LECTURES ON THE MAPPING CLASS GROUP OF A SURFACE

THOMAS KWOK-KEUNG AU, FENG LUO, AND TIAN YANG

Abstract. In these lectures, we give the proofs of two basic theorems on surface topology, namely, the work of Dehn and Lickorish on generating the mapping class group of a surface by Dehn-twists; and the work of Dehn and Nielsen on relating self-homeomorphisms of a surface and automorphisms of the fundamental group of the surface. Some of the basic materials on hyperbolic geometry and large scale geometry are introduced.

Contents

2010 *Mathematics Subject Classification.* Primary: 57N05; Secondary: 57M60, 57S05.
Key words and phrases. Mapping class group, Dehn-Lickorish, Dehn-Nielsen.

Introduction

The purpose of this paper is to give a quick introduction to the mapping class group of a surface. We will prove two main theorems in the theory, namely, the theorem of Dehn-Lickorish that the mapping class group is generated by Dehn twists and the theorem of Dehn-Nielsen that the mapping class group is equal to the outer-automorphism group of the fundamental group. We will present a proof of Dehn-Nielsen realization theorem following the argument of B. Farb and D. Margalit, [De, FM]. Along the way of the proof, we will introduce some of the basic notions and tools used in the surface theory.

This paper is the result of a series of lectures given by the second author in the Center of Mathematical Sciences, Zhejiang University, China during the summer of 2008. It aims at providing a concise understanding of the surface theory, especially suitable for graduate students. The prerequisites for these lectures have been kept to a minimum. Basic knowledge of algebraic topology and Riemannian geometry is all one needs to follow the lectures.

Let Σ be a compact oriented surface. A homeomorphism $\phi : \Sigma \to \Sigma$ is called a *Dehn twist* if its support A, the closure of $\{x \in \Sigma \mid \phi(x) \neq x\}$, is homeomorphic to an annulus under a homeomorphism $\mathcal{I} : \mathbb{S}^1 \times [0, 2] \to A$ and $\mathcal{I}^{-1}\phi\mathcal{I}$ is a 2π-twist on the annulus. A more precise definition will be given later in Definition 1.16.

The goal of these lectures is to prove the following two theorems.

Dehn-Lickorish Theorem. *Suppose $h : \Sigma \to \Sigma$ is an orientation preserving self-homeomorphism of a surface Σ so that $h|_{\partial\Sigma} = id$. Then there exists a finite set of Dehn twists $\phi_1, \ldots, \phi_n$ of Σ such that h is homotopic to the composition $\phi_1 \circ \cdots \circ \phi_n$.*

In fact, Dehn and Lickorish proved a stronger theorem that ϕ_i's can be chosen from a fixed finite set.

Dehn-Nielsen Theorem. *Let Σ be a closed surface of nonzero genus. Then the group of self-homeomorphisms on Σ modulo homotopy is isomorphic to the outer automorphism group of the fundamental group of Σ.*

The theorem may be stated in a simplier form in terms of the notion of mapping class groups, which we will discuss later.

In the first section, we will introduce the notions of mapping class groups and Dehn twists. Simple examples of mapping class groups and basic properties of Dehn twists will also be given. In the second section, how the mapping class group of a surface is generated by Dehn twists will be discussed. Indeed, we will prove the Dehn-Lickorish Theorem. Section three is a discussion of the hyperbolic geometry of surfaces. It consists of a quick introduction of the hyperbolic plane and the geometric interpretation of the fundamental group of a surface. In section four, quasi-isometries and large scale geometry are introduced. Our attention is on facts pertinent to our study of mapping class groups. In the last section, we will prove the Dehn-Nielsen Theorem.

The authors are very grateful to the Center of Mathematical Sciences and the hosts for the wonderful and quiet environment where people could think and work well. The workshop also provided enlightenment to a number of mathematicians and graduate students. Particular thanks should be addressed to the organizers especially Professor Lizhen Ji. We also thank the referee for many helpful suggestions on improving the exposition in this paper.

1. Mapping Class Group

In this section, we will introduce basic notions about mapping class groups and elementary techniques in handling Dehn twists.

Let $\Sigma = \Sigma_{g,r}$ be the compact oriented surface of genus g with r boundary components, $r \geq 0$.

Definition 1.1. The *mapping class group* of Σ is given by

$$\Gamma(\Sigma) :\stackrel{\text{def}}{=\!=} \text{Homeo}(\Sigma)/\simeq,$$

where $\text{Homeo}(\Sigma)$ is the set of homeomorphisms from Σ to Σ and the relation $\simeq$ is homotopy of maps.

Remark. The relationship between homotopic and isotopic homeomorphisms on surfaces was established by Baer. Baer's theorem says that they are the same in the following sense.

Theorem 1.2. (Baer) *Suppose f and g are two homotopic homeomorphisms of a compact oriented surface X so that $f = g$ on the boundary of X. Then f is isotopic to g by an isotopy leaving each point in the boundary of X fixed.*

One may consult [St2, ch. 6] or [Ep] for a proof, which includes a discussion on isotopic curve systems on surfaces. For this reason, we will not address the isotopy issues in this paper.

The theory of mapping class groups plays an important role in the study of low-dimensional topology.

Example 1.3. Let H_g be the genus g handlebody, that is, the regular neighborhood of a wedge of g circles in the 3-space. The boundary of H_g is $\Sigma_{g,0}$. It is well-known that for each closed orientable 3-manifold M, there is a positive number g and an $h \in \Gamma(\Sigma_{g,0})$ such that

$$M = H_g \cup_h H_g\,.$$

Such a decomposition of a 3-manifold is called a *Heegaard Splitting*. In other words, M is obtained by gluing up two copies of H_g by a homeomorphism h of their boundary surfaces $\Sigma_{g,0}$. The resulting manifold M depends only on the homotopy class of h.

Example 1.4. Thurston's Virtual Fibre Conjecture states that any closed hyperbolic 3-manifold M has a finite cover N such that

$$N = \Sigma_{g,0} \times [0,1] / \{ (x,0) \sim (h(x),1) : x \in \Sigma_{g,0} \} ,$$

where h is a homeomorphism on $\Sigma_{g,0}$. In short, N is a surface bundle over the circle. This is one of the main conjectures in 3-manifold theory after the resolution of the Geometrization Conjecture.

Example 1.5. Mapping class groups are needed in several other important areas. For example, in the theory of Lefschetz fibrations of 4-manifolds, [Au, Don, Gom]; in the Teichmüller theory of surfaces in which the mapping class group acts on the Teichmm̈uller space, [BH, Pa]; and in the theory of the Moduli space of algebraic curves and Riemann surfaces, [HL].

We will assume basic facts from topology. For example, the Euler Characteristic $\chi(\Sigma_{g,r})$ of a surface satisfies

$$\chi(\Sigma_{g,r}) = 2 - 2g - r \, .$$

Surfaces $\Sigma = \Sigma_{g,r}$ with $\chi(\Sigma) > 0$ are the sphere $\mathbb{S}^2$ and the disk $\mathbb{D}^2$; for $\chi(\Sigma) = 0$, we have the torus $\mathbb{T}^2$ and the annulus $\mathbb{S}^1 \times [0,1]$; all others are of $\chi(\Sigma) < 0$.

As a beginning, the reader is encouraged to work out the following.

Example 1.6. $\Gamma(\mathbb{S}^2) \cong \mathbb{Z}/(2\mathbb{Z})$ and $\Gamma(\mathbb{D}^2) = \{ \mathrm{id} \}$.

The following was proved by J. Nielsen in his Ph.D. thesis in 1913.

Theorem 1.7. (Nielsen) $\Gamma(\mathbb{T}^2) \cong \mathrm{GL}(2,\mathbb{Z})$.

Proof. We will represent $\mathbb{T}^2$ as the quotient space $\mathbb{R}^2 / \mathbb{Z}^2$. There is a natural homomorphism

$$\rho \colon \Gamma(\mathbb{T}^2) \ \to \ \mathrm{Aut}(H_1(\mathbb{T}^2, \mathbb{Z})) = \mathrm{GL}(2,\mathbb{Z})$$

that takes a homeomorphism class $[h]$ to the induced map h_* on the first homology group.

First, we will show that ρ is surjective, i.e., $\mathrm{Image}(\rho) = \mathrm{GL}(2,\mathbb{Z})$.

