Simplicial Complexes for Topological Data Analysis

Daniel Seitz

June 2020

Contents

1 Simplicial complexes

1.1 Simplices [1]

Before defining what a **simplex** is, we need to discuss the elements which are used to construct them. For this purpose, we define a subset of R^n, $V = \{v_0, ..., v_n\}$ such that the vectors

$$\{v_i - v_0 \mid v_i \in V \wedge v_i \neq v_0\}$$

are linearly independant. Then we call V a **vertex set** with **geometrically independent** elements $v_i \in V$.

We also need to define the **convex hull** $conv(X)$ of a subset X of a real or complex vector space S.

$$conv(X) := \bigcap_{\substack{X \subseteq K \subseteq S \\ K \text{ convex}}} K$$

with a set K being **convex** if, given any two points in K, their direct line segment is contained in K.

Now we define a (geometric) **n-simplex** θ as the **convex hull** of a **geometrically independent** vertex set V with $n+1$ elements. Some examples can be seen in *Figure 1*.

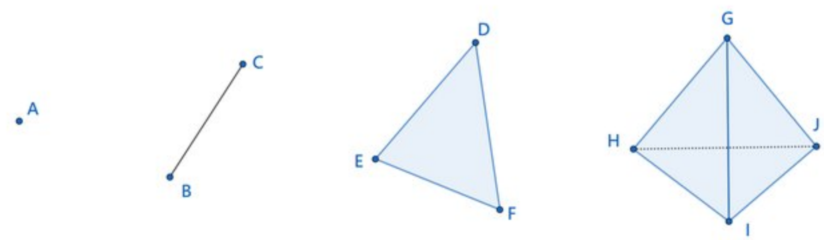

Figure 1 – From left to right: 0-simplex, 1-simplex, 2-simplex, 3-simplex. [5]

Definition: Let $\{v_0, ..., v_n\}$ be a **geometrically independent set** in R^n. We define the **n-simplex** θ spanned by $v_0, ..., v_n$ to be the set of all points $x \in R^n$ such that

$$x = \sum_{i=0}^{n} t_i v_i, \text{ where } \sum_{i=0}^{n} t_i = 1 \text{ and } t_i \geq 0$$

The numbers t_i are called the **barycentric coordinates** of the point x with respect to $v_0, ..., v_n$ (denoted as $t_i(x)$). They can be uniquely determined by x. Given this definition, we describe the simplices spanned by subsets of V as

faces. For example, the 1-simplex in *Figure 1* has 3 **faces**: the 1-simplex itself and the two vertices it consists of.

The **interior** $Int(\theta)$ of a **complex** θ is defined as

$$Int(\theta) := \{x \in \theta \mid t_i(x) > 0\}$$

Intuitively, they are the points of a simplex, excluding its boundary. Looking at the 2-simplex in *Figure 1*, the **interior** is the filled out, light blue area on the inside of the triangle. The 1-simplices between the three vertices are not part of the **interior**.

1.2 Simplicial Complexes [1]

The goal is to represent a space through the usage of **simplices**. Therefore we need a way of combining multiple **simplices**, to construct more complex structures.

Definition: Given a collection K of **simplices** in R^n, two properties have to be fulfilled:

- Any **face** of a **simplex** of K is a **simplex** of K.
- An intersection of two **simplices** must be empty or a common **face** of both.

Then we call K a **simplicial complex**.

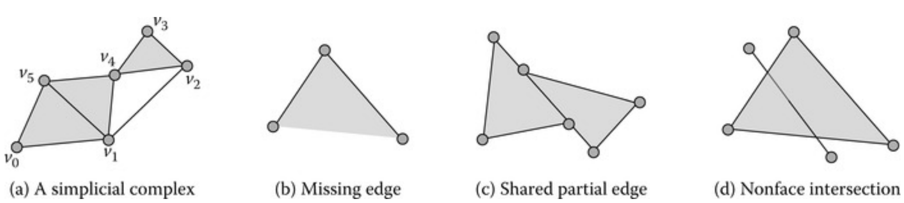

Figure 2 – One valid complex (a); Three invalid complexes (b,c,d). [4]

Subcollections of a complex K that contain all faces of its elements are simplicial complexes themselves. They are called **subcomplexes** of K.
One important **subcomplex** of K is the collection of all simplices of K with dimension at most p. This is called the **p-skeleton** of K and will be denoted $K^{(p)}$. The most important **p-skeleton** is $K^{(0)}$, containing only the vertices of K.

