemma 4.3.7. *Let $G = (V, E)$ be a directed acyclic graph, and let $H \subset E$ be a set of arcs that have no vertex in common. There exists a linear order of the vertices which is an extension of the partial order induced by G, and in which for any $(v, w) \in H$ and $u \in V$ we have: if $v < u < w$, then $v = u$ or $w = u$.*



(a) Complex with pairing (b) Hasse diagram with inverted arrows (c) Extension to linear order

Figure 4.8: The Hasse diagram with matching extended to a linear order.

Proof. Note that the directed acyclic graph and the subset H correspond to a poset P with acyclic vector field H by a minimal extension of the order given by the arcs. Because the vector field is acyclic, it is possible to define a 'quotient poset' Q with elements all pairs H and all critical cells $C = \{v \in V | v \notin e \text{ for all } e \in E\}$. The order in this poset is given by the order of the poset P. It is easy to check that this order is well defined. Now it is generally known that we can extend a poset to a linear order. Any such extension suffices.

4.3.3 Main theorem of discrete Morse theory

The main theorem of discrete Morse theory can be compared to normal Morse theory, where critical points of a vector field are indicative of the topology of the space. Here, too, the critical points are used to infer information about the complex, but the vector field is not assumed continuous, it is assumed to be a discrete vector field. Finding a nice discrete vector field hence gives us a lot of information about the complex.

Theorem 4.3.8. *Let $\mathcal{K}$ be a simplicial complex with an acyclic discrete vector field V, then $\mathcal{K}$ is homotopy equivalent to a cell complex with as many cells of dimension k as there are critical cells of dimension k in V.*

The following basic topological lemma tells us that homotopic attaching maps for a cell give homotopy equivalent spaces. This lemma will be crucial in the proof of the main theorem, as we want to build up the complexes, but only in a homotopy equivalent way. The proof can be found in most standard works that include a section about homotopy.

Lemma 4.3.9. *Let X and Y be homotopy equivalent spaces with homotopy $f : X \to Y$, and let $g : S^{k-1} \to X$ be the attaching map of a k-cell to X. The spaces $X \cup_g D^k$ and $Y \cup_{f \circ g} D^k$ are also homotopy equivalent.*

4.3.8. By Lemma 4.3.7 there exists a linear order of the simplicies of $\mathcal{K}$ which agrees with the adjusted Hasse diagram. For the proof we use induction on the simplices, ordered by this linear order. Note that the first simplex will always be a vertex of the complex, as any higher dimensional cell has at least one incoming edge from below.

Define for this proof the i^{th} complex $\mathcal{K}_i$ to be the complex consisting of all simplices in the first i pairs/critical cells.

Basis The first simplex in the order is critical, otherwise it could not have been the first simplex. The complex consisting of this vertex only, which we call $\mathcal{K}_1$, is obviously homotopy equivalent to a complex with one 0-cell. Call this complex Θ_1.

Hypothesis The complex $\mathcal{K}_i$ is homotopy equivalent to a complex Θ_i with c_k cells of dimension k, for all $i < n$.

Step We prove that $\mathcal{K}_n \simeq \Theta_n$ for a Θ_n that has appropriate numbers of cells in each dimension, if $\mathcal{K}_i \simeq \Theta_i$ for all $i < n$ and Θ_i with similarly appropriate numbers of cells. To do this, we distinguish two cases: $\mathcal{K}_n \setminus \mathcal{K}_{n-1}$ can either be a critical cell, or a pair of cells from the discrete vector field.

In the first case, of the critical cell, we do the following. Say the homotopy equivalence of $\mathcal{K}_{n-1}$ and Θ_{n-1} is given by a map $g : \mathcal{K}_{n-1} \to \Theta_{n-1}$, and that the attaching map of the critical cell (of dimension k) is the map $f : \delta\Delta^k \to \mathcal{K}_{n-1}$. Then, by Lemma 4.3.9, we know that

$$\mathcal{K}_n \cong \mathcal{K}_{n-1} \cup_f \Delta^k \simeq \Theta_{n-1} \cup_{g\circ f} \Delta^k =: \Theta_n,$$

where Θ_n has the correct number of cells in each dimension.

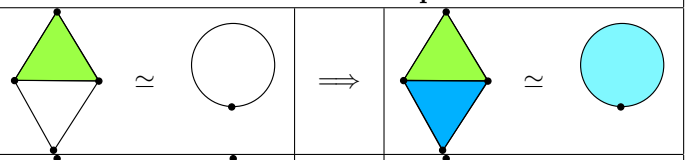
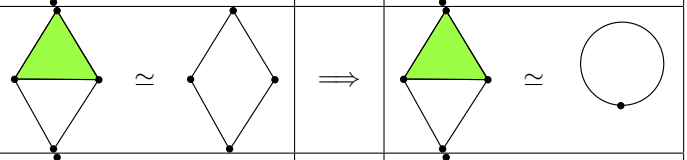
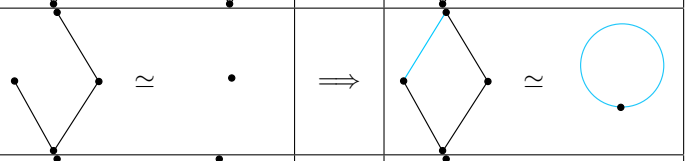
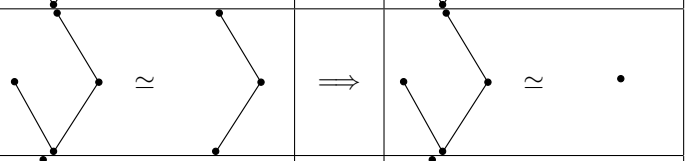
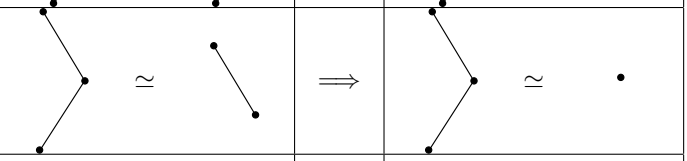
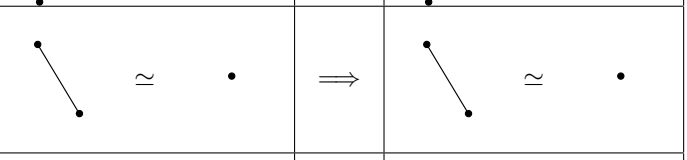
In the second case, of the pair of cells $(\tau \prec \sigma)$, we use an elementary collapse. Note that we can restrict the vector field to the complex $\mathcal{K}_n$. In this vector field, the chosen pair is largest in the extended linear order. This implies that $(\tau \prec \sigma)$ is a free pair in $\mathcal{K}_n$: if $\upsilon > \sigma$, then obviously the pair is not largest in the order, hence σ is maximal; and if $\upsilon > \tau$ and $\upsilon \neq \sigma$, then we get $\upsilon > \sigma$ which gives the same contradiction, so τ is a free face of σ. Now, because the pair is free in $\mathcal{K}_n$, we can do an elementary collapse and get a homotopy equivalent space. This homotopy equivalent space is the space $\mathcal{K}_{n-1}$ which we get by just removing the pair. Hence

$$\mathcal{K}_n \simeq \mathcal{K}_{n-1} \simeq \Theta_{n-1} =: \Theta_n$$

defines the space Θ_n with the right number of cells in each dimension.

Note that also more abstract proofs are possible, which work for arbitrary (?) cell complexes. These however, loose some of the intuitive arguments, and use more abstract algebraic topology. We follow the method with internal collapses because the resulting theorem is strong enough for the purposes in this thesis, and the intuitive/visible arguments are preferable above abstract ones (even sometimes when the abstract ones are shorter).

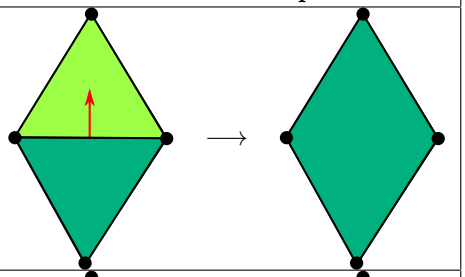
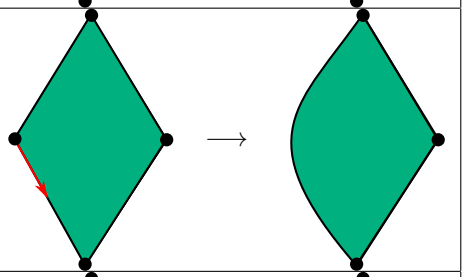
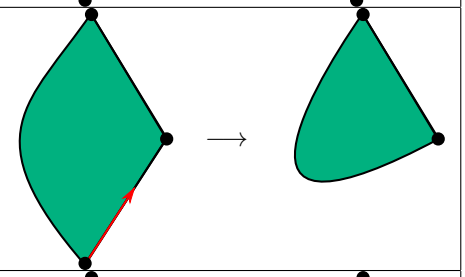
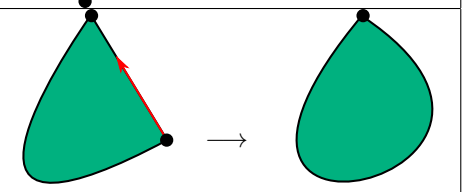
Table 4.1: The inductive elementary collapses and attachments of cells like in the theorem for the matching in Figure 4.8. Read the diagram from the bottom to the top to understand the homotopy equivalence.

Step	Action	Induction step
1	add critical cell 123	
2	collapse (23,234)	
3	add critical cell 24	
4	collapse (2,12)	
5	collapse (1,13)	
6	collapse (3,34)	
7	add critical cell 4	$\emptyset \simeq \emptyset \Longrightarrow \bullet \simeq \bullet$

It is instructive to look at figures 4.1 and 4.2 for an example of the steps in

the theorem. In the first of these figures, we see what actually happens in the theorem. In the second one we see the link with internal collapses. Although it is not directly clear that internal collapses are allowed, the proof of the theorem implies they are allowed in the cases we encounter by going through the steps of the theorem.

Table 4.2: The internal collapses for the matching in Figure 4.8. Note that the collapses give homotopic spaces because they can be produced from homotopic spaces by attaching the critical cells as seen in Figure 4.1.

Step	Internal Collapse
2	
4	
5	
6	

Part II

Random geometry

Chapter 5

Random Geometric Graphs

This chapter gives an overview of random geometric graphs. In particular it contains definitions and a collection of useful results. For a simplicial complex, we know that the 1-skeleton can be seen as a graph. In the same way, random geometric complexes are the 1-skeleta of random geometric complexes that we encounter in next chapter. It is therefore useful to have some results about geometric graphs ready to use, for example about the number and size of components, as those properties are directly useful for studying complexes. Most results from this chapter are from Penrose's "Random gometric graphs."[14]

Random geometric graphs can behave in quite a lot of ways. A good predictor of certain behaviour is the 'limiting density' of the points, i.e. the number of points we expect to see in some area when we take increasingly many points. Because points get connected when they are close together, this density gives us an idea of how connected the vertices of the graph are. If there are only a few points per very large area, we expect to see a lot of very small components. If in any small area, we already expect to see multiple points, these points will likely be connected, and thus we expect multiple large components; and when the density becomes large enough, we even expect to see only one large component. These density limits are called regimes, and we will shortly talk about these.

Lastly, we will look at the expected number of specific kinds of components. These are important for the homology of random complexes that we will study later. The graph components can already tell us a lot about the possibility of it supporting holes.

5.1 Geometric graphs

A geometric graph is a graph built from geometric information: a set of points in a metric space and a fixed radius r. If points are close enough together, then they are connected by an edge. In terms of balls, this means that points whose balls of radius r intersect, get connected (Figure 5.1). Because later, for geometric complexes, the definitions are easiest when we phrase everything in

terms of these balls and intersections, we will do the same here.

Definition 5.1.1. *Let X be a set of points in a metric space and $r \geq 0$ a radius. This information defines a geometric graph $G(X, r)$ with vertex set X and edges $\{x_1, x_2\}$ if and only if $B(x_1, r) \cap B(x_2, r) \neq \emptyset$.*

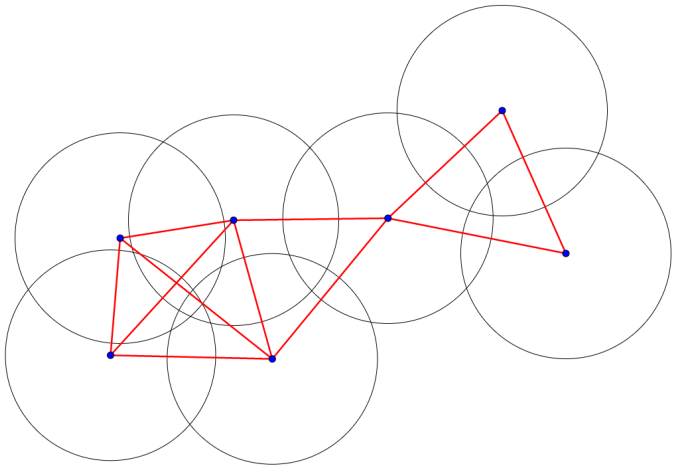


Figure 5.1: The geometric graph $G(X, r)$ for some set of points X and radius r.

This construction might remind one of the Nerve we saw in Chapter 2; in fact, the random geometric graph $G(X, r)$ is the same as the 1-skeleton of the nerve of the cover $\{B(x, r)\}_{x \in X}$. One of the geometric complexes studied in next chapter, in fact, is this nerve. This also explains the relevance of studying the geometric graphs here.

5.1.1 Feasible graphs

It turns out geometric graphs can be restrictive in some sense: for a fixed dimension, there are graphs that cannot be realised as a geometric graph in that dimension. For example: the bipartite graph $K_{1,6}$ cannot be realised in $\mathbb{R}^2$. We therefore would like to be able to indicate if a graph is geometrically realisable in such a way.

Definition 5.1.2. *A graph G is feasible in dimension d if there exists a (finite) set $X \subset \mathbb{R}^d$ and radius r such that $G(X, r) \simeq G$.*

The following proposition is obvious, but useful.

Proposition 5.1.3. *If G is feasible in dimension d then it is feasible in any dimension $d' \geq d$.*

As a generalisation of the example we want to look at a certain class of graphs and in which dimension they are feasible. The class of graphs we are interested in is the class of star graphs.

Definition 5.1.4. *The star graph $\star_n$ is the graph with vertex set $V = [n+1]$ and edges*

$$E = \{(n+1, i) | i \in [n]\}.$$

In other words $\star_n := K_{1,n}$.

In our example before, we claimed that $\star_6$ was not feasible in $\mathbb{R}^2$. In the following proposition, we give a criterion for which star graphs are feasible in a certain dimension. Note that it only gives a rather crude bound, but having this finite bound is enough for some proof later.

Proposition 5.1.5. *For every dimension d there is a natural number N_d such that the star graph $\star_n$ is feasible in dimension d iff $n \leq N_d < \infty$.*

Proof. Fix vertex $n+1$ at the origin. To be connected to this vertex, each other vertex has to be at distance at most $2r$ from the origin, in particular, they must lie in the cube $[-2r, 2r]^d$. Now partition this cube in subsets with diameter smaller than $2r$, we can do this with a finite number of subsets. For example subdivide it regularly in smaller cubes of side length less than $\frac{2r}{\sqrt{d}}$. If two of the vertices are together in one of the cubes, there is an edge between them, which is not allowed. Hence, by the pigeonhole principle, the maximally feasible star graph has less vertices than the cubes in our subdivision.

5.2 Random point processes

To get to our random graphs and complexes, we need a random set of points. The way we 'produce' these is by random point processes. These random point processes pick out a certain number of random points according to a distribution. The number of points might be fixed, or it may also be random. The cases we treat have points that are all identically distributed, and the number of points is either fixed or Poisson distributed. Of course there are more possibilities, but we restrict ourselves to these two point processes, which can be most easily imagined.

5.2.1 Formal definition: measure theoretic

The way to think about point processes may already be clear from the intro above: 'just pick a bunch of random points'. The formal definition, however, is at the first sight not equally clear. To formally define a point process, we need to put some restrictions on the ambient space, and we have to look at two 'layers of measures' (a random measure). At first the formal definition and the intuition do not seem to align, but on a closer look, we can easily see where the connection forms. This section is purely included for the reader who wants to

see how a point process can be defined abstractly, we will not actually use these definitions in the rest of this thesis.

In this section we will look at the formal definition of a point process, and we will see how it aligns with our intuition. Firstly, we need to know how we represent one outcome of the random process, then we see how it is 'made random'.

Outcomes: a set of points?

As we are working with probability theory, it is needed to phrase certain things in terms of measures. The sets of points we want to pick out, are therefore best represented by a counting measure. Suppose we want to pick out a few points in a space S. Then the points we pick can be represented by a counting measure. This can be thought about as a measure with discrete weight one at each point we picked.

Now, we also want the point sets to have some specific properties. In particular we want it to be locally finite. This means that in any compact subset, we want to find at most a finite number of points. Collecting these thoughts, we get the following definition for the set of outcomes we want for our point processes.

Definition 5.2.1. *A locally finite counting measure on the space S with Borel σ-algebra $B(S)$, is a measure μ defined by*

$$\mu(B) = \#\{B \cap P\}$$

for some locally finite set $P \subset S$.

Random outcomes

Now we want to take a random such outcome, i.e. we want a random variable. To get a random variable, we need a measure on the outcomes, and a mapping from a probability space to the this measure space of outcomes. We first define this measure space of outcomes.

Definition 5.2.2. *The space $M(S)$ of locally finite counting measures on S is made measurable by generating the σ-algebra $\rho(M(S))$ on M by all sets necessary to make*

$$\mu \mapsto \mu(B) : M \to \mathbb{N}$$

measurable for all pre-compact sets $B \in B(S)$.

The reason we want a definition as such, is that we in the end are mostly interested in counting the number of points in subsets of the space. If we can do this for enough subsets, we can actually pinpoint the locations of the points, too.

Furthermore, the probability of a set of points being in exactly one position should in most cases be 0, certainly if the distribution of each point is not

discrete. Hence the only way to make sense of a measure of random points, is to see how many there are in a certain subset. This way the probability of a certain outcome can actually be non-zero.

The last step now is to make a random variable, i.e. a random sampling of these outcomes. This is simply done by taking a probability space and a measurable map from the probability space to the measurable space of outcomes.

Definition 5.2.3. *A point process on S is a measurable mapping*

$$(\Omega, F, \mathbb{P}) \xrightarrow{\Phi} (M(S), \rho(M(S))),$$

where M, ρ as defined above and $(\Omega, F, \mathbb{P})$ is a probability space.

5.2.2 Formal definitions: specific kinds

In this thesis, we actually work with slightly less general point processes. This also means that we can take a slightly less abstract definition, which might seem less rigorous. It can, however, be proven that the definitions we will give now, are actually point processes in the formal sense discusses in last section.

Binomial point process

The first of the two random point processes that we consider is the binomial point process. It has a fixed number of points which are distributed independently. In fact this already describes the whole process, giving us the following definition.

Definition 5.2.4. *A binomial point process $\mathcal{X}_n$ ($n \in \mathbb{N}$) in a space S is a random vector $(X_1, \cdots, X_n) \subset S^n$, where each of the points X_i is independent and identically distributed with some density f on S.*

The name binomial point process is easily explained if we look at the distribution of the number of points in some set $U \subset X$. This is binomially distributed with parameter $\int_U f(x)dx$.

It is important to note that, even though the X_i are independently distributed, there is dependence in the process. It seems trivial to note that if there are k points in a subset U, then there are $n - k$ points in $S \setminus U$. In the next section we look at a different kind of point process, where even this trivial dependence is not present.

Poisson point process

Whereas the binomial point process has a fixed number of points, the Poisson point process has a random number of points. This number is a Poisson random number N with some parameter λ, i.e. $\mathbb{P}(N = n) = \frac{\lambda^n}{n!}e^{-\lambda}$. There are multiple equivalent definitions of a Poisson point process, one of which is the following,

which uses explicitly this relation with the binomial point process. In next section we will look at this relation more closely.

Definition 5.2.5. *Let f be a density function. A Poisson point process with parameter $\lambda \geq 0$ and density function f, which is denoted $\mathcal{P}_\lambda$, is defined as $\mathcal{X}_{\mathrm{Poi}(\lambda)}$.*

Note that the expected number of points in such a process is λ. The outcome of such a process and a binomial point process with λ points can hence hardly be distinguished. It will in fact turn out that we can use the Poisson point process to approximate the binomial point process. This is useful because the Poisson point process is easier to work with, as two non-overlapping regions can be considered separately. With this we mean that for $U \cap V = \emptyset$, we have the independence

$$\mathbb{E}\big(|X_n \cap U|\big) = \mathbb{E}\big(\,|X_n \cap U|\,\big|\,|X_n \cap U| = m\big).$$

Of course this independence is not directly clear from the definition. However, if we take the following definition, it is quite clear.

Definition 5.2.6. *Let X be a space and f an intensity function on X, then a Poisson point process $\mathcal{P}_\lambda$ with parameter λ is defined by the following*

- *The number of points $\mathcal{P}_\lambda \cap U$ in $U \subset X$ is distributed as $\mathrm{Poi}(\int_U \lambda f(x)dx)$ for all pre-compact Borel sets U;*
- *if $U_1, \ldots, U_k$ are disjoint, then $\mathcal{P}_\lambda \cap U_i$ are independent for $i \in [k]$.*

This definition is equivalent to the former one, if the intensity function is a density function.

For this definition the comparison with a Poisson arrival process is easy to make: Take as space the real line $\mathbb{R}$ and intensity function $f(x) = \lambda$. The number of arrivals is Poisson distributed with a parameter linear in the length (time) of the interval. In fact, such an arrival process can be viewed as a Poisson point process on $\mathbb{R}$.

The two definitions above seem quite different, but they are equivalent. For a proof we refer to any standard work about point processes.

Proposition 5.2.7. *The two definitions of a Poisson point process above are equivalent.*

Proof. Let the space be X and the distribution of the points be given by the density f on X.

Note firstly that the number of points in both definitions has the same distribution, because $\int_X \lambda f(x)dx = \lambda$. Now if we condition on $\mathcal{P}_\lambda = N$, then in the first definition, we just have a binomial point process $\mathcal{X}_N$. We have to prove that the same holds for the second definition. This follows quite easily from the independence of the second definition and looking at the probability that one of the n points is in a certain subset U. In particular, it is done by proving that $\mathbb{P}(X_i \in U | \#\mathcal{P}_\lambda = n)$ is binomially distributed. 

A short remark

A lot of results hold for both Poisson point processes and binomial point processes, we therefore introduce notation for a point process that does not distinguish between the two. If we write Φ_n we either mean $\mathcal{P}_n$ or $\mathcal{X}_n$.

5.3 Random geometric graphs

Random geometric graphs are now very easily defined. Where we had a geometric graph $G(X, r)$ on a set of points X and a radius r, we just take a random set of points, so either $X = \mathcal{X}_n$ or $\mathcal{P}_\lambda$. These graphs can be interesting of themselves, but they become more interesting if we take increasingly many points and make the radius smaller and smaller. This gives the random graphs a lot of structure that we can study, for example the limiting number of components.

We will here only shortly look at such results, because we will focus on comparable results for random complexes. It is however useful to note already, that the rate at which $n, \lambda \to \infty$ and $r \to 0$ make a great difference in the structure we see in the random graphs. These rates define different regimes, and we will now first look at these.

5.3.1 Regimes

In the study of random geometric graphs we distinguish between different regimes. These regimes are related to the limiting density of the points, i.e. the number of points we expect per certain area. It is easy to imagine that if the expected number of points in a ball of radius $2r_n$ becomes very large, we will have a highly connected graph, and if it goes to zero, most points will be isolated. This expected number can be related to the density of the points as given by the limit

$$\lim_{n\to\infty} nr_n^d$$

when we have an ambient space of dimension d. The exponent d is easily explained by the fact that the density is related to the volume of the ball of radius r_n, which is of the order r_n^d.

In the regime where $nr_n^d \to 0$, which we call the sparse regime, we will see a great number of small components. This regime is therefore also called the sand grain regime. The next regime is called the thermodynamic regime, here $nr_n^d \to r \in (0, \infty)$.

This regime is the most interesting in terms of the form of the components. There is a quite sudden shift in shape if we let r raise from 0 upwards. At a certain value R percolation occurs and if $r > R$ there is suddenly almost surely one infinite component in the limit. This value for R is the critical limit. And the thermodynamic regime is split in a (thermodynamic) subcritical regime, where $\lim_{n\to\infty} nr_n^d < R$; a (thermodynamic) critical regime, where $\lim_{n\to\infty} nr_n^d = R$; and a (thermodynamic) supercritical regime, where $\lim_{n\to\infty} nr_n^d > R$.

Lastly there is the dense regime, in which $\lim_{n\to\infty} nr_n^d = \infty$. In this regime we expect that locally, all nodes are connected. If we look at a point process with uniform intensity on a connected set, we might expect to see one connected component.

5.4 Subgraph count

Certain random complexes (the Vietoris-Rips complex) complexes are determined by their underlying graph, and hence counting certain subgraphs can give us a lot of information about the topology of such complexes. Think for example of the cross-polytope of dimension k, the minimal complex with non-zero homology in degree k. The ability to count such complexes could get us a long way to determining Betti numbers, as we will do in the next chapter. In this section we will state a result about such subgraph counts in random geometric graphs, and we will follow the proof provided by Penrose [14].