Let $A \in \mathrm{GL}(2,\mathbb{Z})$. The integral matrix A can be seen as the linear map (by abuse of language)

$$A\colon \mathbb{R}^2 \ \to\ \mathbb{R}^2, \qquad x \mapsto Ax \quad \text{for } x \in \mathbb{R}^2\,,$$

which preserves the integer lattice, i.e., $A(x+n) = A(x) + A(n)$ with $A(n) \in \mathbb{Z}^2$ for every $x \in \mathbb{R}^2$ and $n \in \mathbb{Z}^2$. This clearly induces a homeomorphism on the quotient

$$\tilde{A}\colon \mathbb{T}^2 \ \to\ \mathbb{T}^2, \qquad [x] \mapsto [Ax]\,.$$

It is easy to verify that $\rho(\tilde{A}) = (\tilde{A})_* = A$ with a suitable identification of n and $A(n) \in \mathbb{Z}^2$ with the standard basis of $\mathbb{R}^2$.

Second, to prove that ρ is injective, we will show $\ker(\rho) = \{\,\mathrm{id}\,\}$.

Let $[h] \in \Gamma(\mathbb{T}^2)$ where $h \in \mathrm{Homeo}(\mathbb{T}^2)$ and $h_* = \mathrm{id}$. Represent $\mathbb{T}^2$ as the quotient space of $[0,1]\times[0,1]$ by gluing $[0,1]\times\{\,0\,\}$ to $[0,1]\times\{\,1\,\}$ and $\{\,0\,\}\times[0,1]$ to $\{\,1\,\}\times[0,1]$. Let a, b denote the curves in $\mathbb{T}^2$ corresponding to quotient classes of $[0,1]\times\{\,0,1\,\}$ and $\{\,0,1\,\}\times[0,1]$, respectively. Let P be the point of intersection of a and b.

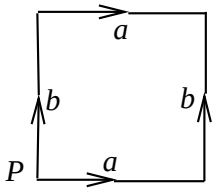

Figure 1.1

Then $H_1(\mathbb{T}^2,\mathbb{Z}) \cong \pi_1(\mathbb{T}^2,P) = \langle a,b\rangle \cong \mathbb{Z}^2$. Also, $\mathbb{T}^2 \setminus \{\,a,b\,\}$ is a disk.

Since $h_* = \mathrm{id}$, we have $h|_a \simeq \mathrm{id}_a$ and $h|_b \simeq \mathrm{id}_b$. Using such facts, we will construct below a continuous function H from $\mathbb{T}^2\times[0,1]$ to $\mathbb{T}^2$ satisfying

- $H|_{\mathbb{T}^2\times\{\,0\,\}} = \mathrm{id}$ and $H|_{\mathbb{T}^2\times\{\,1\,\}} = h$, and
- $H|_{a\times[0,1]}$ is a homotopy of $h|_a$ to id_a, and
- $H|_{b\times[0,1]}$ is a homotopy of $h|_b$ to id_b.

Let $X = (\mathbb{T}^2 \times \{0, 1\}) \cup ((a \cup b) \times [0, 1])$. First of all, we define H from X to $\mathbb{T}^2$ as specified by the above three conditions. Let $\wp : [0, 1]^2 \to \mathbb{T}$ be the quotient map, which induces the map $\wp \times \mathrm{id} : [0, 1]^3 \to \mathbb{T} \times [0, 1]$. Consider the map $H \circ (\wp \times \mathrm{id}) : \partial([0, 1]^3) \to \mathbb{T}^2$. Since $\pi_2(\mathbb{T}^2) = 0$, the map $H \circ (\wp \times \mathrm{id})$ extends to a map $F : [0, 1]^3 \to \mathbb{T}^2$. According to this construction, the map F clearly induces a continuous map $\mathbb{T}^2 \times [0, 1] \to \mathbb{T}^2$. This map is the required homotopy between $\mathrm{id}_{\mathbb{T}^2}$ and h. Hence $[h] = 1 \in \Gamma(\mathbb{T}^2)$. □

Exercise 1.1. A similar proof also shows that if $\chi(\Sigma_{g,r}) \leq 0$ and $h \in \mathrm{Homeo}(\Sigma_{g,r})$ satisfies $h_* = \mathrm{id}$ in $\pi_1(\Sigma_{g,r})$, then $h \simeq \mathrm{id}$ rel $\partial\Sigma_{g,r}$. Here, we need to use $\pi_2(\Sigma_{g,r}) = 0$ for surfaces with $\chi(\Sigma_{g,r}) \leq 0$.

This theorem about $\mathbb{T}^2$ is the key to the understanding of $\Gamma(\Sigma_{g,r})$ for all g, r. Moreover, it can be seen that the essential conditions are that $h_* = \mathrm{id}$ on π_1 and $\pi_k = 0$ for $k \geq 2$.

Definition 1.8. Let $\mathrm{Homeo}^+(\Sigma, \partial\Sigma)$ be the group of all orientation preserving self-homeomorphisms

$$h \colon (\Sigma, \partial\Sigma) \ \to \ (\Sigma, \partial\Sigma), \qquad h|_{\partial\Sigma} = \mathrm{id} \ .$$

The *relative mapping class group* is given by

$$\Gamma^*(\Sigma) :\overset{\text{def}}{=\!=} \mathrm{Homeo}^+(\Sigma, \partial\Sigma) \big/ (\simeq_{\partial\Sigma}) \, ,$$

where $\simeq_{\partial\Sigma}$ is the homotopy relative to $\partial\Sigma$, i.e., homotopy that leaves $\partial\Sigma$ pointwise fixed.

Example 1.9. *Standard Dehn twist on the annulus*

Consider $\Sigma_{0,2} = \mathbb{S}^1 \times [0, 2]$, we will demonstrate a generator for $\Gamma^*(\Sigma_{0,2})$. Let $c = \mathbb{S}^1 \times \{1\}$ be the central curve of the annulus. Define

$$D_c \colon \mathbb{S}^1 \times [0, 2] \ \to \ \mathbb{S}^1 \times [0, 2], \qquad D_c(e^{i\theta}, t) = (e^{i(\theta + \pi t)}, t) \, .$$

Clearly, $D_c \in \mathrm{Homeo}^*(\Sigma_{0,2}, \partial\Sigma_{0,2})$. It is called the *standard Dehn twist* on the annulus along the curve c. Note that it is a twist of one full turn on the annulus leaving the boundary circles pointwise fixed.

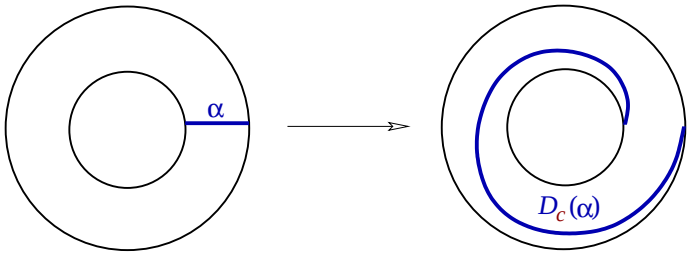

Figure 1.2: The orientation of $\mathbb{S}^1 \times [0,2]$ gives a normal into the paper.

Remark. The above picture is drawn according to traditional convention that a positive turn is counterclockwise. That is equivalent to taking a normal into the paper. In the future, pictures of Dehn twists on surfaces are drawn with respect to the right hand orientation of the front face.

If $\Sigma_{0,2}$ is drawn as a cylinder with outward normal orientation, the Dehn twist along the central meridian curve c is drawn as follows. The rule is that if one traces from either end of α, the image curve $D_c(\alpha)$ turns right into c and follows c, then turns left into the other end of α.

Figure 1.3: Drawn with outward normal orientation.

Exercise 1.2. Show that $D_c \simeq \mathrm{id}$ but $D_c \not\simeq \mathrm{id}$ rel $\partial\Sigma_{0,2}$.

Proposition 1.10. $\Gamma^*(\mathbb{S}^1 \times [0,2]) \simeq \mathbb{Z}$.

Sketch of proof. An outline of the the proof is given here. Details are left as exercises to the reader. For simplicity, let $A = \mathbb{S}^1 \times [0,2]$.

Consider the homomorphism $\varphi : \Gamma^*(A) \to H_1(A, \partial A)$ defined by $\varphi([h]) = [h(\alpha)]$ where α is the arc $e^{2\pi \mathbf{i}} \times [0, 2]$. The homomorphism φ is surjective as shown in Example 1.9. To show that φ is injective, suppose that $[h(\alpha)] = 0 \in H_1(A, \partial A)$. Then the curves α and $h(\alpha)$ are isotopic by an isotopy leaving ∂A pointwise fixed. Using the fact that $A \setminus (\alpha \cap \partial A)$ is an open disk, one obtains a homotopy between h and the identity. □

We will assume certain basic facts from surface theory without proof. Readers may refer to [St, CB, Ro].