To extend the notion of the **interior** in the context of a **complex** K, we define the following terms:

- (Star) $St(v, K)$:= Union of the Interiors of those simplizes of K that have v as a vertex

- (Closed Star) $\overline{St}(v, K) :=$ Union of those simplizes of K that have v as a vertex
- (Link) $Lk(v, K) := \overline{St}(v, K) - St(v, K)$

Intuitively, these three definitions describe the environment of a vertex v in a complex K. Some examples are visualized in Figure 3.

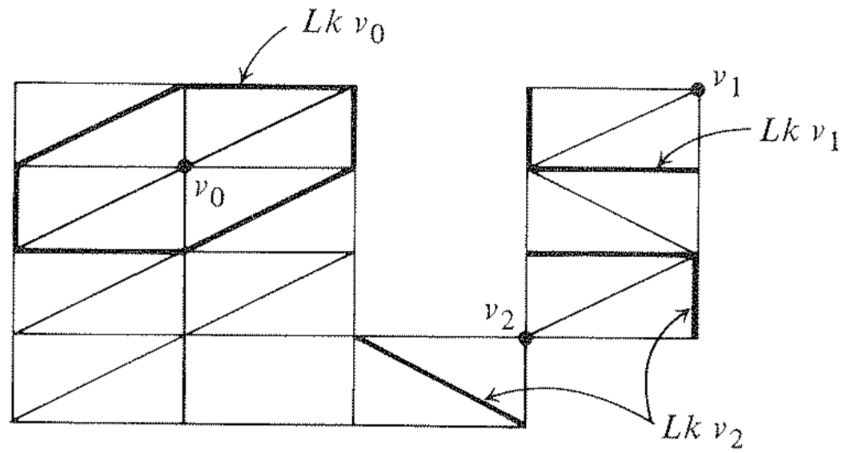

Figure 3 – Simplicial complex with links of v_0, v_1 and v_2 emphasized. [1]

Since we want to be able to utilize **simplicial complexes** in the context of topological data analysis, one needs to define the topology that is associated with a complex K. We call this the **underlying space** $|K|$. It is constructed by taking the subspace topology of the union of the contained simplices in R^n, and is fully characterized by the combinatorial description of the collection of simplices. In the following chapter we define such description.

1.3 Abstract Simplicial Complexes [1][2]

Definition: Let V be a vertex set. An **abstract simplicial complex** A consists of a subset of the power set of V. The elements of A are called **faces** and every subset of a **face** is also an element of A.
The dimension of an **abstract simplex** is defined by its cardinality - 1.
The dimension of an **abstract simplicial complex** is defined by the largest dimension of its simplices.

Example: Call the valid **geometric simplicial complex** from *Figure 2* K. The following **abstract simplicial complex** is the **combinatorial description** of K:

$$A = \{\{v_0, v_1, v_5\}, \{v_0, v_1\}, \{v_0, v_5\}, \{v_1, v_5\}, \{v_1, v_5, v_4\}, \{v_4, v_5\}, \{v_1, v_4\},$$

$$\{v_2, v_3, v_4\}, \{v_3, v_4\}, \{v_2, v_3\}, \{v_2, v_4\}, \{v_1, v_2\}, \{v_0\}, \{v_1\}, \{v_2\}, \{v_3\}, \{v_4\}, \{v_5\}\}$$

It is important to note that the relationship between a (geometric) **simplicial complex** and its **combinatorial description** is not bijective. While there can only be one **abstract complex** per **geometric complex**, different embeddings of the same **abstract complex** can result in drastically different **geometric complexes**.

