October 2007

The Classification Theorem for Compact Surfaces and a Detour on Fractals

Jean H. Gallier
University of Pennsylvania, jean@cis.upenn.edu

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-07-20.

The Classification Theorem for Compact Surfaces and a Detour on Fractals

Abstract

In the words of Milnor himself, the classification theorem for compact surfaces is a formidable result. According to Massey, this result was obtained in the early 1920's and was the culmination of the work of many. Indeed, a rigorous proof requires, among other things, a precise definition of a surface and of orientability, a precise notion of triangulation, and a precise way of determining whether two surfaces are homeomorphic or not. This requires some notions of algebraic topology such as, fundamental groups, homology groups, and the Euler-Poincaré characteristic. Most steps of the proof are rather involved and it is easy to loose track.

The purpose of these notes is to present a fairly complete proof of the classification Theorem for compact surfaces. Other presentations are often quite informal (see the references in Chapter V) and we have tried to be more rigorous. Our main source of inspiration is the beautiful book on Riemann Surfaces by Ahlfors and Sario. However, Ahlfors and Sario's presentation is very formal and quite compact. As a result, uninitiated readers will probably have a hard time reading this book.

Our goal is to help the reader reach the top of the mountain and help him not to get lost or discouraged too early. This is not an easy task!

We provide quite a bit of topological background material and the basic facts of algebraic topology needed for understanding how the proof goes, with more than an impressionistic feeling. We hope that these notes will be helpful to readers interested in geometry, and who still believe in the rewards of serious hiking!

Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-07-20.

The Classification Theorem for Compact Surfaces And A Detour On Fractals

Jean Gallier
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@saul.cis.upenn.edu

October 22, 2007

The Classification Theorem for Compact Surfaces And A Detour On Fractals

Jean Gallier

Abstract. In the words of Milnor himself, the classification theorem for compact surfaces is a formidable result. According to Massey, this result was obtained in the early 1920's and was the culmination of the work of many. Indeed, a rigorous proof requires, among other things, a precise definition of a surface and of orientability, a precise notion of triangulation, and a precise way of determining whether two surfaces are homeomorphic or not. This requires some notions of algebraic topology such as, fundamental groups, homology groups, and the Euler-Poincaré characteristic. Most steps of the proof are rather involved and it is easy to loose track.

The purpose of these notes is to present a fairly complete proof of the classification Theorem for compact surfaces. Other presentations are often quite informal (see the references in Chapter V) and we have tried to be more rigorous. Our main source of inspiration is the beautiful book on Riemann Surfaces by Ahlfors and Sario. However, Ahlfors and Sario's presentation is very formal and quite compact. As a result, uninitiated readers will probably have a hard time reading this book.

Our goal is to help the reader reach the top of the mountain and help him not to get lost or discouraged too early. This is not an easy task!

We provide quite a bit of topological background material and the basic facts of algebraic topology needed for understanding how the proof goes, with more than an impressionistic feeling. We hope that these notes will be helpful to readers interested in geometry, and who still believe in the rewards of serious hiking!

Contents

Chapter 1

Surfaces

1.1 Introduction

Few things are as rewarding as finally stumbling upon the view of a breathtaking landscape at the turn of a path after a long hike. Similar experiences occur in mathematics, music, art, etc. When I first read about the classification of the compact surfaces, I sensed that if I prepared myself for a long hike, I could probably enjoy the same kind of exhilarating feeling.

In the words of Milnor himself, the classification theorem for compact surfaces is a formidable result. According to Massey [11], this result was obtained in the early 1920's, and was the culmination of the work of many. Indeed, a rigorous proof requires, among other things, a precise definition of a surface and of orientability, a precise notion of triangulation, and a precise way of determining whether two surfaces are homeomorphic or not. This requires some notions of algebraic topology such as, fundamental groups, homology groups, and the Euler-Poincaré characteristic. Most steps of the proof are rather involved and it is easy to loose track.

One aspect of the proof that I find particularly fascinating is the use of certain kinds of graphs (cell complexes) and of some kinds of rewrite rules on these graphs, to show that every triangulated surface is equivalent to some cell complex *in normal form* . This presents a challenge to researchers interested in rewriting, as the objects are unusual (neither terms nor graphs), and rewriting is really modulo cyclic permutations (in the case of boundaries). We hope that these notes will inspire some of the researchers in the field of rewriting to investigate these mysterious rewriting systems.

Our goal is to help the reader reach the top of the mountain (the classification theorem for compact surfaces, with or without boundaries (also called borders)), and help him not to get lost or discouraged too early. This is not an easy task! On the way, we will take a glimpse at fractals defined in terms of iterated function systems.

We provide quite a bit of topological background material and the basic facts of algebraic topology needed for understanding how the proof goes, with more than an impressionistic feeling. Having reviewed some material on complete and compact metric spaces, we indulge in a short digression on the Hausdorff distance between compact sets, and the definition of fractals in terms of iterated function systems. However, this is just a pleasant interlude, our main goal being the classification theorem for compact surfaces.

We also review abelian groups, and present a proof of the structure theorem for finitely generated abelian groups due to Pierre Samuel. Readers with a good mathematical background should proceed directly to Section 1.3, or even to Section 2.1.

We hope that these notes will be helpful to readers interested in geometry, and who still believe in the rewards of serious hiking!

Acknowledgement: I would like to thank Alexandre Kirillov for inspiring me to learn about fractals, through his excellent lectures on fractal geometry given in the Spring of 1995. Also many thanks to Chris Croke, Ron Donagi, David Harbater, Herman Gluck, and Steve Shatz, from whom I learned most of my topology and geometry. Finally, special thanks to Eugenio Calabi and Marcel Berger, for giving fascinating courses in the Fall of 1994, which changed my scientific life irrevocably (for the best!).

Basic topological notions are given in Chapter 6. In this chapter, we simply review quotient spaces.

1.2 The Quotient Topology

Ultimately, surfaces will be viewed as spaces obtained by identifying (or gluing) edges of plane polygons and to define this process rigorously, we need the concept of quotient topology. This section is intended as a review and it is far from being complete. For more details, consult Munkres [13], Massey [11, 12], Amstrong [2], or Kinsey [9].

Definition 1.2.1 Given any topological space X and any set Y, for any surjective function $f\colon X \to Y$, we define the *quotient topology on Y determined by f* (also called the *identification topology on Y determined by f*), by requiring a subset V of Y to be open if $f^{-1}(V)$ is an open set in X. Given an equivalence relation R on a topological space X, if $\pi\colon X \to X/R$ is the projection sending every $x \in X$ to its equivalence class $[x]$ in X/R, the space X/R equipped with the quotient topology determined by π is called the *quotient space of X modulo R*. Thus a set V of equivalence classes in X/R is open iff $\pi^{-1}(V)$ is open in X, which is equivalent to the fact that $\bigcup_{[x]\in V}[x]$ is open in X.

It is immediately verified that Definition 1.2.1 defines topologies, and that $f\colon X \to Y$ and $\pi\colon X \to X/R$ are continuous when Y and X/R are given these quotient topologies.

One should be careful that if X and Y are topological spaces and $f\colon X \to Y$ is a continuous surjective map, Y *does not* necessarily have the quotient topology determined by f. Indeed, it may not be true that a subset V of Y is open when $f^{-1}(V)$ is open. However, this will be true in two important cases.

Definition 1.2.2 A continuous map $f\colon X \to Y$ is an *open map* (or simply *open*) if $f(U)$ is open in Y whenever U is open in X, and similarly, $f\colon X \to Y$ is a *closed map* (or simply *closed*) if $f(F)$ is closed in Y whenever F is closed in X.

Then, Y has the quotient topology induced by the continuous surjective map f if either f is open or f is closed. Indeed, if f is open, then assuming that $f^{-1}(V)$ is open in X, we have $f(f^{-1}(V)) = V$ open in Y. Now, since $f^{-1}(Y - B) = X - f^{-1}(B)$, for any subset B of Y, a subset V of Y is open in the quotient topology iff $f^{-1}(Y - V)$ is closed in X. From this, we can deduce that if f is a closed map, then V is open in Y iff $f^{-1}(V)$ is open in X.

Among the desirable features of the quotient topology, we would like compactness, connectedness, arcwise connectedness, or the Hausdorff separation property, to be preserved. Since $f\colon X \to Y$ and $\pi\colon X \to X/R$ are continuous, by Proposition 6.3.4, its version for arcwise connectedness, and Proposition 6.4.8, compactness, connectedness, and arcwise connectedness, are indeed preserved. Unfortunately, the Hausdorff separation property is not necessarily preserved. Nevertheless, it is preserved in some special important cases.

Proposition 1.2.3 *Let X and Y be topological spaces, $f\colon X \to Y$ a continuous surjective map, and assume that X is compact, and that Y has the quotient topology determined by f. Then Y is Hausdorff iff f is a closed map.*

Proof. If Y is Hausdorff, because X is compact and f is continuous, since every closed set F in X is compact, by Proposition 6.4.8, $f(F)$ is compact, and since Y is Hausdorff, $f(F)$ is closed, and f is a closed map. For the converse, we use Proposition 6.4.5. Since X is Hausdorff, every set $\{a\}$ consisting of a single element $a \in X$ is closed, and since f is a closed map, $\{f(a)\}$ is also closed in Y. Since f is surjective, every set $\{b\}$ consisting of a single element $b \in Y$ is closed. If $b_1, b_2 \in Y$ and $b_1 \neq b_2$, since $\{b_1\}$ and $\{b_2\}$ are closed in Y and f is continuous, the sets $f^{-1}(b_1)$ and $f^{-1}(b_2)$ are closed in X, and thus compact, and by Proposition 6.4.5, there exists some disjoint open sets U_1 and U_2 such that $f^{-1}(b_1) \subseteq U_1$ and $f^{-1}(b_2) \subseteq U_2$. Since f is closed, the sets $f(X - U_1)$ and $f(X - U_2)$ are closed, and thus the sets

$$\begin{aligned} V_1 &= Y - f(X - U_1) \\ V_2 &= Y - f(X - U_2) \end{aligned}$$

are open, and it is immediately verified that $V_1 \cap V_2 = \emptyset$, $b_1 \in V_1$, and $b_2 \in V_2$. This proves that Y is Hausdorff. □

Remark: It is easily shown that another equivalent condition for Y being Hausdorff is that

$$\{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\}$$

is closed in $X \times X$.

Another useful proposition deals with subspaces and the quotient topology.

Proposition 1.2.4 *Let X and Y be topological spaces, $f\colon X \to Y$ a continuous surjective map, and assume that Y has the quotient topology determined by f. If A is a closed subset (resp. open subset) of X and f is a closed map (resp. is an open map), then $B = f(A)$ has the same topology considered as a subspace of Y, or as having the quotient topology induced by f.*

Proof. Assume that A is open and that f is an open map. Assuming that $B = f(A)$ has the subspace topology, which means that the open sets of B are the sets of the form $B \cap U$, where $U \subseteq Y$ is an open set of Y, because f is open, B is open in Y, and it is immediate that $f|A\colon A \to B$ is an open map. But then, by a previous observation, B has the quotient topology induced by f. The proof when A is closed and f is a closed map is similar. □

We now define (abstract) surfaces.

1.3 Surfaces: A Formal Definition

Intuitively, what distinguishes a surface from an arbitrary topological space, is that a surface has the property that for every point on the surface, there is a small neighborhood that looks like a little planar region. More precisely, a surface is a topological space that can be covered by open sets that can be mapped homeomorphically onto open sets of the plane. Given such an open set U on the surface S, there is an open set Ω of the plane $\mathbb{R}^2$, and a homeomorphism $\varphi\colon U \to \Omega$. The pair (U, φ) is usually called a *coordinate system*, or *chart*, of S, and $\varphi^{-1}\colon \Omega \to U$ is called a *parameterization* of U. We can think of the maps $\varphi\colon U \to \Omega$ as defining small planar maps of small regions on S, similar to geographical maps. This idea can be extended to higher dimensions, and leads to the notion of a topological manifold.

Definition 1.3.1 For any $m \geq 1$, a *(topological) m-manifold* is a second-countable, topological Hausdorff space M, together with an open cover $(U_i)_{i\in I}$ and a family $(\varphi_i)_{i\in I}$ of homeomorphisms $\varphi_i\colon U_i \to \Omega_i$, where each Ω_i is some open subset of $\mathbb{R}^m$. Each pair (U_i, φ_i) is called a *coordinate system*, or *chart* (or local chart) of M, each homeomorphism $\varphi_i\colon U_i \to \Omega_i$ is called a *coordinate map*, and its inverse $\varphi_i^{-1}\colon \Omega_i \to U_i$ is called a *parameterization* of U_i. For any point $p \in M$, for any coordinate system (U, φ) with $\varphi\colon U \to \Omega$, if $p \in U$, we say that (Ω, φ^{-1}) is a *parameterization of M at p*. The family $(U_i, \varphi_i)_{i\in I}$ is often called an *atlas* for M. A *(topological) surface* is a connected 2-manifold.

Remarks:

(1) The terminology is not universally agreed upon. For example, some authors (including Fulton [7]) call the maps $\varphi_i^{-1}\colon \Omega_i \to U_i$ charts! Always check the direction of the homeomorphisms involved in the definition of a manifold (from M to $\mathbb{R}^m$, or the other way around).

(2) Some authors define a surface as a 2-manifold, i.e., they do not require a surface to be connected. Following Ahlfors and Sario [1], we find it more convenient to assume that surfaces are connected.

(3) According to Definition 1.3.1, m-manifolds (or surfaces) do not have any differential structure. This is usually emphasized by calling such objects *topological* m-manifolds (or *topological* surfaces). Rather than being pedantic, until specified otherwise, we will simply use the term m-manifold (or surface). A 1-manifold is also called a curve.

One may wonder whether it is possible that a topological manifold M be both an m-manifold and an n-manifold for $m \neq n$. For example, could a surface also be a curve? Fortunately, for connected manifolds, this is not the case. By a deep theorem of Brouwer (the invariance of dimension theorem), it can be shown that a connected m-manifold is not an n-manifold for $n \neq m$.

Some readers many find the definition of a surface quite abstract. Indeed, the definition does not assume that a surface is a subspace of any given ambient space, say $\mathbb{R}^n$, for some n. Perhaps, such surfaces should be called "abstract surfaces". In fact, it can be shown that every surface can be embedded in $\mathbb{R}^4$, which is somewhat disturbing, since $\mathbb{R}^4$ is hard to visualize! Fortunately, orientable surfaces can be embedded in $\mathbb{R}^3$. However, it is not necessary to use these embeddings to understand the topological structure of surfaces. In fact, when it comes to higher-order manifolds, (m-manifolds for $m \geq 3$), and such manifolds do arise naturally in mechanics, robotics and computer vision, even though it can be shown that an m-manifold can be embedded in $\mathbb{R}^{2m}$ (a hard theorem due to Whitney), this usually does not help in understanding its structure. In the case $m = 1$ (curves), it is not too difficult to prove that a 1-manifold is homeomorphic to either a circle or an open line segment (interval).

Since an m-manifold M has an open cover of sets homeomorphic with open sets of $\mathbb{R}^m$, an m-manifold is locally arcwise connected and locally compact. By Theorem 6.3.14, the connected components of an m-manifold are arcwise connected, and in particular, a surface is arcwise connected.

An open subset U on a surface S is called a *Jordan region* if its closure $\overline{U}$ can be mapped homeomorphically onto a closed disk of $\mathbb{R}^2$, in such a way that U is mapped onto the open disk, and thus, that the boundary of U is mapped homeomorphically onto the circle, the boundary of the open disk. This means that the boundary of U is a Jordan curve. Since every point in an open set of the plane $\mathbb{R}^2$ is the center of a closed disk contained in that open set, we note that every surface has an open cover by Jordan regions.

Triangulations are a fundamental tool to obtain a deep understanding of the topology of surfaces. Roughly speaking, a triangulation of a surface is a way of cutting up the surface into triangular regions, such that these triangles are the images of triangles in the plane, and the edges of these planar triangles form a graph with certain properties. To formulate this notion precisely, we need to define simplices and simplicial complexes. This can be done in the context of any affine space.

Chapter 2

Simplices, Complexes, and Triangulations

2.1 Simplices and Complexes

A simplex is just the convex hull of a finite number of affinely independent points, but we also need to define faces, the boundary, and the interior, of a simplex.

Definition 2.1.1 Let $\mathcal{E}$ be any normed affine space. Given any $n+1$ affinely independent points $a_0, \ldots, a_n$ in $\mathcal{E}$, the *n-simplex (or simplex) σ defined by* $a_0, \ldots, a_n$ is the convex hull of the points $a_0, \ldots, a_n$, that is, the set of all convex combinations $\lambda_0 a_0 + \cdots + \lambda_n a_n$, where $\lambda_0 + \cdots + \lambda_n = 1$, and $\lambda_i \geq 0$ for all i, $0 \leq i \leq n$. We call n the *dimension* of the n-simplex σ, and the points $a_0, \ldots, a_n$ are the *vertices* of σ. Given any subset $\{a_{i_0}, \ldots, a_{i_k}\}$ of $\{a_0, \ldots, a_n\}$ (where $0 \leq k \leq n$), the k-simplex generated by $a_{i_0}, \ldots, a_{i_k}$ is called a *face* of σ. A face s of σ is a *proper face* if $s \neq \sigma$ (we agree that the empty set is a face of any simplex). For any vertex a_i, the face generated by $a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$ (i.e., omitting a_i) is called the *face opposite* a_i. Every face which is a $(n-1)$-simplex is called a *boundary face*. The union of the boundary faces is the *boundary of* σ, denoted as $\partial\sigma$, and the complement of $\partial\sigma$ in σ is the *interior* $\mathrm{Int}\,\sigma = \sigma - \partial\sigma$ of σ. The interior $\mathrm{Int}\,\sigma$ of σ is sometimes called an *open simplex*.

It should be noted that for a 0-simplex consisting of a single point $\{a_0\}$, $\partial\{a_0\} = \emptyset$, and $\mathrm{Int}\,\{a_0\} = \{a_0\}$. Of course, a 0-simplex is a single point, a 1-simplex is the line segment (a_0, a_1), a 2-simplex is a triangle (a_0, a_1, a_2) (with its interior), and a 3-simplex is a tetrahedron (a_0, a_1, a_2, a_3) (with its interior), as illustrated in Figure 2.1.

We now state a number of properties of simplices, whose proofs are left as an exercise. Clearly, a point x belongs to the boundary $\partial\sigma$ of σ iff at least one of its barycentric coordinates $(\lambda_0, \ldots, \lambda_n)$ is zero, and a point x belongs to the interior $\mathrm{Int}\,\sigma$ of σ iff all of its barycentric coordinates $(\lambda_0, \ldots, \lambda_n)$ are positive, i.e., $\lambda_i > 0$ for all $i, 0 \leq i \leq n$. Then, for every $x \in \sigma$, there is a unique face s such that $x \in \mathrm{Int}\,s$, the face generated by those points a_i for which $\lambda_i > 0$, where $(\lambda_0, \ldots, \lambda_n)$ are the barycentric coordinates of x.

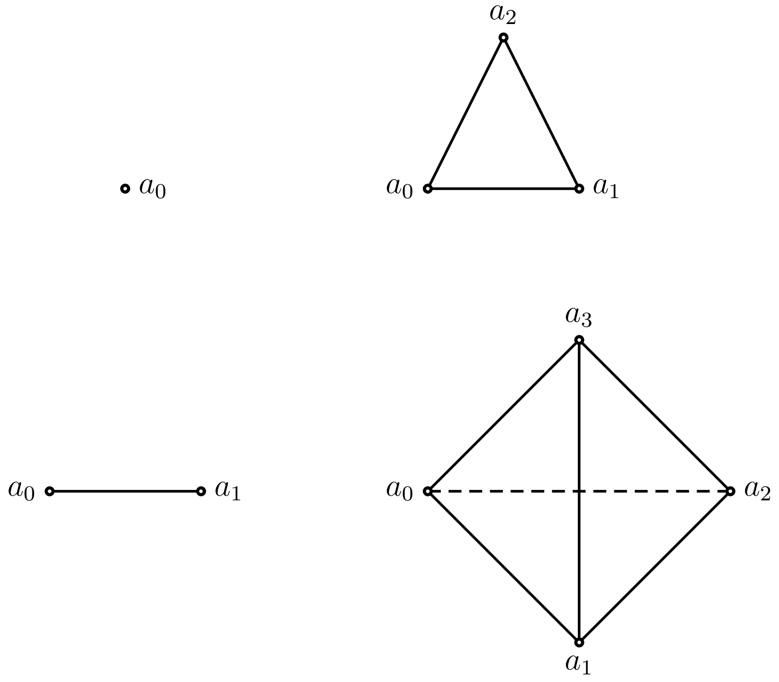

Figure 2.1: Examples of simplices

A simplex σ is convex, arcwise connected, compact, and closed. The interior $\operatorname{Int} \sigma$ of a simplex is convex, arwise connected, open, and σ is the closure of $\operatorname{Int} \sigma$.

For the last property, we recall the following definitions. The *unit n-ball B^n* is the set of points in $\mathbb{A}^n$ such that $x_1^2 + \cdots + x_n^2 \leq 1$. The *unit n-sphere S^{n-1}* is the set of points in $\mathbb{A}^n$ such that $x_1^2 + \cdots + x_n^2 = 1$. Given a point $a \in \mathbb{A}^n$ and a nonnull vector $u \in \mathbb{R}^n$, the set of all points $\{a + \lambda u \mid \lambda \geq 0\}$ is called a *ray emanating from a*. Then, every n-simplex is homeomorphic to the unit ball B^n, in such a way that its boundary $\partial\sigma$ is homeomorphic to the n-sphere S^{n-1}.

We will prove a slightly more general result about convex sets, but first, we need a simple proposition.

Proposition 2.1.2 *Given a normed affine space $\mathcal{E}$, for any nonempty convex set C, the topological closure $\overline{C}$ of C is also convex. Furthermore, if C is bounded, then $\overline{C}$ is also bounded.*

Proof. First, we show the following simple inequality. For any four points $a, b, a', b' \in \mathcal{E}$, for any $\epsilon > 0$, for any λ such that $0 \leq \lambda \leq 1$, letting $c = (1-\lambda)a + \lambda b$ and $c' = (1-\lambda)a' + \lambda b'$, if $\|\mathbf{aa'}\| \leq \epsilon$ and $\|\mathbf{bb'}\| \leq \epsilon$, then $\|\mathbf{cc'}\| \leq \epsilon$.

This is because

$$\mathbf{cc'} = (1-\lambda)\mathbf{aa'} + \lambda\mathbf{bb'},$$

and thus

$$\|\mathbf{cc'}\| \leq (1-\lambda)\|\mathbf{aa'}\| + \lambda\|\mathbf{bb'}\| \leq (1-\lambda)\epsilon + \lambda\epsilon = \epsilon.$$

Now, if $a, b \in \overline{C}$, by the definition of closure, for every $\epsilon > 0$, the open ball $B_0(a, \epsilon/2)$ must intersect C in some point a', the open ball $B_0(b, \epsilon/2)$ must intersect C in some point b', and by the above inequality, $c' = (1-\lambda)a' + \lambda b'$ belongs to the open ball $B_0(c, \epsilon)$. Since C is convex, $c' = (1-\lambda)a' + \lambda b'$ belongs to C, and $c' = (1-\lambda)a' + \lambda b'$ also belongs to the open ball $B_o(c, \epsilon)$, which shows that for every $\epsilon > 0$, the open ball $B_0(c, \epsilon)$ intersects C, which means that $c \in \overline{C}$, and thus that $\overline{C}$ is convex. Finally, if C is contained in some ball of radius δ, by the previous discussion, it is clear that $\overline{C}$ is contained in a ball of radius $\delta + \epsilon$, for any $\epsilon > 0$. □

The following proposition shows that topologically, closed bounded convex sets in $\mathbb{A}^n$ are equivalent to closed balls. We will need this proposition in dealing with triangulations.

Proposition 2.1.3 *If C is any nonempty bounded and convex open set in $\mathbb{A}^n$, for any point $a \in C$, any ray emanating from a intersects $\partial C = \overline{C} - C$ in exactly one point. Furthermore, there is a homeomorphism of $\overline{C}$ onto the (closed) unit ball B^n, which maps ∂C onto the n-sphere S^{n-1}.*

Proof. Since C is convex and bounded, by Proposition 2.1.2, $\overline{C}$ is also convex and bounded. Given any ray $R = \{a + \lambda u \mid \lambda \geq 0\}$, since R is obviously convex, the set $R \cap \overline{C}$ is convex, bounded, and closed in R, which means that $R \cap \overline{C}$ is a closed segment

$$R \cap \overline{C} = \{a + \lambda u \mid 0 \leq \lambda \leq \mu\},$$

for some $\mu > 0$. Clearly, $a + \mu u \in \partial C$. If the ray R intersects ∂C in another point c, we have $c = a + \nu u$ for some $\nu > \mu$, and since $\overline{C}$ is convex, $\{a + \lambda u \mid 0 \leq \lambda \leq \nu\}$ is contained in $R \cap \overline{C}$ for $\nu > \mu$, which is absurd. Thus, every ray emanating from a intersects ∂C in a single point.

Then, the map $f\colon \mathbb{A}^n - \{a\} \to S^{n-1}$ defined such that $f(x) = \mathbf{ax}/\|\mathbf{ax}\|$ is continuous. By the first part, the restriction $f_b\colon \partial C \to S^{n-1}$ of f to ∂C is a bijection (since every point on S^{n-1} corresponds to a unique ray emanating from a). Since ∂C is a closed and bounded subset of $\mathbb{A}^n$, it is compact, and thus f_b is a homeomorphism. Consider the inverse $g\colon S^{n-1} \to \partial C$ of f_b, which is also a homeomorphism. We need to extend g to a homeomorphism between B^n and $\overline{C}$. Since B^n is compact, it is enough to extend g to a continuous bijection. This is done by defining $h\colon B^n \to \overline{C}$, such that:

$$h(u) = \begin{cases} (1 - \|u\|)a + \|u\|g(u/\|u\|) & \text{if } u \neq 0; \\ a & \text{if } u = 0. \end{cases}$$

It is clear that h is bijective and continuous for $u \neq 0$. Since S^{n-1} is compact and g is continuous on S^{n-1}, there is some $M > 0$ such that $\|\mathbf{ag(u)}\| \leq M$ for all $u \in S^{n-1}$, and if $\|u\| \leq \delta$, then $\|\mathbf{ah(u)}\| \leq \delta M$, which shows that h is also continuous for $u = 0$. □

Remark: It is useful to note that the second part of the proposition proves that if C is a bounded convex open subset of $\mathbb{A}^n$, then any homeomorphism $g\colon S^{n-1} \to \partial C$ can be extended to a homeomorphism $h\colon B^n \to \overline{C}$. By Proposition 2.1.3, we obtain the fact that if C is a bounded convex open subset of $\mathbb{A}^n$, then any homeomorphism $g\colon \partial C \to \partial C$ can be extended to a homeomorphism $h\colon \overline{C} \to \overline{C}$. We will need this fact later on (dealing with triangulations).

We now need to put simplices together to form more complex shapes. Following Ahlfors and Sario [1], we define abstract complexes and their geometric realizations. This seems easier than defining simplicial complexes directly, as for example, in Munkres [14].

Definition 2.1.4 An *abstract complex* (for short *complex*) is a pair $K = (V, \mathcal{S})$ consisting of a (finite or infinite) nonempty set V of *vertices*, together with a family $\mathcal{S}$ of finite subsets of V called *abstract simplices* (for short *simplices*), and satisfying the following conditions:

(A1) Every $x \in V$ belongs to at least one and at most a finite number of simplices in $\mathcal{S}$.

(A2) Every subset of a simplex $\sigma \in \mathcal{S}$ is also a simplex in $\mathcal{S}$.

If $\sigma \in \mathcal{S}$ is a nonempty simplex of $n+1$ vertices, then its dimension is n, and it is called an *n-simplex*. A 0-simplex $\{x\}$ is identified with the vertex $x \in V$. The *dimension of an abstract complex* is the maximum dimension of its simplices if finite, and ∞ otherwise.

We will use the abbreviation complex for abstract complex, and simplex for abstract simplex. Also, given a simplex $s \in \mathcal{S}$, we will often use the abuse of notation $s \in K$. The purpose of condition (A1) is to insure that the geometric realization of a complex is locally compact. Recall that given any set I, the real vector space $\mathbb{R}^{(I)}$ freely generated by I is defined as the subset of the cartesian product $\mathbb{R}^I$ consisting of families $(\lambda_i)_{i\in I}$ of elements of $\mathbb{R}$ with finite support (where $\mathbb{R}^I$ denotes the set of all functions from I to $\mathbb{R}$). Then, every abstract complex $(V, \mathcal{S})$ has a geometric realization as a topological subspace of the normed vector space $\mathbb{R}^{(V)}$. Obviously, $\mathbb{R}^{(I)}$ can be viewed as a normed affine space (under the norm $\|x\| = \max_{i\in I}\{x_i\}$) denoted as $\mathbb{A}^{(I)}$.

Definition 2.1.5 Given an abstract complex $K = (V, \mathcal{S})$, its *geometric realization* (also called the *polytope of* $K = (V, \mathcal{S})$) is the subspace K_g of $\mathbb{A}^{(V)}$ defined as follows: K_g is the set of all families $\lambda = (\lambda_a)_{a\in V}$ with finite support, such that:

(B1) $\lambda_a \geq 0$, for all $a \in V$;

(B2) The set $\{a \in V \mid \lambda_a > 0\}$ is a simplex in $\mathcal{S}$;

(B3) $\sum_{a\in V} \lambda_a = 1$.

For every simplex $s \in \mathcal{S}$, we obtain a subset s_g of K_g by considering those families $\lambda = (\lambda_a)_{a\in V}$ in K_g such that $\lambda_a = 0$ for all $a \notin s$. Then, by (B2), we note that

$$K_g = \bigcup_{s\in\mathcal{S}} s_g.$$

It is also clear that for every n-simplex s, its geometric realization s_g can be identified with an n-simplex in $\mathbb{A}^n$.

Given a vertex $a \in V$, we define the *star of* a, denoted as $\mathrm{St}\, a$, as the finite union of the interiors $\overset{\circ}{s}_g$ of the geometric simplices s_g such that $a \in s$. Clearly, $a \in \mathrm{St}\, a$. The *closed star of* a, denoted as $\overline{\mathrm{St}}\, a$, is the finite union of the geometric simplices s_g such that $a \in s$.

We define a topology on K_g by defining a subset F of K_g to be closed if $F \cap s_g$ is closed in s_g for all $s \in \mathcal{S}$. It is immediately verified that the axioms of a topological space are indeed verified. Actually, we can find a nice basis for this topology, as shown in the next proposition.

Proposition 2.1.6 *The family of subsets* U *of* K_g *such that* $U \cap s_g = \emptyset$ *for all by finitely many* $s \in \mathcal{S}$, *and such that* $U \cap s_g$ *is open in* s_g *when* $U \cap s_g \neq \emptyset$, *forms a basis of open sets for the topology of* K_g. *For any* $a \in V$, *the star* $\mathrm{St}\, a$ *of* a *is open, the closed star* $\overline{\mathrm{St}}\, a$ *is the closure of* $\mathrm{St}\, a$ *and is compact, and both* $\mathrm{St}\, a$ *and* $\overline{\mathrm{St}}\, a$ *are arcwise connected. The space* K_g *is locally compact, locally arcwise connected, and Hausdorff.*

Proof. To see that a set U as defined above is open, consider the complement $F = K_g - U$ of U. We need to show that $F \cap s_g$ is closed in s_g for all $s \in \mathcal{S}$. But $F \cap s_g = (K_g - U) \cap s_g = s_g - U$, and if $s_g \cap U \neq \emptyset$, then $U \cap s_g$ is open in s_g, and thus $s_g - U$ is closed in s_g. Next, given any open subset V of K_g, since by $(A1)$, every $a \in V$ belongs to finitely many simplices $s \in \mathcal{S}$, letting U_a be the union of the interiors of the finitely many s_g such that $a \in s$, it is clear that U_a is open in K_g, and that V is the union of the open sets of the form $U_a \cap V$, which shows that the sets U of the proposition form a basis of the topology of K_g. For every $a \in V$, the star $\mathrm{St}\, a$ of a has a nonempty intersection with only finitely many simplices s_g, and $\mathrm{St}\, a \cap s_g$ is the interior of s_g (in s_g), which is open in s_g, and $\mathrm{St}\, a$ is open. That $\overline{\mathrm{St}}\, a$ is the closure of $\mathrm{St}\, a$ is obvious, and since each simplex s_g is compact, and $\overline{\mathrm{St}}\, a$ is a finite union of compact simplices, it is compact. Thus, K_g is locally compact. Since s_g is arcwise connected, for every open set U in the basis, if $U \cap s_g \neq \emptyset$, $U \cap s_g$ is an open set in s_g that contains some arcwise connected set V_s containing a, and the union of these arcwise connected sets V_s is arcwise connected, and clearly an open set of K_g. Thus, K_g is locally arcwise connected. It is also immediate that $\mathrm{St}\, a$ and $\overline{\mathrm{St}}\, a$ are arcwise connected. Let $a, b \in K_g$, and assume that $a \neq b$. If $a, b \in s_g$ for some $s \in \mathcal{S}$, since s_g is Hausdorff, there are disjoint open sets $U, V \subseteq s_g$ such that $a \in U$ and $b \in V$. If a and b do not belong to the same simplex, then $\mathrm{St}\, a$ and $\mathrm{St}\, b$ are disjoint open sets such that $a \in \mathrm{St}\, a$ and $b \in \mathrm{St}\, b$. $\square$

We also note that for any two simplices s_1, s_2 of $\mathcal{S}$, we have

$$(s_1 \cap s_2)_g = (s_1)_g \cap (s_2)_g.$$

We say that a complex $K = (V, \mathcal{S})$ is connected if it is not the union of two complexes $(V_1, \mathcal{S}_1)$ and $(V_2, \mathcal{S}_2)$, where $V = V_1 \cup V_2$ with V_1 and V_2 disjoint, and $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ with $\mathcal{S}_1$ and $\mathcal{S}_2$ disjoint. The next proposition shows that a connected complex contains countably many simplices. This is an important fact, since it implies that if a surface can be triangulated, then its topology must be second-countable.

Proposition 2.1.7 *If $K = (V, \mathcal{S})$ is a connected complex, then $\mathcal{S}$ and V are countable.*

Proof. The proof is very similar to that of the second part of Theorem 6.3.14. The trick consists in defining the right notion of arcwise connectedness. We say that two vertices $a, b \in V$ are *path-connected, or that there is a path from a to b* if there is a sequence $(x_0, \ldots, x_n)$ of vertices $x_i \in V$, such that $x_0 = a$, $x_n = b$, and $\{x_i, x_{i+1}\}$, is a simplex in $\mathcal{S}$, for all $i, 0 \leq i \leq n-1$. Observe that every simplex $s \in \mathcal{S}$ is path-connected. Then, the proof consists in showing that if $(V, \mathcal{S})$ is a connected complex, then it is path-connected. Fix any vertex $a \in V$, and let V_a be the set of all vertices that are path-connected to a. We claim that for any simplex $s \in \mathcal{S}$, if $s \cap V_a \neq \emptyset$, then $s \subseteq V_a$, which shows that if $\mathcal{S}_a$ is the subset of $\mathcal{S}$ consisting of all simplices having some vertex in V_a, then $(V_a, \mathcal{S}_a)$ is a complex. Indeed, if $b \in s \cap V_a$, there is a path from a to b. For any $c \in s$, since b and c are path-connected, then there is a path from a to c, and $c \in V_a$, which shows that $s \subseteq V_a$. A similar reasoning applies to the complement $V - V_a$ of V_a, and we obtain a complex $(V - V_a, \mathcal{S} - \mathcal{S}_a)$. But $(V_a, \mathcal{S}_a)$ and $(V - V_a, \mathcal{S} - \mathcal{S}_a)$ are disjoint complexes, contradicting the fact that $(V, \mathcal{S})$ is connected. Then, since every simplex $s \in \mathcal{S}$ is finite and every path is finite, the number of path from a is countable, and because $(V, \mathcal{S})$ is path-connected, there are at most countably many vertices in V and at most countably many simplices $s \in \mathcal{S}$. □

2.2 Triangulations

We now return to surfaces and define the notion of triangulation. Triangulations are special kinds of complexes of dimension 2, which means that the simplices involved are points, line segments, and triangles.