Example 1.11. For the torus, $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, let a, b be the curves defined by $a = \wp(0 \times \mathbb{R})$ and $b = \wp(\mathbb{R} \times 0)$, where $\wp \colon \mathbb{R}^2 \to \mathbb{T}^2$ is the covering projection. Two homeomorphisms D_a and D_b on $\mathbb{T}^2$ are induced by the following linear transformations (using the same notations) on $\mathbb{R}^2$

$$D_a = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \qquad D_b = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

As a side remark, the homeomorphisms D_a and D_b are in fact Dehn twists on $\mathbb{T}^2$. Since $\mathrm{SL}(2, \mathbb{Z})$ is clearly generated by these two matrices, it follows that D_a and D_b generate the index two subgroup of $\Gamma(\mathbb{T}^2)$ corresponding to orientation preserving homeomorphisms. The curve $D_a(b)$ is shown in the picture.

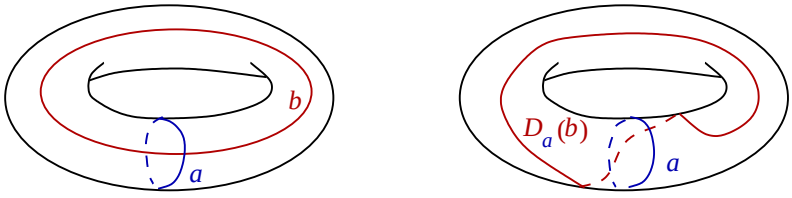

Figure 1.4

Exercise 1.3. Show that $D_a(b) = a - b$ where the curves are also interpreted as homology classes.

Example 1.12. The group $\Gamma(\Sigma_{g,r})$ acts transitively on the r circle boundary components. Thus, if $r \geq 2$, then the mapping class group and the relative mapping class group are different. Note that Dehn twists do not permute boundary components.

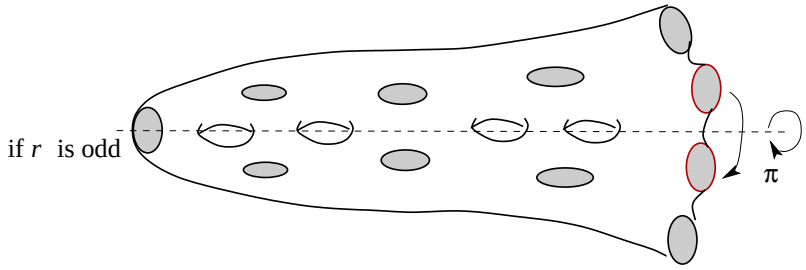

Figure 1.5

The proof of Dehn-Lickorish Theorem depends on the study of simple loops on surfaces. In our context, unless it is specified otherwise, a *simple loop* in a surface Σ is a simple closed curve in the interior $\mathrm{Int}(\Sigma)$. Let us begin with the following.

Definition 1.13. Let Σ be a connected surface. A simple loop $s \subset \mathrm{Int}(\Sigma)$ is *nonseparating* if $\Sigma \setminus s$ is connected. Otherwise, it is separating.

In the case that a simple loop s is separating in a surface $\Sigma_{g,r}$ with boundary, we say that s *bounds the boundary components* $b_1, \ldots, b_k$ of $\Sigma_{g,r}$ if $b_1, \ldots, b_k$ are those boundary components other than s of a connected component of $\Sigma_{g,r} \setminus s$. If s bounds a single boundary component b, we say that s is *parallel to the boundary* b.

Remark. A simple loop s bounds the boundary components $b_1, \ldots, b_k$ of $\Sigma_{g,r}$ if and only if it also bounds $b_{k+1}, \ldots, b_r$. Any simple loop in $\Sigma_{0,r}$ is separating.

Lemma 1.14. *If s, s' are two nonseparating simple loops in $\mathrm{Int}(\Sigma_{g,r})$, then there is an $h \in \mathrm{Homeo}^+(\Sigma_{g,r}, \partial\Sigma_{g,r})$ such that $h(s) = s'$.*

Proof. Since both s and s' are nonseparating, the surfaces Σ_s and $\Sigma_{s'}$ obtained by removing small open neighborhoods of s and s' have genus $g-1$ and $r+2$ boundary

circles. Let us denote the new boundary circles $s_\pm$ and $s'_\pm$ respectively in the surfaces Σ_s and $\Sigma_{s'}$. Since $\Gamma(\Sigma_{g-1,r+2})$ induces transitive actions on the boundary circles, one may choose a homeomorphism h_0 from Σ_s to $\Sigma_{s'}$ such that $h_0(s_\pm) = s'_\pm$ and $h_0|_{\partial\Sigma} = \mathrm{id}$. Now, if one reglues the surfaces Σ_s and $\Sigma_{s'}$ along s and s' respectively, one obtains a homeomorphism h on $\Sigma_{g,r}$, which is induced by h_0 and it satisfies $h(s) = s'$. □

Lemma 1.15. *If s, s' are two simple loops in $\Sigma_{0,r}$ bounding the same boundary components, then there exists an $h \in \mathrm{Homeo}^+(\Sigma_{0,r}, \partial\Sigma_{0,r})$ such that $h(s) = s'$.*

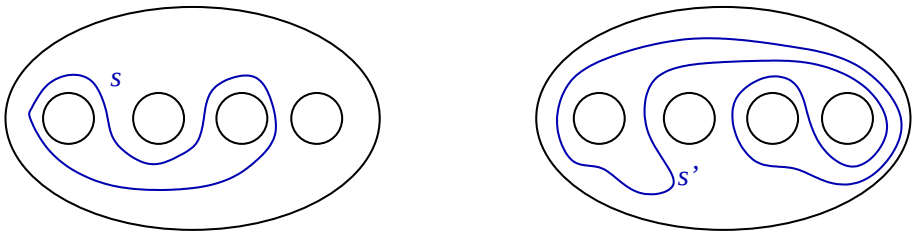

Figure 1.6

Exercise 1.4. Prove this lemma, using the technique similar to that in Lemma 1.14.

Definition 1.16. (Dehn Twist) Let s be a simple loop in $\mathrm{Int}(\Sigma_{g,r})$. A *positive Dehn twist* or simply *Dehn twist* along s, $D_s \colon \Sigma_{g,r} \to \Sigma_{g,r}$, is a homeomorphism defined as follows.

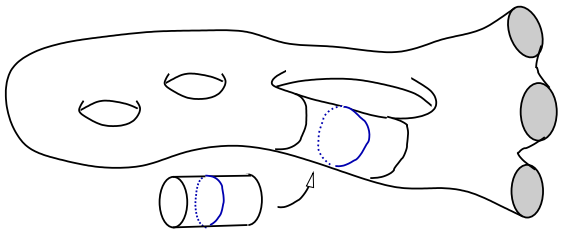

Figure 1.7

Let $c = \mathbb{S}^1 \times \{1\} \in \mathbb{S}^1 \times [0,2] = \Sigma_{0,2}$ and

$$\mathcal{I} \ : \ (\Sigma_{0,2}, c) \hookrightarrow (\mathcal{N}(s), s) \subset \Sigma_{g,r}$$

be an orientation preserving embedding onto a neighborhood $\mathcal{N}(s)$ of s such that the central curve $c \subset \Sigma_{0,2}$ is mapped to $\mathcal{I}(c) = s \subset \Sigma_{g,r}$. Denote D_c the Standard

Dehn twist on the annulus along c (see Example 1.9). Define

$$D_s :\overset{\text{def}}{=\!=} \mathcal{I} \circ D_c \circ \mathcal{I}^{-1}$$

and extend it to $\Sigma_{g,r}$ by the identity map outside $\mathcal{N}(s)$.

It will be important in our study to know how to do computations with Dehn twists. We will finish the section with a discussion on it.

Let a, b be simple loops in a surface that intersect transversely in k points. One way to find $D_a(b)$ is by using the resolution of the intersection points.

Definition 1.17. Let x, y be two smooth arcs in Σ intersecting transversely at a point $p \in \Sigma$. The *resolution of $x \cup y$ at p from x to y* is defined as follows (the picture is drawn with right-hand orientation on the plane).

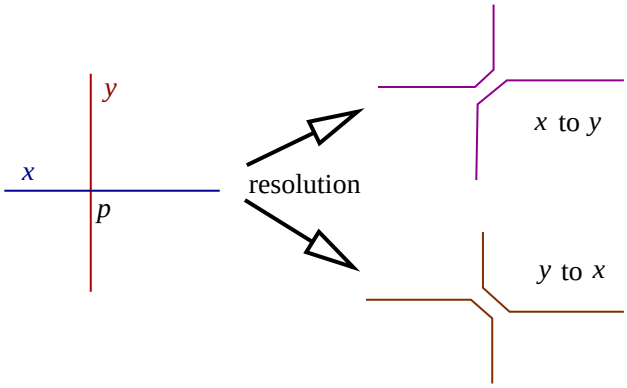

Figure 1.8

Take any orientation of x and determine the orientation of y at the intersection point p such that the orientations on x and y from x to y are consistent with the orientation of Σ. Then resolve the intersection point p according to the orientations on x and y as in the picture above.