5.4.1 Terminology for random geometric graphs

Of course, because we are talking about geometric graphs, all usual terminology for graphs are also used here. This means we have (induced) subgraphs, connected components, cycles, and graph isomorphisms, etc. as usual. There are, however, some terms and notation specific to geometric graphs. In particular, we will look at some stochastic variables for the number of such regular graph structures.

Definition 5.4.1. *Let $G(\Phi_n, r_n)$ be a random graph with parameter n, then we have the following random variables:*

- *The random variable $G_n(H)$ is the number of unlabelled occurrences of H as an induced subgraph of $G(\Phi_n, r_n)$.*
- *The random variable $J_n(H)$ is the number of components in $G(\Phi_n, r_n)$ isomorphic to H.*

To count these numbers, we have indicator variables for a subset $X \subset \Phi_n$:

$$h_H(X) = \begin{cases} 1 & \text{if } G(\Phi_n, r_n)[X] \cong H \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{h}_H = \begin{cases} 1 & \text{if } G(\Phi_n, r_n)[X] \cong H \text{ and } G(\Phi_n, r_n)[X] \text{ is a component} \\ 0 & \text{otherwise} \end{cases}.$$

Note that there is the following way to use the indicator to count:

$$G_n(H) = \sum_{Y \subset \Phi_n} h_H(Y)$$

$$J_n(H) = \sum_{Y \subset \Phi_n} \tilde{h}_H(Y).$$

5.4.2 Subgraph count theorem

In this subsection we will look at the theorem by Penrose about the number of times we expect to find some graph as a subgraph in a random geometric graph.

Theorem 5.4.2. *Suppose* $\lim_{n\to\infty} r_n^d n = 0$ *and* H *is a connected feasible graph on* $k \geq 2$ *nodes. Suppose also that the distribution of the points is given by bounded density function* f*. Then*

$$\lim_{n\to\infty} r^{-d(k-1)} n^{-k} \mathbb{E}(G_n(H)) = \lim_{n\to\infty} r^{-d(k-1)} n^{-k} \mathbb{E}(J_n(H)) = \mu_H,$$

where μ_H *is the following quantity:*

$$k!^{-1} \left(\int_{\mathbb{R}^d} f(x)^k \right) \int_{(\mathbb{R}^d)^{k-1}} h_H(\{0, x_1, \ldots, x_{k-1}\}) d(x_1, \ldots, x_{k-1})$$

The proof of this theorem is given in Penrose's monograph [14]. Instead of repeating this here, we give the proof of a similar theorem about the number of certain subcomplexes in Section 6.3. The proofs of these theorems go along the same lines, so it would be unnecessary to give them both in detail.

Chapter 6

Random Geometric Complexes

In this chapter the constructions of different kinds of geometric complexes are displayed. First we look at the construction of a complex from some set of points. When we are finished with the intricacies of these constructions, we will bring in the random element, just like in last chapter about geometric graphs. Whereas for geometric graphs we only gave one construction, we will here give multiple. After the introduction of the geometric complexes, we continue with the random element. This means we will continue with random point processes, the processes which supply us with the random sets of points from which we construct our random complexes. The use of the random point processes is roughly the same as for random geometric graphs, this section will hence mainly introduce notation.

6.1 Geometric complexes

A geometric complex is essentially just a simplicial complex built from geometric information about the vertices. In this thesis, this geometric information will be given by the distance between points in Euclidean space. There is a simplex on a set of points if the points are close enough together.

Of course there are different choices for such constructions, we could for example connect two points if they are close enough or, alternatively, if they are closest to each other. The first of these corresponds to the Čech and Vietoris-Rips constructions, and the second to the Voronoi construction of complexes. We will only be concerned with the Čech and the Vietoris-Rips complex, which differ only in their two- and higher-dimensional simplices. We will now give specific recipes for these constructions, after which we elaborate on each construction some more.

6.1.1 Čech

The first construction of a simplicial complex using points in a metric space is that of the Čech complex. This construction is very natural if we compare it with the nerve of a cover as defined before (Definition 2.4.1), and is easy to work with whenever the nerve has nice properties.

Definition 6.1.1. *Let $P \subset X$ a set of points in a metric space (X, d), and let $r \geq 0$ be a fixed radius. The Čech complex $\check{\mathcal{C}}(P, r)$ is the simplicial complex where $(x_0, \ldots, x_n)$ forms a simplex iff*

$$\cap_{i=0}^{n} B(x_i, r) \neq \emptyset.$$

In other words, it is the nerve of the cover $\{B(p, r)\}_{p \in P}$.

Even though the Čech complex can contain high dimensional simplices, the nerve theorem tells us that this does not complicate the homotopy type very much. The complex is weakly equivalent to the subset of the ambient space covered by all the balls. This means the homology of the complex is equivalent to the homology of this subspace.

In general, these kind of subspaces will be 'nice', and we can infer some facts about the homology of the complex from the topology of the ambient space. For Euclidean space in particular, we get the following result, which restricts the non-trivial homology of the Čech complex to certain dimensions.

Proposition 6.1.2 (Dimensionality of Čech complexes)**.** *For any Čech complex X built from a (finite) set of points in $\mathbb{R}^d$, we have $H_D(X) = 0$ for all $D \geq d$.*

Proof. If we build a Čech-complex from a set of points $P \subset \mathbb{R}^d$ and radius $r > 0$, then the nerve theorem (4.2.1) tells us that

$$\check{\mathcal{C}}(P, r) = N(\{B(p, r)\}_{p \in P}) \simeq \cup_{p \in P} B(p, r).$$

Because $\cup_{p \in P} B(p, r) \subset \mathbb{R}^d$, we know that $H_D(\cup_{p \in P} B(p, r)) = 0$ for all $D \geq d$. To see this, use the long exact sequence of the pair $(\mathbb{R}^d, \cup_{p \in P} B(p, r))$. This gives the bound on homology we want. ◭

The dimensionality mentioned here of course references the dimension of homology. It is easy to see that we can get spaces with arbitrary high 'topological dimension' as a Čech complex of points in $\mathbb{R}^2$: by clustering $n+1$ points we get an n-simplex.

6.1.2 Vietoris-Rips

In this section we look at the second kind of geometric complex we study, the Vietoris-Rips complex. Whereas the Čech complex has a natural definition as the nerve, the Vietoris-Rips complex is easily defined as a clique complex.

Definition 6.1.3. *The Vietoris-Rips complex is the clique complex of a geometric graph, which we denote*

$$\mathcal{R}(X, r_n) := X(G(X, r_n)).$$

The fact that the Vietoris-Rips complex is not the nerve of some cover, means that we cannot use the arguments for the dimensionality of the Čech complex in this case. Moreover, there is no bound on the dimension of the homology for Vietoris-Rips complexes in any dimension $d > 1$.

Proposition 6.1.4 (Dimensionality of Rips complexes)**.** *For any dimensions $d \geq 2$ and $D \geq 0$, there is a Rips complex X built from a set of points in $\mathbb{R}^d$, such that $H_D(X) \neq 0$.*

Proof. We first prove this proposition for $d = 2$. For each D we give a set of points $P_D \subset \mathbb{R}^d$ and radius r_D such that $\mathcal{R}_D := \mathcal{R}(P_D, r_D) \cong S^D$, and hence, such that $H_D(\mathcal{R}_D) \neq 0$.

Take for P_D the vertices of the regular $(2D+2)$-gon, and for r_D half the distance between two opposite vertices in the $(2D+2)$-gon. We prove by induction that the rips complex $\mathcal{R}_D \triangleq \Omega^D \cong S^D$.

Basis For $D = 0$, $\mathcal{R}_D$ consists of two disconnected vertices, which clearly is homeomorphic to S^0.

Hypothesis Now suppose we know that $\mathcal{R}_{D-1} \cong S^{D-1}$.

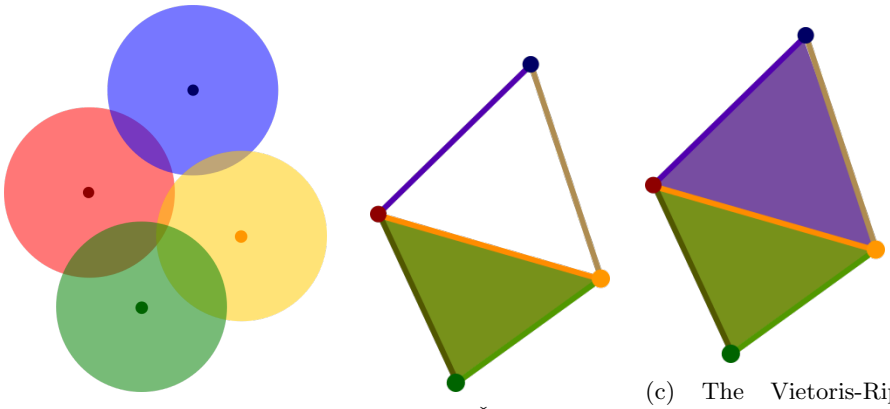
Step In the $(2D+2)$-gon, any two non-opposite vertices will be connected by an edge as the distance between two non-opposite points is less than the distance between opposite points. This defines the complex $\mathbb{R}_D$ for all D. Now note: if we take P_D and remove two opposite vertices, the complex of this resulting set of points (with radius r_D) will be homeomorphic to $\mathcal{R}_{D-1} \cong S^{D-1}$. It is now trivial to see that adding the two missing points will produce the suspension. Hence $\mathcal{R}_D \cong \Sigma\mathcal{R}_{D-1} \cong \Sigma S^{D-1} \cong S^D$.

The cases $d > 2$ can now be constructed in the same way by viewing $\mathbb{R}^2$ as an affine subspace of $\mathbb{R}^d$: all cross polytopes are feasible in any dimension $d \geq 2$.

In particular, the proof of the proposition above tells us that all cross-polytopes are feasible in $\mathbb{R}^d$ for $d \geq 2$ as Vietoris-Rips complexes.

6.1.3 Comparison

We first compare the homology in different dimensions for the geometric complex constructions in the following remark, then we will focus on some properties relating the two kinds of complexes.



(a) The cover consisiting of balls. (b) The Čech complex (c) The Vietoris-Rips complex

Figure 6.1: The Čech complex $\check{\mathcal{C}}(X, r)$ and the Rips complex $\mathcal{R}(X, r)$ for one set X and radius r. Note that one triangle in the Čech complex is not filled, because there is no triple intersection for the balls around the three vertices. Note also that the Čech complex is contained in the Rips complex.

Remark 1. *Comparing propositions 6.1.2 and 6.1.4 we see that the dimensionality of the Rips and the Čech complexes have fundamentally different dimensionality properties. Whereas the Čech complex can have non-trivial homology in all dimensions up to the dimension of the space from which the points are sampled, the Vietoris-Rips complex can have non-trivial homology in all dimensions.*

Even though the dimensions of the Čech and the Vietoris-Rips complexes can be wildly different, there is a simple containment relation between the two. It is quite trivial that $\check{\mathcal{C}}(X, r) \subset \mathcal{R}(X, r)$. This relation is obvious when we realise that the underlying graphs of the complexes are the same. For a good example of a set of points where the Čech complex and the Vietoris-Rips complex do not coincide, see Figure 6.1. Note that in fact the containment we just mentioned holds.

It also quite easy to see that there is another containment relation $\mathcal{R}(X, r) \subset \check{\mathcal{C}}(X, 2r)$. Because for each simplex in the Vietoris-Rips complex, all vertices are in each of the balls $B(v, 2r)$, we can see that certainly the barycenter of the simplex is in each of these balls. Hence the simplex is in the Čech complex with twice the radius.

6.2 Random geometric complexes

Now finally we introduce the structures we will be studying in this thesis, the random geometric complexes. The definition now is quite short as we have thoroughly introduced all the ingredients. The following sentence should be

enough to write down a quite rigorous definition yourself:"A random geometric complex is the Čech or Vietoris-Rips complex on a point process." In short, we will be studying the random geometric complexes $\mathcal{R}(\Phi_n, r_n)$ and $\check{\mathcal{C}}(\Phi_n, r_n)$.

6.2.1 Regimes

Like for graphs, we study limits of properties of these random complexes for an increasing number of vertices. This implies we also distinguish several regimes for random geometric complexes. The regimes we use here are largely the same as for random geometric graphs (Section 5.3.1). However, there seem to be several slight differences in the use of the terms.

The definitions of the sparse, the thermodynamic, and the dense regime correspond precisely to the case of the random geometric graphs. However, the words regarding the critical value sometimes take a different meaning w.r.t. random complexes. There are instances where the subcritical regime is equated with the dense regime, the critical with the thermodynamic and the supercritical with the dense regime (e.g. Kahles paper [10]).

This might be understandable, as the percolation critical value used for random geometric graphs gives a less complete picture when we look at random geometric complexes: connectivity in higher dimensions is also very relevant. Additionally, there seem to be multiple critical values in the thermodynamic regime: we have not only connectivity in the sense of graphs, but also higher dimensional connectivity in the sense of loosing holes in a certain dimension.

In this thesis we choose to take terminology corresponding as much to random geometric graphs as possible. This means we use the terms sparse, thermodynamic and dense regimes.

6.3 Subcomplex count

In this section we look at a theorem suggested by Kahle about the number of occurrences of some complex within the random geometric complex, as promised in Section 5.4. While the theorems were suggested, they did not write them down in this generality in his papers, and they also did not provide poofs. The proofs were said to be analogous to the proofs by Penrose for the case of random geometric graphs. Hence, we follow the proofs by Penrose for graphs and adjust them to the case of random geometric complexes.

6.3.1 Notation

Just like for random graphs, where $G_n(H)$ indicated the number of unlabelled occurrences of H as an induced subgraph of the random graph, we need similar notation for random complexes. Let us start by defining a few indicators which we use for counting subcomplexes, as we did for random graphs.

Definition 6.3.1. *Let $\mathcal{K}$ be a simplicial complex, define the indicator function $\hbar_{\mathcal{K},r} : \mathcal{P}(\mathbb{R}^d) \to \{0,1\}$ by*

$$\hbar_{\mathcal{K},r}(y_1,\dots,y_k) = \begin{cases} 1 & \text{if } \mathcal{K} \triangleq \check{\mathcal{C}}(\{y_1,\dots,y_k\},r) \\ 0 & \text{otherwise} \end{cases}$$

If the radius is not specified, it is understood to be equal to 1; in other words $\hbar_{\mathcal{K}} := \hbar_{\mathcal{K},1}$.

If we write a tilde ($\tilde{\cdot}$) above the indicator, it is one if the condition above is fulfilled and additionally the induced complex is a component.

These indicator functions let us define the analogues of $G_n(H)$ and $J_n(H)$ that we have seen for graphs (Definition 5.4.1).

Definition 6.3.2. *For a complex $\mathcal{K}$, point process Φ_n, and indicator function $\hbar$ as above, we define the following random variables:*

$$G_n(\mathcal{K}) := \sum_{Y \subset \Phi_n} \hbar_{\mathcal{K},r}(Y)$$

$$J_n(\mathcal{K}) := \sum_{Y \subset \Phi_n} \tilde{\hbar}_{\mathcal{K},r}(Y).$$

Note that $G_n(\mathcal{K})$ and $J_n(\mathcal{K})$ indeed count the number of unlabelled induced subcomplexes and unlabelled components isomorphic to $\mathcal{K}$ in a random complex on the points Φ_n. This means that if we sum over ordered subsets $(X_1,\dots,X_k) \subset \Phi_n$, we count each subcomplex $k!$ times.

It is also important to note that the assumed complex construction (Čech) is not reflected in the notation. This choice influences the indicator function significantly, so it should strictly be incorporated in the notation. We, however, choose to suppress this from the notation because it makes it cumbersome, and it is always clear from context whether we use a Čech or a Vietoris-Rips complex.

Now we define a constant relating to the indicator variables, which will be very useful in next section.

Definition 6.3.3. *Let I be an indicator function. We define μ_I to be a constant only dependent on I with the value*

$$\mu_I = \int_{\mathbb{R}^d} f(x)^k \,\mathrm{d}x \int_{(\mathbb{R}^d)^{k-1}} I(0,x_2,\dots,x_k)\,\mathrm{d}x_2 \cdots \mathrm{d}x_k.$$

If $I = \hbar_{\mathcal{K}}$ we also write $\mu_{\hbar_{\mathcal{K}}} =: \mu_{\mathcal{K}}$.

The constant depending on I, d and f that we just defined plays a very special role in the count of subcomplexes. We will see that it defines a kind of limit. This might intuitively already be clear: the integration over the indicator is an approximation of the probability of having a subcomplex of the form

indicated by I, when there is one fixed point; the outer integral over the space then computes the expectation that there is such a complex anywhere.

Note that it is not guaranteed that this value exists for all indicators and all f. In the rest of this thesis, we will always assume that we have such indicator and f. The restrictions are rather mild: it is enough that f be Lebesgue measurable and bounded; and for the indicator functions it can easily be proven that these are integrable, in particular because fixing one of the points, the others cannot be far away for the function to be non-zero.

6.3.2 Induced subomplexes

The following theorem is the equivalent for complexes of theorem of Penrose about the number of induced subgraphs in a random complex. Note that we are talking about induced subcomplexes here. The following theorem holds for complexes on Binomial point processes in the sparse regime.

Theorem 6.3.4 (Penrose Proposition 3.1 and 3.2 [14]; Kahle Theorem 3.6 [10])**.** *Suppose that $\mathcal{K}$ is a feasible connected complex with $k := |\mathcal{K}| \geq 2$, and that $r_n \xrightarrow{n\to\infty} 0$. Suppose also that f is bounded, i.e. $\sup_{x\in\mathbb{R}^d} f(x) =:\| f \|_\infty < \infty$, and the point process is binomial, then*

$$\lim_{n\to\infty} r^{-d(k-1)} n^{-k} \mathbb{E}(G_n(\mathcal{K})) = \lim_{n\to\infty} r^{-d(k-1)} n^{-k} \mathbb{E}(J_n(\mathcal{K})) = \frac{\mu_{\hbar_\mathcal{K}}}{k!}.$$

Note again that we did not specify whether we use the indicator for the Rips or for the Čech construction. This, it turns out, does not matter. In fact, the proof for the Rips complex is exactly the same, as the proof works for any indicator function that is bounded (in the sense that all points should lie at most kr_n away from each other), is translation invariant, and works well with scaling.

The proof uses he fact that the indicator function is only nonzero when all the points are close to each other. If the points are close to each other, the density functions at such points are also close to each other (for a Lebesgue measurable density function). Hence we show that taking $f(x)^k$ instead of $f(x_1)\cdots f(x_k)$ as joint distribution will not change the value. The former is of course easier to work with, and the theorem follows quickly. Note that this corresponds with the intuition given for the constant $\mu_{\hbar_\mathcal{K}}$ in last section.

Before we start the proof, we state a small lemma which we use for the theorem. The proof can be found in Penrose's book and goes by induction on k. The intuition for this statement is that the ball over which we integrate becomes small very quickly compared to the terms within the integral.

Lemma 6.3.5. *[Penrose p. 49 [14]] Suppose $\lim_{n\to\infty} r_n = 0$, $k \geq 2$ and f a Lebesque integrable function on $\mathbb{R}^d$, then the integral*

$$w_n(x) = \int_{B(x,kr_n)^{k-1}} r_n^{d(k-1)} \left| f(x_2)\cdots f(x_k) - f(x)^k \right| \mathrm{d}x_2 \cdots \mathrm{d}x_k$$

goes to 0 if $n \to \infty$.

Using this lemma, we can prove results about the number of occurrences of a certain subcomplex in a random geometric complex. Most importantly, we prove the theorem above, but a few intermediate lemmas are also used later, hence we split up the proof slightly more than Penrose does.

Lemma 6.3.6. *Suppose that $\mathcal{K}$ is a feasible connected complex with $k := |\mathcal{K}| \geq 2$, then*

$$\lim_{n\to\infty} r^{-d(k-1)}\mathbb{E}(\hbar_{\mathcal{K},r}(X_1,\ldots,X_k)) = \mu_{\hbar_{\mathcal{K}}}.$$

This holds for bot Poisson and binomial point processes.

Proof. It is not difficult to write out the definition of this expectation in terms of integrals, and then split it into two parts:

$$\begin{aligned}\mathbb{E}(\hbar_{\mathcal{K},r}(X_1,\ldots,X_k)) &= \int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} \hbar_{\mathcal{K},r_n}(\{x_1,\ldots,x_k\})f(x_1)\cdots f(x_k)\,\mathrm{d}x_1\cdots\mathrm{d}x_k \\ &= \int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} \hbar_{\mathcal{K},r_n}(\{x_1,\ldots,x_k\})f(x_1)^k\,\mathrm{d}x_1\cdots\mathrm{d}x_k \\ &\quad + \int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} \hbar_{\mathcal{K},r_n}(\{x_1,\ldots,x_k\})\left(f(x_1)\cdots f(x_k) - f(x_1)^k\right)\mathrm{d}x_1\cdots\mathrm{d}x_k.\end{aligned}$$

The split is such that we have one term that looks like $\mu_{\hbar_{\mathcal{K}}}$; and a term that will tend to zero, because the indicator is only non-zero when all points lie close to each other, i.e. when the densities at all points lie close to each other. First we look at the first term, to see that a rescaling will give us the limit $\mu_{\hbar_{\mathcal{K}}}$.

Use the change of variables $x_1 = x$ and $x_i = x_1 + ry_i$ (with $\frac{\mathrm{d}y_i}{\mathrm{d}x_i} = r_n^d$ for $2 \leq i \leq k$ and $\frac{\mathrm{d}y_1}{\mathrm{d}x_1} = 1$) on this first term to get

$$r_n^{d(k-1)}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} \hbar_{\mathcal{K},r_n}(\{x, x+r_ny_2,\ldots,x+r_ny_k\})\,\mathrm{d}y_k\cdots\mathrm{d}y_2 f(x)^k\,\mathrm{d}x.$$

Note that $\hbar_{\mathcal{K}}$ is translation invariant and that $\hbar_{\mathcal{K},r}(rX) = \hbar_{\mathcal{K}}(X)$ for any set X where rX denotes scaling of all elements of X by r. Using this we see that this term can be written as

$$r_n^{d(k-1)}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} \hbar_{\mathcal{K}}(\{0, y_2,\ldots,y_k\})\,\mathrm{d}y_k\cdots\mathrm{d}y_2 f(x)^k\,\mathrm{d}x.$$

Now the inner $k-1$ integrals are independent of x, and we can write the first term as

$$r_n^{d(k-1)}\int_{\mathbb{R}^d} f(x)^k\,\mathrm{d}x \cdot \int_{\mathbb{R}^{d(k-1)}} \hbar_{\mathcal{K}}(\{0, y_2,\ldots,y_k\})\,\mathrm{d}y_k\cdots\mathrm{d}y_2.$$

Now notice that this means the first term is equal to

$$\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} \hbar_{\mathcal{K},kr_n}(\{x_1,\ldots,x_k\})f(x_1)^k\,\mathrm{d}x_1\cdots\mathrm{d}x_k = r^{d(k-1)}\mu_{\mathcal{K}}.$$

Now we have to prove that the second term of the expression for $\mathbb{E}(\hbar_{\mathcal{K},r}(X_1,\ldots,X_k))$ multiplied by $r_n^{d(k-1)}$ goes to zero. This 'new' factor is introduced because it is the limit we ultimately want to know for the theorem. We first rewrite the second term, by taking a factor $f(x_1)$ from the brackets, to get for the second term

$$\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}\hbar_{\mathcal{K},r_n}(\{x_1,\ldots,x_k\})\cdot f(x_1)\left(f(x_2)\cdots f(x_k)-f(x_1)^{k-1}\right)\mathrm{d}x_1\cdots\mathrm{d}x_k.$$

Now multiplying by the factor $r_n^{d(k-1)}$, and seeing that the indicator function is only non-zero if all points are within distance kr_n from each other, we find an upper bound

$$\int_{\mathbb{R}^d} f(x_1)\int_{B(x_1,r_n)}\cdots\int_{B(x_1,r_n)} r_n^{d(k-1)}\left|f(x_2)\cdots f(x_k)-f(x_1)^{k-1}\right|\mathrm{d}x_k\cdots\mathrm{d}x_1. \tag{6.1}$$

By Lemma 6.3.5 the inner $k-1$ integrals go to 0 for all Lebesgue points x_1 of f. Because almost all points are Lebesgue points (f is a Lebesgue function), and by the dominated convergence theorem, this proves that the whole second term multiplied by $r_n^{d(k-1)}n^-k$ (6.1) goes to zero.

With the analyses of the two terms combined, we conclude that

$$\lim_{n\to\infty} r^{-d(k-1)}\mathbb{E}(\hbar_{\mathcal{K},r}(X_1,\ldots,X_k)) = \mu_{\hbar_{\mathcal{K}}}.$$

This lemma makes it quite easy to prove the theorem from Penrose's book, as the number of occurrences of a subcomplex is just the sum of the indicator values for different subsets.

Theorem 6.3.4. It is clear that $\mathbb{E}(G_n(\mathcal{K})) = \binom{n}{k}\mathbb{E}(\hbar_{\mathcal{K},r_n}(X_k))$ because: G_n is the sum of all $\hbar_{\mathcal{K},r_n}(Y)$ where $Y\subset X_n$; the indicator $h_{\mathcal{K},r_n}(Y)$ is only non-zero if $|Y|=|\mathcal{K}|=k$; there are $\binom{n}{k}$ such subsets, which are all distributed like X_k; and expectation is linear.