Definition 2.2.1 Given a surface M, a *triangulation of* M is a pair (K, σ) consisting of a 2-dimensional complex $K = (V, \mathcal{S})$ and of a map $\sigma\colon \mathcal{S} \to 2^M$ assigning a closed subset $\sigma(s)$ of M to every simplex $s \in \mathcal{S}$, satisfying the following conditions:

(C1) $\sigma(s_1 \cap s_2) = \sigma(s_1) \cap \sigma(s_2)$, for all $s_1, s_2 \in \mathcal{S}$.

(C2) For every $s \in \mathcal{S}$, there is a homeomorphism φ_s from the geometric realization s_g of s to $\sigma(s)$, such that $\varphi_s(s'_g) = \sigma(s')$, for every $s' \subseteq s$.

(C3) $\bigcup_{s \in \mathcal{S}} \sigma(s) = M$, that is, the sets $\sigma(s)$ cover M.

(C4) For every point $x \in M$, there is some neighborhood of x which meets only finitely many of the $\sigma(s)$.

If (K, σ) is a triangulation of M, we also refer to the map $\sigma\colon \mathcal{S} \to 2^M$ as a triangulation of M and we also say that K is a triangulation $\sigma\colon \mathcal{S} \to 2^M$ of M. As expected, given a triangulation (K, σ) of a surface M, the geometric realization K_g of K is homeomorphic to the surface M, as shown by the following proposition.

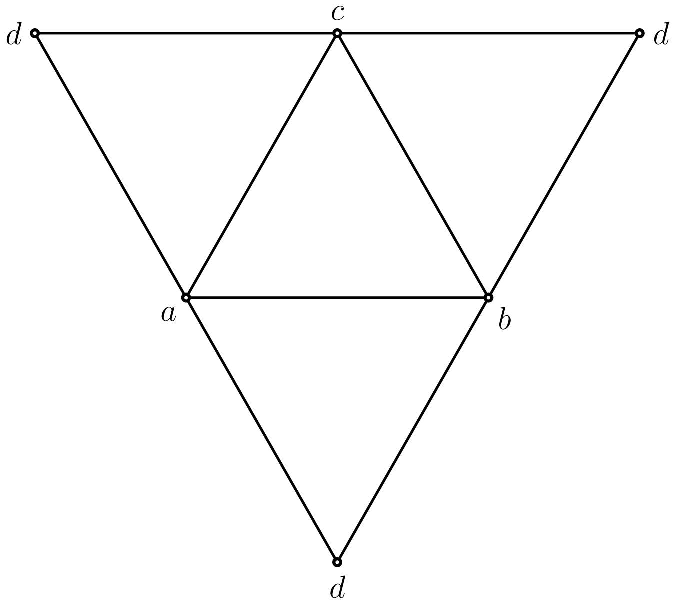

Figure 2.2: A triangulation of the sphere

Proposition 2.2.2 *Given any triangulation $\sigma\colon \mathcal{S} \to 2^M$ of a surface M, there is a homeomorphism $h\colon K_g \to M$ from the geometric realization K_g of the complex $K = (V, \mathcal{S})$ onto the surface M, such that each geometric simplex s_g is mapped onto $\sigma(s)$.*

Proof. Obviously, for every vertex $x \in V$, we let $h(x_g) = \sigma(x)$. If s is a 1-simplex, we define h on s_g using any of the homeomorphisms whose existence is asserted by (C1). Having defined h on the boundary of each 2-simplex s, we need to extend h to the entire 2-simplex s. However, by (C2), there is some homeomorphism φ from s_g to $\sigma(s)$, and if it does not agree with h on the boundary of s_g, which is a triangle, by the remark after Proposition 2.1.3, since the restriction of $\varphi^{-1} \circ h$ to the boundary of s_g is a homeomorphism, it can be extended to a homeomorphism ψ of s_g into itself, and then $\varphi \circ \psi$ is a homeomorphism of s_g onto $\sigma(s)$ that agrees with h on the boundary of s_g. This way, h is now defined on the entire K_g. Given any closed set F in M, for every simplex $s \in \mathcal{S}$,

$$h^{-1}(F) \cap s_g = h^{-1}|_{s_g}(F),$$

where $h^{-1}|_{s_g}(F)$ is the restriction of h to s_g, which is continuous by construction, and thus, $h^{-1}(F) \cap s_g$ is closed for all $s \in \mathcal{S}$, which shows that h is continuous. The map h is injective because of (C1), surjective because of (C3), and its inverse is continuous because of (C4). Thus, h is indeed a homeomorphism mapping s_g onto $\sigma(s)$. □

Figure 2.2 shows a triangulation of the *sphere*.

The geometric realization of the above triangulation is obtained by pasting together the pairs of edges labeled (a, d), (b, d), (c, d). The geometric realization is a tetrahedron.

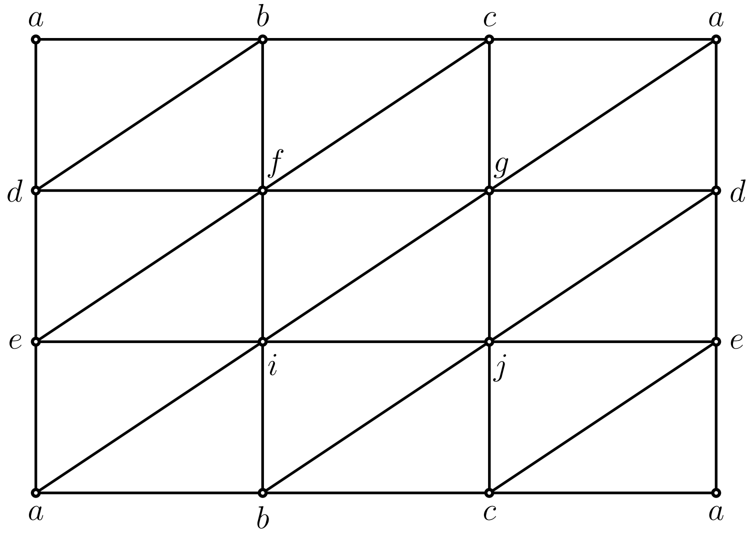

Figure 2.3: A triangulation of the torus

Figure 2.3 shows a triangulation of a surface called a *torus*.

The geometric realization of the above triangulation is obtained by pasting together the pairs of edges labeled (a, d), (d, e), (e, a), and the pairs of edges labeled (a, b), (b, c), (c, a).

Figure 2.4 shows a triangulation of a surface called the *projective plane*.

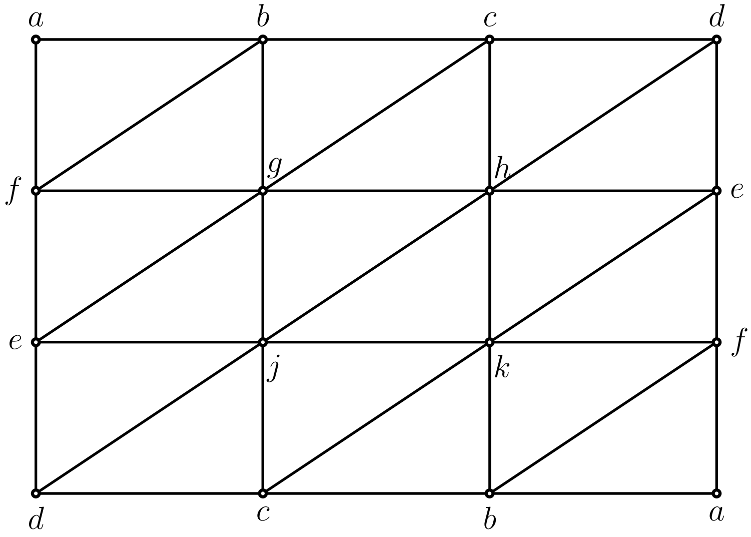

Figure 2.4: A triangulation of the projective plane

The geometric realization of the above triangulation is obtained by pasting together the pairs of edges labeled (a, f), (f, e), (e, d), and the pairs of edges labeled (a, b), (b, c), (c, d).

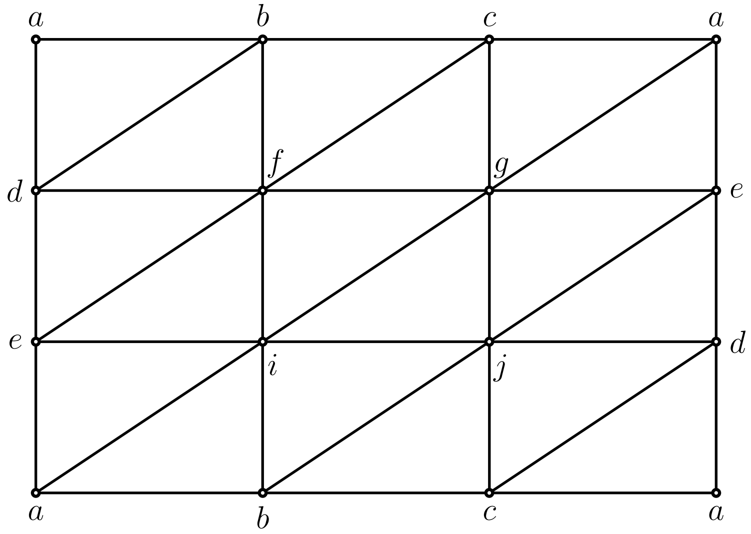

Figure 2.5: A triangulation of the Klein bottle

This time, the gluing requires a “twist”, since the the paired edges have opposite orientation. Visualizing this surface in $\mathbb{A}^3$ is actually nontrivial.

Figure 2.5 shows a triangulation of a surface called the *Klein bottle*.

The geometric realization of the above triangulation is obtained by pasting together the pairs of edges labeled (a, d), (d, e), (e, a), and the pairs of edges labeled (a, b), (b, c), (c, a). Again, some of the gluing requires a “twist”, since some paired edges have opposite orientation. Visualizing this surface in $\mathbb{A}^3$ not too difficult, but self-intersection cannnot be avoided.

We are now going to state a proposition characterizing the complexes K that correspond to triangulations of surfaces. The following notational conventions will be used: vertices (or nodes, i.e., 0-simplices) will be denoted as α, edges (1-simplices) will be denoted as a, and triangles (2-simplices) will be denoted as A. We will also denote an edge as $a = (\alpha_1\alpha_2)$, and a triangle as $A = (a_1a_2a_3)$, or as $A = (\alpha_1\alpha_2\alpha_3)$, when we are interested in its vertices. For the moment, we do not care about the order.

Proposition 2.2.3 *A 2-complex $K = (V, \mathcal{S})$ is a triangulation $\sigma\colon \mathcal{S} \to 2^M$ of a surface M such that $\sigma(s) = s_g$ for all $s \in \mathcal{S}$ iff the following properties hold:*

(D1) Every edge a is contained in exactly two triangles A.

(D2) For every vertex α, the edges a and triangles A containing α can be arranged as a cyclic sequence $a_1, A_1, a_2, A_2, \ldots, A_{m-1}, a_m, A_m$, in the sense that $a_i = A_{i-1} \cap A_i$ for all i, $2 \leq i \leq m$, and $a_1 = A_m \cap A_1$, with $m \geq 3$.

(D3) K is connected, in the sense that it cannot be written as the union of two disjoint nonempty complexes.

Proof. A proof can be found in Ahlfors and Sario [1]. The proof requires the notion of the winding number of a closed curve in the plane with respect to a point, and the concept of homotopy. □

A 2-complex K which satisfies the conditions of Proposition 2.2.3 will be called a *triangulated complex*, and its geometric realization is called a *polyhedron*. Thus, triangulated complexes are the complexes that correspond to triangulated surfaces. Actually, it can be shown that every surface admits some triangulation, and thus the class of geometric realizations of the triangulated complexes is the class of all surfaces. We now give a quick presentation of homotopy, the fundamental group, and homology groups.

Chapter 3

The Fundamental Group, Orientability

3.1 The Fundamental Group

If we want to somehow classify surfaces, we have to deal with the issue of deciding when we consider two surfaces to be equivalent. It seems reasonable to treat homeomorphic surfaces as equivalent, but this leads to the problem of deciding when two surfaces are not homeomorphic, which is a very difficult problem. One way to approach this problem is to forget some of the topological structure of a surface, and look for more algebraic objects that can be associated with a surface. For example, we can consider closed curves on a surface, and see how they can be deformed. It is also fruitful to give an algebraic structure to appropriate sets of closed curves on a surface, for example, a group structure. Two important tools for studying surfaces were invented by Poincaré, the fundamental group, and the homology groups. In this section, we take a look at the fundamental group. Roughly speaking, given a topological space E and some chosen point $a \in E$, a group $\pi(E, a)$ called the fundamental group of E based at a is associated with (E, a), and to every continuous map $f\colon (X, x) \to (Y, y)$ such that $f(x) = y$, is associated a group homomorphism $f_*\colon \pi(X, x) \to \pi(Y, y)$. Thus, certain topological questions about the space E can translated into algebraic questions about the group $\pi(E, a)$. This is the paradigm of algebraic topology. In this section, we will focus on the concepts rather than dwelve into technical details. For a thorough presentation of the fundamental group and related concepts, the reader is referred to Massey [11, 12], Munkres [13], Bredon [3], Dold [5], Fulton [7], Rotman [15]. We also recommend Sato [16] for an informal and yet very clear presentation.

The intuitive idea behind the fundamental group is that closed paths on a surface reflect some of the main topological properties of the surface. Actually, the idea applies to any topological space E. Let us choose some point a in E (a *base point*), and consider all closed curves $\gamma\colon [0, 1] \to E$ based at a, that is, such that $\gamma(0) = \gamma(1) = a$. We can compose closed curves γ_1, γ_2 based at a, and consider the inverse γ^{-1} of a closed curve, but unfortunately, the operation of composition of closed curves is not associative, and $\gamma\gamma^{-1}$ is not the identity in general. In order to obtain a group structure, we define a notion of equivalence of closed curves under continuous deformations. Actually, such a notion can be defined for any two paths with the same origin and extremity, and even for continuous maps.

Definition 3.1.1 Given any two paths $\gamma_1 \colon [0,1] \to E$ and $\gamma_2 \colon [0,1] \to E$ with the same intial point a and the same terminal point b, i.e., such that $\gamma_1(0) = \gamma_2(0) = a$, and $\gamma_1(1) = \gamma_2(1) = b$, a map $F \colon [0,1] \times [0,1] \to E$ is a *(path) homotopy* between γ_1 and γ_2 if F is continuous, and if

$$\begin{aligned} F(t,0) &= \gamma_1(t), \\ F(t,1) &= \gamma_2(t), \end{aligned}$$

for all $t \in [0,1]$, and

$$\begin{aligned} F(0,u) &= a, \\ F(1,u) &= b, \end{aligned}$$

for all $u \in [0,1]$. In this case, we say that γ_1 and γ_2 are *homotopic*, and this is denoted as $\gamma_1 \approx \gamma_2$.

Given any two continuous maps $f_1 \colon X \to Y$ and $f_2 \colon X \to Y$ between two topological spaces X and Y, a map $F \colon X \times [0,1] \to Y$ is a *homotopy* between f_1 and f_2 iff F is continuous and

$$\begin{aligned} F(t,0) &= f_1(t), \\ F(t,1) &= f_2(t), \end{aligned}$$

for all $t \in X$. We say that f_1 and f_2 are homotopic, and this is denoted as $f_1 \approx f_2$.

Intuitively, a (path) homotopy F between two paths γ_1 and γ_2 from a to b is a continuous family of paths $F(t,u)$ from a to b, giving a deformation of the path γ_1 into the path γ_2. It is easily shown that homotopy is an equivalence relation on the set of paths from a to b. A simple example of homotopy is given by reparameterizations. A continuous nondecreasing function $\tau \colon [0,1] \to [0,1]$ such that $\tau(0) = 0$ and $\tau(1) = 1$ is called a *reparameterization*. Then, given a path $\gamma \colon [0,1] \to E$, the path $\gamma \circ \tau \colon [0,1] \to E$ is homotopic to $\gamma \colon [0,1] \to E$, under the homotopy

$$(t,u) \mapsto \gamma((1-u)t + u\tau(t)).$$

As another example, any two continuous maps $f_1 \colon X \to \mathbb{A}^2$ and $f_2 \colon X \to \mathbb{A}^2$ with range the affine plane $\mathbb{A}^2$ are homotopic under the homotopy defined such that

$$F(t,u) = (1-u)f_1(t) + uf_2(t).$$

However, if we remove the origin from the plane $\mathbb{A}^2$, we can find paths γ_1 and γ_2 from $(-1,0)$ to $(1,0)$ that are not homotopic. For example, we can consider the upper half unit circle, and the lower half unit circle. The problem is that the "hole" created by the missing origin prevents continuous deformation of one path into the other. Thus, we should expect that homotopy classes of closed curves on a surface contain information about the presence or absence of "holes" in a surface.

It is easily verified that if $\gamma_1 \approx \gamma_1'$ and $\gamma_2 \approx \gamma_2'$, then $\gamma_1\gamma_2 \approx \gamma_1'\gamma_2'$, and that $\gamma_1^{-1} \approx \gamma_1'^{-1}$. Thus, it makes sense to define the composition and the inverse of homotopy classes.

Definition 3.1.2 Given any topological space E, for any choice of a point $a \in E$ (a *base point*), the *fundamental group (or Poincaré group)* $\pi(E, a)$ *at the base point* a is the set of homotopy classes of closed curves $\gamma\colon [0, 1] \to E$ such that $\gamma(0) = \gamma(1) = a$, under the multiplication operation $[\gamma_1][\gamma_2] = [\gamma_1\gamma_2]$, induced by the composition of closed paths based at a.

One actually needs to prove that the above multiplication operation is associative, has the homotopy class of the constant path equal to a as an identity, and that the inverse of the homotopy class $[\gamma]$ is the class $[\gamma^{-1}]$. The first two properties are left as an exercise, and the third property uses the homotopy

$$F(t, u) = \begin{cases} \gamma(2t) & \text{if } 0 \leq t \leq (1-u)/2; \\ \gamma(1-u) & \text{if } (1-u)/2 \leq t \leq (1+u)/2; \\ \gamma(2-2t) & \text{if } (1+u)/2 \leq t \leq 1. \end{cases}$$

As defined, the fundamental group depends on the choice of a base point. Let us now assume that E is arcwise connected (which is the case for surfaces). Let a and b be any two distinct base points. Since E is arcwise connected, there is some path α from a to b. Then, to every closed curve γ based at a corresponds a close curve $\gamma' = \alpha^{-1}\gamma\alpha$ based at b. It is easily verified that this map induces a homomorphism $\varphi\colon \pi(E, a) \to \pi(E, b)$ between the groups $\pi(E, a)$ and $\pi(E, b)$. The path α^{-1} from b to a induces a homomorphism $\psi\colon \pi(E, b) \to \pi(E, a)$ between the groups $\pi(E, b)$ and $\pi(E, a)$. Now, it is immediately verified that $\varphi \circ \psi$ and $\psi \circ \varphi$ are both the identity, which shows that the groups $\pi(E, a)$ and $\pi(E, b)$ are isomorphic.

Thus, when the space E is arcwise connected, the fundamental groups $\pi(E, a)$ and $\pi(E, b)$ are isomorphic for any two points $a, b \in E$.

Remarks:

(1) The isomorphism $\varphi\colon \pi(E, a) \to \pi(E, b)$ is not canonical, that is, it depends on the chosen path α from a to b.

(2) In general, the fundamental group $\pi(E, a)$ is not commutative.

When E is arcwise connected, we allow ourselves to refer to any of the isomorphic groups $\pi(E, a)$ as *the* fundamental group of E, and we denote any of these groups as $\pi(E)$.

The fundamental group $\pi(E, a)$ is in fact one of several homotopy groups $\pi_n(E, a)$ associated with a space E, and $\pi(E, a)$ is often denoted as $\pi_1(E, a)$. However, we won't have any use for the more general homotopy groups.

If E is an arcwise connected topological space, it may happen that some fundamental groups $\pi(E, a)$ is reduced to the trivial group $\{1\}$ consisting of the identity element. It is easy to see that this is equivalent to the fact that for any two points $a, b \in E$, any two paths from a to b are homotopic, and thus, the fundamental groups $\pi(E, a)$ are trivial for all $a \in E$. This is an important case, which motivates the following definition.

Definition 3.1.3 A topological space E is *simply-connected* if it is arcwise connected and for every $a \in E$, the fundamental group $\pi(E, a)$ is the trivial one-element group.

For example, the plane and the sphere are simply connected, but the torus is not simply connected (due to its hole).

We now show that a continuous map between topological spaces (with base points) induces a homomorphism of fundamental groups. Given two topological spaces X and Y, given a base point x in X and a base point y in Y, for any continuous map $f\colon (X, x) \to (Y, y)$ such that $f(x) = y$, we can define a map $f_*\colon \pi(X, x) \to \pi(Y, y)$ as follows:

$$f_*([\gamma]) = [f \circ \gamma],$$

for every homotopy class $[\gamma] \in \pi(X, x)$, where $\gamma\colon [0, 1] \to X$ is a closed path based at x.

It is easily verified that f_* is well defined, that is, does not depend on the choice of the closed curve γ in the homotopy class $[\gamma]$. It is also easily verified that $f_*\colon \pi(X, x) \to \pi(Y, y)$ is a homomorphism of groups. The map $f \mapsto f_*$ also has the following important two properties. For any two continuous maps $f\colon (X, x) \to (Y, y)$ and $g\colon (Y, y) \to (Z, z)$, such that $f(x) = y$ and $g(y) = z$, we have

$$(g \circ f)_* = g_* \circ f_*,$$

and if $Id\colon (X, x) \to (X, x)$ is the identity map, then $Id_*\colon \pi(X, x) \to \pi(X, x)$ is the identity homomorphism.

As a consequence, if $f\colon (X, x) \to (Y, y)$ is a homeomorphism such that $f(x) = y$, then $f_*\colon \pi(X, x) \to \pi(Y, y)$ is a group isomorphism. This gives us a way of proving that two spaces are not homeomorphic: show that for some appropriate base points $x \in X$ and $y \in Y$, the fundamental groups $\pi(X, x)$ and $\pi(Y, y)$ are not isomorphic.

In general, it is difficult to determine the fundamental group of a space. We will determine the fundamental group of $\mathbb{A}^n$ and of the punctured plane. For this, we need the concept of the winding number of a closed curve in the plane.

3.2 The Winding Number of a Closed Plane Curve

Consider a closed curve $\gamma\colon [0,1] \to \mathbb{A}^2$ in the plane, and let z_0 be a point not on γ. In what follows, it is convenient to identify the plane $\mathbb{A}^2$ with the set $\mathbb{C}$ of complex numbers. We wish to define a number $n(\gamma, z_0)$ which counts how many times the closed curve γ winds around z_0.

We claim that there is some real number $\rho > 0$ such that $|\gamma(t) - z_0| > \rho$ for all $t \in [0,1]$. If not, then for every integer $n \geq 0$, there is some $t_n \in [0,1]$ such that $|\gamma(t_n) - z_0| \leq 1/n$. Since $[0,1]$ is compact, the sequence (t_n) has some convergent subsequence (t_{n_p}) having some limit $l \in [0,1]$. But then, by continuity of γ, we have $\gamma(l) = z_0$, contradicting the fact that z_0 is not on γ. Now, again since $[0,1]$ is compact and γ is continuous, γ is actually uniformly continuous. Thus, there is some $\epsilon > 0$ such that $|\gamma(t) - \gamma(u)| \leq \rho$ for all $t, u \in [0,1]$, with $|u - t| \leq \epsilon$. Letting n be the smallest integer such that $n\epsilon > 1$, and letting $t_i = i/n$, for $0 \leq i \leq n$, we get a subdivision of $[0,1]$ into subintervals $[t_i, t_{i+1}]$, such that $|\gamma(t) - \gamma(t_i)| \leq \rho$ for all $t \in [t_i, t_{i+1}]$, with $0 \leq i \leq n-1$.

For every $i, 0 \leq i \leq n-1$, if we let

$$w_i = \frac{\gamma(t_{i+1}) - z_0}{\gamma(t_i) - z_0},$$

it is immediately verified that $|w_i - 1| < 1$, and thus, w_i has a positive real part. Thus, there is a unique angle θ_i with $-\frac{\pi}{2} < \theta_i < \frac{\pi}{2}$, such that $w_i = \lambda_i(\cos\theta_i + i\sin\theta_i)$, where $\lambda_i > 0$. Furthermore, because γ is a closed curve,

$$\prod_{i=0}^{n-1} w_i = \prod_{i=0}^{n-1} \frac{\gamma(t_{i+1}) - z_0}{\gamma(t_i) - z_0} = \frac{\gamma(t_n) - z_0}{\gamma(t_0) - z_0} = \frac{\gamma(1) - z_0}{\gamma(0) - z_0} = 1,$$

and the angle $\sum \theta_i$ is an integral multiple of 2π. Thus, for every subdivision of $[0,1]$ into intervals $[t_i, t_{i+1}]$ such that $|w_i - 1| < 1$, with $0 \leq i \leq n-1$, we define the *winding number* $n(\gamma, z_0)$*, or index, of* γ *with respect to* z_0, as

$$n(\gamma, z_0) = \frac{1}{2\pi} \sum_{i=0}^{i=n-1} \theta_i.$$

Actually, in order for $n(\gamma, z_0)$ to be well defined, we need to show that it does not depend on the subdivision of $[0,1]$ into intervals $[t_i, t_{i+1}]$ (such that $|w_i - 1| < 1$). Since any two subdivisions of $[0,1]$ into intervals $[t_i, t_{i+1}]$ can be refined into a common subdivision, it is enough to show that nothing is changed is we replace any interval $[t_i, t_{i+1}]$ by the two intervals $[t_i, \tau]$ and $[\tau, t_{i+1}]$. Now, if θ_i' and θ_i'' are the angles associated with

$$\frac{\gamma(t_{i+1}) - z_0}{\gamma(\tau) - z_0},$$

and

$$\frac{\gamma(\tau) - z_0}{\gamma(t_i) - z_0},$$

we have

$$\theta_i = \theta_i' + \theta_i'' + k2\pi,$$

where k is some integer. However, since $-\frac{\pi}{2} < \theta_i < \frac{\pi}{2}$, $-\frac{\pi}{2} < \theta_i' < \frac{\pi}{2}$, and $-\frac{\pi}{2} < \theta_i'' < \frac{\pi}{2}$, we must have $|k| < \frac{3}{4}$, which implies that $k = 0$, since k is an integer. This shows that $n(\gamma, z_0)$ is well defined.

The next two propositions are easily shown using the above technique. Proofs can be found in Ahlfors and Sario [1].

Proposition 3.2.1 *For every plane closed curve $\gamma\colon [0,1] \to \mathbb{A}^2$, for every z_0 not on γ, the index $n(\gamma, z_0)$ is continuous on the complement of γ in $\mathbb{A}^2$, and in fact constant in each connected component of the complement of γ. We have $n(\gamma, z_0) = 0$ in the unbounded component of the complement of γ.*

Proposition 3.2.2 *For any two plane closed curve $\gamma_1\colon [0,1] \to \mathbb{A}^2$ and $\gamma_2\colon [0,1] \to \mathbb{A}^2$, for every homotopy $F\colon [0,1] \times [0,1] \to \mathbb{A}^2$ between γ_1 and γ_2, for every z_0 not on any $F(t,u)$, for all $t, u \in [0,1]$, we have $n(\gamma_1, z_0) = n(\gamma_2, z_0)$.*

Proposition 3.2.2 shows that the index of a closed plane curve is not changed under homotopy (provided that none the curves involved go through z_0). We can now compute the fundamental group of the punctured plane, i.e., the plane from which a point is deleted.

3.3 The Fundamental Group of the Punctured Plane

First, we note that the fundamental group of $\mathbb{A}^n$ is the trivial group. Indeed, consider any closed curve $\gamma\colon [0,1] \to \mathbb{A}^n$ through $a = \gamma(0) = \gamma(1)$, take a as base point, and let a be the constant closed curve reduced to a. Note that the map

$$(t, u) \mapsto (1-u)\gamma(t)$$

is a homotopy between γ and a. Thus, there is a single homotopy class $[a]$, and $\pi(\mathbb{A}^n, a) = \{1\}$.

The above reasoning also shows that the fundamental group of an open ball, or a closed ball, is trivial. However, the next proposition shows that the fundamental group of the punctured plane is the infinite cyclic group $\mathbb{Z}$.

Proposition 3.3.1 *The fundamental group of the punctured plane is the infinite cyclic group $\mathbb{Z}$.*

Proof. Assume that the origin $z = 0$ is deleted from $\mathbb{A}^2 = \mathbb{C}$, and take $z = 1$ as base point. The unit circle can be parameterized as $t \mapsto \cos t + i \sin t$, and let α be the corresponding closed curve. First of all, note that for every closed curve $\gamma\colon [0,1] \to \mathbb{A}^2$ based at 1, there is a homotopy (central projection) $F\colon [0,1] \times [0,1] \to \mathbb{A}^2$ deforming γ into a curve β lying on the unit circle. By uniform continuity, any such curve β can be decomposed as $\beta = \beta_1\beta_2 \cdots \beta_n$, where each β_k either does not pass through $z = 1$, or does not pass through $z = -1$. It is also easy to see that β_k can deformed into one of the circular arcs δ_k between its endpoints. For all k, $2 \leq k \leq n$, let σ_k be one of the circular arcs from $z = 1$ to the initial point of δ_k, and let $\sigma_1 = \sigma_{n+1} = 1$. We have

$$\gamma \approx (\sigma_1\delta_1\sigma_2^{-1}) \cdots (\sigma_n\delta_n\sigma_{n+1}^{-1}),$$

and it is easily seen that each arc $\sigma_k\delta_k\sigma_{k+1}^{-1}$ is homotopic either to α, or α^{-1}, or 1. Thus, $\gamma \approx \alpha^m$, for some integer $m \in \mathbb{Z}$.

It remains to prove that α^m is not homotopic to 1 for $m \neq 0$. This is where we use Proposition 3.2.2. Indeed, it is immediate that $n(\alpha^m, 0) = m$, and $n(1, 0) = 0$, and thus α^m and 1 are not homotopic when $m \neq 0$. But then, we have shown that the homotopy classes are in bijection with the set of integers. □

The above proof also applies to a cicular annulus, closed or open, and to a circle. In particular, the circle is not simply connected.

We will need to define what it means for a surface to be orientable. Perhaps surprisingly, a rigorous definition is not so easy to obtain, but can be given using the notion of degree of a homeomorphism from a plane region. First, we need to define the degree of a map in the plane.

3.4 The Degree of a Map in the Plane

Let $\varphi\colon D \to \mathbb{C}$ be a continuous function to the plane, where the plane is viewed as the set $\mathbb{C}$ of complex numbers, and with domain some open set D in $\mathbb{C}$. We say that φ is *regular at* $z_0 \in D$ if there is some open set $V \subseteq D$ containing z_0 such that $\varphi(z) \neq \varphi(z_0)$, for all $z \in V$. Assuming that φ is regular at z_0, we will define the *degree of* φ *at* z_0.

Let Ω be a punctured open disk $\{z \in V \mid |z - z_0| < r\}$ contained in V. Since φ is regular at z_0, it maps Ω into the punctured plane Ω' obtained by deleting $w_0 = \varphi(z_0)$. Now, φ induces a homomorphism $\varphi_*\colon \pi(\Omega) \to \pi(\Omega')$. From Proposition 3.3.1, both groups $\pi(\Omega)$ and $\pi(\Omega')$ are isomorphic to $\mathbb{Z}$. Thus, it is easy to determine exactly what the homorphism φ_* is. We know that $\pi(\Omega)$ is generated by the homotopy class of some circle α in Ω with center a, and that $\pi(\Omega')$ is generated by the homotopy class of some circle β in Ω' with center $\varphi(a)$. If $\varphi_*([\alpha]) = [\beta^d]$, then the homomorphism φ_* is completely determined. If $d = 0$, then $\pi(\Omega') = 1$, and if $d \neq 0$, then $\pi(\Omega')$ is the infinite cyclic subgroup generated by the class of β^d. We let d be the *degree of* φ *at* z_0, and we denote it as $d(\varphi)_{z_0}$. It is easy to see that this definition does not depend on the choice of a (the center of the circle α) in Ω, and thus, does not depend on Ω.

Next, if we have a second mapping ψ regular at $w_0 = \varphi(z_0)$, then $\psi \circ \varphi$ is regular at z_0, and it is immediately verified that

$$d(\psi \circ \varphi)_{z_0} = d(\psi)_{w_0} d(\varphi)_{z_0}.$$

Let us now assume that D is a region (a connected open set), and that φ is a homeomorphism between D and $\varphi(D)$. By a theorem of Brouwer (the invariance of domain), it turns out that $\varphi(D)$ is also open, and thus, we can define the degree of the inverse mapping φ^{-1}, and since the identity clearly has degree 1, we get that $d(\varphi)d(\varphi^{-1}) = 1$, which shows that $d(\varphi)_{z_0} = \pm 1$.

In fact, Ahlfors and Sario [1] prove that if $\varphi(D)$ has a nonempty interior, then the degree of φ is constant on D. The proof is not difficult, but not very instructive.

Proposition 3.4.1 *Given a region D in the plane, for every homeomorphism φ between D and $\varphi(D)$, if $\varphi(D)$ has a nonempty interior, then the degree $d(\varphi)_z$ is constant for all $z \in D$, and in fact, $d(\varphi) = \pm 1$.*

When $d(\varphi) = 1$ in Proposition 3.4.1, we say that φ is *sense-preserving*, and when $d(\varphi) = -1$, we say that φ is *sense-reversing*. We can now define the notion of orientability.

3.5 Orientability of a Surface

Given a surface F, we will call a region V on F a *planar region* if there is a homeomorphism $h\colon V \to U$ from V onto an open set in the plane. From Proposition 3.4.1, the homeomorphisms $h\colon V \to U$ can be divided into two classes, by defining two such homeomorphisms h_1, h_2 as equivalent iff $h_1 \circ h_2^{-1}$ has degree 1, i.e., is sense-preserving. Observe that for any h as above, if $\overline{h}$ is obtained from h by conjugation (i.e., for every $z \in V$, $\overline{h}(z) = \overline{h(z)}$, the complex conjugate of $h(z)$), then $d(h \circ \overline{h^{-1}}) = -1$, and thus h and $\overline{h}$ are in different classes. For any other such map g, either $h \circ g^{-1}$ or $\overline{h} \circ g^{-1}$ is sense-preserving, and thus, there are exactly two equivalence classes.

The choice of one of the two classes of homeomorphims h as above, constitutes an *orientation* of V. An orientation of V induces an orientation on any subregion W of V, by restriction. If V_1 and V_2 are two planar regions, and these regions have received an orientation, we say that these orientations are *compatible* if they induce the same orientation on all common subregions of V_1 and V_2.

Definition 3.5.1 A surface F is *orientable* if it is possible to assign an orientation to all planar regions in such a way that the orientations of any two overlapping planar regions are compatible.

Clearly, orientability is preserved by homeomorphisms. Thus, there are two classes of surfaces, the orientable surfaces, and the nonorientable surfaces. An example of a nonorientable surface is the Klein bottle. Because we defined a surface as being connected, note that an orientable surface has exactly two orientations. It is also easy to see that to orient a surface, it is enough to orient all planar regions in some open covering of the surface by planar regions.

We will also need to consider bordered surfaces.

3.6 Bordered Surfaces

Consider a torus, and cut out a finite number of small disks from its surface. The resulting space is no longer a surface, but certainly of geometric interest. It is a surface with boundary, or bordered surface. In this section, we extend our concept of surface to handle this more general class of bordered surfaces. In order to do so, we need to allow coverings of surfaces using a richer class of open sets. This is achieved by considering the open subsets of the half-space, in the subset topology.