Note that if the orientation of x is reversed, so is the one of y; and so the resolution does not change. Thus, the resolution depends only on the order of (x, y) and the orientation of the surface Σ.

Lemma 1.18. Computation of $D_a(b)$. *Assume that $|a \cap b| = k \in \mathbb{Z}$ transversely, then $D_a(b)$ is obtained by taking k parallel copies ka of a and resolving all intersections of $(ka) \cap b$ from ka to b.*

Remark. In this case, one has to perform resolution at a total of k^2 intersections. At each intersection, it is sufficient to choose the compatible orientations of the curves locally.

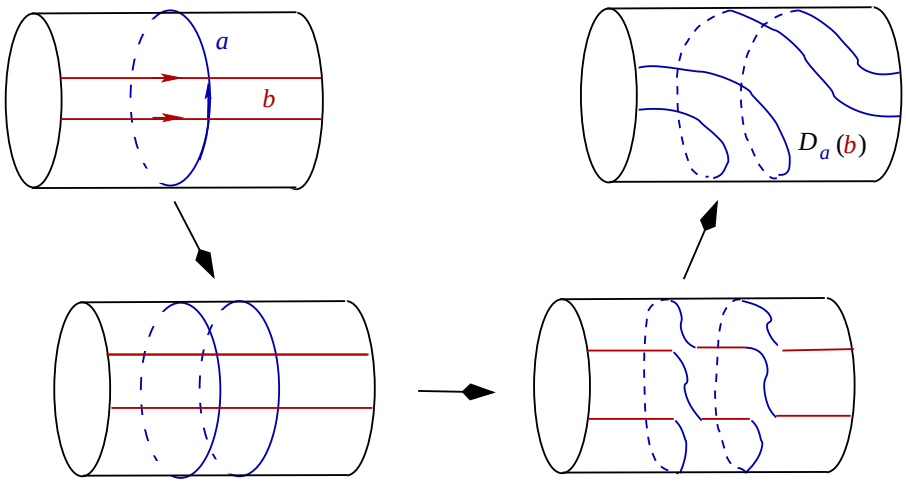

Figure 1.9: An example of $D_a(b)$ with $k = 2$.

Lemma 1.19. *If a, b are simple loops in $\Sigma_{g,r}$ such that they intersect transversely at a single point (i.e., $|a \pitchfork b| = 1$), then $D_a D_b(a) \simeq b$. In addition, $(D_b\, D_a\, D_a\, D_b)(a) \simeq a$ with orientation reversed.*

Proof. Let $\mathcal{I} : \mathbb{S}^1 \times [0, 2] \hookrightarrow \mathcal{N}(a) \subset \Sigma_{g,r}$ be an embedding of the annulus onto a regular neighborhood $\mathcal{N}(a)$ of a such that $b \cap \mathcal{N}(a) = \mathcal{I}(\{\,1\,\} \times [0, 2])$. Moreover, let $a \cap b \in \mathcal{N}(a)$ be shown as in the first picture of Figure 1.10 below. Then, after resolving the intersection $a \cap b$, $D_b(a)$ intersects a at one point in $\mathcal{N}(a)$ as in the second picture. A further resolution at this point gives $D_a D_b(a) \subset \mathcal{N}(a)$ as in the third picture. On the other hand, outside $\mathcal{N}(a)$, the curves $D_b(a)$ and $D_a D_b(a)$ coincide with b. From the third picture, $D_a D_b(a)$ is homotopic to b with a homotopy supported in $\mathcal{N}(a)$.

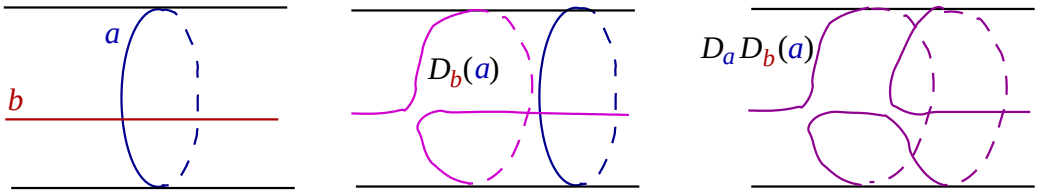

Figure 1.10

The second result is proved similarly and is illustrated by the following pictures in Figure 1.11. In the first picture, $\{p, p'\} = b \cap \partial\mathcal{N}(a)$. After resolving the intersection, we have the second picture where $\{p, p'\} = D_a(b) \cap \partial\mathcal{N}(a)$ and $\{q, q'\} = b \cap \partial\mathcal{N}(a)$. Then, a further resolution leads to the third picture.

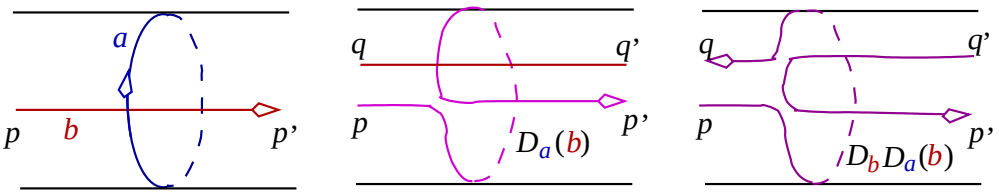

Figure 1.11

It is indicated by the arrows in the last picture that $D_b D_a(b) \simeq a$ with orientation reversed. □

2. Dehn-Lickorish Theorem

The main objective of this section is to prove the following fundamental theorem.

Dehn-Lickorish Theorem. *For any orientable compact surface Σ, the mapping class group relative boundary*

$$\Gamma^*(\Sigma) = \mathrm{Homeo}^+(\Sigma, \partial\Sigma) \big/ (\simeq_{\partial\Sigma})$$

is generated by Dehn twists.

Remark. As it is mentioned before, a Dehn twist does not permute boundary components.

The proof of the Dehn-Lickorish Theorem is essentially following the idea of the torus case in Theorem 1.7. One first establishes by induction a homotopy relative boundary between a homeomorphism and a composition of Dehn twists on a set of simple closed curves. Then this homotopy is extended to the whole surface. For this purpose, we will begin with the study of the action of Dehn twists on curves.

Definition 2.1. Let $a, b \subset \mathrm{Int}(\Sigma)$ be two simple closed curves. We say that $a \sim b$ if there exists a finite sequence of simple closed curves $c_1, \ldots, c_n$ such that

$$D_{c_1}^{\epsilon_1} \circ D_{c_2}^{\epsilon_2} \circ \cdots \circ D_{c_n}^{\epsilon_2}(a) = b\,,$$

where $\epsilon_j \in \mathbb{Z}$, $j = 1, \ldots, n$.

Example 2.2. If $|a \pitchfork b| = 1$, then $a \sim b$. This is shown in Lemma 1.19 in the previous section. Moreover, if a^{-1} denotes the same curve a with orientation reversed, by choosing b with $|a \pitchfork b| = 1$, one has $a \sim a^{-1}$.

Here is a simple way to characterize nonseparating curves (recall Definition 1.13) and see their equivalences. The proof of the lemma is left as an exercise.

Lemma 2.3. *A simple closed curve $a \subset \Sigma$ is nonseparating if and only if there is a simple closed curve $b \subset \Sigma$ such that $b \pitchfork a$ (transversely) at a point.*

Proposition 2.4. (Dehn-Lickorish)

(a) *If a, b are nonseparating, then $a \sim b$.*

(b) *If a, b are separating in $\Sigma_{0,r}$ and they bound the same boundary components of Σ, then $a \sim b$.*

Proof. For (a), let $a, b \subset \mathrm{Int}(\Sigma)$ be nonseparating simple closed curves. We will prove by induction on

$$I(a, b) = \min \left\{ \left| a' \cap b' \right| : a' \simeq a,\ \ b' \simeq b,\ \ a', b' \subset \Sigma \right\} ,$$

where $\simeq$ denotes isotopy between curves in Σ.

We will handle the proof in four cases:

Case 1. $I(a,b)=0$;

Case 2. $|a\cap b|=2$ with opposite algebraic intersection signs at the two intersection points;

Case 3. $I(a,b)\geq 2$ with the same algebraic intersection sign at two adjacent intersection points along the curve a;

Case 4. $I(a,b)\geq 3$ with alternating algebraic intersection signs at three consecutive intersection points along the curve a.

