university of groningen

faculty of mathematics and natural sciences

Contractibility and Self-Intersections of Curves on Surfaces

David de Laat

Bachelor Thesis in Mathematics

August, 2009

Contractibility and Self-Intersections of Curves on Surfaces

Summary

We discuss whether closed curves on closed orientable surfaces are contractible, and for non-contractible curves whether they are homotopic to a curve having no self-intersections. We prove that the minimal number of self-intersections of an (m, n)-torus curve is $\gcd(m, n) - 1$. We discuss Dehn's algorithm for solving the word problem in the fundamental group, which is the algebraic equivalent of the contractibility problem, and Poincaré's solution of the problem concerning intersection-free curves.

Before doing this we develop the theory of curves on surfaces. By triangulating the surfaces we can associate them with normal form schemata and use this to realize the surfaces geometrically. We construct a locally isometric map from the spherical, Euclidean, or hyperbolic covering 2-spaces onto the surfaces and construct a group of isometries on this covering space. This map is used to give a characterization of homotopy and we prove that the fundamental group, consisting of the homotopy equivalence classes, is isomorphic to the covering isometry group. The Cayley graph explains the structure of the fundamental group and is important in the development of Dehn's algorithm. The geometric properties of the covering isometry group are important in Poincaré's approach.

Bachelor Thesis in Mathematics
Author: David de Laat
Supervisor: Gert Vegter
Date: August, 2009

Institute of Mathematics and Computing Science
P.O. Box 407
9700 AK Groningen
The Netherlands

Contents

Introduction

Imagine a rubber band knotted around a ball, a doughnut, or another three-dimensional object with any number of holes. Assume that the band fits tightly around the object, as if it where glued to the surface. When the object is a ball then we know for sure that we can pull the band off the object without having to cut it loose. When the object has one or more holes this might or might not be possible. This question is the content of chapter 4. When a band can not be pulled off the object we can still move it around, varying the number of self-crossings. In chapter 5 we find a way to see whether the minimal number of self-crossings of a band is zero and for the special case where the surface has one hole (a doughnut) we also find an expression for what this minimum is.

To state these problems in a mathematical precise way and to solve them efficiently we need quite a lot of background theory, chapters 1, 2, and 3 give an overview of this theory. Chapter 1 is concerned with closed orientable surfaces and how to represent them by schemata. The closed orientable surfaces are the surfaces corresponding with the surfaces of three-dimensional objects like balls or doughnuts, for simplicity we will restrict ourselves to these surfaces even if a result holds for a more general case.

Then in chapter 2 we introduce geometric surfaces and we use the results of the previous chapter to realize the closed orientable surfaces geometrically. We use this to construct a locally isometric map, denoted the pencil map, from the spherical, Euclidean, or hyperbolic covering space onto the surface and prove that this map is a covering map. Although it is possible to construct such a covering map by topological means only, we choose to use a more geometric approach since we will need the local isometry property in a alter chapter. We use this map to construct the covering isometry group.

In chapter 3 we finally introduce curves which represent the rubber bands. We also introduce homotopy which corresponds with moving the rubber band on the surface without cutting it open. We use the pencil map and the covering surface to give a characterisation for homotopic curves. By doing this it starts to become clear why the theory that we have build up so far is useful when discussing contractibility and self-intersections. We then divide the set of closed curves on the surface into homotopy classes to form the fundamental group and we show that this group is isomorphic to the covering isometry group. Finally we discuss the Cayley graph which forms a bridge between the group structure of the fundamental group and the geometric properties of the covering isometry group.

In the presentation of this theory we focus on our goal of solving the contractibility and self-intersections problems. This means that almost all results are used later on to solve these two problems. We focus on explaining the idea behind the theory and leave out some technical proofs, referring to the literature instead. We use particularly many results from John Stillwell's books [10] and [9].

At this point we have developed enough theory to state and solve the contractibility

problem. We define a contractible curve to be a curve that is homotopic to the constant curve and use the pencil map to give a characterisation of contractible curves. Then we discuss the more computational side of the problem by giving an algorithm that decides, in a finite number of steps, whether a curve is contractible. Finally we describe Dehn's algorithm which uses the Cayley graph to give the steps required to contract a curve.

In the last chapter we start by defining what self-intersections of curves are and prove that contractible curves are homotopic to a curve having no self-intersections. In [9] a proof is given of the theorem that the minimal number of self-intersections of a (m, n)-torus is zero if and only if m and n are relatively prime. We extend this by proving that the minimal number of self-intersections of a (m, n)-torus curve is $\gcd(m, n) - 1$. We finish by describing a method for determining whether a curve on a surface with more then one hole is homotopic to a curve without self-intersections.

Chapter 1

Surfaces and schemata

We start by defining closed orientable surfaces and we will give an idea of what these surfaces are by giving examples and counter examples. Then we continue by defining schemata, which are combinatorial representations of surfaces. We will introduce the so called identification space corresponding to a schema and we will show that, up to homeomorphism, there is exactly one identification space corresponding to each schema. By defining a metric on an identification space we prove that such a space is a surface.

Then we introduce the normal form schemata, which form a subset of all schemata and we will show that the g-th normal form schema represents a surface with g holes. After that we introduce the concept of a triangulation and show that any closed orientable surface can be triangulated. In the last section of this chapter we use this triangulation to show that any closed orientable surface is homeomorphic to the identification space of a normal form schema.

1.1 Closed orientable surfaces

A *surface* is a Hausdorff topological space in which every point has an open neighbourhood homeomorphic to some open subset of the plane. Examples of surfaces are the sphere

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = r^2\} \tag{1.1}$$

and the torus, parametrized by

$$\{((R + r\cos v)\cos u, (R + r\cos v)\sin u, r\sin v) \in \mathbb{R}^3 \mid u, v \in [0, 2\pi]\}, \tag{1.2}$$

with $0 < r < R$. To see that the sphere S is a surface, we note that it is a subset of $\mathbb{R}^3$, so it is a Hausdorff space, and any point $X \in S$ has an open neighbourhood homeomorphic to some open subset of the plane. This can be seen, for instance, by using stereographic projection in the antipodal point of X to map an open neighbourhood of X to the plane.

The double cone is not a surface since the point where the two cones meet does not have an open neighbourhood homeomorphic to some open subset of the plane. Another example of a Hausdorff space that is not a surface is the closed unit disk $D = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$. Points on the boundary of this disk do not have a neighbourhood homeomorphic to some open subset of the plane. The closed unit disk, however, is what is called a surface with boundary.

A *surface with boundary* is a Hausdorff space on which every point has an open neighbour-

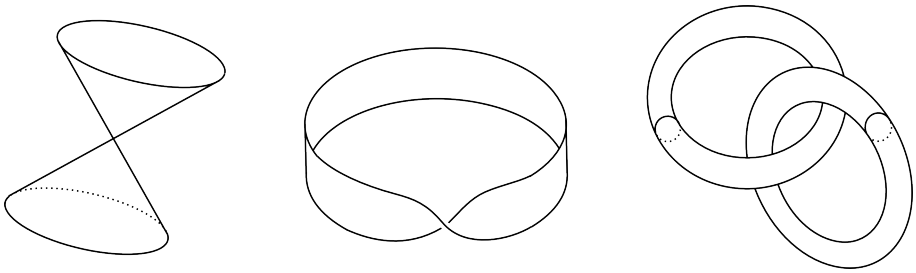


Figure 1.1: The double cone, the Möbius strip, and linked tori

hood homeomorphic to some open subset of the upper half plane $\{(x, y) \in \mathbb{R}^2 \mid y \geq 0\}$. In the closed unit disk example, points in the interior of the disk have a neighbourhood homeomorphic to some open neighbourhood of an interior point of the upper half plane, while points on the unit circle have a neighbourhood homeomorphic to an open neighbourhood of a point on the x-axis. Note that the word 'boundary' here does not refer to the usual topological definition of boundary.

A surface is *compact* when it is compact as a topological space. The sphere is a compact surface since it is a subset of $\mathbb{R}^3$ which means that it is compact if and only if it is closed and bounded, which it clearly is. The sphere without north pole

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\} - \{(0, 0, 1)\}$$

is not compact because it is not closed as a topological space.

We call a surface *orientable* if a consistent concept of clockwise rotation can be defined on the surface in a continuous manner. The sphere, for instance, is an orientable surface since its normal varies continuously when moving over the surface. The Möbius strip, however, is a compact surface with boundary but is not orientable.

When we can connect any two points in a surface by a path that lies in the surface, then the surface is *connected*. The surface consisting of two linked tori, for example, is disconnected.

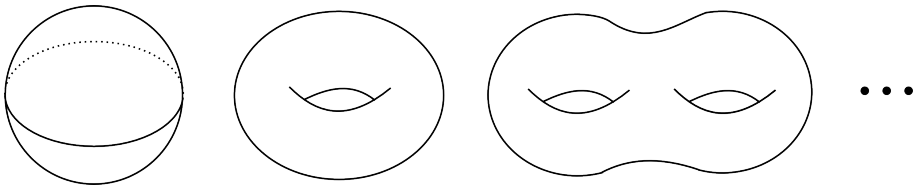


Figure 1.2: The closed orientable surfaces

In this report we will only discuss compact connected orientable surfaces (without boundary), which we will call *closed orientable surfaces* (closed, again, does not have the usual topological meaning). After having seen examples of Hausdorff spaces that are not surfaces, not compact, not connected, not orientable, or have a boundary, we might wonder which spaces *are* closed orientable surfaces. We will see that, up to homeomorphism, each closed orientable surface is a sphere with $g \geq 0$ handles. The sphere with zero handles is usually just called the sphere, the sphere with one handle the torus, and the sphere with two handles the double torus. The number of handles that a surface has is called the *genus* of the surface. We will see that two surfaces with an unequal number of handles are not homeomorphic. That

is, a closed orientable surface is completely determined by its genus. Consequences of this are that there are, up to homeomorphism, only countably many closed orientable surfaces and that it is possible to embed any closed orientable surface in $\mathbb{R}^3$.

1.2 Schemata

A *schema* Π is a finite collection of ordered sets each containing a finite number of labels. Each label occurs exactly twice in the collection and each label can be inverted. An example of a schema is the collection:

$$\Pi = \{(e_3, e_2^{-1}, e_1^{-1}, e_2, e_4), (e_1^{-1}, e_3^{-1}, e_4^{-1})\}. \tag{1.3}$$

Note that, as the example shows, it is not required that for each label its inverse also occurs. The possibilities of a label occurring twice not inverted, once inverted and once not inverted, and twice inverted, are all allowed.

We construct a set P_Π corresponding to a schema Π by constructing the ordered sets of labels $\pi_i \in \Pi$ as disjoint regular polygons in the plane where each edge has equal length. The polygon corresponding to π_i is constructed to have $\#\pi_i$ edges which are assigned, in a clockwise manner, the labels of the corresponding ordered set π_i. When the label is an inverse then the edge is given a clockwise orientation and otherwise it is given an anticlockwise orientation, see Figure 1.3.

The set P_Π is a compact subspace of $\mathbb{R}^2$ but, since it has a boundary, is not a surface. Let S_Π be another space which is the same as P_Π except that we identify the edges that have the same label. We make these identifications according to the orientations of the edges, such that, for instance, the point that is the front of an edge is identified with the point that is the front of the edge bearing the same label. Since each label occurs exactly twice in a schema the edges are identified in pairs. This means that each point in the interior of an edge is identified with precisely one other point that is also an interior edge point. The edge identifications however can imply that more than two vertices are identified to each other. That is, the edge identifications subdivide the set of vertices in so called equivalence classes of identified vertices. The space S_Π is called the *identification space* of Π and consists of:

- interior points X of P_Π,
- identified point pairs $\{X, X'\}$ where X and X' are points in the interior of two edges of the polygons of P_Π that bear the same label, and
- equivalence classes $\{X_1, X_2, \ldots, X_k\}$ where the X_i are vertices of the polygons in P_Π which are identified by the edge identifications.

The surjective non-injective map $\mathscr{I} : P_\Pi \to S_\Pi$ that sends interior points to itself, interior edge points to their identified point pairs and vertices to their vertex equivalence classes is called the *identification map*.

Note that, unlike P_Π, the set S_Π is not a subspace of $\mathbb{R}^2$, so although we have called S_Π the identification space of Π we do not know yet that it is a topological space. We will show that it is a metric space by defining a distance function on it. For this we first need the concept of a polygonal path.

Let $M_1, M_2, \ldots, M_n \in S_\Pi$ for some $n \geq 2$ and let $W_1, W_2, \ldots, W_n$ be points in P_Π where $W_i = M_i$ if M_i is an interior point and $W_i \in M_i$ if M_i is an edge pair or vertex equivalence

class. A *polygonal path* from M_1 to M_n is a set of line segments $w_1, w_2, \dots w_{n-1}$ where w_i connects W_i with W_{i+1}. Since the polygons are convex we know that these line segments are fully contained in P_Π. Note that for fixed M_1 and M_n there are many possible polygonal paths between them, for we can choose the $M_2, \dots, M_{n-1}$ freely and if M_i is an edge pair or a vertex equivalence class we can choose the point $W_i \in M_i$ freely. We define the length of a polygonal path to be the sum of the lengths of the line segments it consists of.

For two points $A, B \in S_\Pi$ we define the distance $d_{S_\Pi}(A, B)$ between them as the infinum of the lengths of all polygonal paths connecting them. It is clear that this function is a metric and thus that S_Π is a metric space.

It is known that polygons with the same number of vertices are homeomorphic. Furthermore it is possible to construct a homeomorphism that maps vertices to vertices and edges to edges, preserving the order and orientations of the edges. From this it follows that any two spaces S_Π and S'_Π corresponding to some schema Π are homeomorphic to each other. This means that for each schema Π there is a topologically unique identification space S_Π.