Definition 3.6.1 The *half-space* $\mathbb{H}^m$ is the subset of $\mathbb{R}^m$ defined as the set

$$\{(x_1, \ldots, x_m) \mid x_i \in \mathbb{R},\, x_m \geq 0\}.$$

For any $m \geq 1$, a *(topological) m-manifold with boundary* is a second-countable, topological Hausdorff space M, together with an open cover $(U_i)_{i\in I}$ of open sets and a family $(\varphi_i)_{i\in I}$ of homeomorphisms $\varphi_i\colon U_i \to \Omega_i$, where each Ω_i is some open subset of $\mathbb{H}^m$ in the subset topology. Each pair (U, φ) is called a *coordinate system*, or *chart*, of M, each homeomorphism $\varphi_i\colon U_i \to \Omega_i$ is called a *coordinate map*, and its inverse $\varphi_i^{-1}\colon \Omega_i \to U_i$ is called a *parameterization* of U_i. The family $(U_i, \varphi_i)_{i\in I}$ is often called an *atlas* for M. A *(topological) bordered surface* is a connected 2-manifold with boundary.

Note that an m-manifold is also an m-manifold with boundary.

If $\varphi_i\colon U_i \to \Omega_i$ is some homeomorphism onto some open set Ω_i of $\mathbb{H}^m$ in the subset topology, some $p \in U_i$ may be mapped into $\mathbb{R}^{m-1} \times \mathbb{R}_+$, or into the "boundary" $\mathbb{R}^{m-1} \times \{0\}$ of $\mathbb{H}^m$. Letting $\partial\mathbb{H}^m = \mathbb{R}^{m-1} \times \{0\}$, it can be shown using homology, that if some coordinate map φ defined on p maps p into $\partial\mathbb{H}^m$, then every coordinate map ψ defined on p maps p into $\partial\mathbb{H}^m$. For $m = 2$, Ahlfors and Sario prove it using Proposition 3.4.1.

Thus, M is the disjoint union of two sets ∂M and $\mathrm{Int}M$, where ∂M is the subset consisting of all points $p \in M$ that are mapped by some (in fact, all) coordinate map φ defined on p into $\partial\mathbb{H}^m$, and where $\mathrm{Int}\, M = M - \partial M$. The set ∂M is called the *boundary* of M, and the set $\mathrm{Int}\, M$ is called the *interior* of M, even though this terminology clashes with some prior topological definitions. A good example of a bordered surface is the Möbius strip. The boundary of the Möbius strip is a circle.

The boundary ∂M of M may be empty, but $\mathrm{Int}\, M$ is nonempty. Also, it can be shown using homology, that the integer m is unique. It is clear that $\mathrm{Int}\, M$ is open, and an m-manifold, and that ∂M is closed. If $p \in \partial M$, and φ is some coordinate map defined on p, since $\Omega = \varphi(U)$ is an open subset of $\partial\mathbb{H}^m$, there is some open half ball B^m_{o+} centered at $\varphi(p)$ and contained in Ω which intersects $\partial\mathbb{H}^m$ along an open ball B^{m-1}_o, and if we consider $W = \varphi^{-1}(B^m_{o+})$, we have an open subset of M containing p which is mapped homeomorphically onto B^m_{o+} in such that way that every point in $W \cap \partial M$ is mapped onto the open ball B^{m-1}_o. Thus, it is easy to see that ∂M is an $(m-1)$-manifold.

In particular, in the case $m = 2$, the boundary ∂M is a union of curves homeomorphic either to circles of to open line segments. In this case, if M is connected but not a surface, it is easy to see that M is the topological closure of $\mathrm{Int}\, M$. We also claim that $\mathrm{Int}\, M$ is connected, i.e. a surface. Indeed, if this was not so, we could write $\mathrm{Int}\, M = M_1 \cup M_2$, for two nonempty disjoint sets M_1 and M_2. But then, we have $M = \overline{M_1} \cup \overline{M_2}$, and since M is connected, there is some $a \in \partial M$ also in $\overline{M_1} \cap \overline{M_2} \neq \emptyset$. However, there is some open set V containing a whose intersection with M is homeomorphic with an open half-disk, and thus connected. Then, we have

$$V \cap M = (V \cap M_1) \cup (V \cap M_2),$$

with $V \cap M_1$ and $V \cap M_2$ open in V, contradicting the fact that $M \cap V$ is connected. Thus, $\mathrm{Int}\, M$ is a surface.

When the boundary ∂M of a bordered surface M is empty, M is just a surface. Typically, when we refer to a bordered surface, we mean a bordered surface with a nonempty border, and otherwise, we just say surface.

A bordered surface M is orientable iff its interior $\mathrm{Int}\, M$ is orientable. It is not difficult to show that an orientation of $\mathrm{Int}\, M$ induces an orientation of the boundary ∂M. The components of the boundary ∂M are called *contours*.

The concept of triangulation of a bordered surface is identical to the concept defined for a surface in Definition 2.2.1, and Proposition 2.2.2 also holds. However, a small change needs to made to Proposition 2.2.3, see Ahlfors and Sario [1].

Proposition 3.6.2 *A 2-complex $K = (V, \mathcal{S})$ is a triangulation $\sigma\colon \mathcal{S} \to 2^M$ of a bordered surface M such that $\sigma(s) = s_g$ for all $s \in \mathcal{S}$ iff the following properties hold:*

(D1) Every edge a such that a_g contains some point in the interior $\mathrm{Int}\, M$ *of M is contained in exactly two triangles A. Every edge a such that a_g is inside the border ∂M of M is contained in exactly one triangle A. The border ∂M of M consists of those a_g which belong to only one A_g. A border vertex or border edge is a simplex σ such that $\sigma_g \subseteq \partial M$.*

(D2) For every non-border vertex α, the edges a and triangles A containing α can be arranged as a cyclic sequence $a_1, A_1, a_2, A_2, \ldots, A_{m-1}, a_m, A_m$, in the sense that $a_i = A_{i-1} \cap A_i$ for all i, with $2 \leq i \leq m$, and $a_1 = A_m \cap A_1$, with $m \geq 3$.

(D3) For every border vertex α, the edges a and triangles A containing α can be arranged in a sequence $a_1, A_1, a_2, A_2, \ldots, A_{m-1}, a_m, A_m, a_{m+1}$, with $a_i = A_i \cap A_{i-1}$ for of all i, with $2 \leq i \leq m$, where a_1 and a_{m+1} are border vertices only contained in A_1 and A_m respectively.

(D4) K is connected, in the sense that it cannot be written as the union of two disjoint nonempty complexes.

A 2-complex K which satisfies the conditions of Proposition 3.6.2 will also be called a *bordered triangulated* 2*-complex*, and its geometric realization a *bordered polyhedron*. Thus, bordered triangulated 2-complexes are the complexes that correspond to triangulated bordered surfaces. Actually, it can be shown that every bordered surface admits some triangulation, and thus the class of geometric realizations of the bordered triangulated 2-complexes is the class of all bordered surfaces.

We will now give a brief presentation of simplicial and singular homology, but first, we need to review some facts about finitely generated abelian groups.

Chapter 4

Homology Groups

4.1 Finitely Generated Abelian Groups

An abelian group is a commutative group. We will denote the identity element of an abelian group as 0, and the inverse of an element a as $-a$. Given any natural number $n \in \mathbb{N}$, we denote

$$\underbrace{a + \cdots + a}_{n}$$

as na, and let $(-n)a$ be defined as $n(-a)$ (with $0a = 0$). Thus, we can make sense of finite sums of the form $\sum n_i a_i$, where $n_i \in \mathbb{Z}$. Given an abelian group G and a family $A = (a_j)_{j \in J}$ of elements $a_j \in G$, we say that G is *generated by A* if every $a \in G$ can be written (in possibly more than one way) as

$$a = \sum_{i \in I} n_i a_i,$$

for some finite subset I of J, and some $n_i \in \mathbb{Z}$. If J is finite, we say that G is *finitely generated by A*. If every $a \in G$ can be written in a *unique manner* as

$$a = \sum_{i \in I} n_i a_i$$

as above, we say that G is *freely generated by A*, and we call A a *basis of* G. In this case, it is clear that the a_j are all distinct. We also have the following familiar property.

If G is a free abelian group generated by $A = (a_j)_{j \in J}$, for every abelian group H, for every function $f \colon A \to H$, there is a unique homomorphism $\widehat{f} \colon G \to H$, such that $\widehat{f}(a_j) = f(a_j)$, for all $j \in J$.

Remark: If G is a free abelian group, one can show that the cardinality of all bases is the same. When G is free and finitely generated by $(a_1, \ldots, a_n)$, this can be proved as follows. Consider the quotient of the group G modulo the subgroup $2G$ consisting of all elements of the form $g + g$, where $g \in G$. It is immediately verified that each coset of $G/2G$ is of the form

$$\epsilon_1 a_1 + \cdots + \epsilon_n a_n + 2G,$$

where $\epsilon_i = 0$ or $\epsilon_i = 1$, and thus, $G/2G$ has 2^n elements. Thus, n only depends on G. The number n is called the *dimension* of G.

Given a family $A = (a_j)_{j \in J}$, we will need to construct a free abelian group generated by A. This can be done easily as follows. Consider the set $F(A)$ of all functions $\varphi\colon A \to \mathbb{Z}$, such that $\varphi(a) \neq 0$ for only finitely many $a \in A$. We define addition on $F(A)$ pointwise, that is, $\varphi + \psi$ is the function such that $(\varphi + \psi)(a) = \varphi(a) + \psi(a)$, for all $a \in A$.

It is immediately verified that $F(A)$ is an abelian group, and if we identify each a_j with the function $\varphi_j\colon A \to \mathbb{Z}$, such that $\varphi_j(a_j) = 1$, and $\varphi_j(a_i) = 0$ for all $i \neq j$, it is clear that $F(A)$ is freely generated by A. It is also clear that every $\varphi \in F(A)$ can be uniquely written as

$$\varphi = \sum_{i \in I} n_i \varphi_i,$$

for some finite subset I of J such that $n_i = \varphi(a_i) \neq 0$. For notational simplicity, we write φ as

$$\varphi = \sum_{i \in I} n_i a_i.$$

Given an abelian group G, for any $a \in G$, we say that a has *finite order* if there is some $n \neq 0$ in $\mathbb{N}$ such that $na = 0$. If $a \in G$ has finite order, there is a least $n \neq 0$ in $\mathbb{N}$ such that $na = 0$, called the *order of* a. It is immediately verified that the subset T of G consisting of all elements of finite order is a subroup of G, called the *torsion subgroup of* G. When $T = \{0\}$, we say that G is *torsion-free*. One should be careful that a torsion-free abelian group is not necessarily free. For example, the field $\mathbb{Q}$ of rationals is torsion-free, but not a free abelian group.

Clearly, the map $(n, a) \mapsto na$ from $\mathbb{Z} \times G$ to G satisfies the properties

$$\begin{aligned}(m + n)a &= ma + na,\\ m(a + b) &= ma + nb,\\ (mn)a &= m(na),\\ 1a &= a,\end{aligned}$$

which hold in vector spaces. However, $\mathbb{Z}$ is not a field. The abelian group G is just what is called a $\mathbb{Z}$*-module*. Nevertheless, many concepts defined for vector spaces transfer to $\mathbb{Z}$-modules. For example, given an abelian group G and some subgroups $H_1, \ldots, H_n$, we can define the *(internal) sum*

$$H_1 + \cdots + H_n$$

of the H_i as the abelian group consisting of all sums of the form $a_1 + \cdots + a_n$, where $a_i \in H_i$. If in addition, $G = H_1 + \cdots + H_n$ and $H_i \cap H_j = \{0\}$ for all i, j, with $i \neq j$, we say that G is the *direct sum of the* H_i, and this is denoted as

$$G = H_1 \oplus \cdots \oplus H_n.$$

When $H_1 = \ldots = H_n = H$, we abbreviate $H \oplus \cdots \oplus H$ as H^n. Homomorphims between abelian groups are $\mathbb{Z}$-linear maps. We can also talk about linearly independent families in G, except that the scalars are in $\mathbb{Z}$. The *rank* of an abelian group is the maximum of the sizes of linearly independent families in G. We can also define (external) direct sums.

Given a family $(G_i)_{i\in I}$ of abelian groups, the *(external) direct sum* $\bigoplus_{i\in I} G_i$ is the set of all function $f\colon I \to \bigcup_{i\in I} G_i$ such that $f(i) \in G_i$, for all $\in I$, and $f(i) = 0$ for all but finitely many $i \in I$. An element $f \in \bigoplus_{i\in I} G_i$ is usually denoted as $(f_i)_{i\in I}$. Addition is defined component-wise, that is, given two functions $f = (f_i)_{i\in I}$ and $g = (g_i)_{i\in I}$ in $\bigoplus_{i\in I} G_i$, we define $(f + g)$ such that

$$(f + g)_i = f_i + g_i,$$

for all $i \in I$. It is immediately verified that $\bigoplus_{i\in I} G_i$ is an abelian group. For every $i \in I$, there is an injective homomorphism $in_i\colon G_i \to \bigoplus_{i\in I} G_i$, defined such that for every $x \in G_i$, $in_i(x)(i) = x$, and $in_i(x)(j) = 0$ iff $j \neq i$. If $G = \bigoplus_{i\in I} G_i$ is an external direct sum, it is immediately verified that $G = \bigoplus_{i\in I} in_i(G_i)$, as an internal direct sum. The difference is that G must have been already defined for an internal direct sum to make sense. For notational simplicity, we will usually identify $in_i(G_i)$ with G_i.

The structure of finitely generated abelian groups can be completely described. Actually, the following result is a special case of the structure theorem for finitely generated modules over a principal ring. Recall that $\mathbb{Z}$ is a principal ring, which means that every ideal $\mathcal{I}$ in $\mathbb{Z}$ is of the form $d\mathbb{Z}$, for some $d \in \mathbb{N}$. For the sake of completeness, we present the following result, whose neat proof is due to Pierre Samuel.

Proposition 4.1.1 *Let G be a free abelian group finitely generated by $(a_1, \ldots, a_n)$, and let H be any subroup of G. Then, H is a free abelian group, and there is a basis $(e_1, ..., e_n)$ of G, some $q \leq n$, and some positive natural numbers $n_1, \ldots, n_q$, such that $(n_1e_1, \ldots, n_qe_q)$ is a basis of H, and n_i divides n_{i+1} for all i, with $1 \leq i \leq q - 1$.*

Proof. The proposition is trivial when $H = \{0\}$, and thus, we assume that H is nontrivial. Let $L(G, \mathbb{Z})$ we the set of homomorphisms from G to $\mathbb{Z}$. For any $f \in L(G, \mathbb{Z})$, it is immediately verified that $f(H)$ is an ideal in $\mathbb{Z}$. Thus, $f(H) = n_h\mathbb{Z}$, for some $n_h \in \mathbb{N}$, since every ideal in $\mathbb{Z}$ is a principal ideal. Since $\mathbb{Z}$ is finitely generated, any nonempty family of ideals has a maximal element, and let f be a homomorphism such that $n_h\mathbb{Z}$ is a maximal ideal in $\mathbb{Z}$. Let $\pi\colon G \to \mathbb{Z}$ be the i-th projection, i.e., π_i is defined such that $\pi_i(m_1a_1 + \cdots + m_na_n) = m_i$. It is clear that π_i is a homomorphism, and since H is nontrivial, one of the $\pi_i(H)$ is nontrivial, and $n_h \neq 0$. There is some $b \in H$ such that $f(b) = n_h$.

We claim that for every $g \in L(G, \mathbb{Z})$, the number n_h divides $g(b)$. Indeed, if d is the gcd of n_h and $g(b)$, by the Bezout identity, we can write

$$d = rn_h + sg(b),$$

for some $r, s \in \mathbb{Z}$, and thus

$$d = rf(b) + sg(b) = (rf + sg)(b).$$

However, $rf + sg \in L(G, \mathbb{Z})$, and thus,

$$n_h\mathbb{Z} \subseteq d\mathbb{Z} \subseteq (rf + sg)(H),$$

since d divides n_h, and by maximality of $n_h\mathbb{Z}$, we must have $n_h\mathbb{Z} = d\mathbb{Z}$, which implies that $d = n_h$, and thus, n_h divides $g(b)$. In particular, n_h divides each $\pi_i(b)$, and let $\pi_i(b) = n_h p_i$, with $p_i \in \mathbb{Z}$.

Let $a = p_1a_1 + \cdots + p_na_n$. Note that

$$b = \pi_1(b)a_1 + \cdots + \pi_n(b)a_n = n_hp_1a_1 + \cdots + n_hp_na_n,$$

and thus, $b = n_ha$. Since $n_h = f(b) = f(n_ha) = n_hf(a)$, and since $n_h \neq 0$, we must have $f(a) = 1$.

Next, we claim that

$$G = a\mathbb{Z} \oplus f^{-1}(0),$$

and

$$H = b\mathbb{Z} \oplus (H \cap f^{-1}(0)),$$

with $b = n_ha$.

Indeed, every $x \in G$ can be written as

$$x = f(x)a + (x - f(x)a),$$

and since $f(a) = 1$, we have $f(x - f(x)a) = f(x) - f(x)f(a) = f(x) - f(x) = 0$. Thus, $G = a\mathbb{Z} + f^{-1}(0)$. Similarly, for any $x \in H$, we have $f(x) = rn_h$, for some $r \in \mathbb{Z}$, and thus,

$$x = f(x)a + (x - f(x)a) = rn_ha + (x - f(x)a) = rb + (x - f(x)a),$$

we still have $x - f(x)a \in f^{-1}(0)$, and clearly, $x - f(x)a = x - rn_ha = x - rb \in H$, since $b \in H$. Thus, $H = b\mathbb{Z} + (H \cap f^{-1}(0))$.

To prove that we have a direct sum, it is enough to prove that $a\mathbb{Z} \cap f^{-1}(0) = \{0\}$. For any $x = ra \in a\mathbb{Z}$, if $f(x) = 0$, then $f(ra) = rf(a) = r = 0$, since $f(a) = 1$, and thus, $x = 0$. Therefore, the sums are direct sums.

We can now prove that H is a free abelian group by induction on the size q of a maximal linearly independent family for H. If $q = 0$, the result is trivial. Otherwise, since

$$H = b\mathbb{Z} \oplus (H \cap f^{-1}(0)),$$

it is clear that $H \cap f^{-1}(0)$ is a subgroup of G and that every maximal linearly independent family in $H \cap f^{-1}(0)$ has at most $q - 1$ elements. By the induction hypothesis, $H \cap f^{-1}(0)$ is a free abelian group, and by adding b to a basis of $H \cap f^{-1}(0)$, we obtain a basis for H, since the sum is direct.

The second part is shown by induction on the dimension n of G. The case $n = 0$ is trivial. Otherwise, since

$$G = a\mathbb{Z} \oplus f^{-1}(0),$$

and since by the previous argument, $f^{-1}(0)$ is also free, it is easy to see that $f^{-1}(0)$ has dimension $n - 1$. By the induction hypothesis applied to its subgroup $H \cap f^{-1}(0)$, there is a basis $(e_2, \ldots, e_n)$ of $f^{-1}(0)$, some $q \leq n$, and some positive natural numbers $n_2, \ldots, n_q$, such that, $(n_2e_2, \ldots, n_qe_q)$ is a basis of $H \cap f^{-1}(0)$, and n_i divides n_{i+1} for all i, with $2 \leq i \leq q-1$. Let $e_1 = a$, and $n_1 = n_h$, as above. It is clear that $(e_1, \ldots, e_n)$ is a basis of G, and that that $(n_1e_1, \ldots, n_qe_q)$ is a basis of H, since the sums are direct, and $b = n_1e_1 = n_ha$. It remains to show that n_1 divides n_2. Consider the homomorphism $g\colon G \to \mathbb{Z}$ such that $g(e_1) = g(e_2) = 1$, and $g(e_i) = 0$, for all i, with $3 \leq i \leq n$. We have $n_h = n_1 = g(n_1e_1) = g(b) \in g(H)$, and thus, $n_h\mathbb{Z} \subseteq g(H)$. Since $n_h\mathbb{Z}$ is maximal, we must have $g(H) = n_h\mathbb{Z} = n_1\mathbb{Z}$. Since $n_2 = g(n_2e_2) \in g(H)$, we have $n_2 \in n_1\mathbb{Z}$, which shows that n_1 divides n_2. □

Using Proposition 4.1.1, we can also show the following useful result.

Proposition 4.1.2 *Let G be a finitely generated abelian group. There is some natural number $m \geq 0$ and some positive natural numbers $n_1, \ldots, n_q$, such that H is isomorphic to the direct sum*

$$\mathbb{Z}^m \oplus \mathbb{Z}/n_1\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/n_q\mathbb{Z},$$

and where n_i divides n_{i+1} for all i, with $1 \leq i \leq q - 1$.

Proof. Assume that G is generated by $A = (a_1, \ldots, a_n)$, and let $F(A)$ be the free abelian group generated by A. The inclusion map $i\colon A \to G$ can be extended to a unique homomorphism $f\colon F(A) \to G$ which is surjective since A generates G, and thus, G is isomorphic to $F(A)/f^{-1}(0)$. By Proposition 4.1.1, $H = f^{-1}(0)$ is a free abelian group, and there is a basis $(e_1, ..., e_n)$ of G, some $p \leq n$, and some positive natural numbers $k_1, \ldots, k_p$, such that $(k_1e_1, \ldots, k_pe_p)$ is a basis of H, and k_i divides k_{i+1} for all i, with $1 \leq i \leq p - 1$. Let r, $0 \leq r \leq p$, be the largest natural number such that $k_1 = \ldots = k_r = 1$, rename k_{r+i} as n_i, where $1 \leq i \leq p - r$, and let $q = p - r$. Then, we can write

$$H = \mathbb{Z}^{p-q} \oplus n_1\mathbb{Z} \oplus \cdots \oplus n_q\mathbb{Z},$$

and since $F(A)$ is isomorphic to $\mathbb{Z}^n$, it is easy to verify that $F(A)/H$ is isomorphic to

$$Z^{n-p} \oplus \mathbb{Z}/n_1\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/n_q\mathbb{Z},$$

which proves the proposition. □

Observe that $\mathbb{Z}/n_1\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/n_q\mathbb{Z}$ is the torsion subgroup of G. Thus, as a corollary of Proposition 4.1.2, we obtain the fact that every finitely generated abelian group G is a direct sum $G = Z^m \oplus T$, where T is the torsion subroup of G, and Z^m is the free abelian group of dimension m. It is easy to verify that m is the rank (the maximal dimension of linearly independent sets in G) of G, and it is called the *Betti number* of G. It can also be shown that q, and the n_i, only depend on G.

One more result will be needed to compute the homology groups of (two-dimensional) polyhedras. The proof is not difficult and can be found in most books (a version is given in Ahlfors and Sario [1]). Let us denote the rank of an abelian group G as $r(G)$.

Proposition 4.1.3 *If*

$$0 \longrightarrow E \xrightarrow{f} F \xrightarrow{g} G \longrightarrow 0$$

is a short exact sequence of homomorphisms of abelian groups and F has finite rank, then $r(F) = r(E) + r(G)$. In particular, if G is an abelian group of finite rank and H is a subroup of G, then $r(G) = r(H) + r(G/H)$.

We are now ready to define the simplicial and the singular homology groups.

4.2 Simplicial and Singular Homology

There are several kinds of homology theories. In this section, we take a quick look at two such theories, simplicial homology, one of the most computational theories, and singular homology theory, one of the most general and yet fairly intuitive. For a comprehensive treatment of homology and algebraic topology in general, we refer the reader to Massey [12], Munkres [14], Bredon [3], Fulton [7], Dold [5], Rotman [15], Amstrong [2], and Kinsey [9]. An excellent overview of algebraic topology, following a more intuitive approach, is presented in Sato [16].

Let $K = (V, \mathcal{S})$ be a complex. The essence of simplicial homology is to associate some abelian groups $H_p(K)$ with K. This is done by first defining some free abelian groups $C_p(K)$ made out of oriented p-simplices. One of the main new ingredients is that every oriented p-simplex σ is assigned a *boundary* $\partial_p\sigma$. Technically, this is achieved by defining homomorphisms

$$\partial_p \colon C_p(K) \to C_{p-1}(K),$$

with the property that $\partial_{p-1} \circ \partial_p = 0$. Letting $Z_p(K)$ be the kernel of ∂_p, and

$$B_p(K) = \partial_{p+1}(C_{p+1}(K))$$

be the image of ∂_{p+1} in $C_p(K)$, since $\partial_p \circ \partial_{p+1} = 0$, the group $B_p(K)$ is a subgroup of the group $Z_p(K)$, and we define the homology group $H_p(K)$ as the quotient group

$$H_p(K) = Z_p(K)/B_p(K).$$

What makes the homology groups of a complex interesting, is that they only depend on the geometric realization K_g of the complex K, and not on the various complexes representing K_g. Proving this fact requires relatively hard work, and we refer the reader to Munkres [14] or Rotman [15], for a proof.

The first step in defining simplicial homology groups is to define oriented simplices. Given a complex $K = (V, \mathcal{S})$, recall that an n-simplex is a subset $\sigma = \{\alpha_0, \ldots, \alpha_n\}$ of V that belongs to the family $\mathcal{S}$. Thus, the set σ corresponds to $(n+1)!$ linearly ordered sequences $s\colon \{1, 2, \ldots, n+1\} \to \sigma$, where each s is a bijection. We define an equivalence relation on these sequences by saying that two sequences $s_1\colon \{1, 2, \ldots, n+1\} \to \sigma$ and $s_2\colon \{1, 2, \ldots, n+1\} \to \sigma$ are equivalent iff $\pi = s_2^{-1} \circ s_1$ is a permutation of even signature (π is the product of an even number of transpositions)

The two equivalence classes associated with σ are called *oriented simplices*, and if $\sigma = \{\alpha_0, \ldots, \alpha_n\}$, we denote the equivalence class of s as $[s(1), \ldots, s(n+1)]$, where s is one of the sequences $s\colon \{1, 2, \ldots, n+1\} \to \sigma$. We also say that the two classes associated with σ are the *orientations of* σ. Two oriented simplices σ_1 and σ_2 are said to have *opposite orientation* if they are the two classes associated with some simplex σ. Given an oriented simplex σ, we denote the oriented simplex having the opposite orientation as $-\sigma$, with the convention that $-(-\sigma) = \sigma$.

For example, if $\sigma = \{a_1, a_2, a_3\}$ is a 3-simplex (a triangle), there are six ordered sequences, the sequences $\langle a_3, a_2, a_1\rangle$, $\langle a_2, a_1, a_3\rangle$, and $\langle a_1, a_3, a_2\rangle$, are equivalent, and the sequences $\langle a_1, a_2, a_3\rangle$, $\langle a_2, a_3, a_1\rangle$, and $\langle a_3, a_1, a_2\rangle$, are also equivalent. Thus, we have the two oriented simplices, $[a_1, a_2, a_3]$ and $[a_3, a_2, a_1]$. We now define p-chains.

Definition 4.2.1 Given a complex $K = (V, \mathcal{S})$, a *p-chain* on K is a function c from the set of oriented p-simplices to $\mathbb{Z}$, such that,

(1) $c(-\sigma) = -c(\sigma)$, iff σ and $-\sigma$ have opposite orientation;

(2) $c(\sigma) = 0$, for all but finitely many simplices σ.

We define addition of p-chains pointwise, i.e., $c_1 + c_2$ is the p-chain such that $(c_1 + c_2)(\sigma) = c_1(\sigma) + c_2(\sigma)$, for every oriented p-simplex σ. The group of p-chains is denoted as $C_p(K)$. If $p < 0$ or $p > \dim(K)$, we set $C_p(K) = \{0\}$.

To every oriented p-simplex σ is associated an *elementary p-chain* c, defined such that,

$c(\sigma) = 1$,

$c(-\sigma) = -1$, where $-\sigma$ is the opposite orientation of σ, and

$c(\sigma') = 0$, for all other oriented simplices σ'.

We will often denote the elementary p-chain associated with the oriented p-simplex σ also as σ.

The following proposition is obvious, and simply confirms the fact that $C_p(K)$ is indeed a free abelian group.

Proposition 4.2.2 *For every complex $K = (V, \mathcal{S})$, for every p, the group $C_p(K)$ is a free abelian group. For every choice of an orientation for every p-simplex, the corresponding elementary chains form a basis for $C_p(K)$.*

The only point worth elaborating is that except for $C_0(K)$, where no choice is involved, there is no canonical basis for $C_p(K)$ for $p \geq 1$, since different choices for the orientations of the simplices yield different bases.

If there are m_p p-simplices in K, the above proposition shows that $C_p(K) = \mathbb{Z}^{m_p}$.

As an immediate consequence of Proposition 4.2.2, for any abelian group G and any function f mapping the oriented p-simplices of a complex K to G, and such that $f(-\sigma) = -f(\sigma)$ for every oriented p-simplex σ, there is a unique homomorphism $\widehat{f}\colon C_p(K) \to G$ extending f.

We now define the boundary maps $\partial_p\colon C_p(K) \to C_{p-1}(K)$.

Definition 4.2.3 Given a complex $K = (V, \mathcal{S})$, for every oriented p-simplex

$$\sigma = [\alpha_0, \ldots, \alpha_p],$$

we define the *bounday $\partial_p \sigma$ of* σ as

$$\partial_p \sigma = \sum_{i=0}^{p} (-1)^i [\alpha_0, \ldots, \widehat{\alpha_i}, \ldots, \alpha_p],$$

where $[\alpha_0, \ldots, \widehat{\alpha_i}, \ldots, \alpha_p]$ denotes the oriented $p-1$-simplex obtained by deleting vertex α_i. The *boundary map* $\partial_p\colon C_p(K) \to C_{p-1}(K)$ is the unique homomorphism extending ∂_p on oriented p-simplices. For $p \leq 0$, ∂_p is the null homomorphism.

One must verify that $\partial_p(-\sigma) = -\partial_p \sigma$, but this is immediate. If $\sigma = [\alpha_0, \alpha_1]$, then

$$\partial_1 \sigma = \alpha_1 - \alpha_0.$$

If $\sigma = [\alpha_0, \alpha_1, \alpha_2]$, then

$$\partial_2 \sigma = [\alpha_1, \alpha_2] - [\alpha_0, \alpha_2] + [\alpha_0, \alpha_1] = [\alpha_1, \alpha_2] + [\alpha_2, \alpha_0] + [\alpha_0, \alpha_1].$$

If $\sigma = [\alpha_0, \alpha_1, \alpha_2, \alpha_3]$, then

$$\partial_3 \sigma = [\alpha_1, \alpha_2, \alpha_3] - [\alpha_0, \alpha_2, \alpha_3] + [\alpha_0, \alpha_1, \alpha_3] - [\alpha_0, \alpha_1, \alpha_2].$$

We have the following fundamental property.

Proposition 4.2.4 *For every complex $K = (V, \mathcal{S})$, for every p, we have $\partial_{p-1} \circ \partial_p = 0$.*

Proof. For any oriented p-simplex $\sigma = [\alpha_0, \ldots, \alpha_p]$, we have

$$\begin{aligned}
\partial_{p-1} \circ \partial_p \sigma &= \sum_{i=0}^{p} (-1)^i \partial_{p-1}[\alpha_0, \ldots, \widehat{\alpha_i}, \ldots, \alpha_p], \\
&= \quad \sum_{i=0}^{p} \sum_{j=0}^{i-1} (-1)^i (-1)^j [\alpha_0, \ldots, \widehat{\alpha_j}, \ldots, \widehat{\alpha_i}, \ldots, \alpha_p] \\
&\quad + \sum_{i=0}^{p} \sum_{j=i+1}^{p} (-1)^i (-1)^{j-1} [\alpha_0, \ldots, \widehat{\alpha_i}, \ldots, \widehat{\alpha_j}, \ldots, \alpha_p] \\
&= 0.
\end{aligned}$$

The rest of the proof follows from the fact that $\partial_p \colon C_p(K) \to C_{p-1}(K)$ is the unique homomorphism extending ∂_p on oriented p-simplices. □

In view of Proposition 4.2.4, the image $\partial_{p+1}(C_{p+1}(K))$ of $\partial_{p+1} \colon C_{p+1}(K) \to C_p(K)$ is a subgroup of the kernel $\partial_p^{-1}(0)$ of $\partial_p \colon C_p(K) \to C_{p-1}(K)$. This motivates the following definition.

Definition 4.2.5 Given a complex $K = (V, \mathcal{S})$, the kernel $\partial_p^{-1}(0)$ of the homomorphism $\partial_p \colon C_p(K) \to C_{p-1}(K)$ is denoted as $Z_p(K)$, and the elements of $Z_p(K)$ are called *p-cycles*. The image $\partial_{p+1}(C_{p+1})$ of the homomorphism $\partial_{p+1} \colon C_{p+1}(K) \to C_p(K)$ is denoted as $B_p(K)$, and the elements of $B_p(K)$ are called *p-boundaries*. The *p-th homology group $H_p(K)$* is the quotient group

$$H_p(K) = Z_p(K)/B_p(K).$$

Two p-chains c, c' are said to be *homologous* if there is some $(p+1)$-chain d such that $c = c' + \partial_{p+1} d$.

We will often omit the subscript p in ∂_p.

At this stage, we could determine the homology groups of the finite (two-dimensional) polyhedras. However, we are really interested in the homology groups of geometric realizations of complexes, in particular, compact surfaces, and so far, we have not defined homology groups for topological spaces.

It is possible to define homology groups for arbitrary topological spaces, using what is called *singular homology*. Then, it can be shown, although this requires some hard work, that the homology groups of a space X which is the geometric realization of some complex K are independent of the complex K such that $X = K_g$, and equal to the homology groups of any such complex.

The idea behind singular homology is to define a more general notion of an n-simplex associated with a topological space X, and it is natural to consider continuous maps from some standard simplices to X. Recall that given any set I, we defined the real vector space $\mathbb{R}^{(I)}$ freely generated by I (just before Definition 2.1.5). In particular, for $I = \mathbb{N}$ (the natural numbers), we obtain an infinite dimensional vector space $\mathbb{R}^{(\mathbb{N})}$, whose elements are the countably infinite sequences $(\lambda_i)_{i\in\mathbb{N}}$ of reals, with $\lambda_i = 0$ for all but finitely many $i \in \mathbb{N}$. For any $p \in \mathbb{N}$, we let $e_i \in \mathbb{R}^{(\mathbb{N})}$ be the sequence such that $e_i(i) = 1$ and $e_i(j) = 0$ for all $j \neq i$, and we let Δ_p be the p-simplex spanned by $(e_0, \ldots, e_p)$, that is, the subset of $\mathbb{R}^{(\mathbb{N})}$ consisting of all points of the form

$$\sum_{i=0}^{p} \lambda_i e_i, \quad \text{with} \quad \sum_{i=0}^{p} \lambda_i = 1, \text{ and } \lambda_i \geq 0.$$

We call Δ_p the *standard p-simplex*. Note that Δ_{p-1} is a face of Δ_p.

Definition 4.2.6 Given a topological space X, a *singular p-simplex* is any continuous map $T\colon \Delta_p \to X$. The free abelian group generated by the singular p-simplices is called the *p-th singular chain group*, and is denoted as $S_p(X)$.

Given any $p + 1$ points $a_0, \ldots, a_p$ in $\mathbb{R}^{(\mathbb{N})}$, there is a unique affine map $f\colon \Delta_p \to \mathbb{R}^{(\mathbb{N})}$, such that $f(e_i) = a_i$, for all i, $0 \leq i \leq p$, namely the map such that

$$f(\sum_{i=0}^{p} \lambda_i e_i) = \sum_{i=0}^{p} \lambda_i a_i,$$

for all λ_i such that $\sum_{i=0}^{p} \lambda_i = 1$, and $\lambda_i \geq 0$. This map is called the *affine singular simplex* determined by $a_0, \ldots, a_p$, and it is denoted as $l(a_0, \ldots, a_p)$. In particular, the map

$$l(e_0, \ldots, \widehat{e_i}, \ldots, e_p),$$

where the hat over e_i means that e_i is omited, is a map from Δ_{p-1} onto a face of Δ_p. We can consider it as a map from Δ_{p-1} to Δ_p (although it is defined as a map from Δ_{p-1} to $\mathbb{R}^{(\mathbb{N})}$), and call it the i-th face of Δ_p.