1.4 Duality of Simplicial Complexes [2]

These definitions give rise to a dual perspective on simplicial complexes, as we can handle them both as topological and combinatorial objects.
Given a (geometric) **simplicial complex** G, there is a unique **abstract simplicial complex** A which describes G combinatorially. We call G the **geometric realization** of A.
Now, if a **geometric complex** G is the **geometric realization** of an **abstract complex** A, their underlying spaces are homeomorphic.
Therefore, whenever a **geometric simplicial complex** is given, we can handle it in a purely **combinatorial** way (profiting from the computational efficiency) while not losing out on any **topological** attributes.

2 Nerv Theorem

Having introduced the notion of a **simplicial complex**, the question arises how one can apply these complexes in data analysis. Naturally, we can use a given dataset X as **vertex set** and construct **complexes** on top of them. One example for such **complex** is the **Čech complex**.

2.1 Čech Complex [2]

Definition: Let X be a set of points in R^n. The **Čech complex** $\check{C}ech_\alpha(X)$ consists of all **n-simplices** $\{x_0,...,x_n\}$ with $x_i \in X$, such that the $n+1$ closed balls $B(x_i, \alpha)$ have a non-empty intersection.

Example: Given a dataset consisting of 7 points in the R^2 (*Figure 4*), we can construct the **closed ball** around each **vertex**(data point). Whenever two or more balls intersect, a new **simplex** is generated. It has to be emphasized that only the simultaneous intersection of multiple balls give rise to higher dimensional **simplices**.

2.2 Nerve [2]

The **Čech complex** is a particular case of a family of complexes associated to covers of a (data) set X. Given a cover $D = (U_i)_{i \in I}$ of X, the **nerve** of D is the abstract simplicial complex $C(D)$ whose vertices are the U_i's, such that

$$\theta = [U_{i0}, ..., U_{in}] \in C(D) \iff \bigcap_{j=0}^{n} U_{ij} \neq \emptyset$$

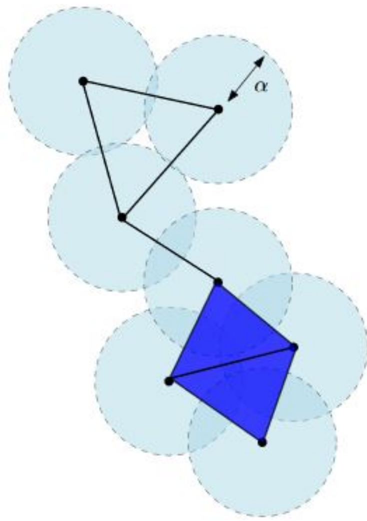

Figure 4 – A Čech complex of dimension 2. [2]

It is implied that the **nerve** can vary heavily, depending solely on the **cover** that is chosen.

Example: Figure 5 visualizes how a **nerve** can be constructed out of an arbitrary **cover**. In this case, the **cover** D consists of the sets U_1, U_2, U_3. These sets give rise to one **0-simplex** each, while their pairwise intersections give rise to **1-simplices**.

We want to make sure that the underlying space of a **nerve** is topologically equivalent to the space of our data. For this purpose we introduce the notion of a **homotopy equivalence**.

2.3 Homotopy [3]

Definition: Given two continuous maps $f, g : X \to Y$, a **Homotopy** between f and g is defined as the continuous function

$$H : X \times [0,1] \to Y \text{ with } H(x,0) = f(x) \ \wedge \ H(x,1) = g(x)$$

Intuitively, this means that f and g can be continuously deformed into one another. (*Figure 6)*

Definition: Two spaces X and Y are **homotopy equivalent** if there exist two continuous maps

