NOTES FOR MATH 5520, SPRING 2011

DOMINGO TOLEDO

1. Outline

This will be a course on the topology and geometry of surfaces. This is a continuation of Math 4510, and we will often refer to the notes for that course [12]. We first recall the definition of a surface. The first two parts are in Definition 6.1 of [12].

Definition 1.1. A topological space S is called:

(1) A *topological surface* if it is a Hausdorff space with a countable basis and it has the property that every $x \in S$ has a neighborhood U which is homeomorphic to an open set in $\mathbb{R}^2$, in other words, there exists a covering $\{U_\alpha\}_{\alpha\in A}$ for some index set A, and for each $\alpha \in A$ there exists a homeomorphism $\phi_\alpha : U_\alpha \to V_\alpha$, where $V_\alpha \subset \mathbb{R}^2$ is open. These homeomorphisms are called *coordinate charts*.
(2) A *differentiable surface* or a *smooth surface* if it is a topological surface and the above homeomorphisms (or coordinate charts) can be chosen to have the following property: whenever $U_\alpha \cap U_\beta \neq \emptyset$, the homeomorphism $\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta)$ is smooth. The maps $\phi_\alpha \circ \phi_\beta^{-1}$ are called the *transition maps* between charts.
(3) A *topological surface with boundary* is a Hausdorff space S with a countable basis with the property that every $x \in S$ has a neighborhood U and a homeomorphism $\phi : U \to V$ where V is an open set in $\mathbb{R}^2_+ = \{(x, y) \in \mathbb{R}^2 : y \geq 0\}$.
(4) If S is a topological surface with boundary and x and ϕ are as in the definition, the point x is called an *interior point* if $\phi(x) \in \{(x, y) : y > 0\}$ and X is called a *boundary point* if $\phi(x) \in \{(x, y) : y = 0\}$.
(5) A *differentiable surface with boundary* or *smooth surface with boundary* is a topological surface with boundary in which all the transition maps $\phi_\alpha \circ \phi_\beta^{-1}$ are smooth.

Remark 1.1. (1) The corresponding definitions in any dimension n, with $\mathbb{R}^2$ replaced by $\mathbb{R}^n$ and $\mathbb{R}^2_+$ replaced by $\mathbb{R}^n_+ = \{(x_1, \ldots, x_n) : x_n \geq 0\}$ are called topological or differentiable n manifolds, or topological or differentiable n-manifolds with boundary.

Date: April 26, 2011.

(2) It is true, but not trivial to prove, that the dimension n is a topological invariant. It is also true but not trivial to prove that homeomorphisms take boundary points to boundary points and interior points to interior points. This last fact we will be able to prove for surfaces ($n = 2$) once we have the fundamental group.
(3) The conditions of Hausdorff and countable basis are needed for correctness of the definition, but it is not clear why they are required. One equivalent formulation of these conditions is to require that S be a *metrizable* space locally homeomorphic to the plane. I chose the formulation above because may be easier to check. We will not prove the equivalence of these two formulations, and for the most part we will forget these conditions since they will be automatic in the examples we consider.

Example 1.1. Here are some examples of smooth surfaces. A good reference is the first chapter of [9]. In all these examples it will be clear that they have a countable basis because $\mathbb{R}^n$ does, and all the examples are subspaces or quotient spaces of subspaces of some $\mathbb{R}^n$, so they still have a countable basis. The Hausdorff property would be easy to check in each case.

(1) The *sphere* $S^2 \subset \mathbb{R}^3$ defined by $S^2 = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}$. It is checked in Example 6.3 of [12] that S^2 is a smooth surface.

(2) The *torus* T^2, that can be defined (up to homeomorphism, or up to diffeomorphism) in one of three equivalent ways: as $S^1 \times S^1$, as the identification space $\mathbb{R}^2/\mathbb{Z}^2$, or as the identification space $[0, 1] \times [0, 1]/((x, 0) \sim (x, 1), (0, y) \sim (1, y))$. See Example 4.4 of [12] for more details. In particular, the neighborhoods pictured in Figures 4.2 and 4.3 of [12] are homeomorphic to disks in $\mathbb{R}^2$, showing that T^2 is a topological surface, and a bit more care shows that it is a smooth surface.

(3) The *Klein Bottle* K of Definition 4.5 of [12], see Figure 4.4. It is the identification space $[0, 1] \times [0, 1]/\sim$, where the identification is $(x, 0) \sim (x, 1)$ and $(0, y) \sim (1, 1 - y)$. Again it is a smooth surface.

(4) The *Möbius band* M of Definition 4.6 of [12]: $M = [0, 1]\times[-1, 1]/((0, y) \sim (1, -y))$. This is an example of a surface with boundary.

(5) The disk $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ and the cylinder $C = [0, 1] \times [-1, 1]/((0, y) \sim (1, y))$ are other examples of surfaces with boundary.

(6) The *projective plane* P^2 is defined to be the quotient of S^2 by the identification $x \sim -x$. In other words, antipodal points on S^2 are identified. Points in P^2 are in one to one correspondence with lines through the origin in $\mathbb{R}^3$, just as points in S^2 are in one to one correspondence with rays through the origin in $\mathbb{R}^3$. To check that P^2 is a smooth manifold, let $p : S^2 \to P^2$ be the projection map, that

to $x \in S^2$ assigns the antipodal pair $\{x, -x\} \in P^2$. The six sets $U_z^{\pm}$, $U_y^{\pm}$, $U_x^{\pm}$, where the corresponding coordinate has the corresponding sign, used in Example 6.3 of [12] to cover S^2 project to 3 sets that cover P^2. Namely since every antipodal pair has a representative (x, y, z) with one of $x, y, z > 0$, the ones with superscript + suffice. Call their images under p U_z, U_y, U_x, and use the same formulas for the transition maps to check that they are smooth.

(7) Another way to describe the projective plane P^2 is as follows: Let $S^2_+ = \{(x, y, z) \in S^2 : z \geq 0$ and ∂S^2_+ denote its boundary $\{z = 0\}$. Consider the map $f : S^2_+ \to P^2$ obtained as the composition:

$$S^2_+ \subset S^2 \to P^2.$$

Then f is surjective, it is injective on the interior of S^2_+ and 2 to one on its boundary. It induces a continuous bijection $g : (S^2_+/\sim) \to P^2$ where $\sim$ is the equivalence relation $x \sim -x$ for all $x \in \partial S^2_+$. It is easy to see that g is a homeomorphism (see Example 2.2 below), so we get that P^2 is homeomorphic to a hemisphere with antipodal points on its boundary identified. Since a hemisphere is homeomorphic to a disk, we get the final description: P^2 is homeomorphic to a disk with antipodal points on its boundary identified.

(8) The *surface of genus* 2 is defined at the end of §4 of [12] as an identification space of an octagon, see Figure 4.6 and the pictures on pp. 300–301 of [6], to justify that the following identifications give the picture of a surface "of genus two" or "with two holes".

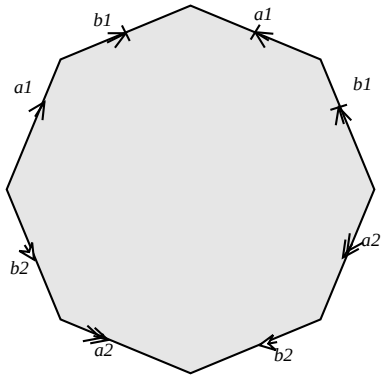


FIGURE 1.1. Surface of Genus Two

(9) *Presenting a surface as a quotient of a polygon*: Before we define the surface of genus g, we explain a convenient notation for describing the quotient space of a polygon by identifying the of edges of its boundary in pairs, using Figure 1.1 as an example. We choose a direction around the perimeter of the polygon, say clockwise, label

the edges to be identified with the same letter, draw similar arrows on each of the two edges that are identified indicating how they are identified: a monotone map identifying tail with tail. Move around the perimeter and write down the symbols for the edges, either a letter, if you travel in the same direction as the arrow, or its inverse, if you travel in the opposite direction. Thus the identification in Figure 1.1 could be written as

$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1},$$

while the identifications in the boundaries of the following four squares, running counterclockwise from the bottom, can be written as

$$aba^{-1}b^{-1} \text{ gives T} = \text{torus } T^2$$
$$aba^{-1}b \text{ gives K} = \text{Klein bottle}$$
$$abb^{-1}a^{-1} \text{ gives S} = \text{sphere } S^2$$
$$abab \text{ gives P} = \text{projective plane } P^2$$

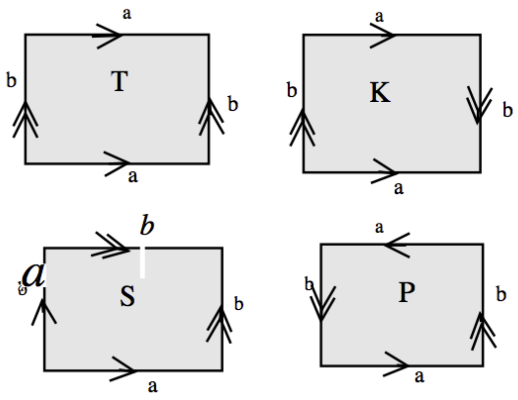


FIGURE 1.2. Four Surfaces as quotients of the square

(10) *Checking that the quotient space is a surface* requires finding for each point in the identification space a neighborhood homeomorphic to an open set in $\mathbb{R}^2$, say, a disk. Disk neighborhoods are immediate for any point in the interior of the polygon. For boundary points that are in the interior of an edge a set that projects to a disk in the quotient space is illustrated in the first picture of Figure 1.3. For the vertices we have to check how many distinct ones there are in the quotient space. In the case of the octagon for the genus two surface all the vertices go to a single point in the quotient space, and set that projects to a disk neighborhood in the quotient space is as in the second picture of Figure 1.3

Observe that the parts of the neighborhood of an interior point of an edge fit not only topologically, but also geometrically, into a

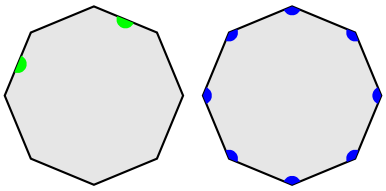


FIGURE 1.3. Identification Space is Locally Homeomorphic to the Plane

disk in the plane, while the parts of the neighborhood of the vertex only fit topologically to a disk. Observe also that figures 4.2 and 4.3 of [12] show that boundary points in the square project to points in the identification space with a disk neighborhood. In that case, even the parts of the neighborhood of the vertex fit geometrically into a disk. The significance of this topological versus geometric fitting will become clear later. A good exercise would be to find the corresponding neighborhoods for the remaining pictures in Figure 1.2 (those labeled K, S, P).

(11) Finally we define, for any $g \geq 1$, the *surface of genus* g, denoted by Σ_g, to be the quotient space of a regular $4g$-gon by the identification labeled

$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}, \tag{1.1}$$

generalizing the identification of Figure 1.1 in the obvious way. We will write Σ_0 for the sphere S^2 and note that Σ_1 is the same as the torus T^2. Thus Σ_g is defined for $g = 0, 1, 2, 3, \ldots$.

Our task is to understand surfaces both topologically and geometrically. We will restrict ourselves mostly to *compact, connected* surfaces (without boundary). We will review compactness in the next section, and connectedness was discussed in Section 5 of [12]. We will also need the concept of *orientability*, which is difficult to define precisely. This will be our temporary definition:

Definition 1.2. A surface S is *non-orientable* if it contains a Möbius band M, meaning that there is a continuous map $\phi : M \to S$ which is a homeomorphism onto its image. Otherwise, S is called *orientable*.

Example 1.2. (1) The Klein bottle K is not orientable. In its description in (3) of Example 1.1, the projection to K of the the set $[0, 1] \times [\frac{1}{4}, \frac{3}{4}]$ is homeomorphic to M.

(2) The projective plane P^2 is not orientable. If we describe P^2 as in (7) of Example 1.1, as the disk $D = \{x^2 + y^2 \leq 1\}$ with identification $(x, y) \sim (-x, -y)$ for $x^2 + y^2 = 1$, then the projection to the quotient

space of the set $\frac{1}{2} \leq x^2 + y^2 \leq 1$ is a Möbius band (see the homework problems).

(3) The sphere S^2 is orientable. This is a reasonable statement, but it is not clear how to prove it rigorously.

(4) More generally, the surfaces Σ_g, $g = 0, 1, 2, \ldots$ of (11) of Example 1.1 are orientable. We will be able to prove this later once we have the concept of fundamental group and can reduce the problem to a more tractable one.

We will (partially) prove the following theorem:

Theorem 1.1. *Every compact, connected, orientable surface is homeomorphic to Σ_g for some $g = 0, 1, 2, \ldots$. Moreover, if Σ_g is homeomorphic to Σ_h, then $g = h$.*

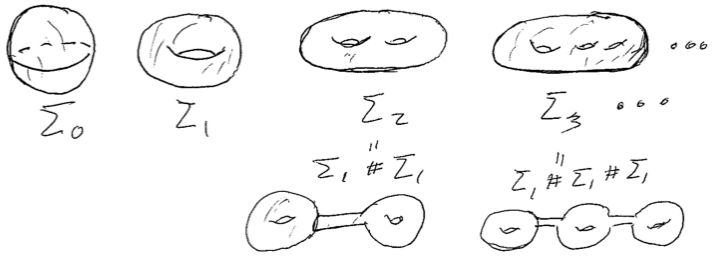


FIGURE 1.4. The Surfaces Σ_g

There is a more comprehensive theorem which includes the non-orientable surfaces. This is best stated in terms of *connected sum* of surfaces. See p. 8 of [9] for a definition of the connected sum operation $S_1 \# S_2$ of surfaces.

Whenever we use this operation we have to keep in mind that it depends on choices: choose a disk in each surface, and a homeomorphism between their boundaries. It turns out the any choices of these three ingredients gives a homeomorphic result. Thus connected sum is an operation on *homeomorphism classes of surfaces*. In fact, in our discussion of classification of surfaces we will always identify homeomorphic surfaces. Connected sum is a *commutative, associative operation with unit* S^2, but no inverses.

We have that $\Sigma_g \# \Sigma_h = \Sigma_{g+h}$, in particular, $\Sigma_g = T^2 \# \ldots \# T^2$ (g summands). Thus every Σ_g, $g \geq 1$, is homeomorphic to a connected sum of tori, see Figure 1.4.

Theorem 1.2. *Every compact, connected surface is homeomorphic to S^2 or to a connected sum of tori or to a connected sum of projective planes.*

1.1. **Summary of Geometric and Topological Classification.** This topological classification is only part of the story. We also want a geometric classification and how it interacts with the topological one. We will state the classifications that we have in mind in the following table. For simplicity only orientable surfaces are considered. Observe that the surfaces are divided into three classes, one contains only the sphere S^2, one only the torus T^2, and the third class contains all the others. (For non-orientable surfaces, add P^2 to the first class, K to the second, and all others to the third class).

Surface	Σ_0	Σ_1	$\Sigma_2, \Sigma_3, \ldots$
genus g	0	1	$2, 3, \ldots$
Euler characteristic χ	2	0	$-2, -4, \ldots$
Fundamental Group π_1	$\{e\}$	$\mathbb{Z}^2$	non-abelian group
Natural Geometry	Spherical	Euclidean	Hyperbolic
Its Gaussian Curvature	1	0	-1
Universal Cover	S^2	$\mathbb{R}^2$	Hyperbolic Plane H^2
Isometries of Univ. Cover	$O(3)$	$E(2)$	$PGL(2, \mathbb{R})$

The purpose of the course is to explain all undefined terms in this table and to explain how they are related. A major goal will be to understand the *Gauss-Bonnet Theorem* that unifies geometry and topology. It says that for any geometry on the surface (to be defined more precisely), its Gaussian curvature function K and the topology of the surface are related by the formula

$$\int_S K \, dA = 2\pi\chi(S), \tag{1.2}$$

where $\chi(S)$ is the Euler characteristic that appears in the above table.

2. Compact Spaces

This section is a quick look at material that should have been in Math 4510 but was not covered there by lack of time.

Definition 2.1. Let X be a topological space.

(1) An *open cover* of X is a collection $\{U_\alpha\}_{\alpha \in A}$ of open sets, indexed by some set A, so that $X = \cup_{\alpha \in A} U_\alpha$.
(2) The space X is called *compact* if every open cover of X has a finite sub-cover, in other words, there exists a finite subset $\{\alpha_1, \ldots, \alpha_n\} \subset A$ so that $X = U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}$.

Remark 2.1. Often the definition is applied to a subspace X of a topological space Y, in the subspace topology. Using the definition of subspace topology, it is easy to see that the definition of compactness of X is equivalent to the following:

(1) An open cover of X is a collection $\{U_\alpha\}_{\alpha\in A}$ of open sets in Y, indexed by some set A, so that $X \subset \cup_{\alpha\in A} U_\alpha$.
(2) The space X is called *compact* if every open cover of X has a finite sub-cover, in other words, there exists a finite subset $\{\alpha_1, \dots, \alpha_n\} \subset A$ so that $X \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}$

Example 2.1. It is not easy to give non-trivial examples of compact spaces, but it is easy to find non-compact ones.

(1) A finite space is compact.
(2) Any indiscrete space is compact: there are only two open sets.
(3) An infinite discrete space is not compact. For example, $\mathbb{Z}$ is not compact; the collection $\{n : n \in \mathbb{Z}\}$ if singleton subsets is an open cover with no finite sub-cover.
(4) $\mathbb{R}$ is not compact: the collection $\{(-n, n) : n \in \mathbb{N}\}$ of open sets covers $\mathbb{R}$ but has no finite sub-cover.

A non-trivial example of a compact space is:

Theorem 2.1. *The inverval* $[0, 1]$ *is compact.*

Proof. We give a quick sketch of the proof, which has to depend on the *completeness* of $\mathbb{R}$. Assume, to get a contradiction, that $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ is a cover of $[0, 1]$ with no finite subcover. Divide $[0, 1]$ into two equal subintervals: $[0, 1] = [0, \frac{1}{2}] \cup [\frac{1}{2}, 1]$. Then at least one of these two subintervals cannot be covered by any finite subcollection of $\mathcal{U}$, choose this interval and call it I_1. Repeat the process: divide I_1 into two equal subintervals, choose one, called I_2, that is not covered by any finite subcollection of $\mathcal{U}$, and so on. In this way we get a sequence $I_1 \supset I_2 \supset I_3 \supset \dots$ where the length of $I_i = 2^{-i} \to 0$. Writng $I_i = [a_i, b_i]$ we have

$$a_1 \le a_2 \le a_2 \le \cdots \le b_3 \le b_2 \le b_1.$$

Let $a = \sup\{a_1, a_2, \dots\}$. Since every b_i is an upper bound for $\{a_1, a_2, \dots\}$, we have $a \le b_i$ for all i, in other words, $a \in \cap_i I_i$. In fact, since the length of I_i converges to 0, $\{a\} = \cap_i I_i$. Pick α so that $z \in U_\alpha$. Since $a_i \to a$ and $b_i \to a$, there exists an i_0 so that $I_i \subset U_\alpha$ for all $i \ge i_0$. But this contradicts the fact that no finite subcollection of $\mathcal{U}$ can cover I_i. □

2.1. **Formal Properties of Compactness.** We list some easy but very useful properties of compactness:

(1) Compactness is a topological property: if X and Y are homeomorphic, then X is compact if and only if Y is compact.
(2) The continuous image of a compact space is compact: If $f : X \to Y$ is continuous and surjective, and X is compact, then Y is compact.

Proof: If $\{U_\alpha\}_{\alpha\in A}$ is an open cover of Y, then $\{f^{-1}(U_\alpha)\}$ is an open cover of X, take a finite subcover indexed by $\{\alpha_1,\dots,a_n\}\subset A$, then $U_{\alpha_1},\dots U_{\alpha_n}$ covers Y.

(3) If X is a metric space and $K\subset X$ is compact, then K is closed and bounded.

Proof: To show bounded, take any $x\in X$ and use the open cover $\{B(x,n)\}_{n\in\mathbb{N}}$. For closed, prove $X\backslash K$ is open by choosing any $x\notin K$ and using the cover of $X\setminus\{x\}$ by the open sets $\{y : d(x,y)>\frac{1}{n}\}$, $n\in\mathbb{N}$.

(4) If X is compact and $f:X\to\mathbb{R}$ is continuous, then f has a maximum and a minimum: there exist $x_1,x_2\in X$ so that $f(x_1)\le f(x)\le f(x_2)$ for all $x\in X$.

Proof: By (2) $f(X)\subset\mathbb{R}$ is compact, and by (3) it is bounded, so $\sup f(X)$ and $\inf f(X)$ exist, and $f(X)$ is closed, so $\sup f(X)\in f(X)$ and $\inf f(X)\in f(X)$.

2.2. **Other useful properties of compactness.**

Theorem 2.2. (1) *If X is compact and $F\subset X$ is closed, then F is compact.*
(2) *If X is Hausdorff and $K\subset X$ is compact, then K is closed.*

Proof. For the first part, if $\mathcal{U}=\{U_\alpha\}$ is an cover of F by open sets in X, then $\{X\setminus F)\}\cup\mathcal{U}$ is an open cover of X, so it has a finite sub-cover, and the sets in it other than $X\setminus F$ cover F.

For the second part, to prove that $X\setminus K$ is open, choose $x\notin K$. Since X is Hausdorff, for each $y\in K$ there exists a neighborhood U_y of x and a neighborhood V_y of y so that $U_y\cap V_y=\emptyset$. Then $\{V_y\}_{y\in K}$ is an open cover of K. Since K is compact, there is a finite sub-cover $V_{y_1},\dots V_{y_n}$. Then $U_{y_1}\cap\dots\cap U_{y_n}$ is a finite intersection of open sets, hence open, contains x and is disjoint from K (check this!), so $X\setminus K$ is open. □

Remark 2.2. The assumption that X is Hausdorff is essential to the second part of the Theorem. An example would be the space $(\mathbb{R},Z)=\mathbb{R}$ with the Zariski topology of [12] Example 3.14. If $A\subset\mathbb{R}$ is any subset, then it is compact (check this!), but an infinite subset other than $\mathbb{R}$ is not closed.

Here are some useful corollaries to the Theorem:

Corollary 2.1. *Suppose $f:X\to Y$ is continuous, X is compact, and Y is Hausdorff. Then*

(1) *f is a closed map.*
(2) *If f is also surjective, then f is an identification.*
(3) *If f is also bijective, then f is a homeomorphism.*

Proof. For the first part, if $F \subset X$ is closed, then F is compact, so $f(F) \subset Y$ is compact. Since Y is Hausdortt, $f(F)$ is closed, so f is a closed map. For the third part, recall that a closed continuous bijection is a homeomorphism: f^{-1} exists and if $F \subset X$ is closed, then $(f^{-1})^{-1}(F) = f(F)$ is closed, so f^{-1} is continuous and f is a homeomorphism. For the second part, for f to be an identification Y must have the quotient topology, which means (since f is continuous) that for $A \subset Y$, if $f^{-1}(A)$ is closed then A is closed. Since f is surjective, $f(f^{-1}(A)) = A$, and f is a closed map, A is closed whenever $f^{-1}(A)$ is. □

Example 2.2. Let us look in more detail at the presentation of the projective plane in part (7) of Example 1.1. The map $g : (S^2_+/\sim) \to P^2$ defined there is continuous and bijective. Since S^2_+ is compact, so is its continuous image $S^2_+/\sim$, and P^2 is Hausdorff. So g is a homeomorphism as asserted.

Another useful fact:

Theorem 2.3. *Suppose X and Y are compact topological spaces. Then $X \times Y$ is compact.*

Proof. See pp. 168–170 of [10] for the proof. □

This, together with Theorem 2.1, gives us a large number of examples of compact spaces:

Corollary 2.2. *(The Heine–Borel Theorem) Let $A \subset \mathbb{R}^n$ be closed and bounded. Then A is compact.*

Proof. By Theorem 2.1, $[0, 1]$ is compact, so is any interval $[a, b] \subset \mathbb{R}$, so is any product $[a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$ (using Theorem 2.3 inductively). If A is bounded, then it is a subset of some such product. If, in addition, A is closed, being a closed subset of a compact space it is compact.

□

3. Topological Classification of Surfaces

We start on the proof of Theorems 1.1 and 1.2. We will need some tools that are useful in many other contexts.

3.1. **Triangulations.** We give two definitions of triangulation of a surface. For simplicity, we will consider only compact surfaces. The first definition follows p.16 of [9]:

Definition 3.1. Let S be a compact surface (or a compact surface with boundary). A *triangulation* of S is a decomposition $S = T_1 \cup \cdots \cup T_k$ satisfying the following conditions:

(1) Each T_i is a closed subspace of S and for each i there is a homeomorphism $\phi_i : T_i' \to T_i$ where $T_i' \subset \mathbb{R}^2$ is a triangle. The images under ϕ_i of the vertices, respectively edges, of T_i' are called the vertices, respectively edges, of T_i.
(2) If $i \neq j$, $T_i \cap T_j$ is either empty, or is a common vertex of T_i and T_j, or is a common edge of T_i and T_j.

This definition says that S is decomposed into simple pieces, and puts a strong restriction on how these pieces intersect. Certain types of intersection are allowed, others are not, see Figure 3.1

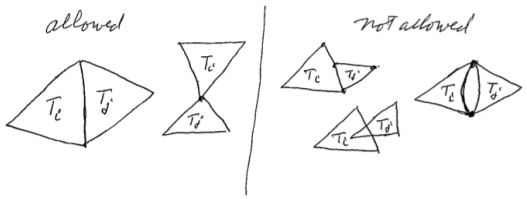


FIGURE 3.1. Intersections of two Triangles

Example 3.1. (1) Take the boundary of a regular tetrahedron in $\mathbb{R}^3$. It is homeomorphic to S^2, and the images under this homeomorphism of the four triangles give a triangulation of S^2. Similarly, the octahedron and the icosahedron give triangulations of S^2, with 8 and 20 triangles respectively.
(2) Figure 3.2 gives two decompositions of the torus T^2 into triangles, the first is *not* a triangulation, the second one is. In the first picture we have drawn in some detail how half of the identification space goes onto the top half of the usual torus of revolution in $\mathbb{R}^3$ and what resulting decomposition into triangles. In the picture in $\mathbb{R}^3$ you can see, for example, how the vertices 3 and 4 are joined by two edges, going around a circle in the torus. In fact, every pair of vertices in this decomposition is joined by two edges.

The condition on intersections in Definition 3.1 says that the vertices determine the triangle (this fails in the first decomposition of Figure 3.2). What this means is that a triangulation is partly a purely combinatorial object. This has been formalized into a very useful concept that we define next. For simplicity we only consider the *finite* situation.

Definition 3.2. (1) A (finite) *simplicial complex* is a finite set K, whose elements are called *vertices*, and a collection of non-empty subsets of K called *simplices*, satisfying the following conditions:
   (a) Every vertex $v \in K$, the set $\{v\} \subset K$ is a simplex.
   (b) If $\sigma \subset K$ is a simplex and $\tau \subset \sigma$, $\tau \neq \emptyset$, then τ is a simplex.

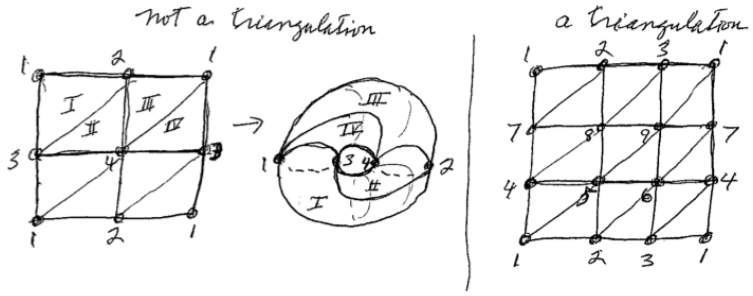


FIGURE 3.2. Decompositions of the Torus into Triangles

Terminology: If $\sigma = \{v_0, \dots v_k\} \subset K$ is a simplex, then $v_0, \dots v_k$ are called *the vertices of* σ and k is called the *dimension* of σ. The *dimension* of K is the maximum dimension of its simplices. We usually use the bracket notation $< v_0, \dots, v_k >$ to indicate that a subset is a simplex.

(2) If K is a simplicial complex, its *geometric realization* $|K|$ is the following space. First, let $\mathbb{R}_K$ denote a real vector space with one basis element for each $v \in K$. If K has m elements, this space is isomorphic to $\mathbb{R}^m$, and in it we can take linear combinations of vertices with real coefficients.
   (a) If $\sigma =< v_0, \dots, v_k >$ is a simplex, let $|\sigma| = \{t_0 v_0 + \dots, t_k v_k : t_0, \dots, t_k \geq 0, t_0 + \dots t_k = 1\} \subset \mathbb{R}_K$.
   (b) Let $|K| = \cup\{|\sigma| : \sigma \text{ a simplex in } K\} \subset \mathbb{R}_K$.

This definition is hard to visualize because we have used a vector space $\mathbb{R}_K$ of dimension the cardinality of K, so we would not be able to easily visualize the definition except when K has only three elements (as we will do next). But this is just a device of going from combinatorics to topological spaces in a well-defined way. Notice that the first part of the definition is purely combinatorial, talking about finite sets and some of their subsets, and that the second part produces a topological space from the data of the first part.

Example 3.2. Suppose $K = \{1, 2, 3\}$ is a set with three elements and we make it into a simplicial complex by requiring that the simplices are

$$< 1, 2 >, < 1, 3 >, < 2, 3 >, < 1 >, < 2 >, < 3 > .$$

It is easy to check that this collection satisfies the definition of a simplcial complex. To form its geometric realization, form a vector space with basis e_1, e_2, e_3. This is the standard $\mathbb{R}^3$, and the three one-dimensional simplices have geometric realization the segments $\overline{e_1 e_2}$, $\overline{e_1 e_3}$ and $\overline{e_2 e_3}$. Thus $|K|$ is the boundary of a triangle, homeomorphic to a circle. See Figure 3.3

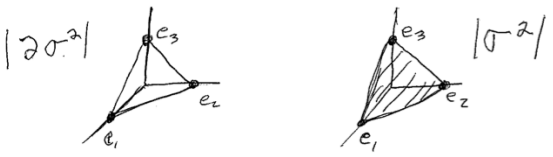


FIGURE 3.3. Geometric Realizations

Example 3.3. Note that in the preceding example the simplices are precisely the proper subsets of K. If we change the definition of the complex to include also the whole set $< 1, 2, 3 >$ as a simplex, then its geometric realization is a triangle $| < 1, 2, 3 > | = \{t_1e_1 + t_2e_2 + t_3e_3 : t_1, t_2, t_3 \geq 0, t_1 + t_2 + t_3 = 1\}$, or, in more usual terms, $\{(x, y, z) : x, y, z \geq 0, x + y + z = 1\}$, see Figure 3.3.

Example 3.4. Extending the last two examples, let's define for $n = 0, 1, 2, \ldots$ the simplicial complex

$$\sigma^n = \{0, 1, \ldots, n\}, \text{ all subsets are simplices,}$$

called the *n-dimensional simplex*, and, for $n = 1, 2, 3, \ldots$, the simplicial complex

$$\partial\sigma^n = \{0, 1, \ldots, n\}, \text{ simplices are the proper subsets,}$$

called the *boundary of the n-simplex*. Then $|\sigma^n|$ is the geometric n-dimensional simplex $\{(t_0, t_1, \ldots, t_n) : t_i \geq 0, t_0 + \cdots + t_n = 1\}$, while $|\partial\sigma^n|$ is its boundary, an $(n-1)$-dimensional space homeomorphic to the unit sphere S^{n-1} in $\mathbb{R}^n$. For $n = 0, 1, 2, 3$, we have that $|\sigma^n|$ is a point, interval, triangle (as in Example 3.3), solid tetrahedron, ... , while for $n = 1, 2, 3$, $|\partial\sigma^n|$ is two points, three segments forming the boundary of a triangle as in Example 3.2, the boundary of a tetrahedron. We see in all these examples that applying Definition 3.2 gives $|\sigma^n|$ and $|\partial\sigma^n|$ as subspaces of $\mathbb{R}^{n+1}$. Looking more closely, we see that they actually lie in the affine-linear subspace (hyperplane) where the sum of the coordinates is 1, which is isomorphic to $\mathbb{R}^n$, so we actually get subspaces of $\mathbb{R}^n$ as they should be. But the realization in $\mathbb{R}^{n+1}$ is more symmetric, therefore, in some sense, more natural.

Remark 3.1. There is an alternative way of defining the geometric realization $|K|$ of a simplicial complex K without using the vector space $\mathbb{R}_K$. We could, for each simplex $\sigma =< v_0, \ldots . v_k >$ define its geometric realization $|\sigma|$ as in (2a) of Definition 3.2. This only requires a vector space of dimension $k + 1$, which can be visualized for k small, say for surfaces. Think of all these geometric realizations σ, for all simplices $\sigma \subset K$ as disjoint, and form the following space

$$(\sqcup_{\sigma \subset K} |\sigma|)/\sim$$

where $\sim$ is the following equivalence relation: if $\tau \subset \sigma$, identify $x \in |\tau|$ with $f(x) \in |\sigma|$ where $f : |\tau| \to |\sigma|$ is the linear map that sends each vertex in $|\tau|$ to the vertex with the same name in $|\sigma|$. There is a map from this space

to $|K| \subset \mathbb{R}_K$ by sending each vertex of each $|\sigma|$ to the basis element with the same name in $\mathbb{R}_K$ and extending by linearity. This gives a continuous bijection

$$(\sqcup_{\sigma \subset K} |\sigma|)/\sim \to |K|$$

hence a homeomorphism. This can be visualized, for the simplicial complex with 4 vertices $0, 1, 2, 3$ and two simplices $< 0, 1, 2 >, < 0, 2, 3 >$ as in Figure 3.4, where the arrows represent the identification maps f defined above.

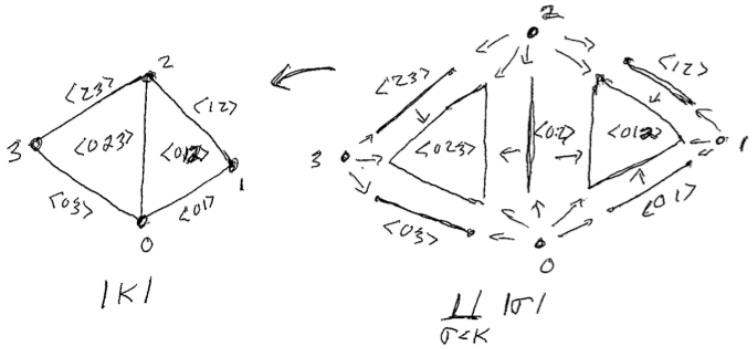


FIGURE 3.4. Assembling the Geometric Realization

Another reason for using the high-dimensional vector space $\mathbb{R}_K$ in the definition of the geometric realization of K (or the alternative definition in the remark above), is to make the following Lemma clear. This Lemma says that the combinatorics completely determines the geometry, in the sense that the intersections of the geometric simplices in $\mathbb{R}_K$ correspond to the intersections of the sets of vertices.

Lemma 3.1. *Let K be a simplicial complex and let σ, τ be simplices in K. Then $|\sigma| \cap |\tau|$ is empty if and only if $\sigma \cap \tau = \emptyset$. Otherwise $\sigma \cap \tau$ is a simplex, and in this case $|\sigma| \cap |\tau| = |\sigma \cap \tau|$.*

Proof. Let $\sigma =< v_0, \ldots, v_k >$ and $\tau =< w_0, \ldots, w_l >$ where $v_0, \ldots, v_k \in K$ are distinct vertices, $w_0, \ldots, w_l \in K$ are distinct vertices, but of course the two sets may intersect. Recall that for a point $x \in \mathbb{R}_K$ we have that

$$x \in |\sigma| \text{ if and only if } x = \Sigma_{i=0}^{k}\, t_i v_i \text{ where } t_i \geq 0 \text{ and } \Sigma t_i = 1,$$

and similarly

$$x \in |\tau| \text{ if and only if } x = \Sigma_{i=0}^{l}\, s_j w_j \text{ where } s_j \geq 0 \text{ and } \Sigma s_j = 1.$$

Suppose first that $\sigma \cap \tau = \emptyset$. Then the set $v_0, \ldots, v_k, w_0, \ldots, w_l$ is linearly independent (being part of a basis), so if $x \in \mathbb{R}_K$ is in $|\sigma| \cap |\tau|$, since it is of both the above forms, linear independence would imply that all the

coefficients t_i, s_j are 0, but this is impossible since they each sum to 1. So $|\sigma| \cap |\tau| = \emptyset$.

Suppose $\sigma \cap \tau \neq \emptyset$. Order the vertices of σ and of τ so that $v_0 = w_0, \ldots, v_m = w_m$ and $\sigma \cap \tau =< v_0, \ldots, v_m >=< w_0, \ldots, w_m >$. If $x \in |\sigma| \cap |\tau|$, then x is a linear combination of both of the displayed forms. By the linear independence of the set $v_0, \ldots, v_m, v_{m+1}, \ldots, v_k, w_{m+1}, \ldots, w_l$ (listing all vertices without repetitions), all coefficients of $v_{m+1}, \ldots, v_k, w_{m+1}, \ldots, w_l$ must vanish, so x must be a linear combination (and with non-negative coefficients adding to one) of the common basis vectors $v_0, \ldots v_m$, in other words, $|\sigma| \cap |\tau| = |\sigma \cap \tau|$ as desired. □

We are know in a position to define "triangulation" of any topological space, not just a surface.

Definition 3.3. Let X be a topological space. A *triangulation of* X means a homeomorphism $\phi : |K| \to X$ for some simplicial complex K.

If S is a compact surface, this definition says that a triangulation of S is a homeomorphism $\phi : |K| \to S$ for some two-dimensional simplicial complex K. This looks like a more precise version of Definition 3.1. It will be instructive to show that both definitions of triangulation of S are equivalent.

3.1.1. *Equivalence of the two definitions.* To see the equivalence, assume that we have a homeomorphism as in Definition 3.3. Then $|K| = |\sigma_1^2| \cup \cdots \cup |\sigma_k^2|$, where $\sigma_1^2, \ldots, \sigma_k^2$ are the two-dimensional simplices of K. Each $|\sigma_i^2|$ is homeomorphic to a triangle in $\mathbb{R}^2$, and if $i \neq j$, then $\sigma_i^2 \cap \sigma_j^2$ is either empty, or a one element set $< v >$, where v is a common vertex, or a two element subset $< v_0, v_1 >$ where v_0, v_1 are common vertices, hence $< v_0, v_1 >$ is a common edge. Then Lemma 3.1 tells us that $|\sigma_i^2| \cap |\sigma_j^2|$ is either empty, or $| < v > |$, or $| < v_0, v_1 > |$. If we write $S = \phi(|\sigma_1^2|) \cup \cdots \cup \phi(|\sigma_k^2|)$, then the sets $T_i = \phi(|\sigma_i^2|)$ give a triangulation in the sense of Definition 3.1.

Conversely, suppose that we have a triangulation $S = T_1, \cup \cdots \cup T_k$ and $\phi : T_i' \to T_i$ as in Definition 3.1. Then there is a well-defined collection of vertices in S. Call a point $v \in S$ a *vertex* if it is the image of a vertex of some T_i' under the corresponding ϕ_i. The second condition of Definition 3.1 implies that if v is a vertex of T_i and is a point of T_j, then it is also a vertex of T_j, so the meaning of vertex is well-defined, independent of the triangle T_i used to define it. The same is true of edges: a pair of vertices determines an edge if they are the vertices of an edge of some T_i, and if they are also vertices of another T_j, they determine the same edge in T_j. Said in another way, the triangles and edges in S are uniquely determined by their vertices, and the inclusions are correct.

Another way of saying this is that, first, if we let K be the collection of vertices in S and call a subset of K a simplex if and only if its elements are

either a single vertex, the vertices of an edge, or the vertices of a triangle, then K is a simplicial complex. To see that there is a homeomorphism from $|K|$ to S, let us consider first the identification space used by Massey on p.19 of [9]. Let us use the symbol $\sqcup$ (instead of $\cup$) to denote *disjoint union*, meaning that we are taking a union of *disjoint* sets. Let

$$T' = T'_1 \sqcup T'_2 \sqcup \cdots \sqcup T'_k$$

and define a map $\Phi : T' \to S$ by letting $\Phi(x) = \phi_i(x)$ if $x \in T'_i$. This is a continuous map since the spaces T'_i are, by assumption, disjoint, and ϕ_i is continuous in T'_i, It is surjective, but of course not a homeomorphism since T' has many connected components. In fact, it is not injective, since each point on an edge in S has two pre-images, and every vertex in S has several pre-images. Next, we define an equivalence relation $\sim$ on T' by declaring $x \sim y$ if and only if $x \in T'_i$, $y \in T'_j$ and $\phi_i(x) = \phi_j(y)$. Then we get a commutative diagram

$$\begin{array}{ccc} T' = T'_1 \sqcup \cdots \sqcup T'_k & \xrightarrow{\Phi} & S \\ \Big\downarrow p & \nearrow \wr & \\ T = T'/\sim & & \end{array} \tag{3.1}$$

The map ϕ defined by this diagram is, by construction, bijective, and, by definition of quotient topology, continuous (see Theorem 4.6 of [12]). Since it is a continuous bijection of a compact space to Hausdorff space, by Corollary 2.1 it is a homeomorphism.

The informal way of stating what we have just proved is that S is obtained from the disjoint union of the triangles T'_i by gluing them along edges.

Finally we bring in the simplicial complex K. Suppose $< v_0, v_1, v_2 >$ is a simplex in K. then v_0, v_1, v_2 are the vertices of a unique triangle T_i is S, which is the image of a triangle $T'_i \subset \mathbb{R}^2$. Let $v'_o, v'_1, v'_2 >$ be the corresponding vertices of T'_i. Define a map $\psi_i : | < v_0, v_1, v_2 > | \to T_i$ by sending each vertex to the corresponding one and then extending by linearity, namely:

$$\psi_i(t_0 v_0 + t_1 v_1 + t_2 v_2) = t_0 v'_0 + t_1 v'_1 + t_2 v'_2,$$

where, of course, $t_0, t_1, t_2 \geq 0$ and $t_0 + t_1 + t_2 = 1$. Observe that v_0, v_1, v_2 are basis elements in the space $\mathbb{R}_K$ where $|K|$ lies, and the linear combinations in the left hand side are in the vector space $\mathbb{R}_K$. The linear combination in the right hand side is in a Euclidean plane $\mathbb{R}^2$ where T'_i lies. Since we are regarding the T'_i as disjoint, we could think of a different Euclidean plane for each i.