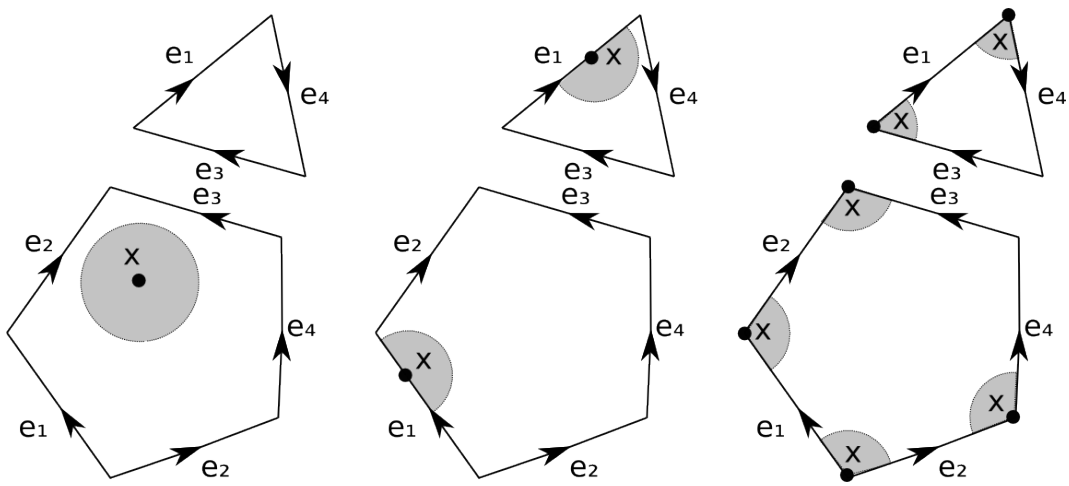


Figure 1.3: Open disks in S_Π

Theorem 1. *The identification space S_Π of a schema Π is a surface.*

Proof. We have already seen that we can define a distance function on the set S_Π such that it becomes a metric space, so it is, in particular, a Hausdorff space. We will now show that each point $X \in S_\Pi$ has an open neighbourhood homeomorphic to an open disk in the plane. If X is an interior point of a polygon then we can take a small open disk around it which is homeomorphic, by the inclusion map, to an open disk in the plane. If X it is a set $\{X_1, X_2\}$ of identified interior edge points then we can take the two half disks around the two points and paste them together to obtain an open planar disk. Note that this pasting of two half disks is just a renaming of the points in the half disks, the distance between any two points stays the same, so this operation can certainly be done by a homeomorphism. If X is a vertex equivalence class $\{X_1, \cdots, X_k\}$ then a small open neighbourhood consists of n slices. Let α_k be the angle of the sharp edge of the k-th slice. By picturing a slice in the upper half plane with its sharp vertex at the origin and one edge on the x-axes we see that the map

$$\phi_k(re^{i\theta}) = re^{i\theta\frac{2\pi}{n\alpha_k}}$$

is a homeomorphism that squeezes the slice so that its sharp angle is $\frac{2\pi}{n}$, see Figure 1.4. The like labeled edges of the n slices can now be pasted as to obtain an open planar disk. □

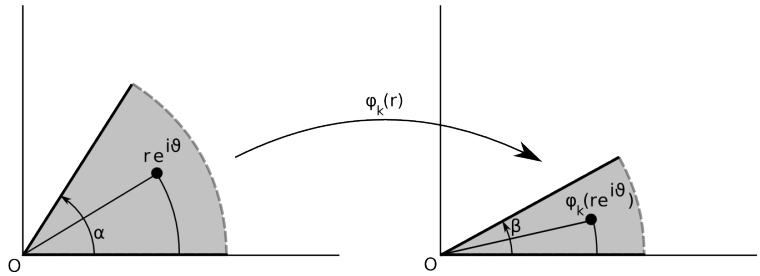


Figure 1.4: Resizing a slice

1.3 Normal forms

The *orientable normal form schemata* are the collections:

$$\begin{aligned} \Pi_0 &= \{(e_1, e_1^{-1})\} \\ \Pi_1 &= \{(e_1, e_2, e_1^{-1}, e_2^{-1})\} \\ &\vdots \\ \Pi_k &= \{(e_1, e_2, e_1^{-1}, e_2^{-1}, \ldots, e_{2k-1}, e_{2k}, e_{2k-1}^{-1}, e_{2k}^{-1})\} \\ &\vdots \end{aligned} \tag{1.4}$$

In the previous section we noted that the operation of pasting two edges, whose labels agree and whose orientations match, can be realized by a homeomorphism. The same holds for the reverse operation of cutting a polygon in two parts and identifying the cutting edges. The polygons P_Π are compact subspaces of $\mathbb{R}^2$, but seeing them as subspaces of $\mathbb{R}^3$ keeps the topology of P_Π, and S_Π after identification, intact. Finally, the operation of smoothly curving the polygons of P_Π can be done by a diffeomorphism, so in particular this keeps the topology of P_Π and thus S_Π intact. We will use these homeomorphisms to show that the identification spaces S_{Π_i} are homeomorphic to the surfaces depicted in Figure 1.4.

The surface S_{Π_0} is a disk with two identified edges. This disk can be folded, and the edges pasted, to obtain the sphere. Note that the orientation of the edges is correct, so the pasting can indeed be realized by a homeomorphism.

The schema Π_1 corresponds to the torus. This can be seen by realizing S_{Π_1} as a square where the opposite edges are identified. This square can be rolled up and pasted to obtain a cylinder. Then we can roll this cylinder up and paste the two remaining edges to obtain the torus, see Figure 1.5.

The next normal form Π_2 corresponds to the double torus. To see this we take S_{Π_2} to be a regular polygon and cut it in half in such a way that by giving the two new edges both the same label c and an opposite orientation we get the schema $\{(e_1, e_2, e_1^{-1}, e_2^{-1}, c), (e_3, e_4, e_3^{-1}, e_4^{-1}, c^{-1})\}$. These two polygons are called handles and we can fold and paste them as shown in Figure 1.6. Finally, we can paste the two remaining edges having label c to obtain the double torus.

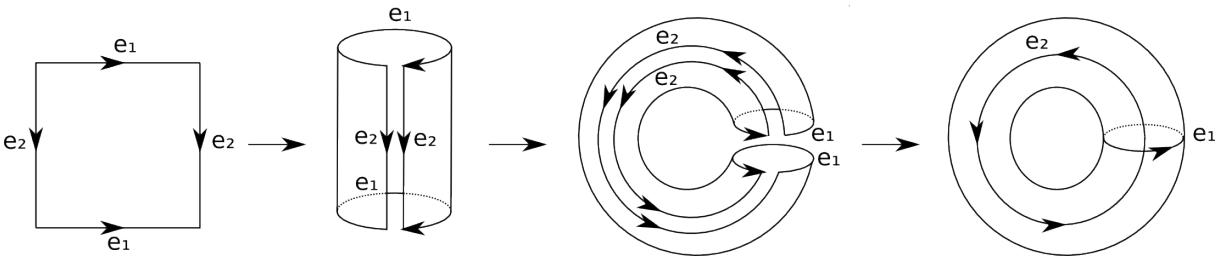


Figure 1.5: Folding and pasting of the torus

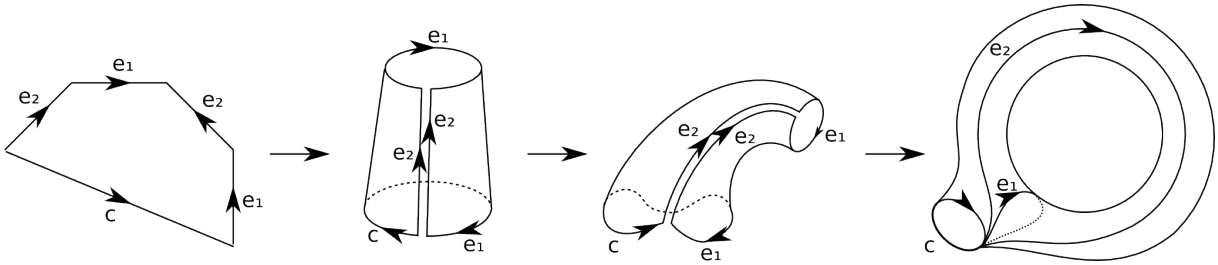


Figure 1.6: Folding and pasting of a handle

The rest of the normal form schemata, Π_k for $k > 2$, correspond to the sphere with k handles. We start with the schema

$$\{(e_1, e_2, e_1^{-1}, e_2^{-1}, \ldots, e_{2k-1}, e_{2k}, e_{2k-1}^{-1}, e_{2k}^{-1})\}.$$

Now we cut of k handles (as shown in Figure 1.7 for the case where $i = 3$) to obtain the schema

$$\{(e_1, e_2, e_1^{-1}, e_2^{-1}, c_1), \ldots, (e_{2k-1}, e_{2k}, e_{2k-1}^{-1}, e_{2k}^{-1}, c_k), (c_1^{-1}, c_2^{-1}, \ldots, c_k^{-1})\}.$$

Each handle can be folded like we did for the double torus. The edge identifications of the original schema imply that all vertices of $(c_1^{-1}, c_2^{-1}, \ldots, c_k^{-1})$ are identified, so we can take all vertices together and fold this k-gon like a table sheet. Now we can paste each handle to one opening of this folded table sheet to obtain the sphere with k handles. Note that although the figure shows this only for Π_3, every step can be done for the general Π_k where $k \geq 3$.

The sphere with g handles. In chapter 1.1 we gave a precise definition of the sphere (by an implicit equation) and the torus (by a parametrization). In general we define the sphere with g handles to be the surface S_{Π_g}. As we have just seen this general definition agrees with the definitions already given for the cases $g = 0$ and $g = 1$. In chapter 2.1 we see that this new definition is very natural when we discuss distances on surfaces. To prove that a surface is completely determined by its genus we have to prove that any surface is homeomorphic to S_{Π_g} for some g, and that S_{Π_i} and S_{Π_j} are not homeomorphic for $i \neq j$. To prove the first part we need a triangulation of the surface which we will discuss in the next section.

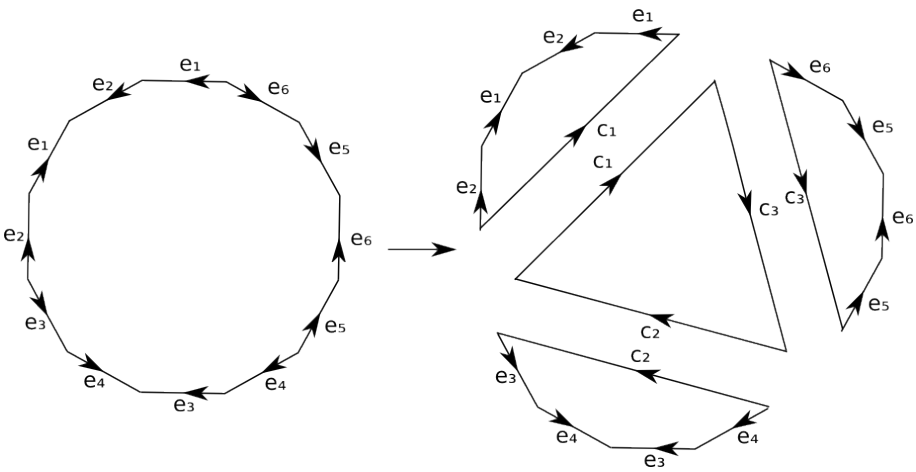


Figure 1.7: Cutting the polygon for a sphere with three handles

1.4 Triangulations

A *triangulation* of a surface S is a subdivision of the surface as a finite union of sets τ_i called faces. Each face is homeomorphic to a closed triangle in the plane. The intersection of two faces is either empty, a single point called a vertex, or is a set homeomorphic to the interval $[0, 1]$ called an edge. The intersection of two edges is either empty or a vertex. Each face has exactly three edges. When two faces share an edge they are called adjacent.

The criterion that each face is homeomorphic to a closed triangle in the plane is not significant for a triangulation. We could as well have said that it must be homeomorphic to a closed disk, since these are topologically the same. What is significant, however, is the criterion that the intersection of two faces is either empty, a vertex or an edge and that each face has exactly three edges.

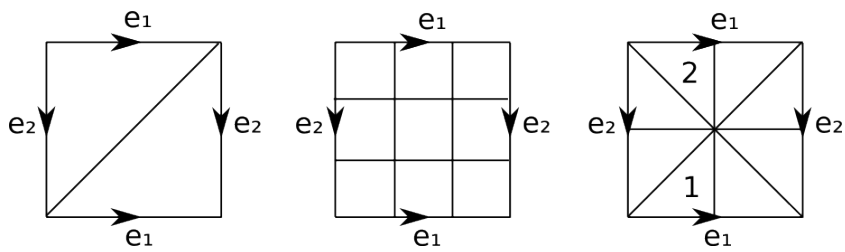


Figure 1.8: Subdivisions of the square

In Figure 1.8 we see three subdivisions of the square. When we identify the sides of these squares to obtain the torus then these subdivisions are no triangulations. The first one is no triangulation because, after identification, the two faces are not homeomorphic to closed disks. The second one is no triangulation because the faces have four edges, and the third one because the intersection of, for instance, face 1 and 2 is a vertex *and* an edge.

In Figure 1.9 we see two subdivisions of the square that *are* triangulations of the torus. This can be checked directly by using the definition of a triangulation. In [4] it is shown that the minimal number of vertices required to triangulate the torus is 7. The second triangulation

in Figure 1.9 is such a minimal triangulation.

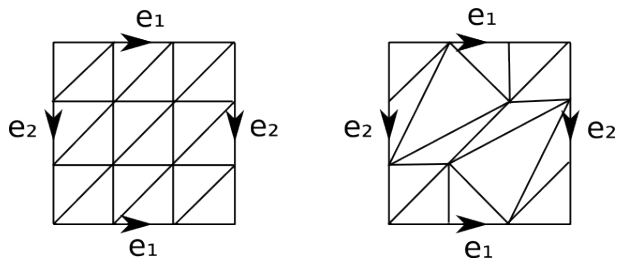


Figure 1.9: Triangulations of the torus

Triangulations of closed orientable surfaces. Having constructed explicit triangulations for the torus we might wonder whether every closed orientable surface S can be triangulated. This is indeed the case and we will show how we can use the fact that S is a compact surface to prove this. This is only a sketch of the proof, the full proof is quite delicate and can be found in [2].

Since S is a surface each point $x \in S$ has an open neighbourhood homeomorphic to an open disk in the plane. The union of these neighbourhoods cover S, but since S is compact there is a finite subset of these neighbourhoods that also cover S. In each of the remaining neighbourhoods we can take a subset that is homeomorphic to a closed disk in such a way that the union of these closed subsets still cover S. It can be shown that these closed subsets can be chosen in such a way that their boundaries intersect each other only finitely often (this is the hard part). When we remove the unnecessary sets from this collection of closed sets we have a covering of S that looks like Figure 1.10. We can now add vertices and edges as shown in figure 1.10 to obtain a triangulation.

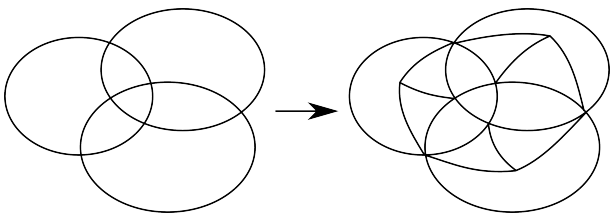


Figure 1.10: Triangulation of a compact surface

1.5 Surfaces as normal forms

In chapter 1.4 we saw that each closed orientable surface S can be triangulated. We will use this triangulation to construct a homeomorphism from S to a polygon in the plane with pairwise identified edges. After that we will see that this polygon is homeomorphic to S_{Π_i} where Π_i is the i-th orientable normal form schema.

Unfolding of a surface. We start by taking a face τ_1 of the triangulation. By the definition of a triangulation there is a homeomorphism ϕ_1 from this face to a planar triangle T_1. We

can choose this homeomorphism in such a way that the three edges of the face τ_1 are mapped to the edges of T_1. Since the surface is orientable we have a concept of clockwise orientation, so we are able to give the edges of T_1 an anticlockwise orientation. Now we take one of the adjacent faces of τ_1 which we denote by τ_2. For this face there is a homeomorphism ϕ_2 to a triangle T_2 in the plane. We can again choose this homeomorphism such that vertices are mapped to vertices, but we also choose it such that ϕ_1 and ϕ_2 agree on the common edge and such that $T_1 \cup T_2$ is convex, which is always possible.