Then, if $T\colon \Delta_p \to X$ is a singular p-simplex, we can form the map

$$T \circ l(e_0, \ldots, \widehat{e_i}, \ldots, e_p)\colon \Delta_{p-1} \to X,$$

which is a singular $p-1$-simplex, which we think of as the i-th face of T. Actually, for $p = 1$, a singular p-simplex $T\colon \Delta_p \to X$ can be viewed as curve on X, and its faces are its two endpoints. For $p = 2$, a singular p-simplex $T\colon \Delta_p \to X$ can be viewed as triangular surface patch on X, and its faces are its three boundary curves. For $p = 3$, a singular p-simplex $T\colon \Delta_p \to X$ can be viewed as tetrahedral "volume patch" on X, and its faces are its four boundary surface patches. We can give similar higher-order descriptions when $p > 3$.

We can now define the boundary maps $\partial_p\colon S_p(X) \to S_{p-1}(X)$.

Definition 4.2.7 Given a topological space X, for every singular p-simplex $T\colon \Delta_p \to X$, we define the *bounday $\partial_p T$ of T* as

$$\partial_p T = \sum_{i=0}^{p} (-1)^i \, T \circ l(e_0, \ldots, \widehat{e_i}, \ldots, e_p).$$

The *boundary map* $\partial_p\colon S_p(X) \to S_{p-1}(X)$ is the unique homomorphism extending ∂_p on singular p-simplices. For $p \leq 0$, ∂_p is the null homomorphism. Given a continuous map $f\colon X \to Y$ between two topological spaces X and Y, the homomorphism $f_{\sharp,p}\colon S_p(X) \to S_p(Y)$ is defined such that

$$f_{\sharp,p}(T) = f \circ T,$$

for every singular p-simplex $T\colon \Delta_p \to X$.

The next easy proposition gives the main properties of ∂.

Proposition 4.2.8 *For every continuous map $f\colon X \to Y$ between two topological spaces X and Y, the maps $f_{\sharp,p}$ and ∂_p commute for every p, i.e.,*

$$\partial_p \circ f_{\sharp,p} = f_{\sharp,p-1} \circ \partial_p.$$

We also have $\partial_{p-1} \circ \partial_p = 0$.

Proof. For any singular p-simplex $T\colon \Delta_p \to X$, we have

$$\partial_p f_{\sharp,p}(T) = \sum_{i=0}^{p} (-1)^i \, (f \circ T) \circ l(e_0, \ldots, \widehat{e_i}, \ldots, e_p),$$

and

$$f_{\sharp,p-1}(\partial_p T) = \sum_{i=0}^{p} (-1)^i \, f \circ (T \circ l(e_0, \ldots, \widehat{e_i}, \ldots, e_p)),$$

and the equality follows by associativity of composition. We also have

$$\begin{aligned}
\partial_p l(a_0, \ldots, a_p) &= \sum_{i=0}^{p} (-1)^i \, l(a_0, \ldots, a_p) \circ l(e_0, \ldots, \widehat{e_i}, \ldots, e_p) \\
&= \sum_{i=0}^{p} (-1)^i \, l(a_0, \ldots, \widehat{a_i}, \ldots, a_p),
\end{aligned}$$

since the composition of affine maps is affine. Then, we can compute $\partial_{p-1}\partial_p l(a_0, \ldots, a_p)$ as we did in Proposition 4.2.4, and the proof is similar, except that we have to insert an l at appropriate places. The rest of the proof follows from the fact that

$$\partial_{p-1}\partial_p T = \partial_{p-1}\partial_p(T_\sharp(l(e_0, \ldots, e_p))),$$

since $l(e_0, \ldots, e_p)$ is simply the inclusion of Δ_p in $\mathbb{R}^{(\mathbb{N})}$, and that ∂ commutes with $T_\sharp$. $\square$

In view of Proposition 4.2.8, the image $\partial_{p+1}(S_{p+1}(X))$ of $\partial_{p+1}\colon S_{p+1}(X) \to S_p(X)$ is a subgroup of the kernel $\partial_p^{-1}(0)$ of $\partial_p\colon S_p(X) \to S_{p-1}(X)$. This motivates the following definition.

Definition 4.2.9 Given a topological space X, the kernel $\partial_p^{-1}(0)$ of the homomorphism $\partial_p\colon S_p(X) \to S_{p-1}(X)$ is denoted as $Z_p(X)$, and the elements of $Z_p(X)$ are called *singular p-cycles*. The image $\partial_{p+1}(S_{p+1})$ of the homomorphism $\partial_{p+1}\colon S_{p+1}(X) \to S_p(X)$ is denoted as $B_p(X)$, and the elements of $B_p(X)$ are called *singular p-boundaries*. The *p-th singular homology group* $H_p(X)$ is the quotient group

$$H_p(X) = Z_p(X)/B_p(X).$$

If $f\colon X \to Y$ is a continuous map, the fact that

$$\partial_p \circ f_{\sharp,p} = f_{\sharp,p-1} \circ \partial_p$$

allows us to define homomorphisms $f_{*,p}\colon H_p(X) \to H_p(Y)$, and it it easily verified that

$$(g \circ f)_{*,p} = g_{*,p} \circ f_{*,p},$$

and that $Id_{*,p}\colon H_p(X) \to H_p(Y)$ is the identity homomorphism, when $Id\colon X \to Y$ is the identity. As a corollary, if $f\colon X \to Y$ is a homeomorphism, then each $f_{*,p}\colon H_p(X) \to H_p(Y)$ is a group isomorphism. This gives us a way of showing that two spaces are not homeomorphic, by showing that some homology groups $H_p(X)$ and $H_p(Y)$ are not isomorphic.

It is fairly easy to show that $H_0(X)$ is a free abelian group, and that if the path components of X are the family $(X_i)_{i\in I}$, then $H_0(X)$ is isomorphic to the direct sum $\bigoplus_{i\in I} \mathbb{Z}$. In particular, if X is arcwise connected, $H_0(X) = \mathbb{Z}$.

The following important theorem shows the relationship between simplicial homology and singular homology. The proof is fairly involved, and can be found in Munkres [14], or Rotman [15].

Theorem 4.2.10 *Given any polytope X, if $X = K_g = K'_g$ is the geometric realization of any two complexes K and K', then*

$$H_p(X) = H_p(K) = H_p(K'),$$

for all $p \geq 0$.

Theorem 4.2.10 implies that $H_p(X)$ is finitely generated for all $p \geq 0$. It is immediate that if K has dimension m, then $H_p(X) = 0$ for $p > m$, and it can be shown that $H_m(X)$ is a free abelian group.

A fundamental invariant of finite complexes is the Euler-Poincaré characteristic.

Definition 4.2.11 Given a finite complex $K = (V, \mathcal{S})$ of dimension m, letting m_p be the number of p-simplices in K, we define the *Euler-Poincaré characteristic* $\chi(K)$ *of* K as

$$\chi(K) = \sum_{p=0}^{m} (-1)^p \, m_p.$$

The following remarkable theorem holds.

Theorem 4.2.12 *Given a finite complex* $K = (V, \mathcal{S})$ *of dimension* m*, we have*

$$\chi(K) = \sum_{p=0}^{m} (-1)^p \, r(H_p(K)),$$

the alternating sum of the Betti numbers (the ranks) of the homology groups of K*.*

Proof. We know that $C_p(K)$ is a free group of rank m_p. Since $H_p(K) = Z_p(K)/B_p(K)$, by Proposition 4.1.3, we have

$$r(H_p(K)) = r(Z_p(K)) - r(B_p(K)).$$

Since we have a short exact sequence

$$0 \longrightarrow Z_p(K) \longrightarrow C_p(K) \xrightarrow{\partial_p} B_{p-1}(K) \longrightarrow 0,$$

again, by Proposition 4.1.3, we have

$$r(C_p(K)) = m_p = r(Z_p(K)) + r(B_{p-1}(K)).$$

Also, note that $B_m(K) = 0$, and $B_{-1}(K) = 0$. Then, we have

$$\begin{aligned}
\chi(K) &= \sum_{p=0}^{m} (-1)^p \, m_p \\
&= \sum_{p=0}^{m} (-1)^p \, (r(Z_p(K)) + r(B_{p-1}(K))) \\
&= \sum_{p=0}^{m} (-1)^p \, r(Z_p(K)) + \sum_{p=0}^{m} (-1)^p \, r(B_{p-1}(K)).
\end{aligned}$$

Using the fact that $B_m(K) = 0$, and $B_{-1}(K) = 0$, we get

$$\begin{aligned}
\chi(K) &= \sum_{p=0}^{m} (-1)^p \, r(Z_p(K)) + \sum_{p=0}^{m} (-1)^{p+1} \, r(B_p(K)) \\
&= \sum_{p=0}^{m} (-1)^p \, (r(Z_p(K)) - r(B_p(K))) \\
&= \sum_{p=0}^{m} (-1)^p \, r(H_p(K)).
\end{aligned}$$

□

A striking corollary of Theorem 4.2.12 (together with Theorem 4.2.10), is that the Euler-Poincaré characteristic $\chi(K)$ of a complex of finite dimension m only depends on the geometric realization K_g of K, since it only depends on the homology groups $H_p(K) = H_p(K_g)$ of the polytope K_g. Thus, the Euler-Poincaré characteristic is an invariant of all the finite complexes corresponding to the same polytope $X = K_g$, and we can say that it is the Euler-Poincaré characteristic of the polytope $X = K_g$, and denote it as $\chi(X)$. In particular, this is true of surfaces that admit a triangulation, and as we shall see shortly, the Euler-Poincaré characteristic in one of the major ingredients in the classification of the compact surfaces. In this case, $\chi(K) = m_0 - m_1 + m_2$, where m_0 is the number of vertices, m_1 the number of edges, and m_2 the number of triangles, in K. We warn the reader that Ahlfors and Sario have flipped the signs, and define the Euler-Poincaré characteristic as $-m_0 + m_1 - m_2$.

Going back to the triangulations of the sphere, the torus, the projective space, and the Klein bottle, it is easy to see that their Euler-Poincaré characteristic is 2 (sphere), 0 (torus), 1 (projective space), and 0 (Klein bottle).

At this point, we are ready to compute the homology groups of finite (two-dimensional) polyhedras.

4.3 Homology Groups of the Finite Polyhedras

Since a polyhedron is the geometric realization of a triangulated 2-complex, it is possible to determine the homology groups of the (finite) polyhedras. We say that a triangulated 2-complex K is orientable if its geometric realization K_g is orientable. We will consider the finite, bordered, orientable, and nonorientable, triangulated 2-complexes. First, note that $C_p(K)$ is the trivial group for $p < 0$ and $p > 2$, and thus, we just have to consider the cases where $p = 0, 1, 2$. We will use the notation $c \sim c'$, to denote that two p-chains are homologous, which means that $c = c' + \partial_{p+1} d$, for some $(p+1)$-chain d.

The first proposition is very easy, and is just a special case of the fact that $H_0(X) = \mathbb{Z}$ for an arcwise connected space X.

Proposition 4.3.1 *For every triangulated 2-complex (finite or not) K, we have $H_0(K) = \mathbb{Z}$.*

Proof. When $p = 0$, we have $Z_0(K) = C_0(K)$, and thus, $H_0(K) = C_0(K)/B_0(K)$. Thus, we have to figure out what the 0-boundaries are. If $c = \sum x_i \partial a_i$ is a 0-boundary, each a_i is an oriented edge $[\alpha_i, \beta_i]$, and we have

$$c = \sum x_i \partial a_i = \sum x_i \beta_i - \sum x_i \alpha_i,$$

which shows that the sum of all the coefficients of the vertices is 0. Thus, it is impossible for a 0-chain of the form $x\alpha$, where $x \neq 0$, to be homologous to 0. On the other hand, we claim that $\alpha \sim \beta$ for any two vertices α, β. Indeed, since we assumed that K is connected, there is a path from α to β consisting of edges

$$[\alpha, \alpha_1], \ldots, [\alpha_n, \beta],$$

and the 1-chain

$$c = [\alpha, \alpha_1] + \ldots + [\alpha_n, \beta]$$

has boundary

$$\partial c = \beta - \alpha,$$

which shows that $\alpha \sim \beta$. But then, $H_0(K)$ is the infinite cyclic group generated by any vertex. □

Next, we determine the groups $H_2(K)$.

Proposition 4.3.2 *For every triangulated 2-complex (finite or not) K, either $H_2(K) = \mathbb{Z}$ or $H_2(K) = 0$. Furthermore, $H_2(K) = \mathbb{Z}$ iff K is finite, has no border and is orientable, else $H_2(K) = 0$.*

Proof. When $p = 2$, we have $B_2(K) = 0$, and $H_2(K) = Z_2(K)$. Thus, we have to figure out what the 2-cycles are. Consider a 2-chain $c = \sum x_i A_i$, where each A_i is an oriented triangle $[\alpha_0, \alpha_1, \alpha_2]$, and assume that c is a cycle, which means that

$$\partial c = \sum x_i \partial A_i = 0.$$

Whenever A_i and A_j have an edge a in common, the contribution of a to ∂c is either $x_i a + x_j a$, or $x_i a - x_j a$, or $-x_i a + x_j a$, or $-x_i a - x_j a$, which implies that $x_i = \epsilon x_j$, with $\epsilon = \pm 1$. Consequently, if A_i and A_j are joined by a path of pairwise adjacent triangles, A_k, all in c, then $|x_i| = |x_j|$. However, Proposition 2.2.3 and Proposition 3.6.2 imply that any two triangles A_i and A_j in K are connected by a sequence of pairwise adjacent triangles. If some triangle in the path does not belong to c, then there are two adjacent triangles in the path, A_h and A_k, with A_h in c and A_k not in c such that all the triangles in the path from A_i to A_h belong to c. But then, A_h has an edge not adjacent to any other triangle in c, so $x_h = 0$ and thus, $x_i = 0$. The same reasoning applied to A_j shows that $x_j = 0$. If all triangles in the path from A_i to A_j belong to c, then we already know that $|x_i| = |x_j|$. Therefore, all x_i's have the same absolute value. If K is infinite, there must be some A_i in the finite sum which is adjacent to some triangle A_j not in the finite sum, and the contribution of the edge common to A_i and A_j to ∂c must be zero, which implies that $x_i = 0$ for all i. Similarly, the coefficient of every triangle with an edge in the border must be zero. Thus, in these cases, $c \sim 0$, and $H_2(K) = 0$.

Let us now assume that K is a finite triangulated 2-complex without a border. The above reasoning showed that any nonzero 2-cycle, c, can be written as

$$c = \sum \epsilon_i x A_i,$$

where $x = |x_i| > 0$ for all i, and $\epsilon_i = \pm 1$. Since $\partial c = 0$, $\sum \epsilon_i A_i$ is also a 2-cycle. For any other nonzero 2-cycle, $\sum y_i A_i$, we can subtract $\epsilon_1 y_1(\sum \epsilon_i A_i)$ from $\sum y_i A_i$, and we get the cycle

$$\sum_{i \neq 1}(y_i - \epsilon_1 \epsilon_i y_1)A_i,$$

in which A_1 has coefficient 0. But then, since all the coefficients have the same absolute value, we must have $y_i = \epsilon_1 \epsilon_i y_1$ for all $i \neq 1$, and thus,

$$\sum y_i A_i = \epsilon_1 y_1(\sum \epsilon_i A_i).$$

This shows that either $H_2(K) = 0$, or $H_2(K) = \mathbb{Z}$.

It remains to prove that K is orientable iff $H_2(K) = \mathbb{Z}$. The idea is that in this case, we can choose an orientation such that $\sum A_i$ is a 2-cycle. The proof is not really difficult, but a little involved, and the reader is referred to Ahlfors and Sario [1] for details. □

Finally, we need to determine $H_1(K)$. We will only do so for finite triangulated 2-complexes, and refer the reader to Ahlfors and Sario [1] for the infinite case.

Proposition 4.3.3 *For every finite triangulated 2-complex K, either $H_1(K) = \mathbb{Z}^{m_1}$, or $H_1(K) = \mathbb{Z}^{m_1} \oplus \mathbb{Z}/2\mathbb{Z}$, the second case occurring iff K has no border and is nonorientable.*

Proof. The first step is to determine the torsion subgroup of $H_1(K)$. Let c be a 1-cycle, and assume that $mc \sim 0$ for some $m > 0$, i.e., there is some 2-chain $\sum x_i A_i$ such that $mc = \sum x_i \partial A_i$. If A_i and A_j have a common edge a, the contribution of a to $\sum x_i \partial A_i$ is either $x_i a + x_j a$, or $x_i a - x_j a$, or $-x_i a + x_j a$, or $-x_i a - x_j a$, which implies that either $x_i \equiv x_j (\mathrm{mod}\, m)$, or $x_i \equiv -x_j (\mathrm{mod}\, m)$. Because of the connectedness of K, the above actually holds for all i, j. If K is bordered, there is some A_i which contains a border edge not adjacent to any other triangle, and thus x_i must be divisible by m, which implies that every x_i is divisible by m. Thus, $c \sim 0$. Note that a similar reasoning applies when K is infinite, but we are not considering this case. If K has no border and is orientable, by a previous remark, we can assume that $\sum A_i$ is a cycle. Then, $\sum \partial A_i = 0$, and we can write

$$mc = \sum (x_i - x_1)\partial A_i.$$

Due to the connectness of K, the above argument shows that every $x_i - x_1$ is divisible by m, which shows that $c \sim 0$. Thus, the torsion group is 0.

Let us now assume that K has no border and is nonorientable. Then, by a previous remark, there are no 2-cycles except 0. Thus, the coefficients in $\sum \partial A_i$ must be either 0 or ± 2. Let $\sum \partial A_i = 2z$. Then, $2z \sim 0$, but z is not homologous to 0, since from $z = \sum x_i \partial A_i$, we would get $\sum (2x_i - 1)\partial A_i \sim 0$, contrary to the fact that there are no 2-cycles except 0. Thus, z is of order 2.

Consider again $mc = \sum x_i \partial A_i$. Since $x_i \equiv x_j (\operatorname{mod} m)$, or $x_i \equiv -x_j (\operatorname{mod} m)$, for all i, j, we can write

$$mc = x_1 \sum \epsilon_i \partial A_i + m \sum t_i \partial A_i,$$

with $\epsilon_i = \pm 1$, and at least some coefficient of $\sum \epsilon_i \partial A_i$ is ± 2, since otherwise $\sum \epsilon_i A_i$ would be a nonnull 2-cycle. But then, $2x_1$ is divisible by m, and this implies that $2c \sim 0$. If $2c = \sum u_i \partial A_i$, the u_i are either all odd or all even. If they are all even, we get $c \sim 0$, and if they are all odd, we get $c \sim z$. Hence, z is the only element of finite order, and the torsion group if $\mathbb{Z}/2\mathbb{Z}$.

Finally, having determined the torsion group of $H_1(K)$, by the corollary of Proposition 4.1.2, we know that $H_1(K) = \mathbb{Z}^{m_1} \oplus T$, where m_1 is the rank of $H_1(K)$, and the proposition follows. □

Recalling Proposition 4.2.12, the Euler-Poincaré characteristic $\chi(K)$ is given by

$$\chi(K) = r(H_0(K)) - r(H_1(K)) + r(H_2(K)),$$

and we have determined that $r(H_0(K)) = 1$ and either $r(H_2(K)) = 0$ when K has a border or has no border and is nonorientable, or $r(H_2(K)) = 1$ when K has no border and is orientable.

Thus, the rank m_1 of $H_1(K)$ is either

$$m_1 = 2 - \chi(K)$$

if K has no border and is orientable, and

$$m_1 = 1 - \chi(K)$$

otherwise. This implies that $\chi(K) \leq 2$.

We will now prove the classification theorem for compact (two-dimensional) polyhedras.

Chapter 5

The Classification Theorem for Compact Surfaces

5.1 Cell Complexes

It is remarkable that the compact (two-dimensional) polyhedras can be characterized up to homeomorphism. This situation is exceptional, as such a result is known to be essentially impossible for compact m-manifolds for $m \geq 4$, and still open for compact 3-manifolds. In fact, it is possible to characterize the compact (two-dimensional) polyhedras in terms of a simple extension of the notion of a complex, called cell complex by Ahlfors and Sario. What happens is that it is possible to define an equivalence relation on cell complexes, and it can be shown that every cell complex is equivalent to some specific normal form. Furthermore, every cell complex has a geometric realization which is a surface, and equivalent cell complexes have homeomorphic geometric realizations. Also, every cell complex is equivalent to a triangulated 2-complex. Finally, we can show that the geometric realizations of distinct normal forms are not homeomorphic.

The classification theorem for compact surfaces is presented (in slightly different ways) in Massey [11] Amstrong [2], and Kinsey [9]. In the above references, the presentation is sometimes quite informal. The classification theorem is also presented in Ahlfors and Sario [1], and there, the presentation is formal and not always easy to follow. We tried to strike a middle ground in the degree of formality. It should be noted that the combinatorial part of the proof (Section 5.2) is heavily inspired by the proof given in Seifert and Threlfall [18]. One should also take a look at Chapter 1 of Thurston [20], especially Problem 1.3.12. Thurston's book is also highly recommended as a wonderful and insighful introduction to the topology and geometry of three-dimensional manifolds, but that's another story.

The first step is to define cell complexes. The intuitive idea is to generalize a little bit the notion of a triangulation, and consider objects made of oriented faces, each face having some boundary. A boundary is a cyclically ordered list of oriented edges. We can think of each face as a circular closed disk, and of the edges in a boundary as circular arcs on the boundaries of these disks. A cell complex represents the surface obtained by identifying identical boundary edges.

Technically, in order to deal with the notion of orientation, given any set X, it is convenient to introduce the set $X^{-1} = \{x^{-1} \mid x \in X\}$ of formal inverses of elements in X. We will say that the elements of $X \cup X^{-1}$ are *oriented*. It is also convenient to assume that $(x^{-1})^{-1} = x$, for every $x \in X$. It turns out that cell complexes can be defined using only faces and boundaries, and that the notion of a vertex can be defined from the way edges occur in boundaries. This way of dealing with vertices is a bit counterintuitive, but we haven't found a better way to present cell complexes. We now give precise definitions.

Definition 5.1.1 A *cell complex* K consists of a triple $K = (F, E, B)$, where F is a finite nonempty set of *faces*, E is a finite set of *edges*, and $B\colon (F \cup F^{-1}) \to (E \cup E^{-1})^*$ is the *boundary function*, which assigns to each oriented face $A \in F \cup F^{-1}$ a cyclically ordered sequence $a_1 \ldots a_n$ of oriented edges in $E \cup E^{-1}$, the *boundary of* A, in such a way that $B(A^{-1}) = a_n^{-1} \ldots a_1^{-1}$ (the reversal of the sequence $a_1^{-1} \ldots a_n^{-1}$). By a cyclically ordered sequence, we mean that we do not distinguish between the sequence $a_1 \ldots a_n$ and any sequence obtained from it by a cyclic permutation. In particular, the successor of a_n is a_1. Furthermore, the following conditions must hold:

(1) Every oriented edge $a \in E \cup E^{-1}$ occurs either once or twice as an element of a boundary. In particular, this means that if a occurs twice in some boundary, then it does not occur in any other boundary.

(2) K is connected. This means that K is not the union of two disjoint systems satisfying the conditions.

It is possible that $F = \{A\}$ and $E = \emptyset$, in which case $B(A) = B(A^{-1}) = \epsilon$, the empty sequence.

For short, we will often say face and edge, rather than oriented face or oriented edge. As we said earlier, the notion of a vertex is defined in terms of faces and boundaries. The intuition is that a vertex is adjacent to pairs of incoming and outgoing edges. Using inverses of edges, we can define a vertex as the sequence of incoming edges into that vertex. When the vertex is not a boundary vertex, these edges form a cyclic sequence, and when the vertex is a border vertex, such a sequence has two endpoints with no successors.

Definition 5.1.2 Given a cell complex $K = (F, E, B)$, for any edge $a \in E \cup E^{-1}$, a *successor* of a is an edge b such that b is the successor of a in some boundary $B(A)$. If a occurs in two places in the set of boundaries, it has a *a pair of successors* (possibly identical), and otherwise it has a *single successor*. A cyclically ordered sequence $\alpha = (a_1, \ldots, a_n)$ is called an *inner vertex* if every a_i has a_{i-1}^{-1} and a_{i+1}^{-1} as pair of successors (note that a_1 has a_n^{-1} and a_2^{-1} as pair of successors, and a_n has a_{n-1}^{-1} and a_1^{-1} as pair of successors). A *border vertex* is a cyclically ordered sequence $\alpha = (a_1, \ldots, a_n)$ such that the above condition holds for all i,

$2 \leq i \leq n-1$, while a_1 has a_2^{-1} as only successor, and a_n has a_{n-1}^{-1} as only successor. An edge $a \in E \cup E^{-1}$ is a *border edge* if it occurs once in a single boundary, and otherwise an *inner edge*.

Given any edge $a \in E \cup E^{-1}$, we can determine a unique vertex α as follows: the neighbors of a in the vertex α are the inverses of its successor(s). Repeat this step in both directions until either the cycle closes, or we hit sides with only one successor. The vertex α in question is the list of the incoming edges into it. For this reason, we say that *a leads to α*. Note that when a vertex $\alpha = (a)$ contains a single edge a, there must be an occurrence of the form aa^{-1} in some boundary. Also, note that if (a, a^{-1}) is a vertex, then it is an inner vertex, and if (a, b^{-1}) is a vertex with $a \neq b$, then it is a border vertex.

Vertices can also characterized in another way which will be useful later on. Intuitively, two edges a and b are equivalent iff they have the same terminal vertex.

We define a relation λ on edges as follows: $a\lambda b$ iff b^{-1} is the successor of a in some boundary. Note that this relation is symmetric. Indeed, if ab^{-1} appears in the boundary of some face A, then ba^{-1} appears in the boundary of A^{-1}. Let Λ be the reflexive and transitive closure of λ. Since λ is symmetric, Λ is an equivalence relation. We leave as an easy exercise to prove that the equivalence class of an edge a is the vertex α that a leads to. Thus, vertices induce a partition of $E \cup E^{-1}$. We say that an edge a is an edge from a vertex α to a vertex β if $a^{-1} \in \alpha$ and $a \in \beta$. Then, by a familiar reasoning, we can show that the fact that K is connected implies that there is a path between any two vertices.

Figure 5.1 shows a cell complex with border. The cell complex has three faces with boundaries abc, bed^{-1}, and adf^{-1}. It has one inner vertex $b^{-1}ad^{-1}$ and three border vertices edf, $c^{-1}be^{-1}$, and $ca^{-1}f^{-1}$.

If we fold the above cell complex by identifying the two edges labeled d, we get a tetrahedron with one face omitted, the face opposite the inner vertex, the endpoint of edge a.

There is a natural way to view a triangulated complex as a cell complex, and it is not hard to see that the following conditions allow us to view a cell complex as a triangulated complex.

(C1) If a, b are distinct edges leading to the same vertex, then a^{-1} and b^{-1} lead to distinct vertices.

(C2) The boundary of every face is a triple abc.

(C3) Different faces have different boundaries.

It is easy to see that a and a^{-1} cannot lead to the same vertex, and that in a face abc, the edges a, b, c are distinct.

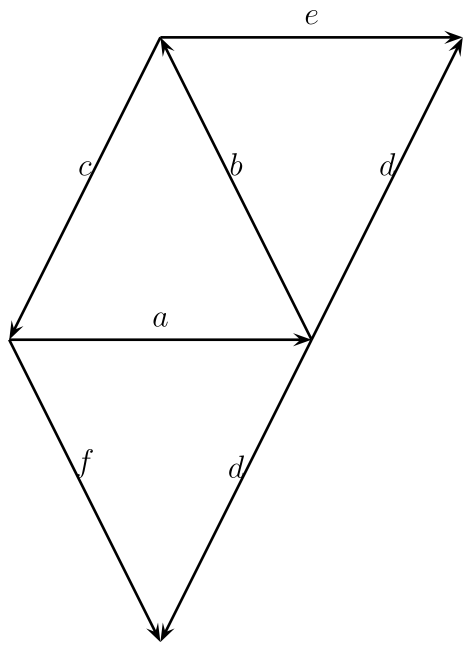

Figure 5.1: A cell complex with border

5.2 Normal Form for Cell Complexes

We now introduce a notion of elementary subdivision of cell complexes which is crucial in obtaining the classification theorem.

Definition 5.2.1 Given any two cells complexes K and K', we say that K' *is an elementary subdivision of* K if K' is obtained from K by one of the following two operations:

(P1) Any two edges a and a^{-1} in K are replaced by bc and $c^{-1}b^{-1}$ in all boundaries, where b, c are distinct edges of K' not in K.

(P2) Any face A in K with boundary $a_1 \ldots a_p a_{p+1} \ldots a_n$ is replaced by two faces A' and A'' in K', with boundaries $a_1 \ldots a_p d$ and $d^{-1} a_{p+1} \ldots a_n$, where d is an edge in K' not in K. Of course, the corresponding replacement is applied to A^{-1}.

We say that a cell complex K' is a *refinement* of a cell complex K if K and K' are related in the reflexive and transitive closure of the elementary subdivision relation, and we say that K and K' are *equivalent* if they are related in the least equivalence relation containing the elementary subdivision relation.

As we will see shortly, every cell complex is equivalent to some special cell complex in normal form. First, we show that a topological space $|K|$ can be associated with a cell complex K, that this space is the same for all cell complexes equivalent to K, and that it is a surface.

Given a cell complex K, we associate with K a topological space $|K|$ as follows. Let us first assume that no face has the empty sequence as a boundary. Then, we assign to each face A a circular disk, and if the boundary of A is $a_1 \ldots a_m$, we divide the boundary of the disk into m oriented arcs. These arcs, in clockwise order are named $a_1 \ldots a_m$, while the opposite arcs are named $a_1^{-1} \ldots a_m^{-1}$. We then form the quotient space obtained by identifying arcs having the same name in the various disks (this requires using homeomorphisms between arcs named identically, etc).

We leave as an exercise to prove that equivalent cell complexes are mapped to homeomorphic spaces, and that if K represents a triangulated complex, then $|K|$ is homeomorphic to K_g.

When K has a single face A with the null boundary, by (P2), K is equivalent to the cell complex with two faces A', A'', where A' has boundary d, and A'' has boundary d^{-1}. In this case, $|K|$ must be homeomorphic to a sphere.

In order to show that the space $|K|$ associated with a cell complex is a surface, we prove that every cell complex can be refined to a triangulated 2-complex.

Proposition 5.2.2 *Every cell complex K can be refined to a triangulated 2-complex.*

Proof. Details are given in Ahlfors and Sario [1], and we only indicate the main steps. The idea is to subdivide the cell complex by adding new edges. Informally, it is helpful to view the process as adding new vertices and new edges, but since vertices are not primitive objects, this must be done via the refinement operations (P1) and (P2). The first step is to split every edge a into two edges b and c where $b \neq c$, using (P1), introducing new border vertices (b, c^{-1}). The effect is that for every edge a (old or new), a and a^{-1} lead to distinct vertices. Then, for every boundary $B = a_1 \ldots a_n$, we have $n \geq 2$, and intuitively, we create a "central vertex" $\beta = (d_1, \ldots, d_n)$, and we join this vertex β to every vertex including the newly created vertices (except β itself). This is done as follows: first, using (P2), split the boundary $B = a_1 \ldots a_n$ into $a_1 d$ and $d^{-1} a_2 \ldots a_n$, and then using (P1), split d into $d_1 d_n^{-1}$, getting boundaries $d_n^{-1} a_1 d_1$ and $d_1^{-1} a_2 \ldots a_n d_n$. Applying (P2) to the boundary $d_1^{-1} a_2 \ldots a_n d_n$, we get the boundaries $d_1^{-1} a_2 d_2$, $d_2^{-1} a_3 d_3, \ldots$, $d_{n-1}^{-1} a_n d_n$, and $\beta = (d_1, \ldots, d_n)$ is indeed an inner vertex. At the end of this step, it is easy to verify that (C2) and (C3) are satisfied, but (C1) may not. Finally, we split each new triangular boundary $a_1 a_2 a_3$ into four subtriangles, by joining the middles of its three sides. This is done by getting $b_1 c_1 b_2 c_2 b_3 c_3$, using (P1), and then $c_1 b_2 d_3$, $c_2 b_3 d_1$, $c_3 b_1 d_2$, and $d_1^{-1} d_2^{-1} d_3^{-1}$, using (P2). The resulting cell complex also satisfies (C1), and in fact, what we have done is to provide a triangulation. □

Next, we need to define cell complexes in normal form. First, we need to define what we mean by orientability of a cell complex, and to explain how we compute its Euler-Poincaré characteristic.

Definition 5.2.3 Given a cell complex $K = (F, E, B)$, an *orientation of* K is the choice of one of the two oriented faces A, A^{-1} for every face $A \in F$. An orientation is *coherent* if for every edge a, if a occurs twice in the boundaries, then a occurs in the boundary of a face A_1 and in the boundary of a face A_2^{-1}, where $A_1 \neq A_2$. A cell complex K is *orientable* if is has some coherent orientation. A *contour* of a cell complex is a cyclically ordered sequence $(a_1, \ldots, a_n)$ of edges such that a_i and a_{i+1}^{-1} lead to the same vertex, and the a_i belong to a single boundary.

It is easily seen that equivalence of cell complexes preserves orientability. In counting contours, we do not distinguish between $(a_1, \ldots, a_n)$ and $(a_n^{-1}, \ldots, a_1^{-1})$. It is easily verified that (P1) and (P2) do not change the number of contours.

Given a cell complex $K = (F, E, B)$, the number of vertices is denoted as n_0, the number n_1 of edges is the number of elements in E, and the number n_2 of faces is the number of elements in F. The Euler-Poincaré characteristic of K is $n_0 - n_1 + n_2$. It is easily seen that (P1) increases n_1 by 1, creates one more vertex, and leaves n_2 unchanged. Also, (P2) increases n_1 and n_2 by 1 and leaves n_0 unchanged. Thus, equivalence preserves the Euler-Poincaré characteristic. However, we need a small adjustment in the case where K has a single face A with the null boundary. In this case, we agree that K has the "null vertex " ϵ. We now define the normal forms of cell complexes. As we shall see, these normal forms have a single face and a single inner vertex.

Definition 5.2.4 A *cell complex in normal form, or canonical cell complex*, is a cell complex $K = (F, E, B)$, where $F = \{A\}$ is a singleton set, and either

(I) $E = \{a_1, \ldots, a_p, b_1, \ldots, b_p, c_1, \ldots, c_q, h_1, \ldots, h_q\}$, and

$$B(A) = a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_p b_p a_p^{-1} b_p^{-1} c_1 h_1 c_1^{-1} \cdots c_q h_q c_q^{-1},$$

where $p \geq 0$, $q \geq 0$, or

(II) $E = \{a_1, \ldots, a_p, c_1, \ldots, c_q, h_1, \ldots, h_q\}$, and

$$B(A) = a_1 a_1 \cdots a_p a_p c_1 h_1 c_1^{-1} \cdots c_q h_q c_q^{-1},$$

where $p \geq 1$, $q \geq 0$.