We can define a map $F : |K| \to S$ by $F(x) = \phi_i \circ \psi_i(x)$ if $x \in |\sigma^2_i|$, where σ^2_i is the simplex of K formed by the vertices of $T_i \subset S$. It is now easy to

check that this map is well-defined, is bijective, and is continuous on each $|\sigma_i^2|$, hence it is continuous. Again, since it is a continuous bijection from a compact space to a Hausdorff space, it is a homeomorphism. Hence the two definitions of triangulation agree.

We could summarize the construction of the map F by the following diagram:

$$\begin{array}{ccccc} |\sigma_1^2| \sqcup \cdots \sqcup |\sigma_k^2| & \xrightarrow{\Psi = \sqcup\psi_i} & T' = T_1' \sqcup \cdots \sqcup T_k' & \xrightarrow{\Phi} & S \\ \downarrow{\scriptstyle p_K} & & \downarrow{\scriptstyle p_{T'}} & \nearrow{\scriptstyle \Phi} & \\ |K| & \xrightarrow{\psi} & T'/\sim & & \end{array} \tag{3.2}$$

The map F we have just defined is the composition $\phi \circ \psi$, where ψ is defined by this diagram: the upper right hand map $\Psi = \sqcup\psi_i$ is the map whose restriction to each $|\sigma_i^2|$ is $\psi_i : |\sigma_k^2| \to T_i'$. The vertical maps p_K, $p_{T'}$ are identification maps, and the map Ψ preserves identifications and descends to a map as shown.

Remark 3.2. It is worth looking in more detail at the meaning of triangulation. Looking first one dimension lower, if we consider the example of the circle S^1 divided into two arcs by two vertices v_1, v_2, then the abstract simplicial complex defined by the vertices is σ^1, thus its geometric realization is an interval. We can defined maps $|\sigma^1| \to S^1$ by mapping the interval to either arc, but neither of these maps is surjective. To triangulate S^1 we need at least three vertices, and the complex $\partial\sigma^2$ gives of course a triangulation with three vertices.

In checking if a decomposition of a surface is a triangulation, it helps to keep in mind that a circle requires at least three vertices. If we look at the decomposition of T^2 in Figure 3.2 we immediately see several topological circles with only two vertices, so this tells us immediately that this is not a triangulation.

Another remark: if we examine the abstract simplicial complex formed by the vertices of the decomposition of the torus in Figure 3.2, we see that it is actually $\partial\sigma^3$, namely all proper subsets of a 4-element set (except that each subset of cardinality 1 or 2 appears twice). The geometric realization of $\partial\sigma^3$ is topologically S^2 which is topologically quite different from T^2.

3.2. **Triangulability of Surfaces.** Not every topological space can be triangulated. For example, since we have assumed that our simplicial complexes K are finite, their geometric realizations $|K|$ are compact (being a finite union of the compact subspaces $|\sigma|$). There are more restrictions, for

example any geometric realization $|K|$ is locally connected. But the following theorem is true, although tedious to prove, and we will not prove it here:

Theorem 3.1. *Let S be a compact surface. Then S can be triangulated.*

We will assume this theorem in deriving the topological classification of surfaces. Since we are not going to prove it, we should, strictly speaking, add the hypothesis "triangulable" to any theorem that we prove assuming this one.

Example 3.5. In addition to the triangulations we have already shown for S^2 and T^2, there is a well-known triangulation of P^2, with six vertices, illustrated in Figure 3.5. Note that there are many topological circles in this triangulation, each divided into 3 segments.

Question: If $p : S^2 \to P^2$ is the quotient map, what is the triangulation $p^{-1}(K)$ of S^2?

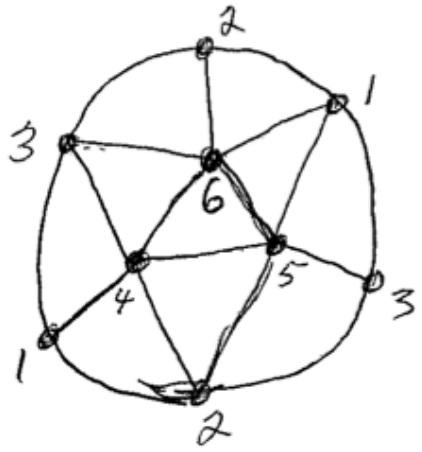


FIGURE 3.5. A Triangulation of P^2

3.3. **The Euler Characteristic.** If K is a simplicial complex, let f_i be the number of simplices of dimension i. The *Euler characteristic* of K is defined to be the number $\chi(K) = f_0 - f_1 + f_2 - \cdots + (-1)^d f_d$, where d is the dimension of K. Thus $\chi(K)$ is the alternating sum of the number of "faces" of K of each dimension.

If K is two-dimensional, the numbers f_i are usually denoted $V = f_0$, $E = f_1$ and $F = f_2$, for the number of vertices, edges and faces respectively. This number turns out to have enormous geometric and topological significance.

It turns out to be a topological invariant of $|K|$. In particular, if S is a surface and K is a triangulation of S, then $\chi(K)$ turns out to depend only on S. We will give an indirect proof of this fact later on. Even though we cannot totally justify that the following definition is correct, we record the terminology in the following definition. We will not used the unproved fact, but in computations we will often say $\chi(S)$ rather than $\chi(K)$.

Definition 3.4. If S is a compact surface, and K is a two dimensional simplicial complex that triangulates S, the *Euler characteristic* of K, denoted $\chi(K)$, is the number $\chi(K) = V-E+F$. (Fact: all triangulations K of S give the same value for $\chi(K)$, and $\chi(S)$ is defined to be $\chi(K)$ for a triangulation K of S).

Example 3.6.
(1) If $S = S^2$ triangulated as $\partial\sigma^3$, the boundary of a tetrahedron, then $V = 4$, $E = 6$ and $F = 2$, so $\chi(\partial\sigma^3) = 2 = \chi(S^2)$. Actually, Euler proved that for any triangulation of S^2 as the boundary of a convex polyhedron, $\chi = 2$, hence the name Euler Characteristic. Check that for the octahedral and icosahedral triangulations of S^2 we also get $\chi = 2$.
(2) Using the triangulation of Figure 3.2, we see that $\chi(T^2) = 2$.
(3) Using the triangulation of Figure 3.5 we see that $\chi(P^2) = 1$. Is there another way of seeing this by using the identification map $p : S^2 \to P^2$?

For a triangulation K of a surface the numbers V, E and F are not independent. We have the following fact:

Lemma 3.2. *If K is a triangulation of a surface, every edge of K is contained in exactly two triangles*

Proof. Let $e = |\sigma|$ be an edge of K for some one-simplex σ of K, and let $x \in e$ be an interior point. We would like to prove that the only way that x can have a neighborhood in $|K|$ homeomorphic to a disk is that e be contained in exactly two triangles. First, we know that an interval cannot be homeomorphic to a disk, so e must be contained in at least one triangle. Then looking at Figure 3.6, it's reasonable that if e is contained in any number of triangles other than two, it cannot have a neighborhood homeomorphic to a disk. We will be able to give a rigorous proof later. □

FIGURE 3.6. Each Edge Contained in Two Triangles

Corollary 3.1. *If K is a triangulation of a surface, then $2E = 3F$. Thus $\chi = V - \frac{1}{2}F = V - \frac{1}{3}E$.*

Proof. Since every triangle has 3 edges and each edge is contained in 2 triangles, $3F$ counts the edges exactly twice, so $3F = 2E$. □

3.4. **Proof of Classification.** The first step in the proof of classification relies on the existence of triangulations. We state the theorem using the model of a disk with its boundary divided into arcs. We could equally well use a the model of a regular polygon and its boundary naturally divided into edges. We will move freely from one model to the other.

Theorem 3.2. *Let S be a compact connected surface. Then S is homeomorphic to the identification space of a disk D whose boundary ∂D is divided into an even number $n = 2m$ of arcs, and these arcs are identified in pairs.*

Proof. We follow Section 7 of the first chapter of [9]. Let K be a triangulation of S, thus there is a homeomorphism $\phi : |K| \to S$. Write $T_1', \ldots, T_k'$ for the geometric realization of the two-dimensional simplices of K and $T_i = \phi(T_i') \subset S$. Choose the ordering of the triangles and an ordered collection $\{e_1, \ldots e_{k-1}\}$ of edges so that T_i has the edge e_i in common with one of $T_1, \ldots T_{i-1}$. This can be done as follows. Choose any triangle, call it T_1, choose a second triangle, call it T_2 that has an edge, call it e_1, in common with T_1. Then choose a triangle, call it T_3, that has an edge, call it e_2, in common with $T_1 \cup T_2$, continue in this way.

Since S is *connected*, we must get all triangles this way. (*Proof*: Otherwise we would stop at some $l < k$. Then $A = T_1 \cup \cdots \cup T_l$ and $B = T_{l+1} \cup \cdots \cup T_k$ would be disjoint closed sets with $S = A \cup B$, contradicting connectedness.)

Define a triangulated space D by

$$D = (T_1' \sqcup \cdots \sqcup T_k')/\sim \tag{3.3}$$

where $x \sim y$ if, for one of the edges e_j just chosen, $\Phi(x), \Phi(y) \in e_j$ and $\Phi(x) = \Phi(y)$, where Φ is as in Diagram 3.1. Referring back to Diagram 3.1, we see that the only difference is that here we have not made all the identifications that the map p makes, but only the ones over the chosen edges $\{e_1, \ldots, e_{k-1}\}$. So Φ descends to a map, let's call it $\phi_1 : D \to S$ which is one to one over all of D except on ∂D which is the pre-image of the edges in S other than the chosen $\{e_1, \ldots, e_{k-1}\}$, on which it is two-to-one (because is edge is contained in two triangles). This means that ∂D is divided into pairs of edges and the interior of each edge in each pair maps homeomorphically to the corresponding edge in D. Thus S is homeomorphic to $D/\sim$ where $x \sim y$ if $\phi_1(x) = \phi_1(y)$. Finally, we need the following Lemma:

Lemma 3.3. *Let D_1, D_2 be disks, let $I_1 \subset \partial D_1$ and $I_2 \subset \partial D_2$ be arcs in the boundary (meaning subspaces of the boundary homeomorphic to the*

unit interval I), let $h : I_1 \to I_2$ be a homeomorphism. Then the space $(D_1 \sqcup D_2)/(x \sim h(x))$ is homeomorphic to a disk.

Proof. Clearly each there is a homeomorphism of each D_i with a half-disk $D_i' = \{x^2 + y^2 \leq 1, y \geq 0 \text{ or } y \leq 0\}$ which takes I_1 to $I_i' = [-1, 1] \times \{0\}$, and it is clear that gluing two half disks by a homeomorphism of this part of the boundary produces a disk, see Figure 3.7 □

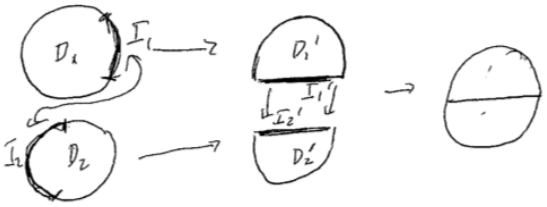


FIGURE 3.7. Gluing Disks Along Arcs in Boundary

To finish the proof of the theorem it only remains to prove that D is homeomorphic to a disk, but this follows by an easy induction from the Lemma and the construction of D: T_1', T_2' are disks with edges (= arcs) e_1', e_1'' in their boundaries identified by a homeomorphism to get $(T_1' \sqcup T_2')/\sim$, so this space is homeomorphic to a disk. Moreover it has an edge e_2' that is identified with an edge e_2'' of T_3' to get the space $(T_1' \sqcup T_2' \sqcup T_3') \sim$, which is then homeomorphic to a disk, and so on until we get to D. □

Remark 3.3. The phrase "arcs identified in pairs" is rather vague. What this means is that the collection of arcs is divided into pairs, the two arcs in each pair are identified by a homeomorphism which is monotone in the sense of the arrows used in the diagrams explained in (9) of Example 1.1. These homeomorphisms are usually not specified. It is a fact that different choices lead to homeomorphic results, but we don't go into any more detail.

Remark 3.4. The converse of this Theorem is also true. The space obtained from a disk by dividing its boundary into an even number of arcs and identifying these arcs in pairs is a compact, connected surface. The verification prodeeds by the same reasoning explained in (9) and (10) of Example 1.1 and illustrated in Figure 1.3. If $x \in \partial D$ is interior to an arc, then a neighborhood of x consists of two half-disks identified into a disk as in the first part of Figure 1.3. If x is a vertex, then a neighborhood in the quotient space is obtained by identifying several sectors of disks into a single disk. The only difference with the second half of Figure 1.3 is that the quotient space may have more than one vertex, as in the example below. A neighborhood of a vertex in the quotient space will still be obtained by identifying sectors along common boundary edges, each edge in two sectors, and the edges being cyclically arranged (see the next example for an illustration of what "cyclically" means), so this identification space will still be a disk.

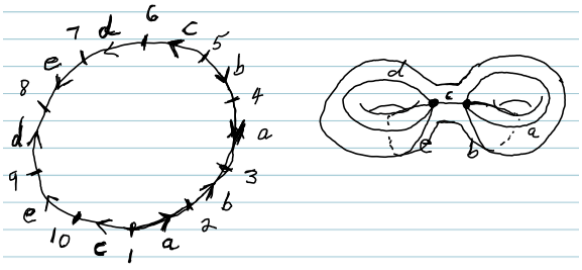


FIGURE 3.8. The Surface of Genus Two as $aba^{-1}b^{-1}cded^{-1}e^{-1}c^{-1}$.

Example 3.7. To see how to find vertices and their neighborhoods, let's look at the example shown if Figure 3.8. If we use the notation explained in (9) of Example 1.1, reading the left hand figure counterclockwise from the bottom, this identification space is described by the symbol $aba^{-1}b^{-1}cded^{-1}e^{-1}c^{-1}$. To see how many vertices there are, start anywhere, say at 1, and see what's equivalent to it: being the tail of c, it is equivalent to 5, as the tail of b it is equivalent to 2, as head of a it is equivalent to 3, as the head of b it is equivalent to 4, as the tail of a it is equivalent to 1 and we are back where we started: $1 \sim 5 \sim 2 \sim 3 \sim 4 \sim 1$, thus we have come in a full cycle and the set $\{1, 5, 2, 3, 4\}$ forms one equivalence class. Similarly we see $6 \sim 10 \sim 7 \sim 8 \sim 9 \sim 10$, again we complete a cycle and $\{6, 10, 7, 8, 9\}$ is an equivalence class. Thus there are two vertices in the identification space. This is a surface of genus two, which can be pictured as in the right half of Figure 3.8. It is a connected sum of two tori, the image of ∂D has two vertices, as pictured. It is a longer presentation of this surface than the standard one (8) of Example 1.1

3.4.1. *Inductive Part of the Proof.* We now quickly sketch a nice inductive argument for going from Theorem 3.2 to the Classification Theorem 1.2. The proof is in the paper [2] by Burgess. We refer to this paper for details.

Following Burgess, let D be a disk and for each even n, divide ∂D into n arcs and let $D(n)$ be the collection of all possible pairs of oriented arcs in this division of ∂D. Let $M(n)$ the class of surfaces obtained from D by so identifying ∂D according to the elements of $D(n)$.

The argument will be by induction on n. To begin the induction:

Lemma 3.4. *If S is a surface of type $M(2)$, then S is homeomorphic to either S^2 or P^2.*

Proof. The proof is clear: the only identifications of type $D(2)$ are aa which gives P^2 and aa^{-1} that gives S^2. □

Then, to be able to carry on the induction, let $A = \{\frac{1}{2} \leq x^2 + y^2 \leq 1\})$ be an annulus, let $C_1 = \{x^2 + y^2 = 1\}$ be one of its boundary components,

divided into n arcs, and let $A(n)$ denote the set of all possible pairs of oriented arcs in this division. Let $B(n)$ be the class of surfaces with boundary obtained from A by identifying the boundary component C_1 according to the elements of $A(n)$. These surfaces have as boundary the other component C_2 of the boundary of A.

Lemma 3.5. *Every surface of type $B(n)$ is homeomorphic to $M \setminus \Delta^0$, where M is a surface of type $M(n)$ and Δ is a disk embedded in M.*

Proof. The proof is clear, since we can fill in the boundary of C_2 of $A/\sim$ with a disk, thus obtaining $D/\sim$. □

The proof then proceeds: Assume that every surface of type $M(m)$ for even $m \leq n-2$ is as in Theorem 1.2, prove that this also holds for every surface of type $M(n)$. This is done by cases, which we now list and illustrate with the example $aba^{-1}b^{-1}cdd^{-1}efe^{-1}fc^{-1}$ of Figure 3.9.

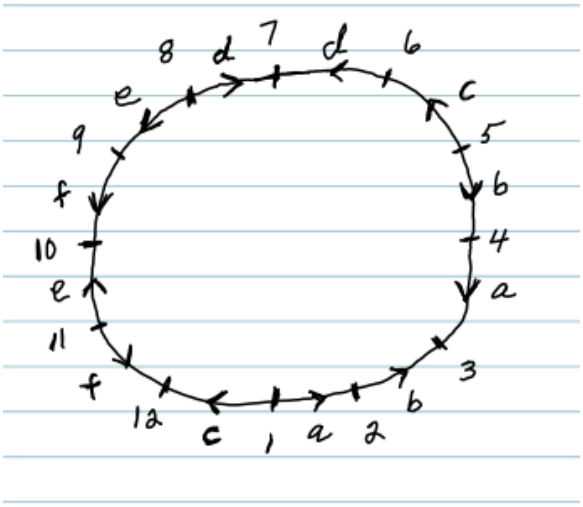


FIGURE 3.9. $aba^{-1}b^{-1}cdd^{-1}efe^{-1}fc^{-1}$

Pick a surface M of type $M(n)$. There are 4 cases to be considered:

(1) There is a *twisted pair*. This means a pair of identifications $\ldots a \ldots a \ldots$ in the same direction, as the pair $\ldots f \ldots f \ldots$ in Figure 3.9. Then making just this identification gives a Möbius band (homeomorphic to $P^2 \setminus \Delta_1^0$) and a neighborhood of its boundary is an annulus $A(n-2)$, hence performing the remaining identifications gives an $N \setminus \Delta_2^0$ for some $N \in M(n-2)$. From this we see that $M = N \# P^2$ for $N \in M(n-2)$, so by induction M is as desired.

In Figure 3.9 we would get $aba^{-1}b^{-1}cdd^{-1}efe^{-1}fc^{-1}$ is a connected sum $P^2 \# N$ where $P^2 = ff$ and $N = aba^{-1}b^{-1}cdd^{-1}ee$. Note

here that the e^{-1} changed to e because of the sense in which we travel along the boundary of the Möbius band. So we get a twisted pair ee where the edges are consecutive, seeing that such an identification gives a Móbius band requires separate argument (see Figure 1.2 of [2])

(2) There are two *separated non-twisted pairs*. This means two pairs , each in opposite directions, each pair separating the elements of the other, as $aba^{-1}b^{-1}$ in Figure 3.9. Then if $n = 4$ we have a T^2, if $n > 4$ we have $T^2\#N$ for some $n \in M(n-4)$.

In Figure 3.9 we would get $T^2\#N$ where $N \in M(n-4)$ is given by $cdd^{-1}efe^{-1}fc^{-1}$ (which is easily seen to be the same as $efe^{-1}f$, a Klein bottle).

(3) There is a *non-twisted pair of adjacent edges*, as dd^{-1} in Figure 3.9. Then it is easy to see that these edges "cancel" and $M = N$ where $N \in M(n-2)$. In Figure 3.9, we get $aba^{-1}b^{-1}cefe^{-1}fc^{-1}$.

(4) There is a *non-twisted, non-adjacent, non-separating pair*, such as $\dots c \dots c^{-1} \dots$ in Figure 3.9. Then performing just the identification of c with c^{-1} yields an annulus with each boundary component divided into arcs, the identifications remain within each component, so $M = N_1\#N_2$ where N_1, N_2 are of type $M(n_1)$, $M(n_2)$ for $n_1 + n_2 = n - 2$. In Figure 3.9 we get $M = T^2\#K$ where $T^2 = aba^{-1}b^{-1}$ and $K = efe^{-1}f$ as before.

A good illustration of this case is Figure 3.8, where $\dots c \dots c^{-1}$ is such a pair, and the right hand half of the picture illustrates the connected sum.

Since this exhausts all possibilities for pairs of edges, the inductive step is complete. This concludes the sketch of the proof. See [2] for more details.

We see that the example of Figure 3.9 can be reduced in several ways, either as $P^2\#P^2\#T^2$ using ff, ee, $aba^{-1}b^{-1}$ by following, say, the sequence of moves $aba^{-1}b^{-1}cdd^{-1}efe^{-1}fc^{-1} \to ff\#aba^{-1}b^{-1}cdd^{-1}eec^{-1} \to ff\#ee\#aba^{-1}b^{-1}cdd^{-1}c^{-1} \to ff\#ee\#aba^{-1}b^{-1}$, where the first two moved are (1), and the third is two applications of (3). Or we could start with (2) and follow $aba^{-1}b^{-1}cdd^{-1}efe^{-1}fc^{-1} \to aba^{-1}b^{-1}\#cdd^{-1}efe^{-1}fc^{-1} \to aba^{-1}b^{-1}\#efe^{-1}f = T^2\#K$, where the second move is two applications of (3), remembering that we go cyclically around the circle to cancel $c \dots c^{-1}$. Note that these two presentations of the surface are not contradictory since $K = P^2\#P^2$ (see HW 1).

3.5. **Comments on Classification and Euler Characteristic.** This will be an informal discussion, since we will assume that the Euler characteristic is independent of the triangulation used to define it. We will also assume that it is a topological invariant. We will not use either of these facts in a serious way, we include them to give a more complete view of the classification theorem just proved.

First we remark that the Euler characteristic of a surface can be computed from any presentation as in Theorem 3.2. If we look at the proof of that theorem, we see that to get the disk D from the triangulation K by assembling the triangles T_i along some of their edges. In each process of the construction we replace two triangles by a single one and erase their common edge. Thus $F - E$ is unchanged at each stage. The remaining edges are on the boundary, and we may simplify the picture by joining two successive edges along a vertex Each such move reduces both E and V by one, so it leaves $V - E$ unchanged. Both of these operations leave $V - E_F$ unchanged, where we of course have a new meaning of these names: a face is no longer a triangle, and an edge is no longer determined by its endpoints. But in any case $\chi = V - E + F$. In the situation of Theorem 3.2, $F = 1$, $E = m = \frac{n}{2}$ (note that the edges are counted in the identification space, thus E is precisely half the number of arcs in ∂D) and V has to be determined by the identifications as explained in Remark 3.4 and in Example 3.7. In particular, we see that in Example 3.7 (see Figure 3.8), $F = 1$, $E = 5$ and $V = 2$, so $\chi = -2$, which is the same answer that we would get from the presentation in (8) of Example 1.1, see Figure 1.1, where $F = 1$, $E = 4$ and $V = 1$. More generally, from the presentation of Σ_g, $g \geq 1$, given in Equation 1.1, we see that

$$\chi(\Sigma_g) = 2 - 2g, \tag{3.4}$$

since $F = 1$, $E = 2g$ and $V = 1$.

If we look at the example of Figure 3.9 we see that $F = 1$, $E = 6$ and we can check, as we did in Example 3.7 that $V = 3$. Thus $\chi = -2$ as in Example 3.7. But the surfaces are not homeomorphic, they are $T^2\#T^2$ and $T^2\#K$.

The difference between $T^2\#T^2$ and $T^2\#K$ is that the first is orientable while the second one is not. Again, this is an informal discussion since we do not have a good definition of orientability. Definition 1.2 allows us to see that $T^2\#K$ is not orientable since K contains Möbius bands, but we don't have a good way of showing that $T^2\#T^2$ does not. The theorem is that *orientability* and *Euler characteristic* determine the topological classification of surfaces. In fact it can be shown that the classification is:

$$\begin{array}{lll} \text{Orientable: } S^2 \text{ or } \Sigma_g = T^2\#\ldots T^2 \ (g \text{ times}) & \chi = 2 - 2g \\ \text{Non-Orientable: } P^2\#\ldots\#P^2 \ (n \text{ summands}) & \chi = 2 - n. \end{array} \tag{3.5}$$

There are other ways of stating the result for non-orientable surfaces, using the fact that $P^2\#P^2\#P^2 = K\#P^2 = T^2\#P^2$, so there are several equivalent ways of representing a non-orientable surface as a connected sum, see [2] and sections 7 and 8 of the first chapter of [9] for more details. We will

deal mostly with orientable surfaces, and use Theorem 1.1 as our list of orientable surfaces. We have proved the first part of that theorem. After we study the fundamental group we will be able to prove the second part.

4. The Fundamental Group

Let X be a connected and locally path connected space. We will assign to X a group that is a topological invariant of X, and which, in fact, will give a lot more topological information on possible continuous maps between spaces. A good reference is Chapter 2 of [9], another good reference is Chapter 1 of [5].

Convention: *In this section we will assume that all topological spaces are connected and locally path connected (thus, in particular, path connected).*

The reason for this convention will be clear after reading this section. For now, let's say that we want path connected spaces because the constructions will involve sending paths from one point to another (see, for example, Theorem 4.3). On the other hand spaces that are path connected but not locally path connected, such as in Figure 4.1 present difficulties: Looked at from afar it looks like a circle, so its fundamental group should be $\mathbb{Z}$ (see Theorem 4.5). On the other hand there are no loops starting at x_0 and going all the way around, so all loops are contractible, so its fundamental group should be trivial. To avoid deciding what is the proper interpretation of this example, we assume local path connectedness.

Figure 4.1. Closing the $\sin(1/x)$ Curve.

4.1. **Homotopy.** The unifying concept is called *homotopy*, which formalizes the notion of deformation: If X, Y are topological spaces and $f, g : X \to Y$ are continuous maps, the following definition makes precise what it means to deform f to g. We will also need a more refined concept: If there happens to be a subset $A \subset X$ on which $f = g$, we may want to keep this equality throughout the deformation.

Definition 4.1. Let X and Y be topological spaces and $f, g : X \to Y$ be continuous maps. We say that f and g are *homotopic*, and write $f \sim g$, if there exists a continuous map $F : X \times I \to Y$ such that

$$F(x,0) = f(x) \text{ and } F(x,1) = g(x) \text{ for all } x \in X. \tag{4.1}$$

The map F is called *a homotopy between f and g*. If, in addition, $A \subset X$ and $f(x) = g(x)$ for all $x \in A$, we say that f and g are *homotopic relative to A*, and write $f \sim g$ rel A, if there exists a homotopy $F : X \times I \to Y$ between f and g that, in addition, satisfies

$$F(x,s) = f(x)(= g(x)) \text{ for all } x \in X \text{ and for all } s \in I. \tag{4.2}$$

(in words, points in A do not move under the homotopy).

We will first apply this definition *paths*. Recall the following terminology:

Definition 4.2. Let X be a connected and path connected topological space, and let $x_0, x_1 \in X$. A *path in X from x_0 to x_1* means a continuous map $\alpha : I \to X$ such that $\alpha(0) = x_0$ and $\alpha(1) = x_1$.

Then $\alpha \sim \alpha'$ rel $\{0,1\}$ means: there exists a continuous map $F : I \times I \to X$ with

$$F(t,0) = \alpha(t),\ F(t,1) = \alpha'(t);\ F(0,s) = x_0,\ F(1,s) = x_1 \text{ for all } s,t. \tag{4.3}$$

We often say $\alpha \sim \beta$ *relative to the endpoints* or simply *rel endpoints*. We can picture this situation as in Figure 4.2.

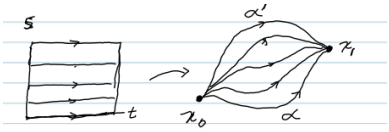


FIGURE 4.2. Homotopic Paths

4.2. **Composition of Paths.** Recall also that we had the concatenation (or composition) of paths (see Definition 5.4 of [12]): If $\alpha, \beta : I \to X$ and $\alpha(1) = \beta(0)$, then $\alpha \cdot \beta$ is defined by

$$\alpha \cdot \beta(t) = \begin{cases} \alpha(2t) & \text{if} \quad 0 \le t \le \frac{1}{2}, \\ \beta(2t-1) & \text{if} \quad \frac{1}{2} \le t \le 1. \end{cases} \tag{4.4}$$

If we have three paths α, β, γ that can be concatenated, meaning that $\alpha(1) = \beta(0)$ and $\beta(1) = \gamma(0)$, then we have two paths $(\alpha \cdot \beta) \cdot \gamma$ and $\alpha \cdot (\beta \cdot \gamma)$, which are different, even though they are reparametrizations of each other. In other words, this operation is not associative. We have some candidates for "units": if $x \in X$, let $\epsilon_x : I \to X$ denote the constant path at x:

$$\epsilon_x(t) = x \text{ for all } t \in I. \tag{4.5}$$

Then $\epsilon_{\alpha(0)} \cdot \alpha \neq \alpha$, even though they travel the same path, one of them constant for half the time and then speeding up, following the other at twice the speed. The same goes of $\alpha \cdot \epsilon_{\alpha(1)} \neq \alpha$, this time the first path stays put for the second half of the interval. Finally we use the notation

$$\alpha^{-1}(t) = \alpha(1-t) \text{ for } t \in I. \tag{4.6}$$

We would like to think of this as an inverse, but $\alpha \cdot \alpha^{-1} \neq \epsilon_{\alpha(0)}$ and $\alpha^{-1} \cdot \alpha \neq \epsilon_{\alpha(1)}$.

Theorem 4.1. *Let X be as above, let $x_0, x_1, x_2, x_3 \in X$ and let $\alpha, \beta, \gamma : I \to X$ be paths from x_0 to x_1, x_1 to x_2 and x_2 to x_3 respectively, and let ϵ_x and α^{-1} be as above. Also, let α', β' be paths from x_0 to x_1 and x_1 to x_2 respectively, such that $\alpha \sim \alpha'$ and $\beta \sim \beta'$ both relative to endpoints. Then:*

1. *$(\alpha \cdot \beta) \cdot \gamma \sim \alpha \cdot (\beta \cdot \gamma)$ rel endpoints.*
2. *$\epsilon_{x_0} \cdot \alpha \sim \alpha$ and $\alpha \cdot \epsilon_{x_1}$ both rel endpoints.*
3. *$\alpha \cdot \alpha^{-1} \sim \epsilon_{x_0}$ and $\alpha^{-1} \cdot \alpha \sim \epsilon_{x_1}$ both rel endpoints.*
4. *$\alpha \cdot \beta \sim \alpha' \cdot \beta'$ rel endpoints.*
5. *$\alpha^{-1} \sim (\alpha')^{-1}$ rel endpoints.*

Proof. For detailed proofs see §2 of Chapter 2 of [9]. We will quickly sketch the proof. For the first 3 statements, we have to find a map $F : I \times I \to X$ so that $F(t,0)$ and $F(t,1)$ are the maps on each side of the $\sim$ sign, and $F(0,s), F(1,s)$ are the appropriate constant maps. The maps on both sides of the $\sim$ sign have images that are either the same (in (1) and (2)) or one contained in the other (as in (3)). In each figure we show a division of $I \times I$ and for each s, we indicate the map on the interval $I \times \{s\}$ that interpolates between the given maps at $s = 0$ and $s = 1$. Each division of the interval is mapped to the unit interval by the unique linear map that maps endpoints to endpoints.

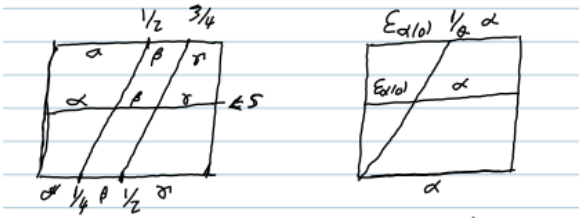


FIGURE 4.3. Homotopy Associativity and Units

For example, in Figure 4.3 we use different parametrizations of the map obtained from α on the first portion of the interval, then β, then γ. Note that in this homotopy it is only the relevant endpoints $x_0 = \alpha(0)$ and $x_3 = \gamma(1)$ that stay fixed, all others move. Figure 4.3 also shows the homotopy for $\epsilon_{x_0} \cdot \alpha \sim \alpha$, the one for $\alpha \sim \alpha \cdot \epsilon_{x_1}$ is symmetric to this one.

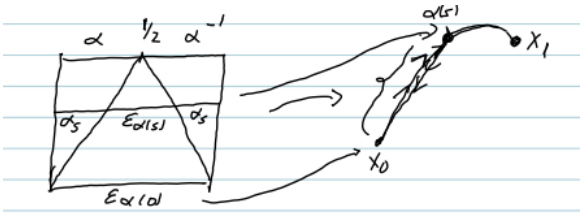


FIGURE 4.4. Homotopy Inverse

Particularly interesting is the homotopy for $\alpha \cdot \alpha^{-1} \sim \epsilon_{x_0}$ shown in Figure 4.4. Here α_s denotes the path $\alpha|_{[0,s]}$. This homotopy can be pictured as traveling along α until reaching $\alpha(s)$, then statying at $\alpha(s)$ for the indicated time, then traveling back to $\alpha(0)$ by way of α^{-1}. This interpolates between being put at $\alpha(0)$ for $s = 0$ and the path $\alpha \cdot \alpha^{-1}$ for $s = 1$. The homotopy for $\alpha \cdot \alpha^{-1}$ is similar.

For (4) and (5), if F is a homotopy between α and α' and G is a homotopy between β and β', (both relative endpoints) then

$$H(t,s) = \begin{cases} F(2t,s) & \text{if} \quad 0 \le t \le \frac{1}{2}, \\ G(2t-1,s) & \text{if} \quad \frac{1}{2} \le t \le 1. \end{cases}$$

is a homotopy between $\alpha \cdot \beta$ and $\alpha' \cdot \beta'$, and $F(1-t,s)$ is a homotopy between α^{-1} and $(\alpha')^{-1}$, both also relative to the endpoints. □

The last two statements of this theorem justify the following Definition:

Definition 4.3. If $\alpha : I \to X$ is a path, write $[\alpha] = \{\alpha' : \alpha \sim \alpha' \text{ rel endpoints}\}$, called the *homotopy class of* α. If $\alpha, \beta : I \to X$ are paths with $\alpha(1) = \beta(0)$ define two operations on homotopy classes:

(1) $[\alpha] \cdot [\beta]$ is defined to be $[\alpha \cdot \beta]$,
(2) $[\alpha]^{-1}$ is defined to be $[\alpha^{-1}]$.

Corollary 4.1. *With the same notation and assumptions as in Theorem 4.1, the operations of Definition 4.3 satisfy:*

(1) *Associativity:* $[\alpha] \cdot ([\beta] \cdot [\gamma]) = ([\alpha] \cdot [\beta]) \cdot [\gamma]$.
(2) *Existence of units:* $[\epsilon_{x_0}] \cdot [\alpha] = [\alpha] \cdot [\epsilon_{x_1}] = [\alpha]$.
(3) *Existence of inverses:* $[\alpha] \cdot [\alpha]^{-1} = [\epsilon_{x_0}]$ *and* $[\alpha]^{-1} \cdot [\alpha] = [\epsilon_{x_1}]$.

4.2.1. *Path Operations and Continuous Maps.* Suppose that X, Y, Z are topological spaces (let's keep them always connected and locally path connected), suppose $f, f_1, f_2 : X \to Y$ and $g : Y \to Z$ are continuous maps. We want to see how the path operations and homotopies behave with respect to maps.

Theorem 4.2. (1) *If* $\alpha, \alpha' : I \to X$ *and* $\alpha \sim \alpha'$ *rel endpoints, then* $f \circ \alpha \sim f \circ \alpha'$ *rel endpoints.*

(2) *If* $\alpha, \beta : I \to X$ *and* $\alpha(1) = \beta(0)$*, then* $f \circ (\alpha \cdot \beta) = (f \circ \alpha) \cdot (f \circ \beta)$.
(3) *If* $f_1(\alpha(0)) = f_2(\alpha(0))$*,* $f_1(\alpha(1)) = f_2(\alpha(1)))$ *and* $f_1 \sim f_2$ *relative to* $\{\alpha(0), \alpha(1)\}$*, then* $f_1 \circ \alpha \sim f_2 \circ \alpha$ *rel endpoints.*

Proof. For (1), if $F : I \times I \to X$ is a homotopy between α and α' relative endpoints, then $f \circ F : I \times I \to Y$ is a homotopy between $f \circ \alpha$ and $f \circ \alpha'$ relative endpoints. Part (2) is clear, and (3) is like (1): If $F : X \times I \to Y$ is a homotopy between f_1 and f_2 relative to $\{\alpha(0), \alpha(1)\}$, in other words, $F(x, 0) = f_1(x)$, $F(x, 1) = f_2(x)$, $F(\alpha(0), s) = f_1(\alpha(0)) = f_2(\alpha(0)))$ and $F(\alpha(1), s) = f_1(\alpha(1))) = f_2(\alpha(1))$ for all $s \in I$, then $G(t, s) = F(\alpha(t), s)$ is a homotopy between $f_1 \circ \alpha$ and $f_2 \circ \alpha$ relative to endpoints. □

The first part of the theorem justifies the following definition:

Definition 4.4. Let $f : X \to Y$ be continuous. Define a map f_* from homotopy classes of paths in X to homotopy classes of paths in Y by $f_*[\alpha] = [f \circ \alpha]$ for any path $\alpha : I \to X$.

Corollary 4.2. *Suppose that* $\alpha, \beta : I \to X$ *are paths from* x_0 *to* x_1 *and* x_1 *to* x_2*, and suppose* $f : X \to Y$ *and* $g : Y \to Z$ *are continuous. Then*

(1) $f_*([\alpha]\dot{[}\beta]) = f_*([\alpha]) \cdot f_*([\beta])$.
(2) $f_*([\epsilon_x]) = [\epsilon_{f(x)}]$ *for all* $x \in X$.
(3) $f_*([\alpha]^{-1}] = (f_*([\alpha]))^{-1}$.
(4) *If* $f_1, f_2 : X \to Y$ *are continuous,* $f_1(x_0) = f_1(x_0)$, $f_1(x_1) = f_2(x_1)$ *and* $f_1 \sim f_2$ *relative to* $\{x_0, x_1\}$*, then* $(f_1)_*([\alpha]) = (f_2)_*([\alpha])$.
(5) $(g \circ f)_*([\alpha]) = g_*(f_*([\alpha]))$.

Proof. Part (1) follows from (2) of Theorem 4.2, (2) is clear, (3) follows from (1), (2) and the usual arguments of uniqueness of inverses, (4) follows from (3) of Theorem 4.2, and (5) is clear. □

4.3. **Definition of the Fundamental Group.** The preceding discussion says that the set of homotopy classes of paths is somewhat like a group under composition (concatenation) of paths, and that a continuous map $f : X \to Y$ induces an operation f_* from classes in X to classes in Y that looks like a homomorphism. These classes do not form a group because the operation $[\alpha] \cdot [\beta]$ is not always defined, it requires that $\alpha(1) = \beta(0)$, and there are as many units $[\epsilon_x]$ as there are points $x \in X$. This structure is called a *groupoid*.

It is easy to get a group out of this situation: Fix a point $x_0 \in X$ and consider only paths $\alpha : I \to X$ that start and end at x_0: $\alpha(0) = \alpha(1) = x_0$. These are called *loops based at* x_0 and their homotopy classes form a group:

Definition 4.5. Let $x_0 \in X$. The *fundamental group of* X *based at* x_0, denoted $\pi_1(X, x_0)$ is defined as

$$\pi_1(X, x_0) = \{[\alpha] : \alpha \text{ is a loop based at } x_0\},$$

with multiplication $[\alpha] \cdot [\beta]$, unit $[\epsilon_{x_0}]$ and inverse $[\alpha]^{-1}$ as above.

Observe that Corollary 4.1 implies that $\pi_1(X, x_0)$ is indeed a group, with unit and inverses as asserted.

Remark 4.1. The notation π_1 is used because there are also groups denoted $\pi_2, \pi_3, \ldots$ which are defined by maps of higher dimensional objects into a space. We will not be concerned with these other groups. The letter "π" refers to Poincaré, who first defined a version of this group.

Example 4.1. Let $X = R^2$ and $x_0 = 0$. Then $\pi_1(\mathbb{R}^2, 0) = \{[\epsilon_0]\}$, the trivial group. The reason is very simple: if $\alpha : I \to \mathbb{R}^2$ is a loop at 0, then $F(t, s) = s\alpha(t)$ is a homotopy of α to ϵ_0 relative to 0. Thus there is only one element in $\pi_1(\mathbb{R}^2, 0)$. The same argument shows that if $C \subset \mathbb{R}^n$ is any convex set and $x_0 \in C$, then $\pi_1(C, x_0)$ is the trivial group.

4.4. **Properties of the Fundamental Group.** We now study some of the basic properties of the fundamental group. First, since its definition involved choosing a point $x_0 \in X$, usually called the *basepoint* of X. It is natural to ask how the group depends on the choice of basepoint.

Theorem 4.3. *Let $x_0, x_1 \in X$ and let $\sigma : I \to X$ be a path from x_0 to x_1. Then the map $\phi_\sigma : \pi_1(X, x_0) \to \pi_1(X, x_1)$ defined by $\phi_\sigma([\alpha]) = [\sigma^{-1}] \cdot [\alpha] \cdot [\sigma]$ is a group isomorphism.*

Proof. Clearly ϕ_σ is a homomorphism, and the map $\phi_{\sigma^{-1}}$ is its inverse. □

Figure 4.5 illustrates the map ϕ_σ. In the second half of the figure we have deformed the path slightly (keeping, of course, x_1 fixed) to better illustrate $[\phi_\sigma(\alpha)]$.

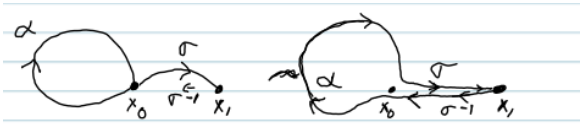


FIGURE 4.5. Changing the Basepoint

Thus different basepoints give isomorphic fundamental groups. This is often stated as "the fundamental group is independent of the basepoint", and this is a good first statement. But, as you study the subject more deeply, you realize that this is a somewhat inaccurate statement, because there can be many isomorphisms, depending on the choice of $[\sigma]$, and sometimes this choice can be important.