We can now choose a third face τ_3 that is adjacent to τ_1 or τ_2 for which we do the same. If τ_3 is adjacent to both τ_1 and τ_2 then we will not try to make the triangle T_3 large enough to be adjacent to both T_1 and T_2 but instead we will identify the two edges that should otherwise be a common edge. Since the surface is connected we can continue this process until there are no faces left. Note that due to the consistent concept of orientation we will, when we traverse the edge path around the final convex polygon, traverse the corresponding edges in opposite directions. Since the maps ϕ_i agree where their domains overlap we can define a map ϕ by $\phi(x) = \phi_i(x)$ where i is an index such that $x \in \tau_i$. Now we have that ϕ is a homeomorphism from $\cup_i \tau_i = S$ to a planar polygon with pairwise identified edges. [3, page 233]

Reduction to a normal form polygon. In [10, page 127] it is shown that the polygon that results from unfolding a closed orientable surface can be reduced, by cutting and pasting, to a polygon with a single vertex cycle. After that it is shown in [10, page 136] that this polygon can be reduced, again by cutting and pasting, to an orientable normal form polygon. The cutting surface S_Π corresponding to this new normal form polygon is, since we only used cutting and pasting, homeomorphic to the original surface. This shows that each closed orientable surface S is homeomorphic to S_{Π_i} for some i.

Chapter 2

Geometry

In this chapter we start by defining the spherical, Euclidean, and hyperbolic 2-spaces, and we will define geometric surfaces. Then we will show that any closed orientable surface can be realized as a geometric surface. We define the geometric concept of completeness and prove that closed orientable surfaces are complete.

After that we use the fact that a closed orientable surface S is a complete geometric surface to construct a locally isometric map, called the pencil map, from the spherical, Euclidean, or hyperbolic 2-space onto S. We also define a covering map and covering space and claim that the pencil map is a covering map. We use the pencil map to introduce the concept of a covering isometry and continue by proving a couple of important properties of covering isometries. We use the covering isometries to prove that the pencil map is a covering map.

2.1 Geometric surfaces

The *Euclidean* 2*-space*, denoted $\mathbb{E}^2$, is the metric space $(\mathbb{R}^2, d_{\mathbb{E}^2})$ where $d_{\mathbb{E}^2}$ is the usual Euclidean distance function:

$$d_{\mathbb{E}^2} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R},\ (x, y) \mapsto ((x_1 - y_1)^2 + (x_2 - y_2)^2)^{\frac{1}{2}}.$$

An open Euclidean disk is a set $D_\epsilon(x) = \{y \in \mathbb{E}^2 \mid d_{\mathbb{E}^2}(x, y) < \epsilon\}$ for some $x \in \mathbb{E}^2$ and some $\epsilon > 0$.

An *isometry* is a map from one metric space onto another that preserves distance. That is, if (M_1, d_1) and (M_2, d_2) are metric spaces and $f : M_1 \to M_2$ is an isometry, then $d_1(x, y) = d_2(f(x), f(y))$ for all $x, y \in M_1$. Isometries are bijective maps: they are injective since if $f(x) = f(y)$ then $d(x, y) = d(f(x), f(y)) = 0$ so $x = y$, and they are surjective by definition. An isometry and its inverse clearly are continuous, so an isometry is a homeomorphism. Two metric spaces are called *isometric* when there exists an isometry between them. From the above it follows that two isometric spaces are, in particular, homeomorphic.

A *Euclidean surface* is a metric space (S, d) such that for any $x \in S$ there is an $\epsilon > 0$ such that $D_\epsilon(x) = \{y \in S \mid d(x, y) < \epsilon\}$ is isometric to a Euclidean disc. *Hyperbolic space*, denoted $\mathbb{H}^2$, is the set $\{(x, y) \in \mathbb{R}^2 \mid y > 0\}$ with the hyperbolic metric defined on it, see [8] for a general introduction to hyperbolic geometry. *Spherical space*, denoted $\mathbb{S}^2$ is the unit sphere with the spherical distance function. By replacing Euclidean space by spherical space or hyperbolic space in the definition of Euclidean surface we get the definitions of spherical

and hyperbolic surfaces.

When it is possible to define a distance function on a surface such that it is a Euclidean, spherical or hyperbolic surface then we say that we can realize the surface *geometrically*, or we just say that the surface is a *geometric surface*. A geometric surface is, in particular, a surface. It is a metric space so it is a Hausdorff space, and any point has an open neighbourhood isometric to an open disk in $\mathbb{E}^2$, $\mathbb{S}^2$, or $\mathbb{H}^2$ which, in turn, is homeomorphic to an open disk in $\mathbb{R}^2$. In the next section we will prove that, conversely, any closed orientable surface can be realized geometrically.

2.2 Geometric realization

In chapter 1.2 we assigned a distance function to the set S_{Π_g} in order to prove that it is a surface. Then we proved that each closed orientable surface is homeomorphic to S_{Π_g} for some g. We will now assign specific distance functions to each of the spaces S_{Π_g} and show that in this way we can realize the closed orientable surfaces geometrically.

We have seen that the space S_{Π_0} is homeomorphic to $\mathbb{S}^2$. Let $\phi : S_{\Pi_0} \to \mathbb{S}^2$ be a homeomorphism between the two spaces and let $d_{\mathbb{S}^2}$ be the spherical distance function. Let $d(x, y) = d_{\mathbb{S}^2}(\phi(x), \phi(y)$ for all $x, y \in S_{\Pi_0}$, d is a distance function on S_{Π_0} and with this distance function S_{Π_0} becomes a spherical surface.

The torus S_{Π_1} is a Euclidean surface. To see this we just have to repeat the proof of Theorem 1 and note that the homeomorphisms used in the proof are in fact isometries. The only tricky part here is the last step where we resize the slices in order to have a total angle sum of 2π, but since the polygon is a square the angle sum of the four angles already is 2π, so we do not need to resize the slices and as such the last step can also be done by an isometry.

The rest of the spaces, S_{Π_g} for $g \geq 2$, are hyperbolic surfaces. We will again use a variation of Theorem 1 but now we take P_{Π_g} to be a regular polygon in the hyperbolic plane with angle sum 2π. We define polygonal paths in the same way except that we take hyperbolic geodesic segments as the straight line segments. We define the distance between two points to be the infinum of all hyperbolic polygonal paths between them. We have already noted that k-gons are homeomorphic, the same holds if the edges are not (Euclidean) straight lines so this identification space S_{Π_g} is, as it should be, homeomorphic to the one given in chapter 1.2.

Small open neighbourhoods of interior points are isometric, by the inclusion map, to open hyperbolic disks and small open neighbourhoods of edge pairs are isometric to open hyperbolic disks by the hyperbolic isometry that pastes the two half disks. Now we also want a neighbourhood of a vertex cycle to be isomorphic to an open hyperbolic disk. For this to be the case we need the angle sum of the polygon to be 2π. Since if this is the case we do not need to rezise the slices and we can use a hyperbolic isometry to paste the slices to obtain a hyperbolic disk [10, page 124]. We will now show that it is always possible to constuct any regular hyperbolic polygon in such a way that the angle sum is 2π.

Construction of regular polygons with angle sum 2π. In the Euclidean plane the only regular polygon with angle sum 2π is the square, we will see that in the hyperbolic 2-space we can construct any regular polygon in such a way that it has angle sum 2π. This explains why the torus is Euclidean and why surfaces of higher genus are hyperbolic.

The area of a hyperbolic triangle is $\pi - \alpha - \beta - \gamma$ where α, β, and γ are the angles of the triangle. We can divide a regular hyperbolic n-gon with angle sum s into n triangles with

angles $\frac{2\pi}{n}$, $\frac{s}{2n}$, and $\frac{s}{2n}$. This means that the area A of a regular hyperbolic n-gon with angle sum s is $n\pi - 2\pi - s$. In order to get angle sum 2π we need the area to be $(n-4)\pi$, which is positive since $n \geq 8$. By taking the center of the polygon to be 0 in the Poincaré disk we see that the diameter of the polygon can be anything between 0 and ∞. It is also clear that the area of the polygon varies with the diameter d between 0 and $n\pi - 2\pi$, since $s \to 0$ when $d \to \infty$. So by the intermediate value theorem there is a diameter such that the area is $(n-4)\pi$ which gives us angle sum 2π [10]s.

Distance on surfaces. This explains why it is nice to define the sphere with k handles to be S_{Π_k}. The distance function follows naturally from the geometric space in which we take the polygon. When we take some closed orientable surface S that is a subset of $\mathbb{R}^3$, for instance as the image of some map $p : [0,1]^2 \to \mathbb{R}^3$, then it is homeomorphic to some S_{Π_k}. The distance function on S_{Π_k} induces some distance function on S, but this distance function will be quite strange. That is, it will be very different from the distance function induced from embedding the surface in $\mathbb{E}^3$.

As an example we can take a look at the torus, as defined in (1.2), with the Euclidean distance function. In Figure 1.5 we see how we obtain the torus in $\mathbb{R}^2$ from folding a square in $\mathbb{E}^2$. Now, the paths e_1 and e_2 on the torus do have the same length, while it looks as if e_2 is longer.

2.3 Completeness

Euclid's second postulate asserts that straight line segments can be continued indefinitely. This means that a straight line segment in the spherical, Euclidean or hyperbolic 2-space, can be extended in either direction in such a way that the length of the half line diverges to infinity. There are geometric surfaces for which this is not true. Take for instance the Euclidean plane without the origin. This is a Euclidean surface, but when we extend the line segment between the points $(1,0)$ and $(2,0)$ in the direction of the origin then its length converges to 2.

A surface for which this postulate *does* hold is called a *complete surface*. We will now prove that all closed orientable surfaces S are complete.

Theorem 2. *A closed orientable surface S is complete.*

Proof. When a surface S is spherical then it is isometric to $\mathbb{S}^2$. A straight line segment l on $\mathbb{S}^2$ lies on a great circle, by following this circle around and around this line segment can clearly be extended in either direction in such a way that its length diverges to infinity.

We have seen that any other closed orientable surface S is homeomorphic to S_{Π_g} for some $g > 0$. A line segment l on S_{Π_g} is the image, under the identification map $\mathscr{I}$, of a set of line segments $L_1, L_2, \ldots, L_k$ on P_{Π_g}. See Figure 2.1 for an example of three of these segments on the polygon for the double torus. When we extend l we will need more and more L_i's in the pre-image. All but maybe the first and last segments connect two edge points of the polygon.

The length of l is the sum of the lengths of the segments V_i. Now assume that the length of l does not diverge to infinity. Since the lengths of the segments V_i are non-negative this means that $\lim_{k\to\infty} V_k = 0$. The only way for this to happen is when these segments come closer and closer to some vertex V of P_{Π_g}, that is, l converges to $\mathscr{I}(V)$.

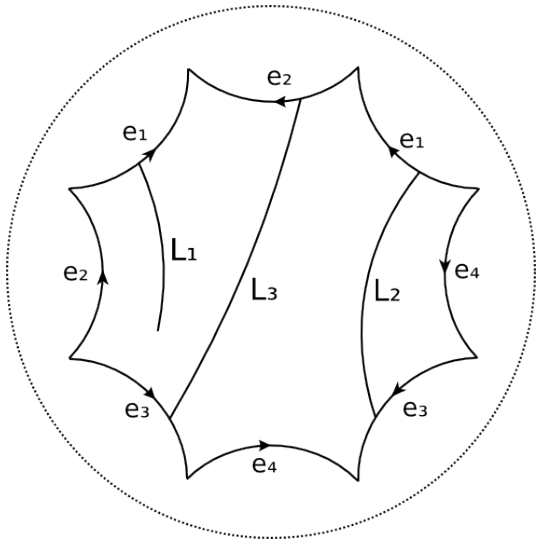


Figure 2.1: Line segments on P_{Π_2}

From chapter 2.1 we know, however, that S_{Π_g} is a geometric surface so $\mathscr{I}(V)$ has an open neighbourhood isometric to an open disk in the Euclidean or hyperbolic plane. We can now use Euclid's second postulate to reach a contradiction. □

2.4 The pencil map

A *covering* for a topological space S is a topological space $\tilde{S}$ together with a map $\rho : \tilde{S} \to S$, such that for all points $x \in S$ there is an open neighbourhood N such that $\rho^{-1}(N)$ is a disjoint union of open sets, each of which is mapped homeomorphically by ρ onto N. Such a space $\tilde{S}$ and map ρ are called a *covering space* and a *covering map* for the space S, respectively.

From the definition it follows that a covering map is surjective, another important property is that is is continuous.

Theorem 3. *A covering map $\rho : \tilde{S} \to S$ is continuous.*

Proof. To prove continuity we have to prove that the inverse image of any open set in S is open. Let $O \subset S$ be some open set. For each $x \in O$ there is an open neighbourhood N_x such that $\rho^{-1}(N_x)$ is a disjoint union open sets, each of which is mapped homeomorphically by ρ onto N_x. From this it follows that the ρ inverse image of the open set $N_x \cap O \subset N_x$ is also open. Now we have that

$$\rho^{-1}(O) = \rho^{-1}(\bigcup_{x \in O}(N_x \cap O)) = \bigcup_{x \in O}(\rho^{-1}(N_x \cap O))$$

and since an arbitrary union of open sets is open we have that $\rho^{-1}(O)$ is open. □

We will show that for any closed orientable surface S there is a covering. For this we are going to construct a map ρ from $\mathbb{S}^2$, $\mathbb{E}^2$ or $\mathbb{H}^2$ onto S.

If the surface S is spherical then there is a homeomorphism ϕ from $\mathbb{S}^2$ to S, this ϕ is a covering map. Now assume that the surface S is Euclidean or hyperbolic and let $\tilde{S}$ be $\mathbb{E}^2$ or

$\mathbb{H}^2$, respectively. In the following proof we will use the previously obtained results that the closed orientable surfaces are geometric and complete surfaces.

Let $O \in S$, since S is a geometric surface there is a $\tilde{O} \in \tilde{S}$ and an isometry $\rho : D_\epsilon(\tilde{O}) \to D_\epsilon(O)$. We now look at the straight half lines that originate from $\tilde{O}$. From the first postulate of Euclid it follows that these lines fill the entire space $\tilde{S}$. Since ρ is an isometry we have that the ρ-images of the parts of the straight lines that lie in the open disk $D_\epsilon(\tilde{O})$ are parts of straight lines that lie in $D_\epsilon(O)$. That is, for each straight half line in $\tilde{S}$ that originates from $\tilde{O}$ there is a unique corresponding straight half line in S that originates from O.