$$f : X \to Y \text{ and } g : Y \to X$$

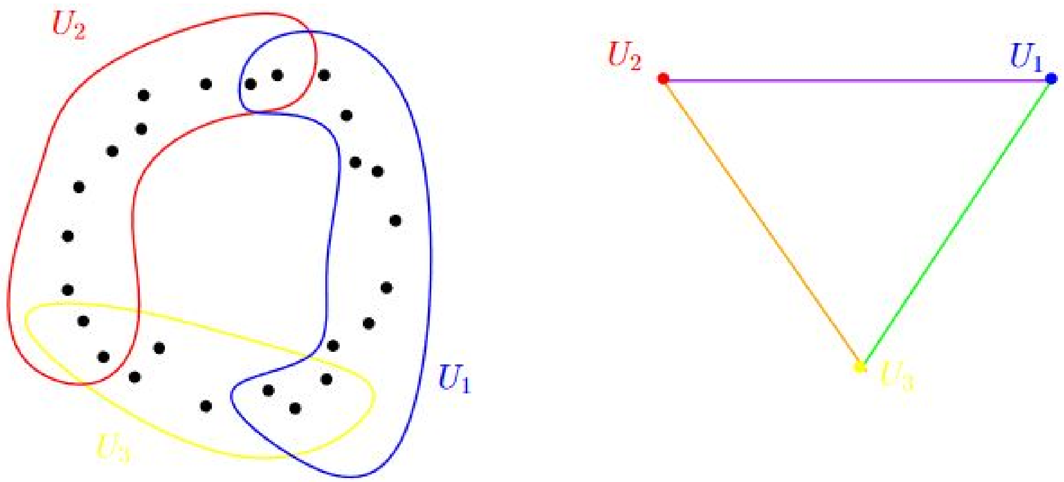

Figure 5 – Generalized nerve construction. [2]

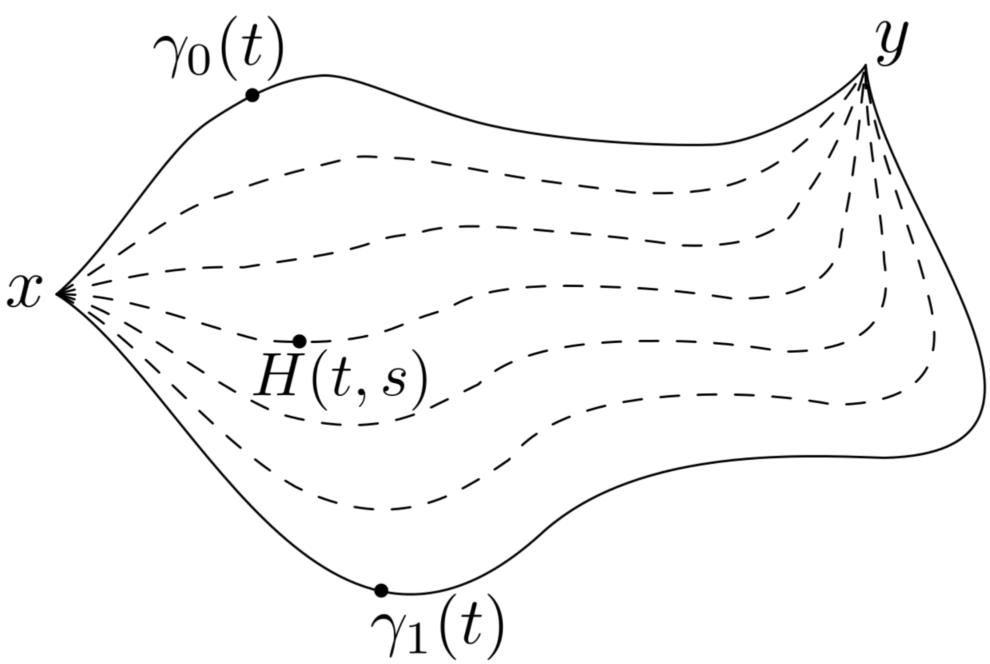

Figure 6 – Two continuous functions $\gamma_0(t), \gamma_1(t)$ that can be continuously deformed into eachother using homotopy $H(t,s)$. [6]

such that $g \circ f$ is **homotopic** to the identity map id_X and $f \circ g$ is **homotopic** to id_Y.

One can think of such **equivalence** of spaces as a continuous squishing and stretching of one into another. This implies that **homeomorphic** spaces are always **homotopy equivalent**, while the converse doesn't apply (**homotopy equivalence** is a weaker version of equivalence). (*Figures 7, 8)*

In TDA, the **homotopy equivalence** of a space to a point in itself is especially important. We call such space **contractible**.