Case 1: Observe that if $I(a,b)=0$, no matter whether $\Sigma\setminus(a\cup b)$ is connected, there is a simple closed curve c such that $|a\cap c|=|b\cap c|=1$. Indeed, assume that $\Sigma\setminus(a\cup b)$ is disconnected, then it has two connected components Σ_1 and Σ_2 because both a,b are nonseparating. For each $j=1,2$, Σ_j has boundary curves a_j,b_j corresponding to a,b respectively. Choose a simple arc c_j to connect a_j to b_j in Σ_j. Then c is formed by c_1,c_2. The construction for connected $\Sigma\setminus(a\cup b)$ is similar (see Figure 2.1 below). In any case, there is a simple loop $c\subset \mathrm{Int}(\Sigma)$ such that $c\sim a$ and $c\sim b$, as mentioned in Example 2.2 above. Thus, $a\sim b$.

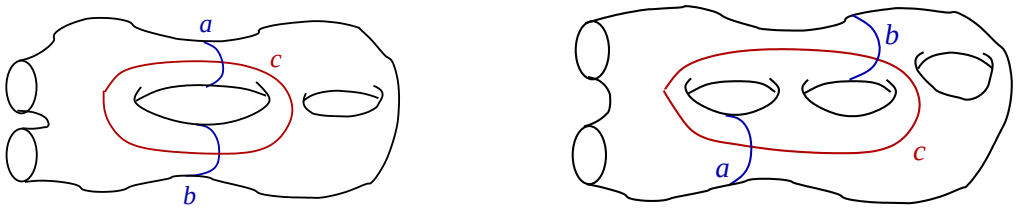

Figure 2.1

Case 2: Suppose $a\cap b=\{p,q\}$ with the signs at the intersections are opposite.

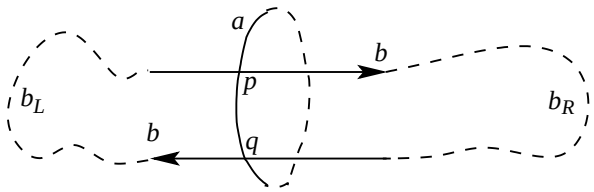

Figure 2.2

The points p, q divide the curve b into segments b_L and b_R as in Figure 2.2. Let b_1 and b_2 be the two boundary components of a regular neighborhood of $a \cup b_R$ such that b_1 and b_2 are not homotopic to a as in Figure 2.3 below.

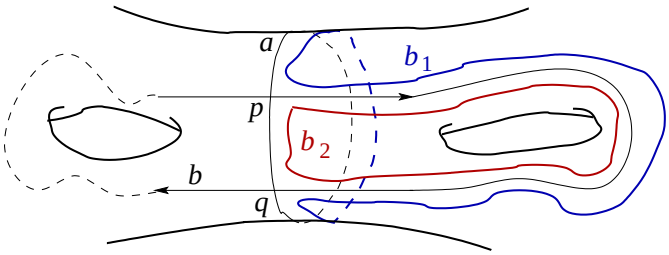

Figure 2.3

The three curves a, b_1, b_2 bound a pair of pants, $\Sigma_{0,3}$ in Σ. Since the curve a is nonseparating, at least one of b_1 or b_2 is also nonseparating. Say, b_1 is nonseparating. Note, $|a \cap b_1| = |b \cap b_1| = 0$. By the first case above, we have $a \sim b_1$ and $b_1 \sim b$ and hence $a \sim b$.

Now, we deal with the remaining two cases which are illustrated in Figure 2.4, in which p, q, r are consecutive intersection points along the curve a.

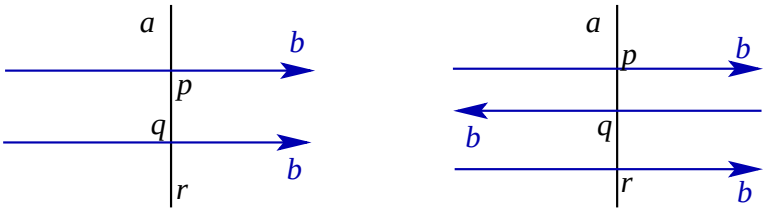

Figure 2.4

Each of the two cases can be handled by replacing the curve b with another nonseparating curve b' that satisfies $b' \sim b$ and $0 < |b' \cap a| < |b \cap a|$. Then, eventually we have $|b' \cap a| = 1$ by induction and so $b' \sim a$.

Case 3: There are two points $p, q \in a \cap b$ that are adjacent along the curve a such that the signs of intersection are the same. In this case, we perform a surgery on the curve b to obtain a curve b' as in Figure 2.5 below.

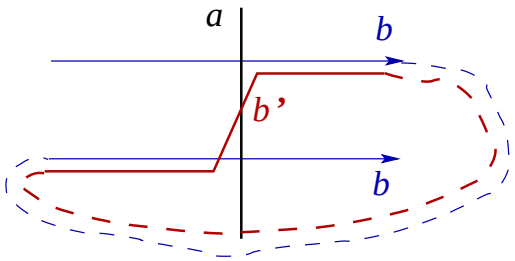

Figure 2.5

Then, $|b' \pitchfork b| = 1$ so b' is nonseparating by Lemma 2.3 and thus $b \sim b'$ by Lemma 1.19. Furthermore, $|b' \cap a| < |b \cap a|$.

Case 4: There are three points $p, q, r \in a \cap b$ that are consecutive along the curve a such that the signs of intersection alternate. We will work on a neighborhood $\mathcal{U}$ of the subarc of the curve a containing $\{ p, q, r \}$, as illustrated in Figure 2.6 below. Note that there are two possible ways of connecting the three subarcs of the curve b as shown. Nevertheless, we will only provide detailed illustrations for the first one. Then, it can be seen that the same proof applies to the other situation.

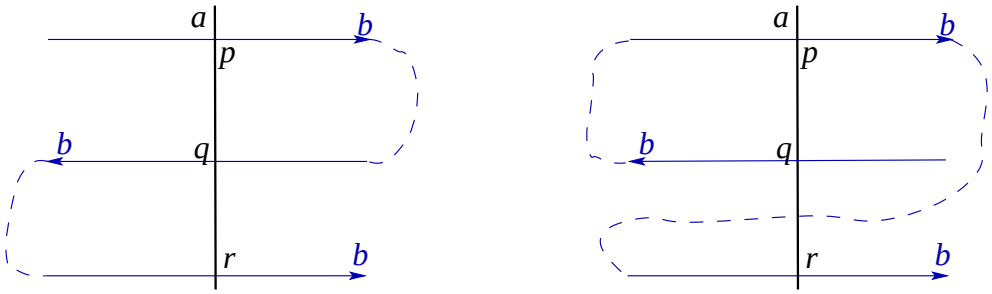

Figure 2.6

For the picture on the left hand side above, we perform a surgery on the curve b to obtain a curve c with $|c \cap b| = 2$ as in Figure 2.7(a) below. Then, to obtain $D_c(b)$, we need to consider the double of c as in Figure 2.7(b).

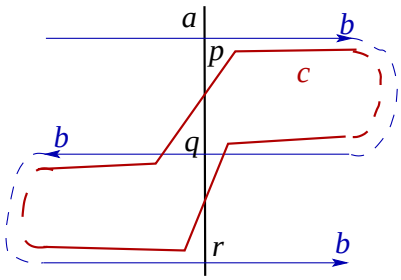

Figure 2.7(a)

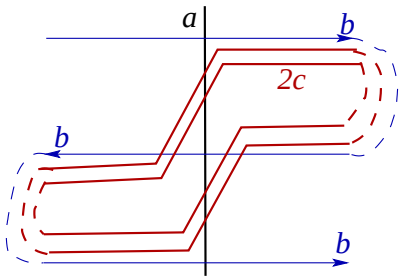

Figure 2.7(b)

Then after the resolution of $b \cap (2c)$, we obtain $b' = D_c(b)$. By definition of the relation $\sim$, we have $b' \sim b$. The curves a, b' within the neighborhood $\mathcal{U}$ are shown as in Figure 2.8(a) below.

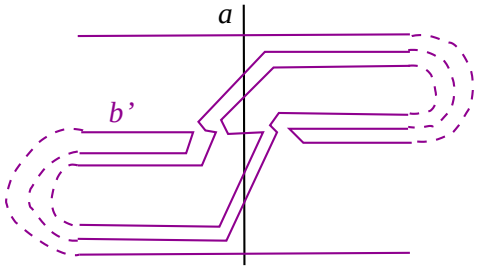

Figure 2.8(a)

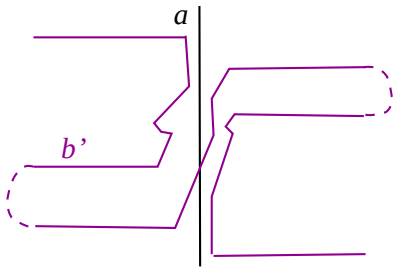

Figure 2.8(b)

Then, up to isotopy supported within a neighborhood of b, the curve b' can be simplified as in Figure 2.8(b). Note that the dotted parts of b in Figure 2.7(a,b) may intersect a, say k times, outside the neighborhood $\mathcal{U}$ of $\{p, q, r\}$. This gives rise to $3k$ points of intersection between a and the dotted parts of b' in Figure 2.8(a). Nevertheless, after the isotopy as in Figure 2.8(b), the dotted part only have k points of intersection with a. Hence, $|b' \cap a| < |b \cap a|$. In addition, the curve b' is the homeomorphic image of a nonseparating loop b, thus b' is also nonseparating.