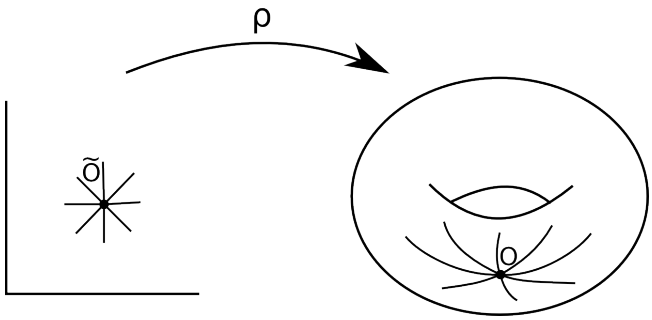


Figure 2.2: The pencil map

We now extend the map ρ so that it maps all of $\tilde{S}$ to S. Let $\tilde{X} \in \tilde{S}$, this point lies on a unique straight line originating from $\tilde{O}$ and has a distance $d_{\tilde{S}}(\tilde{O}, \tilde{X})$ to $\tilde{O}$. We now let $X \in S$ be the point on the corresponding straight half line in S such that $d_S(O, X) = d_{\tilde{S}}(\tilde{O}, \tilde{X})$. This is always possible since we know that S is complete so we can extend each straight line indefinitely.

We now define X to be the ρ-image of $\tilde{X}$. This makes ρ a well defined map from $\tilde{S}$ onto S. Additionally, it is shown in [10, page 36] that this map is onto S and that it is a local isometry.

So we have that when a closed orientable surface S is spherical then we already have a covering map, and when the surface is Euclidean or hyperbolic then we have constructed a locally isometric map ρ from $\mathbb{E}^2$ or $\mathbb{H}^2$ onto S. In the next section we will develop the tools we need to prove that a locally isometric surjective map is a covering, which proves that any closed isometric surface has a covering.

2.5 Covering isometries

In this section we will introduce the covering isometry group, which will give more insight in the geometry of the covering surface. We will give a couple of results about this group which will come to use in later chapters. Additionally we will prove, as claimed in chapter 2.4, that the pencil map ρ is a covering map.

The *isometry group* of a metric space M is the set of isometries from M onto itself with function composition as group operator. In chapter 2.1 we saw that an isometry is bijective and that its inverse is also an isometry. Furthermore, we have that the composition of two isometries $f : M \to M$ and $g : M \to M$ is also an isometry, since $d(g(f(x)), g(f(y))) = d(f(x), f(y)) = d(x, y)$. Using this it is easy to see that the isometry group of a metric space M is a well defined group.

Let G be a subgroup of the isometry group of a metric space M. For an element $X \in M$ we define the G-orbit of X to be the set $G(X) = \{g(X) \mid g \in G\}$, note the ambiguity of the symbol G. A subgroup G is *discontinuous* if none of the G-orbits contain limit points. The group G is *fixed point free* if for any $X \in M$ we have: $g(X) = X$ for some $g \in G$ implies g is the identity. The next theorem shows that if G is fixed point free then for any two points A and B there is at most one element $g \in G$ such that $B = g(A)$.

Theorem 4. *Let G be a fixed point free subgroup of the isometry group of some surface S, for any two $X, Y \in S$ there is at most one isometry $g \in G$ such that $Y = g(X)$.*

Proof. Let $g_1, g_2 \in G$ and assume that $g_1(X) = Y = g_2(X)$. This means that $X = g_2^{-1}(g_1(X))$ but since G is fixed point free this means that $g_2^{-1}g_1$ is the identity element so we have that $g_1 = g_2$. □

The covering isometry group. Given a closed orientable surface S and the pencil map $\rho : \tilde{S} \to S$ we define a *covering isometry* γ to be an isometry of $\tilde{S}$ onto itself such that for any $\tilde{X} \in \tilde{S}$ we have that $\rho(\gamma(\tilde{X})) = \rho(\tilde{X})$. The set of covering isometries of $\tilde{S}$ clearly is a subgroup of the isometry group of $\tilde{S}$. We denote this subgroup by Γ. In [10] it is shown that this subgroup is fixed point free and the following theorem shows that it also is discontinuous.

Theorem 5. *The covering isometry group Γ is discontinuous.*

Proof. If Γ is not discontinuous then there is a $X \in S$ such that $\Gamma(X)$ has a limit point L. Since ρ is a local isometry there is an $\epsilon > 0$ such that ρ is an isometry on $D_\epsilon(L)$. Since $D_\epsilon(L)$ is isometric to an open subset of S each point in $D_\epsilon(L)$ must be mapped to a unique point on S. Since L is a limit point of $\Gamma(X)$ there are infinitely many points of $\Gamma(X)$ in the disk $D_\epsilon(L)$. By the definition of an orbit we have that there is a covering isometry that maps one of these points to another, and by the definition of a covering isometry these two points are mapped to the same point on S. This is a contradiction, so Γ is discontinuous. □

For two points $\tilde{X}, \tilde{Y} \in \tilde{S}$ there clearly are isometries σ in the isometry group of $\tilde{S}$ such that $\tilde{Y} = \sigma(\tilde{X})$. In [10] it is shown that that if $\tilde{X}$ and $\tilde{Y}$ are mapped to the same point on S, then at least one of those isometries is a covering isometry. We use this result to prove that inverse ρ images of points on the surface are orbits of the covering isometry group.

Lemma 6. *The inverse image $\rho^{-1}(\{X\})$ for any $X \in S$ is non-empty and equals $\Gamma(\tilde{X})$ for any $\tilde{X} \in \rho^{-1}(\{X\})$.*

Proof. The pencil map ρ is surjective so its inverse images are non-empty. Let $\tilde{X}$ be some point in $\rho^{-1}(\{X\})$ for some $X \in S$, we will prove that $\rho^{-1}(\{X\}) = \Gamma(\tilde{X})$.

($\subset$) If $\tilde{Y} \in \rho^{-1}(\{X\})$ then $\rho(\tilde{X}) = \rho(\tilde{Y})$ so there is some $\gamma \in \Gamma$ such that $\tilde{Y} = \gamma(\tilde{X})$ which means that $\tilde{Y} \in \Gamma(\tilde{X})$.

($\supset$) If on the other hand $\tilde{Y} \in \Gamma(\tilde{X})$ then there is a $\gamma \in \Gamma$ such that $\tilde{Y} = \gamma(\tilde{X})$, but γ is a covering isometry so by definition we have that $\rho(\tilde{Y}) = \rho(\tilde{X})$ which means that $\tilde{Y} \in \rho^{-1}(\{X\})$. □

We can use this result together with the fact that Γ is discontinuous to prove the following theorem:

Theorem 7. *The pencil map ρ is a covering map.*

Proof. Let $X \in S$, by Lemma 6 the inverse image $\rho^{-1}(\{X\})$ is non-empty and equals $\Gamma(\tilde{X})$ for some $\tilde{X} \in \rho^{-1}(\{X\})$.

Since ρ is a local isometry there is some $\epsilon > 0$ such that ρ is an isometry between $D_\epsilon(\tilde{X})$ and $D_\epsilon(X)$. Since Γ is discontinuous we know that $\Gamma(\tilde{X}) = \rho^{-1}(\{X\})$ does not contain a limit point. This means that we can choose ϵ small enough such that the disks $D_\epsilon(\tilde{Y})$ for all $\tilde{Y} \in \rho^{-1}(\{X\})$ are disjoint.

The open disk $D_\epsilon(\tilde{X})$ gets mapped isometrically onto $D_\epsilon(X)$. But since for any $\tilde{Y} \in \rho^{-1}(\{X\})$ there is a covering isometry $\gamma \in \Gamma$ such that $\tilde{Y} = \gamma(\tilde{X})$ we have that $D_\epsilon(\tilde{Y})$ is also mapped isometrically onto $D_\epsilon(X)$.

So we have that $\rho^{-1}(D_\epsilon(X))$ is a disjoint union of open sets, each of which is mapped isometrically and thus homeomorphically by ρ onto N, so ρ is a covering map. □

Geometric aspects of the covering isometry group. We will now discuss in some more detail what the covering isometry group of the closed orientable surface S looks like. If the genus of S is 0 then the universal covering surface is $\mathbb{S}^2$, and the covering map is a global isometry. This means that the covering isometry group consists just of the identity map.

If S has genus $g = 1$ then the universal covering surface is $\mathbb{E}^2$, the isometry group of $\mathbb{E}^2$ consists of rotations, translations, reflections, and glide reflections. We have seen that the covering isometry group is fixed point free, so it does not contain rotations or reflections, and since S is orientable it does not contain glide reflections either. So the covering isometry groups consist just of translations.

If S has genus $g > 1$ then the universal covering surface is $\mathbb{H}^2$. The hyperbolic translations and hyperbolic glide reflections are the only fixed point free isometries on $\mathbb{H}^2$. Since S is orientable its covering isometry group consists of just the hyperbolic translations. We will now explain in some more detail what hyperbolic translations are.

In the Poincaré disk a hyperbolic translation τ is the composition of two circle inversions in *disjoint* circles that both are orthogonal to the boundary of the disk. These circle arcs are hyperbolic geodesics and since they do not intersect they are called ultra-parallel. The TODO unique geodesic which is orthogonal to both circles is invariant under τ. The equidistant curves to this invariant geodesics, which are not geodesics, are clearly also invariant under τ.

Chapter 3

Curves on surfaces

3.1 Homotopy and curve lifting

In this section we will finally introduce the concept of curves, which will, when they are closed, represent the rubber bands which we talked about in the introduction. We will define the concept of homotopy to represent the operation of moving these bands. The main result of this section is the characterisation of homotopic curves using the covering surface.

A *curve* on a surface S is defined to be a continuous map $p : [0,1] \to S$. The *endpoints* $p(0)$ and $p(1)$ of a curve are respectively called the *origin* and the *terminus*. When we can 'continuously deform' two curves with the same endpoints into each other we call them homotopic. Formally, two curves are *homotopic* if there exists a continuous map $h : [0,1]^2 \to S$ such that:

- $p_1(t) = h(0,t)$ and $p_2(t) = h(1,t)$ for all $t \in [0,1]$, and
- $p_1(0) = h(s,0) = p_2(0)$ and $p_1(1) = h(s,1) = p_2(1)$ for all $s \in [0,1]$.

Universal coverings. A space is called *simply connected* if it is connected and if any two curves whose endpoints coincide are homotopic. When a covering surface is simply connected it is called a *universal covering*. It is clear that any curve p in $\mathbb{R}^2$ can homotopically be 'contracted' to the straight line segment connecting the endpoints of p. From this it follows immediately that $\mathbb{R}^2$ is simply connected. By viewing $\mathbb{H}^2$ as a part of $\mathbb{R}^2$ we also have that $\mathbb{H}^2$ is simply connected. By using stereographic projection we have that any two non space filling curves on $\mathbb{S}^2$ are homotopic. By cutting a space filling curve p on $\mathbb{S}^2$ into smaller parts it can be shown, using the compactness of p, that p is homotopic to a non space filling curve and thus that $\mathbb{S}^2$ is simply connected [10, page 142]. This means that the covering surfaces used in the construction of the pencil map ρ are universal covering surfaces.

Curve lifting. When we have a curve p on a surface S and a curve $\tilde{p}$ on its universal covering surface $\tilde{S}$ such that $\rho \circ \tilde{p} = p$ then $\tilde{p}$ is called a *lift* of p. In [3, page 156] the compactness of curves is used to show that for any curve p on S and for any $\tilde{O} \in \rho^{-1}(\{p(0)\})$ there is a unique lift $\tilde{p}$ for which $\tilde{p}(0) = \tilde{O}$.

The following important theorem will be used extensively to determine whether two curves on a surface are homotopic.

Theorem 8. *Let p_1 and p_2 be curves on a closed orientable surface S whose endpoints coincide and let $\tilde{p}_1$ and $\tilde{p}_2$ be lifts of p_1 and p_2, respectively, whose origins coincide. The curves p_1 and p_1 are homotopic if and only if the termini of $\tilde{p}_1$ and $\tilde{p}_2$ coincide.*

Proof. ($\Rightarrow$) If p_1 and p_2 are homotopic then there is a homotopy h between them. In [3, page 156] it is shown that just as for a curve there is also a uniqe lift for a homotopy. This lift $\tilde{h}$ then is a homotopy between $\tilde{p}_1$ and $\tilde{p}_2$ which is, by definition, only possible if their endpoints coincide.

($\Leftarrow$) If on the other the hand the termini of $\tilde{p}_1$ and $\tilde{p}_2$ coincide then $\tilde{p}_1$ and $\tilde{p}_2$ share the same endpoints. By the definition of a lift, the covering surface is a universal covering surface, which means that it is simply connected. This means that there is a homotopy $\tilde{h}$ between $\tilde{p}_1$ and $\tilde{p}_2$. The map $h = \rho \circ \tilde{h}$ is continuous since it is the composition of two continuous maps and it is easy to check that h is a homotopy between p_1 and p_2. □

3.2 The fundamental group

In this section we will close the curves which we have defined in the previous section and we will group these closed curves according to their homotopy type. We will see that we can associate an element of the covering isometry group to each of these homotopy classes.

Definitions. When the origin and terminus of a curve coincide then it is *closed*, the origin (or terminus) of a closed curve is called the *base point*. The product of two closed curves c_1 and c_2 whose base points coincide is defined as

$$(c_1c_2)(t) = \begin{cases} c_1(2t), & 0 \leq t \leq \frac{1}{2} \\ c_2(2t-1), & \frac{1}{2} \leq t \leq 1 \end{cases}$$

and the inverse of a closed curve c is defined as

$$c^{-1}(t) = c(1-t)$$

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

The homotopy relation is an equivalence relation on the set of closed curves with fixed base point $B \in S$. The equivalence class of a closed curve c is denoted by $[c]$ and the collection of equivalence classes by $\pi_1(S)$. The product of two classes is defined as $[c_1][c_2] = [c_1c_2]$, the inverse as $[c_1]^{-1} = [c_1^{-1}]$, and the identity element as $[1]$, where 1 is the constant curve at the base point. It is straightforward to show directly that this product is well defined and that the set $\pi_1(S)$ is a group under this product [3, page 165]. For a connected surface it does not matter which base point we choose so in particular we have that the fundamental group $\pi_1(S)$ of a closed orientable surface S is unique.

Isomorphism. In chapter 3.1 we saw that two curves c_1 and c_2 are homotopic if and only if the termini of their lifts, each having the same origin $\tilde{O}$, coincide. This means that there is a bijection between the collection $\pi_1(S)$ and the set

$$\{\tilde{c}(1) \mid \tilde{c} \text{ is a lift with origin } \tilde{O} \text{ of a closed curve } c \text{ with base point } O\} \tag{3.1}$$

for some fixed $\tilde{O}$ that lies over O.

Since the curve c is closed the endpoints of a lift are both mapped to the same point by the covering map. This means that for any $\tilde{X}$ in the set of (3.1) there is a covering isometry $\gamma \in \Gamma$ such that $\tilde{X} = \gamma(\tilde{O})$. On the other hand, for any covering isometry γ, there is a closed curve c with base point O such that its lift has endpoints $\tilde{O}$ and $\gamma(\tilde{O})$. This means that the set of (3.1) equals $\Gamma(\tilde{O}) = \{\gamma(\tilde{O}) \mid \gamma \in \Gamma\}$.