Next, if $f : X \to Y$ is a continuous map, and $\alpha \in \pi_1(X, x_0)$, then we have $f_*([\alpha]) \in \pi_1(X, f(x_0))$ defined as in Definition 4.4. The map f_* has the following properties:

Theorem 4.4. *Let $f : X \to Y$ be continuous, and let $f_* : \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$ be as in Definition 4.4. Then*

(1) *f_* is a group homomorphim.*
(2) *If $g : X \to Y$ is continuous, $g(x_0) = f(x_0)$ and $f \sim g$ relative to x_0, then $f_* = g_*$.*
(3) *If $g : Y \to Z$ is continuous, then $(g \circ f)_* = g_* \circ f_* : \pi_1(X, x_0) \to \pi_1(Z, g(f(x_0)))$.*

Proof. Immediate from Corollary 4.2 □

Definition 4.6. If $f : X \to Y$ is continuous, the map $f_* : \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$ is called the *homomorphism induced by f* or simply the *induced homomorphism.*

Corollary 4.3. *If $f : X \to Y$ is a homeomorphism, then the induced homomorphism $f_* : \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$ is an isomorphism.*

Proof. From (3) of Theorem 4.4, we see that $f_* \circ (f^{-1})_*$ is the identity on $\pi_1(Y, f(x_0))$ and $(f^{-1})_* \circ f_*$ is the identity on $\pi_1(X, x_0))$, thus f_* is an isomorphism with inverse $(f^{-1})_*$. □

4.5. **A Useful Lemma (Lebesgue Numbers).** Before proceeding, we prove a useful lemma that will be needed from time to time.

Lemma 4.1. *Let (X, d) be a compact metric space and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be an open cover of X. Then there exists an $\epsilon > 0$ (called a Lebesgue number for $\mathcal{U}$) so that, whenever $x, y \in X$ and $d(x, y) < \epsilon$, there exists $\alpha \in A$ so that $x, y \in U_\alpha$.*

Proof. Since $\mathcal{U}$ is an open cover of X, for each $x \in X$ there is an $\alpha \in A$ and an $\epsilon(x) > 0$ so that $B(x, 2\epsilon(x)) \subset U_\alpha$. The collection $\{B(x, \epsilon(x))\}_{x \in X}$ is an open cover of X. Let $\{B(x_1, \epsilon(1)), \ldots, B(x_n, \epsilon(n))\}$ be a finite subcover, and let $\epsilon = \min\{\epsilon(1), \ldots, \epsilon(n)\}$. Suppose $x, y \in X$ and $d(x, y) < \epsilon$. Then there is an i, $1 \leq i \leq n$, so that $x \in B(x_1, \epsilon(i))$. Since $d(x, y) < \epsilon$, the triangle inequality gives us that $y \in B(x_i, \epsilon(i) + \epsilon) \subset B(x_i, 2\epsilon(i)) \subset U_{\alpha_i}$. □

4.6. **The Fundamental Group of the Circle.** Corollary 4.3 finally gives us a topological invariant of spaces. But in able to use it we need some examples where the group is non-trivial. So far we have only seen Example 4.1 which gives the trivial group.

Let $S^1 \subset \mathbb{R}^2$ be the unit circle centered at the origin, take $(1, 0) \in S^1$ as basepoint.

Theorem 4.5. *The group $\pi_1(S^1, (1, 0))$ is isomorphic to $\mathbb{Z}$.*

Proof. The isomorphism is given by the "winding number" of a path. The proof begins by giving a rigorous construction of the winding number. It relies on familiar properties of the map $p : \mathbb{R} \to S^1$ defined by $p(t) = (\cos(2\pi t), \sin(2\pi t))$. This map is periodic of period 1 and descends to an isomorphism $\mathbb{R}/\mathbb{Z} \to S^1$ (see Examples 4.1 and 4.3 of [12]). If $J \subset \mathbb{R}$ is any interval of length less than one, then $p|J : J \to p(J)$ is a homeomorphism.

We choose the open cover $\mathcal{U} = \{U_1, U_2\}$ of S^1 where $U_1 = \{(x, y) \in S^1 : y > -\frac{1}{\sqrt{2}}\}$ and $U_2 = \{(x, y) \in S^1 : y < \frac{1}{\sqrt{2}}\}$. Let $U_1^0 = (-\frac{1}{8}, \frac{5}{8})$ and $U_2^0 = (-\frac{5}{8}, \frac{1}{8})$. Then $p(U_1^0) = U_1$, $p(U_2^0) = U_2$, and, if for each $i \in \mathbb{Z}$ we define

$$U_1^i = U_0 + i, \quad U_2^i = U_2 + i \quad \text{(translates by i)},$$

see Figure 4.6.

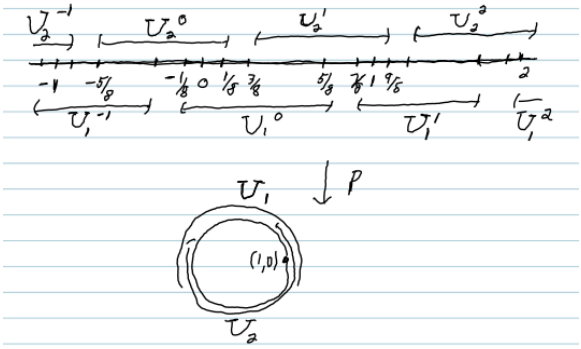


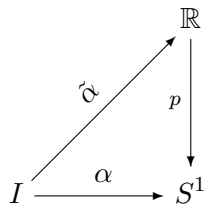
FIGURE 4.6. The Map $p : \mathbb{R} \to S^1$

Lemma 4.2. *The cover $\mathcal{U} = \{U_1, U_2\}$ of S^1 just defined has the property that each component U_1^i, $i \in \mathbb{Z}$, of $p^{-1}(U_1)$ is mapped homeomorphically by p to U_1, and the same for U_2.*

Proof. Clear, since every interval has length $\frac{3}{4} < 1$. □

Let $\alpha : I \to S^1$ be a loop based at $(1, 0)$, so $\alpha(0) = \alpha(1) = (1, 0)$. The first step is to construct a "lift" $\tilde{\alpha} : I \to \mathbb{R}$ of α, that is, a path $\tilde{\alpha}$ such that $p \circ \tilde{\alpha} = \alpha$ and $\alpha(0) = 0$:

(4.7)



To construct $\tilde{\alpha}$, let ϵ be a Lebesgue number (Lemma 4.1) for the open cover $\alpha^{-1}(\mathcal{U})$ of I. Divide I into sub-intervals of length $< \epsilon$, say $[0, t_1], [t_1, t_2]$,

$\dots [t_{n-1}, 1]$, where $0 < t_1 < t_2 < \cdots < 1$ and $t_i - t_{i-1} < \epsilon$. Then, by definition of Lebesgue number, for each i, $\alpha([t_{i-1}, t_i])$ is contained in one of the sets U_1 or U_2. Over each sub-interval $[t_{i-1}, t_i]$ there is no problem in constructing lifts. In other words, if in Diagram 4.7 we replace I by $[t_{i-1}, t_i]$, because on each component of $p^{-1}(U_1)$ or $p^{-1}(U_2)$ the map p is invertible, we can define a lift $\tilde{\alpha} = p^{-1} \circ \alpha$. The only problem is that there are infinitely many components, therefore infinitely many possibilities for p^{-1}. The possibilities have to be chosen in such a way that the choices in successive intervals match at their common endpoint to give a continuous map $\tilde{\alpha} : I \to \mathbb{R}$. These choices can be made as follows:

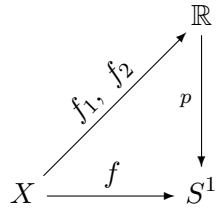
First, $\alpha([0, t_1])$ is contained in one of the two sets U_1, U_2. Choose this set and call it V_1. There is exactly one component of $p^{-1}(V_1)$ that contains 0, choose this component and call it W_1. Define $\tilde{\alpha}(t) = (p|_{W_1})^{-1} \circ \alpha(t)$ for $0 \leq t \leq t_1$. This defines $\tilde{\alpha}$ on the first interval $[0, t_1]$. Then define $\tilde{\alpha}$ on the next interval $[t_1, t_2]$: $\alpha([t_1, t_2])$ is contained in one of U_1, U_2, choose this set and call it V_2. There is exactly one component of $p^{-1}(V_2)$ that contains $\tilde{\alpha}(t_1)$, choose it and call it W_2. Define $\tilde{\alpha}(t) = (p|_{W_2})^{-1} \circ \alpha(t)$ for $t_1 \leq t \leq t_2$. This, combined with the previous definition, defines $\tilde{\alpha}$ on $[0, t_2]$. Continue this way: assume $\tilde{\alpha}$ has been defined on $[0, t_{i-1}]$ so that $p \circ \tilde{\alpha} = \alpha$, define it on $[0, t_i]$ by choosing a set $V_i = U_1$ or U_2 so that $\alpha([t_{i-1}.t_i]) \subset V_i$ and the component W_i of $p^{-1}(V_i)$ that contains $\tilde{\alpha}(t_{i-1})$, defining $\tilde{\alpha}(t) = (p|_{W_i})^{-1} \circ \alpha(t)$ for $t_{i-1} \leq t \leq t_i$, and thereby defining $\tilde{\alpha}$ on $[0, t_i]$ so that $p \circ \tilde{\alpha} = \alpha$. Continuing this way until we get to $t_n = 1$, we get the conclusion:

There exists a continuous lift $\tilde{\alpha}$ of α as in Diagram 4.7 satisfying $\tilde{\alpha}(0) = 0$.

Next, we need to know that this lift is independent of our construction, in fact, *any two continuous lifts $\tilde{\alpha}$ and $\tilde{\alpha}'$ of alpha as in Diagram 4.7 that agree at one point are equal.*

For future use, we record a more general fact:

Lemma 4.3. *Suppose X is a connected space, $f : X \to S^1$ is continuous and $f_1, f_2 : X \to \mathbb{R}$ are two continuous lifts of f:*



If there exists a point $x_0 \in X$ so that $f_1(x_0) = f_2(x_0)$, then $f_1(x) = f_2(x)$ for all $x \in X$.

Proof. Let $A = \{x \in X : f_1(x) = f_2(x)\}$. Since f_1, f_2 are continuous and S^1 is Hausdorff, A is closed. (See Homework for Math 4510). We claim

that A is also open: suppose $x \in A$, that is, $f_1(x) = f_2(x)$. Let $V = U_1$ or U_2 be an open set in the above cover $\mathcal{U}$ of S^1 containing $p(f_1(x)) = p(f_2(x)) = f(x)$, and let W be the component of $p^{-1}(V)$ that contains $f_1(x) = f_2(x)$. By continuity of f_1 and f_2 there exists a neighborhood U of x in X so that $f_1(U) \subset W$ and $f_2(U) \subset W$ (take such a neighborhood for each, and intersect them). Then, for all $y \in U$, $f_1(y) = (p|_W)^{-1} \circ f(y)$ and also $f_2(y) = (p|_W)^{-1} \circ f(y)$, so $f_1(y) = f_2(y)$ for all $y \in U$, so $U \subset A$ and A is open. Since $x_0 \in A$, $A \neq \emptyset$, so the connectedness of X implies that $A = X$. □

Going back to our loop $\alpha : I \to S^1$ and its lift $\tilde{\alpha} : I \to \mathbb{R}$ with $\tilde{\alpha}(0) = 0$, we see that $p(\tilde{\alpha}(1)) = (1, 0)$, thus $\tilde{\alpha}(1) \in p^{-1}(1, 0) = \mathbb{Z}$. This is our definition of winding number:

Definition 4.7. Let $\alpha : I \to S^1$ be a loop based at $(1, 0)$ and let $\tilde{\alpha} : I \to \mathbb{R}$ be its unique continuous lift with $\alpha(0) = 0$. The integer $\tilde{\alpha}(1)$ is called the *winding number* of α and is denoted $w(\alpha)$.

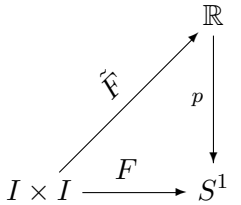
Remark 4.2. We could equally well done the following: construct a lift $\tilde{\alpha}'$ of α with $\tilde{\alpha}'(0)$ being any integer whatsoever (choose any $k \in \mathbb{Z}$ and start the above construction with $\tilde{\alpha}'(0) = k$), get the unique lift $\tilde{\alpha}'$ with $\tilde{\alpha}'(0) = k$. If we shift our old construction by k, we get $\tilde{\alpha} + k$ which is a lift of α that agrees with $\tilde{\alpha}'$ at 0. By uniqueness, $\tilde{\alpha}' = \tilde{\alpha} + k$. In particular, $\tilde{\alpha}'(1) - \tilde{\alpha}'(0) = \tilde{\alpha}(1) - \tilde{\alpha}(0) = w(\alpha)$. As a consequence we see that the winding number $w(\alpha)$ can be defined by taking *any lift* $\tilde{\alpha}$ of α and setting $w(\alpha) = \alpha(1) - \alpha(0)$.

Next, we need to show that $w(\alpha)$ depends only on the homotopy class $[\alpha] \in \pi_1(S^1, (1, 0))$. To do this, suppose α_0, α_1 are loops based at $(1, 0)$ and suppose $F : I \times I \to S^1$ is a homotopy:

$$F(t, 0) = \alpha_0(t),\ F(t, 1) = \alpha_1(t),\ F(0, s) = F(1, s) = (1, 0)$$

for all s, t. Let $\tilde{\alpha}_0 : I \to \mathbb{R}$ be a lift of α_0, say with $\alpha_0(0) = 0$. The next step is to lift F to a homotopy $\tilde{F} : I \times I \to \mathbb{R}$ with $\tilde{F}(t, 0) = \tilde{\alpha}_0(t)$.

(4.8)



We do this in the same way we constructed $\tilde{\alpha}$, but this time using the square $I \times I$ rather than the interval I. Given $F : I \times I \to S^1$ as above, let $\epsilon > 0$ be a Lebesgue number for the open cover $F^{-1}(\mathcal{U})$ of $I \times I$. Divide each factor I into intervals of length less than $\epsilon/\sqrt{2}$, say $0 < t_1 < t_2 \ldots t_{n-1} < 1$ and $0 < s_1 < \cdots < s_{n-1} < 1$ (no need to take the same number in each) so that all $t_i - t_{i-1}$ and $s_i - s_{i-1}$ are less than $\epsilon/\sqrt{2}$. Then, for all i, j, the square $[t_{i-1}, t_i] \times [s_{j-1}, s_j]$ has diameter $< \epsilon$, so any two points in the same square

are in some element of the cover, in other words, $F([t_{i-1}, t_i] \times [s_{j-1}, s_j]) \subset U_1$ or U_2.

We are in the same situation as before: if we replace $I \times I$ by $[t_{i-1}, t_i] \times [s_{j-1}, s_j]$ in Diagram 4.8, since p is invertible on each component of $p^{-1}(U_1)$ or $p^{-1}(U_2)$, there is no problem in lifting F: take $\tilde{F} = p^{-1} \circ F$. But there are infinitely many possibilities for p^{-1}, the problem again is to choose them so that they match along their common intervals to give a continuous map of the square $I \times I$ (and to satisfy $\tilde{F}|_{|i\times 0} = \tilde{\alpha}_0$).

Proceed much as before to make the consistent choices: $F([0, t_1] \times [0, s_1])$ is contained in one of U_1, U_2, choose it and call it $V_{1,1}$. There is a unique component of $p^{-1}(V_{1,1})$ that contains $\tilde{\alpha}_0([0, t_1])$, call it $W_{1,1}$. Define $\tilde{F}(t, s) = (p|_{W_{1,1}})^{-1} \circ F(t, s)$ for $(t, s) \in [0, t_1] \times [0, s_1]$.

Continue by increasing t: $F([t_1, t_2] \times [0, s_1])$ is contained in one of U_1, U_2, choose it, call it $V_{2,1}$. There is a unique component of $p^{-1}(V_{2,1})$ that contains the previously defined $\tilde{F}(t_1 \times [0, s_1])$, call it $W_{2,1}$. Define $\tilde{F}(t, s) = (p|_{W_{2,1}})^{-1} \circ F(t, s)$ for $(t, s) \in [t_1, t_2] \times [0, s_1]$. by the definition of $W_{2,1}$ this agrees on $t_1 \times [0, s_1]$ with the previous definition of $\tilde{F}$, so we get a well defined continuous lift $\tilde{F} : [0, t_2] \times [0, s_1] \to \mathbb{R}$. By construction $p(\tilde{F}(t, 0)) = F(t, 0) = \alpha(t)$ for $t \in [0, t_2]$ and $\tilde{F}(t, 0) = \tilde{\alpha}(t)$ for $t \in [0, t_1]$, thus by the uniqueness lemma (Lemma 4.3), we have that $\tilde{F}(t, 0) = \tilde{\alpha}(t)$ for $t \in [0, t_2]$.

Continue this way defining $\tilde{F} : [0, t_i] \times [0, s_1] \to \mathbb{R}$ until it is defined and continuous on $[0, 1] \times [0, s_1]$. In particular, we get that $\tilde{F}(t, 0) = \tilde{\alpha}(t)$ for all $t \in I$.

Then increase s: $F([0, t_1] \times [s_1, s_2]) \subset U_1$ or U_2 , choose one, call it $V_{1,2}$, there is a unique component of $p^{-1}(V_{1,2})$ that contains $\tilde{F}([0, t_1] \times s_1)$, call it $W_{1,2}$ and define $\tilde{F}(t, s) = (p|_{W_{1,2}})^{-1} \circ F(t, s)$ for $(t, s) \in [0, t_1] \times [s_1, s_2]$, combine it with the previous construction to get a continuous lift on $([0, 1] \times [0, s_1]) \cup ([0, t_1] \times [s_1, s_2])$.

Continue now by increasing t as we did before: at each stage choose $V_{i,2}$ so that the definition $\tilde{F}(t, s) = (p|_{W_{i,2}})^{-1} \circ F(t, s)$ for $(t, s) \in [t_{i-1}, t_i] \times [s_1, s_2]$ agrees with the previously defined $\tilde{F}$ on the vertical segment $t_{i-1} \times [s_2, s_2]$. Then on the horizontal segment $[t_{i-1}, t_i] \times s_1$ the new definition of $\tilde{F}$ agrees with the old one because they are both lifts of $F(t, s_1)$, $t \in [t_{i-1}, t_i]$ that agree at (t_{i-1}, t_i) (another application of Lemma 4.3).

Continue in this fashion until $\tilde{F}$ is defined and continuous on all of $I \times I$ as in Diagram 4.8.

Observe that $\tilde{F}(0, s)$ is a continuous function of s so that $p \circ \tilde{F}(0, s) = F(0, s) = (1, 0)$, so it is a continuous map $I \to p^{-1}(1, 0) = \mathbb{Z}$, hence constant $= \tilde{\alpha}_0(0) = 0$. Hence $\tilde{F}(t, 1)$ is a lift of α_1 starting at 0, so by uniqueness

(Lemma 4.3), we have $\tilde{F}(t,1) = \tilde{\alpha}_1(t)$ for all $t \in I$. Finally $F(1,s)$ is a continuous function of s with values in $p^{-1}(1,0) = \mathbb{Z}$, hence constant, hence $= \tilde{\alpha}_1(1)$ and also $= \tilde{\alpha}_0(1)$, hence $\tilde{\alpha}_1(1) = \tilde{\alpha}_0(1)$, in other words, the winding numbers are equal: $w(\alpha_0) = w(\alpha_1)$ as desired. Thus we can write define the *winding number* of an element of $\pi_1(S^1,(1,0))$ as $w([\alpha]) = w(\alpha)$ for any representative α of $[\alpha]$.

Lemma 4.4. *The map $w : \pi_1(S^1,(1,0)) \to \mathbb{Z}$ is a group isomorphism.*

Proof. We need to check:

(1) w is a homomorphism: $w([\alpha \cdot \beta]) = w([\alpha]) + w([\beta])$. This follows from Remark 4.2: if to lift $\alpha \cdot \beta$ starting at 0, we lift α starting at 0 and follow it by the lift of β starting at $\tilde{\alpha}(1)$. But this lift is exactly $\tilde{\alpha}(1) + \tilde{\beta}$, thus $w(\alpha \cdot \beta) = \widetilde{\alpha \cdot \beta}(1) = \tilde{\alpha}(1) + \tilde{\beta}(1) = w(\alpha) + w(\beta)$.
(2) w is surjective: let $n \in \mathbb{Z}$ and let $\alpha_n(t) = (\cos(2\pi nt), \sin(2\pi nt))$. Then $w(\alpha_n) = n$.
(3) w is injective: suppose $w(\alpha) = 0$. Then $\tilde{\alpha}(0) = \tilde{\alpha}(1) = 0$, so $F(t,s) = s\tilde{\alpha}(t)$ is a homotopy of $\tilde{\alpha}$ to 0 (see Example 4.1), then $p \circ F$ is a homotopy of α to the constant path $(1,0)$, thus $[\alpha] = e \in \pi_1(S^1,(1,0))$.

□

This completes the proof of Theorem 4.5. □

4.7. **Retractions and Deformation Retractions.** Before giving applications of the computation $\pi_1(S^1) \cong \mathbb{Z}$ we introduce some terminology.

Definition 4.8. Let $A \subset X$ be a subspace, and let $i : A \to X$ be the inclusion map $i(a) = a$ for all $a \in A$.

(1) A is called a *retract of* X if there exists a continuous map $r : X \to A$ such that $r \circ i = id_A$. The map r is called a *retraction of X to A*.
(2) A is called a *deformation retract of* X if, in addition, $i \circ r \sim id_X$ relative to A, in other words, there exists a continuous map $F : X \times I \to X$ such that $F(x,0) = i(r(x)), F(x,1) = x$ and $F(a,s) = a$ for all $a \in A$. The map r is called a *deformation retraction* of X to A.

Example 4.2. (1) S^1 is a deformation retract of $\mathbb{R}^2 \setminus \{0\}$. Let $i : S^1 \to \mathbb{R}^2 \setminus \{0\}$ be the inclusion, and define $r : \mathbb{R}^2 \setminus \{0\}$ by

$$r(x) = \frac{x}{||x||}$$

Then $r(i(x)) = x/1 = x$ if $x \in S^1$, so r is a retraction. In addition, if we let

$$F(x,t) = tx + (1-t)\frac{x}{||x||},$$

then $F : \mathbb{R}^2 \setminus \{0\} \times I \to \mathbb{R}^2 \setminus \{0\}$ is continuos (since the straight line segment from x to $x/||x||$ does not go through the origin), $F(x,0) = i(r(s))$, $F(x,1) = x$ and $F(x,t) = x$ for all $x \in S^1$.

(2) By the same reasoning, for any n, the unit sphere S^n is a deformation retract of $\mathbb{R}^{n+1} \setminus \{0\}$.

(3) By the same reasoning, for any a, b, c so that $0 \leq a < b < c$, the sphere of radius b, $S_b = \{||x|| = b\}$ is a deformation retract of the "annulus" $a < ||x|| < c$.

(4) Let $X = \mathbb{R}^2 \setminus \{(0,1)\}$, let A be the x - axis and let $r : X \to A$ be defined by $r(x,y) = (x,0)$. Then $r(x,0) = (x,0)$, so r is a retraction, but r is not a deformation retraction: if it were, then, by Theorem 4.6 the fundamental groups $\pi_1(A)$ and $\pi_1(X)$ would be isomorphic, but we know that $\pi_1(A)$ is the trivial group, and, by part (1), X has a circle as deformation retract, so $\pi_1(X) = \mathbb{Z}$.

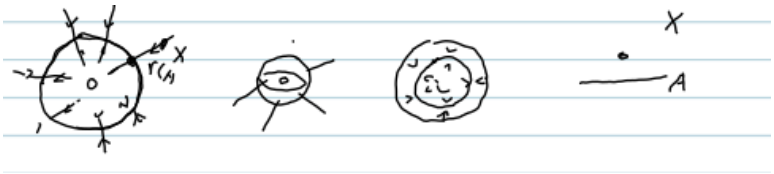


FIGURE 4.7. The Retractions of Example 4.2

Theorem 4.6. (1) *Suppose that $A \subset X$ is a retract, with retraction r, and $x_0 \in A$. Then $r_* \circ i_* = id : \pi_1(A, x_0) \to \pi_1(A, x_0)$. In particular, $i_* : \pi_1(A, x_0) \to \pi_1(X, x_0)$ is injective and $r_* : \pi_1(X, x_0) \to \pi_1(A, x_0)$ is surjective.*

(2) *If $A \subset X$ is a deformation retract, then $i_* : \pi_1(A, x_0) \to \pi_1(X, x_0)$ is an isomorphism, with inverse r_*.*

Proof. For (1), since $r \circ i = id_A$, we have that $r_* \circ i_* = id_A$. This formally implies that i_* is injective: if $i_*(\alpha) = e$, then, on the one hand, $r_*(i_*(\alpha)) = \alpha$, and also $r_*(i_*(\alpha)) = r_*(i_*(e)) = e$, thus $\alpha = e$, thus i_* is injective. Similarly, it formally follows that r_* is surjective: if $\beta \in \pi_1(A, x_0)$, then $\beta = r_*(i_*(\beta))$, so $\beta = r_*(\alpha)$ where $\alpha = i_*(\beta)$.

Part (2) is immediate since, in addition, $i_* \circ r_* = id$, so r_* is a two - sided inverse of i_*. □

4.8. **Applications of Theorem 4.5.** We can now one the example we know of a non-trivial fundamental group to get some topological consequences. Let, as usual, D denote the closed unit disk $D = \{x^2 + y^2 \leq 1\}$ with boundary S^1.

Theorem 4.7. *There is no retraction $r : D \to S^1$.*

Proof. Suppose there were a retraction $r : D \to S^1$. Then $r_* \circ i_* = id$ on $\pi_1(S^1, x_0) \cong \mathbb{Z}$. But, since $\pi_1(D, x_0)$ is the trivial group $\{e\}$, we get that for all $\alpha \in \pi_1(S^1, x_0)$, $i_*(\alpha) = e$, thus $r_* \circ i_*$ is both the zero homomorphism and the identity homomorphism of $\mathbb{Z}$, which is impossible. □

.

The interest of this Theorem is the following famous consequence:

Corollary 4.4. *(The Brouwer Fixed Point Theorem): Let $f : D \to D$ be continuous. Then f has a fixed point. In other words, there exists an $x \in D$ so that $f(x) = x$.*

Proof. Suppose, on the contrary, that there were a continuous map $f : D \to D$ so that $f(x) \neq x$ for all $x \in D$. Then we could construct a retraction $r : D \to S^1$ by letting $r(x)$ be the point of intersection of the directed ray from $f(x)$ to x with the boundary circle S^1, see Figure 4.8 Note that if $x \in S^1$, this ray intersects S^1 at x, so r is the identity on S^1 and we would indeed get a retraction. We do not give a formal proof of the continuity of r, but observe that, parametrizing this ray as $(1-t)x + tf(x)$, $t \geq 0$, the value of t for which the ray meets S^1 is the root of a quadratic equation that is at least 1, so we need to show the continuity of this root as a function of the coefficients. To give a rigorous proof, we can use the quadratic formula, which gives us the continuity provided that the discriminant of this equation is never zero. Checking this is a computation using the Cauchy-Schwarz inequality. □

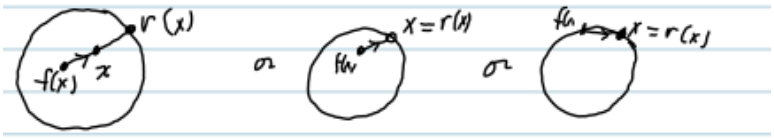


FIGURE 4.8. The Map $p : \mathbb{R} \to S^1$

Remark 4.3. Both the no-retraction theorem (Theorem 4.7) and the Brouwer fixed point theorem are true for the closed unit ball in $\mathbb{R}^n$ for any n. For $n = 1$ this just requires connectedness of the unit interval, for $n = 2$ we used the fundamental group, for $n > 2$ it requires higher dimensional topological invariants.

Theorem 4.8. *Let $H = \{(x, y) \in \mathbb{R}^2 : y \geq 0\}$ be a closed half-plane in $\mathbb{R}^2$, and let $\partial H = \{(x, 0)\}$ (the x-axis). Let $f : H \to H$ be a homeomorphism. Then $f(\partial H) \subset \partial H$. (Consequently, applying the same to f^{-1}, $f(\partial H) = \partial H$, and the restriction of f to ∂H is a homeomorphism.)*

Proof. Let $p = (x_0, 0) \in \partial H$ be arbitrary. We need to show that $f(p) \in \partial H$. We argue as we did in Math 4510 with intervals and their boundaries. Suppose $f(p) = (x_1, y_1)$ is not in ∂H, that is, $y_1 > 0$. Then $f|_{H\setminus\{p\}} : H\setminus\{p\} \to H\setminus\{f(p)\}$ is a homeomorphism. Let's prove that this is impossible by computing their fundamental groups.

Take a basepoint in $H \setminus \{p\}$, say $(x_0, 1)$, and let $\alpha(t) = (x(t), y(t))$ be a loop based at $(x_0, 1)$. Then $F(t, s) = (1 - s)(x(t), y(t)) + s(x_0, 1)$ is a homotopy of α to the constant loop at $(x_0, 1)$. We only need to check that the paths always lie in $H \setminus \{p\}$, which means that its second coordinate is always ≥ 0 and > 0 if its fist coordinate $= x_0$. But the second coordinate of $F(t, s)$ is $(1 - s)y(t) + s$, which is > 0 when $s > 0$ (since $y(t) \geq 0$), and $= y(t)$ for $s = 0$, so it clearly satisfies the required condition for all $(t, s) \in [0, 1] \times [0, 1]$. Therefore $\pi_1(H \setminus \{p\})$ is the trivial group for this basepoint, hence for any basepoint.

Now take a basepoint in $H \setminus \{f(p)\}$, say $(x_1, \frac{y_1}{2})$. Let C be the circle of radius $\frac{y_1}{2}$ centered at (x_1, y_1), and note that C is a deformation retract of $H \setminus \{f(p)\}$: each point $(x, y) \in H$ lies in a unique ray from (x_1, y_1), let $r(x, y)$ be the point of intersection of this ray with C, see Figure 4.9. In formulas

$$r(x, y) = (x_1, y_1) + \frac{y_1}{2\sqrt{(x - x_1)^2 + (y - y_1)^2}} (x - x_1, y - y_1),$$

so r is continuous and $r(x, y) = (x, y)$ for all $(x, y) \in C$, and homotopy

$$F((x, y), s) = (x_1, y_1)+s(x-x_1, y-y_1)+\frac{(1 - s)y_1}{2\sqrt{(x - x_1)^2 + (y - y_1)^2}} (x-x_1, y-y_1)$$

from $i\circ r$ and id: $F((x, y), 0) = r(x, y)$, $F((x, y), 1) = (x, y)$ and $F((x, y), s) = (x, y)$ for all $(x, y) \in C$. Thus, with this basepoint, $\pi_1(H\setminus\{f(p)\}) = \pi_1(C) = \mathbb{Z}$.

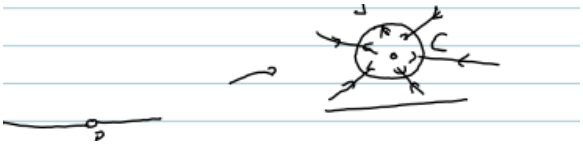


FIGURE 4.9. Boundary of Half Plane is Topologically Invariant

Finally, if we had $f(p) = (x_1, y_1)$ with $y_1 > 0$, we would get that $f_*(\pi_1(H\setminus\{p\}, p_0))) \to \pi_1(H \setminus \{f(p)\})$ would be an isomorphism from a trivial group to a group isomorphic to $\mathbb{Z}$, which is impossible. Therefore we must have $y_1 = 0$, in other words, $f(p) \in \partial H$ as desired. □

This theorem says that the boundary has topological significance, at least in the case of a half-plane. Since any surface with boundary is locally homeomorphic to H, the same should be true of any surface with boundary. Recall

Definition 1.1 for the meaning of surface with boundary S and let ∂S denote the set of its boundary points. We are thinking of topological surfaces with boundary and homeomorphisms. Similar (but easier to prove) statements hold for differentiable surfaces with boundary and diffeomorphisms.

Theorem 4.9. *Let S_1 and S_2 be surfaces with boundary and let $f : S_1 \to S_2$ be a homeomorphism. Then $f(\partial S_1) \subset \partial S_2$ (and, consequently, $f(\partial S_1) = \partial S_2$.)*

Proof. To use the same reasoning as in Theorem 4.8 we need the following lemma:

Lemma 4.5. *Let S be a surface, with or withoug boundary, let q be an interior point of S, let $\phi : U \to D$ be a coordinate chart centered at q, where $D \subset \mathbb{R}^2$ is the open unit disk centered at the origin, and $\phi(q) = 0$. If V is any connected neighborhood of q contained in U ($q \in V \subset U$), then $\pi_1(V \setminus \{q\})$ is an infinite group.*

Proof. By applying the homeomorphism ϕ, we may assume that $U = D$ and $q = 0$. Given any neighborhood V of 0, $V \subset D$, there exists a disk D_r of radius r centered at 0 so that $D_r \subset V$. Let $i : D_r \to V$ and $j : V \to D$ be the inclusion maps. Then $j \circ i$ is the inclusion of D_r in D and $(j \circ i)_* : \pi_1(D_r \setminus \{0\}) \to \pi_1(D \setminus \{0\})$ is an isomorphism, since both $D_r \setminus \{0\}$ and $D \setminus \{0\}$ contain the circle C of radius $\frac{r}{2}$ as a deformation retract and $j \circ i$ is the identity on C. Since $(j \circ i)_* = j_* \circ i_*$, and fixing a point in C as the basepoint in the fundamental groups, we have:

$$\begin{array}{ccccc})\pi_1(D_r \setminus \{0\}) & \xrightarrow{i_*} & \pi_1(V \setminus \{0\}) & \xrightarrow{j_*} & \pi_1(D \setminus \{0\}) \\ \downarrow \cong & & & & \downarrow \cong \\ \pi_1(C) \cong \mathbb{Z} & & \xrightarrow{\qquad id \qquad} & & \pi_1(C) \cong \mathbb{Z} \end{array}$$

therefore i_* is injective (and j_* is surjective), therefore $\pi_1(V \setminus \{0\})$ contains a subgroup isomorphic to $\mathbb{Z}$ (and surjects to a group isomorphic to $\mathbb{Z}$), so it is an infinite group, as asserted, see Figure 4.10. □

We can now prove the theorem: Suppose $f : S_1 \to S_2$ is a homeomorphism, suppose that $p \in \partial S_1$, and suppose that $q = f(p)$ is an interior point of S_2. Take a neighborhood U of q as in the Lemma, then $f^{-1}(U)$ is a neighborhood of p, and there is a neighbohood W of p, $p \in W \subset f^{-1}(U)$ that is the domain of a coordinate chart $\psi : W \to H$ where H is a half-disk $\{x^2 + y^2 < 1, y \geq 0\}$ and $\psi(p) = (0, 0)$. Then, just as in the proof of Theorem 4.8, $\pi_1(H \setminus \{(0, 0)\})$ is the trivial group, thus $\pi_1(W \setminus \{p\})$ is the trivial group. But $f(W) = V$ satisfies the assumptions of the Lemma and f takes $W \setminus \{p\}$ homeomorphically to $V \setminus \{q\}$, which is impossible since

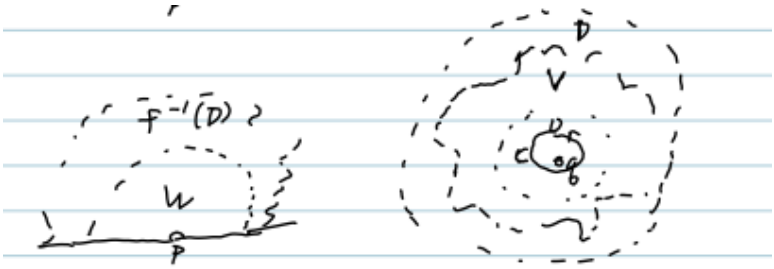


FIGURE 4.10. Fundamental Groups of Deleted Neighborhoods

the fundamental group of the latter is infinite, so, in particular, non-trivial. Therefore $f(p)$ must be a boundary point of S_2. (See Figure 4.10.) □

4.9. **Examples of Induced Homomorphisms.** The applications of the fundamental group to problems in topology, as the ones we have just seen in section 4.8, require the computation of the induced homomorphisms on the fundamental group (Definition 4.6). Facility in computing these homomorphisms is essential in this subject. Remember that the main idea is that to continuous maps $f : X \to Y$ there is an induced homomorphism $f_* : \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$, *and compositions go to compositions*: if $g : Y \to Z$, then $(g \circ f)_* = g_* \circ f_*$. Thus the algebra mirrors the topology. It is this property (called "*functoriality*") that makes the subject work. This is what is used in Theorems 4.6 and 4.7 to prove that certain maps do not exist. It is also used in Lemma 4.5 to prove that certain spaces have non-trivial fundamental group.

Example 4.3. For $n \in \mathbb{Z}$, let $f_n : S^1 \to S^1$ be the map defined (for complex numbers) by $f_n(z) = z^n$, in other words, $f_n(\cos(2\pi t), \sin(2\pi t)) = (\cos(2\pi nt), \sin(2\pi nt))$. Since f_n wraps the circle around itself $|n|$ times (in the reverse direction if $n < 0$), it is reasonable to expect that, under the winding number isomorphism of Lemma 4.4, f_n corresponds to mutiplication by n, in other words, that we have the commutative diagram (4.9)

$$\begin{array}{ccc} \pi_1(S^1,(1,0)) & \xrightarrow{(f_n)_*} & \pi_1(S^1,(1,0)) \\ \downarrow{\scriptstyle w} & & \downarrow{\scriptstyle w} \\ \mathbb{Z} & \xrightarrow{\quad n \quad} & \mathbb{Z} \end{array} \tag{4.9}$$

where w is the winding number isomorphism of Lemma 4.4 and the bottom arrow is the map $\mathbb{Z} \to \mathbb{Z}$ that sends x to nx (multiplication by n).

To prove that this is indeed the case, we have to go back to the definitions. First, we need a generator of $\pi_1(S^1, (1, 0))$, and Lemma 4.4 gives us one, namely the loop α_1, where, for any $n \in \mathbb{Z}$, $\alpha_n(t) = (\cos(2\pi nt), \sin(2\pi nt))$,

$0 \leq t \leq 1$ is a loop in S^1 based at $(1,0)$. Then, by definition, $(f_n)_*[\alpha_1] = [f_n \circ \alpha_1] = [\alpha_n]$ since $f_n(\alpha_1(t)) = f_n(\cos(2\pi t), \sin(2\pi t)) = (\cos(2\pi nt), \sin(2\pi nt))$ $= \alpha_n(t)$. Since, again by Lemma 4.4, $w(\alpha_n) = n$, we get the assertion of Diagram 4.9. Or, in multiplicative notation, $(f_n)_*(\alpha) = \alpha^n$ for all $\alpha \in \pi_1(S^1, (1,0))$.

Example 4.4. Let us look at maps of the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$. Recall that $T^2 = S^1 \times S^!$. From a homework problem (or p. 77 of [9]) we know that $\pi_1(X \times Y) \cong \pi_1(X) \times \pi_1(Y)$, where the isomorphism is given by $\alpha \to ((p_X)_*\alpha, (p_Y)_*\alpha)$ for all $\alpha \in \pi_1(X \times Y)$, where p_X, p_Y are the projections of $X \times Y$ to X and Y respectively.

Putting these facts together with Lemma 4.4, we get the following description of $\pi_1(T^2)$. Let p_1, p_2 be the projections of $S^1 \times S^1$ onto the first and second factors respectively, and let $\alpha : I \to S^1 \times S^1$ be a loop based at $((1,0),(1,0))$ To α we can assign two winding numbers: $w(p_1 \circ \alpha)$ and $w(p_2 \circ \alpha)$. The map $\pi_1(T^2) \to \mathbb{Z}^2$ given by $\alpha \to (w(p_1 \circ \alpha), w(p_2 \circ \alpha))$ is a group isomorphism.

To compute induced homomorphisms it will be easier to represent T^2 as $\mathbb{R}^2/\mathbb{Z}^2 \cong (\mathbb{R}/\mathbb{Z}) \times (\mathbb{R}/\mathbb{Z})$, so we need to describe $\pi_1(T^2)$ also is this context. Let $p : \mathbb{R}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ be the projection to the quotient space. Every loop in T^2 at the point $p(0,0)$ is of the form $p \circ \alpha$, where $\alpha : I \to \mathbb{R}^2$ is a path from $(0,0)$ to some pont $(m,n) \in \mathbb{Z}^2$. The projections p_1 and p_2 of this loop are the projections to $\mathbb{R}/\mathbb{Z}$ of paths in $\mathbb{R}$ from 0 to m, n respectively. This shows that (m,n) is the pari of winding numbers assigned to the loop $p \circ \alpha$ in the above isomorphism. A path with giving pair of winding numbers (m,n) is $p \circ \alpha_{m,n}$, where $\alpha_{m,n}(t) = (mt, nt)$, $0 \leq t \leq 1$, see Figure 4.11

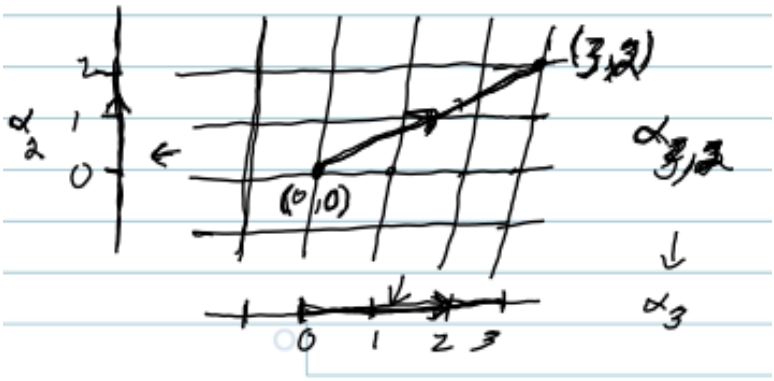


FIGURE 4.11. The loop $\alpha_{3,2}$ in T^2.

If we let 0 denote the point $p(0,0) \in T^2$, we can summarize this discussion:

Lemma 4.6. *The Map $\phi : \mathbb{Z}^2 \to \pi_1(T^2, 0)$ defined by $\phi(m,n) = [p \circ \alpha_{m,n}]$ is a group isomorphism.*

Proof. Clear from the discussion. □

Now to maps of the torus. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \text{where} \quad a, b, c, d \in \mathbb{Z}$$

be an integral matrix. Define a map $f_A : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$f_A(x, y) = (ax + by, cx + dy). \tag{4.10}$$

(if we wrote (x, y) as a column vector, which we hesitate to do for typographical reasons, f_A would be left multiplication by A).