Observe that canonical complexes of type (I) are orientable, whereas canonical complexes of type (II) are not. The sequences $c_i h_i c_i^{-1}$ yield q border vertices (h_i, c_i, h_i^{-1}), and thus q contours (h_i), and in case (I), the single inner vertex

$$(a_1^{-1}, b_1, a_1, b_1^{-1} \ldots, a_p^{-1}, b_p, a_p, b_p^{-1}, c_1^{-1}, \ldots, c_q^{-1}),$$

and in case (II), the single inner vertex

$$(a_1^{-1}, a_1, \ldots, a_p^{-1}, a_p, c_1^{-1}, \ldots, c_q^{-1}).$$

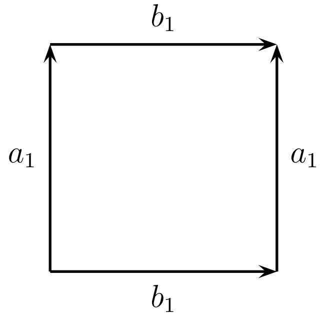

Figure 5.2: A cell complex corresponding to a torus

Thus, in case (I), there are $q+1$ vertices, $2p+2q$ sides, and one face, and the Euler-Poincaré characteristic is $q+1-(2p+2q)+1 = 2-2p-q$, that is

$$\chi(K) = 2-2p-q,$$

and in case (II), there are $q+1$ vertices, $p+2q$ sides, and one face, and the Euler-Poincaré characteristic is $q+1-(p+2q)+1 = 2-p-q$, that is

$$\chi(K) = 2-p-q.$$

Note that when $p = q = 0$, we do get $\chi(K) = 2$, which agrees with the fact that in this case, we assumed the existence of a null vertex, and there is one face. This is the case of the sphere.

The above shows that distinct canonical complexes K_1 and K_2 are inequivalent, since otherwise $|K_1|$ and $|K_2|$ would be homeomorphic, which would imply that K_1 and K_2 have the same number of contours, the same kind of orientability, and the same Euler-Poincaré characteristic.

It remains to prove that every cell complex is equivalent to a canonical cell complex, but first, it is helpful to give more intuition regarding the nature of the canonical complexes.

If a canonical cell complex has the border $B(A) = a_1b_1a_1^{-1}b_1^{-1}$, we can think of the face A as a square whose opposite edges are oriented the same way, and labeled the same way, so that by identification of the opposite edges labeled a_1 and then of the edges labeled b_1, we get a surface homeomorphic to a torus. Figure 5.2 shows such a cell complex.

If we start with a sphere and glue a torus onto the surface of the sphere by removing some small disk from both the sphere and the torus and gluing along the boundaries of the holes, it is as if we had added a handle to the sphere. For this reason, the string $a_1b_1a_1^{-1}b_1^{-1}$ is called a *handle*. A canonical cell complex with boundary $a_1b_1a_1^{-1}b_1^{-1}\cdots a_pb_pa_p^{-1}b_p^{-1}$ can be viewed as the result of attaching p handles to a sphere.

If a canonical cell complex has the border $B(A) = a_1a_1$, we can think of the face A as a circular disk whose boundary is divided into two semi-circles both labeled a_1. The corresponding surface is obtained by identifying diametrically opposed points on the boundary, and thus it is homeomorphic to the projective plane. Figure 5.3 illustrates this situation.

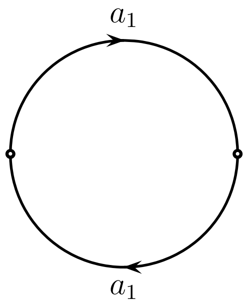

Figure 5.3: A cell complex corresponding to a projective plane

There is a way of performing such an identification resulting in a surface with self-intersection, sometimes called a *cross-cap*. A nice description of the process of getting a cross-cap is given in Hilbert and Cohn-Vossen [8]. A string of the form aa is called a *cross-cap*. Generally, a canonical cell complex with boundary $a_1a_1\cdots a_pa_p$ can be viewed as the result of forming $p\geq 1$ cross-caps, starting from a circular disk with $p-1$ circular holes, and performing the cross-cap identifications on all p boundaries, including the original disk itself.

A string of the form $c_1h_1c_1^{-1}$ occurring in a border can be interpreted as a hole with boundary h_1. For instance, if the boundary of a canonical cell complex is $c_1h_1c_1^{-1}$, splitting the face A into the two faces A' and A'' with boundaries $c_1h_1c_1^{-1}d$ and d^{-1}, we can view the face A' as a disk with boundary d in which a small circular disk has been removed. Choosing any point on the boundary d of A', we can join this point to the boundary h_1 of the small circle by an edge c_1, and we get a path $c_1h_1c_1^{-1}d$. The path is a closed loop, and a string of the form $c_1h_1c_1^{-1}$ is called a *loop*. Figure 5.4 illustrates this situation.

We now prove a combinatorial lemma which is the key to the classification of the compact surfaces. First, note that the inverse of the reduction step (P1), denoted as $(\mathrm{P1})^{-1}$, applies to a string of edges bc provided that $b\neq c$ and (b,c^{-1}) is a vertex. The result is that such a border vertex is eliminated. The inverse of the reduction step (P2), denoted as $(\mathrm{P2})^{-1}$, applies to two faces A_1 and A_2 such that $A_1\neq A_2$, $A_1\neq A_2^{-1}$, and $B(A_1)$ contains some edge d and $B(A_2)$ contains the edge d^{-1}. The result is that d (and d^{-1}) is eliminated. As a preview of the proof, we show that the following cell complex, obviously corresponding to a Möbius strip, is equivalent to the cell complex of type (II) with boundary $aachc^{-1}$. The boundary of the cell complex shown in Figure 5.5 is $abac$.

First using (P2), we split $abac$ into abd and $d^{-1}ac$. Since $abd=bda$ and the inverse face of $d^{-1}ac$ is $c^{-1}a^{-1}d=a^{-1}dc^{-1}$, by applying $(\mathrm{P2})^{-1}$, we get $bddc^{-1}=ddc^{-1}b$. We can now apply $(\mathrm{P1})^{-1}$, getting ddk. We are almost there, except that the complex with boundary ddk has no inner vertex. We can introduce one as follows. Split d into bc, getting $bcbck=cbckb$. Next, apply (P2), getting cba and $a^{-1}ckb$. Since $cba=bac$ and the inverse face of $a^{-1}ckb$ is $b^{-1}k^{-1}c^{-1}a=c^{-1}ab^{-1}k^{-1}$, by applying $(\mathrm{P2})^{-1}$ again, we get $baab^{-1}k^{-1}=aab^{-1}k^{-1}b$, which is of the form $aachc^{-1}$, with $c=b^{-1}$ and $h=k^{-1}$. Thus, the canonical cell complex

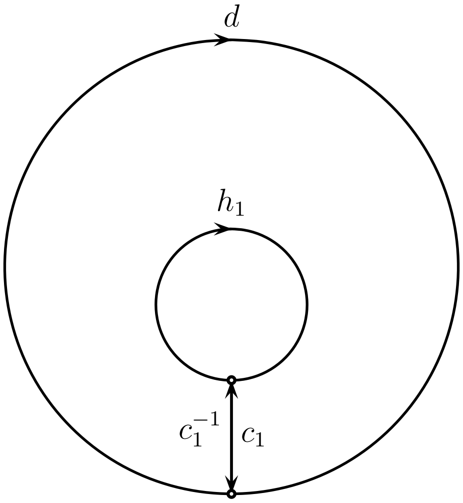

Figure 5.4: A disk with a hole

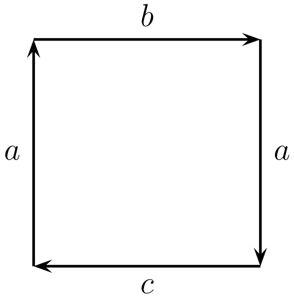

Figure 5.5: A cell complex corresponding to a Möbius strip

with boundary $aachc^{-1}$ has the Möbius strip as its geometric realization. Intuitively, this corresponds to cutting out a small circular disk in a projective plane. This process is very nicely described in Hilbert and Cohn-Vossen [8].

Lemma 5.2.5 *Every cell complex K is equivalent to some canonical cell complex.*

Proof. All the steps are given in Ahlfors and Sario [1], and in a slightly different and more informal manner in Massey [11]. We will only give the keys steps, referring the reader to the above sources for details.

The proof proceeds by steps that bring the original cell complex closer to normal form.

Step 1. Elimination of strings aa^{-1} in boundaries.

Given a boundary of the form $aa^{-1}X$, where X denotes some string of edges (possibly empty), we can use (P2) to replace $aa^{-1}X$ by the two boundaries ad and $d^{-1}a^{-1}X$, where d is new. But then, using (P1), we can contract ad to a new edge c (and $d^{-1}a^{-1}$ to c^{-1}). But now, using $(\mathrm{P2})^{-1}$, we can eliminate c. The net result is the elimination of aa^{-1}.

Step 2. Vertex Reduction.

If $p = 0, q = 0$, there is only the empty vertex, and there is nothing to do. Otherwise, the purpose of this step is to obtain a cell complex with a single inner vertex, and where border vertices correspond to loops. First, we perform step 1 until all occurrences of the form aa^{-1} have been eliminated.

Consider an inner vertex $\alpha = (b_1, \ldots, b_m)$. If b_i^{-1} also belongs to α for all i, $1 \leq i \leq m$, and there is another inner vertex β, since all vertices are connected, there is some inner vertex $\delta \neq \alpha$ directly connected to α, which means that either some b_i or b_i^{-1} belongs to δ. But since the vertices form a partition of $E \cup E^{-1}$, $\alpha = \delta$, a contradiction.

Thus, if $\alpha = (b_1, \ldots, b_m)$ is not the only inner vertex, we can assume by relabeling that b_1^{-1} does not belong to α. Also, we must have $m \geq 2$, since otherwise there would be a string $b_1b_1^{-1}$ in some boundary, contrary to the fact that we performed step 1 all the way. Thus, there is a string $b_1b_2^{-1}$ in some boundary. We claim that we can eliminate b_2. Indeed, since α is an inner vertex, b_2 must occur twice in the set of boundaries, and thus, since b_2^{-1} is a successor of b_1, there are boundaries of the form $b_1b_2^{-1}X_1$ and b_2X_2, and using (P2), we can split $b_1b_2^{-1}X_1$ into $b_1b_2^{-1}c$ and $c^{-1}X_1$, where c is new. Since b_2 differs from b_1, b_1^{-1}, c, c^{-1}, we can eliminate b_2 by $(\mathrm{P2})^{-1}$ applied to $b_2X_2 = X_2b_2$ and $b_1b_2^{-1}c = b_2^{-1}cb_1$, getting $X_2cb_1 = cb_1X_2$. This has the effect of shrinking α. Indeed, the existence of the boundary cb_1X_2 implies that c and b_1^{-1} lead to the same vertex, and the existence of the boundary $b_1b_2^{-1}c$ implies that c^{-1} and b_2^{-1} lead to the same vertex, and if b_2^{-1} does not belong to α, then b_2 is dropped, or if b_2^{-1} belongs to α, then c^{-1} is added to α, but both b_2 and b_2^{-1} are dropped.

This process can be repeated until $\alpha = (b_1)$, at which stage b_1 is eliminated using step 1. Thus, it is possible to eliminate all inner vertices except one. In the event that there was no inner vertex, we can always create one using (P1) and (P2) as in the proof of Proposition 5.2.2. Thus, from now on, we will assume that there is a single inner vertex.

We now show that border vertices can be reduced to the form (h, c, h^{-1}). The previous argument shows that we can assume that there is a single inner vertex α. A border vertex is of the form $\beta = (h, b_1, \ldots, b_m, k)$, where h, k are border edges, and the b_i are inner edges. We claim that there is some border vertex $\beta = (h, b_1, \ldots, b_m, k)$ where some b_i^{-1} belongs to the inner vertex α. Indeed, since K is connected, every border vertex is connected to α, and thus, there is a least one border vertex $\beta = (h, b_1, \ldots, b_m, k)$ directly connected to α by some edge. Observe that h^{-1} and b_1^{-1} lead to the same vertex, and similarly, b_m^{-1} and k^{-1} lead to the same vertex. Thus, if no b_i^{-1} belongs to α, either h^{-1} or k^{-1} belongs to α, which would imply that either b_1^{-1} or b_m^{-1} is in α. Thus, such an edge from β to α must be one of the b_i^{-1}. Then by the reasoning used in the case of an inner vertex, we can eliminate all b_j except b_i, and the resulting vertex is of the form (h, b_i, k). If $h \neq k^{-1}$, we can also eliminate b_i since h^{-1} does not belong to (h, b_i, k), and the vertex (h, k) can be eliminated using $(\mathrm{P1})^{-1}$.

One can verify that reducing a border vertex to the form (h, c, h^{-1}) does not undo the reductions already performed, and thus, at the end of step 2, we either obtain a cell complex with a null inner node and loop vertices, or a single inner vertex and loop vertices.

Step 3. Introduction of cross-caps.

We may still have several faces. We claim that if there are at least two faces, then for every face A, there is some face B such that $B \neq A$, $B \neq A^{-1}$, and there is some edge a both in the boundary of A and in the boundary of B. In this was not the case, there would be some face A such that for every face B such that $B \neq A$ and $B \neq A^{-1}$, every edge a in the boundary of B does not belong to the boundary of A. Then, every inner edge a occurring in the boundary of A must have both of its occurrences in the boundary of A, and of course, every border edge in the boundary of A occurs once in the boundary of A alone. But then, the cell complex consisting of the face A alone and the edges occurring in its boundary would form a proper subsystem of K, contradicting the fact that K is connected.

Thus, if there are at least two faces, from the above claim and using $(\mathrm{P2})^{-1}$, we can reduce the number of faces down to one. It it easy to check that no new vertices are introduced, and loops are unaffected. Next, if some boundary contains two occurrences of the same edge a, i.e., it is of the form $aXaY$, where X, Y denote strings of edges, with $X, Y \neq \epsilon$, we show how to make the two occurrences of a adjacent. Symbolically, we show that the following pseudo-rewrite rule is admissible:

$$aXaY \simeq bbY^{-1}X, \quad \text{or} \quad aaXY \simeq bYbX^{-1}.$$

Indeed, $aXaY$ can be split into aXb and $b^{-1}aY$, and since we also have the boundary

$$(b^{-1}aY)^{-1} = Y^{-1}a^{-1}b = a^{-1}bY^{-1},$$

together with $aXb = Xba$, we can apply $(\mathrm{P2})^{-1}$ to Xba and $a^{-1}bY^{-1}$, obtaining $XbbY^{-1} = bbY^{-1}X$, as claimed. Thus, we can introduce cross-caps.

Using the formal rule $aXaY \simeq bbY^{-1}X$ again does not alter the previous loops and cross-caps. By repeating step 3, we convert boundaries of the form $aXaY$ to boundaries with cross-caps.

Step 4. Introduction of handles.

The purpose of this step is to convert boundaries of the form $aUbVa^{-1}Xb^{-1}Y$ to boundaries $cdc^{-1}d^{-1}YXVU$ containing handles. First, we prove the pseudo-rewrite rule

$$aUVa^{-1}X \simeq bVUb^{-1}X.$$

First, we split $aUVa^{-1}X$ into $aUc = Uca$ and $c^{-1}Va^{-1}X = a^{-1}Xc^{-1}V$, and then we apply $(\mathrm{P2})^{-1}$ to Uca and $a^{-1}Xc^{-1}V$, getting $UcXc^{-1}V = c^{-1}VUcX$. Letting $b = c^{-1}$, the rule follows.

Now we apply the rule to $aUbVa^{-1}Xb^{-1}Y$, and we get

$$\begin{aligned} aUbVa^{-1}Xb^{-1}Y &\simeq a_1bVUa_1^{-1}Xb^{-1}Y \\ &\simeq a_1b_1a_1^{-1}XVUb_1^{-1}Y = a_1^{-1}XVUb_1^{-1}Ya_1b_1 \\ &\simeq a_2^{-1}b_1^{-1}YXVUa_2b_1 = a_2b_1a_2^{-1}b_1^{-1}YXVU. \end{aligned}$$

Iteration of this step preserves existing loops, cross-caps and handles.

At this point, one of the obstacle to the canonical form is that we may still have a mixture of handles and cross-caps. We now show that a handle and a cross-cap is equivalent to three cross-caps. For this, we apply the pseudo-rewrite rule $aaXY \simeq bYbX^{-1}$. We have

$$\begin{aligned} aaXbcb^{-1}c^{-1}Y &\simeq a_1b^{-1}c^{-1}Ya_1c^{-1}b^{-1}X^{-1} = b^{-1}c^{-1}Ya_1c^{-1}b^{-1}X^{-1}a_1 \\ &\simeq b_1^{-1}b_1^{-1}a_1^{-1}Xc^{-1}Ya_1c^{-1} = c^{-1}Ya_1c^{-1}b_1^{-1}b_1^{-1}a_1^{-1}X \\ &\simeq c_1^{-1}c_1^{-1}X^{-1}a_1b_1b_1Ya_1 = a_1b_1b_1Ya_1c_1^{-1}c_1^{-1}X^{-1} \\ &\simeq a_2a_2Xc_1c_1b_1b_1Y. \end{aligned}$$

At this stage, we can prove that all boundaries consist of loops, cross-caps, or handles. The details can be found in Ahlfors and Sario [1].

Finally, we have to group the loops together. This can be done using the pseudo-rewrite rule

$$aUVa^{-1}X \simeq bVUb^{-1}X.$$

Indeed, we can write

$$chc^{-1}Xdkd^{-1}Y = c^{-1}Xdkd^{-1}Ych \simeq c_1^{-1}dkd^{-1}YXc_1h = c_1hc_1^{-1}dkd^{-1}YX,$$

showing that any two loops can be brought next to each other, without altering other successions.

When all this is done, we have obtained a canonical form, and the proof is complete. □

Readers familiar with formal grammars or rewrite rules may be intrigued by the use of the "rewrite rules"

$$aXaY \simeq bbY^{-1}X$$

or

$$aUVa^{-1}X \simeq bVUb^{-1}X.$$

These rules are context-sensitive, since X and Y stand for parts of boundaries, but they also apply to objects not traditionally found in formal language theory or rewrite rule theory. Indeed, the objects being rewritten are cell complexes, which can be viewed as certain kinds of graphs. Furthermore, since boundaries are invariant under cyclic permutations, these rewrite rules apply modulo cyclic permutations, something that I have never encountered in the rewrite rule literature. Thus, it appears that a formal treatment of such rewrite rules has not been given yet, which poses an interesting challenge to researchers in the field of rewrite rule theory. For example, are such rewrite systems confluent, can normal forms be easily found?

We have already observed that identification of the edges in the boundary $aba^{-1}b^{-1}$ yields a torus. We have also noted that identification of the two edges in the boundary aa yields the projective plane. Lemma 5.2.5 implies that the cell complex consisting of a single face A and the boundary $abab^{-1}$ is equivalent to the canonical cell complex $ccbb$. This follows immediately from the pseudo-rewrite rule $aXaY \simeq bbY^{-1}X$. However, it is easily seen that identification of edges in the boundary $abab^{-1}$ yields the Klein bottle. The lemma also showed that the cell complex with boundary $aabbcc$ is equivalent to the cell complex with boundary $aabcb^{-1}c^{-1}$. Thus, intuitively, it seems that the corresponding space is a simple combination of a projective plane and a torus, or of three projective planes.

We will see shortly that there is an operation on surfaces (the connected sum) which allows us to interpret the canonical cell complexes as combinations of elementary surfaces, the sphere, the torus, and the projective plane.

5.3 Proof of the Classification Theorem

Having the key Lemma 5.2.5 at hand, we can finally prove the fundamental theorem of the classification of triangulated compact surfaces and compact bordered surfaces.

Theorem 5.3.1 *Two (two-dimensional) compact polyhedra or compact bordered polyhedra (triangulated compact surfaces or compact bordered surfaces) are homeomorphic iff they agree in character of orientability, number of contours, and Euler-Poincaré characteristic.*

Proof. If $M_1 = (K_1)_g$ and $M_2 = (K_2)_g$ are homeomorphic, we know that M_1 is orientable iff M_2 is orientable, and the restriction of the homeomorphism between M_1 and M_2 to the boundaries ∂M_1 and ∂M_2, is a homeomorphism, which implies that ∂M_1 and ∂M_2 have the same number of arcwise components, that is, the same number of contours. Also, we have stated that homeomorphic spaces have isomorphic homology groups, and by Theorem 4.2.12, they have the same Euler-Poincaré characteristic. Conversely, by Lemma 5.2.5, since any cell complex is equivalent to a canonical cell complex, the triangulated 2-complexes K_1 and K_2, viewed as cell complexes, are equivalent to canonical cell complexes C_1 and C_2. However, we know that equivalence preserves orientability, the number of contours, and the Euler-Poincaré characteristic, which implies that C_1 and C_2 are identical. But then, $M_1 = (K_1)_g$ and $M_2 = (K_2)_g$ are both homeomorphic to $|C_1| = |C_2|$. $\square$

In order to finally get a version of Theorem 5.3.1 for compact surfaces or compact bordered surfaces (not necessarily triangulated), we need to prove that every surface and every bordered surface can be triangulated. This is indeed true, but the proof is far from trivial, and it involves a strong version of the Jordan curve theorem due to Schoenflies. At this stage, we believe that our readers will be relieved if we omit this proof, and refer them once again to Alhfors and Sario [1]. It is interesting to know that 3-manifolds can be triangulated, but that Markov showed that deciding whether two triangulated 4-manifolds are homeomorphic is undecidable (1958). For the record, we state the following theorem putting all the pieces of the puzzle together.

Theorem 5.3.2 *Two compact surfaces or compact bordered surfaces are homeomorphic iff they agree in character of orientability, number of contours, and Euler-Poincaré characteristic.*

We now explain somewhat informally what is the connected sum operation, and how it can be used to interpret the canonical cell complexes. We will also indicate how the canonical cell complexes can be used to determine the fundamental groups of the compact surfaces and compact bordered surfaces.

Definition 5.3.3 Given two surfaces S_1 and S_2, their *connected sum* $S_1 \sharp S_2$ is the surface obtained by choosing two small regions D_1 and D_2 on S_1 and S_2 both homeomorphic to some disk in the plane, and letting h be a homeomorphism between the boundary circles C_1 and C_2 of D_1 and D_2, by forming the quotient space of $(S_1 - \overset{\circ}{D}_1) \cup (S_2 - \overset{\circ}{D}_2)$, by the equivalence relation defined by the relation $\{(a, h(a)) \mid a \in C_1\}$.

Intuitively, $S_1 \sharp S_2$ is formed by cutting out some small circular hole in each surface, and gluing the two surfaces along the boundaries of these holes. It can be shown that $S_1 \sharp S_2$ is a surface, and that it does not depend on the choice of D_1, D_2, and h. Also, if S_2 is a sphere, then $S_1 \sharp S_2$ is homeomorphic to S_1. It can also be shown that the Euler-Poincaré characteristic of $S_1 \sharp S_2$ is given by the formula

$$\chi(S_1 \sharp S_2) = \chi(S_1) + \chi(S_2) - 2.$$

Then, we can give an interpretation of the geometric realization of a canonical cell complex. It turns out to be the connected sum of some elementary surfaces. Ignoring borders for the time being, assume that we have two canonical cell complexes S_1 and S_2 represented by circular disks with borders

$$B_1 = a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_{p_1} b_{p_1} a_{p_1}^{-1} b_{p_1}^{-1}$$

and

$$B_2 = c_1 d_1 c_1^{-1} d_1^{-1} \cdots c_{p_2} d_{p_2} c_{p_2}^{-1} d_{p_2}^{-1}.$$

Cutting a small hole with boundary h_1 in S_1 amounts to forming the new boundary

$$B_1' = a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_{p_1} b_{p_1} a_{p_1}^{-1} b_{p_1}^{-1} h_1,$$

and similarly, cutting a small hole with boundary h_2 in S_2 amounts to forming the new boundary

$$B_2' = c_1 d_1 c_1^{-1} d_1^{-1} \cdots c_{p_2} d_{p_2} c_{p_2}^{-1} d_{p_2}^{-1} h_2^{-1}.$$

If we now glue S_1 and S_2 along h_1 and h_2 (note how we first need to reverse B_2' so that h_1 and h_2 can be glued together), we get a figure looking like two convex polygons glued together along one edge, and by deformation, we get a circular disk with boundary

$$B = a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_{p_1} b_{p_1} a_{p_1}^{-1} b_{p_1}^{-1} c_1 d_1 c_1^{-1} d_1^{-1} \cdots c_{p_2} d_{p_2} c_{p_2}^{-1} d_{p_2}^{-1}.$$

A similar reasoning applies to cell complexes of type (II).

As a consequence, the geometric realization of a cell complex of type (I) is either a sphere, or the connected sum of $p \geq 1$ tori, and the geometric realization of a cell complex of type (II) is the connected sum of $p \geq 1$ projective planes. Furthermore, the equivalence of the cell complexes consisting of a single face A and the boundaries $abab^{-1}$ and $aabb$, shows that the connected sum of two projective planes is homeomorphic to the Klein bottle. Also, the equivalence of the cell complexes with boundaries $aabbcc$ and $aabcb^{-1}c^{-1}$ shows that the connected sum of a projective plane and a torus is equivalent to the connected sum of three projective planes. Thus, we obtain another form of the classification theorem for compact surfaces.

Theorem 5.3.4 *Every orientable compact surface is homeomorphic either to a sphere or to a connected sum of tori. Every nonorientable compact surface is homeomorphic either to a projective plane, or a Klein bottle, or the connected sum of a projective plane or a Klein bottle with some tori.*

If bordered compact surfaces are considered, a similar theorem holds, but holes have to be made in the various spaces forming the connected sum. For more details, the reader is referred to Massey [11], in which it is also shown how to build models of bordered surfaces by gluing strips to a circular disk.

5.4 Application of the Main Theorem: Determining the Fundamental Groups of Compact Surfaces

We now explain briefly how the canonical forms can be used to determine the fundamental groups of the compact (bordered) surfaces. This is done in two steps. The first step consists in defining a group structure on certain closed paths in a cell complex. The second step consists in showing that this group is isomorphic to the fundamental group of $|K|$.

Given a cell complex $K = (F, E, B)$, recall that a vertex α is an equivalence class of edges, under the equivalence relation Λ induced by the relation λ defined such that, $a\lambda b$ iff b^{-1} is the successor of a in some boundary. Every inner vertex $\alpha = (b_1, \ldots, b_m)$ can be cyclically ordered such that b_i has b_{i-1}^{-1} and b_{i+1}^{-1} as successors, and for a border vertex $\alpha = (b_1, \ldots, b_m)$, the same is true for $2 \leq i \leq m-1$, but b_1 only has b_2^{-1} as successor, and b_m only has b_{m-1}^{-1} as successor. An edge from α to β is any edge $a \in \beta$ such that $a^{-1} \in \alpha$. For every edge a, we will call the vertex that a defines the *target* of a, and the vertex that a^{-1} defines the *source* of a. Clearly, a is an edge between its source and its target. We now define certain paths in a cell complex, and a notion of deformation of paths.

Definition 5.4.1 Given a cell complex $K = (F, E, B)$, a *polygon in K* is any nonempty string $a_1 \ldots a_m$ of edges such that a_i and a_{i+1}^{-1} lead to the same vertex, or equivalently, such that the target of a_i is equal to the source of a_{i+1}. The source of the path $a_1 \ldots a_m$ is the source of a_1 (i.e., the vertex that a_1^{-1} leads to), and the target of the path $a_1 \ldots a_m$ is the target of a_m (i.e., the vertex that a_m leads to). The polygon is *closed* if its source and target coincide. The product of two paths $a_1 \ldots a_m$ and $b_1 \ldots b_n$ is defined if the target of a_m is equal to the source of b_1, and is the path $a_1 \ldots a_m b_1 \ldots b_n$. Given two paths $p_1 = a_1 \ldots a_m$ and $p_2 = b_1 \ldots b_n$ with the same source and the same target, we say that p_2 *is an immediate deformation of* p_1 if p_2 is obtained from p_1 by either deleting some subsequence of the form aa^{-1}, or deleting some subsequence X which is the boundary of some face. The smallest equivalence relation containing the immediate deformation relation is called *path-homotopy*.

It is easily verified that path-homotopy is compatible with the composition of paths. Then, for any vertex α_0, the set of equivalence classes of path-homotopic polygons forms a group $\pi(K, \alpha_0)$. It is also easy to see that any two groups $\pi(K, \alpha_0)$ and $\pi(K, \alpha_1)$ are isomorphic, and that if K_1 and K_2 are equivalent cell complexes, then $\pi(K_1, \alpha_0)$ and $\pi(K_2, \alpha_0)$ are isomorphic. Thus, the group $\pi(K, \alpha_0)$ only depends on the equivalence class of the cell complex K. Furthermore, it can be proved that the group $\pi(K, \alpha_0)$ is isomorphic to the fundamental group $\pi(|K|, (\alpha_0)_g)$ associated with the geometric realization $|K|$ of K (this is proved in Ahlfors and Sario [1]). It is then possible to determine what these groups are, by considering the canonical cell complexes.

Let us first assume that there are no borders, which corresponds to $q = 0$. In this case, there is only one (inner) vertex, and all polygons are closed. For an orientable cell complex (of type (I)), the fundamental group is the group presented by the generators $\{a_1, b_1, \ldots, a_p, b_p\}$, and satisfying the single equation

$$a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_p b_p a_p^{-1} b_p^{-1} = 1.$$

When $p = 0$, it is the trivial group reduced to 1. For a nonorientable cell complex (of type (II)), the fundamental group is the group presented by the generators $\{a_1, \ldots, a_p\}$, and satisfying the single equation

$$a_1a_1 \cdots a_pa_p = 1.$$

In the presence of borders, which corresponds to $q \geq 1$, it is easy to see that the closed polygons are products of a_i, b_i, and the $d_i = c_ih_ic_i^{-1}$. For cell complexes of type (I), these generators satisfy the single equation

$$a_1b_1a_1^{-1}b_1^{-1} \cdots a_pb_pa_p^{-1}b_p^{-1}d_1 \cdots d_q = 1,$$

and for cell complexes of type (II), these generators satisfy the single equation

$$a_1a_1 \cdots a_pa_pd_1 \cdots d_q = 1.$$

Using these equations, d_q can be expressed in terms of the other generators, and we get a free group. In the orientable case, we get a free group with $2q+p-1$ generators, and in the nonorientable case, we get a free group with $p+q-1$ generators.

The above result shows that there are only two kinds of complexes having a trivial group, namely for orientable complexes for which $p = q = 0$, or $p = 0$ and $q = 1$. The corresponding (bordered) surfaces are a *sphere*, and a *closed disk* (a bordered surface). We can also figure out for which other surfaces the fundamental group is abelian. This happens in the orientable case when $p = 1$ and $q = 0$, a *torus*, or $p = 0$ and $q = 2$, an *annulus*, and in the nonorientable case when $p = 1$ and $q = 0$, a *projective plane*, or $p = 1$ and $q = 1$, a *Möbius strip*.

It is also possible to use the above results to determine the homology groups $H_1(K)$ of the (bordered) surfaces, since it can be shown that $H_1(K) = \pi(K, a)/[\pi(K, a), \pi(K, a)]$, where $[\pi(K, a), \pi(K, a)]$ is the *commutator subgroup of* $\pi(K, a)$ (see Ahlfors and Sario [1]). Recall that for any group G, the commutator subgroup is the subgroup of G generated by all elements of the form $aba^{-1}b^{-1}$ (the *commutators*). It is a normal subgroup of G, since for any $h \in G$ and any $d \in [G, G]$, we have $hdh^{-1} = (hdh^{-1}d^{-1})d$, which is also in G. Then, $G/[G, G]$ is abelian, and $[G, G]$ is the smallest subgroup of G for which $G/[G, G]$ is abelian.

Applying the above to the fundamental groups of the surfaces, in the orientable case, we see that the commutators cause a lot of cancellation, and we get the equation

$$d_1 + \cdots + d_q = 0,$$

whereas in the nonorientable case, we get the equation

$$2a_1 + \cdots + 2a_p + d_1 + \cdots + d_q = 0.$$

If $q > 0$, we can express d_q in terms of the other generators, and in the orientable case we get a free abelian group with $2p+q-1$ generators, and in the nonorientable case a free abelian group with $p+q-1$ generators. When $q = 0$, in the orientable case, we get a free abelian group with $2p$ generators, and in the nonorientable case, since we have the equation

$$2(a_1 + \cdots + a_p) = 0,$$

there is an element of order 2, and we get the direct sum of a free abelian group of order $p-1$ with $\mathbb{Z}/2\mathbb{Z}$.

Incidentally, the number p is called the *genus* of a surface. Intuitively, it counts the number of holes in the surface, which is certainly the case in the orientable case, but in the nonorientable case, it is considered that the projective plane has one hole and the Klein bottle has two holes. Of course, the genus of a surface is the number of copies of tori occurring in the canonical connected sum of the surface when orientable (which, when $p = 0$, yields the sphere), or the number of copies of projective planes occurring in the canonical connected sum of the surface when nonorientable. In terms of the Euler-Poincaré characteristic, for an orientable surface, the genus g is given by the formula

$$g = (2 - \chi - q)/2,$$

and for a nonorientable surface, the genus g is given by the formula

$$g = 2 - \chi - q,$$

where q is the number of contours.

It is rather curious that bordered surfaces, orientable or not, have free groups as fundamental groups (free abelian groups for the homology groups $H_1(K)$). It is also shown in Massey [11] that every bordered surface, orientable or not, can be embedded in $\mathbb{R}^3$. This is not the case for nonorientable surfaces (with an empty border).

Finally, we conclude with a few words about the Poincaré conjecture. We observed that the only surface which is simply connected (with a trivial fundamental group) is the sphere. Poincaré conjectured in the early 1900's that the same thing holds for compact simply-connected 3-manifolds, that is, any compact simply-connected 3-manifold is homeomorphic to the 3-sphere S^3.

This famous problem is still open! One of the fascinating aspects of the Poincaré conjecture is that one cannot hope to have a classification theory of compact 3-manifolds until it is solved (recall that 3-manifolds can be triangulated, a result of E. Moise, 1952, see Massey [11]). What makes the Poincaré conjecture even more challenging is that a generalization of it was shown to be true by Smale for $m > 4$ in 1960, and true for $m = 4$ by Michael Freedman in 1982. Good luck, and let me know if you crack it!

Chapter 6

Topological Preliminaries

6.1 Metric Spaces and Normed Vector Spaces

This Chapter provides a review of basic topological notions. For a comprehensive account, we highly recommend Munkres [13], Amstrong [2], Dixmier [4], Singer and Thorpe [19], Lang [10], or Schwartz [17]. Most spaces considered will have a topological structure given by a metric or a norm, and we first review these notions. We begin with metric spaces.

Definition 6.1.1 A *metric space* is a set E together with a function $d\colon E \times E \to \mathbb{R}_+$, called a *metric, or distance*, assigning a nonnegative real number $d(x,\, y)$ to any two points $x, y \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(D1) $d(x,\, y) = d(y,\, x)$. (symmetry)

(D2) $d(x,\, y) \geq 0$, and $d(x,\, y) = 0$ iff $x = y$. (positivity)

(D3) $d(x,\, z) \leq d(x,\, y) + d(y,\, z)$. (triangular inequality)

Geometrically, condition (D3) expresses the fact that in a triangle with vertices x, y, z, the length of any side is bounded by the sum of the lengths of the other two sides. From (D3), we immediately get

$$|d(x,\, y) - d(y,\, z)| \leq d(x,\, z).$$

Let us give some examples of metric spaces. Recall that the *absolute value* $|x|$ of a real number $x \in \mathbb{R}$ is defined such that $|x| = x$ if $x \geq 0$, $|x| = -x$ if $x < 0$, and for a complex number $x = a + ib$, as $|x| = \sqrt{a^2 + b^2}$.

Example 6.1 Let $E = \mathbb{R}$, and $d(x,\, y) = |x - y|$, the absolute value of $x - y$. This is the so-called natural metric on $\mathbb{R}$.

Example 6.2 Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). We have the Euclidean metric

$$d_2(x,\, y) = \left(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2\right)^{\frac{1}{2}},$$

the distance between the points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$.

Example 6.3 For every set E, we can define the *discrete metric*, defined such that $d(x,\, y) = 1$ iff $x \neq y$ and $d(x,\, x) = 0$.

Example 6.4 For any $a, b \in \mathbb{R}$ such that $a < b$, we define the following sets:

1. $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, (closed interval)
2. $]a, b[= \{x \in \mathbb{R} \mid a < x < b\}$, (open interval)
3. $[a, b[= \{x \in \mathbb{R} \mid a \leq x < b\}$, (interval closed on the left, open on the right)
4. $]a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$, (interval open on the left, closed on the right)

Let $E = [a, b]$, and $d(x,\, y) = |x - y|$. Then, $([a, b], d)$ is a metric space.