From Theorem 4 it follows that there is a bijection between $\Gamma(\tilde{O})$ and Γ. So we have that there is a bijection between $\pi_1(S)$ and Γ where the homotopy class of a closed curve c is mapped to the covering isometry that maps the origin of a lift of c to its terminus. In [10] it is shown that this bijection is a group isomorphism, which proves the following theorem:

Theorem 9. *Given a closed orientable surface S, the fundamental group $\pi_1(S)$ is isomorphic to the covering isometry group Γ of S, by the isomorphism that maps the homotopy class of a curve c to the covering isometry that maps the origin of a lift of c to its terminus.*

3.3 The Cayley graph

The goal of this section is to describe the group structure of the fundamental group of a closed orientable surface. In order to do this effectively we will start by introducing the concepts of generators, words, relators and group presentations. We have seen that the fundamental group is isomorphic to the covering isometry group of the surface, but although the covering isometry group tells us a lot about the covering surface it does not, in a direct way, help in making the group structure more clear. The Cayley graph will turn out to be a bridge between the geometric aspects of the covering isometry group and the, more abstract, group aspects of the fundamental group.

Generators, words, relators and presentations. A *generating set* for a group G is a subset H such that any $g \in G$ can be expressed as a product of finitely many elements in H and their inverses. A *word* is a finite ordered non-empty set of elements of some generating set. Note that different words can represent the same element of a group. For instance, if a and b are two generators then aa^{-1} and bb^{-1} are different words but both represent the identity element. When two words represent the same group element we call them *equivalent*.

Let F be some subset of G, that does not contain the identity element. The *Cayley graph* $\mathscr{G}$ associated with G and F is the graph such that there is a bijection ϕ from the set of vertices in $\mathscr{G}$ to G and such that there is a directed edge between two vertices $a, b \in G$ if and only if $\phi(a)\phi(b)^{-1} \in F$. The subset F clearly is a generating set for G if and only if the corresponding Cayley graph is connected.

When a word w is equivalent to the identity element of the group then w is a *relator*. Each group G has the trivial relators gg^{-1} and $g^{-1}g$ for each $g \in G$. Two words w_1 and w_2 are equivalent if and only if we can transform w_1 into w_2 by insertions of relators between two consecutive symbols, insertions of relators before the first or after the last symbol, and by removing blocks of consecutive symbols which are equal to a relator.

When we can use a set of relators $r_1, \ldots, r_n$ to transform another relator r into the identity element then we say that we can deduce the relator r from the relators $r_1, \ldots, r_n$. A set of relators $r_1, \ldots, r_n$ is a set of *defining relators* if each relator of a group can be deduced from the relators in this set together with the trivial relators. When a group has an empty set of defining relators, that is, all relators can be deduced from the trivial relators, then it is a *free group*. The Cayley graph of a free group is a tree. Figure 3.1, for instance, shows the Cayley

graph of the free group with two generators. Note that a graph consists just of vertices and of directed connections between these vertices, so when we embed a graph in $\mathbb{R}^2$ we can choose the positions of the vertices and the shapes of the edges freely. This means that Figure 3.1 is just one of many ways of depicting the Cayley graph of the free group with two generators.

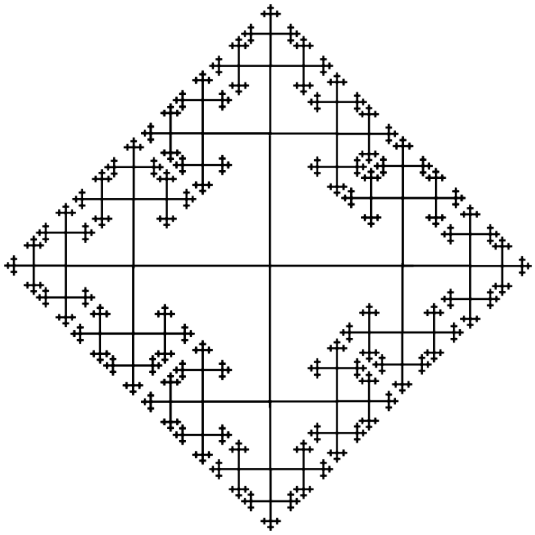


Figure 3.1: The Cayley graph of a free group with two generators

A set of generators H and a set of defining relators R together give a *group presentation* $\langle H|G\rangle$. In [5, page 13] it is shown that for any group there is such a group presentation and that a presentation defines, up to isomorphism, a unique group.

The Cayley graph of the fundamental group. We want to describe the fundamental groups of the closed orientable surfaces in more detail by giving their group presentations. In order to do this we will construct the Cayley graph corresponding to the fundamental group and use this graph to find a generating subset and a set of defining relators.

A surface of genus $g = 0$ has a trivial fundamental group so we will focus on the surfaces of genus $g > 0$. In chapters 1.3 and 2.2 we saw that a surface S_{Π_g}, with $g > 0$, is the image, under the identification map $\mathscr{I}$, of a regular Euclidean or hyperbolic polygon P_{Π_g} with $4g$ edges. In chapter 1.3 we also saw that identifying the edges and then folding the polygon and pasting the identified edges resulted in bringing together all the vertices of the original polygon. We can view the pasted lines as curves with as base point the point where all vertices of the original polygon meet. We denote these curves by $e_1, e_2, \ldots, e_{2g}$. Any set of curves $e_1' \in [e_1], e_2' \in [e_2], \ldots, e_{2g}' \in [e_{2g}]$ which are disjoint, except for their common base point, is called a set of *canonical curves*. We will refer to the set $e_1, e_2, \ldots, e_{2g}$ as *the* set of canonical curves.

Lemma 10. *The graph $\mathscr{G}$ consisting of the lifts of the canonical curves $e_1, e_2, \ldots, e_{2g}$ of a surface S_{Π_g} is an embedding of the Cayley graph associated with $\pi_1(S_{\Pi_g})$ and $H = \{[e_1], [e_2], \ldots, [e_{2g}]\}$.*

Proof. We will first prove that there is a bijection ϕ between the vertices of $\mathscr{G}$ and the elements in $\pi_1(S_{\Pi_g})$. We associate an arbitrary vertex $\tilde{O}$ of $\mathscr{G}$ with the identity element of $\pi_1(S_{\Pi_g})$. Since $\mathscr{G}$ consist of the lifts of the canonical curves, which all have the same base point, we

know that the covering map ρ maps all the vertices of $\mathscr{G}$ to the same point. This means that for any vertex $\tilde{V}$ of $\mathscr{G}$ there is a covering isometry γ such that $\tilde{V} = \gamma(\tilde{O})$ and from Theorem 4 it follows that there is only one covering isometry for which this is true. So we have a bijection between the set of vertices of $\mathscr{G}$ and the covering isometry group and by using the result of Theorem 9 we are done.

Let $\tilde{V}_1$ and $\tilde{V}_2$ be vertices in $\mathscr{G}$ and let γ_1 and γ_2 be the covering isometries such that $\gamma_i(\tilde{O}) = \tilde{V}_i$ for $i = 1, 2$. The covering isometry $\gamma_2\gamma_1^{-1}$ maps $\tilde{V}_1$ to $\tilde{V}_2$ and corresponds, by the bijection between the covering isometry group and the fundamental group, with the equivalence class of any closed curve that lifts to a curve with origin $\tilde{V}_1$ and terminus $\tilde{V}_2$. From this it is clear that $\phi(V_1)\phi(V_2)^{-1} \in H$ if and only if the two vertices are connected by an edge. □

Presentation of the fundamental group. In [10, page 149] it is shown that $\mathscr{G}$ is connected and that it partitions the covering surface into polygons isometric to S_{Π_g}. From the connectedness we know that the set of canonical curves H is a generating set for $\pi_1(S_{\Pi_g})$. From the result that each polygon is isometric to S_{Π_g} it is clear that $e_1e_2e_1^{-1}e_2^{-1}\cdots e_{2g-1}e_{2g}e_{2g-1}^{-1}e_{2g}^{-1}$ is equivalent to the identity word. When the path of a word in the fundamental group runs along more then half the edges of a polygon we can use the above relator to pull it across the polygon and let it run along the other edges of the polygon in opposite direction. In [10, page 151] it is shown that in this way we can use the relator $e_1e_2e_1^{-1}e_2^{-1}\cdots e_{2g-1}e_{2g}e_{2g-1}^{-1}e_{2g}^{-1}$ to reduce any other relator to the identity word, which makes this relator a defining relator. This proves the following theorem:

Theorem 11. $\pi_1(S_{\Pi_g}) = \langle [e_1], [e_2], \ldots, [e_{2g}] \mid [e_1e_2e_1^{-1}e_2^{-1}\cdots e_{2g-1}e_{2g}e_{2g-1}^{-1}e_{2g}^{-1}]\rangle$.

We have already seen that each closed orientable surface has a unique fundamental group. We will now show that the surfaces with an unequal number of handles are not homeomorphic. For this we abelianize $\pi_1(S_{\Pi_g})$ to obtain $\langle [e_1], [e_2], \ldots, [e_{2g}]\rangle$, which is the free group with $2g$ generators. Two free groups with an unequal number of generators are not isomorphic, and when the abelianized versions of two groups are not isomorphic then the original groups are neither, so we have that $\pi_1(S_{\Pi_i})$ is not isomorphic to $\pi_1(S_{\Pi_j})$ for $i \neq j$. Together with the proof that any closed orientable surface is homeomorphic to S_{Π_g} for some g we have finally finished the proof of the following assertion made in chapter 1.1:

Theorem 12. *A closed orientable surface is, up to homeomorphism, completely determined by its genus.*

Chapter 4

Contractibility

4.1 A characterisation of the problem

The problem of deciding whether a closed curve c with base point B is in the homotopy class of the constant curve at B is known as the contractibility problem. We can use the covering surface to give a characterisation of the null-homotopic curves:

Theorem 13. *A closed curve c on a closed orientable surface S is null-homotopic if and only if its lifts are closed.*

Proof. Let c be a closed curve with base point B and let $\tilde{c}$ be a lift with origin $\tilde{O} \in \rho^{-1}(\{B\})$.

($\Rightarrow$) If the curve c is null-homotopic then it is homotopic to the constant curve $t \mapsto B$. Since the covering map ρ is a local bijection, the lift of $t \mapsto B$ with origin $\tilde{O}$ must be the constant path $t \mapsto \tilde{O}$. Using Theorem 8 we have that the terminus of $\tilde{c}$ must be the same as the terminus of $t \mapsto \tilde{O}$ which is $\tilde{O}$, so $\tilde{c}$ is closed.

($\Leftarrow$) If the lift $\tilde{c}$ is closed then, since the universal covering surface is simple, there is a homotopy $\tilde{h}$ between $\tilde{c}$ and the constant curve $t \to \tilde{O}$. It is straightforward that the map $h = \rho \circ \tilde{h}$ is a homotopy between the curve c and the constant curve $t \to \tilde{B}$ so c is null-homotopic. $\square$

In chapter 3.3 we saw that the set $[e_1], [e_2], \ldots [e_{2g}]$ is a generating set for the fundamental group, which means that any closed curve c is homotopic to some product of canonical curves. That is, a closed curve c corresponds with a word in the fundamental group. The curve is null-homotopic if and only if this word is equivalent to the identity word. In the section about Dehn's algorithm we will find out how we can use the relators of the fundamental group to reduce the word of a contractible curve to the identity element. In the next section we will first look into the more computational aspects of the problem of deciding contractibility.

4.2 Edge curves

In chapter 1.4 we saw that each closed orientable surface S can be triangulated. A *closed edge curve* is a closed curve that runs only over the boundary of the faces in such a triangulation. A closed edge curve can be uniquely described by the non-empty ordered set of edges of the triangulation it meets. A closed edge curve is a closed curve by definition, and the next lemma shows that for a given triangulation, up to homotopy, the converse is also true.

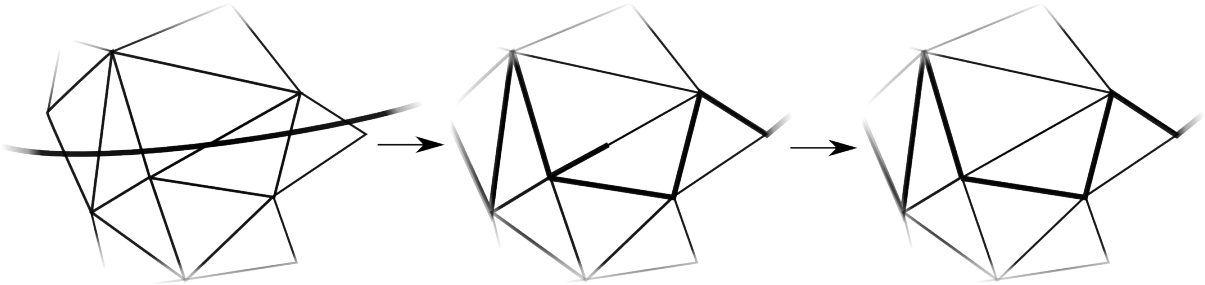


Figure 4.1: Homotopy between a curve and an edge curve

Lemma 14. *Given a triangulation of a surface S, any closed curve c on this surface is homotopic to a closed edge curve of this triangulation.*

Proof. We can cut the closed curve c into curves p_i such that each p_i intersects with the interior of at most one face of the triangulation. This can be done in such a way that the endpoints of a curve p_i lie on the boundary of a face. Each face of the triangulation is, by definition, homeomorphic to a closed triangle in the plane. This implies that a face is simply connected, so any two curves contained in a face, that have the same endpoints, are homotopic to each other. Let p_i' be the curve that connects $p_i(0)$ and $p_i(1)$ and runs over the boundary of a face. From the above we know that p_i and p_i' are homotopic.

By replacing each p_i by p_i' we get a new curve, homotopic to the old one, that runs only over the boundary of the faces. By removing spurs, which can be done homotopically, we get a curve e that can be written as the product of curves e_i, each running over an edge from one vertex to another.

Since there are only finitely many faces in a triangulation, the length of the edges and thus of the curves e_i are bounded from below. From the facts that $[0, 1]$ is compact and that a curve is a continuous map it follows that the image of a curve is compact and thus that the length of a curve is bounded. This means that e is the product of finitely many curves, so it is a closed edge curve. □

We have seen that an edge curve is a curve and that any curve is homotopic to an edge curve. Since we are only interested in curves up to homotopy we may assume that the closed curve c, of which we are to decide whether it is contractible, is an edge curve on some triangulated closed orientable surface. The advantage of this assumption is that this turns the contractibility problem into a combinatorial problem for which, as we will see, a concrete algorithm can be given that solves it.

4.3 Computational solution

Canonical curve crossings. Assume that we have a triangulation of a closed orientable surface and a closed edge curve c on this surface. The first step in the algorithm is to find out which canonical curves it crosses in which order. To do this we also write the canonical curves as edge curves. We will see that it is important that these canonical curves are disjoint, except for their common base point. It is, however, possible that there are not enough faces in the triangulation to achieve this. In [11] it is shown that in this case it is possible to refine the triangulation in such a way that the canonical curves can be given as edge curves of this

triangulation and are disjoint except for their common base point. Furthermore it is shown that the operation of refining the triangulation and constructing the canonical curves can be done in $O(gn)$ time where g is the genus of the surface and n the number of faces in the triangulation. This can be done in such a way that the curve c does not intersect the base point of the canonical curves. We will see in a moment why this is important. Since there are $2g$ canonical curves, the refinement can transform an edge in a sequence of at most $2g$ edges. This means that when the curve c consists of k edges before the refinement it has at most $2gk$ edges after the refinement.