Then $f_A(\mathbb{Z}^2) \subset \mathbb{Z}^2$, so there is a well defined map, still denoted f_A, on $\mathbb{R}^2/\mathbb{Z}^2$, defined by $f_A((x, y) + \mathbb{Z}^2) = f_A(x, y) + \mathbb{Z}^2$. Let's compute the induced homomorphism $(f_A)_* : \pi_1(T^2, 0) \to \pi_1(T^2, 0)$. Take a typical element $[p \circ \alpha_{m,n}] \in \pi_1(T^2, 0)$. Then, by definition, $(f_A)_*([p \circ \alpha_{m,n}]) = [f_A \circ (p \circ \alpha_{m,n})] = [p \circ (f_a \circ \alpha_{m,n})]$, where the first equality is the definition of induced homomorphism, and the second equality is the definition of maps on quotient spaces by taking representatives. Now, $f_A(\alpha_{m,n}(t)) = (amt + bnt, cmt + dnt) = \alpha_{am+bn,cm+dn}(t)$. In other words, under the isomorphism ϕ of Lemma 4.6, $(f_A)_*$ is the same as the restriction of the linear transformation f_A to $\mathbb{Z}^2$, as in Diagram 4.11

$$\begin{array}{ccc} \mathbb{Z}^2 & \xrightarrow{(\phi)} & \pi_1(T^2, 0) \\ \Big\downarrow{\scriptstyle f_A|_{\mathbb{Z}^2}} & & \Big\downarrow{\scriptstyle (f_A)_*} \\ \mathbb{Z}^2 & \xrightarrow{\phi} & \pi_1(T^2, 0) \end{array} \tag{4.11}$$

If the matrix A has determinant ± 1: $ad - bc = \pm 1$, then A^{-1} is also an integral matrix, so $f_{A^{-1}}(\mathbb{Z}^2) \subset \mathbb{Z}^2$. It is easy to check that $(f_A)^{-1} = f_{A^{-1}}$. therefore $(f_A)^{-1}$ gives a well-defined map of T^2 and f_A is a homeomorphism of T^2.

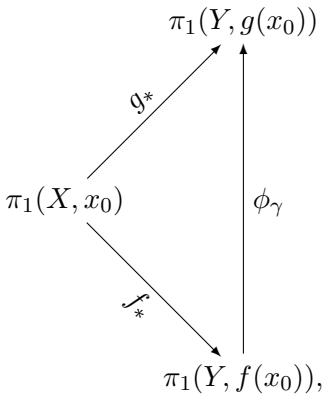
4.10. **Homotopy Equivalence.** We know that the fundamental group is invariant under homeomorphism, but it is also invariant under a weaker equivalence relation between spaces, called *homotopy equivalence*:

Definition 4.9. Let X and Y be topological spaces (we do not need to assume connected). We say that X and Y are *homotopy equivalent* or *have the same homotopy type* if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ so that $g \circ f \sim id_X$ and $f \circ g \sim id_Y$.

Remark 4.4. (1) If instead of saying $g \circ f \sim id_X$ and $f \circ g \sim id_Y$ we had said $g \circ f = id_X$ and $f \circ g = id_Y$ we would have defined the notion of *homeomorphism*. Thus homotopy equivalence is a natural weakening of homeomorphism.

(2) If f and g are as in the definition, it is natural to say that f and g are *homotopy inverses* to each other and that they are *homotopy equivalences*.
(3) Note that it is not assumed that $g \circ f$, $f \circ g$ are homotopic to the identity relative to any basepoint. So, to see what homotopy equivalence says about the fundamental group, we need to study the effect on the fundamental group of homotopies that do not preserve the basepoint.

Theorem 4.10. *Let X, Y be connected and locally path connected, let $f, g : X \to Y$ be continuous, suppose that $F : X \times I \to Y$ is a homotopy from f to g: $F(x,0) = f(x)$, $F(x,1) = g(x)$ (but not assumed to preserve any basepoint), and suppose $x_0 \in X$ is a basepoint. Let $\gamma : I \to Y$ be the path $\gamma(s) = F(x_0, s)$ traced by the basepoint under the homotopy, and let $\phi_\gamma : \pi_1(Y, f(x_0)) \to \pi_1(Y, g(x_0))$ be the isomorphism $\phi_\gamma(\beta) = \gamma^{-1} \cdot \beta \cdot \gamma$ as in Theorem 4.3. Then the induced homomorphisms $f_* : \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$ and $g_* : \pi_1(X, x_0) \to \pi_1(Y, g(x_0))$ are related by $g_* = \phi_\gamma \circ f_*$:*



or, more explicitly, $g_*(\alpha) = \gamma^{-1} \cdot f_*(\alpha) \cdot \gamma$ for all $\alpha \in \pi_1(X, x_0)$.

Proof. Let $\alpha : I \to X$ be a loop based at x_0. Then the map $G : I \times I \to Y$ defined by $G(t,s) = F(\alpha(t), s)$ is a homotopy of $f \circ \alpha$ to $g \circ \alpha$ where the basepoint moves tracing the path γ. To construct a homotopy between $g \circ \alpha$ and $\gamma^{-1} \cdot f \circ \alpha \cdot \gamma$ preserving the basepoint $g(x_0)$ we need a suitable map $p : I \times I \to I \times I$ so that $G \circ p$ is such a homotopy. Figure 4.12 shows, on the left, what the desired map does on the boundary of $I \times I$, and, on the right, how G and $G \circ p$ are related. Its explicit construction is left as an exercise. □

Corollary 4.5. *If $f, g : X \to Y$ are as in the theorem, and, in addition, $f(x_0) = g(x_0) = y_0$, then the map ϕ_γ in Diagram 4.10 is an inner automorphism of $\pi_1(Y, y_0)$, in other words, $\gamma \in \pi_1(Y, y_0)$ and $\phi_\gamma(\beta) = \gamma^{-1} \cdot \beta \cdot \gamma$ conjugation by an element of this group.*

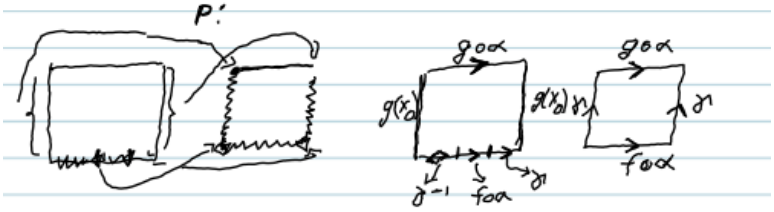


FIGURE 4.12. Effect of Homotopy Moving Basepoint

Proof. If $f(x_0) = g(x_0) = y_0$, then $\gamma(s) = F(x_0, s)$ is a loop based at y_0, thus represents an element of $\pi_1(Y, y_0)$. □

Example 4.5. Let A abd B be tow integral 2 by 2 matrices and let $f_A, f_B : T^2 \to T^2$ be the maps defined in Example 4.4. Since $f_A(0) = f_B(0) = 0$, if f_A were homotopic to f_B, then, for some $\gamma \in \pi_1(T^2, 0)$, we would have that for all $\alpha \in \pi_1(T^2, 0)$, $(f_B)_*(\alpha) = \gamma^{-1} \cdot ((f_A)_*(\alpha)) \cdot \gamma = (f_A)_*(\alpha)$, the last equality because $\pi_1(T^2, 0) = \mathbb{Z}^2$ is abelian. Thus $(f_A)_* = (f_B)_*$, and by Diagram 4.11 we see that $A = B$. Thus f_A is homotopic to f_B if and only if $A = B$. Observe that we are not assuming that the homotopy preserves the basepoint.

Corollary 4.6. *If X and Y are homotopy equivalent, then, for any basepoints in X and Y, $\pi_1(X) \cong \pi_1(Y)$. More precisely, $g_* \circ f_*$ and $f_* \circ g_*$ are isomorphisms obtained as in Diagram 4.10. In particular, f_* and g_* are isomorphisms.*

Proof. Clear. □

Example 4.6.

(1) Any deformation retraction is a homotopy equivalence. This applies to the first three parts of Example 4.2. In particular, S^n and $\mathbb{R}^{n+1} \setminus \{0\}$ are homotopy equivalent.
(2) A solid torus $S^1 \times D$ is homotopy equivalent to S^1, see Figure 4.13.
(3) The solid H_g bounded by the standard picture of the surface Σ_g is homotopy equivalent to a one-dimensional object, as in Figure 4.13
(4) All the previous examples are deformation retractions, and involve spaces of different dimensions (here we are speaking intuitively about dimension, its topological definition and its properties requires some work). Here is an example of a homotopy equivalence, which is not a homeomorphism, between two surfaces with boundary. Let S_1 be a torus T^2 with an open disk removed, and let S_2 be a closed disk with two open disks removed, see Figure 4.14. Then both S_1 and S_2 deformation retract to the "figure 8 (more formally, the one-point union of two circles, see Definition 4.11 below). Then S_1 is homotopy equivalent to S_2, but they are not homeomorphic since ∂S_1 is connected while ∂S_2 has three connected components, and we now know that homeomorphisms respect boundaries (Theorem 4.9)

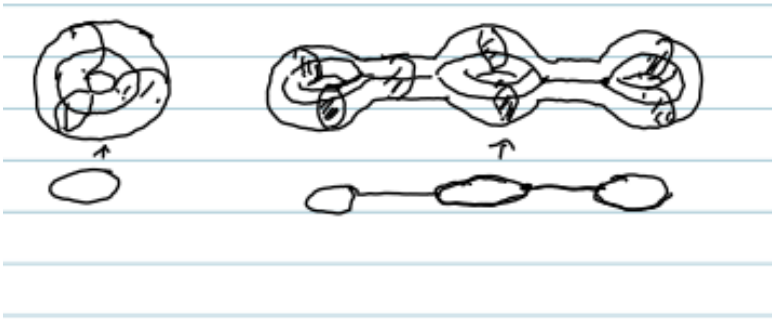


FIGURE 4.13. Solids Bounded by Surfaces

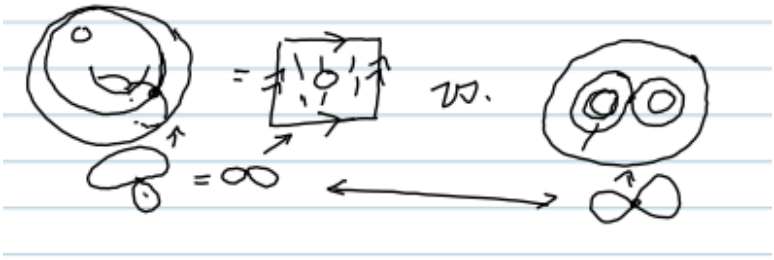


FIGURE 4.14. A Homotopy Equivalence of Surfaces with Boundary

(5) It is a fact that for compact surfaces with empty boundary, homotopy equivalence implies homeomorphism. This follows from classification of surfaces (which we know) and the homotopy invariance of Euler characteristic and orientability (which we have not established).

Finally some terminology:

Definition 4.10. A connected, locally path connected space X is said to be *simply connected* if $\pi_q(X)$ is the trivial group. The space X is said to be *contractible* if it is homotopy equivalent to a point.

Remark 4.5. There are several equivalent characterizations of simply connected spaces: X is simply connected if and only if:

(1) For all x_1, x_2 in X and for any two paths α, β from x_1 to x_2, $\alpha \sim \beta$ relative to the endpoints.
(2) Any continuous map $f : S^1 \to X$ is homotopic to a constant map.
(3) Any continuous map $f : S^1 \to X$ extends to a continuous map $g : D \to X$ of the closed unit disk.

See, for instance, [5, 9] for more details. Moreover, a contractible space is simply connected, but not conversely. We sill see later that S^n is simply

connected for $n \geq 2$, but it is not contractible (an intuitively clear fact that we will not be able to prove here).

Finally, a construction mentioned above:

Definition 4.11. Let X, Y be (connected, locally path connected) spaces, let $x_0 \in X$ and $y_0 \in Y$. The *one point union* of X and Y, denoted $X \vee Y$, is defined to be $X \vee Y = X \sqcup Y / x_0 \sim y_0$.

Example 4.7. The space $S^1 \vee S^1$ is the "figure 8" that we used in Example 4.6, see Figure 4.14. Our next challenge is to compute $\pi_1(S^1 \vee S^1)$.

5. Some Group Theory

We need to develop some group theory in order to describe fundamental groups of spaces. Recall that we want to compute $\pi_1(S^1)$, the fundamental group of the figure eight. Let a and b denote loops going once around each of the two circles, as in Figure 5.1. One plausible property of its fundamental group is that all loops are obtained, up to homotopy, by composing these any number of times, and that the only cancellations are the obvious ones. Here is a precise algebraic structure that describes this idea:

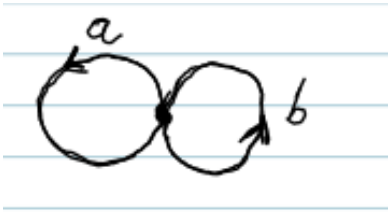


Figure 5.1. $S^1 \vee S^1$

5.1. **Free Groups.** Given two symbols a, b, define a group as follows. Its elements are certain "words" in the "alphabet " a, b, a^{-1}, b^{-1}. A word means a string of these symbols. The empty word is included. A word is called *reduced* if it does not contain any of these sub-strings:

$$aa^{-1}, a^{-1}a, bb^{-1}, b^{-1}b. \tag{5.1}$$

For example, the following are words:

$$bbabb^{-1}a^{-1}b^{-1}b^{-1}aabba^{-1}$$
$$baba^{-1}a^{-1}baabbaabbb$$
$$a^{-1}a^{-1}b^{-1}aaabbababab$$
$$\emptyset$$

where the first is *not reduced*, while the other three are reduced. For each of these words you can trace a loop in $\pi_1(S^1 \vee S^1)$, see Figure 5.1.

Any word can be reduced by erasing any occurence of a substring (5.1) and, if necessary, iterating the process. For example, for the unreduced word above, this process

$$bbabb^{-1}a^{-1}b^{-1}b^{-1}aabba^{-1} \to bbaa^{-1}b^{-1}b^{-1}aabba^{-1} \to bbb^{-1}b^{-1}aabba^{-1}$$

produces a reduced word, called the *reduction* of the original word.

Define a group $F_{a,b}$, called the *free group* on the two generators a, b as follows:

(1) The *elements* of the group are all the reduced words in the alphabet a, b, a^{-1}, b^{-1}, including the empty word.
(2) If $x, y \in F_{a,b}$, their *product* xy is obtained by concatenating the two words x and y and then reducing the concatenation.
(3) The *unit* e is the empty word.
(4) The inverse x^{-1} is obtained by the usual rule: spell x backwards, interchange the letters a, b with a^{-1}, b^{-1}.

Let's look at some examples. iI $x = abaab^{-1}abb$ and $y = baabba^{-1}$, then $xy = abaab^{-1}abbbaabba^{-1}$ while $yx = baabbabaab^{-1}abb$ (reducing the concatenation). In particular $xy \neq yx$, so $F_{a,b}$ is not Abelian. (This can be seen more easily form the fact that $ab \neq ba$ since they are different reduced words.) Since the $abaab^{-1}abbb^{-1}b^{-1}a^{-1}ba^{-1}a^{-1}b^{-1}a^{-1}$ reduces to the empty word, we see that x^{-1}, the inverse of $abaab^{-1}abb$, is $b^{-1}b^{-1}a^{-1}ba^{-1}a^{-1}b^{-1}a^{-1}$ (the usual formula for the inverse).

The group axioms are clear, except for associativity: we will later prove a more general associativity result (Lemma 5.1 below). In the meantime, we will accept that we have defined the free group.

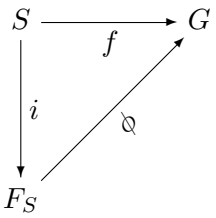
The same reasoning applies to alphabets of any size (which we will keep finite for simplicity):

Definition 5.1. Let $S = \{a_1, \dots a_n\}$ be any finite set. The *free group* on S, denoted F_S or $F_{a_1,\dots,a_n}$, is the collection of all reduced words in the alphabet $a_1 \dots a_n, a_1^{-1}, \dots a_n^{-1}$, including the empty word, and where reduced mean no successive occurrences aa^{-1} of $a^{-1}a$, where a is one of $a_1, \dots a_n$. The operations are defined as for $F_{a,b}$: the product xy is concatenation followed by reduction, the unit e is the empty word, and the inverse x^{-1} is defined by the usual rule.

5.1.1. *The Universal Mapping Property.* Here's a structural characterization of the free group, called a *universal mapping property.*

Theorem 5.1. *Let $S = \{a_1, \dots, a_n\}$, and let $i : S \to F_S$ be the inclusion map: $i(a_j) = a_j$. Suppose G is any group and suppose $f : S \to G$ is a map. Then there is a unique homomorphism $\phi : F_S \to G$ extending f, meaning that $\phi \circ i = f$:*

(5.2)



Proof. Given $f : S \to G$, define ϕ first on the alphabet by

$$\phi(a_j) = f(a_j) \quad \text{and} \quad \phi(a_j^{-1}) = f(a_j)^{-1}, \tag{5.3}$$

where a_j^{-1} is a letter in the alphabet (and, in particular, an element in F_S), and $f(a_j)^{-1}$ is the inverse in the group G of the element $f(a_j) \in G$. Once ϕ is defined on the alphabet, then it can be easily defined on F_S: let $\phi(\emptyset) = e \in G$ and, if $x = x_1 x_2 \dots x_k$ is a word (so each x_l is an a_j or an a_j^{-1}), then define

$$\phi(x_1 \dots x_k) = \phi(x_1) \dots \phi(x_k). \tag{5.4}$$

This defines a homomorphism $\phi : F_S \to G$ because, if $x = x_1 \dots x_k$ and $y = y_1 \dots y_l$ are reduced words, then xy is the reduction of $x_1 \dots x_k y_1 \cdots y_l = x_1 \dots x_{k-j} y_{j+1} \dots y_l$ say (if exactly j cancellations occur in the reduction, that is, $x_k = y_1^{-1}, \dots, x_{k-j+1} = y_j^{-1}$ in F_S, but $x_{k-j} \neq y_{j+1}$. Possibly $j = k = l$ in which case xy is the empty word. Or it may happen that $j = 0$, that is, there are no cancellations). Then

$$\begin{aligned} \phi(xy) = \phi(x_1 \dots x_{k-j} y_{j+1} \dots y_l) &= \phi(x_1) \dots \phi(x_{k-j}) \phi(y_{j+1}) \dots \phi(y_l) \\ &= \phi(x_1) \cdots \phi(x_k) \phi(y_1) \dots \phi(y_l) \ = \phi(x)\phi(y), \end{aligned}$$

where the second equality is the definition (5.4) of ϕ and the third equality is because the same cancellations, if any, $x_k y_1 = e$, etc, that occur in F_S imply $\phi(x_k)\phi(y_1) = e$, etc, in G. This proves the *existence* of ϕ.

To prove the *uniqueness* of ϕ, observe first that any homomorphism extending i must satisfy (5.3) and therefore must also satisfy (5.4). □

Example 5.1.

(1) If $S = \{a\}$ has cardinality one, then F_a is the infinite cyclic group generated by a, The universal mapping property says that to define, for any group G, a homomorphism $F_a \to G$, is equivalent to stating where the generator a goes: choose any $g \in G$, let $\phi(a) = g$, then $\phi(a^n) = g^n \in G$ for any $n \in \mathbb{Z}$.

(2) To see an example were the universal mapping property does *not* hold, look at any finite cyclic group, say the group $\mathbb{Z}_{10}$ of order 10 generated by 1, and let $G = \mathbb{Z}$. If we look at the map $f : \{1\} \to \mathbb{Z}$ defined by $f(1) = 1$, then f does not extend to a homomorphism

$\phi : Z_{10} \to \mathbb{Z}$, since we would have $\phi(2) = 2, \phi(3) = 3, \dots \phi(10) = 10$. But $10 = 0$ in $\mathbb{Z}_{10}$, yet $10 \neq 0$ in $\mathbb{Z}$, so we would get $\phi(0) = 0$ and $\phi(0) = 10$, so ϕ is not defined. The problem is that the relation $10 = 0$ holds in $\mathbb{Z}_{10}$ but does not hold in $\mathbb{Z}$. The meaning of the word *free* is that there are no relations other than the ones imposed by the axioms of group theory. One way of making the meaning of "no relations" precise is the universal mapping property (5.2).

(3) In (5.2), let's take $G = \mathbb{Z}^n$ and $f(a_j) = (0, \dots, 1, \dots 0)$ (the 1 in the j-th posiition). Then the resulting homomorphism $\phi : F_{a_1,\dots,a_n} \to \mathbb{Z}^n$ is given by

$$\phi(x_1 \dots x_k) = (m_1, m_2, \dots, m_n) \tag{5.5}$$

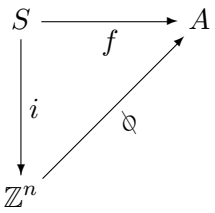
where m_j is the number of occurrences of a_j among the letters $x_1, \dots, x_k$ minus the number of occurrences of a^{-1} among the same letters. This can be seen easily from the fact that all the elements $\phi(x_l)$ commute, this formula holds for each letter and is additive.

(4) Another example of a group that is not free is the group $\mathbb{Z}^2$. Take for G any non-abelian group, say the symmetric group group S_3 on three letters, and any two elements $g, h \in G$ so that $gh \neq hg$, say the transpositions $g = (12)$ and $h = (23)$ in S_3. Define a map $f : \{(1,0), (0,1)\} \to G$ by $f(1,0) = g$ and $f(0,1) = h$ (the standard generators for $\mathbb{Z}^2$). Then f does not extend to a homomorphism $\phi : \mathbb{Z}^2 \to G$ because $\phi(1,1)$ is forced to have two contradictory definitions: $\phi(1,1) = \phi(1,0)\phi(0,1) = gh$ and $\phi(1,1) = \phi(0,1)\phi(1,0) = hg$, but $gh \neq hg$. The problem here is that relations of commutativity hold in $\mathbb{Z}^2$, so $\mathbb{Z}^2$ is not free. These relations go beyond the requirements of group theory, as the existence of non-abelian groups such as S_3 or $F_{a,b}$ show.

(5) The last example shows that $\mathbb{Z}^2$ is not a free group. But, if we stay in the restricted class of Abelian groups, then it, as well as $\mathbb{Z}^n$ for any $n = 1, 2, \dots$ has a freeness property expressed by a universal mapping property similar to (5.2). One way of stating the theorem is to say that $\mathbb{Z}^n$ is *free in the category of Abelian groups*, meaning that it satisfies no relations beyond those dictated by group theory *and commutativity.*

Theorem 5.2. *Let $S = \{a_1, \dots, a_n\}$ be the standard basis for $\mathbb{Z}^n$ as in (3) of Example 5.1, and let $i : S \to \mathbb{Z}^n$ be the inclusion. Then, given any Abelian group A and any map $f : S \to A$, there is a unique group homomorphism $\phi : \mathbb{Z}^n \to A$ extending f, meaning that $\phi(a_j) = f(a_j)$ for $j = 1, \dots n$:*

(5.6)



Proof. A simpler version of that of Theorem 5.1, left as an exercise. □

5.2. **Free Products of Groups.** We need another construction in group theory. Suppose $G_1, \dots G_n$ are groups. We will define a new group as follows:

Definition 5.2. The *free product* of the groups $G_1, \dots G_n$ is the group, denoted $G_1 * \cdots * G_n$,

(1) Its elements are reduced words in the alphabet which is the *disjoint union* $(G_1 \setminus \{e\}) \sqcup \cdots \sqcup (G_n \setminus \{e\})$, including the empty word.
   (a) A word means a string of symbols $x_1 x_2 \dots x_k$ where each x_i belongs to one of the sets $G_j \setminus \{e\}$.
   (b) Note that the indexing by j is essential to the meaning of disjoint union in the definition of the alphabet. We distinguish G_1 and G_2, etc, if they have different indices, even if they are the same group. We may (and will) use products $G * G * \dots$.
   (c) A word is *reduced* if two successive letters x_i and x_{i+1} never belong to the same $G_j \setminus \{e\}$, meaning, as just explained, to the same index j.
   (d) The *length* of the reduced word $x = x_1 \dots x_k$ is defined to be k, and we write $l(x) = k$. The length of the empty word is defined to be 0. Note that $l(x) = 1$ if and only if x is a letter in the alphabet.

(2) The product xy of $x = x_1 \dots x_k$ and $y = y_1 \dots y_l$ is defined by

$$xy = \begin{cases} x_1 \dots x_k y_1 \dots y_l & \text{if } x_k, y_1 \notin \text{ same } G_j, \\ x_1 \dots x_{k-1}(x_k y_1) y_2 \dots y_l & \text{if } x_k, y_1 \in \text{same} G_j \\ & \text{ and } x_k y_1 \neq e, \\ (x_1 \dots x_{k-1})(y_2 \dots y_l) & \text{if } x_k, y_1 \in \text{same } G_j \\ & \text{ and } x_k y_1 = e. \end{cases} \tag{5.7}$$

Note that in the second case the product $x_k y_1$ takes place in the group G_j and produces a letter in the alphabet. Thus in the first case the length $l(xy) = k + l$, in the second case $l(xy) = k + l - 1$. In the third case the definition of xy is reduced to the definition of the product of two words x', y' where $l(x') < l(x)$ and $l(y') < l(y)$, so we could take this as an inductive definition of xy. Of course, in the third case, it means continue applying case 1 or 2 as needed until you get a reduced word.

(3) The *unit* e is the empty word.
(4) The inverse x^{-1} is defined by the usual formula: if $x = x_1 \dots x_k$, then $x^{-1} = x_k^{-1} \dots x_1^{-1}$.

To know that we have indeed defined a group all that is needed is the following lemma:

Lemma 5.1. *The multiplication just defined is associative:* $(xy)z = x(yz)$ *for all* $x, y, z \in G_1 * \cdots * G_n$.

Proof. We sketch a proof by induction on the length $l(y)$, see Proposition 12.5 of Chapter I of [8] for more details. For $l(y) = 0$ there's nothing to prove, and for $l(y) = 1$ it is a quick check using (5.7), there are seven cases to check, left as an exercise. If $l(y) > 1$, then can write $y = y_1 y_2$ where $l(y_1) < l(y)$ and $l(y_2) < l(y)$. Then all the following computations involve associativity with the middle element y_1 or y_2 of strictly smaller length: $(xy)z = (x(y_1y_2))z = ((xy_1)y_2)z = (xy_1)(y_2z) = x(y_1(y_2z)) = x((y_1y_2)z) = x(yz)$. □

Example 5.2. (1) The free group $F_{a_1,\dots,a_n}$ of Definition 5.1 is isomorphic to the free product $\mathbb{Z} * \cdots * \mathbb{Z}$ (n factors). More precisely, $F_{a_1,\dots,a_n}$ is the same as $F_{a_1} * \cdots * F_{a_n}$, recalling, from Example 5.1 that F_a is the infinite cyclic group generated by a, isomorphic to $\mathbb{Z}$. In this description we are using a larger alphabet: $\{a_1^k : k \in \mathbb{Z}\setminus\{0\}\} \sqcup \cdots \sqcup \{a_n^k : k \in \mathbb{Z} \setminus \{0\}\}$ rather than the earlier alphabet $\{a_1, \dots, a_n, a_1^{-1} \dots, a_n^{-1}\}$.

(2) The smallest non-trivial free product that can be formed is $\mathbb{Z}_2 * \mathbb{Z}_2$, the free product of two groups of order 2. Let's call their generators a and b respectively. Then the alphabet is $\{a, b\}$ and, since $a^2 = e$ and $b^2 = e$, we can easily list all the elements of the group according to their lengths: e (the empty word, length 0), a, b (of length 1), ab and ba (of length 2), aba and bab (length 3), $abab$ and $baba$ (length 4), and so on. There are exactly two elements of any given length $l \geq 1$, the element is completely determined by the first letter. It is easy to check that all elements of odd length have order 2, while all elements of even length have infinite order.

There is a good way to picture this group as what is called the *infinite dihedral group* D_∞, see Figure 5.2. This is the group of motions of $\mathbb{R}$ generated by two reflections α, β,where $\alpha(x) = -x$ (reflection in 0) and $\beta(x) = 1 - x$ (reflection in $\frac{1}{2}$). Then $\alpha\beta(x) = -(1-x) = x-1$ is translation by -1 and $\beta\alpha(x) = 1-(-x) = x+1$ is translation by 1. There is a homomorphism $\phi : \mathbb{Z}_2 * \mathbb{Z}_2 \to D_\infty$ defined by $\phi(a) = \alpha$ and $\phi(b) = \beta$, illustrating the universal mapping property of free products in Theorem 5.3 below. It is easy to check that ϕ is an isomorphism, with elements of odd length going to reflections in integers or half-integers, and elements of even length

to translations by an integral amount. These are exactly all the elements of D_∞.

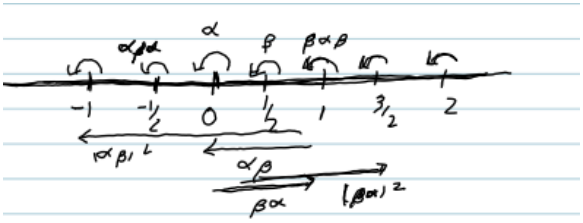


FIGURE 5.2. The Infinite Dihedral Group $D_\infty \cong \mathbb{Z}_2 * \mathbb{Z}_2$

(3) The last example illustrates the fact that non-trivial (meaning at least two factors, both with mote than one element) free products always give infinite groups. The next smallest example would be $\mathbb{Z}_2 * \mathbb{Z}_3$. This is already quite complicated. If we let a generate $\mathbb{Z}_2$ and b generate $\mathbb{Z}_3$, we have e of length 0, a, b, b^2 of length 1, ab, ab^2, ba, b^2a of length 2, $aba, ab^2a, bab, bab^2, b^2ab, b^2ab^2$ of length 3, etc. We should meet this group again in relation to hyperbolic geometry.

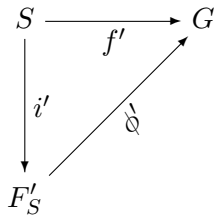
Theorem 5.3. *Let $G_1, \dots G_n$ and G be groups, and suppose given homomorphisms $\phi_j : G_i \to G$ for $j = 1, \dots n$. Let $i_j : G_j \to G_1 * \cdots * G_n$ be the inclusion of each factor in the free product. Then there exists a unique homomorphism $\phi : G_1 * \cdots * G_n \to G$ extending the ϕ_j, meaning that, for each j, $\phi \circ i_j = \phi_j$.*

Proof. This is just like the proof of Theorem 5.1. If $x = x_1 \dots x_k$ is a reduced word and $x_l \in G_{j_l} \backslash \{e\}$, let $\phi(x) = \phi_{j_1}(x_1) \dots \phi_{j_k}(x_k)$ (and $\phi(\text{empty word}) = e$). As before this is well defined and gives the unique extension of $\phi_1 \dots \phi_n$ to $G_1 * \cdots * G_n$. □

Finally one useful fact about the universal mapping properties that we have discussed: they uniquely define the structure of the group in question, including the the inclusion of the distinguished subsets or subgroups.

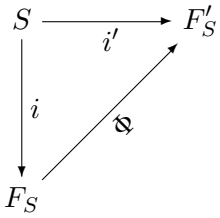
Theorem 5.4. *Fix one of the universal mapping properties of Theorem 5.1, or 5.2 or 5.3. Given any two groups that property, there is a unique isomorphism between them that takes one inclusion to the other. In the case of Theorem 5.1 this means the following: Suppose F'_S is another group such that there is an inclusion $i' : S \to F'_S$ with the property that for any group G and any map $f' : S \to G$ there is a unique homomorphism $\phi' : F'_S \to G$ extending f':*

(5.8)



Then there exists a unique isomorphism $\Phi : F_S \to F'_S$ *so that* $\Phi \circ i = i'$*:*

(5.9)



Similar statements hold for free abelian groups (Theorem 5.2) and free products of groups (Theorem 5.3).

Proof. In the case of the free group, take $G = F'_S$ in (5.2), to get a unique $\Phi : F_S \to F'_S$ as in (5.9). Then take $G = F_S$ in (5.8) to get $\Phi' : F'_S \to F_S$ as in (5.9) but with Φ' in the opposite direction from Φ. Then $\Phi' \circ \Phi : F_S \to F_S$ and $\Phi'(\Phi(i(a))) = \Phi'(i'(a)) = i(a)$ for all $a \in S$, so $\Phi' \circ \Phi$ and id are both homomorphisms of F_S to itself that are the identity on $i(S)$ (in other words, satisfy (5.2) with $G = F_S$ and $f = i$), so, by uniqueness, $\Phi' \circ \Phi = id$. Similarly $\Phi \circ \Phi' = id$, so we have the desired isomorphism. The other cases are similar. □

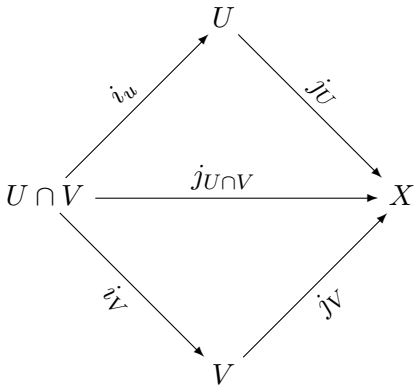
6. Van Kampen's Theorem

Now we state and prove a theorem that allows us to compute the fundamental group of a union of spaces. As usual, X is a connected, locally path connected topological space.

Theorem 6.1. *Suppose* $X = U \cup V$ *where* $U, V \subset X$ *are open, and* U*,* V *and* $U \cap V$ *are* connected and non-empty*. Fix a point* $x_0 \in U \cap V$*. The*

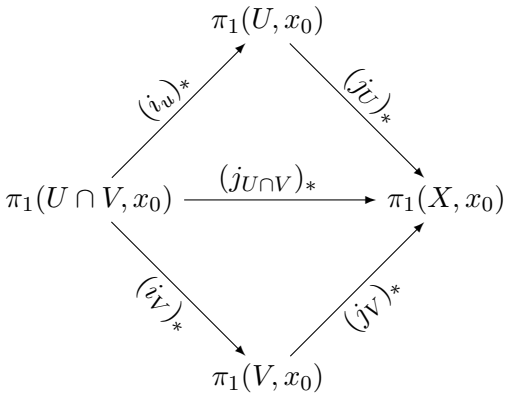
diagram of topological spaces and continuous maps (in this case inclusions)

(6.1)



gives a diagram of fundamental groups and induced homomorphisms:

(6.2)



Then

(1) *The homomorphism of the free product* $\phi : \pi_1(U, x_0) * \pi_1(V, x_0) \to \pi_1(X, x_0)$ *given by Theorem 5.3 is surjective.*
(2) *The kernel* K *of* ϕ *is the smallest normal subgroup of* $\pi_1(U, x_0) * \pi_1(V, x_0)$ *that contains* $\{((i_U)_*\alpha)((i_V)_*\alpha)^{-1} : \alpha \in \pi_1(U \cap V, x_0)\}$.

Remark 6.1. Recall the definition of the homomorphism ϕ from Theorem 5.3: If $\alpha_1 \dots \alpha_k$ is a reduced word representing an element of $\pi_1(U, x_0) * \pi_1(V, x_0)$. then $\phi(\alpha_1 \dots \alpha_k) = \phi_1(\alpha_1) \dots \phi_k(\alpha_k)$ where

$$\phi_i = \begin{cases} (j_U)_* & \text{if } \alpha_i \in \pi_1(U, x_0), \\ (j_V)_* & \text{if } \alpha \in \pi_1(V, x_0). \end{cases}$$

Since we know from Theorem 5.3 that ϕ is well-defined, non-reduced words coud be used as well.

For simplicity, let $G = \pi_1(U, x_0) * \pi_1(V, x_0)$. Since the first assertion says that $\phi : G \to \pi_1(X, x_0)$ is surjective, we must have $\pi_1(X, x_0) = G/K$, where $K = ker(\phi)$, in other words, $\pi_1(X, x_0)$ is a quotient group of G. Diagram 6.2

gives that for all $\alpha \in \pi_1(U \cap V, x_0)$, $(j_U)_*(i_U)_*(\alpha) = (j_V)_*(i_V)_*(\alpha)$ (both equal to $(j_{U\cap V})_*(\alpha)$). Therefore

$$\phi((i_U)_*(\alpha)) = \phi((i_V)_*(\alpha)) \text{ for all } \alpha \in \pi_1(U \cap V, x_o). \tag{6.3}$$

In other words, $\pi_1(X, x_0)$ is a quotient group of G in which (6.3) holds. This is the same as saying that the following set S is contained in K:

$$S = \{((i_U)_*(\alpha))((i_V)_*(\alpha))^{-1} : \alpha \in \pi_1(U \cap V, x_0)\} \subset K. \tag{6.4}$$

The second assertion of the theorem says that K is *as small as possible given the constraint* (6.3) (which is equivalent to the constraint (6.4)). Equivalently, this says that $G/K = \pi_1(X, x_0)$ is *as large as possible given the constraint* (6.3) (equivalently, to constraint (6.4)).

The phrase "K is as small as possible" has he usual meaning: if N is any normal subgroup of G containing S, then $K \subset N$. Equivalently, K is the intersection of all normal subgroups of G containing S. Since K and G/K are , so to speak, "inversely proporional", this is also a reasonable definition of the phrase "G/K is as large as possible", and in many situations this is the best definition of this phrase. But in some situations it may be desirable to have a definition that does not mention K explicitly, so it is reasonable to use the following definition of "as large as possible":

Suppose H is any group such that there is a surjective homomorphism $\psi : G \to H$ satisfying the same constraint (6.3) that ϕ satisfies. Then there is a surjective homomorphism $f : \pi_1(X, x_0) \to H$ so that $f \circ \phi = \psi$. The equivalence of the two formulations is easy to see: if H is as stated, then $H = G/N$ where $N = ker(\psi)$ is normal subgroup of G containing S, hence containing K, and f is the natural map $G/K \to G/N$. This formulation is also equivalent to the universal mapping property used by Massey on p. 114 of [9] to formulate Van Kampen's theorem, while the formulation we have chosen is as used by Hatcher on p. 43 of [5].

Proof of Van Kampen's Theorem. For the first statement, let $\alpha : I \to X$ be a loop based at x_0. We must show that $[\alpha]$ is in the image of ϕ. Let $\epsilon > 0$ be a Lebesgue number for the cover $\alpha^{-1}(U), \alpha^{-1}(V)$ of I, and let $0 = t_0 < t_1 < \cdots < t_n = 1$ be a partition of I so that $t_i - t_{i-1} < \epsilon$ for all i. For each $i = 1, \ldots, n$ choose $W_i = U$ or V so that $\alpha([t_{i-1}, t_i]) \subset W_i$. Since $W_i \cap W_{i+1}$ is, by assumption, connected, and contains x_0, there is a path β_i from x_0 to $\alpha(t_i)$ in $W_i \cap W_{i+1}$, for $i = 1, \ldots, n-1$. Let α_i be the restriction of α to $[t_{i-1}, t_i]$, linearly reparametrized to $[0, 1]$. Then we can write

$$[\alpha] = [\alpha_1 \cdot \beta_1^{-1}] \cdot [\beta_1 \cdot \alpha_2 \cdot \beta_2^{-1}] \cdots [\beta_{n-1} \cdot \alpha_n] = [\gamma_1] \cdots [\alpha_n] \tag{6.5}$$

where γ_i is the loop $\beta_{i-1}^{-1} \cdot \alpha_i \cdot \beta_i$ in W_i (defined by one of the two possible associations, the choice is irrelevant), see Figure 6.1

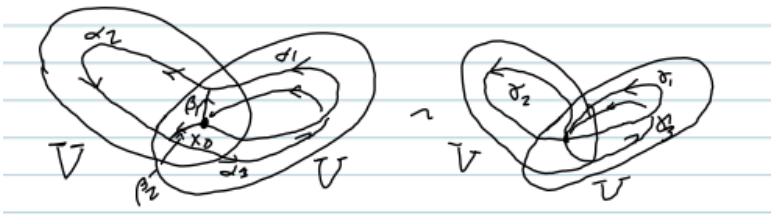


FIGURE 6.1. Surjectivity of $\pi_1(U, x_0) * \pi_1(V, x_0) \to \pi_1(X, x_0)$.

This equation can be read in two ways. If we consider the $[\gamma_i] \in \pi_1(W_i, x_0)$, let's denote it $[\gamma_i]_{W_i}$ in this case, then $[\gamma_1]_{W_1} \cdots [\gamma_n]_{W_n}$ is a word, not necessarily reduced, whose reduction represent an element g of the free product $\pi_1(U, x_0) * \pi_1(V, x_0)$. If we interpret the $[\gamma_i]$ as elements of $\pi_1(X, x_0)$, let's denote them $[\gamma_i]_X$ in this case, then, by definition of ϕ, $[\alpha_i]_X = \phi([\gamma_i]_{W_i})$, and (6.5) then reads $[\alpha]_X = \phi(g)$, thus $[\alpha]$ is in the image of ϕ and the first part of the theorem is proved.

The proof of the second statement is much more involved. We give a quick sketch and refer to [5, 9] for more details. To have a reasonable notation, let

$$G_1 = \pi_1(U, x_0),\ G_2 = \pi_1(V, x_0),\ A = \pi_1(U \cap V, x_0),$$

and

$$i_1 : A \to G_1, \quad i_2 : A \to G_2$$

the induced homomorphisms on fundamental groups. Let

$$S = \{((i_1)_* a)((i_2)_* a)^{-1} : a \in A\}$$

and

$$N = \text{smallest normal subgroup of } G_1 * G_2 \text{ containing } S.$$

Let $\phi : G_1 * G_2 \to \pi_1(X, x_0)$ be the surjective homomorphism given by the first parr. We want to show that if $a_1 \dots a_k \in G_1 * G_2$ and $\phi(a_1 \dots a_k) = e$, then $a_1 \dots a_k \in N$. It will be more convenient to prove the equivalent statement: if $\phi(a_1 \dots a_k) = e$, then $\psi(a_1 \dots a_k) = e$, where $\psi : G_1 * G_2 \to G_1 * G_2/N$ is the natural projection.

We need to know some ways of changing the word $a_1 \dots a_k$ without changing its image $\psi(a_1 \dots a_k)$. Here are two ways:

Move I: In $a_1 \dots a_j \dots a_k$, replace a_j by $a'_j a''_j$ if a_j, a'_j, a''_j are in the same group G_j ($= G_1$ or G_2) and $a_j = a'_j a''_j$ in G_j, and also its inverse operation. This move one does not change the element $a_1 \dots a_k \in G_1 * G_2$.

Move II: In $a_1 \dots a_j \dots a_k$, if $a_j \in G_1$ and $a_j = i_1 a$ for some $a \in A$, replace a_j by $a'_j = i_2 a \in G_2$ (and similar operation if $a_j \in G_2$ and $a_j = i_2 a$). This operation may change the element $a_1 \dots a_k \in G_1 * G_2$, but does not change $\psi(a_1 \dots a_k) \in G_1 * G_2/N$.