Note that the above proof is also valid for the second connection of subarcs of the curve b. It is because intersection resolutions are done within the neighborhood $\mathcal{U}$ of $\{p, q, r\}$ and isotopies are supported in a neighborhood of b.

For the proof of statement (b), note that on $\Sigma_{0,r}$, if $a \cap b = \emptyset$ and if they bound the same boundary components, then a is isotopic to b.

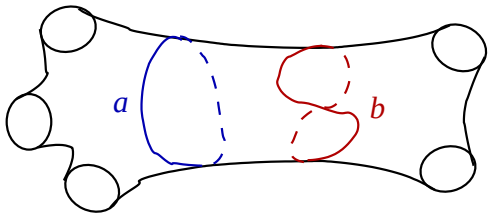

Figure 2.9

Thus, we only need to consider the case that $a \cap b \neq \emptyset$ and $|a \cap b|$ is even. Observe that in this situation, all adjacent intersection points have opposite signs.

Now, we claim that $|a \cap b| \geq 4$. Otherwise, $|a \cap b| = 2$ and then they must bound different boundary components.

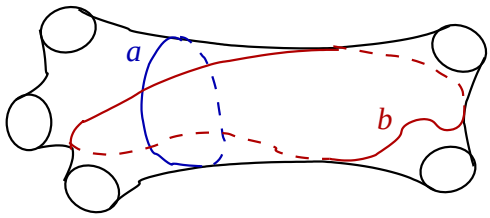

Figure 2.10

Therefore, we have $|a \cap b| \geq 4$ where all adjacent intersection points have opposite signs. This has been dealt with above. Hence $a \sim b$. □

We will need the following basic property of Dehn twists in relation to a homeomorphism on the surface.

Lemma 2.5. *Let $a, b \subset \Sigma$ be simple loops and $\varphi\colon \Sigma \to \Sigma$ be an orientation preserving homeomorphism such that $\varphi(a) = b$. Then $D_b \simeq \varphi \circ D_a \circ \varphi^{-1}$.*

Proof. The proof is left as an exercise. □

Now, we are ready to prove the Dehn-Lickorish Theorem. It will be proved by induction on the number $|\Sigma_{g,r}| = 3g + r$. We will only consider the case that Euler Characteristic $\chi(\Sigma) < 0$. The cases that $\chi(\Sigma) = 0$, i.e., the annulus and the torus, have been dealt with before in Theorem 1.7 and Proposition 1.10.

It should be noted that the starting case is the 3-hole sphere, which is the unique surface with $|\Sigma_{0,3}| = 3$.

Lemma 2.6. *$\Gamma^*(\Sigma_{0,3})$ is generated by Dehn twists along the three boundaries.*

Proof. The proof makes use of the following two facts, of which the proofs will be left as exercises.

FACT 1. If α, β are two embedded arcs in $\mathrm{Int}(\mathbb{D}^2)$ with the same end points, $\alpha(0) = \beta(0)$ and $\alpha(1) = \beta(1)$, then α is isotopic to β by an isotopy leaving $\partial\mathbb{D}^2$ and the end points fixed.

FACT 2. Let α, β be two arcs in the annulus $\Sigma_{0,2}$ having the same starting point in a boundary component and the same terminal point in another boundary component. Then α is isotopic to $D_c^n(\beta)$ for some $n \in \mathbb{Z}$ where c is the central curve in $\Sigma_{0,2}$.

Let $f : \Sigma_{0,3} \to \Sigma_{0,3}$ be a homeomorphism with $f|_{\partial\Sigma_{0,3}} = \mathrm{id}$ and let D_1, D_2, and D_3 are the Dehn twists along the three boundary components b_1, b_2, and b_3 of $\Sigma_{0,3}$ respectively. We are going to show that f is isotopic to $D_1^m D_2^n D_3^\ell$ for some $m, n, \ell \in \mathbb{Z}$. Take an embedded arc $s \in \Sigma_{0,3}$ with $s(0) \in b_1$ and $s(1) \in b_2$. Then $f(s)$ is also an embedded arc joining $s(0)$ to $s(1)$. Let $E = \mathbb{D}^2$ be the quotient space obtained by identifying b_1 to a point and also b_2 to a point. Then the quotient arcs $[s]$ and $[f(s)]$ are both embedded arcs in $\mathrm{Int}(E) = \mathrm{Int}(\mathbb{D}^2)$ joining the quotient points $[b_1]$ and $[b_2]$. By FACT 1, $[s]$ and $[f(s)]$ are isotopic in E. Since the isotopy leaves ∂E fixed, it may be lifted to an isotopy in $\Sigma_{0,3}$. Without loss of generality, by applying isotopic changes, we may assume that $f(s) = s$ on $\Sigma_{0,3} \setminus (N(b_1) \cup N(b_2))$ where $N(b_j)$ is a regular neighborhood of b_j for $j = 1, 2$. Applying FACT 2 on $N(b_1)$ and $N(b_2)$, which are both homeomorphic to $\Sigma_{0,2}$, we may assume that $f|_s = (D_1^m D_2^n)|_s$ for some $m, n \in \mathbb{Z}$.

By a further isotopy of f, we may assume that

$$f|_{N(s \cup b_1 \cup b_2)} = (D_1^m D_2^n)|_{N(s \cup b_1 \cup b_2)}$$

where $N(s \cup b_1 \cup b_2)$ is a regular neighborhood of $s \cup b_1 \cup b_2$. Now, consider the set $A = \Sigma_{0,3} \setminus \mathrm{Int}(N(s \cup b_1 \cup b_2))$, which is an annulus $\Sigma_{0,2}$. The restriction map

$$(D_1^m D_2^n f)|_A \ : \Sigma_{0,3} \setminus \mathrm{Int}(N(s \cup b_1 \cup b_2)) \to \Sigma_{0,3} \setminus \mathrm{Int}(N(s \cup b_1 \cup b_2))$$

is a self-homeomorphism of the annulus A and it equals the identity on ∂A. Thus, by the known fact (Proposition 1.10) that $\Gamma^*(\Sigma_{0,2})$ is generated by the Dehn twist along a boundary component, say b_3, we conclude that the map $(D_1^m D_2^n f)|_A$ is isotopic to D_3^ℓ, for some $\ell \in \mathbb{Z}$, by an isotopy leaving boundary pointwise fixed. Then we simply extend this isotopy to an isotopy on $\Sigma_{0,3}$ by the identity on $N(s \cup b_1 \cup b_2)$. As a result, f is isotopic to $D_1^m D_2^n D_3^\ell$ for $m, n, \ell \in \mathbb{Z}$. □

Lemma 2.7. *Let $s \in \Sigma$ be a simple loop that is neither null-homotopic nor boundary parallel. If Σ' is a connected component obtained from Σ by cutting along s, then $|\Sigma'| < |\Sigma|$.*

The proof is left as an exercise.

Proof of Dehn-Lickorish Theorem. Suppose that the statement of the theorem holds for all Σ with $|\Sigma| < k$ where $k \geq 3$. Let Σ be a surface $\Sigma_{g,r}$ with $|\Sigma| = k$ and $h \in \mathrm{Homeo}^+(\Sigma, \partial\Sigma)$ be a homeomorphism.

Assume first that $\mathrm{genus}(\Sigma) > 0$. Choose a nonseparating loop $s \subset \Sigma$. Consider s and its image $h(s)$. Since both are nonseparating, Proposition 2.4 gives that $h(s) \sim s$. In other words, there is an orientation preserving homeomorphism φ of Σ, which is a composition of Dehn twists, such that $\varphi(s) = h(s)$. By Example 2.2, if s is oriented, we may assume that $h(s)$ and $\varphi(s)$ have the same orientation. Hence, one may further assume that $h|_s = \varphi|_s$, i.e., $(\varphi^{-1} \circ h)|_s = \mathrm{id}$.

Now, cut Σ open along the loop s to obtain a surface Σ' with $|\Sigma'| < |\Sigma|$. In addition, we can regard the homeomorphism $\varphi^{-1} \circ h$ as an element of $\mathrm{Homeo}^+(\Sigma', \partial\Sigma')$. By the induction hypothesis, $\varphi^{-1} \circ h$ is a product of Dehn twists on Σ', which corresponds to a product of Dehn twists on Σ. Thus, h is a product of Dehn twists on Σ.