We can now iterate over the ordered edges of the closed edge curve c and make an ordered list $\mathscr{L}$ of the canonical curves that c crosses. The canonical edge curves are oriented, so when c crosses a canonical curve we can discriminate between the side from which we cross over it. When c approaches the canonical curve from the right side then we append the canonical curve to $\mathscr{L}$ and when it crosses the canonical curve from the left side we will add the inverse of the canonical curve to $\mathscr{L}$. From the fact that the canonical curves are disjoint, except for their common base point, and the fact that c does not intersect this common base point, we have that it is indeed possible to make an *ordered* list of crossings. It is clear that the complexity of this operation is $O(gk)$ and that there are at most $2gk$ elements in $\mathscr{L}$.

The dual graph. Note that the product of the elements in $\mathscr{L}$ is a closed curve that is not homotopic to c. This follows immediately from the fact that the base point of c does not coincide with the common base point of the canonical curves, but there is a deeper difference. Usually, when we trace a curve in the Cayley graph we start at a vertex and then trace along the edges in the graph. Our edge curve c starts in the interior of a polygon of the graph and then crosses over the edges of the graph. Would we transform it homotopically as in chapter 4.2 so that it only runs over the edges of the polygons then this would result in a sequence of canonical curves different from $\mathscr{L}$.

We define the dual of the Cayley graph to be the graph that has a vertex associated with each polygon in the Cayley graph and a directed edge between two vertices if and only if the two corresponding polygons are adjacent. The direction of an edge is towards the vertex for which the common edge between the two adjacent polygons is an anticlockwise edge of the polygon corresponding to the vertex.

It is clear that $\mathscr{L}$ is a word traced out on the dual graph. From the symmetry of the Cayley graph it follows that the dual graph also is the Cayley graph of the surface. That is, the primal and the dual graphs are equal. So although the product of the curves in $\mathscr{L}$ is not homotopic to the orignial curve c we have that c is null-homotopic if and only the product of the elements in $\mathscr{L}$ is.

Testing for contractibility. We have to find out whether the product of this sequence of canonical curves is null-homotopic. The simple approach would be to build a large part of the universal covering surface, start at some vertex in this graph and then trace the elements of $\mathscr{L}$ in the graph. When the final vertex equals the first vertex then the curve is null-homotopic. The disadvantage of this approach is that we would have to build an unnecessarily large part of the covering surface which will greatly increase the complexity of the algorithm.

A more efficient way to do this is to start with some polygon, the first element in $\mathscr{L}$ tells us over which edge the curve will leave this polygon, so we add the adjacent polygon on that edge. We continue in this way until enough adjacent polygons are attached to contain

the entire curve. If the curve ends in the same polygon as where it started then it is null-homotopic. The hard part here is to keep track of the location of the current polygon in relation to the polygon where the curve starts. The solution is to build the Cayley graph in a different than usual shape, where it is easier to track the position of the current polygon in relation to the polygon where the curve starts. We describe this shape in the next section, where we use it for another reason. In [7] it is shown that the complexity of attaching another polygon is $O(g)$. Since there are at most $2gk$ elements in $\mathscr{L}$ the complexity of the operation of testing whether the curve of $\mathscr{L}$ is null-homotopic is $O(g^2k)$. This means that the total complexity of the algorithm is $O(gn+g^2k)$. In [1] it is shown that this bound can be improved to obtain a complexity of $O(n + k \log g)$.

The rubber band problem. The above method can be used to solve the rubber band contractibility problem, as stated in the introduction, by hand. For many objects it is quite easy to use a pen to draw a set of canonical curves on it. We can then give them labels and orientations and use these to obtain an ordered set of directed crossings of the rubber band. Although the product of these crossings is not homotopic to to the rubber band we can still trace it out on a hand drawn Cayley graph of the surface and find out whether this path is closed. If it is then the band is contractible and otherwise it is not.

4.4 Dehn's algorithm

Assume that we have a curve c for which $[c]$ is written as a product of the homotopy classes of the canonical curves $e_1, e_2, \ldots, e_{2g}$, since the set $\{[e_1], \ldots, [e_{2g}]\}$ is a generating set for the fundamental group, this is always possible. Dehn's algorithm tells us how we can use the defining relators of the fundamental group to reduce $[c]$ to the shortest word equivalent to $[c]$. When the curve c is contractible this shortest word will be the identity word.

The torus. The fundamental group of the torus is $\pi_1(S_{\Pi_1}) = \langle e_1, e_2 \mid e_1e_2e_1^{-1}e_2^{-1} = 1\rangle$, which can be seen by looking at the Cayley graph of the torus. We can also write this group as $\langle e_1, e_2 \mid e_1e_2 = e_2e_1\rangle$ which means that the fundamental group of the torus is commutative. From this it follows that we can write any curve $[c]$ on the torus as $[e_1]^m[e_2]^n$, such a curve is called an (m, n)-*torus curve*. By Theorem 13 we have that a curve on the torus is null-homotopic if and only if it is a $(0, 0)$-torus curve.

The curve in Figure 4.2, for instance, corresponds to $e_1e_2e_1e_2^{-1}e_2^{-1}e_1^{-1}e_1^{-1}e_2$, but using the commutativity this is the same as $e_1^0e_2^0$ so this is a $(0, 0)$-torus curve and as such it is null-homotopic.

The hyperbolic surfaces. When we first constructed the Cayley graphs $\mathscr{G}$ for the fundamental groups of the hyperbolic surfaces we took the edges in the graph to be the lifts of the canonical curves of the surface. In this way the graph consisted of regular hyperbolic polygons each with $4g$-edges. Since $\mathscr{G}$ is a graph it does not matter what the shapes of the connected components of $\tilde{S} - \mathscr{G}$ are, they do not necessarily have to be regular hyperbolic polygons.

The idea of Dehn's algorithm is to construct the Cayley graph in a different shape which enables us to develop a systematic way to reduce a null-homotopic closed curve that is traced on this graph.

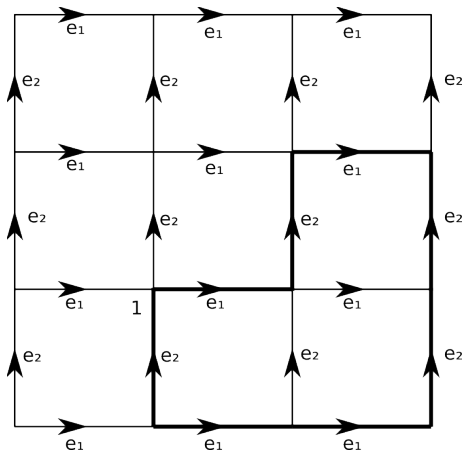


Figure 4.2: A curve on the Cayley graph of the torus

Construction of the graph. Let C_n be a circle with as center the origin and with radius $1 - \frac{1}{2^n}$. Subdivide the first circle C_1 by $4g$ vertices and to each of these vertices add $4g - 2$ edges, called spokes, that connect this vertex with the next circle C_2. The annulus between C_1 and C_2 consists of triangles and quadrilaterals. The quadrilaterals are called bridges and the sets of spokes between two consecutive bridges are called fans. Add vertices to C_2 such that these triangles and quadrilaterals become $4g$-gons. Note that 3 edges meet at the vertices that where already present and that 2 edges meet at the new vertices. We continue by adding a fans, now from the vertices on C_2 to C_3, if two edges meet at a vertex of C_2 then we add a fan consisting of $4g - 2$ edges and when a vertex is the meeting point of 3 edges then we add a fan consisting of $4g - 3$ edges. Continuation of this process yields a tessellation of the Poincaré disk by non-congruent $4g$-gons where $4g$ polygons meet at each vertex.

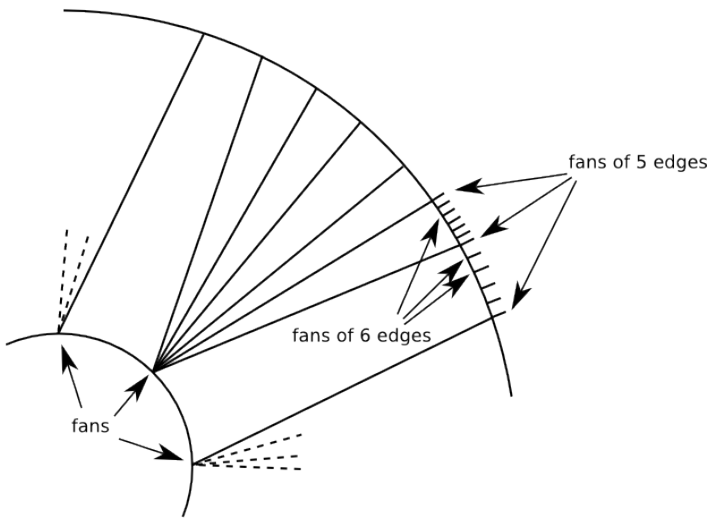


Figure 4.3: The Cayley graph of the double torus

In order to turn this skeleton into the Cayley graph of the fundamental group with as generating set the canonical curves we have to add labels and orientations to the edges. Note that since this graph is just the usual Cayley graph, as constructed in chapter 3.3, but with a different shape, the edges of the polygons still must have, in anticlockwise order, the labels and orientations: $e_1, e_2, e_1^{-1}, e_2^{-1}, \ldots, e_{2g-1}, e_{2g}, e_{2g-1}^{-1}, e_{2g}^{-1}$. Since the polygons are connected, fixing the label and orientation of one edge results in fixing the labels and orientations of all the edges in the graph.

Reduction of the curve. Formally, the input of Dehn's algorithm is a word in the fundamental group and since this group consists of homotopy classe we will denote this word by $[c]$ for some curve c. We will use the relators of the fundamental group to reduce this word. For this we will trace the word on the embedding of the Cayley graph that we have just constructed and we will reduce the path that we obtain in this way. When we view this path as the lift of a curve then we are at the same time contracting a curve.

The algorithm consists of two steps that we execute alternately. The algorithm stops when the word is reduced to the identity word or when it finds out that the word can not be reduced further. In the last case the curve is not null-homotopic and the algorithm gives a shortest word equivalent to $[c]$.

1. Use the trivial relators $e_1e_1^{-1}, e_2e_2^{-1}, \ldots, e_{2g}e_{2g}^{-1}$ to remove the so called 'spurs' of this path. The resulting path does not contain any backtracking. If the word is the identity word then the algorithm terminates, otherwise continue with step 2.

2. Let n be the largest n such that the path meets C_n. We will replace a part p of the symbols in $[c]$ by a set of less symbols.

   If $n = 0$ then let $p = [c]$. If the number of symbols in p is less then or equal to $2g$ then c is not contractible and the algorithm terminates. The part p is a word consisting of a smallest set of symbols equivalent to $[c]$.

   If $n > 0$ then the path enters C_n over a spoke, since this is the largest circle the path meets and since the path does not contain backtracking anymore it has to start moving around the circle clockwise or anticlockwise, wee Figure 4.4. The path can leave C_n over another spoke or it can run around C_n one or more times and then leave through the same spoke. Either way, it has to consecutively run over at least $4g-2$ consecutive edges of one of the polygons for which the spoke over which the path enters C_n is an edge, let p be this part of the path.

   Since p runs over at least $4g-2$ consecutive edges of a polygon it is equal to the product of at least the last $4g-2$ symbols of some cyclic permutation of the relation $e_1e_2e_1^{-1}e_2^{-1}\cdots e_{2g-1}e_{2g}e_{2g-1}^{-1}e_{2g}^{-1}$ or its inverse. By adding the inverse of this cyclic permutation after the last symbol of p in $[c]$ the new word will contain a full relator $e_1e_2e_1^{-1}e_2^{-1}\cdots e_{2g-1}e_{2g}e_{2g-1}^{-1}e_{2g}^{-1}$ which we may remove from the word. Now we continue again with step 1.

The strength of Dehn's algorithm lies in the fact that, from the way the Cayley graph is constructed, it is clear that the second step should either terminate or reduce the number of symbols in the word. From this it is clear that the algorithm always reduces a contractible word to the identity word in a finite number of steps.

Relation to the word problem. The word problem is the problem of finding out whether two words in some group are equivalent. It is an important problem in group theory and in [6] it is shown that there are groups for which the word problem is unsolvable. Two words w_1 and w_2 are equivalent if and only if $w_1^{-1}w_2$ is equivalent to the identity element. This means that Dehn's algorithm gives the solution for the word problem for the specific groups that are isomorphic to the groups generated by $2g$ elements $e_1, \ldots, e_{2g}$ and have defining non-trivial relation $e_1e_2e_1^{-1}e_2^{-1}\cdots e_{2g-1}e_{2g}e_{2g-1}^{-1}e_{2g}^{-1}$. This shows how we can use the Cayley graph, which we found through a topological and geometric approach, to solve an algebraic problem for a special class of groups.

4.5 An example

To illustrate Dehn's algorithm we will take some null-homotopic curve c on the double torus and show how we can use the relations of the fundamental group to reduce the word $[c]$ to the identity word. Let c be a curve homotopic to the following product of canonical curves:

$$e_1e_2^{-1}e_4^{-1}e_1e_2e_1^{-1}e_2^{-1}e_3e_4^{-1}e_1e_2e_1^{-1}e_2^{-1}e_3e_4e_3^{-1}e_4e_4e_3^{-1}e_4^{-1}e_1e_2e_1^{-1}e_1^{-1}$$

Figure 4.4 shows the word $[c]$ traced on the Cayley graph.

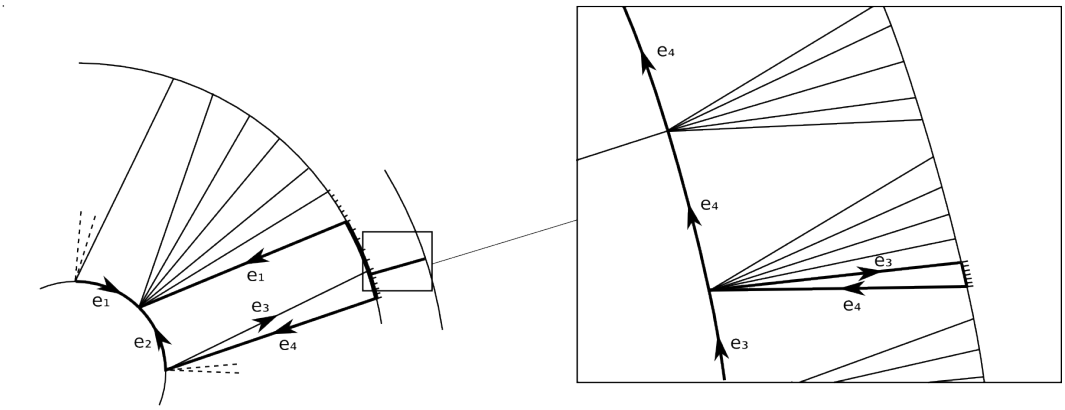


Figure 4.4: A curve on the Cayley graph

From the figure we see that the lift is closed, so c is indeed contractible. We will now apply Dehn's algorithm step by step to reduce this curve.