So take a word $a_1 \dots a_k \in G_1 * G_2$ so that $\phi(a_1 \dots a_k) = e$. Each a_j is represented by a loop (also called a_j) in U or V so that $a_1 \cdots a_k$ is homotopic to x_0. Represent this homotopy class by a loop $a_1 \dots a_k$ where $a_j : [\frac{j-1}{k}, \frac{j}{k}] \to X$. Let $F : I \times I \to X$ be a homotopy of this loop to x_0:

$$F(t,0) = x_0, \quad F(t,1) = a_1 \cdots a_k(t), \quad F(0,s) = F(1,s) = x_0.$$

By the usual Lebesgue number argument plus a bit n=more refinement, we can find subdivisions $0 = t_0 < t_1 < \cdots < t_m = 1$ and $0 = s_0 < s_1 < \cdots < s_n = 1$ of I so that:

(1) The partition $t_0, t_1, \dots, t_m$ of I refines the partition $0, \frac{1}{k}, \frac{2}{k}, \dots, 1$ used to define the loop $a_1 \cdots a_k$.
(2) Let $R_{i,j} = [t_{i-1}, t_i] \times [s_{j-1}, s_j]$. Then for all i, j, $F(R_{i,j}) \subset W_{ij}$, where $W_{ij} = U$ or V.

Let $v_{i,j} = F(t_i, s_j)$, and let $V_{i,j}$ be the intersection of the sets $W_{k,l}$ containing $v_{i,j}$, so $V_{i,j}$ is one of $U, V, U \cap V$. Define paths $a_{i,j}, b_{i,j}$ to be the images under F of the sides of the rectangles as in Figure 6.2.

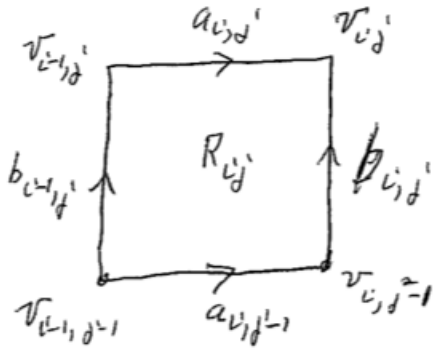


FIGURE 6.2

Then in $W_{i,j}$we have a homotopy relative endpoints:

$$b_{i-1,j} \cdot a_{i,j} \sim a_{i,j-1} \cdot b_{i,j} \quad \text{(rel endpoints)}. \tag{6.6}$$

To convert this to an equation involving loops, choose a path $\gamma_{i,j}$ *lying in* $V_{i,j}$ from x_0 to $v_{i,j}$. Then to the a's and b's we can assign loops α, β by

$$\alpha_{i,j} = \gamma_{i-1,j} \cdot a_{i,j} \cdot \gamma_{i,j}^{-1} \quad \text{and} \quad \beta_{i,j} = \gamma_{i,j-1} \cdot b_{i,j} \cdot \gamma_{i,j}^{-1}.$$

Then (6.6) implies the equation

$$[\beta_{i-1,j}\ \alpha_{i,j}]_{W_{i,j}} = [\alpha_{i,j-1}\ \beta_{i,j}]_{W_{i,j}}, \tag{6.7}$$

the subscript $W_{i,j}$ meaning that the equality takes place in the group $\pi_1(W_{i,j}, x_0)$, which we recall is one of G_1, G_2.

Let's look at the subdivision of $I \times I$. At the bottom edge we have the element $\alpha_{1,0}\alpha_{2,0}\dots\alpha_{m,0} = e$ in $G_1 * G_2$ since each $\alpha_{i,0}$ is the constant path x_0. The last letter $\alpha_{m,0}$ of this word is the same as $\alpha_{m,0}\beta_{m,1}$ since $\beta_{m,1}$ is also the constant path x_0. Looking at (6.7) for the rectangle $R_{m,1}$ (the bottom right hand corner of the subdivision), we get $[\beta_{m-1,1}\alpha_{m,1}]_{W_{m,1}} = [\alpha_{m,0}\beta_{m,1}]_{W_{m,1}}$, which means that, by moves of type I, we can replace the lower edge $\alpha_{1,0}\dots\alpha_{m,0}$ by the path $\alpha_{m,0}\dots\alpha_{m-1,0}\beta_{m-1,1}\alpha_{m,1}$ from the bottom left hand corner of the subdivision to the vertex on the right edge one above the bottom, as pictured in the left of Figure 6.3. Since it only involves Move I, this is an equality in $G_1 * G_2$. (note that $\beta_{m-1,1}\alpha_{m,1} = e$, and all elements with $j = 0$ or $i = m$ are also e since they are the constant loop at x_0).

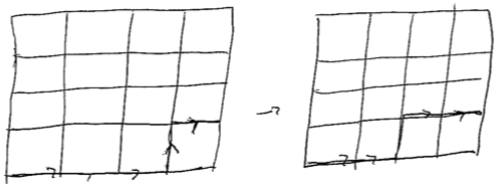


FIGURE 6.3

Now we proceed to the left. We look at $[\beta_{m-1,1}]_{W_{m,1}}$, replace it by $[\beta_{m-1,1}]_{W_{m-1,1}}$, which is a move of type I if $W_{m,1} = W_{m-1,1}$ but is a move of type II otherwise. Then we use the relation (6.7) applied to $R_{m-1,1}$ to "move the path over this rectangle", that is, replace $\alpha_{m,0}\dots\alpha_{m-1,0}\beta_{m-1,1}\alpha_{m,1}$ by $\alpha_{m,0}\dots\alpha_{m-2,0}\beta_{m-2,1}\alpha_{m-1,1}\alpha_{m,1}$ (pictorially, move the β one unit to the left), as shown in Figure 6.3. Since only moves of type I and II are involved, this doesn't change the element of $G_1 * G_2/N$.

Continue this way until you reach the left-hand edge, then move to the next layer on the right hand edge by using $\beta_{m,2} = e$ since again it is the constant loop at x_0, then move to the left as before by "lifting the path" over each rectangle. The general step requires a replacement as shown in Figure 6.4, namely moves of (possibly) type II

$$[\beta_{i,j}]_{W_{i+1,j}} \leftrightarrow [\beta_{i,j}]_{W_{i,j}} \quad \text{and} \quad [\alpha_{i,j-1}]_{W_{i,j-1}} \leftrightarrow [\alpha_{i,j-1}]_{W_{i,j}},$$

followed then by the move (6.7) of type I. Continue this way until we cover the whole subdivision of $I \times I$. Since these moves do not change the element of $G_1 * G_2/N$, we get that the top edge of the rectangle represents the same element of $G_1 * G_2/N$ as the bottom edge:

$$e = \alpha_{1,0}\dots\alpha_{m,0} = \alpha_{1,n}\dots\alpha_{m,n} \quad \text{in} \quad G_1 * G_2/N.$$

Finally the condition above that the partition $t_0, t_1, \dots t_n$ refines the partition $0, \frac{1}{k}\dots 1$ says that $\alpha_{1,n}\dots\alpha_{m,n}$ is a finer factorization of the original

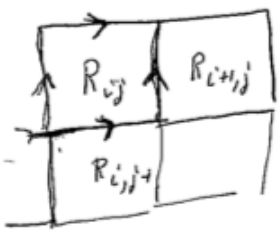


FIGURE 6.4

$a_1 \dots a_k$, thus equivalent to it by moves of type I. Thus we have proved that $a_1 \dots a_k \in N$, in other words, $ker(\phi) \subset N$, concluding the proof of the second statement.

□

6.1. **Applications of Van Kampen's Theorem.** We are now in the position to compute many more examples of fundamental groups. We start with a simple example that only uses the first statement of Van Kampen's theorem.

Example 6.1. *The unit sphere $S^n \subset \mathbb{R}^{n+1}$ is simply connected for $n \geq 2$.*

Proof. Let $U_+ = \{x_{n+1} > -\frac{1}{2}\}$ and $U_- = \{x_{n+1} < \frac{1}{2}\}$ and let $U = U_+ \cap S^n$, $V = U_- \cap S^n$. It is easy to check that, for any $n \geq 1$, U and V are both homeomorphic to an n- disk, in particular, both are simply connected, and $U \cap V$ has the equator $\{(x_1, \dots, x_n, 0) : x_1^2 + \dots x_n^2 = 1\} = S^{n-1}$ as a deformation retract, see Figure 6.5 Since, *for* $n \geq 2$, S^{n-1} is connected, $U \cap V$ is connected, and the first part of Van Kampen's theorem gives a surjection of the trivial group $\{e\} * \{e\}$ to $\pi_1(S^n)$, thus S^n is simply connected for $n \geq 2$. □

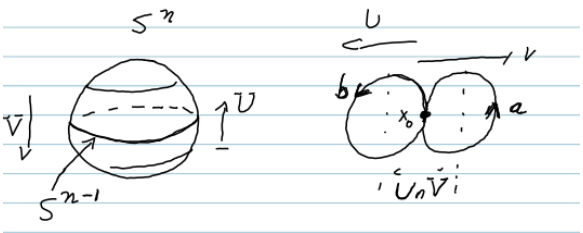


FIGURE 6.5. $\pi_1(S^2)$ and $\pi_1(S^1 \vee S^1)$.

The principle behind this example is the following immediate corollary of Van Kampen's theorem:

Corollary 6.1. *Suppose $X = U \cup V$ where U, V are open and simply connected, and $U \cap V$ is non-empty and connected. Then X is simply connected.*

Proof. The first part of Van Kampen's theorem gives a surjection of the trivial group to $\pi_1(X)$. □

Example 6.2. *$\pi_1(S^1 \vee S^1, x_0) = \mathbb{Z} * \mathbb{Z}$, the free group on two generators $F_{a,b}$, where a, b are loops going once around each of the two circles.*

Proof. Let's take $S^1 \vee S^1$ to be the union X of two circles $C_1, C_2 \subset \mathbb{R}^2$ of radius 1 and center at $(-1, 0)$ and $(0, 1)$ respectively, and $x_0 = (0, 0)$ their intersection. Let $U = X \cap \{x < 1\}$ and $V = X \cap \{x > -1\}$. Then U, V and $U \cap V$ satisfy the hypotheses of Van Kampen's theorem, U has C_1 as a deformation retract, V has C_2 as a deformation retract, and $U \cap V$ has x_0 as deformation retract, see Figure 6.5. Therefore $\pi_1(U, x_0) = \pi_1(V, x_0) = \mathbb{Z}$ and $\pi_1(U \cap V) = \{e\}$, therefore $\mathbb{Z} * \mathbb{Z}$ surjects to $\pi_1(X, x_0)$ with trivial kernel, in other words, $\pi_1(X, x_0) = \mathbb{Z} * \mathbb{Z}$. □

The principle behind this example is the following corollary of Van Kampen's theorem:

Corollary 6.2. *Suppose $X = U \cup V$ where U, V are open, connected, $U \cap V$ is non-empty, (connected) and simply connected, and let $x_0 \in U \cap V$. Then $\pi_1(X, x_0) = \pi_1(U, x_0) * \pi_1(V, x_0)$.*

Proof. Since $\pi_1(U \cap V, x_0) = \{e\}$, the surjective map $\pi_1(U, x_0) * \pi_1(V, x_0) \to \pi_1(X, x_0)$ is also injective. □

Of course, to use this Corollary effectively in examples such as $\pi_1(S^1 \vee S^1)$ we have to choose open sets of the correct homotopy type. We would like to decompose $S^1 \vee S^1$ as the union of the two obvious circles, but these are not open. We choose slightly bigger open sets of the same homotopy type, and with their intersection of the homotopy type of x_0, and then apply the theorem. This type of adjustment is typical of the applications of Theorem 6.1.

Example 6.3. We could use the same reasoning, inductively, to show the following: Let $S^1 \vee \cdots \vee S^1$ (n times) be the one-point union of n circles, in other words, choose a point x_i in each S^1 and define

$$X = S^1 \vee \cdots \vee S^1 = S^1 \sqcup \cdots \sqcup S^1 / (x_1 \sim x_2 \sim \cdots \sim x_n).$$

Then $\pi_1(X) = \mathbb{Z} * \cdots * \mathbb{Z} = F_{a_1, \ldots, a_n}$, the free group on n generators (see Definition 5.1), where a_i is a loop going once around the i-th circle.

Example 6.4. Let P^2 be the projective plane (see (6) and (7) of Example 1.1). *Then $\pi_1(P^2) = \mathbb{Z}_2$, the cyclic group of order two.*

Proof. Present P^2 as the close unit disk D with antipodal points on its boundary S^1 identified. Divide the boundary into two semicircles from $(-1,0)$ to $(0,1)$, label them a. Then, in the notation for presenting surfaces explained in (9) of Example 1.1, P^2 is presented as aa. The half-circle a projects to a loop in P^2. Cover P^2 by two open sets: $U =$ the projection to the quotient space of the annulus $\frac{1}{4} < |x| \leq 1$ and $V =$ projection to quotient space of $|x|, \frac{3}{4}$. Then U is an open Móbius band which deformation retracts to the circle which is the projection of a, and V is an open disk. $U \cap V$ is an annulus that deformation retracts to the circle $|x| = \frac{1}{2}$. Then U and $U \cap V$ are both homotopy equivalent to a circle, thus have infinite cyclic fundamental group, isomorphic to $\mathbb{Z}$, while V is simply connected. We need to take Take the base-point at the point $x_0 = (\frac{1}{2}, 0) \in U \cap V$. Then, by sending a path c from x_0 to $(1,0)$ we can take for generator of $\pi_1(U, x_0)$ the loop cac^{-1} and for generator of $\pi_1(U \cap V)$ the loop b formed by the circle $|x| = \frac{1}{2}$ counterclockwise. Then it is easy to see (and checked in a homework problem) that the homomorphism induced by inclusion sends b to $c \cdot a^2 \cdot c^{-1}$, see Figure 6.6. Theorem 6.1 tells us that $\pi_1(P^2)$ is the quotient of $\pi_1(U, x_0) * \pi_1(V, x_0) = \mathbb{Z} * \{e\} = \mathbb{Z}$ buy the smallest normal subgroup containing twice the generator. Since $\mathbb{Z}$ is abelian, every subgroup is normal, so this is the same as the subgroup $2\mathbb{Z}$, so $\pi_1(P^2) = \mathbb{Z}/2\mathbb{Z} = \mathbb{Z}_2$ as asserted.

□

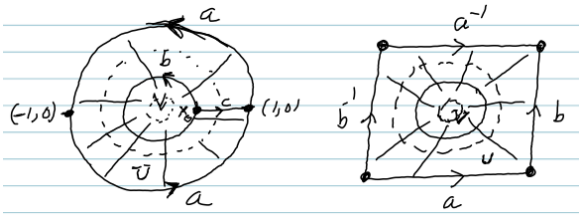


FIGURE 6.6. $\pi_1(P^2)$ and $\pi_1(T^2)$

Remark 6.2. To see shortly a general pattern, it is more suggestive to express the answer in terms of the generator a of $\pi_1(U, (1,0))$: $\pi_1(P^2) =< a > / < a^2 >$ were $< c >$ stands for *the cyclic group generated by c*. All we need to do is to change the base-point to $(1,0)$, the image of $\pi_1(U \cap V, x_0)$ then goes to the subgroup generated by $c^{-1} \cdot b \cdot c = c^{-1} \cdot c \cdot a^2 \cdot c^{-1} \cdot c = a^2$.

Example 6.5. We already know that $\pi_1(T^2) = \mathbb{Z}^2$, see Example 4.4. Let's see that Theorem 6.1 gives the same result. Present T^2 as a quotient of the unit square with identification $aba^{-1}b^{-1}$ and cover T^2 by $U =$ neighborhood of the boundary, and $V =$ interior of the square. Then U deformation retracts to a space homeomorphic to $S^1 \vee S^1$ formed by the two loops a, b in T^2, while V is simply connected, and $U \cap V$ is homotopy equivalent to a circle. So the groups in question in Diagram 6.2 are $\mathbb{Z} = \pi_1(U \cap V, x_0)$,

$F_{a,b} = \mathbb{Z}*\mathbb{Z} = \pi_1(U)$, and $\pi_1(V)$ is trivial. Then, after an argument changing base-points as in the last remark, we see that the image of $\pi_1(U \cap V)$ in $\pi_1(U) = F_{a,b}$ is the subgroup generated by $aba^{-1}b^{-1}$, see Figure 6.6. Thus Theorem 6.1 tells us that $\pi_1(T^2) = F_{a,b}/N$ where N is the smallest normal subgroup of $F_{a,b}$ containing the word $aba^{-1}b^{-1}$. This subgroup N is called the *commutator subgroup* of $F_{a,b}$, and we will study these subgroups shortly. Let's just say for now that there is a map $F_{a,b}/N \to \mathbb{Z}^2$ that is clearly surjective and it must be an isomorphism because we have an independent proof that the loops a and b generate the group $\pi_1(T^2) \cong \mathbb{Z}^2$.

Definition 6.1. Let $a_1, \ldots a_n$ be symbols and let $r_1, \ldots r_m$ be words in the alphabet $a_1, \ldots, a_n$. The symbol

$$< a_1, \ldots, a_n \mid r_1, \ldots r_m >$$

denotes the group

$$G = F_{a_1,\ldots,a_n}/N(r_1, \ldots r_m),$$

where $F_{a_1,\ldots,a_n}$ is the free group on $a_1, \ldots a_n$ and $N(r_1, \ldots r_m)$ is the smallest normal subgroup of $F_{a_1,\ldots,a_n}$ containing the elements $r_1, \ldots, r_m$. We say that $a_1, \ldots a_n$ are *generators of* G, that $r_1, \ldots r_m$ are *relations* in G, and that $< a_1, \ldots, a_n \mid r_1 \ldots, r_m >$ is a *presentation of* G *by generators and relations.*

Example 6.6.
(1) $< a \mid a^2 >$ is a presentation for $\mathbb{Z}_2$, the cyclic group of order 2, as in Example 6.4 and Remark 6.2.
(2) $< a, b \mid a^2, b^2 >$ is a presentation of the infinite dihedral group D_∞ of Example 5.2
(3) $< a, b \mid aba^{-1}b^{-1} >$ is a presentation of the free abelian group in two generators, $\mathbb{Z}^2$, as in Example 6.5.
(4) The same reasoning that we used in Example 6.5 to derive a presentation of $\pi_1(T^2)$ can be used to derive a presentation of $\pi_1(K)$, where K is the Klein bottle: present K as the quotient of the unit square by the identification $aba^{-1}b$ (see Figure 1.2): $\pi_1(K) =< a, b \mid aba^{-1}b >$.
(5) It is now easy to see the pattern, with the same proof, in the case that there is only one vertex in the quotient space: if w is a word in $a_1, \ldots, a_n$ representing identifications on the boundary of a disk (or $2n$-gon) to get a surface X, then $\pi_1(X) =< a_1, \ldots, a_n \mid w >$. For example

$$\pi_1(\Sigma_g) =< a_1, b_1, a_2, b_2, \ldots a_g, b_g \mid a_1 b_1 a_1^{-1} b_1^{-1} \ldots a_g b_g a_g^{-1} b_g^{-1} > . \tag{6.8}$$

where Σ_g is as defined in Example 1.1. The common reason to all these computations of the fundamental group of a surface is: we can cover the quotient space X by two open sets U, V, where U has the homotopy type of a one- point union of n circles, V is simply connected, and $U \cap V$ deformation retracts to a circle. The induced homomorphism $\pi_1(U \cap V) \to \pi_1(U)$ is the homomorphism

$\mathbb{Z} \to F_{a_1,\dots,a_n}$ that sends the generator of $\mathbb{Z}$ to the word w. Thus the situation of Figure 6.6 is typical.

6.2. **Some properties of $\pi_1(\Sigma_g)$.** .

We have given a way of defining groups, called a presentation by generators and relations, see Definition 6.1, and we have given a presentation of $\pi_1(\Sigma_g)$. In general, giving a presentation of a group does not say very much about the group, for instance, is it trivial, abelian, etc. The reason is that it is, in general, not easy to tell what the smallest normal subgroup containing $r_1, \dots r_m$ is. For all we know it could be all of $F_{a_1,\dots,a_n}$. But in the case of the presentation (6.8) a lot more can be said. We study this presenation in more detail.

Theorem 6.2. *Let F_g denote a free group on g generators. Then there are homomorphisms $i : F_g \to \pi_1(\Sigma_g)$ and $j : \pi_1(\Sigma_g) \to F_g$ so that the composition $ji : F_g \to F_g$ is the identity, thus i is injective and j is surjective. In particular, $\pi_1(\Sigma_g)$ is non-abelian if $g > 1$.*

Proof. Recall the presentation (6.8):

$$\pi_1(\Sigma_g) =< a_1, b_1, \cdots, a_g, b_g \mid a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1} >$$

Let $F_g = F_{a_1,\dots,a_g}$. Since $F_{a_1,\dots,a_g}$ is free, there is a unique homomorphism

$$i : F_{a_1,\dots,a_g} \to< a_1, b_1, \cdots, a_g, b_g \mid a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1} >$$

with the property that $i(a_l) = a_l$ for $l = 1, \dots, g$. We want to define a homomorphism

$$j :< a_1, b_1, \cdots, a_g, b_g \mid a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1} >\to F_{a_1,\dots,a_g}$$

by setting $j(a_l) = a_l$ and $j(b_l) = e$ for $l = 1, \dots, g$. We can certainly define a homomorphism $j : F_{a_1,b_1,\dots,a_g,b_g}$ with this property, the question is whether it gives a well defined homomorphism on the quotient group

$$F_{a_1,b_1,\dots,a_g,b_g}/N \to F_{a_1,\dots,a_g},$$

where N is the smallest normal subgroup of $F_{a_1,b_1,\dots,a_g,b_g}$ containing the word $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}$, see Definition 6.1. This happens if and only if $j(N) = \{e\}$. Since the kernel of j is a normal subgroup, by the definition of N this happens if and only if $j(a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}) = e$. But $j(a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}) = a_1 a_1^{-1} \dots a_g a_g^{-1} = e$, so j is well defined on the quotient group. Since $ji(a_l) = a_l$ for $l = 1, \dots, g$, it follows that ji is the identity of $F_{a_1\dots,a_g}$. The remaining statements follow immediately. □

Remark 6.3. There is a geometric picture corresponding to the algebraic proof just given. Using a symmetric model of Σ_g, we can decompose $\Sigma_g = \Sigma_g^+ \cup \Sigma_g^-$, where each $\Sigma_g^\pm$ is homeomorphic to a disk with g holes, so each $\pi_1(\Sigma_g^\pm)$ is a free group F_g. Pick one of the pieces, say Σ_g^+, observe that it is a

retract of Σ_g and that the loops $a_1, \ldots, a_g$ can be chosen to lie in Σ_g^+, while the loops $b_1, \ldots, b_g$ go to trivial loops under the retraction, See Figure 6.7.

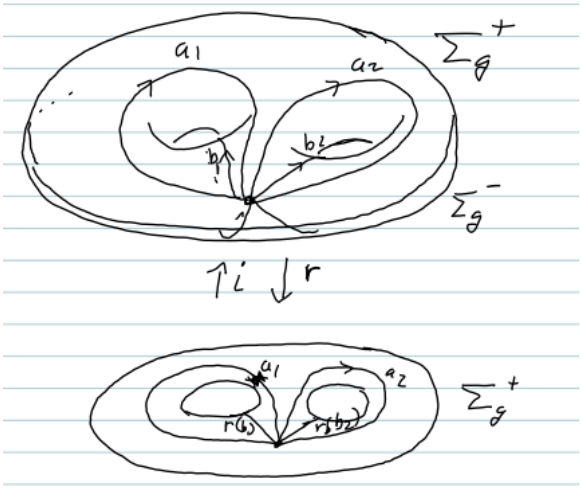
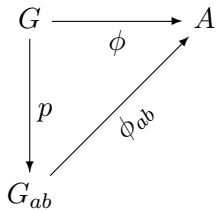


FIGURE 6.7. $F_{a_1,\ldots,a_g}$ is a Subgroup of $\pi_1(\Sigma_g)$.

6.2.1. *Abelianization of a Group.* To finish our discussion of $\pi_1(\Sigma_g)$ we need another concept from group theory. Given any group G, there is an *abelian* group G_{ab}, called the *abelianization* of G, with the following universal property: There is a homomorphism $p : G \to G_{ab}$, and, given any abelian group A and any homomorphism $\phi : G \to A$, there exists a unique homomorphism $\phi_{ab} : G_{ab} \to A$ so that $\phi_{ab} \circ p = \phi$:

(6.9)



An equivalent formulation of this property is to say that G_{ab} is the largest abelian quotient group of G.

To construct G_{ab}, we need the *commutator subgroup* G' of G: Given any group G, let G' be the subgroup generated by $\{ghg^{-1}h^{-1} : g, h \in G\}$. This is a normal subgroup of G because, for any $k \in G$, $k(ghg^{-1}h^{-1})k^{-1} = (kgk^{-1})(khk^{-1})(kgk^{-1})^{-1}(khk^{-1})^{-1}$, therefore $kG'k^{-1} \subset G'$ for all $k \in G$. Thus G/G' is a group and there is a surjective homomorphism $p : G \to G/G'$.

Theorem 6.3. *The group G_{ab} exists and $G_{ab} \cong G/G'$.*

Proof. First, note that for all $g, h \in G$, we have that $p(g)p(h) = p(h)p(g)$ because $p(ghg^{-1}h^{-1}) = e$ by definition of G'. Since p is surjective we get that G/G' is an abelian group. Next, given any abelian group A and any homomorphism $\phi : G \to A$, we have that $G' \subset ker(\phi)$, because, for all $g, h \in G$, since A is abelian, we have $\phi(ghg^{-1}h^{-1}) = \phi(g)\phi(h)\phi(g)^{-1}\phi(h)^{-1} = \phi(g)\phi(g)^{-1}\phi(h)\phi(h)^{-1} = e$. Thus, if $p : G \to G/G'$ is the quotient homomorphism, we see that G/G' has the universal mapping property of G_{ab} in Equation (6.2.2). Putting $A = G_{ab}$ and G/G' for G_{ab} in (6.2.2) we get a map $G/G' \to G_{ab}$, then, putting $A = G/G'$ we get a map $G_{ab} \to G/G'$, and it is easy to check that these are inverses of each other, so G_{ab} and G/G' are isomorphic.

□

Remark 6.4. This construction of G_{ab} makes it clear that it is "functorial" or "natural", meaning that if G and H are groups and $\phi : G \to H$ is a homomorphism, then there is an induced homomorphism $\phi_{ab} : G_{ab} \to H_{ab}$. Moreover, whenever we have compositions defined, then $(\phi \circ \psi)_{ab} = \phi_{ab} \circ \psi_{ab}$. Also $id_{ab} = id$. From this it follow formally that isomorphic groups have isomorphic abelianizations. To construct ϕ_{ab}, we just need to observe that $\phi(G') \subset H'$, since $\phi(ghg^{-1}h^{-1}) = \phi(g)\phi(h)\phi(g)^{-1}\phi(h)^{-1}$. Therefore $\phi : G \to H$ induces a well defined homomorphism $G/G' \to H/H'$. This homomorphism is, by definition, ϕ_{ab}.

Here is one useful example of the abelianization:

Theorem 6.4. *Let F_n be the free group on n generators and let $\mathbb{Z}^n$ be the free abelian group on n generators. Then $(F_n)_{ab} \cong \mathbb{Z}^n$.*

Proof. Let $S = \{a_1, \ldots, a_n\}$ be free generators for F_n, as in Equation (5.2) Then it is easy to check that the inclusion of $p(S)$ in $(F_n)_{ab}$ satisfies the universal mapping property of Equation (6.2.2). Thus the two groups in question are isomorphic. □

Remark 6.5. The computation of $\pi_1(T^2)$ in Example 6.5 gives us a way to visualize the commutator subgroup of the free group $F_{a,b}$ on two generators a, b. Present $T^2 = \mathbb{R}^2/\mathbb{Z}^2$, and also, as in Example 6.5 as $S^1 \vee S^1$, with fundamental group $F_{a,b}$ with relation $aba^{-1}b^{-1}$ corresponding to the boundary of the square. Let X be the pre-image in $\mathbb{R}^2$ of $S^1 \vee S^1$. Thus $X = \{(x, y) : x \in \mathbb{Z} \text{ or } y \in \mathbb{Z}\}$. Then it is easy to see that any loop in X based at the origin projects to a loop in the commutator subgroup $F'_{a,b}$, since it becomes trivial in $\pi_1(T^2)$ which is the abelianization of $F_{a,b}$. Later, when we discuss covering spaces, we will see that $\pi_1(X)$ is actually isomorphic to the kernel of $\pi_1(S^1 \vee S^1) \to \pi_1(T^2)$. The word $aba^{-1}b^{-1}$ corresponds to the loop shown in Figure 6.8, while a conjugate is shown in the same figure, and an arbitrary element of $\pi_1(X)$ is a product of conjugates. Thus

the commutator subgroup $F'_{a,b}$ is quite complicated. Van Kampen's theorem implies that it is the smallest normal subgroup containing $aba^{-1}b^{-1}$, and this picture indicates why it can be difficult to compute this normal subgroup.

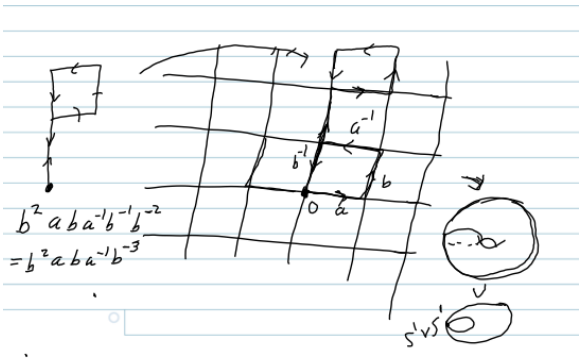
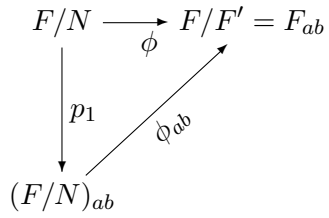


FIGURE 6.8. The Commutator Subgroup of F_2.

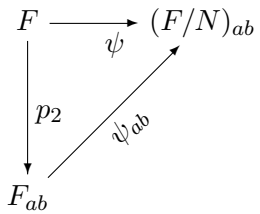
6.2.2. *Topological Invariance of the Genus.* With the concept of abelianization we can finish the proof of Theorem 1.1.

Theorem 6.5. $\pi_1(\Sigma_g)_{ab} \cong \mathbb{Z}^{2g}$.

Proof. Recall the presentation (6.8) of $\pi_1(\Sigma_g)$. Let $F = F_{a_1,b_1,\ldots,a_g,b_g}$ and let N be the smallest normal subgroup of F containing the word $w = a_1b_1a_1^{-1}b_1^{-1}\cdots a_gb_ga_g^{-1}b_g^{-1}$, so that $\pi_1(\Sigma_g) = F/N$. Since $w \in F'$, we see that there is a map $F/N \to F/F' = F_{ab}$. We get diagrams:



and



Since all these maps send the generators $a_1, b_1, \ldots, a_g, b_g$ of F to their images in the respective groups, it is easy to check that $\phi_{ab} \circ \psi_{ab}$ and $\psi_{ab} \circ \phi_{ab}$ are the identity, hence $(F/N)_{ab} \cong F_{ab}$, and, by the previous theorem, $F_{ab} \cong \mathbb{Z}^{2g}$. □

Corollary 6.3. *If Σ_g is homotopy equivalent to Σ_h, then $g = h$.*

Proof. By Remark 6.4, if $\pi_1(\Sigma_G) \cong \pi_1(\Sigma_h)$, then $\mathbb{Z}^{2g} \cong \mathbb{Z}^{2h}$. By standard theory of abelian groups, two free abelian groups are isomorphic if and only if they have the same rank, thus $g = h$. □

Note that this concludes the proof of Theorem 1.1: the only part that was missing was being able to distinguish Σ_g and Σ_h for $g \neq h$.

7. Differential Geometry of Surfaces

Recall, from Definition 1.1 and from Section 6 of [12] what is meant by S being a *smooth surface*. We now turn our attention to geometric concepts that can be defined on smooth surfaces. We will often need to refer to Section 6 of [12] for some of these concepts. In the references there are several books that cover these topics, [4, 7, 11, 13, 14, 15, 16, 17, 18], although often in more detail and with different points of view.

7.1. **Riemannian Metrics on Surfaces.** The concept of Riemannian metric is motivated by the study of the intrinsic geometry of smooth surfaces $S \subset \mathbb{R}^3$, in particular, the intrinsic distance d_S, as defined in Section 6..1 of [12]. Recall that for $x, y \in S$, $d_S(x, y)$ is defined to be

$$\inf\{L(\gamma) : \gamma \text{ a piecewise differentiable path in } S \text{ from } x \text{ to } y\}, \tag{7.1}$$

where $L(\gamma)$ denotes the *length* of γ: if $\gamma : [a, b] \to S$, then

$$L(\gamma) = \int_0^1 |\gamma'(t)| \, dt = \int_0^1 \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2} \, dt, \tag{7.2}$$

which is often abbreviated as

$$L(\gamma) = \int_\gamma \sqrt{dx^2 + dy^2 + dz^2} \; = \int_\gamma ds, \tag{7.3}$$

where $ds^2 = dx^2 + dy^2 + dz^2$.

If the surface S is parametrized by an open set $U \subset \mathbb{R}^2$, meaning that there is a smooth map $\mathbf{x} : U \to \mathbb{R}^3$ so that $\mathbf{x}$ is a homeomorphism from U to S and it is non-singular in the sense that, if (u_1, u_2) are the coordinates of $u \in U$, the partial derivatives $\mathbf{x}_{u_1}$ and $\mathbf{x}_{u_2}$ are linearly independent at each $u \in U$, equivalently, the cross-product $\mathbf{x}_{u_1} \times \mathbf{x}_{u_2} \neq 0$ at each $u \in U$. Let us abbreviate $\mathbf{x}_1 = \mathbf{x}_{u_1}$ and $\mathbf{x}_2 = \mathbf{x}_{u_2}$.

The curve $\gamma(t) = \mathbf{x}(u(t))$ for some curve $u(t), a \leq t \leq b$ in U. Then $\gamma'(t) = u_1'\mathbf{x}_1 + u_2'\mathbf{x}_2$, so

$$\begin{aligned} \gamma'(t) \cdot \gamma'(t) &= (u_1'\mathbf{x}_1 + u_2'\mathbf{x}_2) \cdot (u_1'\mathbf{x}_1 + u_2'\mathbf{x}_2) \\ &= g_{11}(u(t))(u_1')^2 + 2g_{12}(u(t))u_1'u_2' + g_{22}(u(t))(u_2')^2, \end{aligned} \tag{7.4}$$

where the g_{ij} are the smooth functions of u defined on U by

$$g_{ij}(u) = \mathbf{x}_i(u) \cdot \mathbf{x}_j(u), \quad i = 1, 2, \ j = 1, 2. \tag{7.5}$$

Using (7.5), we usually abbreviate (7.4) as

$$ds^2 = g_{11}du_1^2 + 2g_{12}du_1du_2 + g_{22}du_2^2, \tag{7.6}$$

compare with Equations (6.4) and (6.5) of [12] and the discussion of these equations. The length of the curve γ can then be computed in terms of the curve $u(t)$ by

$$L(\gamma) = \int_a^b \, ds = \int_a^b \sqrt{g_{11}(u_1')^2 + 2g_{12}u_1'u_2' + g_{22}(u_2)^2} \, dt. \tag{7.7}$$

The interpretation of Equation (7.6) is the following: The matrix of smooth functions

$$g = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \text{ where } g_{12} = g_{21}, \tag{7.8}$$

is a symmetric, positive definite matrix that gives an inner product on the tangent vectors based at each point $u \in U$, and this inner product is the same as the usual dot product in $\mathbb{R}^3$ of the corresponding tangent vectors to S at $\mathbf{x}(u)$. In fact, Equation (7.4) is the same as

$$(a_1\mathbf{x}_1 + a_2\mathbf{x}_2) \cdot (a_1\mathbf{x}_1 + a_2\mathbf{x}_2) = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \tag{7.9}$$

where the left hand side is the usual dot product of vectors in $\mathbb{R}^3$. These vectors are actually in $T_{\mathbf{x}(u)}S$, the tangent space to S at the point $\mathbf{x}(u)$, and the vectors $\mathbf{x}_1(u), \mathbf{x}_2(u)$ form a basis for this space (recall the assumption that $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent at each $u \in U$). (See Section 7.1.1 below for a quick review of inner products.)

We have seen examples of this, say the parametrization of S^2 by spherical coordinates: $\mathbf{x}(\phi, \theta) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$. To make $\mathbf{x}$ a homeomorphism onto its image we could take $U = (0, \pi) \times (0, 2\pi)$ with image S^2 with a meridian removed. Then $ds^2 = d\phi^2 + \sin^2\phi \, d\theta^2$, see Example 6.6 of [12].

Another interesting example is given by stereographic projection:

Example 7.1. Recall from the homework from Math 4510 the map $\mathbf{x} : \mathbb{R}^2 \to S^2 \setminus \{N\}$, where $N = (0, 0, 1)$ is the north pole, that assigns to $u = (u_1, u_2) \in$

$\mathbb{R}^2$ the point of intersection with S^2 of the straight line segment $\overline{uN}$ from u to N. We have seen the formula

$$\mathbf{x}(u_1, u_2) = (2u_1, 2u_2, u_1^2 + u_2^2 - 1)/(1 + u_1^2 + u_2^2). \tag{7.10}$$

From this it is not hard to get the formulas for the partial derivatives:

$$\begin{aligned} \mathbf{x}_1 &= \frac{2}{(1 + u_1^2 + u_2^2)^2}\ (-u_1^2 + u_2^2 + 1,\ -2u_1u_2,\ 2u_1) \\ \mathbf{x}_2 &= \frac{2}{(1 + u_1^2 + u_2^2)^2}\ (-2u_1u_2,\ u_1^2 - u_2^2 + 1,\ 2u_2). \end{aligned} \tag{7.11}$$

Taking dot products and simplifying find formulas for the $\mathbf{x}_i \cdot \mathbf{x}_j$: $\mathbf{x}_1 \cdot \mathbf{x}_2 = 0$ and $\mathbf{x}_1 \cdot \mathbf{x}_1 = \mathbf{x}_2 \cdot \mathbf{x}_2 = 4/(1 + u_1^2 + u_2^2)^2$. So we get the following expression for g, which, for future reference, we will call g_S (the *spherical metric*):

$$g_S = \frac{4(du_1^2 + du_2^2)}{(1 + u_1^2 + u_2^2)^2} = \frac{4\ du \cdot du}{(1 + |u|^2)^2}, \tag{7.12}$$

where $u = (u_1, u_2)$. We will study the meaning of this expression in more detail once we put this discussion into a more general context.

7.1.1. *Review of Inner Products.* We quickly review some of the linear algebra concepts just used. Refer to any textbook on linear algebra for more details. We will concentrate on two dimensions, but all works the same way in any dimension.

Recall that a symmetric matrix g is said to be *positive definite* if the right hand side of (7.9) is positive for all a_1, a_2 not both 0. Since the left had side of (7.9) clearly has this property, we get that the matrix in the right hand side is positive definite.

Given such a positive definite matrix g, and given two vectors $\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2$ and $\mathbf{b} = b_1\mathbf{e}_1 + b_2\mathbf{e}_2$, in $\mathbb{R}^2$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are the standard basis vectors in $\mathbb{R}^2$, we can define a function $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by assigning to the vectors $\mathbf{a}, \mathbf{b}$ the number $< \mathbf{a}, \mathbf{b} >$ defined by

$$< \mathbf{a}, \mathbf{b} >= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \mathbf{a}^t\ g\ \mathbf{a}, \tag{7.13}$$

where $\mathbf{a}^t$ means the matrix transpose. This function is called an *inner product* in $\mathbb{R}^2$, and every inner product in $\mathbb{R}^2$ is obtained in this way. The definition of *inner product* on a vector space is a real valued function of two vectors $\mathbf{a}, \mathbf{b}$, denoted $< \mathbf{a}, \mathbf{b} >$, that satisfies the following properties:

(1) Is *bilinear*: linear in each variable.
(2) Is *symmetric*: $< \mathbf{a}, \mathbf{b} >=< \mathbf{b}, \mathbf{a} >$ for all $\mathbf{a}, \mathbf{b}$.
(3) Is *positive definite:* $< \mathbf{a}, \mathbf{a} >\geq 0$ for all $\mathbf{a}$ and $< \mathbf{a}, \mathbf{a} >= 0$ only if $\mathbf{a} = 0$.

Given a positive definite 2×2 matrix g, the formula (7.13) clearly defines an inner product on $\mathbb{R}^2$. The first two conditions are clear, and the third is equivalent to the assumption that the matrix g is positive definite. Conversely, if we are given an inner product $< , >$ on $\mathbb{R}^2$, let

$$g = \begin{pmatrix} < \mathbf{e}_1, \mathbf{e}_1 > & < \mathbf{e}_1, \mathbf{e}_2 > \\ < \mathbf{e}_2, \mathbf{e}_1 > & < \mathbf{e}_2, \mathbf{e}_2 > \end{pmatrix}, \tag{7.14}$$

briefly, $g_{ij} =< \mathbf{e}_i, \mathbf{e}_j >$. Using this matrix in (7.13) gives back the inner product $< \mathbf{a}, \mathbf{b} >$. Thus *every* inner product on $\mathbb{R}^2$ is obtained by (7.13), in other words, (7.13) establishes a one to one correspondence between inner products on $\mathbb{R}^2$ and symmetric, positive definite 2×2 matrices.

The standard example of an inner product is the usual dot product in $\mathbb{R}^2$, which correspond to $g = I$, the unit 2×2 matrix. Any inner product allows to define lengths and angles by using the same formulas as for the usual dot product:

Definition 7.1. If $< , >$ is an inner product on $\mathbb{R}^2$ (corresponding to a symmetric positive definite matrix g)), and if $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$, we define:

(1) The *magnitude* of $\mathbf{a}$ (with respect to g), denoted $||\mathbf{a}||$, by $||\mathbf{a}|| = \sqrt{< \mathbf{a}, \mathbf{a} >}$.
(2) The *angle* between $\mathbf{a}$ and $\mathbf{b}$ (with respect to g) to be the number $\angle(\mathbf{a}, \mathbf{b}) \in [0, \pi]$ that satisfies

$$\cos(\angle(\mathbf{a}, \mathbf{b})) = \frac{< \mathbf{a}, \mathbf{b} >}{||\mathbf{a}|| \; ||\mathbf{b}||}. \tag{7.15}$$

Remark 7.1. It is a fact that the Cauchy - Schwarz inequality holds for any inner product, so the number in the right hand side of (7.15) has absolute value at most one, hence it is the cosine of some angle.