Hence we must compute

$$\begin{aligned}\lim_{n\to\infty}\frac{\mathbb{E}(G_n(\mathcal{K}))}{n^k r_n^{d(k-1)}} &= \lim_{n\to\infty}\frac{\binom{n}{k}}{n^k}\frac{\mathbb{E}(\hbar_{\mathcal{K},r}(X_1,\ldots,X_k))}{r_n^{d(k-1)}}\\ &= \lim_{n\to\infty}\frac{\binom{n}{k}}{n^k}\mu_{\mathcal{K}}.\end{aligned}$$

If we let n go to infinity $\binom{n}{k}$ goes to $n^k/k!$, and the result follows easily:

$$\lim_{n\to\infty} r_n^{d(k-1)}n^{-k}\mathbb{E}(G_n(\mathcal{K})) = \mu_{\hbar_{\mathcal{K}}}$$

What rests is to prove the same thing for $J_n(\mathcal{K})$. We do this by using the previous, and seeing that the probability that any occurrence is also a component is big when $nr_n^d \xrightarrow{n\to\infty} 0$.

Suppose k points induce a subcomplex isomorphic to $\mathcal{K}$, then the (conditional) probability that this complex is a component is greater than the conditional probability that all other points lie outside a ball of radius kr from one of the other points. Of course this probability is then bounded below by

$$(1 - \mathrm{Leb}(B(0, kr_n)) \max_{x\in\mathbb{R}^d} f(x))^{n-k} = (1 - \parallel f \parallel_\infty \mathrm{Leb}(B(0,1))kr_n^d)^{n-k}.$$

By the exponential inequality $(1-x)^y > e^{-y\frac{x}{1-x}}$ and the fact that $r_n^d n \xrightarrow{n\to\infty} 0$, this conditional probability goes to 1. Hence we conclude that

$$\lim_{n\to\infty} r_n^{d(k-1)} n^{-k} \mathbb{E}(J_n(\mathcal{K})) = \frac{\mu_{h_\mathcal{K}}}{k!},$$

which finishes the proof of the theorem. ◆

Now the same results hold for random geometric complexes with Poisson point processes. There is a minor change that has to be made in the proof. For Binomial point processes, we can estimate $\mathbb{E}(G_n(\mathcal{K})) = \binom{n}{k}\mathbb{E}(h_{\mathcal{K},r_n}(X_k))$, because for any instance of X_n, there are n points of which we choose k. For a Poisson point process, we do not know the number of points, we only know the expected number. The following lemma (Penrose Theorem 1.6 [14]; Palm theory for Poisson point processes) tells us that for suitable indicator functions, something akin to the binomial case holds; for example $\mathbb{E}(G_n(\mathcal{K})) = \frac{n^k}{k!}\mathbb{E}(h_{\mathcal{K},r_n}(X_k))$. It will be clear that all our indicator functions are nice enough, and that therefore the results above also hold for Poisson point processes.

Lemma 6.3.7 (Penrose Theorem 1.6 [14])**.** *Let $0 < n \in \mathbb{R}$, $k \in \mathbb{N}$ and suppose $I(Y, X)$ is a bounded measurable function defined on all pairs of the form (Y, X) with X a finite subset of R^d and Y a subset of X, satisfying $I(Y, X) = 0$ if $|Y| \neq k$. Then*

$$\mathbb{E}\left(\sum_{Y\subset\mathcal{P}_n} I(Y, X)\right) = \frac{n^k}{k!}\mathbb{E}\Big(I(\mathcal{X}_k, \mathcal{X}_k \cup P_n)\Big),$$

where the sum on the left-hand side is over all subsets Y of the random point set $\mathcal{P}_n$, and on the right-hand side the set $\mathcal{X}_k$ is a copy of binomial point set with k points independent of $\mathcal{P}_n$.

The proof of this lemma is a straightforward summation over all possible values of $|\mathcal{P}_n|$, where we use that $(\mathcal{P}_n|\#\mathcal{P}_n = N)$ and $\mathcal{X}_N$ are identically distributed.

6.3.3 Non-induced subcomplexes

Now we take a slightly different approach to the counting: we do not want to count the occurrences of some *induced* subcomplex, but the number of occurrences of a certain complex as (not necessarily induced) subcomplex. The result will be quite similar, but the notation is hard to keep consistent, so we introduce a new kind of indicator function and counting variables.

Definition 6.3.8. *Let $\mathcal{K}$ be a labeled simplicial complex, i.e. $V(\mathcal{K}) = \{x_1, \dots, x_k\}$. Then $\mathbb{k}_{\mathcal{K},r}$ is the indicator function*

$$\mathbb{k}_{\mathcal{K},r}(y_1, \dots, y_k) = \begin{cases} 1 & \text{if } f(\mathcal{K}) \subset \check{\mathcal{C}}(\{y_1, \dots, y_k\}, r) \\ 0 & \text{otherwise} \end{cases},$$

where f is the map induced by the vertex mapping $x_i \mapsto y_i$. If the radius is not specified, it is understood to be equal to 1; in other words $\mathbb{k}_{\mathcal{K}} := \mathbb{k}_{\mathcal{K},1}$.

This indicator is useful if we want to count the number of (not necessarily induced) subcomplexes of some kind. The labelling is needed here because some subcomplexes will not be counted enough if we only look at a complex and decide whether it contains the subcomplex in any way. For example, let our subcomplex be the k-simplex with one edge attached (i.e. a complex on $k+2$ vertices of which $k+1$ form a simplex, and the $k+2^{\text{th}}$ one is attached by an edge to one of the other vertices); then the number of occurrences of this complex within the full $(k+1)$-simplex should be $(k+1)\binom{k+2}{k+1}$. Without the labelling, only one occurrence can be found, as there is only one subset of the vertices of cardinality $k+2$. Another example can be seen in Figure 6.2, where we look at the labelled and unlabelled occurrences of P_3 in K_3.

The need for a different indicator function also becomes clear if we look at the indicator $\hbar$ in these circumstances. With the indicator function $\hbar$ as before, no occurrence will be found because there is no induced subcomplex of the desired form.

As we now want to count subcomplexes which might not be induced, we also need a new counting variable.

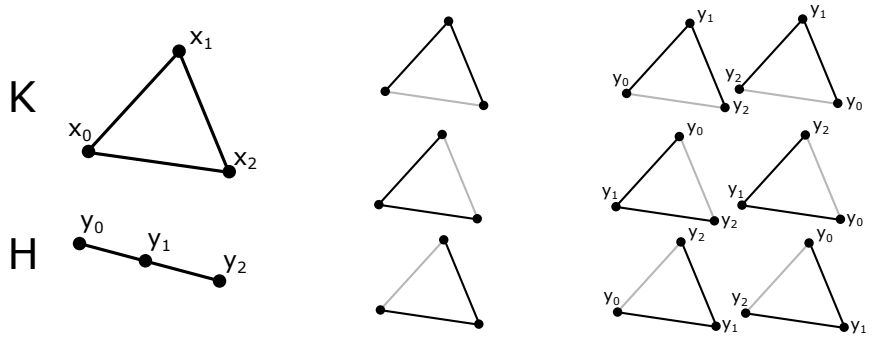
Definition 6.3.9. *Let $\mathcal{K}$ be a simplicial complex, then define $K_n(\mathcal{K})$ to be the stochastic variable*

$$K_n(\mathcal{K}) = \#\{\mathcal{K}' \subset \check{\mathcal{C}}(X_n, r_n) | \mathcal{K}' \triangleq \mathcal{K}\}.$$

With this new notation, we can state the version of previous theorem, where we now count subcomplexes, instead of induced subcomplexes. The proof is mostly the same as before, but we start slightly different, which also explains the factor $1/|\operatorname{Aut}(\mathcal{K})|$.

Theorem 6.3.10. *Suppose that $\mathcal{K}$ is a connected complex with $k := |\mathcal{K}| \geq 2$, and that $r_n \xrightarrow{n\to\infty} 0$. Then*

$$\lim_{n\to\infty} r^{-d(k-1)} n^{-k} \mathbb{E}(K_n(\mathcal{K})) = \frac{\mu_{\mathbb{k}_{\mathcal{K}}}}{|\operatorname{Aut}(\mathcal{K})|}.$$



(a) The complex $\mathcal{K}$ and the counted subcomplex H.

(b) The unlabelled occurrences.

(c) The labelled occurrences.

Figure 6.2: At the top we see $\mathcal{K} = K_3$ the complete graph on three nodes, and $H = P_3$ the path-graph on three nodes. Beneath we see all labelled and unlabelled occurrences of P_3 as subcomplex of K_3. Note that there are two automorphisms of P_3, the identity, and the map which sends 0 to 2, vice versa, and 1 to 1; there are also two times as many labelled as unlabelled occurrences.

Proof. Because any occurrence of $\mathcal{K}$ will already be caught if we look at labeled subcomplexes on k vertices, we can again write the expectation in terms of our indicator function. This time however, the indicator function works with labeled complexes, so instead of $\binom{n}{k}$, there are $\binom{n}{k}k!$ of those.

This seems to imply that as before we get $\mathbb{E}(K_n(\mathcal{K})) = \binom{n}{k}k!\mathbb{E}(K_k(\mathcal{K}))$. However, this is wrong because some relabellings do give the same subcomplex. Therefore we have to correct for this number by a factor that takes the number of such relabellings in account. This of course is exactly $|\operatorname{Aut}\mathcal{K}|$ the number of automorphisms of $\mathcal{K}$, i.e. the number of relabellings that take the complex to itself. Hence we get

$$\mathbb{E}(K_n(\mathcal{K})) = \frac{k!}{|\operatorname{Aut}\mathcal{K}|}\binom{n}{k}\mathbb{E}(K_k(\mathcal{K})).$$

Now the proof goes exactly as before and in the end we get the same result with additional factor $k!/|\operatorname{Aut}\mathcal{K}|$, resulting in

$$\lim_{n\to\infty} r^{-d(k-1)}n^{-k}\mathbb{E}(K_n(\mathcal{K})) = \frac{\mu_{k_\mathcal{K}}}{|\operatorname{Aut}(\mathcal{K})|}$$

as desired. 

For the same reason as before, this holds for Poisson point processes as well.

Part III

Homology of random geometric complexes

The aim of this thesis is to look at random geometric complexes and especially, their topological structure. As a proxy for this structure we in particular look at results regarding the Betti number of these complexes, i.e. the number of holes. In this part, we study a sample of papers with results about the expected Betti number or distribution of the Betti number in random complexes.

Note that we will not be looking at the 0-th homology much in the other chapters, because it is completely determined by the underlying graph. In fact the 0-th homology group is uniquely determined by the number of components of the complex. For the kind of complexes we study, the characterisation of the number of components in the underlying graph is as good as complete. For example, the most interesting case of this characterisation is treated in chapter 13.7 of Penrose's book [14]. Because we want to focus on (new) proofs specifically designed for complexes, and random geometric graphs have already been studied in a lot of detail, we will restrict ourselves to homology structure of dimension 1 or higher.

For the Betti numbers β_k with $k \geq 1$, we treat the results sorted by the relevant regimes. It will turn out that the sparse regime's Betti number β_k is mostly determined by the number of components that look like the minimal complex that can support non-trivial k-cycles. We determined which complexes these were in chapter 3. Most results here are just customised results about the distribution and expectation of the number of these components.

In the thermodynamic regime, the homology grows linearly with the number of points. This can intuitively be explained by the fact that any kind of structure is found linearly in the number of points. In particular, for the number of k-simplices such a result holds, and we use long exact sequences to piece local structures together to see that this implies that the Betti number also grows linearly with the number of points.

In the dense regime, we expect there to be many points in a limited space. It can easily be imagined that these points will form a highly connected structure. This intuition is made precise in a theorem which treats this for a particular kind of point distribution.

Chapter 7

Subcritical regime

In the subcritical regime, the geometric complexes will be highly disconnected. We can see this because random geometric graphs in this regime are highly disconnected and the random geometric complexes have this kind of graph as 1-skeleton. For the complexes, this means we will have a very large collection of small complex components. The size of these components depends on the exact dependence of r_n on n as n goes to infinity.

Obviously, if r_n stays relatively large, the components will be larger. Now we know that non trivial cycles in a certain degree need to be supported on some minimal number of vertices. If the expected number of components with at least that many vertices is 0, we also expect the homology in that degree to be trivial. This is a way to estimate homology. In the rest of this section we look at the expected Betti numbers of random complexes in the subcritical regime, and a threshold for the vanishing of homology.

7.1 Expected Betti number

7.1.1 Vietoris-Rips complex

Theorem 7.1.1 (Kahle [10] Theorem 3.1)**.** *For $d \geq 2$, $k \geq 1$, and $r_n = o(n^{-1/d})$, the expectation of the k-th Betti number $\mathbb{E}(\beta_k)$ of the random Vietoris-Rips complex $\mathcal{R}(X_n, r_n)$ satisfies*

$$\frac{\mathbb{E}(\beta_k)}{n^{2k+2} r^{d(2k+1)}} \to C_k$$

as $n \to \infty$, where C_k is a constant dependent only on k and the underlying distribution of the points.

For the proof e will use the Morse bounds on the Betti number as in Proposition 3.4.5. To use the bound, we introduce stochastic variables corresponding to the variables in the proposition.

Definition 7.1.2. *We denote the stochastic variables where the complex is chosen to be a random geometric complex, corresponding to the counting variables defined in Definition 3.3.11, with capital letters. This means we write $S_{k,n}$ for either $s_k(\check{\mathcal{C}}(X_n, r_n))$ or $s_k(\mathcal{R}(X_n, r_n))$. It will be clear from context which of the two is meant.*

It is clear that these stochastic counting variables are closely related to the counting variables we encountered in Section 6.3. For example $S_{k,n} = G_n(\delta\Delta^{k+1})$ counts the number of hollow $(k+1)$-simplices as induced subcomplex of a random geometric complex.

For $F_{k,n}^{\geq i}$, the relation is not immediately clear, because we put a restriction on the size of the component we want to find the k-simplex in. It is however not so hard to see that we can find an upper bound for these occurrences of k-simplices as subcomplexes of sufficiently large components. We can count the number of occurrences of some 'extended' k-simplices. We have to extend the simplex by adding $i-k$ edges without creating cycles, this ensures that the component has at least $k + (i-k) = i$ vertices. Of course some simplices get counted multiple times, because there are multiple ways of extending the simplex in its component.

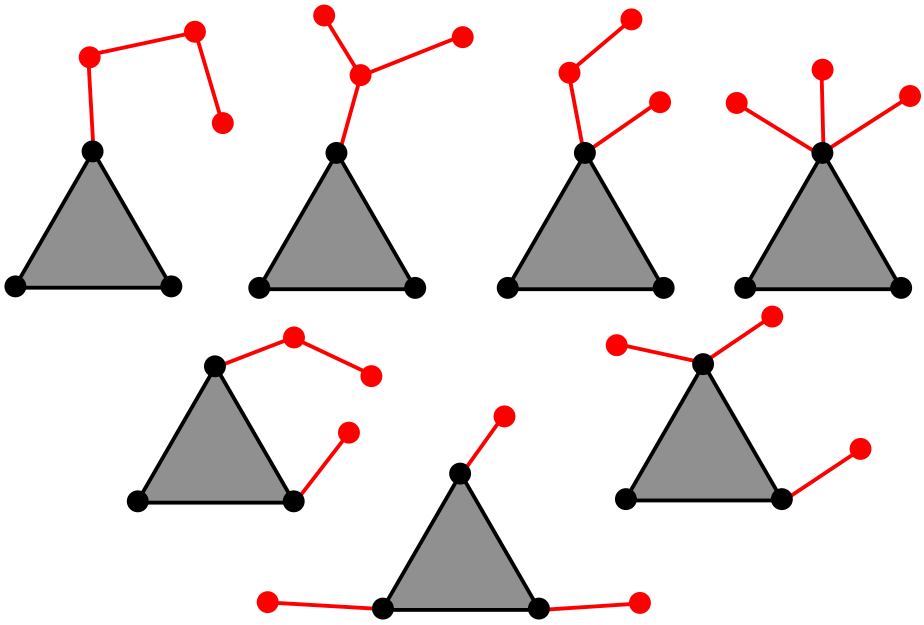


Figure 7.1: All isomorphism classes of the 2-simplex extended by three edges without creating cycles.

Lemma 7.1.3. *Let $k > 1$, and let the possible extensions of Δ^k by $i-k$ edges without creating a cycle be indexed by the set $E_{k,i}$. Let the extensions be denoted*

Δ_ϵ^k *for* $\epsilon \in E$. *then*

$$F_{k,n}^{\geq i} \leq \sum_{\epsilon \in E_{k,i}} K_n(\Delta_\epsilon^k).$$

Proof. Evident from the preceding discussion.

We will now use the Morse inequalities that sandwich the Betti number, and the theorems about the number of occurrences of some subcomplex from Section 6.3 to prove Theorem 7.1.1.

Theorem 7.1.1. We use the bounds

$$\tilde{O}_{k,n} \leq \beta_{k,n} \leq \tilde{O}_{k,n} + F_{k,n}^{2k+3}$$

which we get by applying Proposition 3.4.5 to the stochastic variables we introduced in Definition 7.1.2. Obviously the same bounds hold if we take the expectation of all these variables. It hence suffices to prove the desired limit for the expectation of the left and the right side of the inequality.

For the left side of the inequality we have to know the expected number of occurrences of the k-dimensional cross-polytope Ω^k in $\mathbb{R}^d$, which has $V(\Omega^k) = 2k+2$ vertices. By Theorem 6.3.4 we get

$$\lim_{n\to\infty} \frac{\mathbb{E}(\tilde{O}_{k,n})}{n^{2k+2} r_n^{d(2k+1)}} = \lim_{n\to\infty} \frac{\mathbb{E}(J_n(\Omega^k))}{n^{2k+2} r_n^{d(2k+1)}} = \frac{\mu_{\tilde{h}_{\Omega^k}}}{k!},$$

which means we have a lower bound for the limit:

$$\liminf_{n\to\infty} \frac{\mathbb{E}(\beta_{k,n})}{n^{2k+2} r_n^{d(2k+1)}} \geq \frac{\mu_{\tilde{h}_{\Omega^k}}}{k!}.$$

For the upper bound we have to do a bit more work as we also have the term $F_{k,n}^{\geq 2k+3}$. By Lemma 7.1.3 and linearity of expectation we can write

$$\mathbb{E}(\beta_{k,n}) \leq \mathbb{E}(\tilde{O}_{k,n}) + \mathbb{E}(F_{k,n}^{\geq 2k+3}) \leq \mathbb{E}(\tilde{O}_{k,n}) + \sum_{\epsilon \in E_{k,2k+3}} \mathbb{E}(K_n(\Delta_\epsilon^k)).$$

Using Theorem 6.3.10 we see that

$$\lim_{n\to\infty} \frac{\mathbb{E}(K_n(\Delta_\epsilon^k))}{n^{2k+3} r_n^{d(2k+2)}} = \frac{\mu_{\mathcal{K}_{\Delta_\epsilon^k}}}{|\operatorname{Aut} \Delta_\epsilon^k|}$$

for any $\epsilon \in E_{k,2k+3}$. Because $n r_n^d \to 0$, we certainly get

$$\lim_{n\to\infty} \frac{\mathbb{E}(K_n(\Delta_\epsilon^k))}{n^{2k+2} r_n^{d(2k+1)}} = 0.$$

There is only a finite number of these extensions possible, hence we conclude that

$$\limsup_{n\to\infty} \frac{\mathbb{E}(\beta_{k,n})}{n^{2k+2} r_n^{d(2k+1)}} \leq \frac{\mu_{\tilde{h}_{\Omega^k}}}{k!},$$

which finishes the proof for $k > 1$.

For $k = 1$ we have to take a slightly different approach because the extended 1-simplices used to bound $F_{n,1}^{\geq 5}$ evidently contain more than one 1-simplex, they contain four. Hence we cannot use Lemma lem:ExtendedSimplices, but we have to see that

$$F_{n,1}^{\geq 5} \leq 4 \sum_{\epsilon \in E_{1,5}} K_n(\Delta_\epsilon^1).$$

The rest of the proof is the same as before. 

7.1.2 Čech complex

We derive similar results in the case of Čech random complexes. Most of the theorems and proofs are completely analogous to previous section, among which the following main theorem of this section.

Theorem 7.1.4 (Kahle [10] theorem 3.2)**.** *For $d \geq 2$, $1 \leq k \leq d-1$, and $r_n = o(n^{-1/d})$, the expectation of the kth Betti number $\mathbb{E}(\beta_k)$ of the random Čech complex $\check{\mathcal{C}}(X_n, r_n)$ satisfies*

$$\frac{\mathbb{E}(\beta_k)}{n^{k+2} r^{d(k+1)}} \to D_k$$

as $n \to \infty$, where D_k is a constant dependent only on k and the underlying distribution of the points.

Because the proof is the same as for the case of the Vietoris-Rips complex (Theorem 7.1.1), we will not give the full proof. We will compare the two theorems and point out the small changes that have to be made in the proof.

The statements of the two theorems differ only in a few small aspects. The first thing is the fact that there are stricter bounds on k for the Čech version of the theorem. This is because there is no homology of degree higher that $d-1$ in a Čech complex arising from points in $\mathbb{R}^d$. The restriction on k had to be put in place because a hollow $d+1$ simplex is not feasible in $\mathbb{R}^d$ as a Čech complex, hence the theorems about subcomplex count cannot be used here.

The other difference are the exponents of the factors in the denominator. These arise because in the case of Čech complexes, the minimal k-cycles are supported on hollow simplices, which have $k+2$ vertices, in contrast to the k-cross-polytopes which have $2k+2$ vertices.

The proof does not change much because of these differences: we only have to count hollow simplices instead of cross-polytopes, and can use the same theorems as before to do so. Note lastly, that for the Vietoris-Rips version, we used theorems about subcomplex count, whereas we only need to know the underlying graph. Hence in that case the classical result by Penrose about subgraph counts would have sufficed. In the case of the Čech complexes, this is not the case, because we need more information than only the underlying graph, hence the subcomplex theorems were necessary. These theorems were stated by Kahle in his papers for the cases that he needed them (Only for the hollow simplices), here we used a more general statement of the subcomplex count.

7.2 Betti number distribution: Erratum

Only an expected value might be considered a rather meager result. Hence we now look a bit more closely at the distribution of the Betti numbers. We follow closely the proofs by Kahle and Meckes [11] and their quite extensive erratum [12]. In particular we will look at their theorems about the limiting distribution of the Betti number in the sub-critical regime.

As the title of this section indicates, we will start with the erratum. The choice to start with the erratum is one for clarity. As most of the arguments in the original paper and the proof are similar, it would be confusing to first give the incorrect version. It is much clearer to start with an example of proper use of the techniques as demonstrated in the erratum. An accompanying benefit is that we first see the correct version of the theorem before seeing the incorrect one which features in the original paper. We will see this incorrect version later when we look at the mistakes in the original paper. Now, we first take a look at the corrected theorem and corresponding proof provided in the erratum.

Theorem 7.2.1. *Let $\mathcal{X}_n$ be a binomial point process with n points and bounded density function f. Let $\beta_{k,n}$ denote the k-th Betti number of the random Čech complex $\check{\mathcal{C}}(\mathcal{X}_n, r_n)$. Then we have the following limits in the sparse regime. If one of the following two conditions holds:*

- $\lim_{n\to\infty} n^{k+2} r_n^{d(k+1)} = \infty$ *and* $\lim_{n\to\infty} n^{k+3} r_n^{d(k+2)} = 0$,
- $\lim_{n\to\infty} n^{k+2} r_n^{d(k+1)} = \infty$ *and* $r_n = o(n^{-1/d-\delta})$ *for some* $\delta > 0$,

then

$$\frac{\beta_{k,n} - \mathbb{E}(\beta_{k,n})}{\sqrt{n^{k+2} r_n^{d(k+1)}}} \xrightarrow{D} \mathcal{N}\Big(0, \frac{\mu_{k+2,1}}{(k+2)!}\Big)$$

for $n \to \infty$.

Note that this theorem is specifically about the Čech complex and a binomial point process. The proofs work equally well for Vietoris-Rips complexes on binomial point processes. The changes that have to be made are similar to the ones in previous section: instead of the empty simplex, we need to count cross polytopes, and hence some exponents need to be changed. Unfortunately, the proof does not simply transfer to the Poisson point process case, even though the result is derived using 'Poissonization'. We will point out the step where we run into problems in that case. For completeness we conclude this introduction with the coresponding theorem for the Vietoris-Rips random complexes.

Theorem 7.2.2. *Let $\mathcal{X}_n$ be a binomial point process with n points and bounded density function f. Let $\beta_{k,n}$ denote the k-th Betti number of the random Vietoris-Rips complex $\mathcal{R}(\mathcal{X}_n, r_n)$. Then we have the following limits in the sparse regime. If one of the following two conditions holds:*

- $\lim_{n\to\infty} n^{2k+2} r_n^{d(2k+1)} = \infty$ *and* $\lim_{n\to\infty} n^{2k+3} r_n^{d(2k+2)} = 0$,

- $\lim_{n\to\infty} n^{2k+2} r_n^{d(2k+1)} = \infty$ *and* $r_n = o(n^{-1/d-\delta})$ *for some* $\delta > 0$,

then

$$\frac{\beta_{k,n} - \mathbb{E}(\beta_{k,n})}{\sqrt{n^{2k+2} r_n^{d(2k+1)}}} \xrightarrow{D} \mathcal{N}\Big(0, \frac{\mu_{2k+2,1}}{(2k+2)!}\Big)$$

for $n \to \infty$.