1. The curve does not contain spurs so we continue with step 2.

2. The greatest circle entered by the curve is C_3. We see that the curve does run around an entire polygon between C_2 and C_3 so we can remove the part $e_4^{-1}e_1e_2e_1^{-1}e_2^{-1}e_3e_4e_3^{-1}$ which is a cyclic permutation of the non-trivial defining relator $e_1e_2e_1^{-1}e_2^{-1}e_3e_4e_3^{-1}e_4^{-1}$.

1. This step did not introduce backtracking so we can continue with step 2.

2. The greatest circle entered by the lift is C_2. There are two polygons between C_1 and C_2 for which the curve runs over more then half of the edges. We choose the first polygon so we replace the part $e_4^{-1}e_1e_2e_1^{-1}e_2^{-1}e_3e_4$ by e_3 to obtain $e_1e_2^{-1}e_3e_4e_3^{-1}e_4^{-1}e_1e_2e_1^{-1}e_1^{-1}$.

1. This step again did not introduce backtracking so we continue with step 2.

2. The curve now runs entirely over the other polygon between C_1 and C_2 so we can remove the part $e_2^{-1}e_3e_4e_3^{-1}e_4^{-1}e_1e_2e_1^{-1}$ to obtain $e_1e_1^{-1}$.

1. This last step finally did introduce backtracking, which we can remove by using the relator $e_1e_1^{-1}$ to obtain the identity word.

Chapter 5

Intersections of Curves

5.1 Self-intersections

A *self-intersection* of a curve $p : [0,1] \to S$ is an unordered pair $\{t_1, t_2\}$ for different $t_1, t_2 \in [0,1]$ such that $p(t_1) = p(t_2)$. Such a pair is a self-intersection of a closed curve c if $c(t_1) = c(t_2)$ and if it is different from the pair $\{0,1\}$. Let $\theta(c)$ denote the number of self-intersection of the closed curve c, formally we have that

$$\theta(c) = \#\{\{t_1, t_2\} \mid 0 \leq t_1 < t_2 \leq 1,\ c(t_1) = c(t_2)\} - 1.$$

The closed curve in Figure 5.1 for example has four self-intersections: one at point A and the other three at point B. When a curve or a closed curve contains no self-intersections then it is a *simple curve* or a *simple closed curve*, respectively.

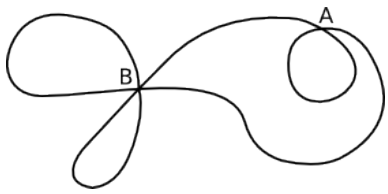


Figure 5.1: A curve with four self-intersections

In this chapter we will look at the problem of determining the minimal number of self-intersections of a curve c when we are allowed to continuously deform the curve by keeping the base point fixed. Let $\Theta(c)$ be this minimum. Formally we have that

$$\Theta(c) = \min_{c' \in [c]} \theta(c').$$

We will first solve the problem for the case where c is null-homotopic.

Lemma 15. *A null-homotopic closed curve c on a closed orientable surface S is homotopic to a simple closed curve.*

Proof. Let B be the base point of the closed curve c. In chapter 2.1 we saw that a closed orientable surface is a geometric surface, so there is an open neighbourhood N of B which is isometric to an open disk in $\mathbb{S}^2$, $\mathbb{E}^2$, or $\mathbb{H}^2$. Euclid's third postulate asserts that there exists

a circle C in N through B. Since c is null-homotopic it is homotopic to the constant curve at the base point B. The neighbourhood N is in particular homeomorphic to an open disk so it is simply connected. This means that the constant curve at the base point is homotopic to a curve that has as image the circle C, which clearly is a simple curve. □

When c is a curve on a spherical surface then it is contractible, so by the above lemma we have that c is homotopic to a simple curve and thus $\Theta(c) = 0$. In the next sections we will find out how to determine $\Theta(c)$ if c is a non null-homotopic curve on surface of genus $g > 0$. The notion of the shortest curve l in the homotopy class $[c]$ as well as the next lemma will turn out to be very helpful in solving this problem.

Lemma 16. *A pair $\{t_1, t_2\} \neq \{0, 1\}$ with $0 \leq t_1 < t_2 \leq 1$ is a self-intersection of a curve c if and only if for any lift $\tilde{c}$ either $\tilde{c}(t_1) = \tilde{c}(t_2)$ or there is another lift $\tilde{c}'$ such that $\tilde{c}(t_1) = \tilde{c}'(t_2)$.*

Proof. ($\Rightarrow$) If $\{t_1, t_2\}$ is a self-intersection of a curve c then $c(t_1) = c(t_2)$. For any lift $\tilde{c}$ we have that $c = \rho \circ \tilde{c}$. Together this gives $\rho(\tilde{c}(t_1)) = c(t_2)$, in other words, $\tilde{c}(t_1) \in \rho^{-1}(\{c(t_2)\})$. This means that either $\tilde{c}(t_1) = \tilde{c}(t_2)$ or there is another lift $\tilde{c}'$ such that $\tilde{c}(t_1) = \tilde{c}'(t_2)$.

($\Leftarrow$) If on the other hand $\tilde{c}(t_1) = \tilde{c}(t_2)$ or there is a lift $\tilde{c}'$ such that $\tilde{c}(t_1) = \tilde{c}'(t_2)$ then $\rho(\tilde{c}(t_1)) = \rho(\tilde{c}(t_2))$ or $\rho(\tilde{c}(t_1)) = \rho(\tilde{c}'(t_2))$. Either way we have that $c(t_1) = c(t_2)$. □

5.2 Self-intersections on the torus

Let c be a non null-homotopic (m, n) torus curve with its base point B coinciding with the base points of the canonical curves e_1 and e_2. Let $\rho : \mathbb{E}^2 \to S$ be a locally isometric covering map such that ρ maps the origin $\tilde{O}$ of $\mathbb{E}^2$ to B. In chapter 2.4 it is shown that this map always exists. We know that we can lift the closed curve c to a curve $\tilde{c}$ on the covering surface and that by fixing the origin of this curve to be $\tilde{O}$ this lift is unique. Let $\tilde{T} = \tilde{c}(1)$ be the terminus of this curve.

The lifts of the canonical curves e_1 and e_2 of the torus provide a tessellation – by squares – of the Euclidean plane. Since the base point of c coincides with the base points of e_1 and e_2 we know that $\tilde{O}$ and $\tilde{T}$ are vertices in this tessellation. From the fact that c is not null-homotopic we know that $\tilde{O} \neq \tilde{T}$.

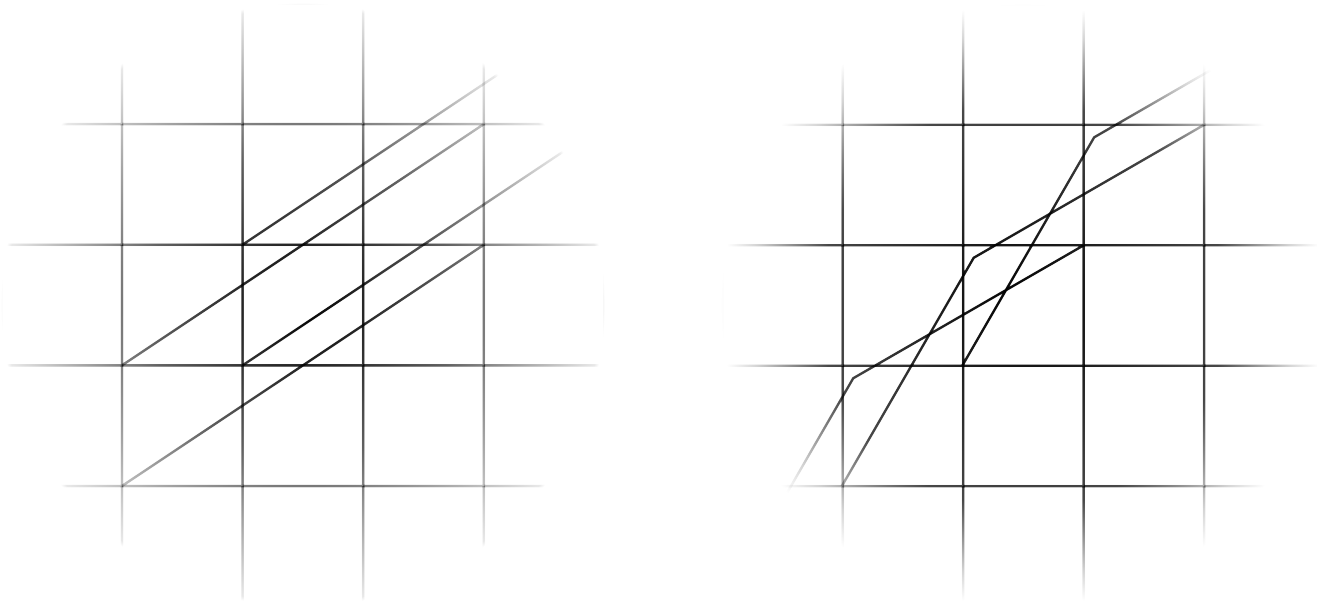
Let l be the shortest (m, n)-torus curve. Since the lift of any curve that is homotopic to c must go from $\tilde{O}$ to $\tilde{T}$ on the covering surface we know that $l = \rho \circ \tilde{l}$, where $\tilde{l}$ is the path $\tilde{l}(t) = t\tilde{T}$.

Let s be the number of vertices of the tessellation in the interior of $\tilde{l}$. Now we claim that $\Theta(c) = s$. We will prove this by showing that both $\Theta(c) \leq s$ and $\Theta(c) \geq s$.

Lemma 17. $\Theta(c) \leq s$

Proof. We show that $\Theta(c) \leq s$ by showing that there is a curve $c' \in [c]$ for which $\theta(c') \leq s$. We will consider the two cases $s = 0$ and $s > 0$ separately.

For the case where $s = 0$ let c' be the shortest curve in $[c]$. Since $s = 0$ we know that the lifts of c' are straight line segments that connect two vertices of the tessellation in such a way that they do not intersect other vertices of the tessellation. Since each lift has the same length and direction (they are translations of each other) we know that the lifts do not intersect, see Figure 5.2(a). If c contains a self-intersection then by Lemma 16 its lifts should intersect, which they do not, so c has no self-intersections and $\theta(c') = 0 \leq s$.



(a) Lifts of the shortest $(3,2)$-torus curve ($s = 0$)

(b) Lifts of a $(3,3)$-torus curve ($s = 2$)

Figure 5.2: Lifts of torus curves

If $s > 0$ we can again look at the shortest curve $l \in [c]$, but now we have that the interior of each lift $\tilde{l}$ intersects with at least one vertex of the tessellation. This means that for each lift of l there is another lift such that these lifts overlap, which by Lemma 16 means that $\theta(l) = \infty$. This means that for the case where $s > 0$ we need to take another approach.

Let $\tilde{l}$ be some fixed lift of l. We will construct a curve c' such that $\theta(c') \leq s$. Let $\tilde{M}$ be a point on the perpendicular bisector of $\tilde{l}$ such that $d(\tilde{M}, \tilde{l}) = \epsilon$ for some $\epsilon > 0$. Let $\tilde{c}'$ be the curve that runs in straight line segments from $\tilde{O}$ to $\tilde{M}$ and then from $\tilde{M}$ to $\tilde{T}$.

The curve c' is simple, so by Lemma 16 we have that a pair $\{t_1, t_2\} \neq \{0, 1\}$ with $0 \leq t_1 < t_2 \leq 1$ is a self-intersection of a curve c only if for any lift $\tilde{c}$ there is another lift $\tilde{c}'$ such that $\tilde{c}(t_1) = \tilde{c}'(t_2)$. By choosing ϵ small enough we have that a lift that does not originate or terminate in a vertex of the tesselation in the interior of $\tilde{l}$, does not intersect $\tilde{c}'$. Furthermore we have that the lifts originating and terminating from a vertex in the interior of $\tilde{l}$ both intersect $\tilde{c}'$, but due to the symmetry this leads to only one intersection pair. Since there are s vertices in the interior of $\tilde{l}$ and since each vertex is responsible for at most one intersection pair we have that $\theta(c') \leq s$. □

The proof of the other inequality $\Theta(c) \geq s$ makes use of the next lemma, which is proven in [9, page 30].

Lemma 18. *If A, B, C, and D are points in cyclic order on the boundary of a polygon $\mathscr{P}$, and p is a simple curve from A to C which elsewhere lies in the interior of $\mathscr{P}$, then p separates B from D in $\mathscr{P}$.*

Lemma 19. $\Theta(c) \geq s$

Proof. We prove that $\Theta(c) \geq s$ by proving that for any $c' \in [c]$ we have that $\theta(c') \geq s$. Let c' be some closed curve in the homotopy class of c and let $\tilde{c}'$ be a lift of c'. We may assume

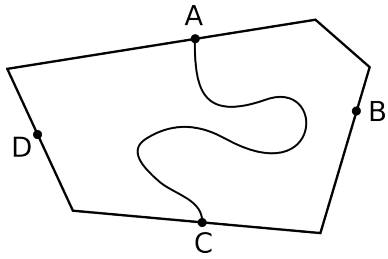


Figure 5.3: Illustration of Lemma 18

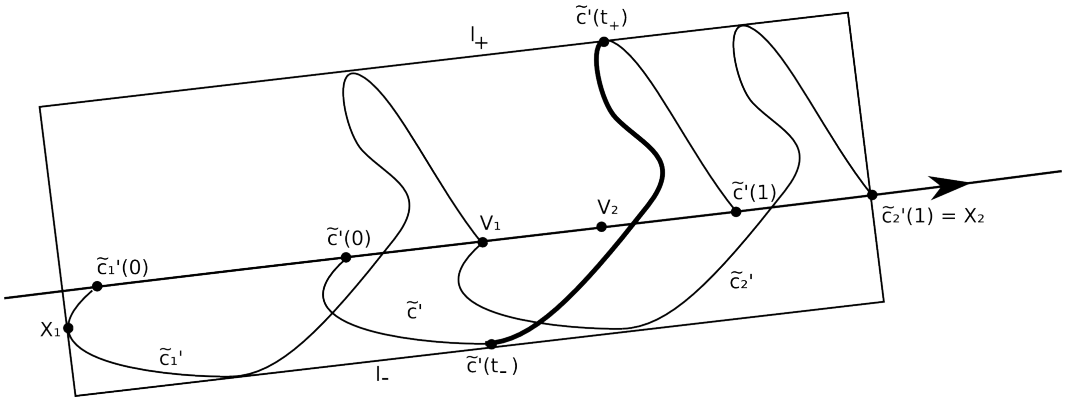


Figure 5.4: Illustration of the proof of Lemma 19

that the curve $\tilde{c}'$ is simple, for if it contains a self-intersection $\{t_1, t_2\} \neq \{0, 1\}$, then we can just remove the part of the curve c' between t_1 and t_2. We may do this because removing a part of the curve can only lower the number of self-intersections so when $\theta(c') \geq s$ holds for the new curve it certainly holds for the origional curve.