We will need the following construction, which again works in any dimension, but we need only in two dimensions: if $A : \mathbb{R}^2 \to \mathbb{R}^2$ is an invertible linear transformation, given by a matrix $A = (a_{ij})$ in the standard basis for $\mathbb{R}^2$, and $< , >$ is an inner product with matrix $g = (< \mathbf{e}_i, \mathbf{e}_j >)$, then we can define a new inner product $<< , >>$ or A^*g, called the *pull-back by* A, by the formula

$$<< \mathbf{a}, \mathbf{b} >>=< A\mathbf{a}, A\mathbf{b} > . \tag{7.16}$$

Theorem 7.1. *The pull-back $<< , >>$ just defined is an inner product on $\mathbb{R}^2$ and its matrix A^*g is given by $A^*g = A^t g A$.*

Proof. It is clear that $<< , >>$ is bilinear and symmetric, to prove positive definite observe that $<< \mathbf{a}, \mathbf{a} >>=< A\mathbf{a}, A\mathbf{a} > \geq 0$, and, since $< , >$ is an inner product, it $= 0$ if and only if $A\mathbf{a} = 0$. Since A is invertible this means $\mathbf{a} = 0$, hence $<< , >>$ is positive definite. Now, $<< \mathbf{a}, \mathbf{b} >>=< A\mathbf{a}, A\mathbf{b} >= (A\mathbf{a})^t g(A\mathbf{a}) = (\mathbf{a}^t A^t) g(A\mathbf{a}) = \mathbf{a}^t (A^t g A)\mathbf{a}$, thus by the formula (7.13) for the

correspondence between matrices and inner products, the matrix of $<<$, $>>$ is as asserted. In the above chain of equalities, the second one is Equation (7.13) while the third is the standard formula for the transpose of the product of two matrices. □

Remark 7.2. As we mentioned, the concept of pull-back works in all generality for *injective* linear maps $A : \mathbb{R}^m \to \mathbb{R}^n$, which of course only exist for $m \leq n$. Equation (7.9) for the two-dimensional inner product obtained from tangent vectors in $\mathbb{R}^3$ is an example of a pull-back by an injective linear map $A : \mathbb{R}^2 \to \mathbb{R}^3$.

7.1.2. *Local Riemannian Metrics.* The structure we have been using to compute lengths of curves for surfaces in $\mathbb{R}^3$ in terms of parametrizations of the surface can be abstracted into the following definition:

Definition 7.2. Let $U \subset \mathbb{R}^2$ be an open set.

(1) A *Riemannian metric on* U is defined in one of the following two equivalent ways:
   (a) An inner product on the tangent vectors to U at each point $u \in U$, varying smoothly with u. If $\mathbf{a}, \mathbf{b}$ are tangent vectors, denote their inner product by $< \mathbf{a}, \mathbf{b} >_u$.
   (b) A symmetric positive definite matrix g of smooth functions on U as in Equation (7.8).

(2) If g is a Riemannian metric on U and $\mathbf{a}, \mathbf{b}$ are tangent vectors to U at u, then the *g-length* $||\mathbf{a}||_u$, of $\mathbf{a}$ and the *g-angle* $\angle(\mathbf{a}, \mathbf{b})_u$ between $\mathbf{a}$ and $\mathbf{b}$ are defined as in Definition 7.1: $||\mathbf{a}||_u = \sqrt{< \mathbf{a}, \mathbf{a} >_u}$ and
$$\cos(\angle(\mathbf{a}, \mathbf{b})_u) = \frac{< \mathbf{a}, \mathbf{b} >_u}{||\mathbf{a}||_u ||\mathbf{b}||_u}.$$

(3) If g is a Riemannian metric on U and $\alpha(t) = (u_1(t), u_2(t)),\ a \leq t \leq b$ is a piecewise smooth curve in U, then the *length of* α (with respect to g) is

$$L(\alpha) = \int_a^b \sqrt{< \alpha'(t), \alpha'(t) >_{\alpha(t)}}\, dt \tag{7.17}$$
$$= \int_a^b \sqrt{g_{11}(u_1')^2 + 2g_{12}u_1'u_2' + g_{22}(u_2')^2}\, dt,$$

and, if U is connected, the *g-distance* or *Riemannian distance* between two points in U is defined to be the infimum of the g - lengths of all piecewise smooth curves in U joining the two points.

(4) If g is a Riemannian metric on U and $D \subset U$ is a domain over which double integrals are defined (rectangles, regions between the graphs of two continuous functions, etc), then $A(D)$, the *area* of D with respect to the metric g, is defined to be

$$A(D) = \iint_D \sqrt{\det(g)}\, du_1\, du_2, \tag{7.18}$$

equivalently, we could define the "area element" dA by

$$dA = \sqrt{\det(g)}\ du_1\ du_2. \tag{7.19}$$

(5) If $D \subset U$ is as above, and $f : D \to \mathbb{R}$ is a continuous function, then we define its integral (with respect to g- area) by

$$\iint_D f\ dA\ =\ \iint_D f(u_1, u_2)\ \sqrt{\det(g)}\ du_1\ du_2. \tag{7.20}$$

Remark 7.3. Some comments on these definitions:

(1) The equivalence between the two versions of the definition of Riemannian metric follows from the discussion of the equivalence between inner products and positive definite matrices in Section 7.1.1. We just need to apply the correspondence at each $u \in U$.
(2) The definition of angles is the straightforward analogue of the usual one in the Euclidean plane.
(3) The definition of length of curves and of intrinsic distance are the straightforward extension of the corresponding definitions for surfaces in $\mathbb{R}^3$.
(4) The same is true of the definition of area: recall that if $\mathbf{x} : U \to S \subset \mathbb{R}^3$ is a parametrization of a surface S, then the usual infinitesimal arguments for the distortion of area give that the area of $\mathbf{x}(D)$ is given by

$$\iint_D ||\mathbf{x}_1 \times \mathbf{x}_2||\ du_1\ du_2,$$

and the usual formulas for magnitudes of cross-products give

$$||\mathbf{x}_1 \times \mathbf{x}_2|| = \sqrt{(\mathbf{x}_1 \cdot \mathbf{x}_1)(\mathbf{x}_2 \cdot \mathbf{x}_2) - (\mathbf{x}_1 \cdot \mathbf{x}_2)^2} = \sqrt{g_{11}g_{22} - g_{12}^2},$$

which, by Equation 7.5, agrees with Equation 7.18 in this case. (Consult any advanced calculus textbook for the formulas for area, as well as for surface integrals).
(5) The definition of integral of a function agrees with the usual definition of surface integral in the case of q parametrized surface $S = \mathbf{x}(D) \subset \mathbb{R}^3$:

$$\iint_S f\ dA = \iint_D f(u_1, u_2)\ ||\mathbf{x}_1 \times \mathbf{x}_2||\ du_1\ du_2.$$

Example 7.2. The basic example of these definitions is the Riemannian metric on U resulting from a smooth, non-singular parametrization $\mathbf{x} : U \to S \subset \mathbb{R}^3$ of a smooth surface $S \subset \mathbb{R}^3$. Non-singular means that the cross-product $\mathbf{x}_1 \times \mathbf{x}_2$ is never zero, that is, $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent at each $u \in U$, thus they give a basis of the tangent plane $T_{\mathbf{x}(u)}S$ of S at $\mathbf{x}(u)$. The Riemannian metric is defined by

$$< \mathbf{a}, \mathbf{b} >_u = (a_1\mathbf{x}_1 + a_2\mathbf{x}_2) \cdot (b_1\mathbf{x}_1 + b_2\mathbf{x}_2) \tag{7.21}$$
$$\text{or} \quad < \mathbf{a}, \mathbf{b} >_u = d_u\mathbf{x}(\mathbf{a}) \cdot d_u\mathbf{x}(\mathbf{b}),$$
$$\text{or, what is the same,} \quad g_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j.$$

In the second line $d_u\mathbf{x}$ denotes the *differential* at u of the map $\mathbf{x}$. This is a linear map that takes tangent vectors to U at u to tangent vectors to S at $\mathbf{x}(u)$, thus $d_u\mathbf{x} : T_uU \to T_{\mathbf{x}(u)}S$. By definition, $d_u\mathbf{x}(\mathbf{e}_1) = \mathbf{x}_1$ and $d_u\mathbf{x}(\mathbf{e}_2) = \mathbf{x}_2$, thus the linearity of $d_u\mathbf{x}$ implies that $d_u\mathbf{x}(\mathbf{a}) = d_u\mathbf{x}(a_1\mathbf{e}_1 + a_2\mathbf{e}_2) = a_1\mathbf{x}_1 + a_2\mathbf{e}_2$, similarly $d_u\mathbf{x}(\mathbf{b}) = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$. Thus, in this class of examples, the interpretation of the inner product $< \mathbf{a}, \mathbf{b} >_u$ is the usual dot product $d_u\mathbf{x}(\mathbf{a}) \cdot d_u\mathbf{x}(\mathbf{b})$ in $T_{\mathbf{x}(u)}S \subset \mathbb{R}^3$. But the inner product given by Riemannian metric need not be obtained in this way.

Example 7.3. The metric $du_1^2 + du_2^2$ is the standard inner product (dot product) of vectors in $\mathbb{R}^2$. This is called the *Euclidean (Riemannian) metric* on U. If f is a *positive* smooth function on U, then the metric $g = f(u)(du_1^2 + du_2^2)$ is called a *conformally Euclidean* metric, because the angle measurements are the same as in the Euclidean metric. If we use the subscript g for the measurements using g and the subscript E for the Euclidean measurements, we get

$$\frac{< \mathbf{a}, \mathbf{b} >_g}{||\mathbf{a}||_g||\mathbf{b}||_g} = \frac{f < \mathbf{a}, \mathbf{b} >_E}{(\sqrt{f}||\mathbf{a}||_E)(\sqrt{f}||\mathbf{b}||_E)} = \frac{< \mathbf{a}, \mathbf{b} >_E}{||\mathbf{a}||_E||\mathbf{b}||_E},$$

hence $\cos\angle(a, b)_g = \cos\angle(\mathbf{a}, \mathbf{b})_E$, so $\angle(a, b)_g = \angle(\mathbf{a}, \mathbf{b})_E$. In particular, we see from Equation (7.12) that the metric resulting from stereographic projection is conformally Euclidean. This is equivalent to the well - known fact that stereographic projection is conformal (preserves angles between curves).

Example 7.4. A very important example is the *Poincaré metric* or *hyperbolic metric* on the unit disk $\{u_1^2 + u_2^2 < 1\}$ defined by a formula with some remarkable similarities to the stereographic formula (7.12) for the spherical metric g_S:

$$g_P = \frac{4(du_1^2 + du_2^2)}{(1 - u_1^2 - u_2^2)^2} = \frac{4\ du \cdot du}{(1 - |u|^2)^2}, \tag{7.22}$$

where $u = (u_1, u_2)$. Observe that this is a conformally Euclidean metric (as defined in Example 7.3) on the disk. It is a fact, which we will not prove (but was proved by Hilbert in the early 1900's) that this metric is *not* the metric of any surface in $\mathbb{R}^3$. We will study this metric in more detail later on.

7.1.3. *Review of Differentials.* We quickly review the notations for derivatives of maps between Euclidean spaces $\mathbb{R}^m$ and $\mathbb{R}^n$. More details can be found in any advanced calculus book. Let $u_1, \ldots u_m$ denote the coordinates of a point $u \in \mathbb{R}^m$ and $v_1, \ldots, v_n$ coordinates of a point $v \in \mathbb{R}^n$.

Let $U \subset \mathbb{R}^m$ be an open set, let $f : U \to \mathbb{R}^n$ be a differentiable map (say smooth, of class C^∞). At each $u \in U$ the map f has a linear approximation, namely a linear map

$$d_u f : \mathbb{R}^m \to \mathbb{R}^n,$$

called the *differential of f at u*, whose value $d_u f(h)$ at a vector h based at u (thought of as a tangent vector to U at u) gives the best linear approximation to $f(u+h) - f(u)$ in the sense that

$$f(u+h) - f(u) = d_u f(h) + o(|h|) \text{ as } \ h \to 0,$$

where $o(|h|)$ denotes a function of h that goes to zero *faster* than $|h|$ (equivalently, *faster than any linear function of h*), meaning $\lim_{h\to 0} \frac{o(|h|)}{|h|} = 0$.

Being a linear transformation $\mathbb{R}^m \to \mathbb{R}^n$, the differential $d_u f$ has a matrix with respect to the standard bases $\mathbf{e}_1, \dots, \mathbf{e}_m$ and $\mathbf{e}_1, \dots, \mathbf{e}_n$ of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. This matrix is the *Jacobian matrix*

$$J(f) = \begin{pmatrix} \frac{\partial v_1}{\partial u_1} & \cdots & \frac{\partial v_1}{\partial u_m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial v_n}{\partial u_1} & \cdots & \frac{\partial v_n}{\partial u_m} \end{pmatrix} \tag{7.23}$$

Thus the columns of the Jacobian matrix give the components of the vector $d_u f(\mathbf{e}_j)$:

$$d_u f(\mathbf{e}_j) = \frac{\partial f}{\partial u_j} = \frac{\partial v_1}{\partial u_j}\mathbf{e}_1 + \dots \frac{\partial v_n}{\partial u_j}\mathbf{e}_n, \text{ for each } j = 1, \dots, m,$$

while the rows of the Jacobian matrix represent the differentials of the component functions v_i of $v = f(u)$,

$$dv_i = \frac{\partial v_i}{\partial u_1} du_1 + \dots + \frac{\partial v_i}{\partial u_m} du_m \ \text{ for each } \ i = 1, \dots, n.$$

Here $du_1, \dots, du_m$ are the differentials of the component functions in $\mathbb{R}^m$ (which, being linear functions, are the same as the u_i, but, when we write d_{u_i} we think of them as operating on vectors based at u). This is the basis for the dual space of $\mathbb{R}^m$ (the space of linear functions on $\mathbb{R}^m$) dual to the standard basis $\mathbf{e}_i$:

$$du_i(\mathbf{e}_j) = \frac{\partial u_i}{\partial u_j} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

7.1.4. *Induced Metrics and Isometries.* Suppose U and V are open subsets of $\mathbb{R}^2$, and $f : U \to V$ is a *diffeomorphism* (a smooth, invertible map with smooth inverse). Suppose we have a Riemannian metric g on V, given by a smoothly varying inner product $< \ , \ >_v$, equivalently, a positive definite matrix $g(v)$ of smooth functions, as in Definition 7.2. Then we get a Riemannian metric on U, called the *pull - back metric*, denoted f^*g, that can be described in one of the following two equivalent ways:

Definition 7.3. The inner product $<<\ ,\ >>$ of the metric f^*g is defined by Equation (7.16), explicitly,

$$<<\mathbf{a},\mathbf{b}>>_u=<d_uf(\mathbf{a}),d_uf(\mathbf{b})>_{f(u)}$$

for all vectors $\mathbf{a},\mathbf{b}$ based at $u\in U$. The matrix of the pull - back metric f^*g is

$$(f^*g)(u)=(J(f)(u))^t\ g(f(u))\ (J(f)(u)),$$

where $J(f)(u)$ is the Jacobian matrix of f at $u\in U$ as in Equation (7.23).

The equivalence of the two definitions is clear from Theorem 7.1 applied to the linear map d_uf whose matrix is the Jacobian matrix of Equation (7.23). Since, by assumption, f is a diffeomorphism, the linear transformation d_uf is an isomorphism, so Theorem 7.1 does indeed apply.

Definition 7.4. Let U and V be open sets in $\mathbb{R}^2$, let h and g be Riemannian metrics on U, V respectively (denote this briefly by: (U,h), (V,g) are Riemannian open sets in $\mathbb{R}^2$), and let $f:U\to V$ be a diffeomorphism. We say that f is an *isometry* if $f^*g=h$.

Remark 7.4. We have earlier defined the concept of *isometry* of metric spaces as a map that preserves distance. We have now defined another concept with the same name. It turns out that there is no conflict. It is clear that if $f:(U,h)\to(V,g)$ is an isometry in the sense just defined, and d_h,d_g denote the intrinsic distance functions as in Definition 7.2, then f is an isometry in the sense of metric spaces, that is, $d_g(f(p),f(q))=d_h(p,q)$. This is clear for quite formal reasons: f gives a one-to-one correspondence between piecewise smooth curves γ in U and $f\circ\gamma$ in V which preserves length: $L(f\circ\gamma)=L(\gamma)$, because, using the chain-rule $(f\circ\gamma)'(t)=d_{\gamma(t)}f\ (\gamma'(t))$, therefore we get

$$\begin{aligned}L(f\circ\gamma)&=\int_a^b<d_{\gamma(t)}f\ (\gamma'(t)),d_{\gamma(t)}f\ (\gamma'(t))>^{\frac{1}{2}}_{f(\gamma(t))}\ dt\\&=\int_a^b<<\gamma'(t),\gamma'(t)>>^{\frac{1}{2}}_{\gamma(t)}\ dt=L(\gamma),\end{aligned}$$

so the two sets of lengths used to define $d_h(p,q)$ and $d_g(f(p),f(q))$ coincide, so they have the same infimum. The converse is also true: any metric isometry $f:(U,d_h)\to(V,d_g)$ is smooth and $f:(U,h)\to(V,g)$ is a Riemannian isometry in the sense just defined. This is harder to prove and we will not prove it here. From now on, the word *isometry*, in any Riemannian context where it would make sense, will always mean isometry as in Definition 7.4.

Example 7.5. A very familiar example of isometry results from changing variables to polar coordinates. Let $U\subset\mathbb{R}^2$ be any open subset on which the map $f(r,\theta)=(r\cos\theta,r\sin\theta)$ is a diffeomorphism onto its image, for instance, $(0,\infty)\times(-\pi,\pi)$. The easiest and most practical way, in this and most examples, to compute the pull-back metric is to use differential

notation: Let $dx^2 + dy^2$ denotes the Euclidean metric in $\mathbb{R}^2$, then

$$\begin{aligned} f^*(dx^2 + dy^2) &= (d(r\cos\theta))^2 + (d(r\sin\theta))^2 \\ &= (\cos\theta\ dr - r\sin\theta\ d\theta)^2 + (\sin\theta\ dr + r\cos\theta\ d\theta)^2 \\ &= (\cos^2\theta + \sin^2\theta)dr^2 + 2(-r\sin\theta\cos\theta + r\cos\theta\sin\theta)\ drd\theta \\ &\quad + r^2(\sin^2\theta + \cos^2\theta)\ d\theta^2 \quad = \quad dr^2 + r^2 d\theta^2, \end{aligned}$$

which is the familiar expression for arclength in polar coordinates. We have carried the computation in detail to illustrate how they can be done. An equivalent way to do the calculation would of course be to use the Jacobian matrix

$$J = J(f) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

thus f^*g has matrix

$$J^t I J = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix} \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}$$

(where I is the unit matrix, the matrix of $dx^2 + dy^2$), which gives the matrix for $dr^2 + r^2 d\theta^2$.

Note that the definition of area from Equation (7.19) gives the familiar formula for area in polar coordiantes:

$$dA = \sqrt{\det \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}}\ drd\theta = rdrd\theta,$$

7.1.5. *Two Important Examples.* To give an idea of how Riemannian metrics can be used, we study in more detail the metric spherical metric g_S of Example 7.1, see Equation (7.12), and the Poincaré metric g_P of Example 7.4, see Equation (7.22).

Let's see what conclusions we can get from the formula (7.12), forgetting, for the moment, that it is describing the unit sphere S^2:

Theorem 7.2. *The metric $g_s = 4(du_1^2 + du_2^2)/(1 + u_1^2 + u_2^2)^2 = 4(du \cdot du)(1 + |u|^2)^2$ on $\mathbb{R}^2$, where $u = (u_1, u_2)$, has the following properties:*

1. *It is conformally Euclidean: angle measurements in g_S are the same as in $du_1^2 + du_2^2$*
2. *It is rotationally symmetric about the origin and is also invariant under reflections on lines through the origin, meaning that, if A is an orthogonal 2×2 matrix, then $A^* g_S = g_S$.*
3. *Its restriction to $\mathbb{R}^2 \setminus \{0\}$ is invariant under inversion in the unit circle $|u| = 1$: if $\phi : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}^2 \setminus \{0\}$ is defined by $\phi(u) = u/|u|^2$, then $\phi^* g_S = g_S$.*

(4) *Let $C_r = \{|u| = r\}$ be the Euclidean circle of radius r. Then its length in g_S is $L(C_r) = \frac{4\pi r}{1+r^2}$. In particular $L(C_r) \le 2\pi$ and $= 2\pi$ if and only if $r = 1$. Also, $L(C_{1/r}) = L(C_r)$ in agreement with (3). See Figure 7.1*
(5) *If γ is a line through the origin: $\gamma(t) = tu$ for some fixed $u \in \mathbb{R}^2 \setminus \{0\}$ and $-\infty < t < \infty$, then $L(\gamma) = 2\pi$.*
(6) *The area of $\mathbb{R}^2$ in the metric g_S is 4π.*

Proof. Statement (1) was observed in Example 7.3. For (2) observe that if A is an orthogonal matrix, by Theorem 7.1 the matrix of $A^*(du_1^2 + du_2^2)$ is $A^t I A = A^t A = I$ where I is the unit matrix. Therefore $A^* g_S = 4A^*(du \cdot du)(1 + |Au|^2)^2 = 4(du \cdot du)(1 + |u|^2)^2 = g_S$ because $|Au| = |u|$ for an orthogonal matrix A. The proof of (3) is a homework problem. For (4), parametrize the circle C_r by $u(t) = r(\cos t, \sin t)$, $0 \le t \le 2\pi$. Then

$$L(C_r) = \int_0^{2\pi} \frac{2|u'(t)|}{1 + |u(t)|^2}\, dt = \int_0^{2\pi} \frac{2r}{1 + r^2}\, dt = \frac{4\pi r}{1 + r^2}.$$

Differentiating, we see that its maximum is at $r = 1$, where the value is 2π, see Figure 7.1. For (5), using the rotational symmetry from (2) it is enough to consider one line through the origin, for instance, the u_1-axis, parametrized as $\gamma(t) = (t, 0)$, $-\infty < t < \infty$ and its length is

$$L(\gamma) = \int_{-\infty}^{\infty} \frac{2\, dt}{1 + t^2}\, dt = 2 \arctan(t)|_{-\infty}^{\infty} = 2\pi.$$

Finally, to compute the area, we need to use the formula (7.18) for area:

$$A(\mathbb{R}^2) = \iint_{\mathbb{R}^2} \frac{4\, du_1 du_2}{(1 + |u|^2)^2} = \int_0^{2\pi} \int_0^{\infty} \frac{4r dr d\theta}{(1 + r^2)^2} = 2\pi \Big(\frac{-2}{1 + r^2} \Big)_0^{\infty} = 4\pi.$$

□

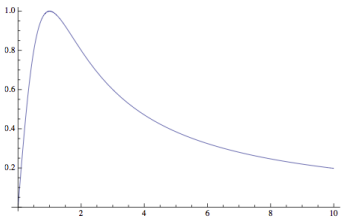


FIGURE 7.1. The Graph of $L(C_r)$.

Remark 7.5. Note that the properties listed in this Theorem are all familiar properties of the spherical metric. Note the following points:

(1) Recall that a 2×2 orthogonal matrix A is of one of the two forms

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \text{ or } A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix},$$

see Section 2.1 of [12], formulas (2.1) and (2.2). The first is orientation preserving (its determinant is 1) and is a counterclockwise rotation about the origin by an angle θ, while the second one is orientation reversing (its determinant is -1) and is a reflection about the line through the origin making an angle $\theta/2$ with the x-axis.

(2) These and the inversion symmetry in (3) are the only symmetries of S^2 that are visible in the formula for g_S. But we know that S^2 has many more symmetries. By the nature of stereographic projection, we only see the symmetries that preserve the z-axis. The symmetries of (2) preserve the north and south poles, while (3) interchanges them.

(3) The circles C_r correspond to parallels in S^2, and only C_1 corresponds to a great circle (the equator). We see in Figure 7.1 the unique maximum of the circumferences of these parallels, and that they shrink to a point as you approach one of the poles.

(4) The straight lines through the origin correspond to meridians: great circles through the north and south poles.

Remark 7.6. The Poincaré metric $g_P = 4(du \cdot du)/(1-|u|^2)^2$ of Example 7.4 shares the conformal property of (1) and the symmetry property of (2) of Theorem 7.2, but not the inversion symmetry of (3). We will soon see that, as in the case of g_S, it has many more symmetries that are not visible in the formula for g_P. In fact, the Poincaré metric, along with the spherical metric g_S and the Euclidean metric g_E share the property of being the most symmetric metrics possible in two dimensions. We will see that each has a three-dimensional group of isometries. For the Poincaré metric we do not have the analogues of (4), (5), and (6) of Theorem 7.2: The function $L(C_r)$ is not bounded and has no critical points, the rays through the origin have infinite length, and the area is infinite, just as it happens in the Euclidean metric of $\mathbb{R}^2$.

7.1.6. *Global Riemannian Metrics.* Now we are in a position to define Riemannian metric on any smooth surface. The definition we give is rather primitive, but will be a workable definition.

Definition 7.5. Let S be a smooth surface as in Definition 1.1. Let S be covered by coordinate charts $\phi_\alpha : U_\alpha \to V_\alpha$, where $V_\alpha \subset \mathbb{R}^2$ is open, and let $\phi_{\alpha\beta} = \phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta)$ denote the transition maps, which are smooth. By a *Riemannian metric* g on S we mean:

(1) A Riemannian metric g_α on each $V_\alpha \subset \mathbb{R}^2$.

(2) The Riemannian metrics g_α are compatible in the following sense: whenever $U_\alpha \cap U_\beta \neq \emptyset$, we have that $g_\beta = \phi_{\alpha\beta}^* g_\beta$.

Example 7.6. Let $S = S^2$ covered by two open sets $U_1 = S^2 \setminus \{(0,0,1)\}$ and $U_2 = S^2 \setminus \{(0,0,-1)\}$. For $i = 1, 2$, let $\phi_i : U_i \to \mathbb{R}^2 = V_i$ be stereographic projection. Let $g_1 = g_2 = g_S$, the spherical metric of Example 7.1, see formula (7.12). Then we know that $\phi_{12} = \phi_{21} = \phi : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}^2 \setminus \{0\}$ is inversion in the unit circle. This is the map that in polar coordinates would be given by $\phi(r, \theta) = (\frac{1}{r}, \theta)$. It is a homework problem to check that ϕ is an isometry.

Remark 7.7. This last example is just describing the usual Riemannian metric on S^2, as a surface in $\mathbb{R}^3$, in terms of a cover of S^2 by coordinate charts. Thus the conditions of Definition 7.5 are clear since we already know the object that we are trying to describe, so the existence of the local metrics g_α and their compatibility are both clear in this situation. Ideally a Riemannian metric on a surface would be a structure that we know in some other way, that locally is isometric to some metric in the plane. In this sense Definition 7.5 is a primitive one, in that it does not really define the object, it just says how to describe it locally, and what is needed from the local descriptions for a global object to exist.

Example 7.7. Let $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ be the torus, and let $p : \mathbb{R}^2 \to T^2$ be the projection, that assigns to $(x, y) \in \mathbb{R}^2$ its equivalence class $p(x, y) = \{(x + m, y + n) : m, n \in \mathbb{Z}\} \in T^2$. Define the following subsets of $\mathbb{R}^2$:

$$V_{0,0} = \{(x, y) : \frac{1}{8} < x < \frac{7}{8}, \frac{1}{8} < y < \frac{7}{8}\},$$

an open square centered at the point $(\frac{1}{2}, \frac{1}{2})$. Then let

$$V_{1,0} = V_{0,0} + (\frac{1}{2}, 0), \quad V_{0,1} = V_{0,0} + (0, \frac{1}{2}), \text{ and } V_{1,1} = V_{0,0} + (\frac{1}{2}, \frac{1}{2}),$$

each of which is a translate of $V_{0,0}$. For $i, j = 0, 1$, let $U_{i,j} = p(V_{i,j}) \subset T^2$., see Figure 7.2. Then it is easy to check that for each i, j, $p : V_{i,j} \to U_{i,j}$ is a homeomorphism, and that the four open sets $U_{i,j}$ cover T^2. Let $\phi_{i,j} = p|^{-1}_{V_{i,j}} : U_{i,j} \to V_{i,j}$. The $(U_{i,j}, \phi_{ij})$ give four coordinate charts covering T^2. Let us look at the transition functions. Take, for example, the set $U_{0,0} \cap U_{1,0}$. It is not connected, it has two connected components W_1, W_2, that is, $U_{0,0} \cap U_{1,0} = W_1 \cup W_2$ where

$$W_1 = p((\frac{1}{8}, \frac{3}{8}) \times (\frac{1}{8}, \frac{7}{8})) \text{ and } W_2 = p((\frac{5}{8}, \frac{7}{8}) \times (\frac{1}{8}, \frac{7}{8})),$$

therefore $\phi_{0,0}(U_{0,0} \cap U_{1,0}) = \phi_{0,0}(W_1) \cup \phi_{0,0}(W_2)$, where

$$\begin{aligned} \phi_{0,0}(W_1) &= (\frac{1}{8}, \frac{3}{8}) \times (\frac{1}{8}, \frac{7}{8}), \\ \phi_{0,0}(W_2) &= (\frac{5}{8}, \frac{7}{8}) \times (\frac{1}{8}, \frac{7}{8}) \end{aligned} \tag{7.24}$$

are both subsets of $V_{0,0}$, while $\phi_{1,0}(U_{0,0} \cap U_{1.0}) = \phi_{1,0}(W_1) \cup \phi_{1,0}(W_2)$, where

$$\phi_{0,1}(W_1) = (\frac{9}{8}, \frac{11}{8}) \times (\frac{1}{8}, \frac{7}{8}), \tag{7.25}$$
$$\phi_{0,1}(W_2) = (\frac{5}{8}, \frac{7}{8}) \times (\frac{1}{8}, \frac{7}{8})$$

are both subsets of $V_{1,0}$. Finally the transition function $\phi_{1,0} \circ \phi_{0,0}^{-1}$ restricted to the first component of (7.24) is translation by $(1, 0)$:

$$\phi_{1,0} \circ \phi_{0,0}^{-1}|_{((\frac{1}{8}, \frac{3}{8}) \times (\frac{1}{8}, \frac{7}{8}))}(x, y) = (x + 1, y),$$

while its restriction to the second component of (7.24) is the identity. Similarly the restriction of $\phi_{0,0} \circ \phi_{1,0}^{-1}$ to the first component of (7.25) is translation by $(-1, 0)$, while it is the identity on the second component, see Figure 7.2.

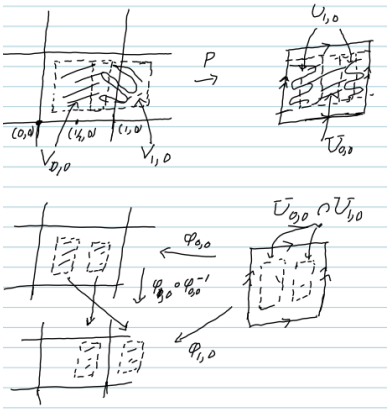


FIGURE 7.2. Transition Functions for T^2.

In the same way we can check that each intersection $U_{i,j} \cap U_{k,l}$ has several components (at most 4) and the transition functions are, on each component, translation by a vector (m, n) where $m, n \in \mathbb{Z}$ (possibly $(m, n) = (0, 0)$ as we have seen), and where the vector (m, n) can be different in each component.

In any case, all the transition functions are isometries of the Euclidean metric. Thus, if we let $g^{0,0}, g^{1,0}$, etc, denote the restriction of the standard Euclidean metric $dx^2 + dy^2$ to $V_{0,0}, V_{0,1}$, etc, (we use superscripts to label the metrics in each chart to avoid conflict with the subscripts used in defining the coefficients of a metric) this collection of local metrics defines a Riemannian metric on T^2 in the sense of Definition 7.5. This metric is *locally Euclidean* in the sense that it is locally isometric to the Euclidean metric on $\mathbb{R}^2$.

Example 7.8. We can easily generalize the preceding example to a description of *all* Riemannian metrics on T^2. Use the same charts $\phi_{i,j} : U_{i,j} \to V_{i,j}$, $i, j = 0, 1$ as in the last example. Let G be any Riemannian metric on $\mathbb{R}^2$ that is *invariant under translation by* $\mathbb{Z}^2$. This means that $G = G_{11}dx^2 + 2G_{12}dxdy + G_{22}dy^2$ where G_{11}, G_{12}, G_{22} are *doubly periodic*

functions on $\mathbb{R}^2$: $G_{ij}(x+m, y+n) = G_{ij}(x,y)$ holds for all $(x,y) \in \mathbb{R}^2$ and for all $(m,n) \in \mathbb{Z}^2$. Then, for all $(m,n) \in \mathbb{Z}^2$, if $T_{(m,n)} : \mathbb{R}^2 \to \mathbb{R}^2$ is translation by (m,n): $T_{(m,n)}(x,y) = (x+m, y+n)$, then $T^*_{(m,n)}G = G$ because $T^*_{(m,n)}G = G_{11}(x+m,y+n)d(x+m)^2 + 2G_{12}(x+m,y+n)d(x+m)d(y+n) + G_{22}(x+m,y+n)d(y_n)^2 = G_{11}(x,y)dx^2 + 2G_{12}(x,y)dxdy + G_{22}(x,y)dy^2 = G$.

Given such a $\mathbb{Z}^2$-translation invariant metric G on $\mathbb{R}^2$, since the transition functions of the collection of charts of Example 7.7 are, on each connected component, translations by elements of $\mathbb{Z}^2$, then we can proceed as follows: For $i,j = 0,1$ let $G^{i,j} = G|_{V_{i,j}}$ (again superscripts to avoid a clash of notation). Then the collection $G^{i,j}$ satisfy Definition 7.5 and thus we get a Riemannian metric on T^2. It is easy to see that the process can be reversed and that this process establishes a *one-to-one correspondence between Riemannian metrics on $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ and $\mathbb{Z}^2$-invariant Riemannian metrics on $\mathbb{R}^2$.*

Remark 7.8. The last three examples illustrate all the examples of Riemannian metrics that we will need: either the metric of a surface in $\mathbb{R}^3$ as $S^2 \subset \mathbb{R}^3$, or the Riemannian metric on a quotient surface $S/\sim$ where S has a Riemannian metric which is compatible with the equivalence relation. We will see more examples later.

7.1.7. *Length and Area for Global Riemannian Metrics.* We need to see that the concepts of Definition 7.2 still make sense. We will avoid defining tangent vectors and restrict ourselves to length and area,

Suppose S is a surface, for simplicity we will assume *compact* and covered by a *finite* collection of charts $\phi_\alpha : U_\alpha \to V_\alpha$, and with metrics g_α on each V_α as in Definition 7.5. In addition, for the definition of integrals, is is convenient to assume that S is *oriented* in the following sense: *the Jacobian determinants of all the transition functions $\phi_{\alpha\beta}$ are positive.*

Let $\gamma : [a,b] \to S$ be a piecewise smooth curve. Define $L(\gamma)$, the *length* of γ, as follows: Using the usual Lebesgue number argument for the cover $\{\gamma^{-1}(U_\alpha)\}$, find $a = t_0 < t_1 < t_2 \cdots < t_n = b$ so that, for each i, $\gamma([t_{i-1}, t_i]) \subset U_{\alpha_i}$, let L_i be the length, in the metric g_{α_i}, of the curve $\phi_{\alpha_i} \circ \gamma : [t_{i-1}, t_i] \to V_{\alpha_i}$, and then define $L(\gamma)$ by

$$L(\gamma) = L_1 + L_2 + \cdots + L_n. \tag{7.26}$$

We need to check that this definition is independent of all choices. This follows easily from the following considerations:

(1) Suppose one of the intervals $[t_{i-1}, t_i]$ is mapped by γ to $U_\alpha \cap U_\beta$. If we compute L_i by using $\phi_\alpha \circ \gamma$ or $\phi_\beta \circ \gamma$ we get the same answer because the transition function $\phi_{\alpha\beta} = \phi_\alpha \circ \phi_\beta^{-1}$ takes $\phi_\beta \circ \gamma$ to $\phi_\alpha \circ \gamma$ and by (2) of Definition 7.5, $\phi_{\alpha\beta}$ is an isometry, so these two curves have the same length and we get the same answer for L_i.

(2) If we use two different partitions of the interval, take a common refinement, use a longer sum in (7.26) and apply the reasoning additivity of length and the reasoning just used to get $L(\gamma)$ to be well defined.
(3) If we use two different collections of charts, take the collection which is their union and apply the previous reasoning.

To define area and, more generally, integrals of functions, we need a trick called the *partition of unity subordinate to a collection of charts*. This means the following: a collection of smooth functions ρ_α with the property that there is a (finite) collection of charts $\phi_\alpha : U_\alpha \to V_\alpha$ as in Defintion 7.5 so that the *support* of each ρ_α is contained in U_α, $\rho \geq 0$, *and*

$$\sum_\alpha \rho_\alpha = 1. \tag{7.27}$$

The *support* of a real valued function f means the *closure* of $\{x : f(x) \neq 0\}$.

There is a standard construction of partitions of unity that can be quickly summarized as follows: We can assume that the charts have the following properties: for each α there is an open subset $U'_\alpha \subset \overline{U'_\alpha} \subset U_\alpha$ so that the $\{U'_\alpha\}$ covers S, that the $V_\alpha \subset \mathbb{R}^2$ are diffeomorphic to disks D of radius 2 and that the images $\phi_\alpha(U'_\alpha)$ are concentric disks $D' \subset D$ of radius 1. We can construct a function $\psi : D \to \mathbb{R}$ with support $\overline{D'}$ by taking the "bump function" of Lemma 6.1 (pictured in Figure 6.3) of [12], putting it on a radius of D with its maximum at the center and support exactly at distance one (that is, using $(a, b) = (-1, 1)$ in Lemma 6.1 of [12]), then extending it in a rotationally symmetric way to D. Then let $\psi_\alpha : S \to \mathbb{R}$ be defined as follows: on U_α, $\psi_\alpha = \psi \circ \phi_\alpha$, and then extend it to be zero on the rest of S. This gives a smooth function on S because it is smooth on U_α and its support is contained in U_α. Finally, for each α in the index set A of the cover, let

$$\rho_\alpha = \frac{\psi_\alpha}{\sum_{\alpha' \in A} \psi_{\alpha'}}.$$

Then these functions, by construction, are ≥ 0, have the required support, and satisfy (7.27)

If $f : S \to \mathbb{R}$ is a smooth function, then, for each α, $\rho_\alpha f$ has support in U_α, so we can define a function $f_\alpha : V_\alpha \to \mathbb{R}$ by $f_\alpha = (\rho f) \circ (\phi_\alpha^{-1})$. Then we can define

$$\iint_S f \, dA = \sum_\alpha \iint_{V_\alpha} f_\alpha \, \sqrt{\det g_\alpha} \, du_1 du_2. \tag{7.28}$$

Again, we have to check, by refinement and change of variable formulas, that this is independent of all choices. Finally, the *area* of S is $A(S) = \iint_S 1 \, dA$.

7.2. **Geodesics.** We next quickly review the definition and basic properties of geodesics from Section 6.2 of [12] and also show that the same definitions and properties hold in any Riemannian metric on a surface.

7.2.1. *Review of First Variation Formula and Geodesic Equation.* Recall form Section 6.2.1 of [12] the First Variation Formula for arclength, in the context of a parametrized surface $\mathbf{x} : U \to S \subset \mathbb{R}^3$, where $\mathbf{x}$ is smooth and $\mathbf{x}_1 \times \mathbf{x}_2 \neq 0$, so that $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent at each point and form a basis for the tangent plane $T_{\mathbf{x}(u)}S$ at each $u \in U$. We considered a smooth curve $\gamma : [0, L_0] \to S$, parametrized by arclength, and a variation

$$\tilde{\gamma} : [0, L_0] \times (-\epsilon, \epsilon) \to S \text{ with } \tilde{\gamma}(s, 0) = \gamma(s) \text{ for all } s \in [0, L_0].$$

We computed the length $L(t)$ of the curve $\tilde{\gamma}(\ , t)$, and found the following formulas for $L'(t)$. With the view of generalizing them to any Riemannian metric, we will write each formula in two ways: first, just as they appeared in [12], using the dot product in $\mathbb{R}^3$ and ordinary derivatives, and then using the Riemannian inner product $<\ ,\ >$ and *covariant derivatives*. Recall that these were defined in Definition 6.3 of [12]: if $V(s)$ is a smooth vector field along a smooth curve $\gamma(s)$ in S, then its *covariant derivative* $\frac{DV}{Ds}$, also denoted $D_{\gamma'}V$, is, by definition

$$D_{\gamma'}V = \frac{DV}{Ds} = V'(s)^T = \text{ the tangential component of } V'(s). \tag{7.29}$$

Here are the formulas, written in the two notations. First

$$L(t) = \int_0^{L_0} (\tilde{\gamma}_s(s,t) \cdot \tilde{\gamma}_s(s,t))^{1/2}\, ds = \int_0^{L_0} <\tilde{\gamma}_s(s,t), \tilde{\gamma}_s(s,t)>^{1/2}\, ds. \tag{7.30}$$

Differentiating with respect to t:

$$\frac{dL}{dt} = \int_0^{L_0} \frac{1}{2}(\tilde{\gamma}_s(s,t) \cdot \tilde{\gamma}_s(s,t))^{-1/2}(2\ \tilde{\gamma}_{st}(s,t) \cdot \tilde{\gamma}_s(s,t))\, ds = \int_0^{L_0} \frac{1}{2} <\tilde{\gamma}_s(s,t), \tilde{\gamma}_s(s,t)>^{-1/2} (2\ < \frac{D}{Dt}\ \tilde{\gamma}_s(s,t), \tilde{\gamma}_s(s,t)>)\, ds. \tag{7.31}$$

Note that in this formula we replaced $\tilde{\gamma}_{st}$ by its tangential component $\frac{D}{Dt}\tilde{\gamma}_s$ because we are taking its dot product with the tangential vector $\tilde{\gamma}_s$, so only the tangential component of $\tilde{\gamma}_{st}$ contributes to the formula: $\tilde{\gamma}_{st} \cdot \tilde{\gamma}_s = \tilde{\gamma}_{st}^T \cdot \tilde{\gamma}_s$. The same reasoning applies to all the formulas below where we replace ordinary derivatives by covariant ones.