A very important thing to note, is that the proofs in next sections are just extended versions of the proofs in the erratum. This means most of it is rewritten with more details added and some minor notational flaws amended. In no way do I claim to have thought of these proofs. Although this might be true for more parts of this thesis, it is most striking in this part, so I feel I have to point this out. Note however, that I have tried to copy as little as possible from the erratum itself, although this is impossible for certain equations.

The added value of my rewriting the proofs is the following. I can more easily compare the erratum with the original paper, which has some major flaws. As the proof techniques are quite similar, it is easier to point out the flaws in the paper. Secondly, the proofs contain nice techniques which should be present in a thesis about random complexes, and a detailed application of the technique is vital to understand it. There are minor (notational) mistakes in the erratum, which can most easily be found when looking at the proofs in a bit more detail than is done in the erratum itself.

7.2.1 Strategy

Variables and constants

Just like in previous section, we need some variables to count the Betti number of the random complexes. This means we introduce a few new random 'counting' variables.

Firstly, we use $\tilde{S}_{k,n}$ as before. This will be the main variable used in the approximation of the Betti number, as we are still in the sparse regime. The number of hollow simplices is not enough to accurately show that the distributions are similar. We need some new variables that approximate the Betti number 'better'. This 'better' becomes negligible in the limit $n \to \infty$.

The second kind of counting variable we use counts the Betti number of certain complexes. The random variable $X_{i,j,k,n}$ counts the number of complexes with i vertices and k-th Betti number j. It should be clear that with this notation:

$$\beta_{k,n} = \sum_{i=k+2}^{\infty} \sum_{j=0}^{\infty} j X_{i,j,k,n}.$$

This full sum is hard to work with, and because we will have relatively little

large complexes, we also consider the truncated Betti number:

$$\beta_{k,n} = \sum_{i=k+2}^{m} \sum_{j=0}^{\infty} j X_{i,j,k,n}.$$

where we only add the Betti numbers of components with m or less vertices. the choice of m is important in the proof, because if it is large, then the difference between $\beta_{k,n}$ and $\tilde{\beta}_{k,n}$ is very small. In particular, for the proofs we set $m := \lfloor 1 + \frac{1}{d\delta} \rfloor$.

Besides the constant m, we also define a range of other constants which are well-defined by Lemma 6.3.6. To define these, we first define the indicator corresponding to $X_{i,j,k,n}$: write $\tilde{h}_{i,j,k,A,n}(Y, X)$ for the indicator function which gives value one if:

- $|Y| = i$,
- the k-th Betti number of the complex spanned by Y equals j,
- the left-most point of Y lies in A, if this parameter is absent, we set $A = \mathbb{R}^d$.

And as before $\tilde{h}$ is one if additionally the set forms a component.

With this indicator the following limit becomes interesting:

$$\mu_{i,j,k,A} := \lim_{n\to\infty} r_n^{d(i-1)} \mathbb{E}(\tilde{h}_{i,j,k,A,n}(X_1, \ldots, X_i)).$$

Note that here we added a parameter $A \subset \mathbb{R}^d$ which indicates the left most point is in A. As we have not defined this yet, here follows a definition.

Definition 7.2.3. *Let $X \subset \mathbb{R}^d$ be a finite subset of $\mathbb{R}^d$ with standard basis, then the left-most point of X, denoted* $\mathrm{LMP}(X)$*, is the first point of X when ordered lexicographically.*

This parameter can be added to all counting variables as we have defined before. All previous results stay valid if we consistently add the parameter.

The reason we added this parameter is because some things are easier proved when we can first bound the searching area, and then expand to $\mathbb{R}^d$. Having a bounded region for the left-most point assures all other points are also nearby in relevant cases (where the indicator is non-zero).

Proof idea

The part of the erratum we study is about only one theorem, and the proof is mostly linear. In the erratum, this linear route is broken by first doing the last step and then filling in the details. We just follow the path till the end. That means we will here first give an overview of the proof, as otherwise it is too long to see what is happening. The proof can be subdivided in the following steps:

- Poissonize the problem, i.e. look at the same problem with a Poisson point process instead of a binomial point process. The following steps all regard the Poissonized problem unless noted differently.

- Compute mean and variance limits for useful random (counting) variables, which approximate the Betti number. In particular, we will look at the number of hollow simplices $\tilde{S}^P_{k,n}$, and the 'truncated' Betti number of the subcomplex consisting of all components of some maximum size. We say truncated because we sum over compoinents of certain sizes to get he betti number, and for this runcated one we stop at certain size.

- Prove a central limit theorem (CLT) for the number of hollow simplices, for this we need to be in the regime $n^{k+2}r_n^{d(k+1)} \to \infty$.

- Prove that this distribution is a good approximation for the distribution of the truncated Betti number. This is because in the regime $nr_n^d \to 0$ there are many more components of size exactly $k+2$ than of size bigger than $k+2$. So the hollow simplices, which have $k+2$ vertices will 'dominate' the Betti number.

- De-Poissonize again by proving a CLT for the truncated Betti number in the binomial case.

- Show that the truncated Betti number is a good approximation for the (real) Betti number in the binomial case. To do this we need to show that the expected difference $\mathbb{E}(|\beta_{k,n} - \tilde{\beta}_{k,n}|)$ is small relative to the scaling of the normalization we use in the CLT for $\beta_{k,n}$.

All these steps are treated in separate subsections.

7.2.2 Limiting behaviour of means and variances

We first look at the limits of the means, which basically follow from the result about counting subcomplexes. The important difference being that we count a lot of different kinds of subcomplexes. Of these only a few dominate the expectation. Notably, these are often the empty simplexes. Keep in mind that we are working with a Poissonized version of the problem here.

Lemma 7.2.4 (Lemma 2.6 of the erratum [12]). *Let $nr_n^d \xrightarrow{n\to\infty} 0$ and let the variables and constants be as defined in Section 7.2.1. Then we have the*

following limits for the means of the variables:

$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\tilde{S}^P_{k,A,n}\right)}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu_{k+2,1,A}}{(k+2)!}$$
$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\tilde{\beta}^P_{k,A,n}\right)}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu_{k+2,1,A}}{(k+2)!}$$
$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\sum_{\substack{k+3\le i\le m\\ j\le 0}} X^P_{i,j,n,A}\right)}{n^{k+3} r_n^{d(k+2)}} = \frac{\sum_{j=1}^{\binom{k+3}{k+1}} j\mu_{k+3,j,A}}{(k+3)!}.$$

Proof. Using the indicator $\tilde{h}_{i,j,k,A,n}$, we note that we can use the result for counting subcomplexes with this slightly different indicator function to see that

$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\sum_{j\ge 0} jX_{i,j,k,A,n}\right)}{n^i r_n^{d(i-1)}} = \frac{\sum_{j\ge 0} j\mu_{i,j,k,A}}{i!}.$$

Now we use a bound on the Betti number of finite complexes to see that the limit is in fact finite. In particular, a complex on i vertices has at most $\binom{i}{k+1}$ k-cells, hence such a complex has k-th Betti number at most $\binom{i}{k+1}$. So for fixed i, we actually have

$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\sum_{j\ge 0} jX_{i,j,k,A,n}\right)}{n^i r_n^{d(i-1)}} = \frac{\sum_{j=1}^{\binom{i}{k+1}} j\mu_{i,j,k,A}}{i!}.$$

Next, we sum over such expectations for different values of i. The equation above tells us that the term with smallest i larger than $k+1$ will dominate for large n. (Terms with $i \le k+1$ do not contribute as there are no complexes of size i with non-zero k-th Betti number.) Here, we are interested in the following examples of this:

$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\sum_{j\ge 0} jX_{k+2,j,k,A,n}\right)}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu_{k+2,1,k,A}}{(k+2)!},$$
$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\sum_{i=k+2}^{l}\sum_{j\ge 0} jX_{i,j,k,A,n}\right)}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu_{k+2,1,k,A}}{(k+2)!},$$
$$\lim_{n\to\infty} \frac{\mathbb{E}\left(\sum_{i=k+3}^{l}\sum_{j\ge 0} jX_{i,j,k,A,n}\right)}{n^{k+3} r_n^{d(k+2)}} = \frac{\sum_{j=1}^{\binom{k+3}{k+1}} j\mu_{k+3,j,k,A}}{(k+3)!},$$

for all $l \le k+2$. For $i = k+2$, we know that the only complex on i vertices with non-zero homology, is the empty simplex. This means that $\mu_{k+2,j,k,A}$ is non-zero only if $j = 1$. This has already been incorporated in the equation above. Setting $l = m$, we have the limits of the lemma.

Next we use this result about the means to derive a result about the variances. In this part we rely on the properties of Poisson point processes. The idea of the proof is to rewrite the variance into a part as large as the mean, and a rest part. This rest part is then proven to be small using the spatial independence of Poisson point processes.

Lemma 7.2.5 (Lemma 2.6 of the erratum [12])**.** *Let $nr_n^d \xrightarrow{n\to\infty} 0$ and let the variables and constants be as defined in Section 7.2.1. Then we have the following limits for the variances of the variables:*

$$\lim_{n\to\infty} \frac{\operatorname{Var}\left(\tilde{S}^P_{k,A,n}\right)}{n^{k+2}r_n^{d(k+1)}} = \frac{\mu_{k+2,1,A}}{(k+2)!}$$

$$\lim_{n\to\infty} \frac{\operatorname{Var}\left(\tilde{\beta}^P_{k,A,n}\right)}{n^{k+2}r_n^{d(k+1)}} = \frac{\mu_{k+2,1,A}}{(k+2)!}$$

$$\lim_{n\to\infty} \frac{\operatorname{Var}\left(\sum_{\substack{k+3\le i\le m\\ j\le 0}} X^P_{i,j,n,A}\right)}{n^{k+3}r_n^{d(k+2)}} = \frac{\sum_{j=1}^{\binom{k+3}{k+1}} j^2\mu_{k+3,j,A}}{(k+3)!}.$$

Proof. We analyse the variance of $\sum_{i=p}^{q}\sum_{j\ge 0} jX_{i,j,k,A,n}$, for $p \in \{k+2, k+3\}$. The proof works in all these cases, and has our interests as special cases.

We first just rewrite he second moment, which gives us the following

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{i=p}^{q}\sum_{j\ge 0} jX_{i,j,k,A,n}\right)^2\right) =& \mathbb{E}\left(\left(\sum_{i=p}^{q}\sum_{j\ge 0}\sum_{Y\subset\mathcal{P}_n} j\tilde{h}_{i,j,k,A,n}(Y,\mathcal{P}_n)\right)\right.\\ &\left.\left(\sum_{i'=p}^{q}\sum_{j'\ge 0}\sum_{Y'\subset\mathcal{P}_n} j'\tilde{h}_{i',j',k,A,n}(Y',\mathcal{P}_n)\right)\right)\\ =& \mathbb{E}\left(\sum_{i,i'=p}^{q}\sum_{j,j'\ge 0}\sum_{Y,Y'\subset\mathcal{P}_n} jj'\tilde{h}_{i,j,k,A,n}(Y,\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(Y',\mathcal{P}_n)\right).\end{aligned}$$

Now note that $\tilde{h}_{i,j,k,A,n}(Y,\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(Y',\mathcal{P}_n)$ can only be non-zero if

- either $Y = Y'$, $i = i'$ and $j = j'$,
- or $Y \cap Y' = \emptyset$,

otherwise, one of the Y does not support a component and its indicator is 0. We can hence further rewrite the second moment to

$$\begin{aligned}&\mathbb{E}\left(\sum_{i=p}^{q}\sum_{j\ge 0}\sum_{Y\subset\mathcal{P}_n} j^2\tilde{h}_{i,j,k,A,n}(Y,\mathcal{P}_n)\right)\\ &+\mathbb{E}\left(\sum_{i,i'=p}^{q}\sum_{j,j'\ge 0}\sum_{Y,Y'\subset\mathcal{P}_n} jj'\tilde{h}_{i,j,k,A,n}(Y,\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(Y',\mathcal{P}_n)\chi_{Y\cap Y'=\emptyset}\right).\end{aligned}$$

The first of these terms will dominate the variance, so note that these actually give the limits of the lemma for similar reasons to previous lemma. For $p = k+2$ the term j^2 is again not present because for a complex on $k+2$ vertices, β_k can only be one (if the complex is a hollow simplex) or zero.

The second term and the square of the mean are rewritten using Lemma 6.3.7 and further analysed together.

$$\begin{aligned}
&\mathrm{Var}(\sum_{i=p}^{q}\sum_{j\geq 0} jX_{i,j,k,A,n}) \\
&\quad =\mathbb{E}\left(\left(\sum_{i=p}^{q}\sum_{j\geq 0} jX_{i,j,k,A,n}\right)^2\right) - \mathbb{E}\left(\sum_{i=p}^{q}\sum_{j\geq 0} jX_{i,j,k,A,n}\right)^2 \\
&\quad =\mathbb{E}\left(\sum_{i=p}^{q}\sum_{j\geq 0}\sum_{Y\subset\mathcal{P}_n} j^2\tilde{h}_{i,j,k,A,n}(Y,\mathcal{P}_n)\right) \\
&\qquad + \sum_{i,i'=p}^{q}\sum_{j,j'\geq 0}\frac{n^{i+i'}jj'}{i!i'!}\Big[\mathbb{E}\left(\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n)\right) \\
&\qquad - \mathbb{E}\left(\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{P}_n)\right)\mathbb{E}\left(\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_{i'}\cup\mathcal{P}_n)\right)\Big].
\end{aligned}$$

where $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ are independent. We now look closer at the part between square brackets, starting by taking everything into one expectation. To do this we need to introduce an copy $\mathcal{P}'_n$ of a Poisson point process with density f, independent of the $\mathcal{P}_n$ we already have.

$$\begin{aligned}
&\mathbb{E}\Big(\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n)\Big) \\
&-\mathbb{E}\Big(\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{P}_n)\Big)\mathbb{E}\Big(\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_{i'}\cup\mathcal{P}_n)\Big) \\
&\quad = \mathbb{E}\Big(\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n) \\
&\quad - \tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_{i'}\cup\mathcal{P}'_n)\Big)
\end{aligned}$$

Now we can rewrite the part within this expectation and split it in three parts giving

$$\begin{aligned}
&\mathbb{E}\left[\left(\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n) - \tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{P}_n)\right)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n)\right] \\
&\quad + \mathbb{E}\left[\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{P}_n)\left(\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_i\cup\mathcal{X}_{i'}\cup\mathcal{P}_n) - \tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_{i'}\cup\mathcal{P}_n)\right)\right] \\
&\quad + \mathbb{E}\left[\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i,\mathcal{X}_i\cup\mathcal{P}_n)\left(\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_{i'}\cup\mathcal{P}_n) - \tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'},\mathcal{X}_{i'}\cup\mathcal{P}'_n)\right)\right] \\
&= E_1 + E_2 + E_3,
\end{aligned}$$

where E_1, E_2 and E_3 correspond to the first second and third term of this expression. To check that this is equal to what we had before, just work out the

brackets and combine the three expectations. We still want to show that this expression is relatively small, so we look at all these three terms separately and compute their order in n.

The first term equals zero as the random expression within the expectation is uniquely zero: If the difference is non-zero, then the addition of some point of $\mathcal{X}_{i'}$ to $\mathcal{X}_i \cup \mathcal{P}_n$, should change this component. This means $\mathcal{X}_{i'}$ is connected to $\mathcal{X}_i$, and hence $\mathcal{X}_{i'}$ does not span a component in $\mathcal{X}_i \cup \mathcal{X}_{i'} \cup \mathcal{P}_n$. This means that $\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_i \cup \mathcal{X}_{i'} \cup \mathcal{P}_n) = 0$, and we conclude there is no case in which the expression in the first expectation is non-zero, i.e. $E_1 = 0$.

The second term can actually be non-zero. To get this, we need

$$\begin{aligned}\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i, \mathcal{X}_i \cup \mathcal{P}_n) &= 1,\\ \tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_{i'} \cup \mathcal{P}_n) &= 1,\\ \tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_i \cup \mathcal{X}_{i'} \cup \mathcal{P}_n) &= 0,\end{aligned}$$

because, for the difference we cannot get value 1 as $\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_i \cup \mathcal{X}_{i'} \cup \mathcal{P}_n) = 1$ implies $\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_i \cup \mathcal{P}_n) = 1$. Now note that the probability of this event is bounded by the probability that $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ together form one connected complex. Picking one point to start with, the other $i + i' - 1$ points should then lie close enough to that point for the event to happen. Hence, the probability that the variable in the second expectation is non-zero, is bounded by

$$\left(2^d (i + i' - 1)^d r_n^d \parallel f \parallel_\infty \theta_d\right)^{i+i'-1} = c_{d,f,i,i'} r_n^{d(i+i'-1)}$$

and we conclude that $|E_2| \leq c_{d,f,i,i'} r_n^{d(i+i'-1)}$

The third term is the most interesting one, as we need a property of the Poisson point process here. If the sets $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ are sufficiently far from each other, knowledge of the interaction of $\mathcal{P}_n$ with $\mathcal{X}_i$ does not influence the interaction of $\mathcal{P}_n$ with $\mathcal{X}_{i'}$. Hence, in that case the distributions of $\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i, \mathcal{X}_i \cup \mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_{i'} \cup \mathcal{P}_n)$ and $\tilde{h}_{i,j,k,A,n}(\mathcal{X}_i, \mathcal{X}_i \cup \mathcal{P}_n)\tilde{h}_{i',j',k,A,n}(\mathcal{X}_{i'}, \mathcal{X}_{i'} \cup \mathcal{P}'_n)$ are the same and the expression in the expectation equals 0. Hence there can only be a non-zero contribution of E_3 in the case that $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ are so close together that they influence each other: i.e. if

$$\left(\bigcup_{x \in \mathcal{X}_i} B(x, 2r_n)\right) \cap \left(\bigcup_{x \in \mathcal{X}_{i'}} B(x, 2r_n)\right) \neq \emptyset$$

The probability of this event is bounded by the event that both $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ are connected, and that these connected parts are at most $4r_n$ distance away from each other. This probability then is bounded by the probability that, picking one point, all other points are at most distance $2((i + i' - 1) + 1)r_n$ away from that point. This probability is bounded by

$$\left(2^d (i + i' - 1 + 1)^d r_n^d \parallel f \parallel_\infty \theta_d\right)^{i+i'-1} = c'_{d,f,i,i'} r_n^{d(i+i'-1)},$$

where we note that the exponent is still $i + i' - 1$, because it depends on the number of points, not the distance. We conclude that $|E_3| \leq = c'_{d,f,i,i'} r_n^{d(i+i'-1)}$

We now return to the variance. Writing $\mathrm{Var}(\sum_{i=p}^{q} \sum_{j\geq 0} jX_{i,j,k,A,n}) = \mathbb{E}(\sum_{i=p}^{q} \sum_{j\geq 0} j^2 X_{i,j,k,A,n}) + E$, we have

$$|E| \leq \sum_{i,i'=p}^{q} \sum_{j,j'\leq 1} \frac{n^{i+i'} jj'}{i!i'!} (c_{d,f,i,i'} + c'_{d,f,i,i'}) r_n^{d(i+i'-1)}.$$

Because for different n we keep summing over the same ranges of i and j, and because $nr_n^d \to 0$, we can bound this by

$$C_{d,f,k,p,q}(nr_n^d)^(k+2) n^{k+2} r_n^{d(k+1)}.$$

Combining this with what we knew about the first part of the variance, we get

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathrm{Var}(\sum_{i=p}^{q} \sum_{j\geq 0} jX_{i,j,k,A,n})}{n^{k+2} r_n^{d(k+1)}} &= \frac{\sum_{j\geq 0} j^2 \mu_{p,j,k,A}}{(p!)} + \lim_{n\to\infty} C_{d,f,k,p,q}(nr_n^d)^{k+2} \\ &= \frac{\sum_{j\geq 0} j^2 \mu_{p,j,k,A}}{(p!)} + 0.\end{aligned}$$

For $p = k + 2$, we again have only one possibility for a complex with non-zero Betti number, namely the empty simplex. This gives the results we want if we make suitable choices for p and q.

In the lemmas above, we have added the notion of the left-most point. The proofs of the lemmas are not influenced by this addition, so it is not immediately clear why it is important. The reason for adding the left-most point is not clear from the erratum alone. We have to return to the original paper to understand this, as the left-most point features in the proof of the lemma in the next section.

7.2.3 CLT for $\tilde{S}_{k,n}^P$

The next part in the proof is to get to a central limit theorem for $\tilde{S}_{k,n}^P$. To get there, we first need to get a CLT for $\tilde{S}_{k,A,n}^P$, with A a bounded region. This allows us to scale up to a CLT for $\tilde{S}_{k,n}^P$ by setting $A = [-K, K]^d$ and letting $K \to \infty$. We will need the following normal approximation, which can be found in Penrose's book.

Theorem 7.2.6 (Theorem 2.4 in [14])**.** *Let $\{\xi_i\}_{i\in I}$ be a finite collection of random variables. Suppose that the dependency graph has maximum degree $D-1$, and that $\mathbb{E}(\xi_i) = 0$ for each i. Set $W = \sum_i \xi_i$, and suppose that $\mathbb{E}(W^2) = 1$, then for all $t \in R$ we have*

$$|\mathbb{P}(W \leq t) - \Phi(t)| \leq \frac{2}{\sqrt[4]{2\pi}} \sqrt{D^2 \sum_{i\in I} \mathbb{E}(|\xi_i|^3)} + 6\sqrt{D^3 \sum_{i\in I} \mathbb{E}(|\xi_i|^4)}.$$

By application of this result, we can get a CLT for $\tilde{S}^P_{k,A,n}$ for a bounded set A. We use the boundedness of A to write $\tilde{S}^P_{k,A,n}$ as a finite sum which we can analyse.

Lemma 7.2.7 (Part of Theorem 3.15 in [11])**.** *Let $A \subset \mathbb{R}^d$ be a bounded set, and for $nr_n^d \to 0$ and $n^{k+2}r_n^{d(k+1)} \to \infty$, we have the following weak limit*

$$\frac{\tilde{S}^P_{k,A,n} - \mathbb{E}(\tilde{S}^P_{k,A,n})}{\sqrt{n^{k+2}r_n^{d(k+1)}}} \xrightarrow{D} \mathcal{N}\left(0, \frac{\mu_{k+2,1,A}}{(k+2)!}\right)$$

The rate of convergence is given by

$$\left|\mathbb{P}\left(\frac{\tilde{S}^P_{k,A,n} - \mathbb{E}(\tilde{S}^P_{k,A,n})}{\sqrt{(n^{k+2}r_n^{d(k+1)})}} \le t\right) - \mathbb{P}\left(\mathcal{N}\left(0, \frac{\mu_{k+2,1,A}}{(k+2)!}\right) \le t\right)\right| \le c(n^{k+2}r_n^{d(k+2)})^{-1/4}$$

Proof. Let $\{Q_{i,n}\}_{i\in\mathbb{N}}$ together with $B := \mathbb{R} \setminus (\cup_{i\in\mathbb{N}} Q_{i,n})$ be a partition of $\mathbb{R}^d$ in open cubes with side length r_n and their collected boundaries B. Also, let $I_{A,n} \subset \mathbb{N}$ be the finite subset of indices i such that $Q_{i,n} \cap A \neq \emptyset$. To see that $I_{A,n}$ is indeed finite or each n, we just note that A is bounded. If we denote the indicator for an empty k-dimensional simplex component with LMP in A and radius r as $\tilde{h}_{k,A,n}$, then we can write

$$\tilde{S}^P_{k,A,n} = \sum_{Y\subset\mathcal{P}_n} \tilde{h}_{k,A\cap B,n}(Y,\mathcal{P}_n) + \sum_{i\in I_{A,n}} \sum_{Y\subset\mathcal{P}_n} \tilde{h}_{k,A\cap Q_{i,n},n}(Y,\mathcal{P}_n).$$

We want to use he theorem by looking at all parts of the partition separately, so we define

$$\xi_{i,n} := \sum_{Y\subset\mathcal{P}_n} \tilde{h}_{k,A\cap Q_{i,n},n}(Y,\mathcal{P}_n)$$
$$\xi_{B,n} := \sum_{Y\subset\mathcal{P}_n} \tilde{h}_{k,A\cap B}(Y,\mathcal{P}_n)$$
$$\Xi_{j,n} := \frac{\xi_j - \mathbb{E}(\xi_j)}{\sqrt{\mathrm{Var}(\tilde{S}^P_{k,A,n})}}$$

for $j \in J := I_A \cup \{B\}$. Note $\mathbb{E}(\Xi_{j,n}) = 0$ and that the denominator is just the limit for $\mathrm{Var}(\tilde{S}^P_{k,n})$. Note also that by definition

$$W_i := \sum_{j\in J} \Xi_{j,n} = \frac{\tilde{S}^P_{k,A,n} - \mathbb{E}(\tilde{S}^P_{k,A,n})}{\sqrt{\mathrm{Var}(\tilde{S}^P_{k,A,n})}}$$

which has $\mathrm{Var}(W_i) = 1$. Hence we can apply the theorem to the variables $\{\Xi_j\}_{j\in J}$. To get a nice result, we just have to bound $\mathbb{E}(|\Xi_j|^3)$, $\mathbb{E}(|\Xi_j|^4)$ and the maximum degree of the dependency graph D. By the definition of X_j, it is enough to find bounds on $\mathbb{E}(|x_j|^p)$. Let $Q^d_{i,n} \subset \mathbb{R}^d$ denote the set of points of

distance less than d from $Q_{i,n}$, and let $Z_{i,n}$ denote he Poisson variable counting the number of points in $Q_{i,n}^{2r_n}$. This variable has parameter

$$\lambda_{i,n} := n \int_{Q_{i,n}^{2r_n}} f dx \leq (\| f \|_\infty 7r_n)^d.$$

With these definitions we can clearly write $|\xi_{i,n}| \leq \binom{Z_{i,n}}{k+2}$, because for each empty simplex with LMP in A there should be $k+2$ points in this region. This means there is a constant $c_{d,f,k,p}$ such that

$$\begin{aligned}\mathbb{E}\left(|\Xi_{j,n}|^p\right) &\leq \mathbb{E}\left(\binom{Z_{i,n}}{k+2}^p\right) \\ &\leq \sum_{l=k+2}^{\infty} \binom{l}{k+2}^p \frac{e^{-\lambda_{i,n}}\lambda_{i,n}^l}{l!} \\ &\leq c_{d,f,k,p}(nr_n^d)^{k+2}.\end{aligned}$$

For $|\Xi_B|^p$ we just note that $\mathbb{P}(\Xi_B \neq 0) = 0$, which gives 'bound' $\mathbb{E}(|\Xi_B|^p) = 0$.