Since $\theta(c')$ is non-negative we see that $\theta(c') \geq s$ is certainly true for the case where $s = 0$. Now assume that $s > 0$ and let $\tilde{V}_1, \tilde{V}_2, \ldots, \tilde{V}_s$ be the vertices of the tessellation that intersect with the interior of the straight line segment connecting $\tilde{c}'(0)$ and $\tilde{c}'(1)$.

Let l be an oriented straight line through $\tilde{c}'(0)$ and $\tilde{c}'(1)$. Define the height y of a point $X \in \mathbb{E}^2$ as

$$y(X) = \begin{cases} 0 & X \text{ lies on } l \\ d_{\mathbb{E}^2}(X, l) & X \text{ lies to the left of } l \\ -d_{\mathbb{E}^2}(X, l) & X \text{ lies to the right of } l \end{cases}$$

The map $y \circ \tilde{c}'$ is continuous since it is the composition of two continuous maps, and the set $[0, 1]$ is compact, so the maximum and minimum heights

$$y_+ = \max_{t \in [0,1]} y(c(t)) \quad \text{and} \quad y_- = \min_{t \in [0,1]} y(c(t))$$

are well defined.

Let $t_\pm \in [0, 1]$ be two numbers such that $y(\tilde{c}'(t_\pm)) = y_\pm$ and such that $y_- < y(\tilde{c}'(t)) < y_+$ for all t between t_- and t_+. Let $l_\pm$ be lines through $\tilde{c}'(t_\pm)$, respectively, that are parallel to l.

Now consider two lifts $\tilde{c}'_1$ and $\tilde{c}'_2$ of c' that have the point V_1 as its terminus and origin, respectively. When we translate $\tilde{c}'$ parallel to l over distance $d(\tilde{c}'(0), V_1)$ then we obtain $\tilde{c}'_2$,

and when we do the same over the distance $-d(\tilde{c}'(1), V_1)$ then we obtain $\tilde{c}'_1$. This means that the maximum and minimum heights of points in $\tilde{c}'_i([0,1])$ for $i = 1, 2$ are the same as those of $\tilde{c}'$ which means that $\tilde{c}'_1$ and $\tilde{c}'_2$ lie in the strip between the lines $l_\pm$.

Now take a rectangle $\mathscr{P}$ with two of its sides lying on l_+ and l_-, and the other two sides chosen such that $\tilde{c}'$, $\tilde{c}_1$, and $\tilde{c}_2$ lie in $\mathscr{P}$ but such that $\mathscr{P}$ is as small as possible. Let X_1 and X_2 be two points where $\tilde{c}_1$ and $\tilde{c}_2$ intersect the edges of $\mathscr{P}$ which are not contained in $l_\pm$, see Figure 5.4. The points $\tilde{c}'(t_-)$, X_1, $\tilde{c}'(t_+)$, and X_2 lie in cyclic order on the boundary of $\mathscr{P}$. Furthermore we have that $\tilde{c}'$ restricted to $[\min\{t_-, t_+\}, \max\{t_-, t_+\}]$ is a simple path whose interior does not intersect the boundary of $\mathscr{P}$. So using Lemma 18 we have that $\tilde{c}'$ must intersect $\tilde{c}'_1$ or $\tilde{c}'_2$ and since the lifts clearly do not intersect on an endpoint the interior of two lifts must intersect.

This leads to an intersection pair $\{t_1, t_2\}$. We can do the same for the other vertices V_i, and since the distance over which we have to translate $\tilde{c}'$ to obtain $\tilde{c}'_1(0)$ and $\tilde{c}'_2(0)$ is different for each V_i it is clear that each V_i leads to a unique intersection pair. So there are at least s different intersection pairs and as such we have that $\theta(c') \geq s$. □

Theorem 20. *If c is an (m, n) torus curve then* $\Theta(c) = \gcd(m, n) - 1$.

Proof. Let $\tilde{l}$ be the lift with origin $(0, 0)$, of the shortest curve in $[c]$. This means that $\tilde{l}$ is the straight line segment connecting $(0, 0)$ and (m, n). We have seen that $\Theta(c)$ equals the number of vertices of the tessellation in the interior of $\tilde{l}$.

Let (k, l) be the first vertex we meet by traversing $\tilde{l}$ from its origin $(0, 0)$ to its terminus (m, n). From the symmetry of the tessellation it follows that for any $q \in \mathbb{N}$ the point (qk, ql) is a vertex, and the interior of the line segment with endpoints (qk, ql) and $((q+1)k, (q+1)l)$ contains no vertices. This means that there is a q such that $(m, n) = (qk, ql)$ and the interior of the line segment with endpoints $(0, 0)$ and (m, n) contains $q - 1$ vertices.

From the equality $(m, n) = (qk, ql)$ it follows that q is a divisor of both m and n, we will now show that it is also the greatest common divisor. Assume that there is a $r > q$ such that $r|m$ and $r|n$. Then there are k' and l' such that $(m, n) = (rk', rl')$, but since $r > q$ this means that $k' < k$ and $l' < l$ which is in contradiction with the fact that (k, l) is the vertex on l closest to the origin. So the number of vertices on l is $q - 1 = \gcd(m, n) - 1$. □

5.3 Simple curves on hyperbolic surfaces

In this section we will consider the problem of deciding whether a closed curve c on a surface S of genus $g > 1$ is homotopic to a simple curve or not. The lifts of the canonical curves $e_1, \ldots, e_{2g}$ provide a tessellation of the covering surface by hyperbolic $4g$-gons. To see whether $\Theta(c) = 0$ we will again look at the shortest curve $l \in [c]$. The lift $\tilde{l}$ of l is a hyperbolic geodesic segment between two vertices in the tessellation. We can solve the problem in two steps:

1. Prove that c is homotopic to a simple curve if and only if l is simple.

2. Find out whether l is simple.

We will start by briefly describing step 2 after which the remaining part of the section will be devoted to the proof of the statement of step 1.

Testing whether l is simple. When we want to use Lemma 16 to test whether l is simple we would have to draw each of the infinitely many lifts of l, which is not possible. Instead we take one polygon $\mathscr{P}$ and one lift whose origin lies in $\mathscr{P}$. When the lift reaches the boundary of the polygon then it enters $\mathscr{P}$ again at the corresponding edge. Figure 2.1, used to illustrate the completeness proof in chapter 2.3, shows what this looks like for the double torus. The curve l does have a self-intersection if and only if two or more of the geodesic segments intersect each other.

Proving the first statement. The 'if' part of 1 is trivial and to prove the 'only if' part we will prove that if l is not simple then c is not homotopic to a simple curve. If l is not simple then considering Lemma 16 there must be two lifts of l that have an intersection point. For the torus this meant that the lifts would overlap, in the hyperbolic case, however, there is the other possibility that two lifts can cross each other.

When two lifts of l overlap then we can use the same method as in the proof of Lemma 19, to show that the lifts of any curve homotopic to c must intersect at a point other than their endpoints. Instead of repeating the proof we will state the adjustments that we need to make for the hyperbolic case. Instead of letting l be an oriented Euclidean line through $\tilde{c}(0)$ and $\tilde{c}(1)$ we let it be the oriented hyperbolic geodesic through those points, and instead of the parallel lines $l_\pm$ we take the hyperbolic equidistants, which are not hyperbolic geodesics. Finally, we use hyperbolic translation parallel to the geodesic l to obtain the curves $\tilde{c}_1$ and $\tilde{c}_2$. Since there is only one hyperbolic translation that maps two different points to two other points, all lying on the same geodesic, we know that this translation is a covering isometry, so $\tilde{c}_1$ and $\tilde{c}_2$ are lifts of c. With these adjustments the proof works for the case of overlapping lifts in the hyperbolic space, we finish with the case where the two geodesic segments cross each other:

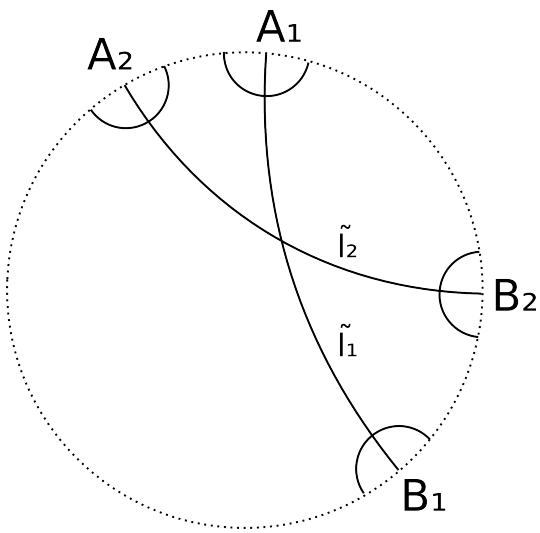


Figure 5.5: Illustration by Lemma 21

Lemma 21. *Let c be a closed curve on a closed orientable surface S of genus $g > 1$. If there are lifts $\tilde{c}_1$ and $\tilde{c}_2$ such that the geodesic $\tilde{l}_1$ through $\tilde{c}_1(0)$ and $\tilde{c}_1(1)$ intersects the geodesic $\tilde{l}_2$ though $\tilde{c}_2(0)$ and $\tilde{c}_2(1)$ then c contains a self-intersection.*

Proof. The geodesics $\tilde{l}_i$ for $i = 1, 2$ depicted in the Poincaré disk have limit points that lie on the boundary of this disk. Denote these limit points by A_1, B_1, A_2, and B_2, see Figure 5.5. We take a disk around each of these limit points and let $\mathscr{P}$ be the Poincaré disk including its boundary excluding the interiors of the disks around the four limit points.

In [9, page 193] it is stated that the terminus of the lift of c_1 that has origin $\tilde{c}_1(1)$, also lies on $\tilde{l}_1$. This means that the covering isometry which is the hyperbolic translation with $\tilde{l}_1$ as invariant geodesis such that $\gamma_1(\tilde{c}_1(0)) = \tilde{c}(1)$ maps the lift $\tilde{c}_1$ to this new lift.

Now we will view the Poincaré disk as a part of the Euclidean plane. When we continue the process of translation, in both directions along $\tilde{l}_1$, the points in the image of a translated copy of $\tilde{c}_1$ will converge to the points A_1 and B_1, respectively. This means that it is possible to choose the disks around A_2 and B_2 small enough such that none of these copies (nor the original curve $\tilde{c}_1$) intersect with either of these two disks. We can do the same for the curves $\tilde{c}_2$, also choosing the disks around A_1 and B_1 small enough such that none of the translations of $\tilde{c}_2$ intersect with one of these two disks.

After finitely many times translating, forward or backward, the curve $\tilde{c}_1$ along $\tilde{l}_1$, it will intersect with the disk around A_1 and B_1, respectively. This means that the curve that is the product of all these translated curves, where we follow the last two curves only to the point where they intersect the disk around the limit point, is a curve that connects two points that lie on $\partial\mathscr{P}$.

We can remove the loops in the curve to obtain a simple curve, this is allowed since this operation will not introduce intersections. Since the lines $\tilde{l}_1$ and $\tilde{l}_2$ cross each other we know that the limit points A_1, A_1, B_1, and B_2 lie in cyclic order on $\partial\mathscr{P}$. This means that we can use Lemma 16 to deduce that one of the translated copies of the curve $\tilde{c}_1$ must intersect one of the copies of the curve $\tilde{c_2}$ which proves the theorem. Note that strictly speaking we cannot use this lemma since $\partial\mathscr{P}$ is not a polygon, but it would be easy to transform $\mathscr{P}$ such that $\partial\mathscr{P}$ is a polygon, or to use the Jordan curve theorem in order to prove a similar lemma that also holds for non-polygons. $\square$

Conclusion

The conclusion concerning the contractibility problem is that we have a simple characterization of contractible curves on closed orientable surfaces. There also is a concrete algorithm to decide whether an edge curve is contractible. Finally, we have an algorithm that solves the word problem for the fundamental groups of the closed orientable surfaces.

The most important conclusion concerning self-intersection of curves on closed orientable surfaces is that the minimal number of self-intersections of an (m, n)-torus curve is $\gcd(m, n) - 1$. Furthermore we have shown how to use the hyperbolic covering surface to see whether a curve on a closed orientable surface of genus $g > 1$ is homotopic to a simple curve.

Concerning the theory that we discussed before considering these two problems we can conclude that the representation of closed orientable surfaces by the identification spaces of orientable normal form schemata is useful in proving that each closed orientable surface is, up to homeomorphism, completely determined by its genus. This representation is also useful in the geometric realization of the surfaces and thus in the construction of the pencil map.

We can conclude that the covering surface and covering map are very useful in discussing properties of closed curves on surfaces. Finally, we can conclude that the geometric properties of the geometric covering surfaces and pencil map are not needed for all results, for instance, the results about contractibility do not need geometry at all, but can be very useful for other results, for instance in the proofs about self-intersections.

Bibliography

[1] T. K. Dey. A new technique to compute polygonal schema for 2-manifolds with application to null-homotopy detection. In *SCG '94: Proceedings of the tenth annual symposium on Computational geometry*, pages 277–284, New York, NY, USA, 1994. ACM.

[2] P.H. Doyle and D.A. Moran. A short proof that compact 2-manifolds can be triangulated. *Inventiones Mathematicae*, 5:160 – 162, 1968.

[3] W. Fulton. *Algebraic Topology: A First Course.* Number 153 in Graduate Texts in Mathematics. Springer, 1995.

[4] F. H. Lutz. Császár's torus. *Electronic Geometric Models*, (2001.02.069), 2001.

[5] W. Magnus, A. Karrass, and D. Solitar. *Combinatorial Group Theory.* Dover, 1975.

[6] P. S. Novikov. *On the algorithmic unsolvability of the word problem in group theory.* Trudy Mat. Inst. im. Steklov. no. 44. Izdat. Akad. Nauk SSSR, Moscow, 1955.

[7] H. Schipper. Determining contractibility of curves. In *SCG '92: Proceedings of the eighth annual symposium on Computational geometry*, pages 358–367, New York, NY, USA, 1992. ACM.

[8] S. Stahl. *A Gateway to Modern Geometry, The Poincaré Half-Plane.* Jones and Bartlett publishers, Sudbury, Massachusetts, 2008.

[9] J. Stillwell. *Classical Topology and Combinatorial Group Theory.* Number 72 in Graduate Texts in Mathematics. Springer-Verlag, New York, NY, 1980.

[10] J. Stillwell. *Geometry of Surfaces.* Universitext. Springer-Verlag, New York, NY, 1992.

[11] G. Vegter and C. K. Yap. Computational complexity of combinatorial surfaces. In *SCG '90: Proceedings of the sixth annual symposium on Computational geometry*, pages 102–111, New York, NY, USA, 1990. ACM.