Evaluating (7.31) at $t = 0$ and using that $|\gamma_s| = 1$ we get

$$\frac{dL}{dt}(0) = \int_0^{L_0} \tilde{\gamma}_{st}(s,0) \cdot \tilde{\gamma}_s(s,0)\, ds = \int_0^{L_0} < \frac{D}{Dt}\tilde{\gamma}_s(s,0), \tilde{\gamma}_s(s,0) >\, ds. \tag{7.32}$$

Then, integrating by parts, using the formula

$$(\tilde{\gamma}_t(s,0) \cdot \tilde{\gamma}_s(s,0))_s = \tilde{\gamma}_{ts}(s,0) \cdot \tilde{\gamma}_s(s,0) \; + \; \tilde{\gamma}_t(s,0) \cdot \tilde{\gamma}_{ss}(s,0) \text{ or} \tag{7.33}$$

$$< \tilde{\gamma}_t(s,0), \tilde{\gamma}_s(s,0) >_s = < \frac{D}{Ds}\tilde{\gamma}_t(s,0), \tilde{\gamma}_s(s,0) > \; + \; < \tilde{\gamma}_t(s,0), \frac{D}{Ds}\tilde{\gamma}_s(s,0) >,$$

where, in order to apply (7.33) to (7.32), we need to use the equality of mixed second partial derivatives:

$$\tilde{\gamma}_{st}(s,t) = \tilde{\gamma}_{ts}(s,t) \;\text{ or }\; \frac{D}{Dt}\tilde{\gamma}_s(s,t) = \frac{D}{Ds}\tilde{\gamma}_t(s,t), \tag{7.34}$$

where the second part follows from the first since $\tilde{\gamma}_{st} = \tilde{\gamma}_{ts}$ implies equality of tangential components: $\tilde{\gamma}_{st}^T = \tilde{\gamma}_{ts}^T$.

Using (7.34) to change the order of differentiation in (7.32) and then using (7.33) to integrate by parts the interchanged version of (7.32), we get the *First Variation Formula*

$$\frac{dL}{dt}(0) = V(s) \cdot \gamma'(s)|_0^{L_0} \; - \; \int_0^{L_0} \; V(s) \cdot \gamma''(s)^T \, ds = \tag{7.35}$$

$$< V(s), \gamma'(s) > |_0^{L_0} \; - \; \int_0^{L_0} \; < V(s), \frac{D\gamma'}{Ds}(s) > \, ds,$$

where $V(s) = \tilde{\gamma}_t(s,0)$ is the *variation vector field* .

From this equation we proved Theorem 6.5 of [12]: if $L'(0) = 0$ for *all variations* $\tilde{\gamma}$ *that keep the endpoints fixed*, then γ satisfies the *geodesic equation*

$$\frac{D\gamma'}{Ds} = 0. \tag{7.36}$$

To study this equation more closely, in particular, to apply the standard theory of second-order ordinary differential equations, we wrote it down more explicitly in local coordinates. Namely, if $\gamma(s) = \mathbf{x}(u_1(s), u_2(s))$, then $\gamma'(s) = u_1'\mathbf{x}_1 + u_2'\mathbf{x}_2$, so

$$\gamma'' = u_1''\mathbf{x}_1 + u_2''\mathbf{x}_2 + (u_1')^2\mathbf{x}_{11} + 2u_1'u_2'\mathbf{x}_{12} + (u_2')^2\mathbf{x}_{22}.$$

Since the tangential component of γ'' vanishes if and only if $\gamma'' \cdot \mathbf{x}_1 = 0$ and $\gamma'' \cdot \mathbf{x}_2 = 0$, we get that γ satisfies (7.36) if and only if

$$u_1''\mathbf{x}_1 \cdot \mathbf{x}_1 + u_2''\mathbf{x}_2 \cdot \mathbf{x}_1 + (u_1')^2\mathbf{x}_{11} \cdot \mathbf{x}_1 + 2u_1'u_2'\mathbf{x}_{12} \cdot \mathbf{x}_1 + (u_2')^2\mathbf{x}_{22} \cdot \mathbf{x}_1 = 0$$
$$u_1''\mathbf{x}_1 \cdot \mathbf{x}_2 + u_2''\mathbf{x}_2 \cdot \mathbf{x}_2 + (u_1')^2\mathbf{x}_{11} \cdot \mathbf{x}_2 + 2u_1'u_2'\mathbf{x}_{12} \cdot \mathbf{x}_2 + (u_2')^2\mathbf{x}_{22} \cdot \mathbf{x}_2 = 0,$$

which, in our current notation, can be written as

$$\begin{aligned} g_{11}u_1'' + g_{12}u_2'' + \Gamma_{11,1}u_1'^2 + 2\Gamma_{12,1}u_i'u_2' + \Gamma_{22,1}u_2'^2 = 0 \\ g_{21}u_1'' + g_{22}u_2'' + \Gamma_{11,2}u_1'^2 + 2\Gamma_{12,2}u_i'u_2' + \Gamma_{22,2}u_2'^2 = 0, \end{aligned} \tag{7.37}$$

where the $g_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$ as before and the $\Gamma_{ij,k} = \mathbf{x}_{ij} \cdot \mathbf{x}_k$ are some smooth coefficient functions.

7.2.2. *The Geodesic Equations in any Riemannian Metric.* Let us examine the Equations (7.37) more closely. We have that $\Gamma_{11,1} = \mathbf{x}_{11} \cdot \mathbf{x}_1 = \frac{1}{2}(\mathbf{x}_1 \cdot \mathbf{x}_1)_1 = \frac{1}{2}\frac{\partial g_{11}}{\partial u_1}$ which we abbreviate as $\frac{1}{2}(g_{11})_1$. Continuing in this way we find the following formulas for the $\Gamma_{ij,k}$:

$$\begin{aligned} \Gamma_{11,1} = \frac{1}{2}(g_{11})_1, \quad \Gamma_{12,1} = \frac{1}{2}(g_{11})_2, \\ \Gamma_{22,1} = (g_{21})_2 - \frac{1}{2}(g_{22})_1, \quad \Gamma_{11,2} = (g_{12})_1 - \frac{1}{2}(g_{11})_2 \\ \Gamma_{12,2} = \frac{1}{2}(g_{22})_1, \quad \Gamma_{22,2} = \frac{1}{2}(g_{22})_2. \end{aligned} \tag{7.38}$$

which can be summarized as $\Gamma_{ij,k} = \frac{1}{2}((g_{jk})_i + (g_{ik})_j - (g_{ij})_k)$. Observe, in particular, the symmetry $\Gamma_{ij,k} = \Gamma_{ji,k}$

The point of these formulas is that the coefficients of the geodesic equations (7.37) *depend only on the functions g_{ij} and their derivatives $(g_{ij})_k$, therefore they make sense in any Riemannian metric.* This illustrates the general principle that we will use several times: *Any concept or formula that is well defined for surfaces in $\mathbb{R}^3$, and that is* intrinsic *in the sense that it involves only the Riemannian metric, is also well defined and holds for any Riemannian metric on any smooth surface.*

We give three illustrations of this principle, one in each of the following three sections:

7.2.3. *Covariant Derivatives and Variation Formula for any Riemannian Metric.* The derivation of the first variation formula for a surface in $\mathbb{R}^3$ used not just the geodesic equation, but also the covariant derivatives and various properties of these derivatives. Since length makes sense in any Riemannian metric, all parts of the reasoning should also make sense. We give more details.

Given any local Riemannian metric g, we not only have the geodesic equations (7.37), but we have covariant derivatives and the whole derivation of the first variation formula (and the second variation formula later on). We can, as in Equation (6.12) of [12], solve (7.37) for u_1'', u_2'' by multiplying by the matrix g^{-1} (inverse to g) and get the system

$$\begin{aligned} u_1'' + \Gamma^1_{11}(u_1')^2 + 2\Gamma^1_{12}u_1'u_2' + \Gamma^1_{22}(u_2')^2 = 0 \\ u_2'' + \Gamma^2_{11}(u_1')^2 + 2\Gamma^2_{12}u_1'u_2' + \Gamma^2_{22}(u_2')^2 = 0 \end{aligned} \tag{7.39}$$

where the coefficient functions Γ^k_{ij} have some explicit expression in the coefficients $\Gamma_{ij,k}$ of (7.37) and the entries of the inverse matrix g^{-1}. We can view the left hand side as the definition of $\frac{D\gamma'}{Ds}$, and take one more step: look at this formula as a special case of the definition of the covariant derivative $\frac{DV}{Ds}$ of *any* vector field $V(s)$ along γ. If $\gamma(s) = (u_1(s), u_2(s))$ and $V(s) = \xi_1(s)\mathbf{e}_1 + \xi_2(s)\mathbf{e}_2$ is a vector based at $\gamma(s)$, then we replace the Equation (7.29) that only makes sense for a surface in $\mathbb{R}^3$ by the expression that we would get for this same object if we were to compute it in a local parametrization, that is,

$$\begin{aligned}\frac{DV}{Ds} &= \Big(\xi_1' + \Gamma^1_{11}\xi_1 u_1' + \Gamma^1_{12}\xi_1 u_2' + \Gamma^1_{21}\xi_2 u_1' + \Gamma^1_{22}\xi_2 u_2'\Big)\mathbf{e}_1 \\ &\quad + \Big(\xi_2' + \Gamma^2_{11}\xi_1 u_1' + \Gamma^2_{12}\xi_1 u_2' + \Gamma^2_{21}\xi_2 u_1' + \Gamma^2_{22}\xi_2 u_2'\Big)\mathbf{e}_2.\end{aligned} \tag{7.40}$$

Note that (7.40) reduces to (7.39) when $V = \gamma'$, thus $\xi_i = u_i'$. This reduction relies on the symmetry

$$\Gamma^k_{ij} = \Gamma^k_{ji} \tag{7.41}$$

which results from the symmetry $\Gamma_{ij,k} = \Gamma_{ji,k}$ noted after (7.38).

In order to derive the first variation formula (8.6), we need Equation (7.31) that in turn requires the identity

$$\frac{d}{dt} < V(t), W(t) > = < \frac{DV}{Dt}, W(t) > + < V(t), \frac{DW}{Dt} > \tag{7.42}$$

for any two vector fields along a curve $\gamma(t)$. Next, for Equation (7.34) we need the identity

$$\frac{D\gamma_s}{Dt} = \frac{D\gamma_t}{Ds} \tag{7.43}$$

for any mapping $\gamma : R \to U$ where R is a rectangle with coordinates s, t. These two formulas hold if we define the covariant derivatives by Equation (7.40). In more detail, (7.42) holds because, if the two vector fields V, W have components ξ_1, η_i respectively, if gives an identity for $\frac{d}{dt}\sum(g_{ij}\xi_i\eta_j)$ that involves derivatives of the g_{ij} and can be seen to be equivalent to the definition (7.38) of the $\Gamma_{ij,k}$ in terms of these derivatives. Formula (7.43) follows from the symmetry of the second partial derivatives *and* the symmetry $\Gamma^k_{ij} = \Gamma^k_{ji}$ of (7.41).

Finally, we record the identity

$$\frac{D(fV)}{Dt} = f'V + f\frac{DV}{Dt}, \tag{7.44}$$

where f is any smooth function along γ. This easily verified identity is not needed for the first variation formula but that we will need below for the second variation and Jacobi's differential equation.

With these formulas we can derive the first variation formula for any Riemannian metric on any $U \subset \mathbb{R}^2$. Moreover, these formulas are invariant

under isometry: An isometry between $f : (U, h) \to (V, g)$ between two Riemannian metrics carries the structures from one to those of the other. This is the basic principle that allows us to extend the local definitions on open sets in $\mathbb{R}^2$ to any smooth surface S by using a covering by coordinate charts and giving the definition chart by chart. This works because the transition functions are isometries. This is what we did in Section 7.1.7 to check that the length of curves is well-defined.

To check that vector quantities, like the covariant derivative, are well defined requires more elaborate formulas than for scalar quantities like the length of a curve. For example, if V is a vector field along a curve γ in U, then the covariant derivatives D^h along γ for h and D^g along $f \circ \gamma$ for g are related by

$$D^g_{df(\gamma')}(df(V)) = df(D^h_{\gamma'}V), \quad \text{where } df = d_{\gamma(s)}f \text{ and } V = V(\gamma(s)).$$

Here it is convenient to use the first (and sometimes more accurate) notation in (7.29) for the covariant derivative. The textbooks in differential geometry develop the mechanism for verifying such formulas, but we will not go into these details. But, as in Remark 7.8, we will not need this generality.

7.2.4. *Existence and Uniqueness of Geodesics.* Here is one useful example of the principle just explained. In any Riemannian metric the system (7.39) of second order ODE's has the same properties studied in Section 6.2 of [12]. In particular, given any Riemannian metric on $U \subset \mathbb{R}^2$, any point $p \in U$ and any tangent vector $\mathbf{v}$ to U at p, there exists unique geodesic $\gamma(s, p, \mathbf{v})$, defined perhaps only for $|s|$ small, but depending smoothly on all variables $s, p, \mathbf{v}$, so that $\gamma(0, p, \mathbf{v}) = p$ and $\gamma_s(0, p, \mathbf{v}) = \mathbf{v}$.

Moreover, the uniqueness implies the identity (6.13) of [12]: $\gamma(rs, p, \mathbf{v}) = \gamma(s, p, r\mathbf{v})$, and, just as in Theorem 6.7 of [12], we get the following consequences: there is some $b > 0$ so that $\gamma(1, p, \mathbf{v})$ is defined on the ball $B^g(0, b)$ in the tangent plane to U at p, the superscript g to emphasize that the size is being measured by the inner product g on this space. We therefore can define the *exponential map* $\exp_p : B^g(0, b) \to U$ by $\exp_p(\mathbf{v}) = \gamma(1, p, \mathbf{v})$ and there is an $\epsilon > 0$ so that the restriction of $\exp_p$ to $B^g(0, \epsilon)$ is a diffeomorphism onto its image.

The geodesic equation, and thus its solutions, must be invariant under isometries: if $f : (U, g) \to (V, h)$ is an isometry, then, using superscripts to denote the metrics, we have the identities

$$\begin{aligned} &\gamma^h(s, f(p), d_p f(\mathbf{v})) = f(\gamma^g(s, p, \mathbf{v})) \text{ and} \\ &\text{therefore } \exp^h_{f(p)}(d_f(\mathbf{v})) = f(\exp^g_p(\mathbf{v})) \end{aligned} \tag{7.45}$$

This invariance has a useful consequence:

Theorem 7.3. *Let $U \subset \mathbb{R}^2$ be an open set, let g be a Riemannian metric on U, let $f : (U, g) \to (U, g)$ be an isometry, and let F be its fixed-point set: $F = \{p \in U : f(p) = p\}$. Then, for every $p \in F$, there is a neighborhood V of p in U and a neighborhood V' of 0 in T_pU so that $V \cap F = \exp_p(V' \cap F')$, where $F' = \{\mathbf{v} \in T_pU : d_pf(\mathbf{v}) = \mathbf{v}\}$ is the fixed-point set of d_pf, where T_pU is the space of tangent vectors to U at p.*

Proof. Let $p \in F$ and choose $\epsilon > 0$ so that $\exp_p$ maps $B^g(0, \epsilon)$ diffeomorphically to its image, as explained above. Let $V' = B^g(0, \epsilon)$ and $V = \exp(V')$.

Suppose $q \in V$. then $q = \exp_p(\mathbf{v})$ for a unique $\mathbf{v} \in V'$ and (7.45) implies that $f(q) = \exp_p(d_p(\mathbf{v}))$. Therefore $f(q) = q$ if and only if $d_pf(\mathbf{v}) = \mathbf{v}$. □

Corollary 7.1. *Let S be a smooth, connected surface with a Riemannian metric, let $f : S \to S$ be an isometry, and let C be a connected component of its fixed point set F. Then either*

1. *C consists of a single point.*
2. *C is a non-singular curve in S and is a geodesic.*
3. *$C = S$.*

Proof. Let $p \in C$, take a coordinate chart U so that $p \in U$, thus reducing to $U \subset \mathbb{R}^2$, and apply the Theorem. Since the fixed-point set F" of d_pf is a linear subspace of T_pU, it is either 0, a line L through 0 or all of T_pU. So locally we have one of the three possibilities of the Corollary. Here, non-singular curve means that locally it is like a line in the plane, which is he case here, and note that $C \cap V$ is a geodesic in this case.

If $F' = 0$ we have clearly $C = p$. If F' is a line, let $q \in C$, take a path from p to q in C, cover it by finitely many $V_1, \ldots, V_k$ as in the Theorem, so that $V_i \cap V_{i+1} \neq \emptyset$, and argue that for all such V_i we must be in case (2) since V_1 is and $V_i \cap V_{i+1}$ is of the same case as V_i and V_{i+1}.

If $F' = T_pU$, and $q \in C$, again send a path from p to q and argue in exactly the same way to get that $F' = T_qU$, thus $F = id$, equivalently, since S is connected, $F = S$. □

Example 7.9. Let g_S be the spherical metric of Equation (7.12). By Theorem 7.2 we know that the isometries fixing the origin of the Euclidean metric are isometries of g_S. In particular, for each line L through the origin, the reflection in this line is an isometry of g_S, and has L as its fixed-poiint set. Therefore every line L through the origin is a geodesic of g_S, meaning that, when reparametrized by arclength it satisfies the geodesic equation. These lines parametrize the great circles through the north and south poles of S^2, hence we know independently that the must be geodesics.

Similarly, for the Poincaré metric g_S of Equation (7.22) on $\{|u| < 1\}$ also has the reflections in the lines L through the origin as isometries (same proof

as for g_S in Theorem 7.2), thus the intersections of this lines with the disk are fixed point sets of isometries, hence geodesics.

Similarly, we know from Theorem 7.2 that the circle $|u| = 1$ is the fixed point set of another isometry: inversion in the unit circle.

Finally, another useful consequence of Theorem 7.3. Note that we have not defined tangent spaces to arbitrary surfaces, nor the differential of a smooth map of surfaces. But the condition that $d_p f = id$ could be defined as $d_p f = id$ in any coordinate chart containing p. This condition makes sense since it is independent of the chart.

Corollary 7.2. *Let S be a smooth, connected surface with a Riemannian metric, and suppose $f : S \to S$ is an isometry, and suppose there is a point $p \in S$ so that $d_p f = id$. Then $f = id$.*

Proof. Since $d_p f = id$, by (7.45) we see that $f = id$ in some neighborhood of p. Then Corollary 7.1 only leaves the possibility $F = S$, that is, $f = id$. □

7.2.5. *Geodesic Equations in Geodesic Polar Coordinates.* Recall that in Section 6.2.4 of [12] we defined geodesic polar coordinates (also called normal coordinates) and used them to study the minimizing properties of geodesics. More precisely, given any point P in a surface (and now we can say in any surface with a Riemannian metric) we can introduce, in some neighborhood of P, a polar coordinates system (r, θ) centered at P in which the metric g takes the form

$$ds^2 = dr^2 + f(r, \theta)^2 d\theta^2 \tag{7.46}$$

where $f(r, \theta)$ is a positive smooth function that satisfies

$$f(r, \theta) = r - \frac{K(P)}{6} r^3 + O(r^4), \tag{7.47}$$

see Theorem 6.2 of [12] (where the function we call f is denoted by g). The significance of the number $K(P)$, which we called the *Gaussian curvature*, will be discussed shortly.

This was all derived in Sections 6.2 and 6.3 of [12] for surfaces in $\mathbb{R}^3$ by using the first variation formula and the theory of second order ODE's. In particular we proved Theorem 6.7 to derive the existence and of normal coordianates, and Gauss's Lemma, Theorem 6.8, to derive the special form (7.46) of the metric in geodesic polar coordinates. The discussion of the last section (7.2.3) means that *the same results hold for any Riemannian metric on any surface.*

It is clear from (7.46) that the geodesics through the pole P, in other words, through the origin $r = 0$ of the polar coordinate system, are given by the lines $\theta = const.$ The geodesics not through the center are harder to determine. We will use Equations (7.37) to get some information on

these other geodesics. Note that (7.46) mean that $g_{11} = 1, g_{12} = 0$ and $g_{22} = f^2$. Thus in the geodesic equations (7.37) we get that any coefficient that involves only g_{12} or derivatives of g_{11} or g_{12} has to vanish. Thus the only $\Gamma_{ij,k}$ that do not vanish are $\Gamma_{22,1}, \Gamma_{21,2}$ and $\Gamma_{22,2}$. A short computation gives that the equations simplify to

$$r'' - ff_r(\theta')^2 = 0 \tag{7.48}$$
$$f\theta'' + 2f_r r'\theta' + f_\theta(\theta')^2 = 0.$$

This is the only explicit computation of the geodesic equation that we will need.

8. The Gauss-Bonnet Theorem

In this Chapter we will prove the Gauss-Bonnet theorem that will unify much of the course. This is the formula (1.2) that involves relates the topology and geometry of surfaces. We have now defined all the individual elements of the formula, but the definition of K is, at the moment, not very illuminating. We begin by giving the original definition.

8.1. **The Gauss Map and the Gaussian Curvature.** Let $S \subset \mathbb{R}^3$ be a smooth surface. At every point $p \in S$ there are two unit normal vectors to S, differing by sign. Choose one, call it $\mathbf{N}(p)$, and assume that it is possible to choose $\mathbf{N}(p)$ so that it is a continuous function of p. This happens to be equivalent to orientability, so it may not always be possible (say, for a Möbius band). But locally it is always possible. If $\mathbf{x} : U \to S$ is a regular parametrization then (following the notation of Section 7.1)), the partial derivatives $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent at each point, in other words, the cross product $\mathbf{x}_1 \times \mathbf{x}_2 \neq 0$ and we can take $\mathbf{N} = \mathbf{x}_1 \times \mathbf{x}_2/|\mathbf{x}_1 \times \mathbf{x}_2|$. For the present purposes we cajn say that to *orient* a surface it means to make a continuous choice of normal vector field $\mathbf{N}$. We can move the normal vectors to be based at the origin and this gives the *Gauss map*:

Definition 8.1. Let $S \subset \mathbb{R}^3$ be a smooth, oriented surface.

(1) The *Gauss map* of S is the map $\mathbf{N} : S \to S^2$ that assigns to $p \in S$ the unit normal vector $\mathbf{N}(p)$. Explicitly, if $\mathbf{x} : U \to S$ is a parametrization of (part of) S, we choose
$$\mathbf{N} = \frac{\mathbf{x}_1 \times \mathbf{x}_2}{|\mathbf{x}_1 \times \mathbf{x}_2|}.$$
(2) The *Gaussian curvature* $K(p)$ of S at p is defined to be the signed area distortion of $\mathbf{N}$ at p, namely
$$\pm \lim_{r \to 0} \frac{A(\mathbf{N}(D_r))}{A(D_r)}, \tag{8.1}$$
where D_r is a geodesic disk of radius r centered at p, choosing the $+$ sign if $\mathbf{N}$ is orientation preserving at p, and the $-$ sign otherwise.

Example 8.1. (1) Let $S = S^2(R)$ be the sphere $|\mathbf{x}| = R$ of radius R centered at the origin, Then the Gauss map is $\mathbf{N}(\mathbf{x}) = \frac{\mathbf{x}}{R}$, the map is orientation-preserving, and the ratios in (8.1) are $1/R^2$ since the area measurements in $S^2(R)$ are R^2 times those in $S^2 = S^2(1)$. Thus the Gaussian curvature $K \equiv 1/R^2$ for $S^2(R)$.

(2) If S is contained in a plane in $\mathbb{R}^3$ then $\mathbf{N}$ is a constant map and $K \equiv 0$.

(3) More generally, if S is a "cylinder" $\mathbf{x}(u_1, u_2) = (x(u_1), y(u_1), u_2)$ on a plane curve $\gamma(s) = (x(s), y(s))$ as in (6.21) of [12], then, assuming $\gamma'(s) \equiv 1$, we get

$$\mathbf{N}(u_1, u_2) = \frac{(x'(u_1), y'(u_1), 1)}{\sqrt{2}},$$

which is a curve in S^2. Thus the Gauss map has one-dimensional image, encloses no area and the numerator in (8.1) vanishes, thus $K \equiv 0$ also in this case, see Figure 8.1.

(4) Even though it is difficult from (8.1) to find the exact value of K, it is easy to see when $K > 0$, $K \equiv 0$ or $K =< 0$. We have seen examples of the first two situations, and a saddle illustrates the third, see Figure 8.1.

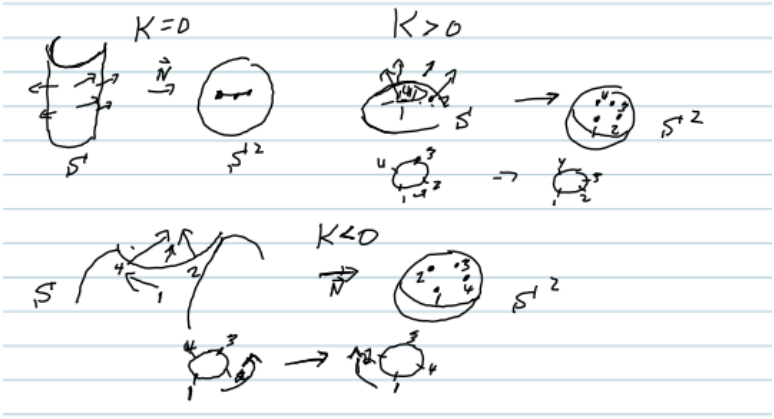


FIGURE 8.1. sign of the Gaussian Curvature.

To get an explicit formula for K, recall the formulas for area from Remark 7.3. The numerator in (8.1) is $\iint_D |\mathbf{N}_1 \times \mathbf{N}_2| \, du_1 du_2$ (for suitable domains $D \subset U$) while the denominator is $\iint_D |\mathbf{x}_1 \times \mathbf{x}_2| \, du_1 du_2$. It follows that their ratio converges, as the domains D shrink to a point $u \in U$, to the ratio of the integrands at u, in other words,

$$K(u) = \pm \frac{|\mathbf{N}_1 \times \mathbf{N}_2|}{|\mathbf{x}_1 \times \mathbf{x}_2|} = \pm \frac{|\mathbf{N}_1 \times \mathbf{N}_2|}{\sqrt{\det g}}, \tag{8.2}$$

where the second equation results from the computation discussed in Remark 7.3 (and taken as the definition of dA in Equation (7.19) of Definition 7.2).

It remains to determine the sign $\pm$. Note that the orientation of S is given by the vector $\mathbf{x}_1 \times \mathbf{x}_2$ which is parallel to N, and the orientation in S^2 given by the map $\mathbf{N}$ is that of $\mathbf{N}_1 \times \mathbf{N}_2$. So the two orientations agree if $\mathbf{N}_1 \times \mathbf{N}_2$ is a positive multiple of $\mathbf{N}$, and disagree if this multiple is negative. Observing that this multiple is $(\mathbf{N}_1 \times \mathbf{N}_2) \cdot \mathbf{N}$ and that its absolute value is $|\mathbf{N}_1 \times \mathbf{N}_2|$, we see that the correct signed denominator is the "scalar triple product" $(\mathbf{N}_1 \times \mathbf{N}_2) \cdot \mathbf{N}$ and we get the final formula for K:

$$K = \frac{(\mathbf{N}_1 \times \mathbf{N}_2) \cdot \mathbf{N}}{\sqrt{\det g}}. \tag{8.3}$$

8.1.1. *Gauss's Theorema Egregium.* In Remark 6.4 and Section 6.4 of [12] we quickly explained the distinction between the *intrinsic* and the *extrinsic* geometry of surfaces in $\mathbb{R}^3$. Extrinsic geometry is basically the study of the *shape* of the surface in $\mathbb{R}^3$ while intrinsic geometry is the study of the Riemannian metric, not distinguishing two surfaces of different shapes if their Riemannian metrics are isometric. For example, the cylinders of Example 8.1 have infinitely many different shapes, as many as the shapes of the plane curves γ used to define them, but they are all isometric to a subset of the plane, so they are all intrinsically equal.

There is a very detailed theory of the extrinsic geometry of surfaces, base on the following technique: at each $p \in S$, write S as the graph of a function f (depending on p)) from the tangent plane T_pS to the normal line at p (multiples of $\mathbf{N}(p)$). This is a quadratic form, called the *second fundamental form of S at p*. (In the classical terminology, the Riemannian metric is called the *first fundamental form*). Properly interpreted, the second fundamental form is the differential $d\mathbf{N}$ of the Gauss map, and the Gauss curvature is the determinant of $d\mathbf{N}$.

Thus the definition of K in Definition 8.1 is *extrinsic*. Gauss discovered that K is actually *intrinsic* and called this the "Theorema Egregium" (usually translated as "Remarkable Theorem'). The book [4] is a good source for its history, which goes as follows:

(1) Gauss discovered the following formula: Let Δ be a geodesic triangle in S with interior angles α, β, γ. Then

$$\iint_\Delta K \; dA = \alpha + \beta + \gamma - \pi. \tag{8.4}$$

(2) Since every member of this equation, except K, is clearly intrinsic, it follows that K must be intrinsic.
(3) Since K is intrinsic, it must have an explicit expression in terms of the g_{ij}.

(4) To find such an expression, Gauss first tried an easier special case, geodesic normal coordinates $dr^2 + f(r,\theta)^2\, d\theta^2$ as in Equation (7.46). He found the formula

$$K = -\frac{f_{rr}}{f}. \tag{8.5}$$

(5) He then found a formula valid in any coordinate system.

Gauss, however, did not publish his results in this way, his famous published paper presents these developments in exactly the opposite order, see [4] for more information. Most books in differential geometry follow this reverse order of presentation.

Here we will prove that K is intrinsic and the formula (8.4) in a different way: will use the second variation formula to prove that K is intrinsic, then will derive (8.5), and (8.4) will follow easily from this and the equations (7.48) for geodesics not through the pole.

8.2. **The Second Variation Formula.** We know that geodesics are critical points of the length function, but not necessarily minima. To better understand the nature of critical points we know, from calculus, that we should look at second derivatives at critical points. This is called the *second variation formula*. We will follow the notation of Section 7.2.1.

We now consider a *geodesic* $\gamma : [0, L_0] \to S$, parametrized by arclength, and a variation

$$\tilde{\gamma} : [0, L_0] \times (-\epsilon, \epsilon) \to S \text{ with } \tilde{\gamma}(s, 0) = \gamma(s) \text{ for all } s \in [0, L_0]. \tag{8.6}$$

To simplify some formulas we will assume that the variation is *normal* meaning that $< \tilde{\gamma}_t(s,0), \tilde{\gamma}_s(s,0) >\equiv 0$. We will choose a unit vector $\mathbf{n}$ along γ perpendicular to the tangent vector γ', and, if we have chosen an orientation on S, we will assume that the pair $\gamma', \mathbf{n}$ is positively oriented, that is, $\mathbf{n}$ is obtained by a counterclockwise rotation of γ' by a right angle. In this notation we get that the *variation vector field* $V(s)$ satisfies

$$V(s) = \tilde{\gamma}_t(s,0) = f(s)\mathbf{n}(s) \tag{8.7}$$

for some smooth real-valued function f defined along γ.

We start with the same formula (7.30)for the length $L(t)$ and get the formula (7.31) for $L'(t)$, but, in order to find $L''(0)$ we cannot go on to (7.32) which sets $t = 0$ at this point, we have to first take one more derivative.

In order to do this, we work with (7.31): we replace $\frac{D\tilde{\gamma}_s}{Dt}$ with $\frac{D\tilde{\gamma}_t}{Ds}$ in the integrand, then integrate by parts, obtaining the formula

$$L'(t) =< \tilde{\gamma}_s, \tilde{\gamma}_s >^{\frac{1}{2}} < \tilde{\gamma}_t, \tilde{\gamma}_s > |_0^{L_0} \tag{8.8}$$
$$+ \int_0^{L_0} < \tilde{\gamma}_s, \tilde{\gamma}_s >^{-\frac{3}{2}} < \frac{D\tilde{\gamma}_s}{Ds}, \tilde{\gamma}_s >< \tilde{\gamma}_t, \tilde{\gamma}_s > \; ds$$
$$- \int_0^{L_0} < \tilde{\gamma}_s, \tilde{\gamma}_s >^{-\frac{1}{2}} < \tilde{\gamma}_t, \frac{D\tilde{\gamma}_s}{Ds} > \; ds$$

Next we take $L''(t)$ and observe that most terms will vanish when setting $t = 0$: any term of the derivative of the first line would contain a factor of $< V, \gamma' >= 0$ by assumption, or of $< \frac{DV}{Ds}, \gamma' >= 0$ by differentiating $< V, \gamma' >= 0$. Similarly every term of the derivative of the second line has a factor either as just discussed, or one containing $\frac{D\gamma'}{Ds} = 0$ again by assumption. The same applies to two of the three terms of the third line. In summary, setting $t = 0$ we get

$$L''(0) = -\int_0^{L_0} < \tilde{\gamma}_t(s,0), \frac{D}{Dt}\frac{D\tilde{\gamma}_s}{Ds}(s,0) > \; ds. \tag{8.9}$$

Next, in order to interpret this formula, we would like to change the order of the derivatives in order to integrate by parts, just as we did before to interpret $L'(0)$. The question is: if $V(s,t)$ is a vector field along $\tilde{\gamma}(s,t)$, how are $\frac{D}{Dt}\frac{DV}{DS}$ and $\frac{D}{Ds}\frac{DV}{Dt}$ related? It turns out they are not equal, let's see what happens.

Recall that $\frac{DV}{Ds} = V_s^T = V_s - (V \cdot \mathbf{N})\mathbf{N}$, therefore

$$\begin{aligned}\frac{D}{Dt}\frac{DV}{Ds} &= (V_s - ((V_{st} \cdot \mathbf{N}) + (V_s \cdot \mathbf{N}_t))\mathbf{N} - (V_s \cdot \mathbf{N})\mathbf{N}_t)^T \\ &= (V_s - ((V_{st} \cdot \mathbf{N}) + (V_s \cdot \mathbf{N}_t))\mathbf{N})^T - (V_s \cdot \mathbf{N})\mathbf{N}_t,\end{aligned}$$

because $\mathbf{N}_t$ is tangential: differentiate $\mathbf{N} \cdot \mathbf{N} = 1$ to get $\mathbf{N}_t \cdot \mathbf{N} = 0$. Next, observe that the terms in parenthesis are symmetric in s, t, therefore so is its tangential component, therefore we get

$$\frac{D}{Dt}\frac{DV}{Ds} - \frac{D}{Ds}\frac{DV}{Dt} = -(V_s \cdot \mathbf{N})\mathbf{N}_t + (V_t \cdot \mathbf{N})\mathbf{N}_s = (V \cdot \mathbf{N}_s)\mathbf{N}_t - (V \cdot \mathbf{N}_t)\mathbf{N}_s,$$

the last expression resulting from the identities $V_s \cdot \mathbf{N} + V \cdot \mathbf{N}_s = (V \cdot \mathbf{N})_s = 0$ and the similar identity in t.

Next, we need to recognize this last expression as $(\mathbf{N}_s \times \mathbf{N}_t) \times V$. Since $\mathbf{N}_s \times \mathbf{N}_t = ((\mathbf{N}_s \times \mathbf{N}_t) \cdot \mathbf{N})\mathbf{N}$, this looks somewhat similar to the numerator of (8.3), except that $\tilde{\gamma}$ is not necessarily a parametrization of S, it can happen that $\tilde{\gamma}_s \times \tilde{\gamma}_t$ vanishes, and, when it doesn't vanish, it may be a positive or negative multiple of $\mathbf{N}$. The correct interpretation of equations (8.2) and (8.3) in this situation, to avoid dividing by 0 and always having the correct sign, is

$$\mathbf{N}_s \times \mathbf{N}_t = K \; (\tilde{\gamma}_s \times \tilde{\gamma}_t), \tag{8.10}$$

which gives us the formula

$$\frac{D}{Dt}\frac{DV}{Ds} - \frac{D}{Ds}\frac{DV}{Dt} = K\ (\tilde{\gamma}_s \times \tilde{\gamma}_t) \times V. \tag{8.11}$$

Let us apply this formula to (8.9) by choosing $V(s,t) = \tilde{\gamma}_s(s,t)$. Observe that $(\tilde{\gamma}_s \times \tilde{\gamma}_t) \times \gamma_s(s,0) = (\tilde{\gamma}_s \times \tilde{\gamma}_t) \times \gamma'(s) = f(s)\mathbf{n}(s)$ where F and $\mathbf{n}$ are as defined in (8.7). Therefore (8.11) becomes

$$\Big(\frac{D}{Dt}\frac{D\tilde{\gamma}_s}{Ds} - \frac{D}{Ds}\frac{D\tilde{\gamma}_s}{Dt}\Big)(s.0) = Kf\mathbf{n}. \tag{8.12}$$

Using this formula, we can change the expression $\frac{D}{Dt}\frac{D\tilde{\gamma}_s}{Ds}(s,0)$ in the integrand of (8.9), to $\frac{D}{Ds}\frac{D\tilde{\gamma}_s}{Dt}(s,0) + K(s)f(s)\mathbf{n}(s)$, and then change the first term to $\frac{D}{Ds}\frac{D\tilde{\gamma}_t}{Ds}(s,0) = f''(s)\mathbf{n}(s)$ by two uses of the identity (7.44) and the fact that $\frac{D\mathbf{n}}{Ds} = 0$ (which follows from the fact that its two components $< \frac{D\mathbf{n}}{Ds}, \mathbf{n} >=< \frac{D\mathbf{n}}{Ds}, \gamma' >= 0$, obtained by differentiating $< \mathbf{n}, \mathbf{n} >= 1$ and $< \mathbf{n}, \gamma' >= 0$ and using $\frac{D\gamma'}{Ds} = 0$). Then rewrite (8.9) as

$$L''(0) = -\int_0^{L_0} (f'' + Kf)f\ ds, \tag{8.13}$$

and, integrating by parts,

$$L''(0) = \int_0^{L_0} (f')^2\ ds\ -\ \int_0^{L_0} Kf^2\ ds. \tag{8.14}$$

From this we see that, since every quantity, other than K, appearing in these expressions is intrinsic, we get the rough form of Gauss's Theorema Egregium:

Corollary 8.1. *The Gaussian curvature K is intrinsic.*

Also, the following is immediate from (8.14):

Corollary 8.2. *If $K(p) \geq 0$ for all $p \in S$, then $L''(0) > 0$ for all non-trivial normal variations $\tilde{\gamma}$ that keep the endpoints fixed.*

Proof. The first term of (8.14) is ≥ 0 and $= 0$ if and only if f is constant. If $K \geq 0$ the second term is also ≥ 0, hence the sum $L''(0) \geq 0$ and $= 0$ if and only if f is constant. For a normal variation with fixed endpoints must have $f(0) = f(L_0) = 0$, hence $f \equiv 0$ if $L''(0) = 0$. □

8.2.1. *Jacobi's Equation and Theorema Egregium in Geodesic Polar Coordinates.* We now derive Gauss's formula (8.5) for K. We will not logically use the second variation formulas (8.13) and (8.14) but these formulas provide the motivation for the arguments that follow.

Suppose now that the variation $\tilde{\gamma}(s,t)$ of (8.6) is *always normal and always by geodesics*, that is, for *all* t, the curve $\tilde{\gamma}(\ ,t)$ is a geodesic, and for *all* s, t

we have $\tilde{\gamma}_t(s,t) \perp \tilde{\gamma}_s(s,t)$. Write $\mathbf{n}(s,t)$ for the unit normal of the geodesic $\tilde{\gamma}(\ ,t)$ at the point (s,t), chosen counterclockwise in the given orientation, and let $f(s,t)$ be defined by $\tilde{\gamma}_t(s,t) = \tilde{f}(s,t)\mathbf{n}(s,t)$, which makes sense since the variation is always normal.

Theorem 8.1. *Let $\tilde{\gamma}(s,t)$ be a variation of a geodesic $\gamma(s) = \tilde{\gamma}(s,0)$ that is always normal and always geodesic as just defined, and let $\tilde{\gamma}_t(s,t) = \tilde{f}(s,t)\mathbf{n}(s,t)$ as above. Then the function $f(s) = \tilde{f}(s,0)$ satisfies Jacobi's differential equation*

$$f'' + Kf = 0,$$

where $K = K(\gamma(s))$.

Proof. Since $\tilde{\gamma}(s,t)$ is a geodesic for all t we have that $\frac{D\tilde{\gamma}_s}{Ds}(s,t) = 0$, hence $\frac{D}{Dt}\frac{D\tilde{\gamma}_s}{Ds}(s,t) = 0$, hence, by (8.12) $\frac{D}{Ds}\frac{D\tilde{\gamma}_s}{Dt}(s,t) + K(\tilde{\gamma}(s,t))\tilde{f}(s,t)\mathbf{n}(s,t) = 0$. Changing $\frac{D\tilde{\gamma}_s}{Dt}$ to $\frac{D\tilde{\gamma}_t}{Ds}$ by (7.43) and then $\frac{D}{Ds}\frac{D\tilde{\gamma}_t}{Ds} = \frac{D}{Ds}\frac{D(\tilde{f}\mathbf{n})}{Ds} = \tilde{f}''\mathbf{n}$, this last identity as in the derivation of (8.13). Finally, setting $t = 0$, we obtain $(f'' + Kf)\mathbf{n} = 0$, hence the desired equation. □

The application we have in mind is the following natural example of an always normal, always geodesic variation:

Corollary 8.3. *Let $p \in S$, let U be a neighborhood of p on which a system of geodesic polar coordinates (r,θ) is defined, (as in Section 7.2.5 above or theorem 6.9 of* [12]*), thus $ds^2 = dr^2 + f(r,\theta)^2\, d\theta^2$ (as in (7.46)). Then, for each θ, the function $f(r,\theta)$ is the unique solution of Jacobi's Equation*

$$f_{rr}(r,\theta) + K(r,\theta)f(r,\theta) = 0$$

satisfying the initial conditions $f(0,\theta) = 0$ and $f_r(0,\theta) = 1$.

Proof. For each θ the curve (r,θ) is a geodesic (parametrized by arclength), so this is an always geodesic variation of any of its members. By Gauss's Lemma (Theorem 6.8 of [12]) it is an always normal variation and the variation vector field is $f(r,\theta)\mathbf{n}(r,\theta)$. Therefore, for any fixed θ, $f(r,\theta)$ satisfies Jacobi's Equation.

The expansion (7.47): $f(r,\theta) = r - \frac{K(0)}{6}r^3 + O(r^4) = r + O(r^3)$ shows that the stated initial conditions are satisfied. □

Corollary 8.4. *Gauss's Theorema Egregium in geodesic polar coordinates:*

$$K(r,\theta) = -\frac{f_{rr}}{f}(r,\theta),$$

This identity also holds at the origin of the coordinate system, in the sense that $\lim_{r\to 0}(-\frac{f_{rr}}{f}) = K(0)$.