It is easy to see that we can bound D by $17^d + 1$: if there was to be any dependence of simplices with left most point in one of the cubes and another, then the cubes may be at at most distance $8r_n$ from each other. This means that all $\Xi_{i'}$ on which Ξ_i depends correspond to the cubes in $Q_{i,n}^{8r_n}$, of which there are at most 17^d. Not forgetting Ξ_B, we have $D \leq 17^d + 1 = D_\infty$.

Lastly, we note that because A is bounded, $|I_A| \leq ar_n^{-d}$ for some constant a. Applying the theorem for each n and letting $n \to \infty$ gives

$$\begin{aligned}\left|\mathbb{P}\left(\frac{\tilde{S}_{k,A,n}^P - \mathbb{E}(\tilde{S}_{k,A,n}^P)}{\sqrt{\mathrm{Var}(\tilde{S}_{k,A,n}^P)}} \leq t\right) - \Phi(t)\right| \leq& \frac{2}{\sqrt[4]{2\pi}}\sqrt{D_\infty^2 a r_n^{-d} c_{d,f,k,3}\frac{(nr_n^d)^{k+2}}{\mathrm{Var}(\tilde{S}_{k,A,n}^P)^3}} \\ &+ 6\sqrt{D_\infty^3 a r_n^{-d} c_{d,f,k,4}\frac{(nr_n^d)^{k+2}}{\mathrm{Var}(\tilde{S}_{k,A,n}^P)^4}} \\ \leq& C_{d,f,k}\sqrt{r_n^{-d}\frac{(nr_n^d)^{k+2}}{(n^{k+2}r_n^{d(k+1)})^3}} + C'_{d,f,k}\sqrt{r_n^{-d}\frac{(nr_n^d)^{k+2}}{(n^{k+2}r_n^{d(k+1)})^4}} \\ \leq& C_{d,f,k}\sqrt{\frac{n^{k+2}r_n^{d(k+1)}}{(n^{k+2}r_n^{d(k+1)})^3}} + C'_{d,f,k}\sqrt{\frac{n^{k+2}r_n^{d(k+1)}}{(n^{k+2}r_n^{d(k+1)})^4}} \\ \leq& C''_{d,f,k,A}\left(n^{k+2}r_n^{d(k+1)}\right)^{-1},\end{aligned}$$

and because $n^{k+2}r_n^{d(k+1)} \to \infty$ we have our result (using also the limit for the variance). ◆

Now, as said above, we move to $A = \mathbb{R}^d$. The proof relies on the fact that $\mu_{i,j,[-K,K]^d} \to \mu_{i,j,\mathbb{R}^d}$ as $K \to \infty$. We will not give the proof in detail,

because it is done quite well in the paper, it consists mainly of straightforward computations.

Lemma 7.2.8 (Part of Theorem 3.15 in [11]). *Suppose $nr_n^d \to 0$ and $n^{k+2}r_n^{d(k+1)} \to \infty$ as $n \to \infty$, then we have the following weak limit:*

$$\frac{\tilde{S}_{k,n}^P - \mathbb{E}(\tilde{S}_{k,n}^P)}{\sqrt{n^{k+2}r_n^{d(k+1)}}} \xrightarrow{D} \mathcal{N}\left(0, \frac{\mu_{k+2,1}}{(k+2)!}\right).$$

With this lemma, we have our CLT for $\tilde{S}_{k,n}^P$. In the next section we will see how this CLT implies a CLT for $\tilde{\beta}_{k,n}^P$.

7.2.4 CLT for $\tilde{\beta}_{k,n}^P$

In the end we want results about the Betti number, not about the number of hollow simplices. Hence we use the fact that $\tilde{S}_{k,n}^P$ is a good approximation or $\tilde{\beta}_{k,n}^P$ and the CLT or $\tilde{S}_{k,n}^P$ to derive a CLT for $\tilde{\beta}_{k,n}^P$.

Lemma 7.2.9. *Suppose $nr_n^d \to 0$ and $n^{k+2}r_n^{d(k+1)} \to \infty$ as $n \to \infty$, then we have the following weak limit:*

$$\frac{\tilde{\beta}_{k,n}^P - \mathbb{E}(\tilde{\beta}_{k,n}^P)}{\sqrt{n^{k+2}r_n^{d(k+1)}}} \xrightarrow{D} \mathcal{N}\left(0, \frac{\mu_{k+2,1}}{(k+2)!}\right).$$

Proof. Note again that $\tilde{\beta}_{k,n}^P = \tilde{S}_{k,n}^P + R_{k,n}^P$, where

$$R_{k,n}^P := \sum_{i=k+3}^{m} \sum_{j \geq 1} j X_{i,j,k,n}.$$

Now fix $t \in \mathbb{R}$ and $\epsilon > 0$, we will prove that

$$\left| \mathbb{P}\left(\frac{\tilde{\beta}_{k,n}^P - \mathbb{E}(\tilde{\beta}_{k,n}^P)}{n^{k+2}r_n^{d(k+1)}} \right) - \mathbb{P}\left(\sqrt{\frac{\mu_{k+2,1}}{(k+2)!}} \mathcal{N}(0,1) \leq t \right) \right| \leq \epsilon$$

using the CLT we have for $\tilde{S}_{k,n}^P$. We set the following notation for the normalised variables:

$$\mathcal{B}_{k,n} := \frac{\tilde{\beta}_{k,n}^P - \mathbb{E}(\tilde{\beta}_{k,n}^P)}{\sqrt{n^{k+2}r_n^{d(k+1)}}}$$

$$\mathcal{R}_{k,n} := \frac{R_{k,n}^P - \mathbb{E}(R_{k,n}^P)}{\sqrt{n^{k+2}r_n^{d(k+1)}}}$$

$$\mathcal{S}_{k,n} := \frac{\tilde{S}_{k,n}^P - \mathbb{E}(\tilde{S}_{k,n}^P)}{\sqrt{n^{k+2}r_n^{d(k+1)}}}.$$

Note that we have the same kind of relation between the variables as before, that is $\mathcal{B}_{k,n} = \mathcal{S}_{k,n} + \mathcal{R}_{k,n}$. We are interested in $\mathbb{P}(\mathcal{B} \leq t)$, so we write

$$\begin{aligned}\mathbb{P}(\mathcal{B}_{k,n} \leq t) &= \mathbb{P}(\mathcal{S}_{k,n} + \mathcal{R}_{k,n} \leq t) \\ &\leq \mathbb{P}(\mathcal{S}_{k,n} \leq t+\epsilon, \mathcal{R}_{n,k<\epsilon}) \quad +\mathbb{P}(\mathcal{S}_{k,n} \geq t+\epsilon, \mathcal{R}_{k,n} < -\epsilon) \\ &\leq \mathbb{P}(\mathcal{S}_{k,n} \leq t+\epsilon) \qquad\qquad +\mathbb{P}(|\mathcal{R}_{k,n}| > \epsilon).\end{aligned}$$

Now we use Lemma 7.2.5 to bound the variance of $\mathcal{R}_{k,n}$.

$$\begin{aligned}\lim_{n\to\infty} \operatorname{Var}(\mathcal{R}_{k,n}) &= \lim_{n\to\infty} \operatorname{Var}\left(\frac{R^P_{k,n} - \mathbb{E}(^P_{k,n})}{\sqrt{n^{k+2} r_n^{d(k+1)}}}\right) \\ &= \lim_{n\to\infty} \frac{\operatorname{Var}(\mathbb{R}_{k,n})}{n^{k+2} r_n^{d(k+1)}} \frac{n r_n^d}{n r_n^d} \\ &= \lim_{n\to\infty} \frac{\operatorname{Var}(\mathbb{R}_{k,n})}{n^{k+3} r_n^{d(k+2)}} n r_n^d \\ &= \frac{\sum_{j=1}^{\binom{k+3}{k+1}} j^2 \mu_{k+3,j}}{(k+3)!} \lim_{n\to\infty} n r_n^d \qquad = 0.\end{aligned}$$

From this, we conclude that for n large enough we have $\operatorname{Var}(\mathcal{R}_{k,n}) \leq \epsilon^3$. Hence we can use Chebyshev's inequality to see that for large enough n

$$\mathbb{P}(|\mathcal{R}_{k,n}| > \epsilon) \leq \frac{\operatorname{Var}(\mathcal{R}_{k,n})}{\epsilon^2} \leq \epsilon.$$

Now using the CLT we had for $\tilde{S}^P_{k,n}$ and letting $\epsilon \to 0$, we have the following limit for the upper bound

$$\limsup_{n\to\infty} \mathbb{P}(\mathcal{B}_{k,n} \leq t) \leq \mathbb{P}(\sqrt{\frac{\mu_{k+2,1}}{(k+2)!}}\mathcal{N}(0,1) \leq t).$$

For the lower bound we do the following. First we find a lower bound or $\mathbb{P}(\mathcal{B}_{k,n} \leq t)$ in terms of $\mathcal{S}_{k,n}$, $\mathcal{R}_{k,n}$ and ϵ:

$$\begin{aligned}\mathbb{P}(\mathcal{B}_{k,n} \leq t) &\geq \mathbb{P}(\mathcal{S}_{k,n} \leq t-\epsilon, \mathcal{R}_{k,n} < t) \\ &\geq \mathbb{P}(\mathcal{S}_{k,n} \leq t-\epsilon) \qquad\qquad -\mathbb{P}(\mathcal{S}_{k,n} \leq t-\epsilon, \mathcal{R}_{k,n} > \epsilon) \\ &\geq \mathbb{P}(\mathcal{S}_{k,n} \leq t-\epsilon) \qquad\qquad -\mathbb{P}(|\mathcal{R}_{k,n}| > \epsilon).\end{aligned}$$

The rest of the argument is the same and we get lower bound

$$\liminf_{n\to\infty} \mathbb{P}(\mathcal{B}_{k,n} \leq t) \geq \mathbb{P}\left(\sqrt{\frac{\mu_{k+2,1}}{(k+2)!}}\mathcal{N}(0,1) \leq t\right).$$

Hence, we conclude that

$$\frac{\tilde{\beta}^P_{k,n} - \mathbb{E}(\tilde{\beta}^P_{k,n})}{\sqrt{n^{k+2} r_n^{d(k+1)}}} \xrightarrow{D} \mathcal{N}\left(0, \frac{\mu_{k+2,1}}{(k+2)!}\right).$$

7.2.5 De-Poissonisation: a CLT for $\tilde{\beta}_{k,n}$

We now de-Poissonise last result to get to the case of a binomial point process. This is done using the following theorem from Penrose's book.

Theorem 7.2.10 (Theorem 2.12 of [14]). *Suppose that for each $n \in \mathbb{N}$, $K_n(X)$ is a real valued functional on finite sets $X \subset \mathbb{R}^d$. Suppose that*

- $\frac{1}{n}\operatorname{Var}(K_n(P_n)) \to \sigma^2$, *and*
- $\frac{1}{\sqrt{n}}(K_n(P_n) - \mathbb{E}(K_n(P_n))) \xrightarrow{D} \sigma^2\mathcal{N}(0,1)$.

Suppose that there are constants $\alpha \in \mathbb{R}$, $\beta > 0$, and $\gamma > 1/2$ such that the increments $R_{q,n} := K_n(\mathcal{X}_{q+1}) - K_n(\mathcal{X}_q)$ satisfy

1. $\lim_{n\to\infty}(\sup_{n-n^\gamma \le q \le n+n^\gamma} |\mathbb{E}(R_{q,n}) - \alpha|) = 0$;
2. $\lim_{n\to\infty}(\sup_{n-n^\gamma \le q \le q' \le ; n+n^\gamma} |\mathbb{E}(R_{q,n}R_{q',n}) - \alpha^2|) = 0$;
3. $\lim_{n\to\infty}(\frac{1}{\sqrt{n}}\sup_{n-n^\gamma \le q \le n+n^\gamma} \mathbb{E}(R_{q,n}^2)) = 0$;
4. $\mathbb{P}(|K_n(X_n)| \le \beta(n+q)^\beta) = 1$.

Then $\alpha \le \sigma$ and as $n \to \infty$, $1/n \operatorname{Var}(K_n(\mathcal{X}_n)) \to \sigma^2 - \alpha^2$ and

$$\frac{1}{\sqrt{n}}\Big(K_n(\mathcal{X}_n) - \mathbb{E}(K_n(\mathcal{X}_n))\Big) \xrightarrow{D} \sqrt{\sigma^2 - \alpha^2}\mathcal{N}(0,1).$$

Before we start with the proof of the CLT for the truncated Betti number, we first have to prove an additional small proposition. This proposition is needed to fix an oversight in the proof in the erratum.

Proposition 7.2.11. *Let $X = \{x_0, \dots, x_q, x_{q+1}\} \subset \mathbb{R}^d$ be a finite subset of d dimensional Euclidean space, and r a radius for constructing a random complex. Suppose the complex (either Čech or Vietoris-Rips) on X is connected. Then there is some constant N_d depending on the dimension d, such that the complex on $X \setminus \{x_{q+1}\}$ has at most N_d components.*

Proof. Let $\{C_i\}_{i\in I}$ be the set of components C_i in the complex on $X \setminus \{x_{q+1}\}$. Each of these components is connected to x_{q+1} in the complex on X because X is connected. Hence for each component C_i there is a vertex $y_{C_i} \in X \setminus \{x_{q+1}\}$ at distance at most $2r$ from x_{q+1}, and $d(y_{C_i}, y_{C_j}) > 2r$ for all i, j. This means the induced subgraph on $\{y_{C_i}\} \cup_{x_{q+1}}$ is isomorphic to the star graph $\star_{|I|+1}$. By Proposition 5.1.5 this star graph is only feasible if $|I| + 1$ is smaller than some constant $N_d < \infty$. Hence the maximum number of components is also finitely bounded. ◆

Now we actually prove the CLT for $\tilde{\beta}_{k,n}$. This means we have to think of a useful functional K, and we then check the conditions of the theorem.

Theorem 7.2.12. *Suppose $nr_n^d \to 0$ and $n^{k+3}r_n^{d(k+2)} \to \infty$ as $n \to \infty$, then we have the following weak limit:*

$$\frac{\tilde{\beta}_{k,n} - \mathbb{E}(\tilde{\beta}_{k,n})}{\sqrt{(n^{k+2}r_n^{d(k+1)})}} \xrightarrow{D} \mathcal{N}\left(0, \frac{\mu_{k+2,1}}{(k+2)!}\right).$$

Proof. We prove that we can apply Theorem 7.2.10 with $\alpha = 0$, $\gamma = 1$ and functional

$$K_n(X) := \frac{1}{\sqrt{n^{k+1}r_n^{d(k+1)}}} \sum_{Y \subset X} \sum_{i=k+2}^{m} \sum_{j \geq 1} j\tilde{h}_{i,j,k,n}(Y, X).$$

Note that we have the following relation between this functional and our relevant counting variables:

$$\frac{1}{\sqrt{n}}(K_n(\mathcal{P}_n) - \mathbb{E}(K_n(\mathcal{P}_n))) = \frac{\tilde{\beta}_{k,n}^P - \mathbb{E}(\tilde{\beta}_{k,n}^P)}{\sqrt{n^{k+2}r_n^{d(k+1)}}}$$
$$\frac{1}{\sqrt{n}}(K_n(\mathcal{X}_n) - \mathbb{E}(K_n(\mathcal{X}_n))) = \frac{\tilde{\beta}_{k,n} - \mathbb{E}(\tilde{\beta}_{k,n})}{\sqrt{n^{k+2}r_n^{d(k+1)}}}.$$

By the central limit theorem for $\tilde{\beta}_{k,n}^P$ (Theorem 7.2.9) and the limit for the variance of this variable we found in Lemma 7.2.5, we can already easily see that the first two bullets of the de-Poissonisation theorem are satisfied. This means we still have to prove the bullets regarding the increments. To work towards that goal, we first introduce a rescaled version of the increments:

$$\begin{aligned} D_{q,n} :&= \sqrt{n^{k+1}r_n^{d(k+1)}} R_{q,n} \\ &= \sum_{Y \subset \mathcal{X}_{q+1}} \left(\sum_{i=k+2}^{m} \sum_{j \geq 1} j\tilde{h}_{i,j,k,n}(Y, \mathcal{X}_{q+1}) \right) - \sum_{Y \subset \mathcal{X}_q} \left(\sum_{i=k+2}^{m} \sum_{j \geq 1} j\tilde{h}_{i,j,k,n}(Y, \mathcal{X}_q) \right). \end{aligned}$$

Note now, that this $D_{n,q}$ is just the truncated(!) k-th Betti number of the component of X_{q+1} minus the truncated k-th Betti number of the complex resulting by taking the component of X_{q+1} and removing all simplices containing X_{q+1} from it. Because these are truncated Betti numbers with maximum complex size m, each component can at most contribute $\binom{m}{k+1}$. Hence, for the positive part of the difference, the component of X_{q+1} contributes at most $\binom{m}{k+1}$. For the negative part we have to be careful! Whereas each component contributes at most $\binom{m}{k+1}$, the complex we get by removing X_{q+1} from its component may have multiple components. Fortunately, this number of components is bounded by a constant N_d depending only on d by Proposition 7.2.11. Hence we can see that $|D_{n,q}|$ has upper bound $N_d\binom{m}{k+1}$.

Now we will prove the limits of the increments as needed for the application of Penrose's theorem. So for now we assume $0 \leq q \leq 2n$, that is we assume $n - n^\gamma \leq q \leq n + n^\gamma$ and set $\gamma = 1$.

Note that for $D_{n,q}$ to be non-zero, we need that the component of X_{q+1} has at least $k+1$ other vertices. This probability is bounded by

$$\begin{aligned}
&\binom{q}{k+1}\mathbb{P}(k+1 \text{ within distance } 2(k+1)r_n \text{ from } X_{q+1}) \\
&\qquad \leq \binom{q}{k+1}(\parallel f \parallel_\infty \theta_d(2(k+1)r_n)^d)^{k+1} \\
&\qquad \leq \binom{2n}{k+1}(\parallel f \parallel_\infty \theta_d(2(k+1)r_n)^d)^{k+1} \\
&\qquad \leq C_{d,k,f}(nr_n^d)^{k+1},
\end{aligned}$$

where the factor $\binom{q}{k+1}$ is explained by the fact that we have a binomial point process here. Together with the bound on $|D_{q,n}|$, this gives us the bound $\mathbb{E}(|D_{q,n}|) \leq N_d\binom{m}{k+1}C_{d,k,f}(nr_n^d)^{k+1} = C'_{d,k,f,m}(nr_n^d)^{k+1}$ on the expectation. We now use this to find the first limit:

$$\begin{aligned}
\lim_{n\to\infty}\Big(\sup_{0\leq q\leq 2n}|\mathbb{E}(R_{q,n})|\Big) &= \lim_{n\to\infty}\Big(\sup_{0\leq q\leq 2n}\frac{|\mathbb{E}(D_{q,n})|}{\sqrt{n^{k+1}r_n^{d(k+1)}}}\Big) \\
&\leq \lim_{n\to\infty}\Big(\sup_{0\leq q\leq 2n}\frac{\mathbb{E}(|D_{q,n}|)}{\sqrt{n^{k+1}r_n^{d(k+1)}}}\Big) \\
&\leq \lim_{n\to\infty}\Big(\sup_{0\leq q\leq 2n}\frac{C'_{d,k,f,m}(nr_n^d)^{k+1}}{\sqrt{n^{k+1}r_n^{d(k+1)}}}\Big) \\
&= \lim_{n\to\infty}\Big(C'_{d,k,f,m}\sqrt{(nr_n^d)^{k+1}}\Big) \\
&= 0
\end{aligned}$$

Now we continue by looking at $\mathbb{E}(D_{q,n}D_{q',n})$ with $q < q'$. We proceed in the same way, by first bounding $|D_{q,n}D_{q',n}|$ and then the probability that it is non-zero. By the same reasoning as before, each of $D_{q,n}$ and $D_{q',n}$ is bounded by $N_d\binom{m}{k+1}$, hence $|D_{q,n}D_{q',n}| \leq N_d^2\binom{m}{k+1}^2$.

The probability that the product $D_{q,n}D_{q',n}$ is nonzero, is treated in a really short way in the erratum, of which I am not sure why it is correct. They need to prove that the probability of $X_{q'+1}$ being in the component of X_{q+1} is bounded above by cr_n^d for some constant c. Because the size of this component is only bounded by $q'+1 < n$, I do not see why the constant is independent of n. Hence we give intuition in a slightly different way.

The probability that the product $D_{q,n}D_{q',n}$ is nonzero, is smaller than the probability that the component of X_{q+1} in the larger complex (on $\mathfrak{X}_{q+1}$) has at least $k+3$ vertices, plus the probability that the components of $X_{q'+1}$ and of X_{q+1} are both of size at least $k+2$ given that the components are not the same. The first of these probabilities is bounded by $c'_{d,k,f}(nr_n^d)^{k+2}$, and the second by something in the order of $c''_{d,f,k}(nr_n^d)^{2k+2}$.

$$\begin{aligned}|\mathbb{E}(D_{q,n}D_{q',n})| &\leq \mathbb{E}(|D_{q,n}D_{q',n}|)\\ &\leq N_d^2\binom{m}{k+1}^2 c'_{d,k,f}(nr_n^d)^{k+2} + c''_{d,k,f}(nr_n^d)^{2k+2}\\ &\leq C_{d,f,k,m}(n^{k+2}r_n^{d(k+2)})^+(nr_n^d)^{2k+2})\end{aligned}$$

Now we can look at the corresponding limit for the differences R:

$$\begin{aligned}\lim_{n\to\infty}\Big(\sup_{0\leq q<q'\leq 2n}|\mathbb{E}(R_{q,n}R_{q',n})|\Big) &\leq \lim_{n\to\infty}\Big(\sup_{0\leq q<q'\leq 2n}\frac{|\mathbb{E}(D_{q,n}D_{q',n})|}{n^{k+1}r_n^{d(k+1)}}\Big)\\ &\leq \lim_{n\to\infty}\Big(\sup_{0\leq q<q'\leq 2n}\frac{C_{d,f,k,m}(n^{k+2}r_n^{d(k+2)})^+(nr_n^d)^{2k+2})}{n^{k+1}r_n^{d(k+1)}}\Big)\\ &= \lim_{n\to\infty}\Big(\sup_{0\leq q<q'\leq 2n}C_{d,f,k,m}(nr_n^d+(nr_n^d)^{k+1})\Big)\\ &= \lim_{n\to\infty}\Big(C'_{d,f,k,m}(nr_n^d+(nr_n^d)^{k+1})\Big)\\ &= 0.\end{aligned}$$

Thirdly, we check the corresponding limit or $q = q'$. As above, we can see that

$$\mathbb{E}(D_{q,n}^2) \leq N_d^2\binom{m}{k+1}^2 C_{d,k,f}(nr_n^d)^{k+1}$$

and hence, cancelling the scalings in $R_{q,n}$, we get

$$\frac{1}{\sqrt{n}}\mathbb{E}(R_{q,n}^2) \leq \frac{N_d^2\binom{m}{k+1}^2 C_{d,k,f}}{\sqrt{n}} \xrightarrow{n\to\infty} 0.$$

Finally, we look at the polynomial boundedness. The erratum calls this condition trivially satisfied, we will take a closer look. Note that by the upper bound $\binom{q}{k+1}$ for the Betti number β_k of any complex with q nodes, we have

$$\begin{aligned}|H_n(\mathcal{X}_q)| &\leq \frac{\binom{q}{k+1}}{n^{k+1}r_n^{d(k+1)}}\\ &= \binom{q}{k+1}\frac{\sqrt{n}}{n^{k+2}r_n^{d(k+1)}}.\end{aligned}$$

And then, using facts about the regime, we can further bound

$$\begin{aligned}|H_n(\mathcal{X}_q)| &\leq \binom{q}{k+1}\frac{n}{n^{k+2}r_n^{d(k+1)}}\\ &\leq q^{k+1}n\\ &\leq (q+n)^{k+2}\end{aligned}$$

for n large enough. Now take for the constant β, the maximum of the $k+2$ and of βs for all cases where n is smaller than large enough. For all these cases there exists such a β, hence the condition is satisfied.