We will need to define the notion of proximity in order to define convergence of limits and continuity of functions. For this, we introduce some standard "small neighborhoods".

Definition 6.1.2 Given a metric space E with metric d, for every $a \in E$, for every $\rho \in \mathbb{R}$, with $\rho > 0$, the set

$$B(a, \rho) = \{x \in E \mid d(a,\, x) \leq \rho\}$$

is called the *closed ball of center a and radius ρ*, the set

$$B_0(a, \rho) = \{x \in E \mid d(a,\, x) < \rho\}$$

is called the *open ball of center a and radius ρ*, and the set

$$S(a,\, \rho) = \{x \in E \mid d(a,\, x) = \rho\}$$

is called the *sphere of center a and radius ρ*. It should be noted that ρ is finite (i.e. not $+\infty$). A subset X of a metric space E is *bounded* if there is a closed ball $B(a, \rho)$ such that $X \subseteq B(a, \rho)$.

Clearly, $B(a, \rho) = B_0(a, \rho) \cup S(a,\, \rho)$.

In $E = \mathbb{R}$ with the distance $|x - y|$, an open ball of center a and radius ρ is the open interval $]a - \rho, a + \rho[$. In $E = \mathbb{R}^2$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the disk of center a and radius ρ, excluding the boundary points on the circle. In $E = \mathbb{R}^3$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the sphere of center a and radius ρ, excluding the boundary points on the sphere.

One should be aware that intuition can be misleading in forming a geometric image of a closed (or open) ball. For example, if d is the discrete metric, a closed ball of center a and radius $\rho < 1$ consists only of its center a, and a closed ball of center a and radius $\rho \geq 1$ consists of the entire space!

If $E = [a, b]$, and $d(x, y) = |x - y|$, as in example 4, an open ball $B_0(a, \rho)$, with $\rho < b - a$, is in fact the interval $[a, a + \rho[$, which is closed on the left.

We now consider a very important special case of metric spaces, normed vector spaces.

Definition 6.1.3 Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm on* E is a function $\| \|\colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (scaling)

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (convexity inequality)

A vector space E together with a norm $\| \|$ is called a *normed vector space*.

From (N3), we easily get

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Given a normed vector space E, if we define d such that

$$d(x, y) = \|x - y\|,$$

it is easily seen that d is a metric. Thus, every normed vector space is immediately a metric space. Note that the metric associated with a norm is invariant under translation, that is,

$$d(x + u, y + u) = d(x, y).$$

For this reason, we can restrict ourselves to open or closed balls of center 0.

Let us give some examples of normed vector spaces.

Example 6.5 Let $E = \mathbb{R}$, and $\|x\| = |x|$, the absolute value of x. The associated metric is $|x - y|$, as in example 1.

Example 6.6 Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the Euclidean norm $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

and the *sup*-norm $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

Some work is required to show the convexity inequality for the Euclidean norm, but this can be found in any standard text. Note that the Euclidean distance is the distance associated with the Euclidean norm. The following proposition is easy to show.

Proposition 6.1.4 *The following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):*

$$\begin{aligned} \|x\|_\infty &\le \|x\|_1 \le n\|x\|_\infty, \\ \|x\|_\infty &\le \|x\|_2 \le \sqrt{n}\|x\|_\infty, \\ \|x\|_2 &\le \|x\|_1 \le \sqrt{n}\|x\|_2. \end{aligned}$$

In a normed vector space, we define a closed ball or an open ball of radius ρ as a closed ball or an open ball of center 0. We may use the notation $B(\rho)$ and $B_0(\rho)$.

We will now define the crucial notions of open sets and closed sets, and of a topological space.

Definition 6.1.5 Let E be a metric space with metric d. A subset $U \subseteq E$ is an *open set* in E if either $U = \emptyset$, or for every $a \in U$, there is some open ball $B_0(a, \rho)$ such that, $B_0(a, \rho) \subseteq U$.[1] A subset $F \subseteq E$ is a *closed set* in E if its complement $E - F$ is open in E.

The set E itself is open, since for every $a \in E$, every open ball of center a is contained in E. In $E = \mathbb{R}^n$, given n intervals $[a_i, b_i]$, with $a_i < b_i$, it is easy to show that the open n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i < x_i < b_i,\ 1 \le i \le n\}$$

is an open set. In fact, it is possible to find a metric for which such open n-cubes are open balls! Similarly, we can define the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \le x_i \le b_i,\ 1 \le i \le n\},$$

which is a closed set.

The open sets satisfy some important properties that lead to the definition of a topological space.

Proposition 6.1.6 *Given a metric space E with metric d, the family $\mathcal{O}$ of open sets defined in Definition 6.1.5 satisfies the following properties:*

(O1) For every finite family $(U_i)_{1\le i\le n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e. $\mathcal{O}$ is closed under finite intersections.

(O2) For every arbitrary family $(U_i)_{i\in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i\in I} U_i \in \mathcal{O}$, i.e. $\mathcal{O}$ is closed under arbitrary unions.

[1] Recall that $\rho > 0$.

(O3) $\emptyset \in \mathcal{O}$*, and* $E \in \mathcal{O}$*, i.e.* $\emptyset$ *and* E *belong to* $\mathcal{O}$*.*

Furthermore, for any two distinct points $a \neq b$ *in* E*, there exist two open sets* U_a *and* U_b *such that,* $a \in U_a$*,* $b \in U_b$*, and* $U_a \cap U_b = \emptyset$*.*

Proof. It is straightforward. For the last point, letting $\rho = d(a,\, b)/3$ (in fact $\rho = d(a,\, b)/2$ works too), we can pick $U_a = B_0(a, \rho)$ and $U_b = B_0(b, \rho)$. By the triangle inequality, we must have $U_a \cap U_b = \emptyset$. $\square$

The above proposition leads to the very general concept of a topological space.

One should be careful that in general, the family of open sets is not closed under infinite intersections. For example, in $\mathbb{R}$ under the metric $|x - y|$, letting $U_n =]-1/n, +1/n[$, each U_n is open, but $\bigcap_n U_n = \{0\}$, which is not open.

6.2 Topological Spaces, Continuous Functions, Limits

Motivated by Proposition 6.1.6, a topological space is defined in terms of a family of sets satisfing the properties of open sets stated in that proposition.

Definition 6.2.1 Given a set E, a *topology on* E *(or a topological structure on* E*)*, is defined as a family $\mathcal{O}$ of subsets of E called *open sets*, and satisfying the following three properties:

(1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e. $\mathcal{O}$ is closed under finite intersections.

(2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e. $\mathcal{O}$ is closed under arbitrary unions.

(3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e. $\emptyset$ and E belong to $\mathcal{O}$.

A set E together with a topology $\mathcal{O}$ on E is called a *topological space*. Given a topological space $(E, \mathcal{O})$, a subset F of E is a *closed set* if $F = E - U$ for some open set $U \in \mathcal{O}$, i.e. F is the complement of some open set.

It is possible that an open set is also a closed set. For example, $\emptyset$ and E are both open and closed. When a topological space contains a proper nonempty subset U which is both open and closed, the space E is said to be *disconnected*. Connected spaces will be studied in Section 6.3.

A topological space $(E, \mathcal{O})$ is said to satisfy the *Hausdorff separation axiom (or* T_2*-separation axiom)* if for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$. When the T_2-separation axiom is satisfied, we also say that $(E, \mathcal{O})$ is a *Hausdorff space*.

As shown by Proposition 6.1.6, any metric space is a topological Hausdorff space, the family of open sets being in fact the family of arbitrary unions of open balls. Similarly, any normed vector space is a topological Hausdorff space, the family of open sets being the family of arbitrary unions of open balls. The topology $\mathcal{O}$ consisting of all subsets of E is called the *discrete topology*.

Remark: Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough "small" open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed). Nevertheless, non-Hausdorff topological spaces arise naturally in algebraic geometry. But even there, some substitute for separation is used.

One of the reasons why topological spaces are important is that the definition of a topology only involves a certain family $\mathcal{O}$ of sets, and not **how** such family is generated from a metric or a norm. For example, different metrics or different norms can define the same family of open sets. Many topological properties only depend on the family $\mathcal{O}$ and not on the specific metric or norm. But the fact that a topology is definable from a metric or a norm is important, because it usually implies nice properties of a space. All our examples will be spaces whose topology is defined by a metric or a norm.

By taking complements, we can state properties of the closed sets dual to those of Definition 6.2.1. Thus, $\emptyset$ and E are closed sets, and the closed sets are closed under finite unions and arbitrary intersections. It is also worth noting that the Hausdorff separation axiom implies that for every $a \in E$, the set $\{a\}$ is closed. Indeed, if $x \in E - \{a\}$, then $x \neq a$, and so there exist open sets U_a and U_x such that $a \in U_a$, $x \in U_x$, and $U_a \cap U_x = \emptyset$. Thus, for every $x \in E - \{a\}$, there is an open set U_x containing x and contained in $E - \{a\}$, showing by (O3) that $E - \{a\}$ is open, and thus that the set $\{a\}$ is closed.

Given a topological space $(E, \mathcal{O})$, given any subset A of E, since $E \in \mathcal{O}$ and E is a closed set, the family $\mathcal{C}_A = \{F \mid A \subseteq F,\ F \text{ a closed set}\}$ of closed sets containing A is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection $\bigcap \mathcal{C}_A$ of the sets in the family $\mathcal{C}_A$ is the smallest closed set containing A. By a similar reasoning, the union of all the open subsets contained in A is the largest open set contained in A.

Definition 6.2.2 Given a topological space $(E, \mathcal{O})$, given any subset A of E, the smallest closed set containing A is denoted as $\overline{A}$, and is called the *closure, or adherence of* A. A subset A of E is *dense in* E if $\overline{A} = E$. The largest open set contained in A is denoted as $\overset{\circ}{A}$, and is called the *interior of* A. The set $\mathrm{Fr}\, A = \overline{A} \cap \overline{E - A}$, is called the *boundary (or frontier) of* A. We also denote the boundary of A as ∂A.

Remark: The notation $\overline{A}$ for the closure of a subset A of E is somewhat unfortunate, since $\overline{A}$ is often used to denote the set complement of A in E. Still, we prefer it to more cumbersome notations such as $\mathrm{clo}(A)$, and we denote the complement of A in E as $E - A$.

By definition, it is clear that a subset A of E is closed iff $A = \overline{A}$. The set $\mathbb{Q}$ of rationals is dense in $\mathbb{R}$. It is easily shown that $\overline{A} = \overset{\circ}{A} \cup \partial A$ and $\overset{\circ}{A} \cap \partial A = \emptyset$. Another useful characterization of $\overline{A}$ is given by the following proposition.

Proposition 6.2.3 *Given a topological space $(E, \mathcal{O})$, given any subset A of E, the closure $\overline{A}$ of A is the set of all points $x \in E$ such that for every open set U containing x, then $U \cap A \neq \emptyset$.*

Proof. If $A = \emptyset$, since $\emptyset$ is closed, the proposition holds trivially. Thus, assume that $A \neq \emptyset$. First, assume that $x \in \overline{A}$. Let U be any open set such that $x \in U$. If $U \cap A = \emptyset$, since U is open, then $E - U$ is a closed set containing A, and since $\overline{A}$ is the intersection of all closed sets containing A, we must have $x \in E - U$, which is impossible. Conversely, assume that $x \in E$ is a point such that for every open set U containing x, then $U \cap A \neq \emptyset$. Let F be any closed subset containing A. If $x \notin F$, since F is closed, then $U = E - F$ is an open set such that $x \in U$, and $U \cap A = \emptyset$, a contradiction. Thus, we have $x \in F$ for every closed set containing A, that is, $x \in \overline{A}$. □

Often, it is necessary to consider a subset A of a topological space E, and to view the subset A as a topological space. The following proposition shows how to define a topology on a subset.

Proposition 6.2.4 *Given a topological space $(E, \mathcal{O})$, given any subset A of E, let*

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

be the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. The following properties hold.

(1) The space $(A,\mathcal{U})$ is a topological space.

(2) If E is a metric space with metric d, then the restriction $d_A \colon A \times A \to \mathbb{R}_+$ of the metric d to A defines a metric space. Furthermore, the topology induced by the metric d_A agrees with the topology defined by $\mathcal{U}$, as above.

Proof. Left as an exercise. □

Proposition 6.2.4 suggests the following definition.

Definition 6.2.5 Given a topological space $(E, \mathcal{O})$, given any subset A of E, the *subspace topology on A induced by $\mathcal{O}$* is the family $\mathcal{U}$ of open sets defined such that

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

is the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. We say that $(A,\mathcal{U})$ has the *subspace topology*. If (E, d) is a metric space, the restriction $d_A \colon A \times A \to \mathbb{R}_+$ of the metric d to A is called the *subspace metric*.

For example, if $E = \mathbb{R}^n$ and d is the Euclidean metric, we obtain the subspace topology on the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\}.$$

One should realize that every open set $U \in \mathcal{O}$ which is entirely contained in A is also in the family $\mathcal{U}$, but $\mathcal{U}$ may contain open sets that are not in $\mathcal{O}$. For example, if $E = \mathbb{R}$ with $|x - y|$, and $A = [a, b]$, then sets of the form $[a, c[$, with $a < c < b$ belong to $\mathcal{U}$, but they are not open sets for $\mathbb{R}$ under $|x - y|$. However, there is agreement in the following situation.

Proposition 6.2.6 *Given a topological space $(E, \mathcal{O})$, given any subset A of E, if $\mathcal{U}$ is the subspace topology, then the following properties hold.*

(1) If A is an open set $A \in \mathcal{O}$, then every open set $U \in \mathcal{U}$ is an open set $U \in \mathcal{O}$.

(2) If A is a closed set in E, then every closed set w.r.t. the subspace topology is a closed set w.r.t. $\mathcal{O}$.

Proof. Left as an exercise. □

The concept of product topology is also useful. We have the following proposition.

Proposition 6.2.7 *Given n topological spaces $(E_i, \mathcal{O}_i)$, let $\mathcal{B}$ be the family of subsets of $E_1 \times \cdots \times E_n$ defined as follows:*

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

and let $\mathcal{P}$ be the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. Then, $\mathcal{P}$ is a topology on $E_1 \times \cdots \times E_n$.

Proof. Left as an exercise. □

Definition 6.2.8 Given n topological spaces $(E_i, \mathcal{O}_i)$, the *product topology on $E_1 \times \cdots \times E_n$* is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$.

If each $(E_i, \| \ \|_i)$ is a normed vector space, there are three natural norms that can be defined on $E_1 \times \cdots \times E_n$:

$$\begin{aligned}
\|(x_1, \ldots, x_n)\|_1 &= \|x_1\|_1 + \cdots + \|x_n\|_n, \\
\|(x_1, \ldots, x_n)\|_2 &= \left(\|x_1\|_1^2 + \cdots + \|x_n\|_n^2\right)^{\frac{1}{2}}, \\
\|(x_1, \ldots, x_n)\|_\infty &= \max\{\|x_1\|_1, \ldots, \|x_n\|_n\}.
\end{aligned}$$

It is easy to show that they all define the same topology, which is the product topology. One can also verify that when $E_i = \mathbb{R}$, with the standard topology induced by $|x - y|$, the topology product on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm.

Definition 6.2.9 Two metrics d_1 and d_2 on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E (i.e., they define the same family $\mathcal{O}$ of open sets). Similarly, two norms $\| \|_1$ and $\| \|_2$ on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E.

Remark: Given a topological space $(E, \mathcal{O})$, it is often useful, as in Proposition 6.2.7, to define the topology $\mathcal{O}$ in terms of a subfamily $\mathcal{B}$ of subsets of E. We say that a family $\mathcal{B}$ of subsets of E is a *basis for the topology* $\mathcal{O}$ if $\mathcal{B}$ is a subset of $\mathcal{O}$ and if every open set U in $\mathcal{O}$ can be obtained as some union (possibly infinite) of sets in $\mathcal{B}$ (agreeing that the empty union is the empty set). It is immediately verified that if a family $\mathcal{B} = (U_i)_{i\in I}$ is a basis for the topology of $(E, \mathcal{O})$, then $E = \bigcup_{i\in I} U_i$, and the intersection of any two sets $U_i, U_j \in \mathcal{B}$ is the union of some sets in the family $\mathcal{B}$ (again, agreeing that the empty union is the empty set). Conversely, a family $\mathcal{B}$ with these properties is the basis of the topology obtained by forming arbitrary unions of sets in $\mathcal{B}$.

A *subbasis for* $\mathcal{O}$ is a family $\mathcal{S}$ of subsets of E, such that the family $\mathcal{B}$ of all finite intersections of sets in $\mathcal{S}$ (including E itself, in case of the empty intersection) is a basis of $\mathcal{O}$.

We now consider the fundamental property of continuity.

Definition 6.2.10 Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. For every $a \in E$, we say that f *is continuous at* a if for every open set $V \in \mathcal{O}_F$ containing $f(a)$, there is some open set $U \in \mathcal{O}_E$ containing a, such that $f(U) \subseteq V$. We say that f *is continuous* if it is continuous at every $a \in E$.

Define a *neighborhood of* $a \in E$ as any subset N of E containing some open set $O \in \mathcal{O}$ such that $a \in O$. Now, if f is continuous at a and N is any neighborhood of $f(a)$, there is some open set $V \subseteq N$ containing $f(a)$, and since f is continuous at a, there is some open set U containing a, such that $f(U) \subseteq V$. Since $V \subseteq N$, the open set U is a subset of $f^{-1}(N)$ containing a, and $f^{-1}(N)$ is a neighborhood of a. Conversely, if $f^{-1}(N)$ is a neighborhood of a whenever N is any neighborhood of $f(a)$, it is immediate that f is continuous at a. Thus, we can restate Definition 6.2.10 as follows:

The function f is continuous at $a \in E$ iff for every neighborhood N of $f(a) \in F$, then $f^{-1}(N)$ is a neighborhood of a.

It is also easy to check that f is continuous on E iff $f^{-1}(V)$ is an open set in $\mathcal{O}_E$ for every open set $V \in \mathcal{O}_F$.

If E and F are metric spaces defined by metrics d_1 and d_2, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } d_1(a, x) \leq \eta, \text{ then } d_2(f(a), f(x)) \leq \epsilon.$$

Similarly, if E and F are normed vector spaces defined by norms $\| \|_1$ and $\| \|_2$, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - f(a)\|_2 \leq \epsilon.$$

It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, and $f: E \to F$ is a function, for every nonempty subset $A \subseteq E$ of E, we say that f *is continuous on* A if the restriction of f to A is continuous with respect to $(A, \mathcal{U})$ and $(F, \mathcal{O}_F)$, where $\mathcal{U}$ is the subspace topology induced by $\mathcal{O}_E$ on A.

Given a product $E_1 \times \cdots \times E_n$ of topological spaces, as usual, we let $\pi_i: E_1 \times \cdots \times E_n \to E_i$ be the projection function such that, $\pi_i(x_1, \ldots, x_n) = x_i$. It is immediately verified that each π_i is continuous.

Given a topological space $(E, \mathcal{O})$, we say that a point $a \in E$ is *isolated* if $\{a\}$ is an open set in $\mathcal{O}$. Then, if $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, any function $f: E \to F$ is continuous at every isolated point $a \in E$. In the discrete topology, every point is isolated. In a nontrivial normed vector space $(E, \| \|)$ (with $E \neq \{0\}$), no point is isolated. To show this, we show that every open ball $B_0(u, \rho,)$ contains some vectors different from u. Indeed, since E is nontrivial, there is some $v \in E$ such that $v \neq 0$, and thus $\lambda = \|v\| > 0$ (by (N1)). Let

$$w = u + \frac{\rho}{\lambda + 1} v.$$

Since $v \neq 0$ and $\rho > 0$, we have $w \neq u$. Then,

$$\|w - u\| = \left\| \frac{\rho}{\lambda + 1} v \right\| = \frac{\rho \lambda}{\lambda + 1} < \rho,$$

which shows that $\|w - u\| < \rho$, for $w \neq u$.

The following proposition is easily shown.

Proposition 6.2.11 *Given topological spaces $(E, \mathcal{O}_E)$, $(F, \mathcal{O}_F)$, and $(G, \mathcal{O}_G)$, and two functions $f: E \to F$ and $g: F \to G$, if f is continuous at $a \in E$ and g is continuous at $f(a) \in F$, then $g \circ f: E \to G$ is continuous at $a \in E$. Given n topological spaces $(F_i, \mathcal{O}_i)$, for every function $f: E \to F_1 \times \cdots \times F_n$, then f is continuous at $a \in E$ iff every $f_i: E \to F_i$ is continuous at a, where $f_i = \pi_i \circ f$.*

One can also show that in a metric space (E, d), the norm $d: E \times E \to \mathbb{R}$ is continuous, where $E \times E$ has the product topology, and that for a normed vector space $(E, \| \|)$, the norm $\| \|: E \to \mathbb{R}$ is continuous.

Given a function $f\colon E_1 \times \cdots \times E_n \to F$, we can fix $n-1$ of the arguments, say $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, and view f as a function of the remaining argument,

$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),$$

where $x_i \in E_i$. If f is continuous, it is clear that each f_i is continuous.

One should be careful that the converse is false! For example, consider the function $f\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined such that,

$$f(x,y) = \frac{xy}{x^2+y^2} \quad \text{if } (x,y) \neq (0,0), \quad \text{and} \quad f(0,0) = 0.$$

The function f is continuous on $\mathbb{R} \times \mathbb{R} - \{(0,0)\}$, but on the line $y = mx$, with $m \neq 0$, we have $f(x,y) = \frac{m}{1+m^2} \neq 0$, and thus, on this line, $f(x,y)$ does not approach 0 when (x,y) approaches $(0,0)$.

The following proposition is useful for showing that real-valued functions are continuous.

Proposition 6.2.12 *If E is a topological space, and $(\mathbb{R}, |x-y|)$ the reals under the standard topology, for any two functions $f\colon E \to \mathbb{R}$ and $g\colon E \to \mathbb{R}$, for any $a \in E$, for any $\lambda \in \mathbb{R}$, if f and g are continuous at a, then $f+g$, λf, $f \cdot g$, are continuous at a, and f/g is continuous at a if $g(a) \neq 0$.*

Proof. Left as an exercise.

Using Proposition 6.2.12, we can show easily that every real polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

Definition 6.2.13 Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. We say that f *is a homeomorphism between E and F* if f is bijective, and both $f\colon E \to F$ and $f^{-1}\colon F \to E$ are continuous.

One should be careful that a bijective continuous function $f\colon E \to F$ is not necessarily an homeomorphism. For example, if $E = \mathbb{R}$ with the discrete topology, and $F = \mathbb{R}$ with the standard topology, the identity is not a homeomorphism. Another interesting example involving a parametric curve is given below. Let $L\colon \mathbb{R} \to \mathbb{R}^2$ be the function, defined such that,

$$\begin{aligned} L_1(t) &= \frac{t(1+t^2)}{1+t^4}, \\ L_2(t) &= \frac{t(1-t^2)}{1+t^4}. \end{aligned}$$

If we think of $(x(t), y(t)) = (L_1(t), L_2(t))$ as a geometric point in $\mathbb{R}^2$, the set of points $(x(t), y(t))$ obtained by letting t vary in $\mathbb{R}$ from $-\infty$ to $+\infty$, defines a curve having the shape of a “figure eight”, with self-intersection at the origin, called the “lemniscate of Bernoulli”. The map L is continuous, and in fact bijective, but its inverse L^{-1} is not continuous. Indeed, when we approach the origin on the branch of the curve in the upper left quadrant (i.e., points such that, $x \leq 0$, $y \geq 0$), then t goes to $-\infty$, and when we approach the origin on the branch of the curve in the lower right quadrant (i.e., points such that, $x \geq 0$, $y \leq 0$), then t goes to $+\infty$.

We also review the concept of limit of a sequence. Given any set E, a *sequence* is any function $x\colon \mathbb{N} \to E$, usually denoted as $(x_n)_{n\in\mathbb{N}}$, or $(x_n)_{n\geq 0}$, or even as (x_n).

Definition 6.2.14 Given a topological space $(E, \mathcal{O})$, we say that *a sequence* $(x_n)_{n\in\mathbb{N}}$ *converges to some* $a \in E$ if for every open set U containing a, there is some $n_0 \geq 0$, such that, $x_n \in U$, for all $n \geq n_0$. We also say that *a is a limit of* $(x_n)_{n\in\mathbb{N}}$.

When E is a metric space with metric d, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $d(x_n,\, a) \leq \epsilon$, for all $n \geq n_0$.

When E is a normed vector space with norm $\| \, \|$, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $\|x_n - a\| \leq \epsilon$, for all $n \geq n_0$.

The following proposition shows the importance of the Hausdorff separation axiom.

Proposition 6.2.15 *Given a topological space* $(E, \mathcal{O})$*, if the Hausdorff separation axiom holds, then every sequence has at most one limit.*

Proof. Left as an exercise.

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit b iff it converges to the same limit b in any equivalent metric (and similarly for equivalent norms).

We still need one more concept of limit for functions.

Definition 6.2.16 Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, let A be some nonempty subset of E, and let $f\colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, we say that $f(x)$ *approaches b as x approaches a with values in A* if for every open set $V \in \mathcal{O}_F$ containing b, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U \cap A) \subseteq V$. This is denoted as

$$\lim_{x\to a, x\in A} f(x) = b.$$

First, note that by Proposition 6.2.3, since $a \in \overline{A}$, for every open set U containing a, we have $U \cap A \neq \emptyset$, and the definition is nontrivial. Also, even if $a \in A$, the value $f(a)$ of f at a plays no role in this definition. When E and F are metric space with metrics d_1 and d_2, it can be shown easily that the definition can be stated as follows:

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } d_1(x, a) \leq \eta, \text{ then } d_2(f(x), b) \leq \epsilon.$$

When E and F are normed vector spaces with norms $\| \|_1$ and $\| \|_2$, it can be shown easily that the definition can be stated as follows:

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - b\|_2 \leq \epsilon.$$

We have the following result relating continuity at a point and the previous notion.

Proposition 6.2.17 *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f \colon E \to F$ be a function. For any $a \in E$, the function f is continuous at a iff $f(x)$ approaches $f(a)$ when x approaches a (with values in E).*

Proof. Left as a trivial exercise.

Another important proposition relating the notion of convergence of a sequence to continuity, is stated without proof.

Proposition 6.2.18 *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f \colon E \to F$ be a function.*

(1) If f is continuous, then for every sequence $(x_n)_{n \in \mathbb{N}}$ in E, if (x_n) converges to a, then $(f(x_n))$ converges to $f(a)$.

(2) If E is a metric space, and $(f(x_n))$ converges to $f(a)$ whenever (x_n) converges to a, for every sequence $(x_n)_{n \in \mathbb{N}}$ in E, then f is continuous.

Remark: A special case of Definition 6.2.16 shows up in the following case: $E = \mathbb{R}$, and F is some arbitrary topological space. Let A be some nonempty subset of $\mathbb{R}$, and let $f \colon A \to F$ be some function. For any $a \in A$, we say that f *is continuous on the right at a* if

$$\lim_{x \to a, x \in A \cap [a, +\infty[} f(x) = f(a).$$

We can define continuity on the left at a in a similar fashion.

We now turn to connectivity properties of topological spaces.

6.3 Connected Sets

Connectivity properties of topological spaces play a very important role in understanding the topology of surfaces. This section gathers the facts needed to have a good understanding of the classification theorem for compact (bordered) surfaces. The main references are Ahlfors and Sario [1] and Massey [11, 12]. For general backgroud on topology, geometry, and algebraic topology, we also highly recommend Bredon [3] and Fulton [7].

Definition 6.3.1 A topological space $(E, \mathcal{O})$ is *connected* if the only subsets of E that are both open and closed are the empty set and E itself. Equivalently, $(E, \mathcal{O})$ is connected if E cannot be written as the union $E = U \cup V$ of two disjoint nonempty open sets U, V, if E cannot be written as the union $E = U \cup V$ of two disjoint nonempty closed sets. A subset $S \subseteq E$ is *connected* if it is connected in the subspace topology on S induced by $(E, \mathcal{O})$. A connected open set is called a *region*, and a closed set is a *closed region* if its interior is a connected (open) set.

Intuitively, if a space is not connected, it is possible to define a continuous function which is constant on disjoint "connected components" and which takes possibly distinct values on disjoint components. This can be stated in terms of the concept of a locally constant function. Given two topological spaces X, Y, a function $f\colon X \to Y$ is *locally constant* if for every $x \in X$, there is an open set $U \subseteq X$ such that $x \in X$ and f is constant on U.

We claim that a locally constant function is continuous. In fact, we will prove that $f^{-1}(V)$ is open for every subset $V \subseteq Y$ (not just for an open set V). It is enough to show that $f^{-1}(y)$ is open for every $y \in Y$, since for every subset $V \subseteq Y$,

$$f^{-1}(V) = \bigcup_{y \in V} f^{-1}(y),$$

and open sets are closed under arbitrary unions. However, either $f^{-1}(y) = \emptyset$ if $y \in Y - f(X)$ or f is constant on $U = f^{-1}(y)$ if $y \in f(X)$ (with value y), and since f is locally constant, for every $x \in U$, there is some open set $W \subseteq X$ such that $x \in W$ and f is constant on W, which implies that $f(w) = y$ for all $w \in W$, and thus that $W \subseteq U$, showing that U is a union of open sets, and thus is open. The following proposition shows that a space is connected iff every locally constant function is constant.

Proposition 6.3.2 *A topological space is connected iff every locally constant function is constant.*

Proof. First, assume that X is connected. Let $f\colon X \to Y$ be a locally constant function to some space Y, and assume that f is not constant. Pick any $y \in f(Y)$. Since f is not constant, $U_1 = f^{-1}(y) \neq X$, and of course $U_1 \neq \emptyset$. We proved just before Proposition 6.3.2 that $f^{-1}(V)$ is open for every subset $V \subseteq Y$, and thus $U_1 = f^{-1}(y) = f^{-1}(\{y\})$ and

$U_2 = f^{-1}(Y - \{y\})$ are both open, nonempty, and clearly $X = U_1 \cup U_2$ and U_1 and U_2 are disjoint. This contradicts the fact that X is connected, and f must be constant.

Assume that every locally constant function $f\colon X \to Y$ to a Hausdorff space Y is constant. If X is not connected, we can write $X = U_1 \cup U_2$, where both U_1, U_2 are open, disjoint, and nonempty. We can define the function $f\colon X \to \mathbb{R}$ such that $f(x) = 1$ on U_1 and $f(x) = 0$ on U_2. Since U_1 and U_2 are open, the function f is locally constant, and yet not constant, a contradiction. □

The following standard proposition characterizing the connected subsets of $\mathbb{R}$ can be found in most topology texts (for example, Munkres [13], Schwartz [17]). For the sake of completeness, we give a proof.

Proposition 6.3.3 *A subset of the real line $\mathbb{R}$ is connected iff it is an interval, i.e., of the form $[a, b]$, $]\,a, b]$, where $a = -\infty$ is possible, $[a, b[$, where $b = +\infty$ is possible, or $]a, b[$, where $a = -\infty$ or $b = +\infty$ is possible.*

Proof. Assume that A is a connected nonempty subset of $\mathbb{R}$. The cases where $A = \emptyset$ or A consists of a single point are trivial. We show that whenever $a, b \in A$, $a < b$, then the entire interval $[a, b]$ is a subset of A. Indeed, if this was not the case, there would be some $c \in]a, b[$ such that $c \notin A$, and then we could write $A = (\,]-\infty, c[\,\cap A) \cup (\,]c + \infty[\,\cap A)$, where $]-\infty, c[\cap A$ and $]c + \infty[\cap A$ are nonempty and disjoint open subsets of A, contradicting the fact that A is connected. It follows easily that A must be an interval.

Conversely, we show that an interval I must be connected. Let A be any nonempty subset of I which is both open and closed in I. We show that $I = A$. Fix any $x \in A$, and consider the set R_x of all y such that $[x, y] \subseteq A$. If the set R_x is unbounded, then $R_x = [x, +\infty[$. Otherwise, if this set is bounded, let b be its least upper bound. We claim that b is the right boundary of the interval I. Because A is closed in I, unless I is open on the right and b is its right boundary, we must have $b \in A$. In the first case, $A \cap [x, b[= I \cap [x, b[= [x, b[$. In the second case, because A is also open in I, unless b is the right boundary of the interval I (closed on the right), there is some open set $]b - \eta, b + \eta[$ contained in A, which implies that $[x, b + \eta/2] \subseteq A$, contradicting the fact that b is the least upper bound of the set R_x. Thus, b must be the right boundary of the interval I (closed on the right). A similar argument applies to the set L_y of all x such that $[x, y] \subseteq A$, and either L_y is unbounded, or its greatest lower bound a is the left boundary of I (open or closed on the left). In all cases, we showed that $A = I$, and the interval must be connected. □

A characterization on the connected subsets of $\mathbb{R}^n$ is harder, and requires the notion of arcwise connectedness. One of the most important properties of connected sets is that they are preserved by continuous maps.

Proposition 6.3.4 *Given any continuous map $f\colon E \to F$, if $A \subseteq E$ is connected, then $f(A)$ is connected.*

Proof. If $f(A)$ is not connected, then there exist some nonempty open sets U, V in F such that $f(A) \cap U$ and $f(A) \cap V$ are nonempty and disjoint, and

$$f(A) = (f(A) \cap U) \cup (f(A) \cap V).$$

Then, $f^{-1}(U)$ and $f^{-1}(V)$ are nonempty and open since f is continuous, and

$$A = (A \cap f^{-1}(U)) \cup (A \cap f^{-1}(V)),$$

with $A \cap f^{-1}(U)$ and $A \cap f^{-1}(V)$ nonempty, disjoint, and open in A, contradicting the fact that A is connected. $\square$

An important corollary of Proposition 6.3.4 is that for every continuous function $f\colon E \to \mathbb{R}$, where E is a connected space, then $f(E)$ is an interval. Indeed, this follows from Proposition 6.3.3. Thus, if f takes the values a and b where $a < b$, then f takes all values $c \in [a, b]$. This is a very important property.

Even if a topological space is not connected, it turns out that it is the disjoint union of maximal connected subsets, and these connected components are closed in E. In order to obtain this result, we need a few lemmas.

Lemma 6.3.5 *Given a topological space E, for any family $(A_i)_{i\in I}$ of (nonempty) connected subsets of E, if $A_i \cap A_j \neq \emptyset$ for all $i, j \in I$, then the union $A = \bigcup_{i\in I} A_i$ of the family $(A_i)_{i\in I}$ is also connected.*

Proof. Assume that $\bigcup_{i\in I} A_i$ is not connected. Then, there exists two nonempty open subsets U and V of E such that $A \cap U$ and $A \cap V$ are disjoint and nonempty, and such that

$$A = (A \cap U) \cup (A \cap V).$$

Now, for every $i \in I$, we can write

$$A_i = (A_i \cap U) \cup (A_i \cap V),$$

where $A_i \cap U$ and $A_i \cap V$ are disjoint, since $A_i \subseteq A$ and $A \cap U$ and $A \cap V$ are disjoint. Since A_i is connected, either $A_i \cap U = \emptyset$ or $A_i \cap V = \emptyset$. This implies that either $A_i \subseteq A \cap U$ or $A_i \subseteq A \cap V$. However, by assumption, $A_i \cap A_j \neq \emptyset$, for all $i, j \in I$, and thus, either both $A_i \subseteq A \cap U$ and $A_j \subseteq A \cap U$, or both $A_i \subseteq A \cap V$ and $A_j \subseteq A \cap V$, since $A \cap U$ and $A \cap V$ are disjoint. Thus, we conclude that either $A_i \subseteq A \cap U$ for all $i \in I$, or $A_i \subseteq A \cap V$ for all $i \in I$. But this proves that either

$$A = \bigcup_{i\in I} A_i \subseteq A \cap U,$$

or

$$A = \bigcup_{i\in I} A_i \subseteq A \cap V,$$

contradicting the fact that both $A \cap U$ and $A \cap V$ are disjoint and nonempty. Thus, A must be connected. □

In particular, the above lemma applies when the connected sets in a family $(A_i)_{i \in I}$ have a point in common.