For a surface Σ with genus$(\Sigma) = 0$, we may take a simple loop $s \subset \Sigma$ such that it bounds two boundary components. Then we may apply the same argument as above on s and $h(s)$ and conclude by induction. □

3. Hyperbolic Plane and Hyperbolic Surfaces

In this section, we will first recall briefly the geometry of the hyperbolic plane and hyperbolic structures on surfaces. Then we will discuss the relationship between the fundamental group of a surface, the deck transformation group, and the geometric action of groups on surfaces. These notions will be frequently used in the proof of the Dehn-Nielsen Theorem.

3.1. **A Crash Introduction to the Hyperbolic Plane.**

We will mainly use the upper half plane model of the Hyperbolic Plane. The upper half plane is given by

$$\mathbb{H}^2 = \{z \in \mathbb{C} \mid \operatorname{Im}(z) > 0\}, \qquad \operatorname{Im}(z) = y \text{ where } z = x + \mathbf{i}y, \quad x, y \in \mathbb{R}.$$

Its boundary in the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ is denoted $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ and is called the *circle of infinity*.

The *hyperbolic metric* on $\mathbb{H}^2$ is the Riemannian metric defined by the symmetric 2-tensor

$$ds^2 = \frac{dx^2 + dy^2}{y^2} = \frac{-4dzd\overline{z}}{(z - \overline{z})^2}.$$

The area form of the metric is $\dfrac{dx \wedge dy}{y^2}$. If the length of a smooth curve $\gamma \subset \mathbb{H}^2$ is denoted by $L(\gamma)$, then

$$L(\gamma) = \int_a^b \frac{|\gamma'(t)|}{\operatorname{Im}(\gamma(t))}\, dt.$$

There is a way to identify the group of orientation preserving isometries of the hyperbolic plane, $\operatorname{Isom}^+(\mathbb{H}^2)$, with the matrix group $\mathrm{PSL}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R})/\{\pm I\}$. Let us start with three specific isometries. Let $f_\lambda(z) = \lambda z$ where $0 < \lambda \in \mathbb{R}$ and $g_\mu(z) = z + \mu$ where $\mu \in \mathbb{R}$. Then by a simple calculation, both f_λ and g_μ are orientation preserving hyperbolic isometries. In addition, the map $h(z) = \dfrac{-1}{z}$

preserves the hyperbolic metric. Indeed, let $w = h(z)$. Then, the pull-back metric $h^*(ds^2)$ is given by

$$h^*(ds^2) = \frac{-4\,dw\,d\overline{w}}{(w-\overline{w})^2} = \frac{-4d(\frac{1}{z})d(\frac{1}{\overline{z}})}{(\frac{1}{z}-\frac{1}{\overline{z}})^2}$$
$$= -4\frac{1/(z^2\overline{z}^2)dzd\overline{z}}{(\overline{z}-z)^2/(z^2\overline{z}^2)} = \frac{-4\,dz\,d\overline{z}}{(z-\overline{z})^2}.$$

Thus $h : \mathbb{H}^2 \to \mathbb{H}^2$ is an orientation preserving isometry.

A Möbius transformation

$$F(z) = \frac{az+b}{cz+d}, \qquad \text{where} \qquad ad-bc = 1,$$

preserves the upper half plane if and only if it has real coefficients a, b, c, d.

Now it is well known that each Möbius transformation F above is a composition of f_λ, g_λ and h for $\lambda \in \mathbb{R}_{>0}$, $\mu \in \mathbb{R}$. Thus, $F \in \mathrm{Isom}^+(\mathbb{H}^2)$. It follows that $\mathrm{PSL}(2,\mathbb{R})$ acts on $\mathbb{H}^2$ as a group of isometries by the correspondence

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \frac{az+b}{cz+d}.$$

We leave it as an exercise for the readers to verify that the group of all orientation preserving isometries of $\mathbb{H}^2$ is $\mathrm{PSL}(2,\mathbb{R})$. *Hint:* show that $\mathrm{SL}(2,\mathbb{R})$ acts transitively on the unit tangent vectors in $\mathbb{H}^2$.

We will give a classification of the orientation preserving isometries of $\mathbb{H}^2$ by looking at the number of fixed points of the isometry in $\overline{\mathbb{H}^2}$, where $\overline{\mathbb{H}^2}$ is the closure of $\mathbb{H}^2$ in the Riemann sphere $\overline{\mathbb{C}}$.

Let $\gamma \in \mathrm{Isom}^+(\mathbb{H}^2)$ be $\gamma(z) = \frac{az+b}{cz+d}$, $a, b, c, d \in \mathbb{R}$, $ad-bc = 1$. Let $z \in \overline{\mathbb{H}^2}$ be a fixed point of γ, by definition, it means that

$$\frac{az+b}{cz+d} = z, \quad \text{i.e.,}$$

$$cz^2 + (d-a)z - b = 0. \tag{1}$$

Note that Equation (1) is quadratic if $c \neq 0$ and is linear if $c = 0$. This equation always has a solution z in the Riemann sphere. Note that its conjugate $\overline{z}$ is also

a solution to (1) because of the fact that $a, b, c, d \in \mathbb{R}$. We may thus assume that z is a solution to (1) in $\overline{\mathbb{H}^2}$.

If $c \neq 0$ and the discriminant $\Delta = (d-a)^2 + 4bc < 0$, then the point z lies in $\mathbb{H}^2$. In this case, γ has a unique fixed point in $\mathbb{H}^2$. If $c \neq 0$ and $\Delta = 0$, then the quadratic equation has a unique real solution $\frac{a-d}{2c}$, which is again the only fixed point of γ lying in $\overline{\mathbb{R}}$.

If $c \neq 0$ and $\Delta > 0$, then the quadratic equation has two distinct real solutions, which are the two fixed points of γ lying in $\overline{\mathbb{R}}$.

Finally, if $c = 0$, $d - a \neq 0$, then $\gamma(z) = \frac{a}{d}z + \frac{b}{d}$. Thus, γ has two fixed points $\frac{b}{d-a}$ and ∞. If $c = 0$, $d - a = 0$, then $\gamma(z) = z + \frac{b}{d}$, and ∞ is the unique fixed point for $\gamma \neq \mathrm{id}$.

To summarize the discussion above, we have the following classification.

Definition 3.1. We say that $\gamma \in \mathrm{Isom}^+(\mathbb{H}^2)$ is of

- *elliptic* type if $\gamma \neq \mathrm{id}$ and it has a fixed point in $\mathbb{H}^2$;
- *parabolic* type if γ has no fixed point in $\mathbb{H}^2$ and only one fixed point in $\overline{\mathbb{H}^2}$;
- *hyperbolic* type if γ has no fixed point in $\mathbb{H}^2$ and two distinct fixed points in $\overline{\mathbb{H}^2}$.

It is clear from the definition that conjugate isometries, γ and $\sigma\gamma\sigma^{-1}$, where $\gamma, \sigma \in \mathrm{Isom}^+(\mathbb{H}^2)$, are of the same type.

Below is an algebraic characterization of the types of isometries.

Proposition 3.2. *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{R})$ *be a matrix representative of* $\gamma \in \mathrm{Isom}^+(\mathbb{H}^2)$, $\gamma \neq \mathrm{id}$, *then*

(1) γ *is of elliptic type if and only if* $|\mathrm{tr}(A)| < 2$;
(2) γ *is of parabolic type if and only if* $|\mathrm{tr}(A)| = 2$;

(3) *γ is of hyperbolic type if and only if* $|\mathrm{tr}(A)| > 2$.

Proof. We use the same notations introduced above. That is,

$$\gamma(z) = \frac{az+b}{cz+d}, \qquad a,b,c,d \in \mathbb{R}\,, \quad ad-bc = 1\,;$$

and $z \in \overline{\mathbb{H}^2}$ is a fixed point of γ in $\overline{\mathbb{H}^2}$ such that Equation (1) holds.

Let us begin with the case that $c \neq 0$. In this case, $|\,\mathrm{tr}(A)|^2 = 4 + \Delta$ where $\Delta = (d-a)^2 + 4bc$ is the discriminant of the quadratic equation (1). Since $c \neq 0$, any fixed point of γ in $\overline{\mathbb{H}^2}$ is a root of the quadratic equation (1). Now $\Delta < 0$ if and only if the equation has no real root, i.e., γ has a unique fixed point in $\mathbb{H}^2$. This shows statement (1). Next, $\Delta = 0$ if and only if the equation has a a unique real root, i.e., γ has a unique fixed point in $\mathbb{R}$. This verifies statement (2). Finally, $\Delta > 0$ if and only if the equation has two distinct real roots, i.e., γ has two distinct fixed points in $\mathbb{R}$. This shows statement (3).