By the de-Poissonisation theorem, we have the CLT for $\beta_{k,n}$ we wanted, because we have proven that the conditions hold for $\alpha = 0$.

7.2.6 From $\tilde{\beta}$ to β

Lastly, we have to use our knowledge of the distribution of our approximation variable $\tilde{\beta}_{k,n}$, to derive a CLT or $\beta_{k,n}$. Because large components are very rare, the difference $\beta_{k,n} - \tilde{\beta}_{k,n}$ is very small.

Lemma 7.2.13. *Suppose $r_n = o(n^{-1/d-\delta})$, then we have the following limit:*

$$\frac{\mathbb{E}(|\beta_{k,n} - \tilde{\beta}_{k,n}|)}{\sqrt{n^{k+2} r_n^{d(k+1)}}}.$$

Proof. Remember that $\tilde{\beta}_{k,n}$ is a truncated version of $\beta_{k,n}$, which means we have

$$\begin{aligned} |\beta_{k,n} - \tilde{\beta}_{k,n}| &= \sum_{i=m+1}^{n} \sum_{j\geq 1} j X_{i,j,k,n} \\ &= \sum_{i=m+1}^{S} \sum_{j\geq 1} j X_{i,j,k,n} + \sum_{i=S+1}^{n} \sum_{j\geq 1} j X_{i,j,k,n} \end{aligned}$$

where we split up the sum at $S := \lceil \frac{2}{d\delta} + 1 \rceil$. The next step is to bound $\mathbb{E}\left(\sum_{j\geq 0} j X_{i,j,k,n}\right)$. To do this, we note that every subset of i nodes can at most contribute $\binom{i}{k+1}$ to this sum, and only if these points are connected. For checking whether the points are connected, we can just check if they are close enough to each other and whether there is a spanning tree for these points in the complex. For a fixed set of i vertices there are i^{i-2} such spanning trees to check. This gives the following bound.

$$\mathbb{E}\left(\sum_{j\geq 0} j X_{i,j,k,n}\right) \leq \binom{n}{i} i^{i-2} (\theta_d (2(i-1) \parallel f \parallel_\infty r_n)^d)^{i-1} \binom{i}{k+1} \leq C_{d,f,k,i} n^i r_n^{d(i-1)}.$$

For the two sums in the difference $\beta - \tilde{\beta}$, this means the dominating term is the one with smallest i, just like for the limits of the means and variances of the variables in this section. In short, we have

$$\sum_{i=m+1}^{S} \mathbb{E}\left(\sum_{j\geq 0} j X_{i,j,k,n}\right) = O(n^{m+1} r_n^{dm})$$

$$\sum_{i=S+1}^{n} \mathbb{E}\left(\sum_{j\geq 0} j X_{i,j,k,n}\right) = O(n \cdot n^{S+1} r_n^{dm}).$$

Now we have to use $r_n = o(n^{1/d-\delta})$, for clarity, we rewrite this as a limit first

$$\begin{aligned} r_n n^{1/d+\delta} &\to 0 \\ r_n^d n^{1+\delta d} &\to 0 \\ r_n^{dm} n^{m+m\delta d} &\to 0 \end{aligned}$$

With this and the fact that $m\delta d < 1 + \delta d$, we can see that the first of the sums is in fact of order $O(n^{-d\delta})$. In a similar way, we can see that the second sum is o order $O(n^{2-d\delta S}) = O(n^{-\delta d})$. Because $\delta d > 0$ we can easily see that $\mathbb{E}(|\beta_{k,n} - \tilde{\beta}_{k,n}|) \to 0$ and the statement follows immediately. ◆

This small difference is used in the proof of the actual theorem, together with the CLT for $\tilde{\beta}_{k,n}$.

Theorem 7.2.14. *If $n^{k+2} r_n^{d(k+1)} \to \infty$ as $n \to \infty$ and $r_n = o(n^{1/d-\delta})$ then*

$$\frac{\beta_{k,n} - \mathbb{E}(\beta_{k,n})}{\sqrt{n^{k+2} r_n^{d(k+1)}}} \xrightarrow{D} \mathcal{N}(0, \frac{\mu_{2+k,1}}{(k+2)!}).$$

Proof. Using the lemma above, we proceed in exactly the same way as in the proof of Lemma 7.2.9. Here, instead of Chebyshev's inequality, we use a Markov inequality. ◆

One might guess that we could do something similar for he Poisson point process. This however, turns out to be more difficult, because the difference between β and its approximation is harder to estimate in that case: instead of summing till n, we sum until N with possibility $\mathbb{P}(|\mathcal{P}_n| = N)$. It is, however, probably possible to estimate this difference anyway.

7.3 Betti number distribution: Kahle & Meckes

Last section talks about the limiting distribution of the Betti number in the dense regime. This section does the same, but we treat the result from the original paper. It is very important to note that this statement has not been proven yet! The erratum exists mainly because the original proof is incorrect. In this section we will look at the differences between the original statement and the corrected statement we treated in last section. Furthermore, we will see why the proof is incorrect. First we give the statement of the original paper.

Conjecture 7.3.1 (Theorem 3.2 (iii) of [11]; Theorem 2.2 of [12])**.** *Suppose $n^{k+2} r_n^{d(k+1)} \xrightarrow{n\to\infty} \infty$ and $n r_n^d \xrightarrow{n\to\infty} 0$, then the Betti number $\beta_{k,n}$ of the random Čech complex $\check{\mathcal{C}}(\mathcal{X}_n, r_n)$ satisfies*

$$\frac{\beta_{k,n} - \mathbb{E}(\beta_{k,n})}{\mathrm{Var}(\beta_{k,n})} \xrightarrow{n\to\infty} \mathcal{N}(0,1).$$

The proof in the paper is quite long and uses a multitude of lemmas and theorems. The main strategy is quite similar to the one taken in the erratum. However, here we have not one approximation variable, but we have an upper and a lower bound on the Betti number. The paper gives central limit theorems for both bounds and tries to deduce a CLT for the Betti number. There is a small flaw in the arguments used for this step. We will see later where exactly. This also means that all results we treat are correctly proven, unless stated otherwise.

7.3.1 Comparison of the statements

Note that the theorem as stated in the erratum has one additional restriction compared to the theorem in the original paper [11], most notably the condition $r_n = o(n^{-1/d-\delta})$ has been added for part of the regime. The proof as given in the paper turns out not to work in this part of the sparse regime. In the other part of the regime this proof does work. This is the reason the additional assumption $r_n = o(n^{-1/d-\delta})$ lacks for this part of the regime.

The other difference is that the erratum has a specific factor for the normalisation, whereas the original denotes the variance of $\beta_{k,n}$ as factor of normalisation. In the paper nothing is proven about this variance, although the variances of the upper and lower bound are given. This seems to be the reason we have to have a specific factor.

7.3.2 Notation for additional variables

As before, we will use additional notation for some stochastic variables. These are mainly of the type we saw before: subcomplex counts. As before, $S_{k,n}$ is the number of empty $(k+1)$-simplices (i.e. copies of $\delta\Delta^{k+1}$) in $\check{\mathcal{C}}(X_n, r_n)$; and similarly, $\tilde{S}_{k,n}$ is the number of isolated such empty simplices.

Now we add the variables $Y_{k,n}$ and $Z_{k,n}$ which give the number of k-simplices with 2 added edges and vertices, or with one added path of length 2 respectively. These are used to estimate the number of k-simlices in components with $k+3$ or more vertices. Collectively, $S_{k,n} + Y_{k,n} + Z_{k,n}$ forms the upper bound of the Betti number the paper uses, this means we also write

$$W_{k,n} := S_{k,n} + Y_{k,n} + Z_{k,n}.$$

Also, because the proof uses a Poissonization, we need to introduce $S^P_{k,n}$, $\tilde{S}^P_{k,n}$ etc., the corresponding stochastic variables in the case where the Čech complex $\check{\mathcal{C}}(P_n, r_n)$ is produced with a Poisson point process P_n.

7.3.3 Proof strategy

The proof goes by a sort of sandwiching argument. The notation introduced in last section gives us the inequality

$$\tilde{S}_{k,n} \leq \beta_{k,n} \leq S_{k,n} + Y_{k,n} + Z_{k,n}, \tag{7.1}$$

which is justified by simple counting arguments in simplicial homology (with $\beta_{k,n}$ the random variable counting the k-th Betti number of the complex on n vertices).In the proof we use this inequality for the stochastic variables to 'sandwich' their distributions. Both the upper and the lower bound variable tend to a normal distribution if they are normalised and n goes to infinity. With a few extra assumptions on the shifts and the scalings of the normalisations, this is enough to ensure that the sandwiched normalised variable will also tend to a normal distribution.

7.3.4 Upper bound

The upper bound variable $W_{k,n}$ is proven to approximate a Poisson distribution with mean tending to infinity. The normalised version of this variable will therefore tend to a normal distribution. They prove this using the following well known approximation lemma which is derived with the Chen-Stein method in a paper by Arratia, Goldstein, and Gordon [1].

Lemma 7.3.2 ([1])**.** *Let $(\zeta_i, i \in I)$ be a finite collection of Bernoulli random variables with dependency graph $(I, \sim)$. Let the variables $p_i := \mathbb{E}(\zeta_i)$ and $p_{ij} := \mathbb{E}(\zeta_i\zeta_j)$ denote the expectations of the variables and let $X := \sum_{i\in I}\zeta_i$ and $\lambda := \sum_{i\in I} p_i$ denote the sums of all the variables and of their expected values respectively. Then we have the following bound for the distance of total variation:*

$$d_{TV}(X, \mathrm{Poi}(\lambda)) \leq \min(3, \lambda^{-1}) \left(\sum_{i\in I} \sum_{j\sim i; i\neq j} p_{ij} + \sum_{i\in I} \sum_{j\sim i} p_i p_j \right).$$

This lemma is used by taking as Bernoulli variables the indicators that certain subsets of points form certain subcomplexes. In practice, we take indicators such that

$$\begin{aligned} S_{k,n} &= \sum_{1\leq i_0<\cdots<i_{k-1}\leq n} \xi_I \\ Y_{k,n} &= \sum_{\substack{1\leq i_0<\cdots<i_{k-2}\leq n \\ i_{k-1}, i_k \notin \{i_0,\ldots,i_{k-2}\} \\ i_{k-1}\neq i_k}} \sum_{0\leq p\leq q\leq k-2} \gamma_I^{p,q} \\ Z_{k,n} &= \sum_{\substack{1\leq i_0<\cdots<i_{k-2}\leq n \\ i_{k-1}, i_k \notin \{i_0,\ldots,i_{k-2}\} \\ i_{k-1}\neq i_k}} \sum_{0\leq p\leq k-2} \eta_I^p, \end{aligned}$$

where ξ_I is the indicator that $\{X_i\}_{i\in I}$ forms an empty simplex; $\gamma_I^{p,q}$ is the indicator that $\{X_{i_0}, \ldots, X_{i_{k-2}}\}$ forms a (full) simplex, and there are edges from X_{i_p} to $X_{i_{k-1}}$ and from X_{i_q} to X_{i_k} (for $I = \{i_0, \ldots, i_k\}$); and η_I^p is the indicator that $\{X_{i_0}, \ldots, X_{i_{k-2}}\}$ forms a (full) simplex, and there are edges from X_{i_p} to $X_{i_{k-1}}$ and from $X_{i_{k-1}}$ to X_{i_k} (for $I = \{i_0, \ldots, i_k\}$).

It is clear that these kind of variables can only be dependent if their relevant point sets have points in common. Observe also that when there is a common point, all points should be close together for both indicators to be one. These observations make it easy to compute upper bounds for p_i and p_{ij}. The large sum in the upper bound of the lemma is bounded smaller than a constant (dependent only on d, k and f) times $n^{k+3}r_n^{d(k+2)}$. Then they use that $\lambda = \mathbb{E}(S_{k,n} + Y_{k,n} + Z_{k,n}) \sim n^{k+2}r_n^{d(k+1)}\frac{\mu}{k!}$ (by counting subcomplexes) to see that

$$\begin{aligned} d_{TV}(W_{k,n}, \mathrm{Poi}(\mathbb{E}(W_{k,n}))) &\leq n^{-(k+2)}r_n^{-d(k+1)}\frac{k!}{\mu}n^{k+3}r_n^{d(k+2)} \\ &= C_{d,k,f}nr_n^d, \end{aligned}$$

which goes to 0 as $n \to \infty$. To complete the argument for the central limit theorem for $W_{k,n}$, we have to note that d_{TV} is invariant under simultaneous shifts and scalings of the random variables, hence

$$d_{TV}\left(\frac{W_{k,n} - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}}, \frac{\mathrm{Poi}(\mathbb{E}(W_{k,n})) - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}}\right) \leq C_{d,k,f}nr_n^d$$

and because $\frac{\mathrm{Poi}(\mathbb{E}(W_{k,n})) - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}} \xrightarrow{D} \mathcal{N}(0,1)$, we have a central limit theorem for the upper bound:

$$\frac{W_{k,n} - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}} \xrightarrow{D} \mathcal{N}(0,1).$$

In the original paper, there was one additional step which led them to conclude that there was a slightly different central limit theorem, namely

$$\frac{W_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} \xrightarrow{D} \mathcal{N}(0,1).$$

This is correct as long as $d_{TV}(\mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n})), \mathrm{Poi}(\mathbb{E}(W_{k,n})))$ goes to zero, which it does if $|\mathbb{E}(\tilde{S}_{k,n}) - \mathbb{E}(W_{k,n})|$ goes to zero. Unfortunately, this is not necessarily the case.

Note that $W_{k,n} - \tilde{S}_{k,n} = S_{k,n} - \tilde{S}_{k,n} + Y_{k,n} + Z_{k,n}$. Note also that, when $nr_n^d \to 0$, also $S_{k,n} - \tilde{S} \to 0$, see the proof of the subcomplexes count for an explanation. This means that

$$W_{k,n} - \tilde{S}_{k,n} \sim (Y_{k,n} + Z_{k,n}) \sim cn^{k+3}r_n^{d(k+2)},$$

for some constant c. Hence, only if $n^{k+3}r_n^{d(k+2)} \to 0$, the central limit theorem for $W_{k,n}$ as given in the paper holds. We will see later that the actual central limit theorem for $W_{k,n}$ in the whole sparse regime, is not enough to prove the full theorem. Concluding, we have the following lemma:

Lemma 7.3.3. *Let $nr_n^d \to 0$, $n^{k+2}r_n^{d(k+1)} \to \infty$, and let $W_{k,n}$ and $\tilde{S}_{k,n}$ be as above, then*

$$\frac{W_{k,n} - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}} \xrightarrow{D} \mathcal{N}(0,1).$$

If additionally $n^{k+3}r_n^{d(k+2)} \to 0$, then

$$\frac{W_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} \xrightarrow{D} \mathcal{N}(0,1).$$

7.3.5 Lower bound

As noted before, for the corresponding lower bound we distinguish two regimes, in which very different ways of proving a central limit theorem are used. Remember that in both regimes, we assume $r_n = o(n^{-1/d})$. The regimes are:

- $\lim_{n\to\infty} n^{k+3}r_n^{d(k+2)} = 0$;
- $\liminf_{n\to\infty} n^{k+3}r_n^{d(k+2)} > 0$.

The first case is properly dealt with in the original paper. For the second regime, the correct proof is given in the erratum, where we slightly change the assumption for r_n as a function of n: the relation between becomes $r_n = o(n^{-1/d-\delta})$ for some $\delta > 0$. We will first follow the approach of the original paper for both regimes.

First case: $\lim_{n\to\infty} n^{k+3}r_n^{d(k+2)} = 0$

For the first regime something akin to the case of the upper bound is done: they prove that the total variational distance $d_{TV}(\tilde{S}_{k,n}, \mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n})))$ is bounded by $c_{d,k,f}n^{k+3}r_n^{d(k+2)}$. Hence in the first regime we see that

$$d_{TV}(\tilde{S}_{k,n}, \mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n}))) \xrightarrow{n\to\infty} 0.$$

Because $n^{k+2}r_n^{d(k+1)} \xrightarrow{n\to\infty} \infty$, we also know that $\mathbb{E}(\tilde{S}_{k,n}) \xrightarrow{n\to\infty} \infty$. Now we use that $(\mathrm{Poi}(\lambda) - \lambda)/\sqrt{\lambda}$ will tend to a standard normal distribution if $\lambda \to \infty$, to see that

$$\frac{\tilde{S}_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} \to \mathcal{N}(0,1)$$

as $n \to \infty$, just like for the upper bound. Because the proof in the paper leaves out some details and intuition, we give the proof of this bound on the distance of total variation.

Lemma 7.3.4 (Lemma 3.8 from [11])**.** *Let $\tilde{S}_{k,n}$ be the random counting variable as before, then for sufficiently large n we have*

$$d_{TV}(\tilde{S}_{k,n}, \mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n}))) < c_{d,k,f}n^{k+3}r_n^{d(k+2)}.$$

for some constant $c_{d,f,k}$ dependent only on d, f and k.

Proof. First note that we can prove that

$$d_{TV}(S_{k,n}, \mathrm{Poi}(\mathbb{E}(S_{k,n}))) < c'_{d,k,f} n r_n^d$$

with Lemma 7.3.2 and the same indicator variables ξ_I as in the proof for the central limit theorem of $W_{k,n}$ in Lemma 7.3.3.

Now we prove that

$$d_{TV}(S_{k,n}, \tilde{S}_{k,n}) \leq c''_{d,f,k} n^{k+3} r_n^{d(k+2)},$$

using the fact that $S_{k,n} - \tilde{S}_{k,n}$ is just the number of empty k-dimensional simplices that are not isolated in the complex. Not being isolated means there is at least one more point connected to this empty simplex. Hence $S_{k,n} - \tilde{S}_{k,n}$ is bounded above by the number of $k+3$-vertex subsets of which $k+2$ form an empty simplex. The probability that $k+3$ points form such a subcomplex is bounded by

$$\Big((2r_n)^d \theta_d \parallel f \parallel_\infty \Big)^{k+1} \Big((4r_n)^d \theta_d \parallel f \parallel_\infty \Big) \leq 4^{d(k+2)} \theta_d^{k+2} \parallel f \parallel_\infty^{k+2} r_n^{d(k+2)}$$

where $\theta_d := \mathrm{Leb}(B(0,1) \subset \mathbb{R}^d)$ and $\parallel f \parallel_\infty := \sup_{x \in \mathbb{R}^d} f(x)$. This bound holds because 'all points should lie close together': Pick one of the points of the empty simplex, then all $k+1$ other points of the empty simplex should lie at distance at most $2r_n$ away from that point, the remaining point must be connected to one of the $k+2$ points, so it must be at most $4r_n$ away from the chosen point.

The number of such subsets with the one point labelled is $\binom{n}{k+3}(k+3)$ which gives the following upper bound

$$\begin{aligned}
\mathbb{E}(S_{k,n} - \tilde{S}_{k,n}) &\leq \binom{n}{k+3}(k+3)\Big(4^{d(k+2)} \theta_d^{k+2} \parallel f \parallel_\infty^{k+2} r_n^{d(k+2)}\Big) \\
&\sim \Big(\frac{4^{d(k+2)} \theta_d^{k+2} \parallel f \parallel_\infty^{k+2}}{(k+3)!}\Big) n^{k+3} r_n^{d(k+2)} \\
&= c''_{d,f,k} n^{k+3} r_n^{d(k+2)}.
\end{aligned}$$

Now we can compute an upper bound for the total variational distance:

$$\begin{aligned}
d_{TV}(S_{k,n}, \tilde{S}_{k,n}) &= \sup_{A \text{Borel}} \Big| \mathbb{P}\Big(S_{k,n} \in A\Big) - \mathbb{P}\Big(\tilde{S}_{k,n} \in A\Big) \Big| \\
&= \sup_{A \text{Borel}} \Big| \mathbb{P}\Big(S_{k,n} \in A, S_{k,n} \neq \tilde{S}_{k,n}\Big) - \mathbb{P}\Big(\tilde{S}_{k,n} \in A, S_{k,n} \neq \tilde{S}_{k,n}\Big) \\
&\quad + \mathbb{P}\Big(S_{k,n} \in A, S_{k,n} = \tilde{S}_{k,n}\Big) - \mathbb{P}\Big(\tilde{S}_{k,n} \in A, S_{k,n} = \tilde{S}_{k,n}\Big) \Big| \\
&= \sup_{A \text{Borel}} \Big| \mathbb{P}\Big(S_{k,n} \in A, S_{k,n} \neq \tilde{S}_{k,n}\Big) - \mathbb{P}\Big(\tilde{S}_{k,n} \in A, S_{k,n} \neq \tilde{S}_{k,n}\Big) \Big| \\
&\leq \sup_{A \text{Borel}} \mathbb{P}\Big(S_{k,n} \neq \tilde{S}_{k,n}\Big) \\
&= \mathbb{P}\Big(S_{k,n} \neq \tilde{S}_{k,n}\Big) \\
&\leq \mathbb{E}\Big(S_{k,n} - \tilde{S}_{k,n}\Big)
\end{aligned}$$

which is smaller than $c''_{d,f,k} n^{k+3} r_n^{d(k+2)}$ for sufficiently big n.

Now finally we use that $d_{TV}(\mathrm{Poi}(a), \mathrm{Poi}(b)) \leq |a-b|$ and $|\mathbb{E}(S_{k,n}) - \mathbb{E}(\tilde{S}_{k,n})| \leq c''_{d,f,k} n^{k+3} r_n^{d(k+2)}$ for sufficiently large n to see that

$$d_{TV}(\mathrm{Poi}(\mathbb{E}(S_{k,n})), \mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n}))) \leq c''_{d,f,k} n^{k+3} r_n^{d(k+2)}$$

for sufficiently large n. Hence, with the triangle inequality for d_{TV}, we conclude that for sufficiently large n

$$\begin{aligned} d_{TV}(\tilde{S}_{k,n}, \mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n}))) \leq & d_{TV}(\tilde{S}_{k,n}, S_{k,n}) + d_{TV}(S_{k,n}, \mathrm{Poi}(\mathbb{E}(S_{k,n}))) \\ & + d_{TV}(\mathrm{Poi}(\mathbb{E}(S_{k,n})), \mathrm{Poi}(\mathbb{E}(\tilde{S}_{k,n}))) \\ \leq & c''_{d,f,k} n^{k+3} r_n^{d(k+2)} + c'_{d,k,f} n r_n^d + c''_{d,f,k} n^{k+3} r_n^{d(k+2)} \\ \leq & c_{d,f,k} n^{k+3} r_n^{d(k+2)}. \end{aligned}$$

Second case: $\liminf_{n\to\infty} n^{k+3} r_n^{d(k+2)} > 0$

The proof for the second regime goes by Poissonisation and de-Poissionisation, just like in the proof of the erratum. In fact, a large part is exactly the same. The first step they take for this case is determining that

$$\frac{\tilde{S}^P_{k,n} - \mathbb{E}(\tilde{S}^P_{k,n})}{\sqrt{\mathrm{Var}(\tilde{S}^P_{k,n})}} \to \mathcal{N}(0,1)$$

in this regime (Lemma 7.2.8).

Using this result, they de-Poissonise using the same method as in the erratum for $\tilde{\beta}_{k,n}$. Here again, the application of the de-Poissonisation theorem is quite straightforward, by computation of a few limits. This results in the following theorem.

Theorem 7.3.5. *Suppose $nr_n^d \to 0$ and $n^{k+2} r_n^{d(k+1)} \to \infty$, then*

$$\frac{\tilde{S}_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathrm{Var}(\tilde{S}_{k,n})}} \to \mathcal{N}(0,1)$$

7.3.6 Central limit theorem for β_k

Now we have the CLT's for the upper and lower bound on β_k, we want a CLT for β_k, too. In the first regime, this is quite easy, and we will first give this argument. After, we will treat the case of the second regime. Here, the paper goes too quickly and omits the argument. This is problematic because without additional arguments, we cannot conclude that there is a CLT or β_k. We will treat this regime more careful.

First regime

Remember that the first regime was the regime in which we had he following limits:

$$\begin{aligned} nr_n^d &\to 0 \\ n^{k+2} r_n^{d(k+1)} &\to \infty \\ n^{k+3} r_n^{d(k+2)} &\to 0, \end{aligned}$$

when $n \to \infty$. With Lemma 7.3.4 and Lemma 7.3.3 from the preceding sections, this gives us the central limit theorems for the upper and lower bound of β_k in the inequality $\tilde{S}_{k,n} \leq \beta_{k,n} \leq W_{k,n}$. We can hence 'normalize' all three variables in the inequality in the same way to get

$$\begin{array}{ccccc} \frac{\tilde{S}_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} & \leq & \frac{\beta_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} & \leq & \frac{W_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} \\ {\scriptstyle n\to\infty}\downarrow{\scriptstyle \text{Lemma 7.3.4}} & & {\scriptstyle n\to\infty}\downarrow & & {\scriptstyle n\to\infty}\downarrow{\scriptstyle \text{Lemma 7.3.3}} \\ \mathcal{N}(0,1) & & ? & & \mathcal{N}(0,1) \end{array} .$$

Even though this already suggests that the limit for $\beta_{k,n}$ should also be a standard normal distribution, we will go through the arguments. We do this because in the second regime, we encounter a similar situation, where the argument does not work.