Lemma 6.3.6 *If A is a connected subset of a topological space E, then for every subset B such that $A \subseteq B \subseteq \overline{A}$, where $\overline{A}$ is the closure of A in E, the set B is connected.*

Proof. If B is not connected, then there are two nonempty open subsets U, V of E such that $B \cap U$ and $B \cap V$ are disjoint and nonempty, and

$$B = (B \cap U) \cup (B \cap V).$$

Since $A \subseteq B$, the above implies that

$$A = (A \cap U) \cup (A \cap V),$$

and since A is connected, either $A \cap U = \emptyset$, or $A \cap V = \emptyset$. Without loss of generality, assume that $A \cap V = \emptyset$, which implies that $A \subseteq A \cap U \subseteq B \cap U$. However, $B \cap U$ is closed in the subspace topology for B, and since $B \subseteq \overline{A}$ and $\overline{A}$ is closed in E, the closure of A in B w.r.t. the subspace topology of B is clearly $B \cap \overline{A} = B$, which implies that $B \subseteq B \cap U$ (since the closure is the smallest closed set containing the given set). Thus, $B \cap V = \emptyset$, a contradiction. □

In particular, Lemma 6.3.6 shows that if A is a connected subset, then its closure $\overline{A}$ is also connected. We are now ready to introduce the connected components of a space.

Definition 6.3.7 Given a topological space $(E, \mathcal{O})$ we say that two points $a, b \in E$ are *connected* if there is some connected subset A of E such that $a \in A$ and $b \in A$.

It is immediately verified that the relation "a and b are connected in E" is an equivalence relation. Only transitivity is not obvious, but it follows immediately as a special case of Lemma 6.3.5. Thus, the above equivalence relation defines a partition of E into nonempty disjoint *connected components*. The following proposition is easily proved using Lemma 6.3.5 and Lemma 6.3.6.

Proposition 6.3.8 *Given any topological space E, for any $a \in E$, the connected component containing a is the largest connected set containing a. The connected components of E are closed.*

The notion of a locally connected space is also useful.

Definition 6.3.9 A topological space $(E, \mathcal{O})$ is *locally connected* if for every $a \in E$, for every neighborhood V of a, there is a connected neighborhood U of a such that $U \subseteq V$.

As we shall see in a moment, it would be equivalent to require that E has a basis of connected open sets.

There are connected spaces that are not locally connected, and there are locally connected spaces that are not connected. The two properties are independent.

Proposition 6.3.10 *A topological space E is locally connected iff for every open subset A of E, the connected components of A are open.*

Proof. Assume that E is locally connected. Let A be any open subset of E, and let C be one of the connected components of A. For any $a \in C \subseteq A$, there is some connected neigborhood U of a such that $U \subseteq A$, and since C is a connected component of A containing a, we must have $U \subseteq C$. This shows that for every $a \in C$, there is some open subset containing a contained in C, and C is open.

Conversely, assume that for every open subset A of E, the connected components of A are open. Then, for every $a \in E$ and every neighborhood U of a, since U contains some open set A containing a, the interior $\overset{\circ}{U}$ of U is an open set containing a, and its connected components are open. In particular, the connected component C containing a is a connected open set containing a and contained in U. □

Proposition 6.3.10 shows that in a locally connected space, the connected open sets form a basis for the topology. It is easily seen that $\mathbb{R}^n$ is locally connected. Another very important property of surfaces, and more generally manifolds, is to be arcwise connected. The intuition is that any two points can be joined by a continuous arc of curve. This is formalized as follows.

Definition 6.3.11 Given a topological space $(E, \mathcal{O})$, an *arc (or path)* is a continuous map $\gamma\colon [a, b] \to E$, where $[a, b]$ is a closed interval of the real line $\mathbb{R}$. The point $\gamma(a)$ is the *initial point* of the arc, and the point $\gamma(b)$ is the *terminal point* of the arc. We say that γ *is an arc joining* $\gamma(a)$ *and* $\gamma(b)$. An arc is a *closed curve* if $\gamma(a) = \gamma(b)$. The set $\gamma([a, b])$ is the *trace* of the arc γ.

Typically, $a = 0$ and $b = 1$. In the sequel, this will be assumed.

One should not confuse an arc $\gamma\colon [a, b] \to E$ with its trace. For example, γ could be constant, and thus, its trace reduced to a single point.

An arc is a *Jordan arc* if γ is a homeomorphism onto its trace. An arc $\gamma\colon [a, b] \to E$ is a *Jordan curve* if $\gamma(a) = \gamma(b)$, and γ is injective on $[a, b[$. Since $[a, b]$ is connected, by Proposition 6.3.4, the trace $\gamma([a, b])$ of an arc is a connected subset of E.

Given two arcs $\gamma\colon [0, 1] \to E$ and $\delta\colon [0, 1] \to E$ such that $\gamma(1) = \delta(0)$, we can form a new arc defined as follows.

Definition 6.3.12 Given two arcs $\gamma\colon [0, 1] \to E$ and $\delta\colon [0, 1] \to E$ such that $\gamma(1) = \delta(0)$, we can form their *composition (or product)* $\gamma\delta$, defined such that

$$\gamma\delta(t) = \begin{cases} \gamma(2t) & \text{if } 0 \leq t \leq 1/2; \\ \delta(2t - 1) & \text{if } 1/2 \leq t \leq 1. \end{cases}$$

The *inverse* γ^{-1} *of the arc* γ is the arc defined such that $\gamma^{-1}(t) = \gamma(1-t)$, for all $t \in [0,1]$.

It is trivially verified that Definition 6.3.12 yields continuous arcs.

Definition 6.3.13 A topological space E is *arcwise connected* if for any two points $a, b \in E$, there is an arc $\gamma\colon [0,1] \to E$ joining a and b, i.e., such that $\gamma(0) = a$ and $\gamma(1) = b$. A topological space E is *locally arcwise connected* if for every $a \in E$, for every neighborhood V of a, there is an arcwise connected neighborhood U of a such that $U \subseteq V$.

The space $\mathbb{R}^n$ is locally arcwise connected, since for any open ball, any two points in this ball are joined by a line segment. Manifolds and surfaces are also locally arcwise connected. It is easy to verify that Proposition 6.3.4 also applies to arcwise connectedness. The following theorem is crucial to the theory of manifolds and surfaces.

Theorem 6.3.14 *If a topological space E is arcwise connected, then it is connected. If a topological space E is connected and locally arcwise connected, then E is arcwise connected.*

Proof. First, assume that E is arcwise connected. Pick any point a in E. Since E is arcwise connected, for every $b \in E$, there is a path $\gamma_b\colon [0,1] \to E$ from a to b, and so

$$E = \bigcup_{b \in E} \gamma_b([0,1])$$

a union of connected subsets all containing a. By Lemma 6.3.5, E is connected.

Now assume that E is connected and locally arcwise connected. For any point $a \in E$, let F_a be the set of all points b such that there is an arc $\gamma_b\colon [0,1] \to E$ from a to b. Clearly, F_a contains a. We show that F_a is both open and closed. For any $b \in F_a$, since E is locally arcwise connected, there is an arcwise connected neighborhood U containing b (because E is a neighborhood of b). Thus, b can be joined to every point $c \in U$ by an arc, and since by the definition of F_a, there is an arc from a to b, the composition of these two arcs yields an arc from a to c, which shows that $c \in F_a$. But then $U \subseteq F_a$, and thus F_a is open. Now assume that b is in the complement of F_a. As in the previous case, there is some arcwise connected neighborhood U containing b. Thus, every point $c \in U$ can be joined to b by an arc. If there was an arc joining a to c, we would get an arc from a to b, contradicting the fact that b is in the complement of F_a. Thus, every point $c \in U$ is in the complement of F_a, which shows that U is contained in the complement of F_a, and thus, that the the complement of F_a is open. Consequently, we have shown that F_a is both open and closed, and since it is nonempty, we must have $E = F_a$, which shows that E is arcwise connected. □

If E is locally arcwise connected, the above argument shows that the connected components of E are arcwise connected.

It is not true that a connected space is arcwise connected. For example, the space consisting of the graph of the function

$$f(x) = \sin(1/x),$$

where $x > 0$, together with the portion of the y-axis, for which $-1 \leq y \leq 1$, is connected, but not arcwise connected.

A trivial modification of the proof of Theorem 6.3.14 shows that in a normed vector space E, a connected open set is arcwise connected by polygonal lines (i.e., arcs consisting of line segments). This is because in every open ball, any two points are connected by a line segment. Furthermore, if E is finite dimensional, these polygonal lines can be forced to be parallel to basis vectors.

We now consider compactness.

6.4 Compact Sets

The property of compactness is very important in topology and analysis. We provide a quick review geared towards the study of surfaces, and for details, refer the reader to Munkres [13], Schwartz [17]. In this section, we will need to assume that the topological spaces are Hausdorff spaces. This is not a luxury, as many of the results are false otherwise.

There are various equivalent ways of defining compactness. For our purposes, the most convenient way involves the notion of open cover.

Definition 6.4.1 Given a topological space E, for any subset A of E, an *open cover* $(U_i)_{i\in I}$ *of* A is a family of open subsets of E such that $A \subseteq \bigcup_{i\in I} U_i$. An *open subcover* of an open cover $(U_i)_{i\in I}$ of A is any subfamily $(U_j)_{j\in J}$ which is an open cover of A, with $J \subseteq I$. An open cover $(U_i)_{i\in I}$ of A is *finite* if I is finite. The topological space E is *compact* if it is Hausdorff and for every open cover $(U_i)_{i\in I}$ of E, there is a finite open subcover $(U_j)_{j\in J}$ of E. Given any subset A of E, we say that A is *compact* if it is compact with respect to the subspace topology. We say that A is *relatively compact* if its closure $\overline{A}$ is compact.

It is immediately verified that a subset A of E is compact in the subspace topology relative to A iff for every open cover $(U_i)_{i\in I}$ of A by open subsets of E, there is a finite open subcover $(U_j)_{j\in J}$ of A. The property that every open cover contains a finite open subcover is often called the *Heine-Borel-Lebesgue* property. By considering complements, a Hausdorff space is compact iff for every family $(F_i)_{i\in I}$ of closed sets, if $\bigcap_{i\in I} F_i = \emptyset$, then $\bigcap_{j\in J} F_j = \emptyset$ for some finite subset J of I.

Definition 6.4.1 requires that a compact space be Hausdorff. There are books in which a compact space is not necessarily required to be Hausdorff. Following Schwartz, we prefer calling such a space *quasi-compact*.

Another equivalent and useful characterization can be given in terms of families having the finite intersection property. A family $(F_i)_{i\in I}$ of sets has the *finite intersection property* if $\bigcap_{j\in J} F_j \neq \emptyset$ for every finite subset J of I. We have the following proposition.

Proposition 6.4.2 *A topological Hausdorff space E is compact iff for every family $(F_i)_{i\in I}$ of closed sets having the finite intersection property, then $\bigcap_{i\in I} F_i \neq \emptyset$.*

Proof. If E is compact and $(F_i)_{i\in I}$ is a family of closed sets having the finite intersection property, then $\bigcap_{i\in I} F_i$ cannot be empty, since otherwise we would have $\bigcap_{j\in J} F_j = \emptyset$ for some finite subset J of I, a contradiction. The converse is equally obvious. □

Another useful consequence of compactness is as follows. For any family $(F_i)_{i\in I}$ of closed sets such that $F_{i+1} \subseteq F_i$ for all $i \in I$, if $\bigcap_{i\in I} F_i = \emptyset$, then $F_i = \emptyset$ for some $i \in I$. Indeed, there must be some finite subset J of I such that $\bigcap_{j\in J} F_j = \emptyset$, and since $F_{i+1} \subseteq F_i$ for all $i \in I$, we must have $F_j = \emptyset$ for the smallest F_j in $(F_j)_{j\in J}$. Using this fact, we note that $\mathbb{R}$ is *not* compact. Indeed, the family of closed sets $([n, +\infty[\,)_{n\geq 0}$ is decreasing and has an empty intersection.

Given a metric space, if we define a *bounded subset* to be a subset that can be enclosed in some closed ball (of finite radius), any nonbounded subset of a metric space is not compact. However, a closed interval $[a, b]$ of the real line is compact.

Proposition 6.4.3 *Every closed interval $[a, b]$ of the real line is compact.*

Proof. We proceed by contradiction. Let $(U_i)_{i\in I}$ be any open cover of $[a, b]$, and assume that there is no finite open subcover. Let $c = (a + b)/2$. If both $[a, c]$ and $[c, b]$ had some finite open subcover, so would $[a, b]$, and thus, either $[a, c]$ does not have any finite subcover, or $[c, b]$ does not have any finite open subcover. Let $[a_1, b_1]$ be such a bad subinterval. The same argument applies, and we split $[a_1, b_1]$ into two equal subintervals, one of which must be bad. Thus, having defined $[a_n, b_n]$ of length $(b - a)/2^n$ as an interval having no finite open subcover, splitting $[a_n, b_n]$ into two equal intervals, we know that at least one of the two has no finite open subcover, and we denote such a bad interval as $[a_{n+1}, b_{n+1}]$. The sequence (a_n) is nondecreasing and bounded from above by b, and thus, by a fundamental property of the real line, it converges to its least upper bound α. Similarly, the sequence (b_n) is nonincreasing and bounded from below by a, and thus, it converges to its greatest lowest bound β. Since $[a_n, b_n]$ has length $(b - a)/2^n$, we must have $\alpha = \beta$. However, the common limit $\alpha = \beta$ of the sequences (a_n) and (b_n) must belong to some open set U_i of the open cover, and since U_i is open, it must contain some interval $[c, d]$ containing α. Then, because α is the common limit of the sequences (a_n) and (b_n), there is some N such that the intervals $[a_n, b_n]$ are all contained in the interval $[c, d]$ for all $n \geq N$, which contradicts the fact that none of the intervals $[a_n, b_n]$ has a finite open subcover. Thus, $[a, b]$ is indeed compact. □

It is easy to adapt the argument of Proposition 6.4.3 to show that in $\mathbb{R}^m$, every closed set $[a_1, b_1] \times \cdots \times [a_m, b_m]$ is compact. At every stage, we need to divide into 2^m subpieces instead of 2.

The following two propositions give very important properties of the compact sets, and they only hold for Hausdorff spaces.

Proposition 6.4.4 *Given a topological Hausdorff space E, for every compact subset A and every point b not in A, there exist disjoint open sets U and V such that $A \subseteq U$ and $b \in V$. As a consequence, every compact subset is closed.*

Proof. Since E is Hausdorff, for every $a \in A$, there are some disjoint open sets U_a and V_b containing a and b respectively. Thus, the family $(U_a)_{a\in A}$ forms an open cover of A. Since A is compact there is a finite open subcover $(U_j)_{j\in J}$ of A, where $J \subseteq A$, and then $\bigcup_{j\in J} U_j$ is an open set containing A disjoint from the open set $\bigcap_{j\in J} V_j$ containing b. This shows that every point b in the complement of A belongs to some open set in this complement, and thus that the complement is open, i.e., that A is closed. □

Actually, the proof of Proposition 6.4.4 can be used to show the following useful property.

Proposition 6.4.5 *Given a topological Hausdorff space E, for every pair of compact disjoint subsets A and B, there exist disjoint open sets U and V such that $A \subseteq U$ and $B \subseteq V$.*

Proof. We repeat the argument of Proposition 6.4.4 with B playing the role of b, and use Proposition 6.4.4 to find disjoint open sets U_a containing $a \in A$ and V_a containing B. □

The following proposition shows that in a compact topological space, every closed set is compact.

Proposition 6.4.6 *Given a compact topological space E, every closed set is compact.*

Proof. Since A is closed, $E - A$ is open, and from any open cover $(U_i)_{i\in I}$ of A, we can form an open cover of E by adding $E - A$ to $(U_i)_{i\in I}$, and since E is compact, a finite subcover $(U_j)_{j\in J} \cup \{E - A\}$ of E can be extracted, such that $(U_j)_{j\in J}$ is a finite subcover of A. □

Remark: Proposition 6.4.6 also holds for quasi-compact spaces, i.e., the Hausdorff separation property is not needed.

Putting Proposition 6.4.5 and Proposition 6.4.6 together, we note that if X is compact, then for every pair of disjoint closed sets A and B, there exist disjoint open sets U and V such that $A \subseteq U$ and $B \subseteq V$. We say that X is a *normal* space.

Proposition 6.4.7 *Given a compact topological space E, for every $a \in E$, for every neighborhood V of a, there exists a compact neighborhood U of a such that $U \subseteq V$.*

Proof. Since V is a neighborhood of a, there is some open subset O of V containing a. Then the complement $K = E-O$ of O is closed, and since E is compact, by Proposition 6.4.6, K is compact. Now, if we consider the family of all closed sets of the form $K \cap F$, where F is any closed neighborhood of a, since $a \notin K$, this family has an empty intersection, and thus there is a finite number of closed neighborhood $F_1, \ldots, F_n$ of a, such that $K \cap F_1 \cap \cdots \cap F_n = \emptyset$. Then, $U = F_1 \cap \cdots \cap F_n$ is a compact neigborhood of a contained in $O \subseteq V$. □

It can be shown that in a normed vector space of finite dimension, a subset is compact iff it is closed and bounded. For $\mathbb{R}^n$, this is easy.

In a normed vector space of infinite dimension, there are closed and bounded sets that are not compact!

More could be said about compactness in metric spaces, but we will only need the notion of Lebesgue number, which will be discussed a little later. Another crucial property of compactness is that it is preserved under continuity.

Proposition 6.4.8 *Let E be a topological space, and F be a topological Hausdorff space. For every compact subset A of E, for every continuous map $f\colon E \to F$, the subspace $f(A)$ is compact.*

Proof. Let $(U_i)_{i\in I}$ be an open cover of $f(A)$. We claim that $(f^{-1}(U_i))_{i\in I}$ is an open cover of A, which is easily checked. Since A is compact, there is a finite open subcover $(f^{-1}(U_j))_{j\in J}$ of A, and thus, $(U_j)_{j\in J}$ is an open subcover of $f(A)$. □

As a corollary of Proposition 6.4.8, if E is compact, F is Hausdorff, and $f\colon E \to F$ is continuous and bijective, then f is a homeomorphism. Indeed, it is enough to show that f^{-1} is continuous, which is equivalent to showing that f maps closed sets to closed sets. However, closed sets are compact, and Proposition 6.4.8 shows that compact sets are mapped to compact sets, which, by Proposition 6.4.4, are closed.

It can also be shown that if E is a compact nonempty space and $f\colon E \to \mathbb{R}$ is a continuous function, then there are points $a, b \in E$ such that $f(a)$ is the minimum of $f(E)$ and $f(b)$ is the maximum of $f(E)$. Indeed, $f(E)$ is a compact subset of $\mathbb{R}$, and thus a closed and bounded set which contains its greatest lower bound and its least upper bound.

Another useful notion is that of local compactness. Indeed, manifolds and surfaces are locally compact.

Definition 6.4.9 A topological space E is *locally compact* if it is Hausdorff and for every $a \in E$, there is some compact neighborhood A of a.

From Proposition 6.4.7, every compact space is locally compact, but the converse is false. It can be shown that a normed vector space of finite dimension is locally compact.

Proposition 6.4.10 *Given a locally compact topological space E, for every $a \in E$, for every neighborhood N of a, there exists a compact neighborhood U of a, such that $U \subseteq N$.*

Proof. For any $a \in E$, there is some compact neighborhood V of a. By Proposition 6.4.7, every neigborhood of a relative to V contains some compact neighborhood U of a relative to V. But every neighborhood of a relative to V is a neighborhood of a relative to E, and every neighborhood N of a in E yields a neighborhood $V \cap N$ of a in V, and thus for every neighborhood N of a, there exists a compact neighborhood U of a such that $U \subseteq N$. □

It is much harder to deal with noncompact surfaces (or manifolds) than it is to deal with compact surfaces (or manifolds). However, surfaces (and manifolds) are locally compact, and it turns out that there are various ways of embedding a locally compact Hausdorff space into a compact Hausdorff space. The most economical construction consists in adding just one point. This construction, known as the *Alexandroff compactification*, is technically useful, and we now describe it and sketch the proof that it achieves its goal.

To help the reader's intuition, let us consider the case of the plane $\mathbb{R}^2$. If we view the plane $\mathbb{R}^2$ as embedded in 3-space $\mathbb{R}^3$, say as the xOy plane of equation $z = 0$, we can consider the sphere Σ of radius 1 centered on the z-axis at the point $(0, 0, 1)$, and tangent to the xOy plane at the origin (sphere of equation $x^2+y^2+(z-1)^2 = 1$). If N denotes the north pole on the sphere, i.e., the point of coordinates $(0, 0, 2)$, then any line D passing through the north pole and not tangent to the sphere (i.e., not parallel to the xOy plane), intersects the xOy plane in a unique point M, and the sphere in a unique point P other than the north pole N. This, way, we obtain a bijection between the xOy plane and the punctured sphere Σ, i.e., the sphere with the north pole N deleted. This bijection is called a *stereographic projection*. The Alexandroff compactification of the plane consists in putting the north pole back on the sphere, which amounts to adding a single point at infinity ∞ to the plane. Intuitively, as we travel away from the origin O towards infinity (in any direction!), we tend towards an ideal point at infinity ∞. Imagine that we "bend" the plane so that it gets wrapped around the sphere, according to stereographic projection. A simpler example consists in taking a line and getting a circle as its compactification. The Alexandroff compactification is a generalization of these simple constructions.

Definition 6.4.11 Let $(E, \mathcal{O})$ be a locally compact space. Let ω be any point not in E, and let $E_\omega = E \cup \{\omega\}$. Define the family $\mathcal{O}_\omega$ as follows:

$$\mathcal{O}_\omega = \mathcal{O} \cup \{(E - K) \cup \{\omega\} \mid K \text{ compact in } E\}.$$

The pair $(E_\omega, \mathcal{O}_\omega)$ is called the *Alexandroff compactification (or one point compactification) of* $(E, \mathcal{O})$.

The following theorem shows that $(E_\omega, \mathcal{O}_\omega)$ is indeed a topological space, and that it is compact.

Theorem 6.4.12 *Let E be a locally compact topological space. The Alexandroff compactification E_ω of E is a compact space such that E is a subspace of E_ω, and if E is not compact, then $\overline{E} = E_\omega$.*

Proof. The verification that $\mathcal{O}_\omega$ is a family of open sets is not difficult but a bit tedious. Details can be found in Munkres [13] or Schwartz [17]. Let us show that E_ω is compact. For every open cover $(U_i)_{i\in I}$ of E_ω, since ω must be covered, there is some U_{i_0} of the form

$$U_{i_0} = (E - K_0) \cup \{\omega\}$$

where K_0 is compact in E. Consider the family $(V_i)_{i\in I}$ defined as follows:

$$\begin{aligned} V_i &= U_i \quad \text{if} \quad U_i \in \mathcal{O}, \\ V_i &= E - K \quad \text{if} \quad U_i = (E - K) \cup \{\omega\}, \end{aligned}$$

where K is compact in E. Then, because each K is compact and thus closed in E (since E is Hausdorff), $E - K$ is open, and every V_i is an open subset of E. Furthermore, the family $(V_i)_{i\in(I-\{i_0\})}$ is an open cover of K_0. Since K_0 is compact, there is a finite open subcover $(V_j)_{j\in J}$ of K_0, and thus, $(U_j)_{j\in J\cup\{i_0\}}$ is a finite open cover of E_ω.

Let us show that E_ω is Hausdorff. Given any two points $a, b \in E_\omega$, if both $a, b \in E$, since E is Hausdorff and every open set in $\mathcal{O}$ is an open set in $\mathcal{O}_\omega$, there exist disjoint open sets U, V (in $\mathcal{O}$) such that $a \in U$ and $b \in V$. If $b = \omega$, since E is locally compact, there is some compact set K containing an open set U containing a, and then, U and $V = (E - K) \cup \{\omega\}$ are disjoint open sets (in $\mathcal{O}_\omega$) such that $a \in U$ and $b \in V$.

The space E is a subspace of E_ω because for every open set U in $\mathcal{O}_\omega$, either $U \in \mathcal{O}$ and $E \cap U = U$ is open in E, or $U = (E - K) \cup \{\omega\}$ where K is compact in E, and thus $U \cap E = E - K$, which is open in E, since K is compact in E, and thus closed (since E is Hausdorff). Finally, if E is not compact, for every compact subset K of E, $E - K$ is nonempty, and thus, for every open set $U = (E - K) \cup \{\omega\}$ containing ω, we have $U \cap E \neq \emptyset$, which shows that $\omega \in \overline{E}$, and thus that $\overline{E} = E_\omega$. □

Finally, in studying surfaces and manifolds, an important property is the existence of a countable basis for the topology. Indeed, this property guarantees the existence of triangulations of surfaces, a crucial property.

Definition 6.4.13 A topological space is called *second-countable* if there is a countable basis for its topology, i.e., if there is a countable family $(U_i)_{i\geq 0}$ of open sets such that every open set of E is a union of open sets U_i.

It is easily seen that $\mathbb{R}^n$ is second-countable, and more generally, that every normed vector space of finite dimension is second-countable. It can also be shown that if E is a locally compact space that has a countable basis, then E_ω also has a countable basis (and in fact, is metrizable). We have the following properties.

Proposition 6.4.14 *Given a second-countable topological space E, every open cover $(U_i)_{i\in I}$ of E contains some countable subcover.*

Proof. Let $(O_n)_{n\geq 0}$ be a countable basis for the topology. Then, all sets O_n contained in some U_i can be arranged into a countable subsequence $(\Omega_m)_{m\geq 0}$ of $(O_n)_{n\geq 0}$, and for every Ω_m, there is some U_{i_m} such that $\Omega_m \subseteq U_{i_m}$. Furthermore, every U_i is some union of sets Ω_j, and thus, every $a \in E$ belongs to some Ω_j, which shows that $(\Omega_m)_{m\geq 0}$ is a countable open subcover of $(U_i)_{i\in I}$. □

As an immediate corollary of Proposition 6.4.14, a locally connected second-countable space has countably many connected components.

In second-countable Hausdorff spaces, compactness can be characterized in terms of accumulation points (this is also true of metric spaces).

Definition 6.4.15 Given a topological Hausdorff space E, given any sequence (x_n) of points in E, a point $l \in E$ is an *accumulation point (or cluster point)* of the sequence (x_n) if every open set U containing l contains x_n for infinitely many n.

Clearly, if l is a limit of the sequence (x_n), then it is an accumulation point, since every open set U containing a contains all x_n except for finitely many n.

Proposition 6.4.16 *A second-countable topological Hausdorff space E is compact iff every sequence (x_n) has some accumulation point.*

Proof. Assume that every sequence (x_n) has some accumulation point. Let $(U_i)_{i \in I}$ be some open cover of E. By Proposition 6.4.14, there is a countable open subcover $(O_n)_{n \geq 0}$ for E. Now, if E is not covered by any finite subcover of $(O_n)_{n \geq 0}$, we can define a sequence (x_m) by induction as follows:

Let x_0 be arbitrary, and for every $m \geq 1$, let x_m be some point in E not in $O_1 \cup \cdots \cup O_m$, which exists, since $O_1 \cup \cdots \cup O_m$ is not an open cover of E. We claim that the sequence (x_m) does not have any accumulation point. Indeed, for every $l \in E$, since $(O_n)_{n \geq 0}$ is an open cover of E, there is some O_m such that $l \in O_m$, and by construction, every x_n with $n \geq m+1$ does not belong to O_m, which means that $x_n \in O_m$ for only finitely many n, and l is not an accumulation point.

Conversely, assume that E is compact, and let (x_n) be any sequence. If $l \in E$ is not an accumulation point of the sequence, then there is some open set U_l such that $l \in U_l$ and $x_n \in U_l$ for only finitely many n. Thus, if (x_n) does not have any accumulation point, the family $(U_l)_{l \in E}$ is an open cover of E, and since E is compact, it has some finite open subcover $(U_l)_{l \in J}$, where J is a finite subset of E. But every U_l with $l \in J$ is such that $x_n \in U_l$ for only finitely many n, and since J is finite, $x_n \in \bigcup_{l \in J} U_l$ for only finitely many n, which contradicts the fact that $(U_l)_{l \in J}$ is an open cover of E, and thus contains all the x_n. Thus, (x_n) has some accumulation point. $\square$

Remark: It should be noted that the proof that if E is compact, then every sequence has some accumulation point, holds for any arbitrary compact space (the proof does not use a countable basis for the topology). The converse also holds for metric spaces. We prove this converse since it is a major property of metric spaces. It is also convenient to have such a characterization of compactness when dealing with fractal geometry.

Given a metric space in which every sequence has some accumulation point, we first prove the existence of a *Lebesgue number*.

Lemma 6.4.17 *Given a metric space E, if every sequence (x_n) has an accumulation point, for every open cover $(U_i)_{i\in I}$ of E, there is some $\delta > 0$ (a Lebesgue number for $(U_i)_{i\in I}$) such that, for every open ball $B_0(a,\epsilon)$ of diameter $\epsilon \leq \delta$, there is some open subset U_i such that $B_0(a,\epsilon) \subseteq U_i$.*

Proof. If there was no δ with the above property, then, for every natural number n, there would be some open ball $B_0(a_n, 1/n)$ which is not contained in any open set U_i of the open cover $(U_i)_{i\in I}$. However, the sequence (a_n) has some accumulation point a, and since $(U_i)_{i\in I}$ is an open cover of E, there is some U_i such that $a \in U_i$. Since U_i is open, there is some open ball of center a and radius ϵ contained in U_i. Now, since a is an accumulation point of the sequence (a_n), every open set containing a contain a_n for infinitely many n, and thus, there is some n large enough so that

$$1/n \leq \epsilon/2 \quad \text{and} \quad a_n \in B_0(a, \epsilon/2),$$

which implies that

$$B_0(a_n, 1/n) \subseteq B_0(a,\epsilon) \subseteq U_i,$$

a contradiction. □

By a previous remark, since the proof of Proposition 6.4.16 implies that in a compact topological space, every sequence has some accumulation point, by Lemma 6.4.17, in a compact metric space, every open cover has a Lebesgue number. This fact can be used to prove another important property of compact metric spaces, the uniform continuity theorem.

Definition 6.4.18 Given two metric spaces (E, d_E) and (F, d_F), a function $f\colon E \to F$ is *uniformly continuous* if for every $\epsilon > 0$, there is some $\eta > 0$, such that, for all $a, b \in E$,

$$\text{if} \quad d_E(a,b) \leq \eta \quad \text{then} \quad d_F(f(a), f(b)) \leq \epsilon.$$

The *uniform continuity theorem* can be stated as follows.

Theorem 6.4.19 *Given two metric spaces (E, d_E) and (F, d_F), if E is compact and $f\colon E \to F$ is a continuous function, then it is uniformly continuous.*

Proof. Consider any $\epsilon > 0$, and let $(B_0(y, \epsilon/2))_{y\in F}$ be the open cover of F consisting of open balls of radius $\epsilon/2$. Since f is continuous, the family

$$(f^{-1}(B_0(y, \epsilon/2)))_{y\in F}$$

is an open cover of E. Since, E is compact, by Lemma 6.4.17, there is a Lebesgue number δ such that for every open ball $B_0(a,\eta)$ of diameter $\eta \leq \delta$, then $B_0(a,\eta) \subseteq f^{-1}(B_0(y, \epsilon/2))$, for some $y \in F$. In particular, for any $a, b \in E$ such that $d_E(a,b) \leq \eta = \delta/2$, we have $a, b \in B_0(a, \delta)$, and thus $a, b \in f^{-1}(B_0(y, \epsilon/2))$, which implies that $f(a), f(b) \in B_0(y, \epsilon/2)$. But then, $d_F(f(a), f(b)) \leq \epsilon$, as desired. □

We now prove another lemma needed to obtain the characterization of compactness in metric spaces in terms of accumulation points.

Lemma 6.4.20 *Given a metric space E, if every sequence (x_n) has an accumulation point, then for every $\epsilon > 0$, there is a finite open cover $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ of E by open balls of radius ϵ.*

Proof. Let a_0 be any point in E. If $B_0(a_0, \epsilon) = E$, the lemma is proved. Otherwise, assume that a sequence $(a_0, a_1, \ldots, a_n)$ has been defined, such that $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ does not cover E. Then, there is some a_{n+1} not in $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$, and either

$$B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_{n+1}, \epsilon) = E,$$

in which case the lemma is proved, or we obtain a sequence $(a_0, a_1, \ldots, a_{n+1})$ such that $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_{n+1}, \epsilon)$ does not cover E. If this process goes on forever, we obtain an infinite sequence (a_n) such that $d(a_m, a_n) > \epsilon$ for all $m \neq n$. Since every sequence in E has some accumulation point, the sequence (a_n) has some accumulation point a. Then, for infinitely many n, we must have $d(a_n, a) \leq \epsilon/3$, and thus, for at least two distinct natural numbers p, q, we must have $d(a_p, a) \leq \epsilon/3$ and $d(a_q, a) \leq \epsilon/3$, which implies $d(a_p, a_q) \leq 2\epsilon/3$, contradicting the fact that $d(a_m, a_n) > \epsilon$ for all $m \neq n$. Thus, there must be some n such that

$$B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon) = E.$$

□

A metric space satisfying the condition of Lemma 6.4.20 is sometimes called *precompact* (or *totally bounded*). We now obtain the *Weierstrass-Bolzano* property.

Theorem 6.4.21 *A metric space E is compact iff every sequence (x_n) has an accumulation point.*

Proof. We already observed that the proof of Proposition 6.4.16 shows that for any compact space (not necessarily metric), every sequence (x_n) has an accumulation point. Conversely, let E be a metric space, and assume that every sequence (x_n) has an accumulation point. Given any open cover $(U_i)_{i \in I}$ for E, we must find a finite open subcover of E. By Lemma 6.4.17, there is some $\delta > 0$ (a Lebesgue number for $(U_i)_{i \in I}$) such that, for every open ball $B_0(a, \epsilon)$ of diameter $\epsilon \leq \delta$, there is some open subset U_j such that $B_0(a, \epsilon) \subseteq U_j$. By Lemma 6.4.20, for every $\delta > 0$, there is a finite open cover $B_0(a_0, \delta) \cup \cdots \cup B_0(a_n, \delta)$ of E by open balls of radius δ. But from the previous statement, every open ball $B_0(a_i, \delta)$ is contained in some open set U_{j_i}, and thus, $\{U_{j_1}, \ldots, U_{j_n}\}$ is an open cover of E. □

Another very useful characterization of compact metric spaces is obtained in terms of Cauchy sequences. Such a characterization is quite useful in fractal geometry (and elsewhere). First, recall the definition of a Cauchy sequence, and of a complete metric space.

Definition 6.4.22 Given a metric space (E, d), a sequence $(x_n)_{n \in \mathbb{N}}$ in E is a *Cauchy sequence* if the following condition holds:

for every $\epsilon > 0$, there is some $p \geq 0$, such that, for all $m, n \geq p$, then $d(x_m, x_n) \leq \epsilon$.

If every Cauchy sequence in (E, d) converges, we say that (E, d) is a *complete metric space*.

First, let us show the following easy proposition.

Proposition 6.4.23 *Given a metric space E, if a Cauchy sequence (x_n) has some accumulation point a, then a is the limit of the sequence (x_n).*

Proof. Since (x_n) is a Cauchy sequence, for every $\epsilon > 0$, there is some $p \geq 0$, such that, for all $m, n \geq p$, then $d(x_m, x_n) \leq \epsilon/2$. Since a is an accumulation point for (x_n), for infinitely many n, we have $d(x_n, a) \leq \epsilon/2$, and thus for at least some $n \geq p$, have $d(x_n, a) \leq \epsilon/2$. Then, for all $m \geq p$,

$$d(x_m, a) \leq d(x_m, x_n) + d(x_n, a) \leq \epsilon,$$

which shows that a is the limit of the sequence (x_n). □

Recall that a metric space is *precompact* (or *totally bounded*) if for every $\epsilon > 0$, there is a finite open cover $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ of E by open balls of radius ϵ. We can now prove the following theorem.