Definition: A topological space $X \neq \emptyset$ is **contractible** if there exists a continuous map $H : X \times [0, 1] \to X$ and a fixed point $p \in X$ such that

$$\forall x \in X:\ H(x, 0) = x\ \wedge\ \forall x \in X:\ H(x, 1) = p$$

Intuitively, such space can be "squished" into a single point. It is implied that **contractible** spaces must not contain "holes". (Figure 9)

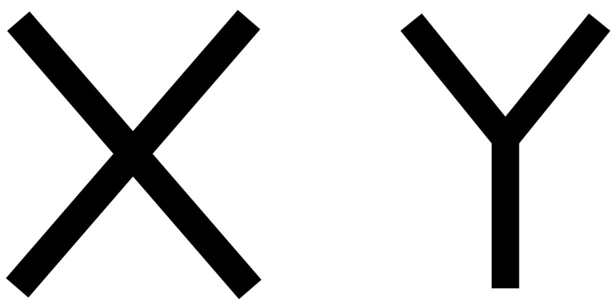

Figure 7 – Two **homotopy equivalent** spaces.

Figure 8 – Three spaces that are not pairwise **homotopy equivalent**. [3]

Lemma: Let V be a topological vector space over R; let $X \subset V$ be a **convex subset**. Then X is **contractible**.

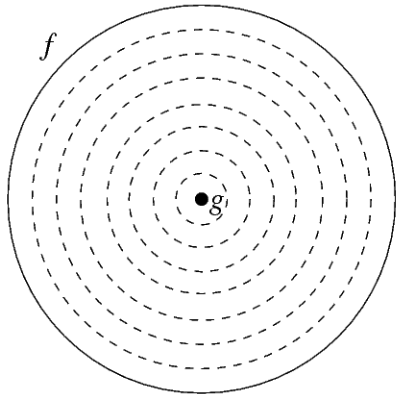

Figure 9 – **Contractible** disc in R^2. [7]

Proof: Let $x_0 \in X$. Define continuous map $H : X \times [0,1] \to X$ by:

$$H(x,t) = tx_0 + (1-t)x$$

This yields a **homotopy** between the identity map id_X and the constant map $c_{x_0}(x) = x_0$. Due to the assumption of **convexity** for X, H takes values in X. H is a continuous function, since it is polynomial seperately in x,t and:

$$H(x,0) = id_X \ \wedge \ H(x,1) = c_{x_0}(x)$$

This proves that H is a homotopy between id_X and $c_{x_0}(x)$. Therefore X is **contractible**. □

2.4 The Nerve Theorem [2]

Theorem: Let $D = (U_i)_{i \in I}$ be a **cover** of a topological space X by open sets U_i such that the intersection of any subcollection of the U_i's is either empty or contractible.
Then X and the **nerve** $C(D)$ of X are **homotopy equivalent**.

Now, since our data is (typically) embedded in R^n, we can chose our cover D to consist of convex (and therefore contractible) subsets of R^n. It follows, that the nerve $C(D)$ and the union of cover elements are **homotopy equivalent**.

This can be applied to the Čech complex of a dataset X, by constructing our **cover** D out of the balls $B(x, \alpha)$ for $\forall x \in X$. Since the balls are **convex** subsets of R^n, the Čech complex is **homotopy equivalent** to the union of the balls.

In conclusion, we can encode the topology of a continuous space into abstract combinatorial structures for efficient computation.

3 Simplicial Approximation Theorem

The goal of topological data analysis is to extract information from a dataset containing (ideally) vast amounts of samples (e.g a set of images). With knowledge of **simplicial complexes** and the **nerve theorem**, one can now represent a sample (e.g an image) as a combinatorial structure, while preserving their topology. The question arises, how one can guarantee that similar samples give rise to similar **complexes**.

For this reason, we introduce an array of concepts that enable us to understand the very important **simplicial approximation theorem**. One of its use cases is the derivation of topological similarities between **complexes**.
Analogous to general topology and its introduction of similarities between spaces, we begin by defining maps between **simplicial complexes**.