The inequality above yields the following inequality on the level of cumulative distribution functions. We also indicate the limit for when $n \to \infty$.

$$\begin{array}{ccccc} \mathbb{P}\Big(\frac{W_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} < t\Big) & \leq & \mathbb{P}\Big(\frac{\beta_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} < t\Big) & \leq & \mathbb{P}\Big(\frac{\tilde{S}_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}} < t\Big) \\ {\scriptstyle n\to\infty}\downarrow{\scriptstyle \text{Lemma 7.3.4}} & & {\scriptstyle n\to\infty}\downarrow & & {\scriptstyle n\to\infty}\downarrow{\scriptstyle \text{Lemma 7.3.3}} \\ \Phi(t) & \leq & \Phi(t) & \leq & \Phi(t) \end{array} .$$

Because this inequality concerns actual sequences of real numbers, we conclude that the middle limit also converges to $\Phi(t)$. This proves the weak convergence of $\frac{\beta_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}}$ to a standard normal distribution. Hence in this regime we have a CLT for β_k.

This almost proves the first part of Theorem 7.2.1, save for the fact that there the normalisation is done with $\mathbb{E}(\beta_{k,n})$ for the translation and $\mathrm{Var}(\beta_{k,n})$ for the rescaling. As long as $\mathrm{Var}(\beta_{k,n}) \sim \mathbb{E}(\tilde{S}_{k,n})$, this does not pose a problem. However, whether this is true is not proven in the paper. In the erratum, they do not talk about the variance of β_k anymore.

Second regime: CLT sandwich

In the second regime, just like in the first, there are CLT's for the upper and lower bounds. In the paper they claim that this is enough to conclude that

these is a CLT for the sandwiched variable. The only clue they leave is the argument they gave for the similar case of the Betti numbers of Erdős-Rényi type clique complexes. We now look at this argument applied to the case of Čech complexes. In particular we give some conditions under which we can actually conclude that a sandwiched variable has a CLT when the upper and lower bound have a CLT.

The following is a correct and complete version of the argument given by Kahle and Meckes in their paper under Claim 2.5. Only the first three conditions were given there, we added the fourth and the fifth.

Proposition 7.3.6. *Let $X_n \leq Y_n \leq Z_n$ be a relation for 3 sequences of $\mathbb{R}$-valued random variables. If there are $\sigma_n, \tau_n \geq 0$ and $\mu_n, \nu_n \in R$ (for all $n \in \mathbb{N}$) such that*

1. $\lim_{n\to\infty} \frac{\sigma_n}{\tau_n} = 1$;
2. $\frac{X_n-\mu_n}{\sigma_n} \xrightarrow{D} N(0,1)$;
3. $\frac{Z_n-\nu_n}{\tau_n} \xrightarrow{D} N(0,1)$;
4. $|\frac{\mu_n}{\sigma_n} - \frac{\nu_n}{\tau_n}| \to 0$;
5. $\lim_{n\to\infty} \mathbb{E}(X_n)|\frac{1}{\tau_n} - \frac{1}{\sigma_n}| = 0$.

then we also have a CLT for Y_n, i.e. $\frac{Y_n-\nu_n}{\tau_n} \xrightarrow{D} \mathcal{N}(0,1)$.

Proof. Note that we can translate and rescale the variables, all in the same way, to get

$$\mathbb{P}\left(\frac{Z_n-\nu_n}{\tau_n} \leq t\right) \leq \mathbb{P}\left(\frac{Y_n-\nu_n}{\tau_n} \leq t\right) \leq \mathbb{P}\left(\frac{X_n-\nu_n}{\tau_n} \leq t\right).$$

The LHS tends to $\Phi(t)$ as $n \to \infty$ (condition 3). We want to prove the same for the RHS.

Let $\epsilon > 0$, and rename $P_n := \frac{X_n-\nu_n}{\tau_n}$ and $Q_n := \frac{X_n-\mu_n}{\sigma_n}$. So P_n is the normalised version of X_n with the same translation and scaling as for the other two variables, and Q_n is the normalised X_n for which we have a CLT. We want to see that there is only a small difference between P_n and Q_n. To that end we note that

$$\begin{aligned}\mathbb{P}(P_n \leq t) &= \mathbb{P}(P_n \leq t, Q_n \leq t+\epsilon) + \mathbb{P}(P_n \leq t, Q_n > t+\epsilon) \\ &\leq \mathbb{P}(Q_n \leq t+\epsilon) + \mathbb{P}(|P_n - Q_n| > \epsilon).\end{aligned}$$

where $\mathbb{P}(Q_n \leq t+\epsilon) \to \Phi(t+\epsilon)$ (condition 2), and the other term goes to 0 as we will now prove.

Note that, by condition 4, for large enough n

$$\left|\frac{\mu_n}{\sigma_n} - \frac{\nu_n}{\tau_n}\right| < \frac{\epsilon}{2}.$$

We first use this, and then Markov's inequality to see:

$$\begin{aligned}
\mathbb{P}(|P_n - Q_n| > \epsilon) &= \mathbb{P}\left(\left|\frac{X_n - \nu_n}{\tau_n} - \frac{X_n - \mu_n}{\sigma_n}\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(X_n\left|\frac{1}{\tau_n} - \frac{1}{\sigma_n}\right| + \left|\frac{\mu_n}{\sigma_n} - \frac{\nu_n}{\tau_n}\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(X_n\left|\frac{1}{\tau_n} - \frac{1}{\sigma_n}\right| > \epsilon/2\right) \\
&\leq \frac{2}{\epsilon}\mathbb{E}(X_n)\left|\frac{1}{\tau_n} - \frac{1}{\sigma_n}\right|
\end{aligned}$$

The term $\left(\mathbb{E}(X_n)\left|\frac{1}{\tau_n} - \frac{1}{\sigma_n}\right|\right)$ is not present in the version of the paper by Kahle and Meckes, and makes condition 5 a necessary condition here.

Now we can see that for any fixed $\epsilon > 0$, this probability $\mathbb{P}(|P_n - Q_n| > \epsilon)$ tends to 0 (conditions 1 and 5). Hence for fixed $\epsilon > 0$ we have upper bound $\Phi(t + \epsilon)$ for the probability $\mathbb{P}(P_n \leq t)$. Now let $\epsilon \to 0$ to see that $\mathbb{P}(P_n \leq t) \to \Phi(t)$, and hence

$$\frac{Y_n - \nu_n}{\tau_n} \xrightarrow{D} \mathcal{N}(0, 1).$$

Now the crucial part, the use of this proposition to get a CLT for $\beta_{k,n}$. Theorem 7.3.5 and Lemma 7.3.3 give us he following weak convergences:

$$\begin{aligned}
\frac{\tilde{S}_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathrm{Var}(\tilde{S}_{k,n})}} &\xrightarrow{D} \mathcal{N}(0, 1) \\
\frac{W_{k,n} - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}} &\xrightarrow{D} \mathcal{N}(0, 1)
\end{aligned}$$

as $n \to \infty$ in this part of the regime. The obvious use of the proposition would be to set

$$\begin{aligned}
X_n &= \tilde{S}_{k,n} & Z_n &= W_{k,n} \\
\mu_n &= \mathbb{E}(\tilde{S}_{k,n}) & \nu_n &= \mathbb{E}(W_{k,n}) \\
\sigma_n &= \sqrt{\mathrm{Var}(\tilde{S}_{k,n})} & \tau_n &= \sqrt{\mathbb{E}(W_{k,n})}
\end{aligned}$$

We then still have to check the five conditions from the proposition. Unfortunately, it turns out that the last two (new) conditions might not hold.

We do know that

$$c\mathbb{E}(W_{k,n}) \sim \mathbb{E}(\tilde{S}_{k,n}) \sim \mathrm{Var}(\tilde{S}_{k,n}),$$

for some constant c. But these are facts about quotients, not about differences! Hence, this is not enough to salvage the situation because the difference between the means (in the numerators) and the scalings of the normalisation might be too large. This is also the only, and very cryptic, reason given in the erratum for its existence. More on this in next section.

7.3.7 Need for Erratum

In this section we look at the reasons there needs to be an erratum to the paper. Most of these reasons have been touched on in previous sections, but for clarity we repeat them here as quick reminder.

Scaling for normalisation

In the third part of the theorem, we need to change the rescaling of the normalisation to $\sqrt{n^{k+2}r_n^{d(k+1)}}$. This is because we actually do not know if the variance of $\beta_{k,n}$ is as large as that of $\tilde{S}_{k,n}$.

In the paper, it seems that this is assumed because there is convergence to a normal distribution. But, because we only have weak convergence, we do not know whether higher moments (than just the expectation) also converge.

Naturally, this goes for the first part of the regime as well as for the second part. Only, in the second part of the regime, we need to look out or a few more things.

Sandwiching in the second regime

This has been treated extensively in Section 7.3.6, so we just repeat the issues because this problem actually runs through the whole argument.

The issue is two-part: firstly the limiting distribution of $W_{k,n}$ is not treated correctly in the second regime, as the normalisation cannot just be changed (Lemma 7.3.3). Secondly, for this central limit theorem for $W_{k,n}$, the sandwiching argument does not work (Section 7.3.6).

The reason that sandwiching does not work is that the difference of the means (in the right proportion) is $\mathbb{E}(\tilde{S}_{k,n}) - c\mathbb{E}(W_{k,n}) \sim Cn^{k+2}r_n^{d(k+1)}$, and the normalisation is of order $\sqrt{n^{k+2}r_n^{d(k+1)}}$. Hence the difference between

$$\frac{W_{k,n} - \mathbb{E}(W_{k,n})}{\sqrt{\mathbb{E}(W_{k,n})}} \text{ and } \frac{W_{k,n} - \mathbb{E}(\tilde{S}_{k,n})}{\sqrt{\mathbb{E}(\tilde{S}_{k,n})}}$$

is too large to conclude that the second of the two approaches a standard normal variable.

Actually, if $W_{k,n}$ had the central limit theorem as stated in the paper, the sandwiching could still work by an argument similar to that for the first regime and the fact that $\mathbb{E}(\tilde{S}_{k,n}) \sim \mathrm{Var}(\tilde{S}_{k,n})$.

7.4 Vanishing thresholds

Using last two sections, we can easily deduce some results about the vanishing of Betti numbers. Vanishing of Betti numbers might be considered interesting because it conveys that there are no 'holes' of certain dimension in the complex.

In Kahles paper [10], theorem 3.10 states the following:

Theorem 7.4.1. *For $d \geq 2$, $k \geq 1$, $r_n = o(n^{-1/d})$, and a random Čech complex $\check{\mathcal{C}}(n, r_n)$ we have he following:*

- *If $r_n = o(n^{-k+2/d(k+1)})$, then a.a.s. $H_k(\check{\mathcal{C}}(n, r_n)) = 0$;*
- *If $r_n = \omega(n^{-k+2/d(k+1)})$ and $\lim_{n\to\infty} n^{k+3} r_n^{k+2} = 0$, or if $r_n = \omega(n^{-k+2/d(k+1)})$ and $\lim_{n\to\infty} n^{k+3} r_n^{k+2} > 0$ and $r_n = o(n^{-1/d-\delta}$ for some $\delta > 0$, then a.a.s. $H_k(\check{\mathcal{C}}(n, r_n)) \neq 0$;*

The first part of the proof is quite trivial when we consider the subcomplex count theorem: the expected number of components of size $k + 2$ and larger goes to 0. Hence, because minimal nontrivial cycles need at least components of size $k + 2$ or larger, the probability that there is nonzero homology becomes 0.

The proof of the second part, uses their theorem about the limiting distribution of the Betti number we saw in the previous sections. It is easy to see by Theorem 7.1.4 that the expected Betti number tends to infinity as $n \to \infty$. Furthermore, the variance and the expectation of the limiting distribution are of the same order of magnitude. It is then easy to see that the probability of non-zero homology becomes 1.

Note that this theorem now uses the restrictions of the erratum for the second part. In their paper, this is not corrected for as far as I could see.

Chapter 8

Critical regime

The results in this section are about the critical or thermodynamic regime. In this regime percolation of the random graph occurs. It is hence unsurprising that the homology of the complexes in this regime becomes more complex (and more difficult to compute). Technically, this means we have to take in account the structure of the complex as a whole instead of as a collection of small components. In the computations, we do this by subdividing the complex in nice parts and using the Mayer-Vietoris sequence in combination with an approximation of the Betti number by count of cells.

The results in this section are mostly from the two papers by Yogeshwaran et al. [16], and by Trinh Khanh Duy [4]. The first one is quite general in setting, but restricts itself to (stationary) point processes in a window $[-K, K]^d$. The second one expands to binomial and Poisson point processes with a more general density function, but does not contain central limit theorems as the first does.

We will explain the results of the papers regarding binomial and Poisson point processes, and give the ideas for some of the proofs. We will restrict ourselves to limits of the expectation.

8.1 Expected Betti number in a window

We first look at the result in the paper of Yogshwaran et al., as these are later generalised by Duy. In short, the results say that the expected Betti number in the thermodynamic regime is linear in he number of points. Yogeshwaran proves this for a Čech complex on uniformly distributed points in a box $[-K, K]^d$. The result is as follows, where we have rewritten it so that instead of enlarging the window, we increase the number of points in a fixed window.

Theorem 8.1.1 (Lemma 3.3 from [16]). *Let $nr_n^d \to \lambda \in (0, \infty)$ and $0 \leq k \leq d-1$. Write $\beta_{n,k}$ for the k-th Betti number of the complex $\check{\mathcal{C}}(\Phi_n, r_n)$ with Φ_n either a binomial or a Poisson point process. Then there exists a constant $C_{d,k,\lambda}$*

such that

$$\lim_{n\to\infty} \frac{\mathbb{E}(\beta_k)}{n} \to C.$$

Note that Kahle (Theorem 4.1 [10]) also proves that $\mathbb{E}(\beta_k) = \Theta(n)$, and does this for more general density functions. However, his theorem does not give an explicit limit as we will see in the theorem by Duy, of which the proof idea is similar to the proof in Yogeshwaran.

We will now look at the proof of Theorem 8.1.1. We mainly give the idea and the most vital steps, for the details we refer to the paper.

8.1.1 Proof sketch

The idea is to use the bound on the Betti number we saw in Lemma 3.3.12. This was the following bound: for a complex $\mathcal{K}$ with subcomplex $\mathcal{K}'$ we have

$$|\beta_k(\mathcal{K}) - \beta_k(\mathcal{K}')| \leq \sum_{j=k}^{k+1} \#\{j\text{-simplices in } \mathcal{K} \setminus \mathcal{K}'\}.$$

To use this, we have to pick smart choices for $\mathcal{K}$ and $\mathcal{K}'$.

Notably, $\mathcal{K}$ will of course be the full random complex. For $\mathcal{K}'$ we choose a subcomplex that consists of a lot of smaller subcomplexes: Let $\{Q_{i,t}\}_{i\in[m^d]}$ be a partition of the window in m^d 'cubes' $[0,t)^d$ of side length t. Denote with $\Phi_{Q_{i,t}} = \Phi_n \cap Q_{i,t}$ the subset of points that lie in the square $Q_{i,t}$, then we pick for $\mathcal{K}'$ the complex

$$\cup_{i\in[m^d]}\check{\mathcal{C}}(\Phi_{Q_{i,t}}, r_n).$$

We estimate $\beta_k(\mathcal{K}')$ for t such that $Q_{i,t}$ contains a 'suitable number of points', and we note that

$$\beta_k(\mathcal{K}') = \sum_{i\in[m^d]} \beta_k(\check{\mathcal{C}}(\Phi_{Q_{i,t}}, r_n)).$$

In this case, the expectation for each of these summands is the same, and approximately linear in the number of points. Counting the relevant number of squares we can conclude that the Betti number of $\mathcal{K}'$ is of order $\Theta(n)$. The argument for this is actually fairly non-trivial, but not so interesting for the intuition regarding this proof.

The crucial step in the proof is now to realise that the number of j-simplices that we have not yet counted, have all their vertices close to the 'boundaries', i.e. they are in the set

$$B := \cup_{i\in[m^d]}\{x \in \mathbb{R}^d : d(x, \delta(Q_{i,t})) < 2r_n\}.$$

An illustration of this situation can be found in Figure **??**. Because the volume of this set is small, the number of these simplices is also small. This makes it possible to estimate the Betti number of the complex by looking at the Betti number of $\mathcal{K}'$, which we determined was of order $\Theta(n)$.

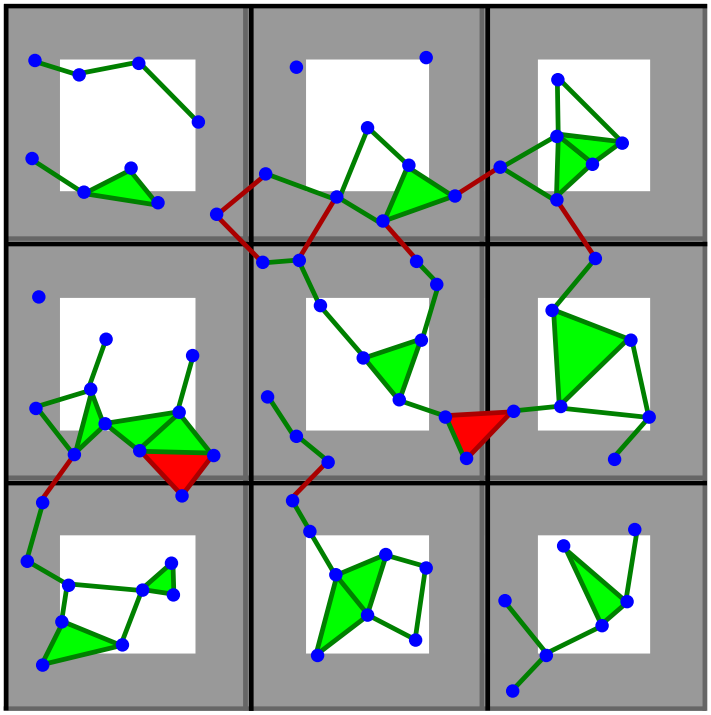


Figure 8.1: The situation for the proof. The grey/Black boxes are the $Q_{i,t}$. The grey shaded region is the region B close to the boundary. The green subcomplex is $\mathcal{K}'$ and the red simplices are the simplices in $\mathcal{K} \setminus \mathcal{K}'$. Note that the vertices of the red simplices all lie in B.

8.1.2 Vietoris-Rips

In the paper it is said that most of their results can probably also be proven for other complexes, like the Vietoris-Rips complex. This is true for this statement, the reason being that the only Čech complex specific part is the bound on the number of simplices. The paper gives such a bound for the Čech complex, but it only uses the fact that there should be a certain number of points close enough (k other points for a k-simplex). Hence this bound can actually also be used for Vietoris-Rips complex.

Theorem 8.1.2. *Let $nr_n^d \to \lambda \in (0, \infty)$ and $0 \leq k$. Write $\beta_{n,k}$ for the k-th Betti number of the complex $\mathcal{R}(\Phi_n, r_n)$ with Φ_n either a binomial or a Poisson point process with uniform density over a window $[-K, K]^d$. Then there exists a constant $C_{d,k,\lambda}$ such that*

$$\lim_{n\to\infty} \frac{\mathbb{E}(\beta_k)}{n} \to C.$$

Note that saying this is all that is treated in the paper would do it great

injustice. There are many more much more general results in it, that also merit study. Time constraints are the main reason for not looking at these in this thesis.

8.2 Expected Betti number with density f

8.2.1 Kahle's theorem

For a weak result which we can prove easily, we turn to Kahle's paper theorem 4.1 [10], which estimates the Betti number in the thermodynamic regime by again counting empty simplices or cross-polytopes.

It is a basic result from Penrose's book that, in the thermodynamic regime, any kind of component occurs linearly in the number of points (Proposition 3.3 from [14]). Similarly, the induced subgraphs we count to estimate he number of k-simplices also occur in the order of $O(n)$ times. This means $\beta_k = \Theta(n)$.

It is however interesting to know with what constant this is. For this we can turn to the result by Duy, which we look at shortly in the next section.

8.2.2 Duy's Remark

The result by Duy is interesting because it is slightly more general in terms of possible density functions. However, we will not really look at the proof, we will only give an idea of the reason it is true, which uses the idea of the proof in previous section. The result is again linearity of the expectation in the thermodynamic regime.

Theorem 8.2.1 (Theorem 1.3 and 1.5 of Duy [4])**.** *Let $nr_n^d \to \lambda \in (0, \infty)$ and $0 \leq k$. Write $\beta_{n,k}$ for the k-th Betti number of the complex $\mathcal{R}(\Phi_n, r_n)$ with Φ_n either a binomial or a Poisson point process with bounded and Riemann integrable density function with bounded support. Then there exists a constant $C_{d,k,\lambda}$ such that*

$$\lim_{n\to\infty} \frac{\mathbb{E}(\beta_k)}{n} \to C.$$

Remember that in the last proof, we saw that we could estimate

$$\beta_k(\mathcal{K}') = \sum_{i\in[m^d]} \beta_k(\check{\mathcal{C}}(\Phi_{Q_{i,t}}, r_n)).$$

In that case, the expectation for each of these summands was the same, but dependent on the density of the points. Now, we have a non-uniform density function. The trick to doing the proof in this case is to let these boxes become infinitesimally small. So in the previous case we had some constant density in each box giving a constant for the limit of the expectation. Now we sum over squares in which the density is variable, and the limit becomes an integral 'summing' the density over all these squares.

To do this, we actually first need to know that for some constant $\hat{\beta}_k$ and $nr_n^d \to \lambda$

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{\beta_k(\Phi_n, r_n)}{n} = \hat{\beta}_k(\lambda)\right) = 1.$$

This is also proved by Yogeshwaran, but we decided not to focus on that result. What we can note however, is that these constants $\hat{\beta}_k$, that depend on $f(x)$, actually define the limit of the expectation by integration.

8.3 Vanishing of homology

We shortly return to the definition of the regimes used for random complexes as compared to regimes of random geometric graphs. As said in Section 6.2.1 the definitions might slightly diverge in the thermodynamic regime. In particular, we noted that the definitions of sub-critical, critical and super-critical were different. In the case of random geometric graphs, there is a clear (critical) limit at which percolation occurs. This limit is less interesting for geometric complexes as studied for their shape, because percolation only means there is one large component, and nothing is being said about the other small components.

It should be noted that for geometric complexes, there also exist interesting critical limits. These critical values for relate to connectivity. There is a critical limiting value for which the complexes/graphs have only one component. This corresponds to the 'vanishing' of the 0-th Betti number (actually, it becomes one, but this is called vanishing here). For higher Betti numbers, similar critical values exist. This has been proven in the case of a point distribution on a flat torus by Bobrowski and Weinberger in a yet to be published paper. [2]

Chapter 9

Supercritical regime

As mentioned earlier, in the super-critical regime the density of the points becomes very large. So large that it is very likely that all points become highly connected. In this section we look at this connectivity for the random Čech and the random Vietoris-Rips complexes. Unlike in the other regimes, the results and the proofs for these two complexes are quite different. This is because the Čech complex is limited by the implications of the Nerve theorem, where the Vietoris-Rips complex is not. Most importantly, we look at two results from Kahle's paper [10] about connectivitity of the complexes in certain regimes.

9.1 Contractible Čech complexes

For the point processes in this section, we assume a uniform distribution of the points on a smoothly bounded convex subset K of $\mathbb{R}^d$. If the cover defining the random Čech complex is such that it covers all of K, we will be able to conclude, using the nerve lemma, that the complex is contractible. The way we ensure that the cover actually covers K is by the following lemma, where we define a set of boxes.

Proposition 9.1.1. *Let K be compact smoothly bounded convex subset of $\mathbb{R}^d$. Then there exists a length $L > 0$ such that if $\lambda < L$, there is a partition of $\mathbb{R}^d$ in boxes of side length λ such that the following holds. Denoting by S_K the set of boxes fully contained in K: for every point $k \in K$ we have $k \in s$ for some $s \in S_K$, or the box which contains k is a adjacent to a box in S_K.*

The following proof of this proposition is in no way rigorous, but the strategy should be transferable to a rigorous proof.

Proof sketch. Let us oriente the boxes so that the sides are normal to respective basis vectors. Note that, because K is convex, we have exactly one minimal and one maximal point in the direction of each basis vector. In these points, the tangent plane is normal to the respective basis vector, i.e. it is parallel to

the side of the box. In any other point k, the tangent plane is not parallel to any side of the box.

In the second type of point there is a corner of the box (in $\mathbb{R}^2$ there is exactly one) that is on the side of the tangent plane containing K (which exists because K is convex). It does not matter how we shift the partition, as long as we keep the point in this box. If $\lambda < \lambda_k$ is small enough, the box touching this corner is completely contained in S_K. For this we need that the boundary of K is continuously differentiable, i.e. the tangent plane is actually a good approximation of the boundary.

For the remaining points, we use the same reasoning regarding the tangent plane, but note that for each size of the box, there might be some shifts that are not suitable. Fix the shift in all directions but the one in the direction of the tangent basis vector. Now shift in the direction of the tangent basis vector such that for both points (maximum and minimum for this direction), the intersection of the boundary of K and the box does not touch the 'tangent side' of the box. If λ_k is small enough, the tangent plane will not diverge from the actual boundary more than half the side length. This means the possible shifts are of an interval greater than half the side length, and these intervals will intersect for the minimal and maximal point. Hence for these points there also is a minimal side length λ_k such that there is a shift of he partition which fulfils the condition.