Theorem 6.4.24 *A metric space E is compact iff it is precompact and complete.*

Proof. Let E be compact. For every $\epsilon > 0$, the family of all open balls of radius ϵ is an open cover for E, and since E is compact, there is a finite subcover $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ of E by open balls of radius ϵ. Thus, E is precompact. Since E is compact, by Theorem 6.4.21, every sequence (x_n) has some accumulation point. Thus, every Cauchy sequence (x_n) has some accumulation point a, and by Proposition 6.4.23, a is the limit of (x_n). Thus, E is complete.

Now, assume that E is precompact and complete. We prove that every sequence (x_n) has an accumulation point. By the other direction of Theorem 6.4.21, this shows that E is compact. Given any sequence (x_n), we construct a Cauchy subsequence (y_n) of (x_n) as follows: Since E is precompact, letting $\epsilon = 1$, there exists a finite cover $\mathcal{U}_1$ of E by open balls of radius 1. Thus, some open ball B_o^1 in the cover $\mathcal{U}_1$ contains infinitely many elements from the sequence (x_n). Let y_0 be any element of (x_n) in B_o^1. By induction, assume that a sequence of open balls $(B_o^i)_{1 \leq i \leq m}$ has been defined, such that every ball B_o^i has radius $\frac{1}{2^i}$, contains infinitely many elements from the sequence (x_n), and contains some y_i from (x_n) such that

$$d(y_i, y_{i+1}) \leq \frac{1}{2^i},$$

for all i, $0 \leq i \leq m-1$. Then, letting $\epsilon = \frac{1}{2^{m+1}}$, because E is precompact, there is some finite cover $\mathcal{U}_{m+1}$ of E by open balls of radius ϵ, and thus of the open ball B_o^m. Thus, some open ball B_o^{m+1} in the cover $\mathcal{U}_{m+1}$ contains infinitely many elements from the sequence (x_n), and we let y_{m+1} be any element of (x_n) in B_o^{m+1}. Thus, we have defined by induction a sequence (y_n) which is a subsequence of (x_n), and such that

$$d(y_i, y_{i+1}) \leq \frac{1}{2^i},$$

for all i. However, for all $m, n \geq 1$, we have

$$d(y_m, y_n) \leq d(y_m, y_{m+1}) + \cdots + d(y_{n-1}, y_n) \leq \sum_{i=m} n\frac{1}{2^i} \leq \frac{1}{2^{m-1}},$$

and thus, (y_n) is a Cauchy sequence. Since E is complete, the sequence (y_n) has a limit, and since it is a subsequence of (x_n), the sequence (x_n) has some accumulation point. □

If (E, d) is a nonempty complete metric space, every map $f\colon E \to E$ for which there is some k such that $0 \leq k < 1$ and

$$d(f(x), f(y)) \leq kd(x, y)$$

for all $x, y \in E$, has the very important property that it has a unique fixed point, that is, there is a unique $a \in E$ such that $f(a) = a$. A map as above is called a *contracting mapping*. Furthermore, the fixed point of a contracting mapping can be computed as the limit of a fast converging sequence.

The fixed point property of contracting mappings is used to show some important theorems of analysis, such as the implicit function theorem, and the existence of solutions to certain differential equations. It can also be used to show the existence of fractal sets defined in terms of iterated function systems, a topic that we intend to discuss later on. Since the proof is quite simple, we prove the fixed point property of contracting mappings. First, observe that a contracting mapping is (uniformly) continuous.

Proposition 6.4.25 *If (E, d) is a nonempty complete metric space, every contracting mapping $f\colon E \to E$ has a unique fixed point. Furthermore, for every $x_0 \in E$, defining the sequence (x_n) such that $x_{n+1} = f(x_n)$, the sequence (x_n) converges to the unique fixed point of f.*

Proof. First, we prove that f has at most one fixed point. Indeed, if $f(a) = a$ and $f(b) = b$, since

$$d(a, b) = d(f(a), f(b)) \leq kd(a, b)$$

and $0 \leq k < 1$, we must have $d(a, b) = 0$, that is, $a = b$.

Next, we prove that (x_n) is a Cauchy sequence. Observe that

$$\begin{aligned}
d(x_2, x_1) &\leq kd(x_1, x_0),\\
d(x_3, x_2) &\leq kd(x_2, x_1) \leq k^2 d(x_1, x_0),\\
\cdots &\quad \cdots\\
d(x_{n+1}, x_n) &\leq kd(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
d(x_{n+p}, x_n) &\leq d(x_{n+p}, x_{n+p-1}) + d(x_{n+p-1}, x_{n+p-2}) + \cdots + d(x_{n+1}, x_n)\\
&\leq (k^{p-1} + k^{p-2} + \cdots + k + 1)k^n d(x_1, x_0)\\
&\leq \frac{k^n}{1-k}\, d(x_1, x_0).
\end{aligned}$$

We conclude that $d(x_{n+p}, x_n)$ converges to 0 when n goes to infinity, which shows that (x_n) is a Cauchy sequence. Since E is complete, the sequence (x_n) has a limit a. Since f is continuous, the sequence $(f(x_n))$ converges to $f(a)$. But $x_{n+1} = f(x_n)$ converges to a, and so $f(a) = a$, the unique fixed point of f. □

Note that no matter how the starting point x_0 of the sequence (x_n) is chosen, (x_n) converges to the unique fixed point of f. Also, the convergence is fast, since

$$d(x_n, a) \leq \frac{k^n}{1-k}\, d(x_1, x_0).$$

The Hausdorff distance between compact subsets of a metric space provides a very nice illustration of some of the theorems on complete and compact metric spaces just presented. It can also be used to define certain kinds of fractal sets, and thus, we indulge into a short digression on the Hausdorff distance.

Definition 6.4.26 Given a metric space (X, d), for any subset $A \subseteq X$, for any $\epsilon \geq 0$, define the *ϵ-hull* of A, as the set

$$V_\epsilon(A) = \{x \in X,\ \exists a \in A \mid d(a, x) \leq \epsilon\}.$$

Given any two nonempty bounded subsets A, B of X, define $D(A, B)$, *the Hausdorff distance between A and B*, as

$$D(A, B) = \inf\{\epsilon \geq 0 \mid A \subseteq V_\epsilon(B) \text{ and } B \subseteq V_\epsilon(A)\}.$$

Note that since we are considering nonempty bounded subsets, $D(A, B)$ is well defined (i.e., not infinite). However, D is not necessarily a distance function. It is a distance function if we restrict our attention to nonempty compact subsets of X. We let $\mathcal{K}(X)$ denote the set of all nonempty compact subsets of X. The remarkable fact is that D is a distance on $\mathcal{K}(X)$, and that if X is complete or compact, then so it $\mathcal{K}(X)$. The following theorem is taken from Edgar [6].

Theorem 6.4.27 *If (X, d) is a metric space, then the Hausdorff distance D on the set $\mathcal{K}(X)$ of nonempty compact subsets of X is a distance. If (X, d) is complete, then $(\mathcal{K}(X), D)$ is complete, and if (X, d) is compact, then $(\mathcal{K}(X), D)$ is compact.*

Proof. Since (nonempty) compact sets are bounded, $D(A, B)$ is well defined. Clearly, D is symmetric. Assume that $D(A, B) = 0$. Then, for every $\epsilon > 0$, $A \subseteq V_\epsilon(B)$, which means that for every $a \in A$, there is some $b \in B$ such that $d(a, b) \leq \epsilon$, and thus, that $A \subseteq \overline{B}$. Since B is closed, $\overline{B} = B$, and we have $A \subseteq B$. Similarly, $B \subseteq A$, and thus, $A = B$. Clearly, if $A = B$, we have $D(A, B) = 0$. It remains to prove the triangle inequality. If $B \subseteq V_{\epsilon_1}(A)$ and $C \subseteq V_{\epsilon_2}(B)$, then

$$V_{\epsilon_2}(B) \subseteq V_{\epsilon_2}(V_{\epsilon_1}(A)),$$

and since

$$V_{\epsilon_2}(V_{\epsilon_1}(A)) \subseteq V_{\epsilon_1+\epsilon_2}(A),$$

we get

$$C \subseteq V_{\epsilon_2}(B) \subseteq V_{\epsilon_1+\epsilon_2}(A).$$

Similarly, we can prove that

$$A \subseteq V_{\epsilon_1+\epsilon_2}(C),$$

and thus, the triangle inequality follows.

Next, we need to prove that if (X, d) is complete, then $(\mathcal{K}(X), D)$ is also complete. First, we show that if (A_n) is a sequence of nonempty compact sets converging to a nonempty compact set A in the Hausdorff metric, then

$$A = \{x \in X \mid \text{there is a sequence } (x_n) \text{ with } x_n \in A_n \text{ converging to } x\}.$$

Indeed, if (x_n) is a sequence with $x_n \in A_n$ converging to x and (A_n) converges to A, then for every $\epsilon > 0$, there is some x_n such that $d(x_n, x) \leq \epsilon/2$, and there is some $a_n \in A$ such that $d(a_n, x_n) \leq \epsilon/2$, and thus, $d(a_n, x) \leq \epsilon$, which shows that $x \in \overline{A}$. Since A is compact, it is closed, and $x \in A$. Conversely, since (A_n) converges to A, for every $x \in A$, for every $n \geq 1$, there is some $x_n \in A_n$ such that $d(x_n, x) \leq 1/n$, and the sequence (x_n) converges to x.

Now, let (A_n) be a Cauchy sequence in $\mathcal{K}(X)$. It can be proven that (A_n) converges to the set

$$A = \{x \in X \mid \text{there is a sequence } (x_n) \text{ with } x_n \in A_n \text{ converging to } x\},$$

and that A is nonempty and compact. To prove that A is compact, one proves that it is totally bounded and complete. Details are given in Edgar [6].

Finally, we need to prove that if (X, d) is compact, then $(\mathcal{K}(X), D)$ is compact. Since we already know that $(\mathcal{K}(X), D)$ is complete if (X, d) is, it is enough to prove that $(\mathcal{K}(X), D)$ is totally bounded if (X, d) is, which is fairly easy. □

In view of Theorem 6.4.27 and Theorem 6.4.25, it is possible to define some nonempty compact subsets of X in terms of fixed points of contracting maps. We will see later on how this can be done in terms of iterated function systems, yielding a large class of fractals.

Finally, returning to second-countable spaces, we give another characterization of accumulation points.

Proposition 6.4.28 *Given a second-countable topological Hausdorff space E, a point l is an accumulation point of the sequence (x_n) iff l is the limit of some subsequence (x_{n_k}) of (x_n).*

Proof. Clearly, if l is the limit of some subsequence (x_{n_k}) of (x_n), it is an accumulation point of (x_n).

Conversely, let $(U_k)_{k\geq 0}$ be the sequence of open sets containing l, where each U_k belongs to a countable basis of E, and let $V_k = U_1 \cap \cdots \cap U_k$. For every $k \geq 1$, we can find some $n_k > n_{k-1}$ such that $x_{n_k} \in V_k$, since l is an accumulation point of (x_n). Now, since every open set containing l contains some U_{k_0}, and since $x_{n_k} \in U_{k_0}$ for all $k \geq 0$, the sequence (x_{n_k}) has limit l. □

Remark: Proposition 6.4.28 also holds for metric spaces.

As promised, we show how certain fractals can be defined by iterated function systems, using Theorem 6.4.27 and Theorem 6.4.25.

Chapter 7

A Detour On Fractals

7.1 Iterated Function Systems and Fractals

A pleasant application of the Hausdorff distance and of the fixed point theorem for contracting mappings is a method for defining a class of "self-similar" fractals. For this, we can use iterated function systems.

Definition 7.1.1 Given a metric space (X, d), an *iterated function system*, for short, an *ifs*, is a finite sequence of functions $(f_1, \ldots, f_n)$, where each $f_i \colon X \to X$ is a contracting mapping. A nonempty compact subset K of X is an *invariant set (or attractor)* for the ifs $(f_1, \ldots, f_n)$ if

$$K = f_1(K) \cup \cdots \cup f_n(K).$$

The major result about ifs's is the following.

Theorem 7.1.2 *If (X, d) is a nonempty complete metric space, every iterated function system $(f_1, \ldots, f_n)$ has a unique invariant set A which is a nonempty compact subset of X. Furthermore, for every nonempty compact subset A_0 of X, this invariant set A if the limit of the sequence (A_m), where $A_{m+1} = f_1(A_m) \cup \cdots \cup f_n(A_m)$.*

Proof. Since X is complete, by Theorem 6.4.27, the space $(\mathcal{K}(X), D)$ is a complete metric space. The theorem will follow from Theorem 6.4.25, if we can show that the map $F \colon \mathcal{K}(X) \to \mathcal{K}(X)$ defined such that

$$F(K) = f_1(K) \cup \cdots \cup f_n(K),$$

for every nonempty compact set K, is a contracting mapping. Let A, B be any two nonempty compact subsets of X, and consider any $\eta \geq D(A, B)$. Since each $f_i \colon X \to X$ is a contracting mapping, there is some λ_i, with $0 \leq \lambda_i < 1$, such that

$$d(f_i(a), f_i(b)) \leq \lambda_i d(a, b),$$

for all $a, b \in X$. Let $\lambda = \max\{\lambda_1, \ldots, \lambda_n\}$. We claim that

$$D(F(A), F(B)) \leq \lambda D(A, B).$$

For any $x \in F(A) = f_1(A) \cup \cdots \cup f_n(A)$, there is some $a_i \in A_i$ such that $x = f_i(a_i)$, and since $\eta = D(A, B)$, there is some $b_i \in B$ such that

$$d(a_i, b_i) \leq \eta,$$

and thus,

$$d(x, f_i(b_i)) = d(f_i(a_i), f_i(b_i)) \leq \lambda_i d(a_i, b_i) \leq \lambda\eta.$$

This show that

$$F(A) \subseteq V_{\lambda\eta}(F(B)).$$

Similarly, we can prove that

$$F(B) \subseteq V_{\lambda\eta}(F(A)),$$

and since this holds for all $\eta \geq D(A, B)$, we proved that

$$D(F(A), F(B)) \leq \lambda D(A, B)$$

where $\lambda = \max\{\lambda_1, \ldots, \lambda_n\}$. Since $0 \leq \lambda_i < 1$, we have $0 \leq \lambda < 1$, and F is indeed a contracting mapping. □

Theorem 7.1.2 justifies the existence of many familiar "self-similar" fractals. One of the best known fractals is the *Sierpinski gasket*.

Example 7.1 Consider an equilateral triangle with vertices a, b, c, and let f_1, f_2, f_3 be the dilatations of centers a, b, c and ratio 1/2. The Sierpinski gasket is the invariant set of the ifs (f_1, f_2, f_3). The dilations f_1, f_2, f_3 can be defined explicitly as follows, assuming that $a = (-1/2, 0)$, $b = (1/2, 0)$, and $c = (0, \sqrt{3}/2)$. The contractions f_a, f_b, f_c are specified by

$$\begin{aligned} x' &= \frac{1}{2}x - \frac{1}{4}, \\ y' &= \frac{1}{2}y, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{2}x + \frac{1}{4}, \\ y' &= \frac{1}{2}y, \end{aligned}$$

and

$$\begin{aligned} x' &= \frac{1}{2}x, \\ y' &= \frac{1}{2}y + \frac{\sqrt{3}}{4}. \end{aligned}$$

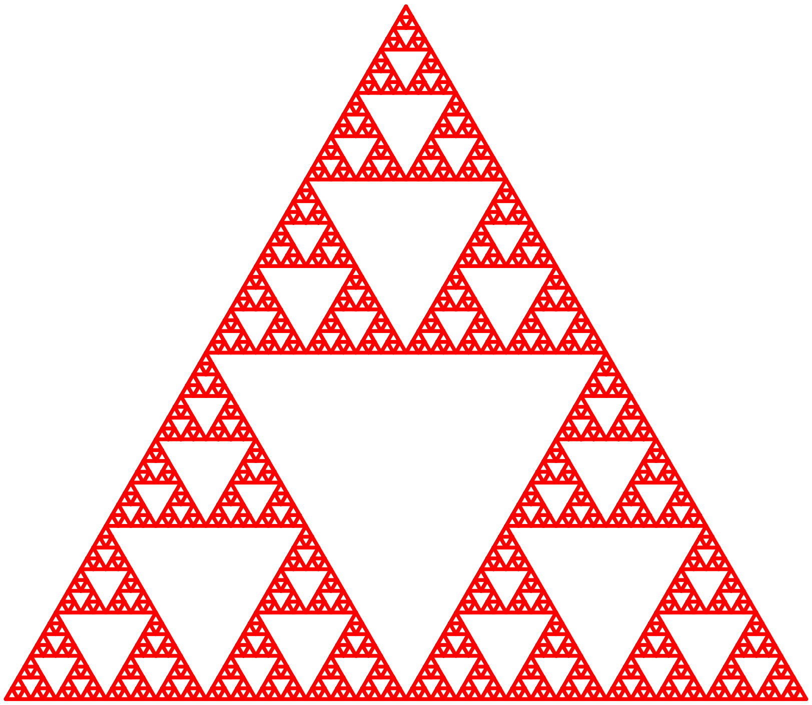

Figure 7.1: The Sierpinski gasket

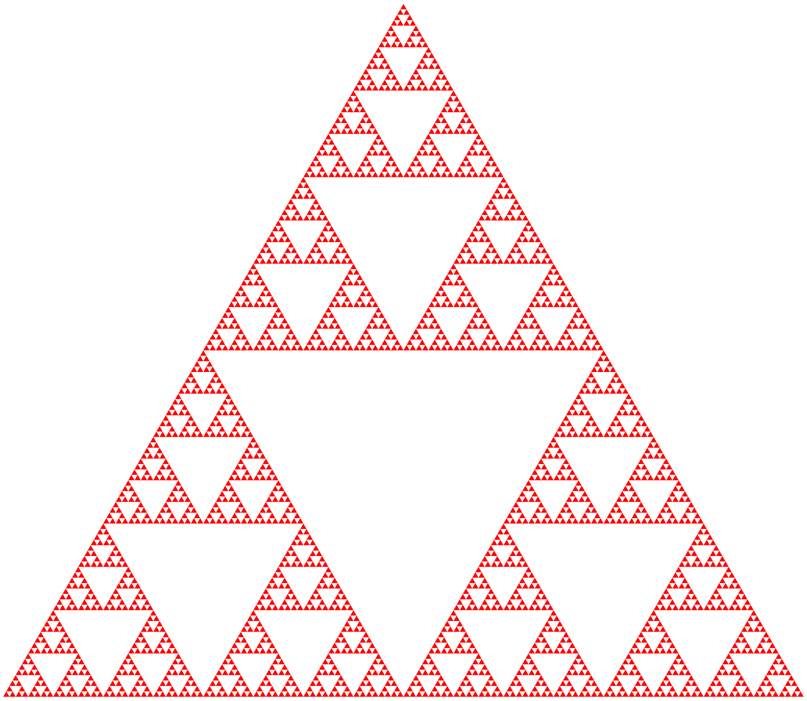

Figure 7.2: The Sierpinski gasket, version 2

We wrote a *Mathematica* program that iterates any finite number of affine maps on any input figure consisting of combinations of points, line segments, and polygons (with their interior points). Starting with the edges of the triangle a, b, c, after 6 iterations, we get the picture shown in Figure 7.1.

It is amusing that the same fractal is obtained no matter what the initial nonempty compact figure is. It is interesting to see what happens if we start with a solid triangle (with its interior points). The result after 6 iterations is shown in Figure 7.2.

The convergence towards the Sierpinski gasket is very fast. Incidently, there are many other ways of defining the Sierpinski gasket.

A nice variation on the theme of the Sierpinski gasket is the *Sierpinski dragon*.

Example 7.2 The Sierpinski dragon is specified by the following three contractions:

$$\begin{aligned} x' &= -\frac{1}{4}x - \frac{\sqrt{3}}{4}y + \frac{3}{4}, \\ y' &= \frac{\sqrt{3}}{4}x - \frac{1}{4}y + \frac{\sqrt{3}}{4}, \end{aligned}$$

$$\begin{aligned} x' &= -\frac{1}{4}x + \frac{\sqrt{3}}{4}y - \frac{3}{4}, \\ y' &= -\frac{\sqrt{3}}{4}x - \frac{1}{4}y + \frac{\sqrt{3}}{4}, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{2}x, \\ y' &= \frac{1}{2}y + \frac{\sqrt{3}}{2}. \end{aligned}$$

The result of 7 iterations starting from the line segment $(-1, 0), (1, 0))$, is shown in Figure 7.3.

This curve converges to the boundary of the Sierpinski gasket.

A different kind of fractal is the *Heighway dragon*.

Example 7.3 The Heighway dragon is specified by the following two contractions:

$$\begin{aligned} x' &= \frac{1}{2}x - \frac{1}{2}y, \\ y' &= \frac{1}{2}x + \frac{1}{2}y, \end{aligned}$$

$$\begin{aligned} x' &= -\frac{1}{2}x - \frac{1}{2}y, \\ y' &= \frac{1}{2}x - \frac{1}{2}y + 1. \end{aligned}$$

It can be shown that for any number of iterations, the polygon does not cross itself. This means that no edge is traversed twice, and that if a point is traversed twice, then this point is the endpoint of some edge. The result of 13 iterations, starting with the line segment $((0, 0), (0, 1))$, is shown in Figure 7.4.

The Heighway dragon turns out to fill a closed and bounded set. It can also be shown that the plane can be tiled with copies of the Heighway dragon.

Another well known example is the *Koch curve*.

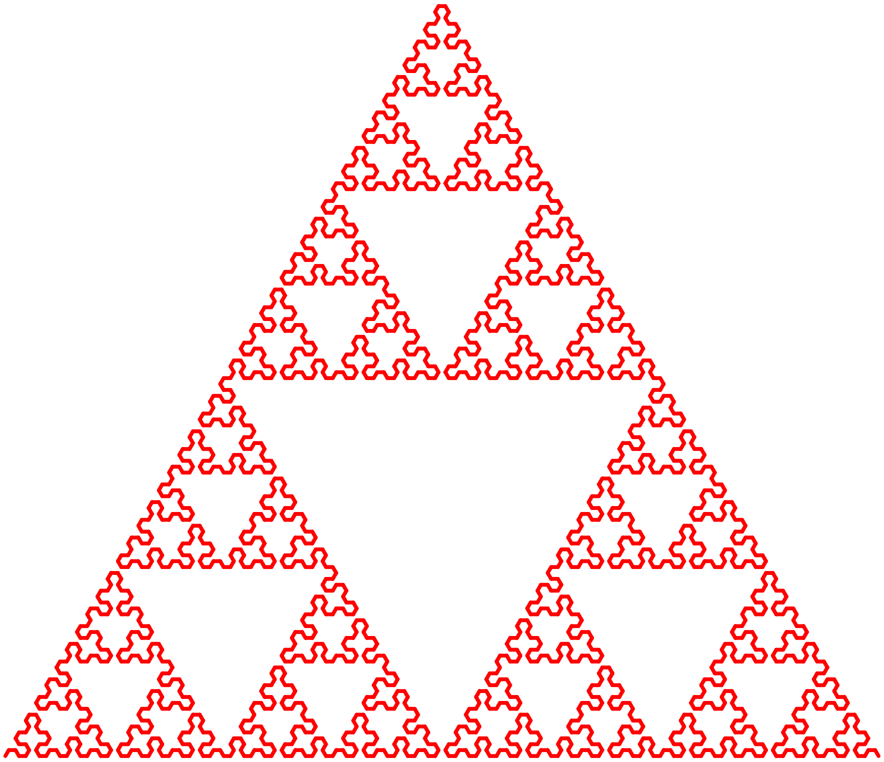

Figure 7.3: The Sierpinski dragon

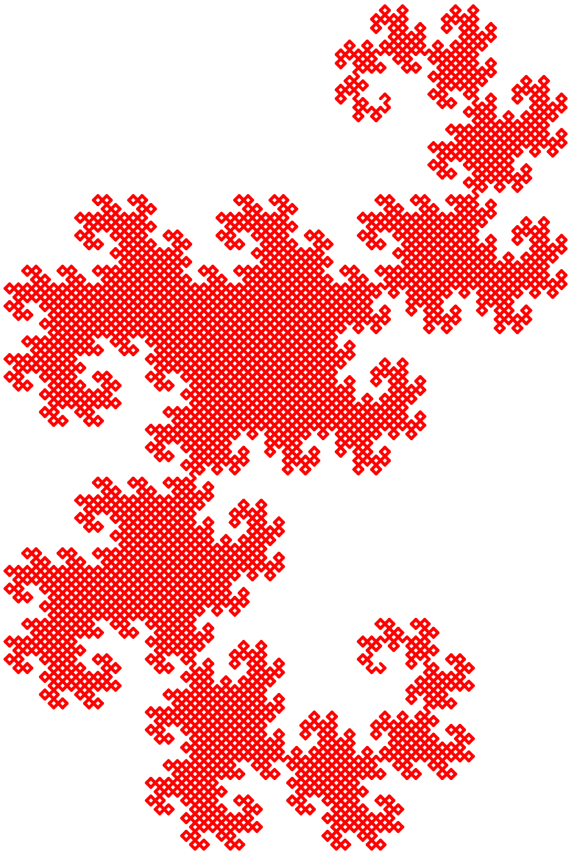

Figure 7.4: The Heighway dragon

Example 7.4 The Koch curve is specified by the following four contractions:

$$\begin{aligned} x' &= \frac{1}{3}x - \frac{2}{3}, \\ y' &= \frac{1}{3}y, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{6}x - \frac{\sqrt{3}}{6}y - \frac{1}{6}, \\ y' &= \frac{\sqrt{3}}{6}x + \frac{1}{6}y + \frac{\sqrt{3}}{6}, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{6}x + \frac{\sqrt{3}}{6}y + \frac{1}{6}, \\ y' &= -\frac{\sqrt{3}}{6}x + \frac{1}{6}y + \frac{\sqrt{3}}{6}, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{3}x + \frac{2}{3}, \\ y' &= \frac{1}{3}y, \end{aligned}$$

The Koch curve is an example of a continuous curve which is nowhere differentiable (because it “wiggles” too much). It is a curve of infinite length. The result of 6 iterations, starting with the line segment $((-1, 0), (1, 0))$, is shown in Figure 7.5.

The curve obtained by putting three Kock curves together on the sides of an equilateral triangle is known as the *snowflake curve* (for obvious reasons, see below!).

Example 7.5 The snowflake curve obtained after 5 iterations is shown in Figure 7.6.

The snowflake curve is an example of a closed curve of infinite length bounding a finite area.

We conclude with another famous example, a variant of the *Hilbert curve*.

Example 7.6 This version of the Hilbert curve is defined by the following four contractions:

$$\begin{aligned} x' &= \frac{1}{2}x - \frac{1}{2}, \\ y' &= \frac{1}{2}y + 1, \end{aligned}$$

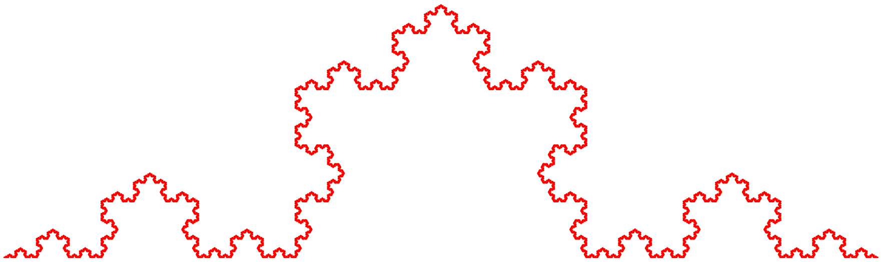

Figure 7.5: The Koch curve

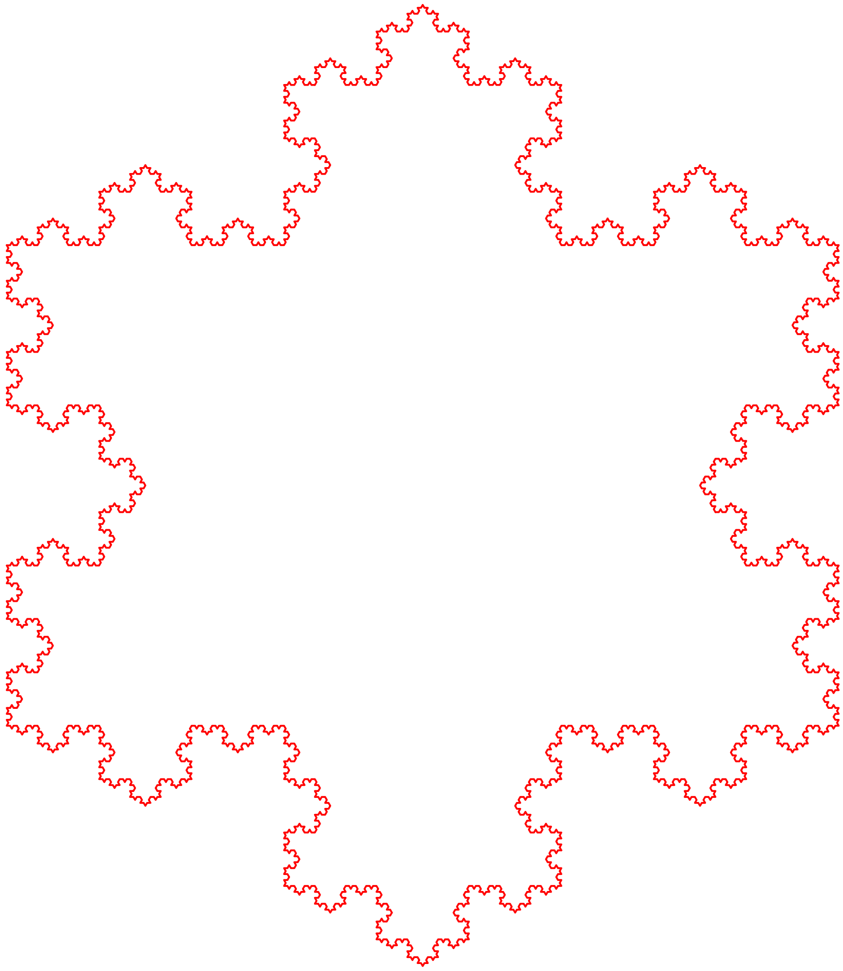

Figure 7.6: The snowflake curve

$$\begin{aligned} x' &= \frac{1}{2}x + \frac{1}{2}, \\ y' &= \frac{1}{2}y + 1, \end{aligned}$$

$$\begin{aligned} x' &= -\frac{1}{2}y + 1, \\ y' &= \frac{1}{2}x + \frac{1}{2}, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{2}y - 1, \\ y' &= -\frac{1}{2}x + \frac{1}{2}, \end{aligned}$$

This continuous curve is a space-filling curve, in the sense that its image is the entire unit square. The result of 6 iterations, starting with the two lines segments $((-1, 0), (0, 1))$ and $((0, 1), (1, 0))$, is shown in Figure 7.7.

For more on iterated function systems and fractals, we recommend Edgar [6].

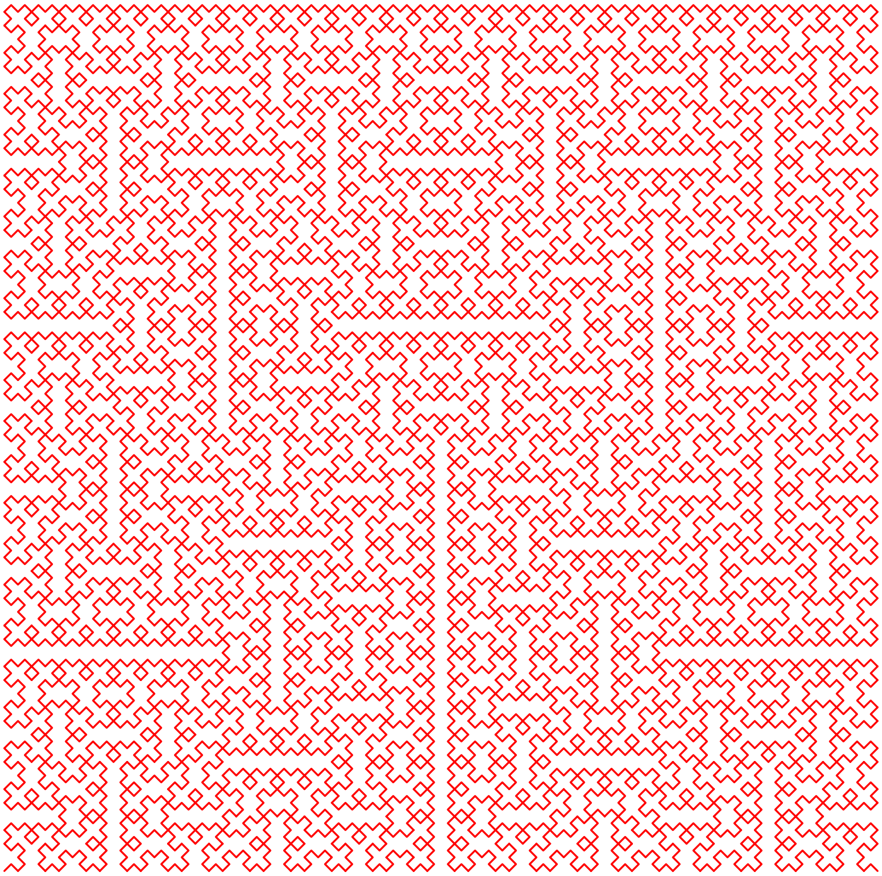

Figure 7.7: A Hilbert curve

Bibliography

[1] Lars V. Ahlfors and Leo Sario. *Riemann Surfaces.* Princeton Math. Series, No. 2. Princeton University Press, 1960.

[2] Mark A. Amstrong. *Basic Topology.* UTM. Springer, first edition, 1983.

[3] Glen E Bredon. *Topology and Geometry.* GTM No. 139. Springer Verlag, first edition, 1993.

[4] Jacques Dixmier. *General Topology.* UTM. Springer Verlag, first edition, 1984.

[5] Albrecht Dold. *Lectures on Algebraic Topology.* Springer, second edition, 1980.

[6] Gerald A. Edgar. *Measure, Topology, and Fractal Geometry.* Undergraduate Texts in Mathematics. Springer Verlag, first edition, 1992.

[7] William Fulton. *Algebraic Topology, A first course.* GTM No. 153. Springer Verlag, first edition, 1995.

[8] D. Hilbert and S. Cohn-Vossen. *Geometry and the Imagination.* Chelsea Publishing Co., 1952.

[9] L. Christine Kinsey. *Topology of Surfaces.* UTM. Springer Verlag, first edition, 1993.

[10] Serge Lang. *Undergraduate Analysis.* UTM. Springer Verlag, second edition, 1997.

[11] William S. Massey. *Algebraic Topology: An Introduction.* GTM No. 56. Springer Verlag, second edition, 1987.

[12] William S. Massey. *A Basic Course in Algebraic Topology.* GTM No. 127. Springer Verlag, first edition, 1991.

[13] James R. Munkres. *Topology, a First Course.* Prentice Hall, first edition, 1975.

[14] James R. Munkres. *Elements of Algebraic Topology.* Addison-Wesley, first edition, 1984.

[15] Joseph J. Rotman. *Introduction to Algebraic Topology.* GTM No. 119. Springer Verlag, first edition, 1988.

[16] Hajime Sato. *Algebraic Topology: An Intuitive Approach.* MMONO No. 183. AMS, first edition, 1999.

[17] Laurent Schwartz. *Analyse I. Théorie des Ensembles et Topologie.* Collection Enseignement des Sciences. Hermann, 1991.

[18] H. Seifert and W. Threlfall. *A Textbook of Topology.* Academic Press, first edition, 1980.

[19] Isadore M. Singer and John A. Thorpe. *Lecture Notes on Elementary Topology and Geometry.* UTM. Springer Verlag, first edition, 1976.

[20] Williams P. Thurston. *Three-Dimensional Geometry and Topology.* Princeton Math. Series, No. 35. Princeton University Press, 1997.