3.1 Simplicial Maps [1]

Definition: Let K and L be complexes and let $f : K^{0)} \to L^{(0)}$ be a map. Suppose that whenever the vertices $v_0, ..., v_n$ of K span a simplex of K, the points $f(v_0), ..., f(v_n)$ are vertices of a simplex of L. Then f can be extended to a continuous map $g : |K| \to |L|$ such that

$$x = \sum_{i=0}^{n} t_i v_i \implies g(x) = \sum_{i=0}^{n} t_i f(v_i)$$

We call g the **simplicial map** induced by the **vertex map** f. The composite of two simplicial maps is also simplicial.

Example: Given **simplicial complexes** K, L with a **vertex map** f (*Figure 10*), we extend f to the **simplicial map** g. Intuitively, a sense of continuity arises through the **barycentric coordinates**. Just as x lies right in the middle between **vertices** a and b, its image $g(x)$ lies right in the middle between $f(a)$ and $f(b)$.

Definition: Let $h : |K| \to |L|$ be a continuous map. We say h satisfies the **star condition** with respect to K and L if for each vertex v of K, there is a vertex w of L such that

$$h(St(v, K)) \subset St(w, L)$$

Example: Let $h : |K'| \to |L|$ (*Figure 11*) satisfy the **star condition**. If one choses a vertex in K' (e.g k), its **star** (e.g the open path between l and j) has to lie in the **star** of a vertex in L (e.g $St(L', L)$).

The condition implies a first idea of two **complexes** being "similar", by describing a **map** that always puts a **star** from K into one in L. This idea also proves to be fundamental for our requested theorem, because we can use it to "approximate" continuous maps between **simplicial complexes**.

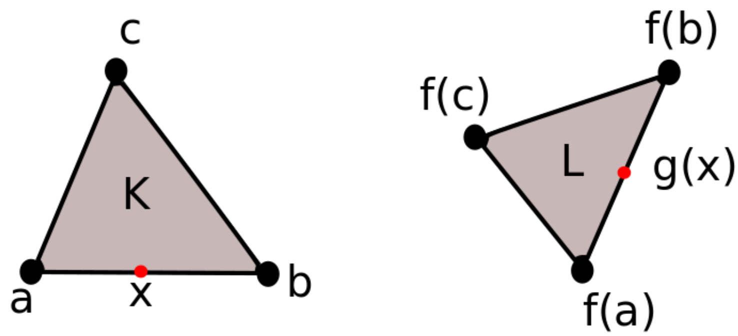

Figure 10 – Two **simplicial complexes** K, L with **vertex map** f and **simplicial map** g.