Finally, because K is bounded, the boundary of K is compact. Hence the minimum value of all the $\lambda_k > 0$ is attained. Because all interior points of K will now automatically also be in a box together with a boundary point, or in a box fully contained in K, we have proven the proposition.

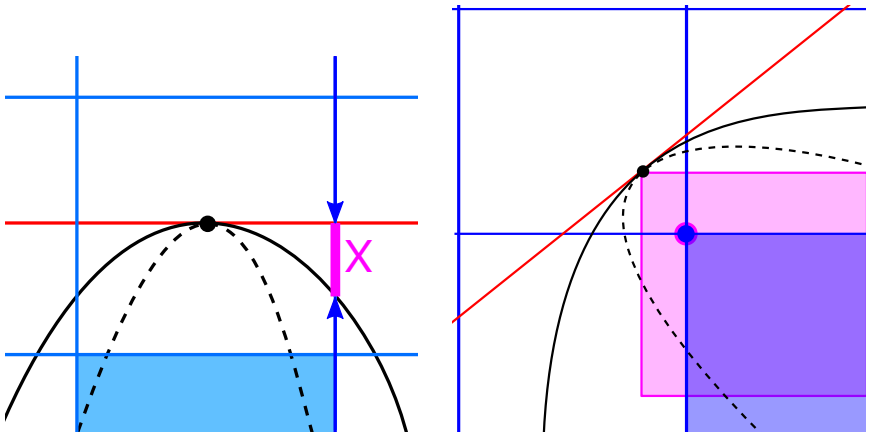
Note that this proof is here limited to a convex and smoothly bounded K. It is, however, not hard to imagine that the convexity might not be needed here. This would also give a slight generalisation of the following theorem about the contractibility of complexes, which we talk about after the statement.

Theorem 9.1.2 (Theorem 6.1 [10])**.** *For a uniform distribution on a smoothly bounded convex subset $K \subset \mathbb{R}^d$, there exists a constant c_K such that if $r_n \geq c_K(\log n/n)^{1/d}$, then the random Čech complex $\check{\mathcal{C}}(X_n, r_n)$ is a.a.s. contractible.*

The proof heavily relies on the Nerve theorem (Theorem 4.2.1). The idea is to make the cover, which defines $\mathbb{C}(X_n, r_n)$, a.a.s. cover the convex set K. If this happens, then by the Nerve theorem: the nerve of the cover (i.e. $\mathbb{C}(X_n, r_n)$) is homotopy equivalent to the union of the cover, which 'is' the convex set K. This means the complex is contractible, because the convex set is.

Now the slight generalisation we mentioned, is that we do not have to consider contractability as an end result, but rather the homotopy equivalence of K and the complex for slightly more general K.

Proof. We can assume that $r_n \to 0$ as $n \to \infty$, as otherwise it is clear that the balls will a.a.s. cover the convex body. Hence we will assume $r_n \to 0$ and $r_n \geq c(\log n/n)^{1/d}$.



(a) Minimal/maximal case (b) Remaining cases

Figure 9.1: An illustration of the proof of Proposition 9.1.1 in $\mathbb{R}^2$. Left we see the case of the maximal/minimal point, right the case of the remaining points. Red lines indicate tangent planes; Blue are the relevant boxes, the shaded blue box is the box fully contained in K; The black lines indicate the border of K, the solid line is the case where we have made the boxes sufficiently small, the dotted black one is an illustration of the case where he box is too large. For the minimal/maximal case the purple segment indicates the forbidden shift (for the solid black border). For the remaining cases, the purple region indicates the possible positions of the bottom-right corner.

For the proof, we partition $\mathbb{R}^d$ in small boxes, such that if each of the boxes completely contained in K contains a point, then the balls around the points cover the whole body. For this partition we use $\lambda\mathbb{Z}^d$, the cubical lattice in $\mathbb{R}^d$ with side lengths $\lambda_n > 0$. Note that if we take $\lambda_n = r_n/(2\sqrt{d})$, the distance from any point to another point in an adjacent box is maximally $2\sqrt{d\lambda_n^2} = 2\lambda_n\sqrt{d} = r_n$. Hence if we have a point in a box, the ball with radius r around it will indeed cover all adjacent boxes, too.

Now, define S_K to be the set of all boxes completely contained in K. Using the proposition above, it is clear that if every completely contained box contains a point, then K is covered by the balls

We now ask: What is the limiting probability that every completely contained box gets a point? To answer this, we just estimate the probability that some box does not get a point. We first note that the probability that any one box $B \in S_K$ does not get a point is

$$\begin{aligned} P_0 := \mathbb{P}(X_n \cap B = \emptyset) &= (1 - \mu(B)/\mu(K))^n \\ &= (1 - \lambda_n^d/\mu(K))^n, \end{aligned}$$

because the distribution of points is uniform (explains $\mu(B)/\mu(K)$) and the points are distributed independently (explains the n-th power). By the calculus inequality $(1-x)^n \le e^{-nx}$ we can bound this probability by

$$\begin{aligned} P_0 \le e^{-n\lambda_n^d/\mu(K)} &= e^{-n(r_n/(2\sqrt{(d)})^d)/\mu(K)} \\ &= e^{-nr_n^d/(2^d d^{d/2}\mu(K))} \\ &= e^{-nr_nC}, \end{aligned}$$

where $C = 1/(2^d d^{d/2}\mu(K))$ is a constant independent of n.

Now we use $r \ge c_K(\log n/n)^{1/d}$ to see that

$$\begin{aligned} P_0 &\le e^{-nr_nC} \\ &\le e^{-nc_K(\log n/n)^{1/d}C} \\ &= n^{-Cc_k^d} \end{aligned}$$

We now bound the number $|S_K|$ of boxes fully contained in K by an upper bound N using an easy volume computation. The body K has a well defined finite volume $\mu(K)$ and each box has volume λ^d, hence an upper bound for the number of boxes fully contained in K is

$$N = \mu(K)/\lambda^d + O(1/\lambda^{d-1}) = (1 + o(1))/Cr^d$$

Now use a very crude bound $P_f < |S_K|P_0$ for the probability P_f that some box in S_K does not contain a point of X_n. This gives

$$P_f \le \frac{(1+o(1))}{Cr^d} n^{-Cc_k^d} \le \frac{(1+o(1))}{Cc_k^d \log n/n} n^{-Cc_k^d} = \frac{(1+o(1))}{Cc_k^d \log n} n^{1-Cc_k^d}.$$

If we want P_f to go to 0 as $n \to \infty$, it is clear that it suffices to ensure the $n^{1-Cc_k^d}$ factor is less or equal to 1. This is easily done by choosing $c_k \geq C^{-1/d}$. This makes K a.a.s. covered by the balls, and hence the Čech complex a.a.s. contractible.

9.2 k-Connected Vietoris-Rips complexes

For Rips complexes a similar result holds. We have to replace contractible by k-connected (all homotopy groups up until the k-th are trivial), however. This is because the dimensionality of Rips complexes is fundamentally different from Čech complexes as noted in propositions 6.1.2 and 6.1.4.

Theorem 9.2.1. *[Theorem 6.5 [10]] For a uniform distribution on a smoothly bounded convex subset $K \subset \mathbb{R}^d$ and a fixed integer $k > 0$, there exists a constant $c_{K,k}$ such that if $r_n \geq c_K(\log n/n)^{1/d}$, then the random Rips complex $\mathcal{R}(X_n, r_n)$ is a.a.s. k-connected.*

The proof of this theorem is also quite different from the case of the Čech complex, because we cannot use the nerve theorem. We instead have to make use of discrete Morse theory, and the following geometric lemma. Again, all lemmas and theorems in this section are due to Kahle; for the bigger proofs I filled in some details which were missing in the original paper.

Lemma 9.2.2. *There exists a constant $\epsilon_d > 0$ such that the following holds for all $r > 0$. Let $l \geq 1$, and let $\{y_0, \ldots, y_k\} \subset \mathbb{R}^d$ be an l-tuple of points such that*

$$\| y_0 \| \leq \| y_1 \| \leq \cdots \leq \| y_l \| .$$

If $\| y_0 - y_1 \| > r$ and $\| y_i - y_j \| \leq r$ for every other $0 \leq i \leq j \leq l$, then the intersection

$$I = \bigcap_{i=1}^{l} B(y_i, r) \cap B(0, \| y_1 \|)$$

satisfies $\mu(I) \geq \epsilon_d r^d$.

Intuitively the conditions on y_0 ensure that all the points lie sufficiently close to each other and to the origin, so that the mutual intersection of balls will also lie close enough (with at least part of it within $\| y_1 \|$) to the origin. Additionally, these conditions are needed to make sense of the role of y_0 in the discrete vector field we will construct.

Because I could not reconstruct some of the arguments that were skipped in the paper, I give a modified version. The proof of this lemma is basically a straightforward calculation on a scaled ($r = 1$) version of the problem.

Proof. The following proof is for the case $r = 1$, the full statement follows by rescaling.

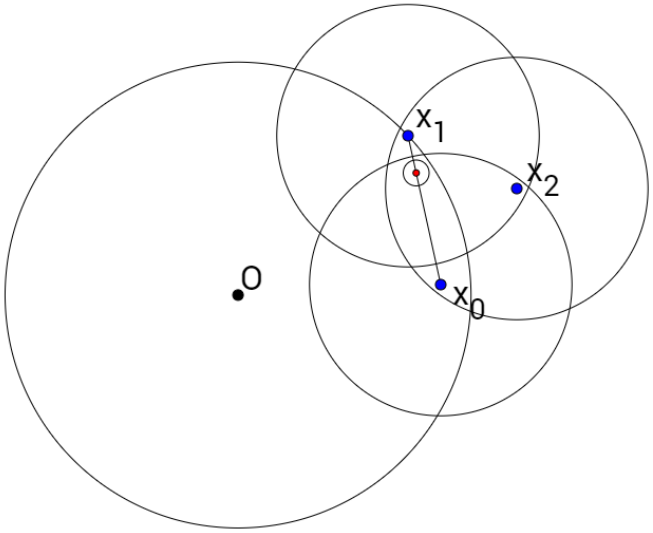


Figure 9.2: The setting of the geometric lemma with a bit of the idea for the proof. The proof assures that the ball of radius $\sqrt{13}/4$ around point x_m will lie within all balls of radius 1 around x_i for $i > 0$.

Let us start by separately handling the case $l = 1$, as the following proof only works for for $l > 1$. The case $l = 1$ is obvious because $\| y_1 \| > 1/2$ (otherwise the conditions are no met), which gives the bound $\mu(I) \geq \mu(B(0, 1/2))$.

Now for $l > 1$: Let $y_m = (y_0 + 3y_1)/4$ be a point on the line segment y_0y_1. We will show that the distance from y_m to any of the points y_i for $1 \leq i \leq l$ is bounded smaller than 1. This means that there is always some ball around y_m of size bounded away from 0 contained in $\cap_{i=1}^{l} B(y_i, 1)$. From this we deduce that the volume of I as in the lemma is bounded away from 0.

Now let θ_j denote the angle between $y_0 - y_j$ and $y_1 - y_j$ for $2 \leq j \leq l$. The law of cosines gives that

$$\begin{aligned}(y_0 - y_j) \cdot (y_1 - y_j) &= \| y_0 - y_j \| \| y_1 - y_j \| \cos\theta_j \\ &= \frac{1}{2}(\| y_0 - y_j \|^2 + \| y_1 - y_j \|^2 - \| y_0 - y_1 \|^2) \\ &< \frac{1}{2},\end{aligned}$$

because per assumption $\| y_0 - y_j \|^2 \leq 1$, $\| y_1 - y_j \|^2 \leq 1$ and $\| y_0 - y_1 \|^2 > 1$.

Using this we compute the distance between y_m and y_j.

$$\begin{aligned}
\| y_m - y_j \|^2 &= (y_m - y_j) \cdot (y_m - y_j) \\
&= \left(\frac{y_0 + 3y_1}{4} - \frac{4y_j}{4} \right) \cdot \left(\frac{y_0 + 3y_1}{4} - \frac{4y_j}{4} \right) \\
&= \left(\frac{y_0 - y_j}{4} + 3\frac{y_1 - y_j}{4} \right) \cdot \left(\frac{y_0 - y_j}{4} + 3\frac{y_1 - y_j}{4} \right) \\
&= \frac{1}{16} \left(\| y_0 - y_j \|^2 + 9 \| y_1 - y_j \|^2 + 6(y_0 - y_j) \cdot (y_1 - y_j) \right) \\
&< \frac{1}{16}(1 + 9 + 6(\frac{1}{2})) \\
&= \frac{13}{16},
\end{aligned}$$

hence

$$\| y_m - y_j \| < \frac{\sqrt{13}}{4}.$$

Now set $\rho = 1 - \frac{\sqrt{13}}{4}$. Then obviously $B(y_m, \rho) \subset B(y_j, 1)$ for any $2 \leq j \leq l$. Now we also want $B(y_m, \rho) \subset B(y_1, 1)$; by the triangle inequality (taking any point y_j, $2 \leq j \leq l$) $\| y_0 - y_1 \| < 2$, hence $\| y_m - y_1 \| < 1/2$. This gives us $B(y_m, \sigma) \subset B(y_1, 1)$ for any $\sigma < 1/2$ and in particular for $\sigma = \rho$. With these inclusions we also have

$$B(y_m, \rho) \cap B(0, \| y_1 \|) \subset \bigcap_{i=1}^{l} B(y_i, 1) \cap B(0, \| y_1 \|).$$

By the triangle inequality $\| y_m \| < \| y_1 \|$ and hence

$$\begin{aligned}
\mu \left(\bigcap_{i=1}^{l} B(y_i, 1) \cap B(0, \| y_1 \|) \right) &\leq \mu \left(B(y_m, \rho) \cap B(0, \| y_1 \|) \right) \\
&\leq \mu \left(B(y_1, \rho) \cap B(0, \| y_1 \|) \right).
\end{aligned}$$

This last volume is bounded below by a volume greater than 0 because $\| y_1 \| > 1/2$ as we saw in the case $l = 1$. Take ϵ_d to be this lower bound (greater than 0) and the lemma follows.

In the original proof, the choice or y_m was $(y_0+y_1)/2$, instead of $(y_0+3y_1)/4$. With this choice the ball $B(y_m, \rho)$ (with corresponding ρ) might not be contained in $B(y_1, 1)$, as there are situations possible where $\| y_1 - y_m \| \approx 1$. The bigger problem is that the lemma they actually want to use is slightly different, as in the end we will be working with points chosen from a convex smoothly bounded set $K \subset \mathbb{R}^d$. Hence the important volume is not $\mu(I)$, but $\mu(I \cap K)$. The author of the paper probably left out an easy argument to fix these problems. As said earlier, I could not find this argument and have worked around it differently. I feel the missing argument must use the scale of the problem, together with

convexity: at 'small scale' everything looks like euclidean space, or euclidean half-space in a nice way.

Lemma 9.2.3. *Let K be a convex smoothly bounded subset of $\mathbb{R}^d$. There exists a constant $\epsilon_{d,K} > 0$ such that the following holds for all $0 < r < R$ for some constant $R > 0$. Let $l \geq 1$, and let $\{y_0, \ldots, y_k\} \subset K$ be an l-tuple of points such that*

$$\| y_0 \| \leq \| y_1 \| \leq \cdots \leq \| y_l \| .$$

If $\| y_0 - y_1 \| > r$ and $\| y_i - y_j \| \leq r$ for every other $0 \leq i \leq j \leq l$, then the intersection

$$I = \bigcap_{i=1}^{l} B(y_i, r) \cap B(0, \| y_1 \|)$$

satisfies $\mu(I \cap K) \geq \epsilon_d r^d$.

The proof is initially the same, but is slightly different following the statement

$$B(y_m, \rho) \cap B(0, \| y_1 \|) \subset \bigcap_{i=1}^{l} B(y_i, 1) \cap B(0, \| y_1 \|).$$

From then on we cannot just move our ball of radius ρ, as the neighbourhood actually matters here.

Proof. Assume w.l.o.g. that the origin is contained in K. As in the previous lemma, we have $\rho_r > 0$ and

$$B(y_m, \rho) \cap B(0, \| y_1 \|) \subset \bigcap_{i=1}^{l} B(y_i, 1) \cap B(0, \| y_1 \|).$$

Now we focus on the intersection of $B(y_m, \rho_r)$ and K, and try to find a lower bound for its volume. Note that the worst case scenario is for when y_m lies on the boundary of K. Using the fact that Gaussian curvature is continuous on the smooth boundary, and that the boundary is compact, we see that there is a maximum curvature on the boundary somewhere.

The Gaussian curvature at a point is $1/r^2$, where r is the radius of the sphere touching the boundary at the point. Denote the minimal radius over the whole boundary by r' This means that the intersection $B(y_m, \rho_r) \cap K$ will in the limit be at least as large $B(O, r') \cap B(r'e_0, \rho)$, where e_0 is a unit vector. This approaches $\mu(B(O, \rho_r))/2$ as $r \to 0$.

We conclude that for every $\epsilon' > 0$ there is a $R > 0$ such that if $r < R$, then $\mu(K \cap I) > \epsilon'_d r^d$. ◆

The geometric lemma above is needed to prove facts about a gradient vector field on the Rips complex. In particular it gives us an estimation of how likely we will get a critical k-cell. The proof of the theorem comes down to the definition of a discrete vector field with certain properties that are easy to work with. The lemma then gives us an estimation of the number of critical k-cells and hence

by the Morse theorem, of the homology of the complex. We will show that the expected number of cells of dimension k or lower goes to 0, except for one 0-cell. This ensures that the complex will a.a.s. be k-connected.

Proof of Theorem 9.2.1. The points X_n are in general position, hence no two points will have the same distance to the origin. Use this fact to index the points such that $\| x_1 \| < \cdots < \| x_n \|$. Now define a discrete vector field on $\mathcal{R}(X_n, r_n)$.

If for a cell $S = \{x_{i_1}, \ldots, x_{i_j}\}$ there exists $i_0 < i_1$ such that $\{x_{i_0}\} \cup S$ is a cell in the complex, then pair S with $\{x_{i_0} \cup S\}$ for the smallest such i_0. It is easy to see that this gives a pairing on the cells of the complex such that we get a discrete vector field: no cell can be paired with more than one other cell.

To be able to use discrete Morse theory, we need a gradient vector field, i.e. a vector field without cycles. In this case we can easily see no such paths exist, because the indices that define subsequent simplices 'decrease': going to a lower dimensional cell we loose any vertex, but going to a higher dimensional cell we may only add a vertex with index smaller than all the others. Obviously this prevents us from finding a cycle.

We will now bound the probabilities that a subset $F = \{x_{i_1}, \ldots, x_{i_{k+1}}\}$ of $k+1$ points forms a k-cell and that such a cell is unpaired (critical) in our vector field:

$$\begin{aligned} p_k &:= \mathbb{P}(F \in \mathcal{R}(X_n, r)) \\ p_c &:= \mathbb{P}(F \text{ critical} | F \in \mathcal{R}(X_n, r)). \end{aligned}$$

If we are given one of the $k+1$ points, and they form a k-simplex, the other k points should all lie within a ball of radius r around this first point. Because the other k points are independent, and the probability that one point lands in such a ball is bounded above by $\mu(B(0,r)) = O(r^d)$, we know that

$$p_k = O(r^{dk}).$$

Now for the probability of such a k-cell F to be critical. Note that for F to be critical, there may be no point $x_a \in X_n$ with $a < i_1$ within distance r from all the vertices of F, because in that case F would be paired with $x_a \cup F$; there also needs to be a point x_{i_0} with $i_0 < i_1$ such that $x_{i_0} \cup (F \setminus \{x_{i_1}\})$ is a cell in the complex, because otherwise there would be a pairing of F with $F \setminus \{x_{i_1}\}$. Set A to be the event that for F there is such a point x_{i_0} with $\{x_{i_0}\} \cup F \notin \mathcal{R}(X_n, r)$ and $x_{i_0} \cup (F \setminus \{x_{i_1}\}) \in \mathcal{R}(X_n, r)$, then

$$p_c \leq \mathbb{P}(F \text{ critical} | F \in \mathcal{R}(X_n, r) \text{ and } A) =: p_c',$$

because A is an event necessary for F to be critical.

Note that the event A puts us precisely in the situation of the geometric lemma, with $y_j := x_{i_j}$. Because the existence of x_{i_0} excludes the possibility of F being paired with $F \setminus \{x_{i_1}\}$, the only way for F to be non-critical is the existence of some x_a with $a < i_1$ (i.e. $x_a \in B(0, \| x_{i_1} \|)$) and $\| x_a - x_{i_j} \| < r$ for all $1 \leq j \leq k+1$ (i.e. $x_a \in \cap_{i=1}^{k+1} B(x_{i_j}, r)$). The location where this x_a should

be (for F to be paired) is precisely the set I of the geometric lemma of which the volume is bounded.

Because K is convex and smoothly bounded, we will always have that $\mu(I \cap K) \geq \epsilon_d r^d$ (by Lemma 9.2.3). Hence the probability that none of the other $n-(k+2)$ points will fall in I is

$$p'_c = (1 - \frac{\mu(K \cap I)}{\mu(K)})^{n-k-2} \leq (1 - \frac{\epsilon_d}{\mu(K)} r^d)^{n-k-2}.$$

We can now bound this above by seeing that

$$(1 - \frac{\epsilon_d}{\mu(K)} r^d)^{n-k-2} \leq e^{-\frac{\epsilon_d}{\mu(K)} r^d (n-k-2)} = O(e^{-cnr^d}),$$

where c is a constant such that $0 < c < \frac{\epsilon_d}{\mu(K)}$.

Let C_k denote the number of critical k-faces in the complex, then (by linearity of expectation) we have

$$\mathbb{E}(C_k) \leq \binom{n}{k+1} p_f p_c = O\left(\binom{n}{k+1} \left(\frac{nr^d}{n}\right)^k e^{-cnr^d}\right).$$

Now simplifying we get

$$O\left(\binom{n}{k+1} \left(\frac{nr^d}{n}\right)^k e^{-cnr^d}\right) = O\left((nr^d)^k e^{-cnr^d} n\right).$$

Note that by the main theorem of discrete Morse theory, this means that

$$\mathbb{E}(\beta_k) = O\left((nr^d)^k e^{-cnr^d} n\right). \tag{9.1}$$

We are not done however, because we want something stronger: we want to prove that the complex is k-connected. We now use the regime $r \geq c_k (\log n/n)^{1/d}$ to see that for suitable c_k the expected number of k'-faces with $1 \leq k' \leq k$ goes to zero:

$$\mathbb{E}(C_{k'}) = O\left((nr^d)^k e^{-cnr^d} n\right) = O\left((c_k^d \log n)^k n^{1-cc_k^d}\right),$$

for all $1 \leq k' \leq k$.
If $c_k > (1/c)^{1/d}$ (which we can choose) then $(c_k^d \log n)^k n^{1-cc_k^d} \to 0$ if $n \to \infty$, hence a.a.s. we will have no cells in dimensions $1 \leq k' \leq k$. In dimension 0 we keep one critical 0-cell because we need one for each component.

Discrete Morse theory (Theorem 4.3.8) then says that the complex is homotopy equivalent to a complex with a 0-cell, no k'-cells for $1 \leq k' \leq k$ and maybe cells in higher dimension. This proves that the complex is k-connected. The theorem hence holds with constant $c_k > \left(\frac{1}{c}\right)^{1/d}$, where c is a constant such that $0 < c < \frac{\epsilon_d}{\mu(K)}$.

Bibliography

[1] Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, pages 9–25, 1989.

[2] Omer Bobrowski and Shmuel Weinberger. On the vanishing of homology in random Čech complexes. *arXiv preprint arXiv:1507.06945*, 2015.

[3] Karol Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae*, 35(1):217–234, 1948.

[4] Trinh Khanh Duy. A remark on the convergence of betti numbers in the thermodynamic regime. *arXiv preprint arXiv:1606.06518*, 2016.

[5] Robin Forman. Morse theory for cell complexes. *Advances in mathematics*, 134(1):90–145, 1998.

[6] Allen Hatcher. Algebraic topology. 2002. *Cambridge UP, Cambridge*, 606(9).

[7] Philip S Hirschhorn. Notes on homotopy colimits and homotopy limits (a work in progress). 2014.

[8] Jakob Jonsson. *Simplicial complexes of graphs*. Springer, 2008.

[9] Matthew Kahle. Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658–1671, 2009.

[10] Matthew Kahle. Random geometric complexes. *Discrete & Computational Geometry*, 45(3):553–573, 2011.

[11] Matthew Kahle and Elizabeth Meckes. Limit theorems for betti numbers of random simplicial complexes. *Homology, Homotopy and Applications*, 15(1):343–374, 2013.

[12] Matthew Kahle and Elizabeth Meckes. Erratum to "limit theorems for betti numbers of random simplicial complexes". *Homology, Homotopy and Applications*, 18(1), 2016.

[13] Dimitry Kozlov. *Combinatorial algebraic topology*, volume 21. Springer Science & Business Media, 2007.

[14] Mathew Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in probability*. Oxford University Press, 2003.

[15] Marten Scheffer, SH Hosper, ML Meijer, Brian Moss, and Erik Jeppesen. Alternative equilibria in shallow lakes. *Trends in ecology & evolution*, 8(8):275–279, 1993.

[16] D Yogeshwaran, Eliran Subag, and Robert J Adler. Random geometric complexes in the thermodynamic regime. *Probability Theory and Related Fields*, pages 1–36, 2016.