Proof. The formula away from the origin is clear. At the origin, even though the coordinates become singular, the expansion (7.47): $f(r,\theta) = r - \frac{K(0)}{6} r^3 + O(r^4) = r + O(r^3)$ gives

$$\frac{f_{rr}}{f} = \frac{-K(0)r + O(r^2)}{r + O(r^3)} = -K(0) + O(r) \to -K(0) \text{ as } \ r \to 0.$$

□

8.2.2. *Definition of Gaussian Curvature for any Riemannian Metric.* Having proved that K is intrinsic for surfaces in $\mathbb{R}^3$, it is reasonable to say that K can be defined for any Riemannian metric. It is enough to define it for metrics (U, g) on open subsets of $\mathbb{R}^2$ in a manner invariant under isometries. The favorite definition at present is to start from formula (8.11). Take as ususal coordinates u_1, u_2 on U and consider a vector field $V(u_1, u_2)$ on U. Then one has to prove that

$$\frac{D}{Du_2}\frac{DV}{Du_1} - \frac{D}{Du_1}\frac{DV}{Du_2} = K \sqrt{\det g}\, V^\perp \tag{8.15}$$

for some smooth function K, where $V^\perp$ is the tangent vector to U at u that is perpendicular to V and obtained by counterclockwise rotation. (The problem is to show that the right hand side is of the form $f\ V^\perp$ for some smooth function f, then we can divide this function by the positive function $\sqrt{\det g}$ and call the result K). This is the fact and can be derived from the equations (7.42), (7.43) and (7.44).) Once this fact is verified, we can *define K by (8.15) and all intrinsic formulas involving K that we have derived so far are valid in any Riemannian metric.*

We will adopt this definition, and observe that the formula of Corollary 8.4 remains valid, in paarticular, we also have $K(0) = \lim_{r\to 0}(-f_{rr}/f)$ that we used in [12].

8.2.3. *Metrics of Constant Curvature.* We can view Corollary 8.3 as a method of constructing a Riemannian metric with prescribed curvature. This is very practical in case that K is a constant. It is convenient to normalize to three values of this constant: $K \equiv 0,\ 1,\ -1$. We obtain the following:

(1) $K \equiv 0$: We need to solve $f_{rr} = 0$ with initial conditions $f(0) = 0$ and $f_r(0) = 1$. The solution is $f(r) = r$ and the expression for the metric is

$$dr^2 + r^2\ d\theta^2, \tag{8.16}$$

which is valid in all of $\mathbb{R}^2$ and is the familiar expression for the Euclidean metric $dx^2 + dy^2$ in polar coordinates.

(2) $K \equiv 1$: We need to solve $f_{rr} + f = 0$ with initial conditions $f(0) = 0$ and $f_r(0) = 1$. The solution is $f(r) = \sin r$ and the expression for

the metric is

$$dr^2 + \sin^2 r \; d\theta^2 \tag{8.17}$$

which gives a positive definite metric only on the disk $\{r < \pi\}$ and is the familiar expression for the metric of the unit sphere S^2 in spherical coordinates, as in Section 7.1 above, or, in more detail, in Example (6.6) of [12] (see formula (6.7)), where r is the angle from the north pole, which is the same as arclength on the great circles through the north pole.

This metric is isometric to the spherical metric g_S of (7.12). In fact, writing g_S in polar coordinates ρ, θ, where $u_1 = \rho\cos\theta$ and $u_2 = \rho\sin\theta$, then g_S becomes

$$g_S = \frac{4(d\rho^2 + \rho^2 \; d\theta^2)}{(1+\rho^2)^2},$$

and the maps $\phi(\rho, \theta) = (2\arctan\rho, \theta)$ and $\psi(r, \theta) = (\tan\frac{r}{2}, \theta)$ are isometries and inverses to each other. (These computations are equivalent to the computations in homework problems of radii and circumferences in g_S of circles centered at 0.)

(3) $K \equiv -1$: We need to solve $f_{rr} - f = 0$ with initial conditions $f(0) = 0$ and $f_r(0) = 1$. The solution is $f(r) = \sinh r$ and the expression for the metric is

$$dr^2 + \sinh^2 r \; d\theta^2, \tag{8.18}$$

which is valid in all of $\mathbb{R}^2$ and is isometric to the Poincaré metric g_P of Example 7.4. To see an explicit isometry, introduce polar coordinates ρ, θ in the unit disk: $u_1 = \rho\cos\theta$ and $u_2 = \rho\sin\theta$ as above, then the formula (7.22) becomes

$$g_P = \frac{4(d\rho^2 + \rho^2 \; d\theta^2)}{(1-\rho^2)^2},$$

and we get inverse isometries $\phi(\rho, \theta) = (\ln\frac{1+\rho}{1-\rho}, \theta) = (2\tanh^{-1}(\rho), \theta)$ and $\psi(r, \theta) = (\tanh\frac{r}{2}, \theta)$, whose verification is equivalent to homework problems.

Remark 8.1. For arbitrary positive constants K, (8.17) would change to

$$dr^2 + \Big(\frac{1}{\sqrt{K}} \; \sin(\sqrt{K}r)\Big)^2 \; d\theta^2,$$

and, for a negative constant K,(8.18) would change to

$$dr^2 + \Big(\frac{1}{\sqrt{-K}} \; \sinh(\sqrt{-K}r)\Big)^2 \; d\theta^2,$$

which, in either case, are asymptotic to $dr^2 + r^2 \; d\theta^2$ as $|K| \to \infty$.

8.3. **Total Curvature or Geodesic Triangles.** We now prove Gauss's formula (8.4). We assume given a smooth surface S with a Riemannian metric, but restrict ourselves to a geodesic triangle Δ lying in a small enough open set U on which a system of geodesic polar coordinates r, θ, centered at one of the vertices, is defined:

Theorem 8.2. *Let S be a smooth surface with a Riemannian metric, let Δ be a geodesic triangle with vertices P, Q, R, and assume that Δ is contained in the domain of a geodesic polar coordinate system centered at P. Let α, β, γ be the interior angles of Δ at the points P, Q, R respectively. Then*

$$\iint_\Delta K \; dA = \alpha + \beta + \gamma - \pi. \tag{8.19}$$

Proof. Using the definition (**??**) of dA, and observing that $\sqrt{\det g} = f(r, \theta)$, we get

$$\iint_\Delta K \; dA = \int_{\theta_1}^{\theta_2} \int_0^{r(\theta)} K(r, \theta) f(r, \theta) \; dr d\theta,$$

where $\theta = \theta_1$ and $\theta = \theta_2$ are the equations of the geodesic rays containing the sides $\overline{PQ}$ and $\overline{PR}$ of Δ, and $(r(\theta), \theta)$, $\theta_1 \le \theta \le \theta_2$, is a parametrization of the third side $\overline{QR}$ of Δ, see Figure **??**.

Using Corollary 8.4 we get

$$\int_{\theta_1}^{\theta_2} \int_0^{r(\theta)} (-f_{rr}) \; dr d\theta = \int_{\theta_1}^{\theta_2} (-f_r(r(\theta), \theta) + f_r(0, \theta)) \; d\theta.$$

Since Corollary 8.3 gives $f_r(0, \theta) = 1$, this becomes

$$\int_{\theta_1}^{\theta_2} (-f_r(r(\theta), \theta) + 1) \; d\theta = \alpha - \int_{\theta_1}^{\theta_2} f_r(r(\theta), \theta) \; d\theta.$$

To evaluate this last integral we need the following lemma:

Lemma 8.1. *Let $(r(s), \theta(s))$ be a geodesic, parametrized by arclength, in the domain of the coordinate system used above, and suppose it does not pass through the center (θ not constant). Let $\phi(s)$ be the angle at $(r(s), \theta(s))$ that its tangent vector $(r'(s), \theta'(s))$ makes with the forward tangent vector to the ray $\theta = \theta(s)$ from the origin, see Figure* **??**. *Then*

$$\frac{d\phi}{d\theta} = -f_r(r(\theta), \theta).$$

Proof. Since the tangent vector to the geodesic is (r', θ'), the tangent vector to the ray from the origin is $(1, 0)$ and the length squared $r'^2 + f^2\theta'^2 = 1$, we get

$$\cos\phi = r' \quad \text{and} \quad \sin\phi = f\theta',$$

the second resulting from $\sin\phi$ being the cosine of the angle between (r', θ') and $(0, 1)$, and these vectors have inner product $f^2\theta'$ and magnitudes 1, f respectively.

Differentiating the first of these equations, then using the first differential equation $r'' - ff_r(\theta')^2 = 0$ of (7.48) and the second of the above two equations, we get

$$-\sin\phi\ \phi' = r'' = ff_r\theta'^2 = -f\theta'\phi',$$

and dividing the last equality by $f\theta'$ (which is valid by the assumption that the geodesic is not through the pole, therefore neither f nor θ' vanishes), we get

$$\frac{d\phi}{d\theta} = \frac{\phi'}{\theta'} = -f_r.$$

□

We can now resume the computation of the integral:

$$\begin{aligned}\alpha - \int_{\theta_1}^{\theta_2} f_r(r(\theta),\theta)\ d\theta &= \alpha + \phi(\theta_2) - \phi(\theta_1)\\ &= \alpha + \gamma - (\pi - \beta) = \alpha + \beta + \gamma - \pi,\end{aligned}$$

see Figure 8.2. □

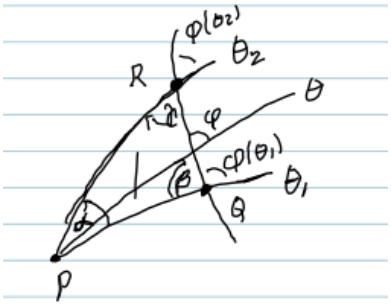


FIGURE 8.2. Angles of the Geodesic Triangle $\Delta = PQR$.

8.4. **The Gauss - Bonnet Theorem.** Theorem 8.2 was stated and proved for small geodesic triangles. But it holds for triangles of any size by subdividing into small triangles to which Theorem 8.2 applies, and adding all the contributions. The result is again the formula (8.4). A similar formula holds for any geodesic n-gon P with interior angles $\alpha_1, \ldots, \alpha_n$:

$$\iint_P K\ dA = \alpha_1 + \cdots + \alpha_n - (n-2)\pi, \tag{8.20}$$

where the term $(n-2)\pi$ in the right hand side is the sum of interior angles of a Euclidean n-gon. This formula also follows from Theorem 8.2 by subdividing P into sufficiently small triangles and adding, using a variation of the argument in the proof of Theorem 8.3 below. There are also formulas for arbitrary regions, not necessarily with geodesic sides, involving extra terms of integrals of the geodesic curvatures of the edges.

We will not pursue these directions, but will prove another theorem, promised in (1.2), and that follows from Theorem 8.2 by the method of subdivision. This theorem relates topology and curvature of the compact surfaces. For simplicity, we will state it only for orientable ones, but it also holds for non-orientable surfaces.

Theorem 8.3. *Let S be a compact oriented smooth surface with a Riemannian metric, and let $K : S \to \mathbb{R}$ be the Gaussian curvature of this metric. Then*

$$\iint_S K \, dA = 2\pi\chi(S) = 4\pi(1-g), \tag{8.21}$$

where $\chi(S)$ is the Euler characteristic (see Definition 3.4) and g is the genus of S (as in Theorem 1.1).

Remark 8.2. Strictly speaking, we should have written χ(triangulation of S) in the right - hand side, rather than $\chi(S)$. With enough care in the proof (which we only hint at) this theorem could be used to prove that χ is independent of the triangulation (since the left-hand side is).

Proof. We will prove this theorem *under the assumption that S has a triangulation $S = T_1 \cup \cdots \cup T_n$ (as in Definition 3.1), where all the T_i are geodesic triangles* (meaning that the edges are geodesic). Also, assume that *each T_i is contained in the domain of a geodesic polar coordinate system centered at one vertex.* This can actually be proved from the existence of triangulations, with subdivision and approximation arguments, but we will not do this here. We will just say that it is best to do this using Definition 3.3, in terms of a simplicial complex C with geometric realization $|C|$ (see Definition 3.2) and a homeomorphism $\phi : |C| \to S$. There is a process of subdviding a complex C to get a new complex C', called its *barycentric subdvision* that makes the simplices smaller. Using this process and approximations, one can first assume that the map ϕ is smooth on each simplex, then subdivide enough times to make the simplices small enough so they lie in the domain of polar coordinate systems, and replace the simplices by geodesic one.

So write $S = T_1 \cup \cdots \cup T_n$ where each T_i is a geodesic triangle with interior angles $\alpha_i, \beta_i, \gamma_i$, and each T_i is small enough so that the proof of Theorem 8.2 is valid. Then

$$\iint_S K \, dA \; = \; \sum_{i=1}^{n} \iint_{T_i} K \, dA = \sum_{i=1}^{n} (\alpha_i + \beta_i + \gamma_i - \pi).$$

Writing, as in Chapter 3, V, E, F for the number of vertices, edges, triangles (faces) in the triangulation, this is

$$\sum_{i=1}^{F} (\alpha_i + \beta_i + \gamma_i) - F\pi$$

since $n = F$. The first term can be reorganized as a sum over the vertices $p_1, \ldots, p_V$ and gives

$$\sum_{j=1}^{V} (\text{sum of the angles at } \ p_j \text{ of the triangles with vertex } p_j) = 2\pi V$$

because, for each j, these angles cover a neighborhood of p_j, so their sum is the same as the sum of the angles between their tangent vectors in the Euclidean space of tangent vectors at p_j, so this sum is 2π. Therefore

$$\iint_S K \, dA = 2\pi V - \pi F = 2\pi(V - \frac{3}{2}F + F) = 2\pi(V - E + F) = 2\pi\chi(S),$$

using the identity $2E = 3F$ of Corollary 3.1. Finally, $\chi(S) = 2 - 2g$ is Equation (3.4).

□

9. The Geometries of Constant Curvature

We look more closely at the Riemannian metrics of constant curvature. We gave models for them in Section 8.2.3 by giving their geodesic polar coordinate expressions. In all these formulas there is a distinguished point, the pole, and the metric is rotationally symmetric around that point, and also has reflectional symmetry about all the geodesics passing through the pole. But we know, both in the case $K \equiv 0$ of Euclidean geometry and $K \equiv 1$ of spherical geometry, that there are more symmetries that are not visible in the formulas of Section **??**. It is reasonable to expect the same to be true of the case $K \equiv -1$ of the Poincaré metric. This metric will also be called the *hyperbolic metric* and its geometry *hyperbolic geometry*.

To discuss the symmetries more precisely, we have to consider more than just the value of the curvature, since there can be several Riemannian metrics with the same curvature. For example, a proper open subset of the Euclidean plane $\mathbb{R}^2$ has $K \equiv 0$ but is not isometric to all of $\mathbb{R}^2$ and it will not be invariant by all translations of $\mathbb{R}^2$. One way of excluding these examples is to talk about *complete* metrics, meaning either *complete as a metric space* or, what turns out to be equivalent for Riemannian metrics, is the condition of *geodesic completeness*, meaning that *all geodesics are defined for all time*. This excludes proper open subsets of $\mathbb{R}^2$, and similar examples.

Another situation of different spaces with the same curvature occurs, for example, the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ and all of $\mathbb{R}^2$ both have $K \equiv 0$, the sphere S^2 and the projective plane P^2 both have metrics with $K \equiv 1$. What distinguishes $\mathbb{R}^2$ and S^2 in this situation is that they are *simply connected*.

The examples of Section 8.2.3 are *complete, simply connected surfaces of constant curvature*. It can be proved that, up to isometry, they are the only ones. In the case of $K \equiv 1$, we have to interpret the formula (8.17) as

defining a metric on S^2 by presenting S^2 as the quotient space of the disk $\{r \leq \pi\} \subset \mathbb{R}^2$ by identifying its boundary to a point.

9.1. **Groups of Isometries.** We know from Section 2.1 of [12] that the group of isometries of $\mathbb{R}^2$ is the Euclidean group $E(2)$ consisting of all affine linear transformations $x \to Ax + b$ where $A \in O(2)$ is an orthogonal 2×2 matrix and $b \in \mathbb{R}^2$. Recall the homeomorphism (Remark 2.6 of [12]) of $E(2)$ with the topological space $\mathbb{R}^2 \times O(2)$, which has two connected components, each homeomorphic to $\mathbb{R}^2 \times S^1$. Thus the group $E(2)$ is three-dimensional. This is the largest possible dimension of the group of isometries of a Riemannian metric on a surface. We do not want to make this precise, but intuitively it can be seen as follows: If G is the group of isometries of S, then for any $p \in S$, the subgroup G_p of isometries fixing the point p is at most one-dimensional, because the differential at p gives a homomorphism of this group to the group, isomorphic ot $O(2)$, of linear isometries of the tangent space T_pS. By Corollary 7.2 this homomorphism is injective, hence the dimension of G_p is at most one. Since S is two dimensional, this leaves at most two more possible dimensions for G, hence at most three dimensions in total.

Similarly, the group of isometries of S^2 is the group $O(3)$ of orthogonal 3×3 matrices, see Definition 2.4 of [12]. It is clear that any matrix $A \in O(3)$ gives an isometry of R^2 fixing the origin, hence leaves S^2 invariant and if gives an isometry of S^2. This can be seen as follows. For any $x \in S^2$, the tangent space $T_xS^2 = x^\perp = \{y \in \mathbb{R}^3 : x \cdot y = 0\}$ the orthogonal complement of x, and orthogonal matrices preserve orthogonality. Hence $A|_{x^\perp} : x^\perp = T_xS^2 \to (Ax)^\perp = T_{Ax}S^2$ is an isometry.

Theorem 9.1. *Let G be the group of isometries of S^2. The restriction map $\rho : O(3) \to G$ defined by $\rho(A) = A|_{S^2}$ is a group isomorphism.*

Proof. We have just remarked that for all $A \in O(3)$, $\rho(A)$ is an isometry of S^2, so $\rho : O(3) \to G$, and it is clear that ρ is homomorphism. If $\rho(A) = id$ then, since S^2 contains bases for $\mathbb{R}^3$, $A = I$, thus ρ is injective. So it is harmless to simplify notation and not distinguish A and $\rho(A)$. To prove that ρ is surjective, fix a point $x_0 \in S^2$. Given any $g \in G$, that is, given any isometry $g : S^2 \to S^2$, there exists $A \in O(3)$ so that $Ax_0 = g(x_0)$. Then $f = A^{-1}g$ is an isometry of S^2 fixing x_0. Then $d_{x_0}f$ is an isometry of $T_{x_0}S^2$, and (using an orthonormal basis for $\mathbb{R}^3$ with x_0 as one of its members, the other two then being an orthonormal basis for $x_o^\perp = T_{x_0}S^2$), we can find $B \in O(3)$ with $Bx_0 = x_0$ and $B|_{x_0^\perp} = d_{x_0}f$. Then $B^{-1}f$ is an isometry of S^2 that fixed x_0 and is the identity on $T_{x_0}S^2$, hence, by Corollary 7.2 we have that $B^{-1}f = -id$, hence $f = B$, hence $A^{-1}g = B$ or $g = AB \in O(3)$ and ρ is surjective. □

Remark 9.1. It can be proved that $O(3)$ has two connected components, distinguished by the value ± 1 of the determinant. Composition with a

fixed reflection gives a bijection between the subgroup $SO(3)$ of rotations (orthogonal matrices with det $= 1$) and the coset of orthogonal matrices with det $= -1$. It is not hard to work out the topology of $SO(3)$: it is homeomorphic to the *projective space* P^3, defined to be $P^3 = S^3/(x \sim -x) = B^3/(x \sim -x \text{if} x \in \partial B^3)$ (in analogy with the definition and alternative description of P^2 in Example 1.1) This can be seen as follows: a rotation is determined by its its axis, its angle of rotation, and the sense of rotation. This can be encoded by a vector $\mathbf{v} \in \mathbb{R}^3$ chosen to point along the axis and related to the direction of rotation by the right-hand rule. The magnitude $|\mathbf{v}|$ is chosen to be the angle α of rotation, and can choose $0 \leq \alpha \leq \pi$ by choosing the direction of $\mathbf{v}$ appropriately: α and $\mathbf{v}$ give the same rotation as $2\pi - \alpha$ and $-\mathbf{v}$. In this way we get a continuous, surjective map $p : \overline{B(0,\pi)} \to SO(3)$, that is injective on the open ball $B(0,\pi)$ and identifies $\mathbf{v}$ and $-\mathbf{v}$ if $|\mathbf{v}| = 1$. This gives a homeomorphism between $\overline{B(0,\pi)}/(\mathbf{v} \sim -\mathbf{v}$ if $|\mathbf{v}| = \pi)$ and $SO(3)$. In particular, we see that $SO(3)$ is connected and that $O(3)$ has two connected components.

To study the group of isometries of the Poincaré metric, it is convenient to have other models of hyperbolic geometry.

9.2. **Möbius Transformations and Hyperbolic Geometry.** Let $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$ and let f be the transformation

$$f(z) = \frac{az + b}{cz + d} \tag{9.1}$$

it is defiined for $cz + d \neq 0$, that is, $z \neq -\frac{d}{c}$, but it is harmless to say that $f(-\frac{d}{c}) = \infty$ and that $f : \mathbb{C}^+ = \mathbb{C} \cup \{\infty\} \to \mathbb{C}^+$ and that $f(\infty) = \frac{a}{c}$. We can identify $\mathbb{C}^+$ with S^2 by stereographic projection, view these as transformations of S^2. These are *conformal* transformations of the spherical metric and of the usual Euclidean metric of $\mathbb{C}$, see **??**. The basic fact here is that f is differentiable in the complex sense, that is, it is a complex analytic function, and such functions are conformal. These are the basic facts we need about Möbius transformations, see, for example, Section 3 of Chapter 3 of [1] for more details.

9.2.1. *Basic Properties of Möbius Transformations.* Let G_1 denote the group $GL(2, \mathbb{C})$ of invertible 2×2 complex matrices

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a, b, c, d \in \mathbb{C}, \quad \det(A) = ad - bc \neq 0. \tag{9.2}$$

Let f_A be the transformation of $\mathbb{C}^+$ given by formula (9.1). Then

(1) For all $A \in G_1$, f_A is a conformal transformation of $\mathbb{C}^+$ (it preserves angles between curves). This is a consequence of the fact that it is a complex analytic function, that is, it is differentiable in the complex sense.

(2) The derivative of f_A is

$$f'_A(z) = \frac{ad - bc}{(cz + d)^2}. \tag{9.3}$$

(3) Let G_2 denote the group (under composition) of conformal transformations of $\mathbb{C}^+$. Then, the map $G_1 \to G_2$ given by the assignment $A \to f_A$ is a group homomorphism : $f_{AB} = f_A \circ f_B$. Its kernel is $\{\lambda I : \lambda \in \mathbb{C}, \lambda \neq 0\}$, the group of non-zero multiples of the unit matrix. In particular, for all $A \in G_1$, f_A is invertible and $f_{A^{-1}} = (f_A)^{-1}$.

(4) Let $C \subset \mathbb{C}^+$ be either a circle or a straight line (= circle through ∞). Then $f_A(C)$ is either a circle or a straight line. Briefly: f_A takes circles to circles. **Warning**: *But they need not take centers of circles to centers.* We will see many examples later.

(5) G_2 is the group of all transformations of $\mathbb{C}^+$ that take circles to circles and preserve orientation.

9.2.2. *Two Models of Hyperbolic Geometry.* We have introduced the Poincaré metric g_P. Let us call it the *disk model* of hyperbolic geometry. We could write $D = \{w \in \mathbb{C} : |w| < 1\}$ where $w = u + iv$ and $dw = du + idv$. Then, in complex notation,

$$g_P = \frac{4|dw|^2}{(1 - |w|^2)^2} \ = \frac{4(du^2 + dv^2)}{(1 - (u^2 + v^2))^2} \tag{9.4}$$

We want the *upper half-plane* model, defined as follows:

Definition 9.1. The *upper half plane model* of hyperbolic geometry is the open set $H = \{z = x + iy \in \mathbb{C} : y > 0\}$, with Riemannian metric

$$g_H = \frac{|dz|^2}{y^2} = \frac{dx^2 + dy^2}{y^2}. \tag{9.5}$$

To see that (D, g_P) and (H, g_H) are isometric, we need to find a map. The theory of Möbius transformations easily provides such maps. For example, let $h^+_{\mathbb{C}} \to \mathbb{C}^+$ be defined by

$$h(z) = \frac{z - i}{z + i}. \tag{9.6}$$

It is clear that $h(i) = 0$ and that, if $z = x \in \mathbb{R}$, $|h(x)| = |(x - i0/(x + ii))| = 1$, so $h(\mathbb{R}) \subset S^1$, thus $h(\mathbb{R}) = S^1$ (by invertibility and the circle-preserving property), therefore the connected component H of $\mathbb{C}^+ \setminus \{\mathbb{R}\}$ containing i must be mapped to the component D of $\mathbb{C}^+ \setminus \{S^1\}$ containing $0 = h(i)$, and diffeomeorphically because of the invertibility of h. It remains to check:

Theorem 9.2. *The transformation $h : H \to D$ defined in (9.6) is an isometry: $h^*(g_P) = g_H$.*

Proof. Let $w = h(z)$. From Equation (9.3) $dw = h'(z)\ dz = 2i(z+1)^{-2}\ dz$, so

$$h^* g_P = \frac{4|h'(z)|^2|dz|^2}{(1-|h(z)|^2)^2} = \frac{16|z+i|^{-4}|dz|^2}{(\frac{(|z+i|^2-|z-i|^2)}{|z+i|^2})^2} = \frac{16|dz|^2}{(4y)^2} = g_H,$$

the equality next to last being $|z+i|^2 - |z-i|^2 = (x^2+(y+1)^2) - (x^2 + (y-1)^2) = 4y$. □

The value of this identification is that H has a visible set of isometries:

Theorem 9.3. *Let $a, b \in \mathbb{R}$ and $a > 0$. Then the transformation $f_{a,b} : H \to H$ defined by $f_{a,b}(z) = az + b$ is an isometry of (H, g_H). In particular, H is* homogeneous*: given any two points $z_1, z_2 \in H$ there exists an isometry $f : H \to H$ with $f(z_1) = z_2$.*

Proof. Let $z = x+iy$ with $y > 0$. then $f_{a,b}(z) = (ax+b)+i(ay)$ has positive imaginary part since $a > 0$, so $f_{a,b} : H \to H$. To show it is an isometry, compute:

$$f_{a,b}^*(g_H) = \frac{d(ax+b)^2 + d(ay)^2}{(ay)^2} = \frac{a^2(dx^2+dy^2)}{a^2y^2} = \frac{dx^2+dy^2}{y^2} = g_H.$$

Given any point $z_0 = x_0 + iy_0 \in H$, $f_{y_0,x_0}(i) = y_0 i + x_0 = z_0$, thus there is a transformation taking i to any $z_0 \in H$. To take z_1 to z_2, take z_1 to i and i to z_2, that is, use $f = f_{y_2,x_2} \circ f_{y_1,x_1}^{-1}$. □

With this information we can find all the geodesics in hyperbolic geometry, both in the H-model and the D-model.

Theorem 9.4.
(1) *The geodesics in H are the vertical lines and the semi-circles with center on the real axis $\mathbb{R}$.*
(2) *The geodesics in D are the straight lines through the origin and the circle arcs perpendicular to the boundary circle S^1.*
(3) *Given any two points $P, Q \in H$, there is a unique geodesic segment from P to Q. The length of this segment is $d(P,Q)$, the hyperbolic distance between P and Q. (Therefore the same statement holds in D).*

Proof. We know that the geodesics through the origin in D are the straight line segments through the origin. Since these meet the boundary at right angles, the isometry $h^{-1} : D \to H$ takes these segments to circles through i perpendicular to the real axis. These are exactly the semicircles with center on the real axis, including the imaginary axis (which is $h^{-1}(D \cap \mathbb{R})$). Since the transformations $f_{a,b}$ take i to any other point $b + ai \in H$, and they take the geodesics through i to those through $ai + b$, we see that we get all the vertical lines and all the semicircles centered in the real axis in this way. Since we have found geodesics through every point and every direction, we have found all geodesics.

This proves part (1). Part (3) is then checked by elementary geometry: if P and Q are on the same vertical line, then the vertical segment between them is the unique geodesic connecting them, which therefore must realize the distance. Otherwise, there is a unique semi-circle centered on the real axis connecting P and Q, thus we reach the same conclusion.

Finally, the images by h of all the geodesics we have found in H gives us all geodesics in D. These must be circle arcs perpendicular to the boundary circle, since all the geodesics in H have the similar property. See Figure **??**. There are several beautiful renderings by Escher of the Poincaré disk model that show geodesics and some of the isometries of D, for example Figure 9.1. □

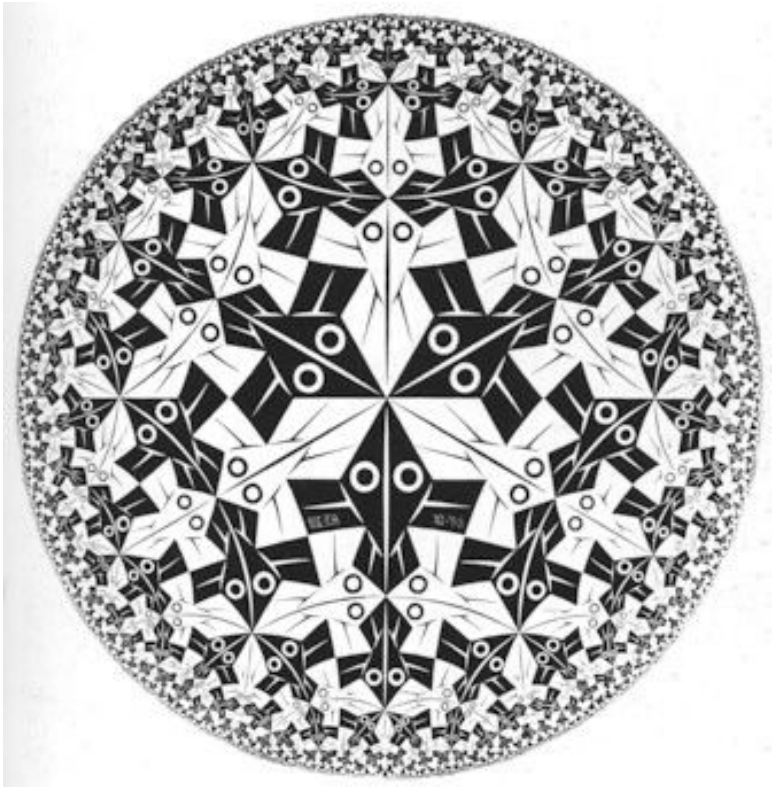


FIGURE 9.1. Escher's Circle Limit 1.

We have, in the Poincaré model D, the isometries f_θ defined by $f_\theta(z) = e^{i\theta}z$, rotations about the origin, and now in the upper half plane model H we have the $f_{a,b}$ just defined (fixing ∞). So we have transformations depending on the three parameters a, b, θ, and we expect a three-dimensional group of symmetries, so it is reasonable to expect that we have found all isometries (by composing the above). The only problem is that we have them in different models. To bring them to one model, in H we could use

the $f_{a,b}$ and the $h^{-1} \circ f_\theta \circ h$, or, in D, the f_θ and the $h \circ f_{a,b} \circ h^{-1}$, with h as in (9.6).

Theorem 9.5. *Let A be as in (9.2) with $a, b, c, d \in \mathbb{R}$ and $\det(A) = ad - bc > 0$. Then $f_A : H \to H$ is an isometry, and every orientation preserving isometry of (H, g_H) is obtained this way.*

Proof. Let A have real coefficients. Then $f_A(\mathbb{R}) \subset \mathbb{R}$, so $f_A(H)$ the connected component of $\mathbb{C}^+ \setminus \mathbb{R}$ containing i, which is either H or the lower half plane $L = \{y < 0\}$. Now

$$f_A(i) = \frac{ai + b}{ci + d} = \frac{(ai + b)(-ci + d)}{c^2 + d^2} = \frac{(ac + bd) + i(ad - bc)}{c^2 + d^2},$$

which has positive imaginary part if and only if $ad - bc > 0$. Thus $f_A(i) \in H$, so $f_A(H) = H$.

To check that f_A is an isometry of g_H is a computation along the lines of Theorem 9.2. An alternative proof is given in a homework problem.

Suppose $f_A(i) = i$. This means $a + bi = i(ci + d) = d - ci$, thus $a = d$ and $b = -c$, in other words

$$A = \begin{pmatrix} d & -c \\ c & d \end{pmatrix}, \tag{9.7}$$

in particular $\det A = c^2 + d^2$ and, using the formula (9.3), we have

$$f_A'(i) = \frac{c^2 + d^2}{(ci + d)^2} = e^{-2i\theta} \text{ where } \cos\theta = \frac{d}{c^2 + d^2},\ \sin\theta = \frac{c}{c^2 + d^2}.$$

This implies that we can get every rotation of T_iH, the space of tangent vectors to H at i by a suitable f_A, for instance, to get rotation by α we can take $\theta = -\alpha/2$. Therefore we get every rotation about the origin in T_iH from a suitable f_A fixing i. then, exactly the same argument as in the proof of the corresponding theorem for S^2, Theorem 9.1, we get that every orientation-preserving isometry of H is obtained in this way.

□

Remark 9.2.

(1) If we know all the orientation preserving isometries, then all isometries are obtained by composing these with one fixed orientation-reversing isometry, for instance, reflection in the y-axis: $x + iy \to x - iy$. Or, noting that both $z \to \bar{z}$ and f_A for $a, b, c, d \in \mathbb{R}$ *but* $\det A = ad - bc < 0$ interchange the upper and lower half-planes, we see that $z \to f_A(\bar{z}) = \overline{f_A(z)}$ preserves H and is orientation reversing, and we obtain them all this way. Thus all isometries of H are either of the form $z \to f_A(z)$ for real A with $\det A > 0$ or $z \to f_A(\bar{z})$ for real A with $\det A < 0$.

(2) Just as in the case of isometries of S^2 we get a map $H \times S^1$ to the group of orientation preserving isometries of H by sending (b, a) with $a > 0$ and $\theta \in [0, \pi]$ to the composition $f_{a,b} f_A$ where A is

as in (9.7), normalizing $c^2 + d^2 = 1$ and choosing θ as indicated there. Since $f_A = f_{-A}$, there is a choice of θ in $[0, \pi]$ that gives us f_A, and we get a map $\{(b, a) : a > 0\} \times ([0, \pi]/0 \sim \pi)$ to the group of orientation preserving isometries is a bijection (actually a homeomorphism). In particular, this group is homotopy equivalent to S^1, therefore connected. Thus the full group of isometries has two connected components, as in the case of $\mathbb{R}^2$ and S^2, and, as is the case for $\mathbb{R}^2$ (but not for S^2), each component is homeomorphic to the cartesian product of the space with S^1.

(3) There are two standard notations for the group of orientation preserving isometries of H. One is $PGL^+(2, \mathbb{R})$, meaning: take the group 2×2 real matrices with positive determinant, denoted $GL^+(2, \mathbb{R})$, and take the quotient by the normal subgroup of scalar matrices The letter "P" stands for "projective", meaning that you only look at how the group acts on lines through the origin (projective space), where the scalar matrices act trivially. The other is $PSL(2, \mathbb{R})$, meaning: take the group of 2×2 real matrices A with $\det A = 1$, usually denoted $SL(2, \mathbb{R})$, the *special linear group*, and take the quotient by the normal subgroup $\pm I$. This subgroup is the intersection with $SL(2, \mathbb{R})$ of the subgroup of scalar matrices in $GL^+(2, \mathbb{R})$.

One common feature of the three geometries of constant curvature is the following property, sometimes called *two point homogeneity.*

Theorem 9.6. *Let X be one of $\mathbb{R}^2$, S^2 or H with its metric of constant curvature. Given any four points $P, Q, P', Q' \in X$ such that $d(P, Q) = d(P', Q')$ there exists an orientation preserving isometry $f : X \to X$ such that $f(P) = P'$ and $f(Q) = Q'$. The isometry f is unique except in the case $X = S^2$ and P, Q are antipodal points.*

Proof. Let $D = d(P, Q)$ and let $\gamma : [0, D] \to X$ be the geodesic arc from P to Q realizing the distance. This arc is unique except in the case $X = S^2$ and P, Q are antipodal points. Let $\gamma_1 : [0, D] \to X$ be the geodesic arc from P' to Q' realizing the distance. Then, as we have seen in each case, there exists an orientation preserving isometry $f : X \to X$ so that $f(P) = P'$ and $d_P f(\gamma'(0)) = \gamma_1'(0)$. Since geodesics are determined by these initial conditions, must have $f \circ \gamma = \gamma_1$, which gives $f(\gamma(D)) = \gamma_1(D)$ or $f(Q) = Q'$. The proof of the uniqueness statement is left as an exercise.

□

9.3. **Existence of Metrics of Constant Curvature.** Now comes the final theorem:

Theorem 9.7. *Let S be a compact, connected, orientable surface. Then S has a Riemannian metric of constant Gaussian curvature.*

Proof. Using the classification theorem for surfaces, Theorem 1.1, we only need to consider $S^2(=\Sigma_0)$ and the surfaces Σ_g or $g = 1, 2, 3, \ldots$.

Case 1 :The sphere S^2: we know that it has a metric of constant curvature 1, so there is nothing to prove.

Case 2: The torus $T^2 = \Sigma_1$. We know from Example 7.7 that it has a Riemannian metric. This metric was constructed to be $dx^2 + dy^2$ on each coordinate chart , so it is locally isometric to the Euclidean plane $\mathbb{R}^2$. Since K is invariant under isometries and $K \equiv 0$ for the Euclidean plane $\mathbb{R}^2$, it follows that this metric on T^2 has $K \equiv 0$.

What also has to be checked is that the four pieces around the vertex fit into a Euclidean square. They do because the angles at each vertex of the square is a right angle. The same would work more generally using a parallelogram rather than a square, the sectors around the four corners will still fit isometrically into a Euclidean neighborhood of a point because the sum of the angles is 2π (for any Euclidean quadrilateral, in particular, any Euclidean parallelogram).

Case 3: The surfaces Σ_g of higher genus $g \geq 2$. We will concentrate in the case $g = 2$ that gives all the essential ideas.

First of all, we know, from the Gauss - Bonnet Theorem 8.3 that $K < 0$, let us normalize the situation to $K \equiv -1$. Recalling from (8) to (10) of Example 1.1 the presentation of the surface Σ_2 as a quotient of the octagon by identifications on the boundary, see Figure 1.1, and the definition of topological charts on the surface as in Figure 1.3 , it is not hard to see that the identifications on the charts centered at the interior points of edges,as in the first part of Figure 1.3 can be done using Euclidean geometry. But the identifications needed for the charts around a vertex, as in the second part of Figure 1.3 cannot be done in Euclidean geometry, since the sum of the angles of an octagon is 6π, so the pieces cannot metrically fit into a disk.

This is where hyperbolic geometry is needed. Gauss - Bonnet not only gives us a necessary condition, but the formula (8.20) tells us that for $K \equiv -1$ the sum of interior angles of a hyperbolic octagon is less than 6π. We need to find octagons whose interior angle sum is 2π, in other words, by (8.20), of area 4π, compare also with Theorem 8.3, which would also give area 4π to the quotient if $g = 2$.

The easiest way would be to use regular octagons. We want one with area 4π and interior angles all $\pi/4$. That one exists follows from a continuity argument: The maximum area is 6π, which is not attained by any finite octagon but is attained by the "asymptotic octagon" of Figure 9.2

There are also octagons of very small area, whose angles are therefore close to the Euclidean angles, as in Figure 9.3.

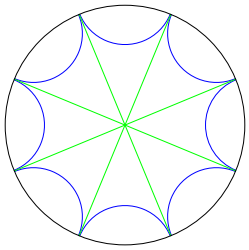


FIGURE 9.2. Regular Asymptotic Octagon, Area 6π

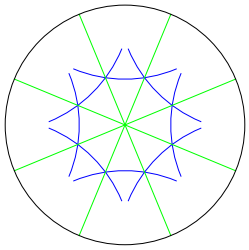


FIGURE 9.3. Regular Octagon of Small Area

By the intermediate value theorem there must be a regular octagon of area 4π, therefore interior angles $\pi/4$, as in Figure 9.4

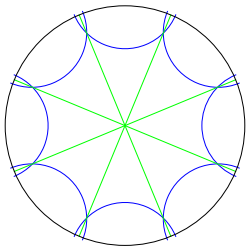


FIGURE 9.4. Regular Octagon of Area 4π

Choose this regular octagon of area 4π. We want to make the identifications as in Figure 1.1. Since all the sides have equal lengths, by Theorem 9.6 we can make the identifications by hyperbolic isometries. Let us call these isometries a_1, b_1, a_2, b_2, as in Figure **??**. ..... □

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, Mc Graw Hill, 1979.
[2] C. E. Burgess, Classification of Surfaces, American Math Monthly **92** (1985), 349–354.
[3] M. A. Do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice-Hall, 1976.
[4] K. F. Gauss, *General Investigations of Curved Surfaces*, Dover, 2005.

[5] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2002, available at `http://www.math.cornell.edu/ hatcher/AT/ATchapters.html`
[6] D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelsea Pub. Co.
[7] W. Klingenberg, *A Course in Differential Geometry*, Springer, 1978.
[8] S. Lang, *Algebra*, Third Edition, Springer, 2002.
[9] W. S. Massey, *Algebraic Topology: an Introduction*, Springer.
[10] B. Mendelson, *Introduction to Topology*, Dover Publications, 1990.
[11] R. S. Millman and G. D. Parker, *Elements of Differential Geometry*, Prentice Hall, 1977.
[12] Notes for Math 4510, Fall 2010, `http://www.math.utah.edu/ toledo/4510notes.pdf`
[13] B. O'Neill, *Elementary Differential Geometry*, Academic Press, 1966.
[14] J. Oprea, *Differential Geometry and its Applications*, Prentice-Hall 1997.
[15] A. Pressley, *Elementary Differential Geometry*, Springer, 2002.
[16] T. Shifrin, *Differential Geometry: A First Course in Curves and Surfaces*, notes available at `http://www.math.uga.edu/ shifrin/ShifrinDiffGeo.pdf`.
[17] J. Stillwell, *Geometry of Surfaces*, Universitext, Springer-Verlag, 1992.
[18] J. A. Thorpe, *Elementary Topics in Differential Geometry*, Springer, 1979.