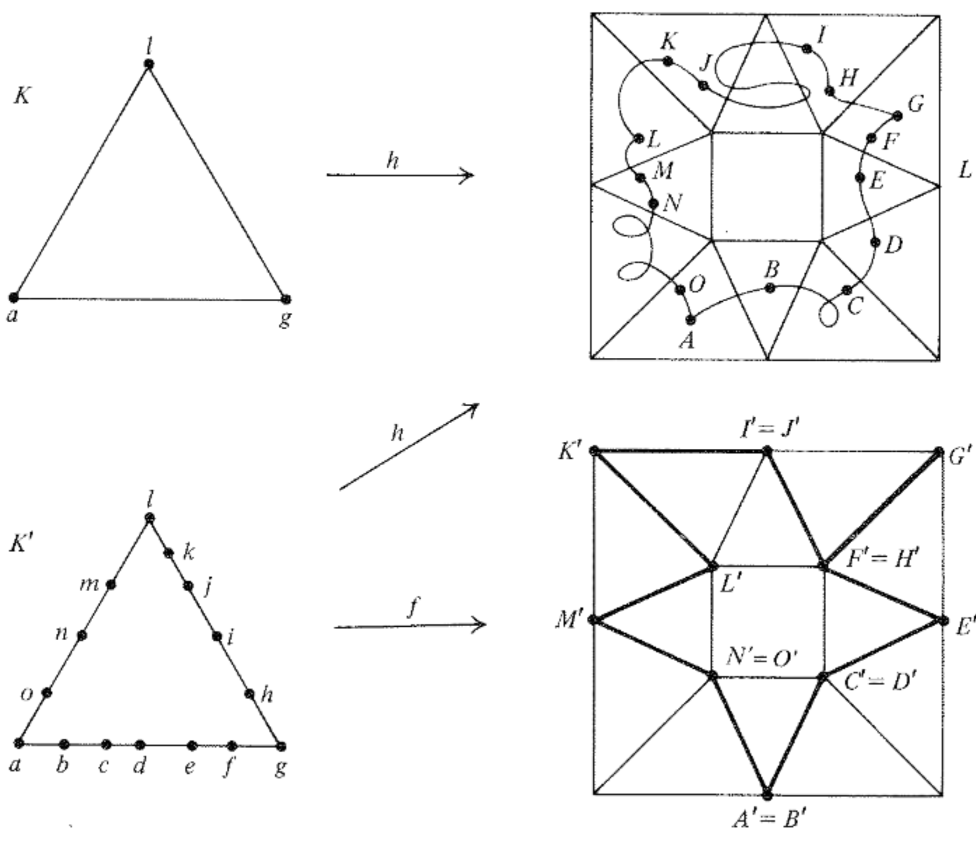

Figure 11 – Three **complexes** K, K', L with continuous function $h : |K| \to |L|$ and **simplicial approximation** $f : K' \to L$. K, K' are to be seen as collections of **1-simplices**, while L is a collection of **2-simplices**. A capitalized vertex name in L corresponds to the image of a lowercase vertex in K or K'. (e.g $h(d) = D$ and $f(d) = D'$). [1]

3.2 Simplicial Approximations [1]

Definition: Let $h : |K| \to |L|$ be a continuous map. If $f : K \to L$ is a simplicial map such that

$$\forall v \in K : \ h(St(v, K)) \subset St(f(v), L)$$

then f is called a **simplicial approximation** to h. It also follows, that f and h are **homotopic**.

Example: $f : K' \to L$ (*Figure 11*) is a **simplicial approximation** to $h : |K'| \to |L|$.

Intuitively, the map f is in some way "close" to h. We can formalize this in the following corollary.

Corollary: let $f : K \to L$ be a simplicial approximation to $h : |K| \to |L|$. Given $x \in |K|$, there is a simplex θ of L such that $h(x) \in Int(\theta)$ and $f(x) \in \theta$.

Now one might ask, if there also exists a **simplicial approximation** to $h : |K| \to |L|$ in *Figure 11*. Even though K and K' are rather similar, this can be disproven rather quickly. Since $h(St(l, K))$ does not lie in any one **star** of a vertex in L, the necessary condition is not fulfilled.
Their differences in granularity are exactly the reason that only one can be used to construct an **approximation** of h. Motivated by this, we introduce the notion of a **subdivision** of a **complex**.

3.3 Barycentric Subdivision [1]

Definition: Let K be a geometric complex in R^n. A complex K' is said to be a **subdivision** of K if:

- Each simplex of K' is contained in a simplex of K.
- Each simplex of K equals the union of *finitely* many simplices of K'.

This implies, that the union of the simplices of K' equals the union of the simplices of K. $|K|$ and $|K'|$ are also equivalent.
It follows, that the **subdivision** K'' of a **subdivision** K' is a **subdivision** of the original K.

Example: K' is a **subdivision** of K. (*Figure 11*)

While there are many ways of **subdividing** a **complex**, we are specifically interested in the **barycentric subdivision**.

Definition: Let θ be a n-simplex with the vertices $v_0, ..., v_n$. The **barycenter** $\bar{\theta}$ is defined to be the point

$$\bar{\theta} = \sum_{i=0}^{n} \frac{1}{n+1} v_i$$

It is the point of $Int(\theta)$ all of whose barycentric coordinates with respect to the vertices of θ are equal.

Example: σ is the **barycenter** of the **2-simplex** in K. (*Figure 12*)

Definition Let K be a complex. We define a sequence of subdivisions of the skeletons of K as follows: Let $L_0 = K^{(0)}$ be the 0-skeleton of K. In general, if L_p is a subdivision of the **p-skeleton** of K, let L_{p+1} be the subdivision of the $p+1$ skeleton obtained by starring L_p from the barycenters of the $p+1$ simplices of K. The union of the complexes L_p is a subdivision of K. It is called the **first barycentric subdivision** of K, and denoted $sd(K)$. Multiple consecutive **barycentric subdivisions** are denoted as $sd^n(K)$, with n being the number of subdivisions.

This construct can be used to split a simplicial complex into indefinitely smaller parts. (*Figure 12*)

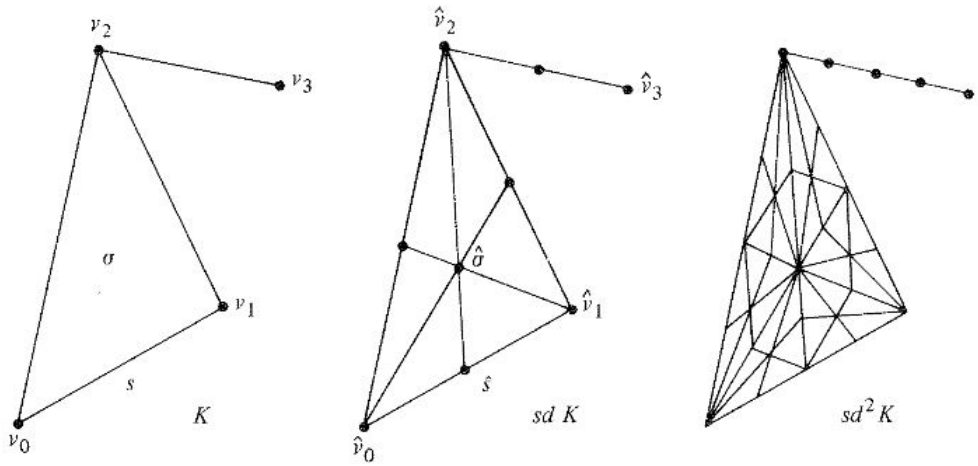

Figure 12 – A **complex** K and its **first** and **second barycentric subdivision**. [1]

Lemma: Given a finite complex K, given a metric for $|K|$, and given $\epsilon > 0$, there is a n such that each simplex of $sd^n(K)$ has diameter less than ϵ.

With all of these tools, we can now define the **simplicial approximation theorem**.

3.4 The Simplicial Approximation Theorem [1]

Theorem: Let K and L be complexes; let K be finite. Given a continuous map $h : |K| \to |L|$, there is a n such that h has a simplicial approximation $f : sd^n(K) \to$ Ł.

Proof: Cover $|K|$ by the open sets $h^{-1}(St(w, L))$, as w ranges over the vertices of L. Now given this open covering D of the compact metric space K, there is a number λ such that any set of diameter less than λ lies in one of the elements of D; such a number is called a **Lebesgue Number** for D. Its existence is guaranteed by the **Lebesgue number lemma**.
Choose n so that each simplex in $sd^n(K)$ has diameter less than $\frac{\lambda}{2}$. Then each star of a vertex in $sd^n(K)$ has diameter less than λ, so it lies in one of the sets $h^{-1}(St(w, L))$. Then $h : |K| \to |L|$ satisfies the star condition relative to $sd^n(K)$ and L, and the desired simplicial approximation exists. $\square$

Example: Given two samples A and B of a dataset X, we can construct the **Čech complex** for a fixed α on both samples. If there is a **continuous map** $h : |A| \to |B|$, we can **subdivide** A and find a **simplicial approximation** $f : sd^n(A) \to B$. This idea can be used to prove the **homotopy equivalence** of the **Čech complexes**.

References

[1] Munkres, James R. *Elements of algebraic topology.* CRC Press, 2018.

[2] arXiv:1710.04019v1 [math.ST] 11 Oct 2017

[3] Hatcher, Allen *Algebraic Topology.* Cambridge University Press, 2002.

[4] Zomorodian, Afra. "Computational topology." *Algorithms and theory of computation handbook 2.3* 2009.

[5] arXiv:1908:06856v1 [eess.SP] 9 Aug 2019

[6] Archibald, Homotopy curves, 2009, https://de.m.wikipedia.org/wiki/Datei:Homotopy_curves.svg

[7] Yonatan, Homotopie Bsp, 2005, https://commons.wikimedia.org/wiki/File:Homotopie_Bsp.png