In the case that $c = 0$, we have $\Delta = \mathrm{tr}(A) - 4 = (d-a)^2 \geqslant 0$ and $ad = ad - bc = 1$. Now $|\mathrm{tr}(A)| = 2$ if and only if $a = d \neq 0$, in which case $\gamma(z) = z + \frac{b}{a}$ is parabolic. This shows statement (2). Finally, $|\mathrm{tr}(A)| > 2$ if and only if $a \neq d$, in which case $\gamma(z) = \frac{a}{d}z + \frac{b}{d}$ is hyperbolic. This verifies statement (3). □

From this proposition, it is easy to verify that $f_\lambda(z) = \lambda z$ with $0 < \lambda \in \mathbb{R} \setminus \{1\}$ is hyperbolic; $g_\mu(z) = z + \mu$ with $\mu \in \mathbb{R}$ is parabolic; and $h(z) = -1/z$ is elliptic.

Our next task is to find all the geodesics in the hyperbolic plane. A *geodesic* is a curve with the local distance minimizing property. We will adopt the terminology that a hyperbolic *geodesic line* is a geodesic in $\mathbb{H}^2$ that is isometric to $\mathbb{R}$. Also, a *geodesic ray* is a subset of a geodesic line isometric to $[0, \infty)$. Since isometries preserve geodesics, we will find one geodesic and obtain others as images of it under isometries. Here is a typical geodesic in the upper half plane model found by simple calculations.

Example 3.3. The positive y-axis $Y = \{\, \mathbf{i}y \mid y > 0 \,\}$ is a geodesic in $\mathbb{H}^2$.

Let $z(t)$, $t \in [0,1]$, parametrize a path from $\mathbf{i}a$ to $\mathbf{i}b$ in $\mathbb{H}^2$ where $b > a > 0$. Write $z(t) = x(t) + \mathbf{i}y(t)$, where $y(t) > 0$. Then $z'(t) = x'(t) + \mathbf{i}y'(t)$ and the length of the path is given by the integral

$$\int_0^1 \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)}\, dt \geq \int_0^1 \frac{\sqrt{y'(t)^2}}{y(t)} dt \geq \left| \int_0^1 \frac{y'(t)}{y(t)} dt \right| = \ln \frac{b}{a}\,.$$

Note that the above equalities hold if and only if $x(t) = 0$ and $y'(t) \geq 0$. That is the same as that $z(t) \in Y$ and $y(t)$ is monotonic increasing. Thus the positive y-axis is a geodesic. Moreover, one obtains explicitly the distance between $\mathbf{i}a$ and $\mathbf{i}b$, that is,

$$d_{\mathbb{H}^2}(\mathbf{i}a, \mathbf{i}b) = \ln\left(\frac{b}{a}\right)\,.$$

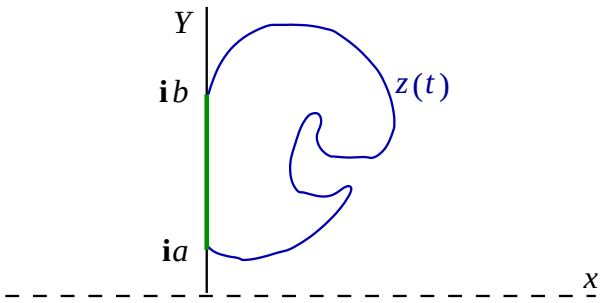

Figure 3.1

Definition 3.4. The *cross ratio* of four complex numbers $z_1, z_2, z_3, z_4 \in \mathbb{C}$ is defined to be $(z_1, z_2, z_3, z_4) :\overset{\text{def}}{=\!=} \dfrac{z_1 - z_3}{z_1 - z_4} \cdot \dfrac{z_2 - z_4}{z_2 - z_3}$.

The distance between $\mathbf{i}a$ and $\mathbf{i}b$ can be expressed in terms of cross ratio by

$$d_{\mathbb{H}^2}(\mathbf{i}a, \mathbf{i}b) = \ln(\mathbf{i}a, \mathbf{i}b, \infty, 0)\,.$$

Let $z \neq w \in \mathbb{H}^2$ and α be the geodesic line passing through z, w. Then there is a unique Möbius transformation $\varphi \in \mathrm{Isom}^+(\mathbb{H}^2)$ taking $\mathbf{i}a$, $\mathbf{i}b$, Y to z, w, α respectively. Let $\widetilde{z} = \varphi(0)$ and $\widetilde{w} = \varphi(\infty)$ in $\overline{\mathbb{R}}$ (see Figure 3.2 below). Then $\varphi(Y)$ is the geodesic in $\mathbb{H}^2$ ending at $\widetilde{z}$ and $\widetilde{w}$. More precisely, its closure in $\overline{\mathbb{H}^2}$ intersects $\overline{\mathbb{R}}$ at $\widetilde{z}$ and $\widetilde{w}$. Furthermore, since φ preserves the cross ratio,

$$(\star) \qquad d_{\mathbb{H}^2}(z, w) = \ln(z, w, \widetilde{w}, \widetilde{z}), \qquad z, w \in \mathbb{H}^2\,.$$

A direct calculation shows that $\mathrm{PSL}(2,\mathbb{R})$ acts transitively on $UT\mathbb{H}^2$, the unit tangent bundle of $\mathbb{H}^2$. Indeed, for all $z = x + \mathbf{i}y \in \mathbb{H}^2$, $\begin{pmatrix} \sqrt{y} & \frac{x}{\sqrt{y}} \\ 0 & \frac{1}{\sqrt{y}} \end{pmatrix}$ maps $\mathbf{i}$ to z, which shows that $\mathrm{PSL}(2,\mathbb{R})$ acts transitively on $\mathbb{H}^2$, and the subgroup $\left\{ \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \mid \theta \in (0, 2\pi] \right\}$ of $\mathrm{PSL}(2,\mathbb{R})$ acts transitively on the unit tangent vectors at $\mathbf{i}$ as rotations. As a consequence, we have the following corollary whose proof is left as an exercise.

Corollary 3.5. *Each geodesic line in $\mathbb{H}^2$ is either a vertical line or a semicircle perpendicular to the x-axis.*

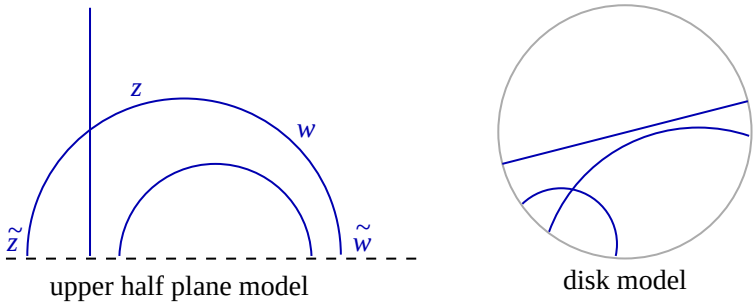

Figure 3.2

Remark. There is an analogous discussion using the disk model. One simply replaces $\mathbb{H}^2$ by $\mathbb{D}^2$ and $\mathrm{PSL}(2,\mathbb{R})$ by the isometry group

$$\left\{ z \mapsto e^{\mathbf{i}\theta} \left(\frac{z-a}{\overline{a}z - 1} \right) \mid a \in \mathbb{C},\ |a| < 1 \right\}.$$

In our pictures, we often use disk model whenever it is more illustrative.

A geodesic line α in $\mathbb{H}^2$ is determined by its end points x, y in $\partial\mathbb{H}^2$. For instance, the end points of the positive y-axis Y are $0, \infty$.

Definition 3.6. Let $\gamma \in \mathrm{Isom}^+(\mathbb{H}^2)$ be of hyperbolic type. The *axis* of γ, denoted by $\mathrm{Axis}(\gamma)$, is the geodesic line with the two ends equal to the fixed points of γ.

By definition, we have $\mathrm{Axis}(\sigma\gamma\sigma^{-1}) = \sigma(\mathrm{Axis}(\gamma))$ for $\sigma \in \mathrm{Isom}^+(\mathbb{H}^2)$. Since γ sends $\mathrm{Axis}(\gamma)$ to a geodesic line with the same end points, $\gamma(\mathrm{Axis}(\gamma)) = \mathrm{Axis}(\gamma)$. On the other hand, if $\mathrm{Axis}(\gamma_1) = \mathrm{Axis}(\gamma_2)$, then γ_1 and γ_2 have the same fixed

points. After a conjugation, one may assume their fixed points to be 0 and ∞. It is then easy to show that there exists $\gamma \in \mathrm{Isom}^+(\mathbb{H}^2)$ of hyperbolic type and $m, n \in \mathbb{Z}$ such that $\gamma_1 = \gamma^m$ and $\gamma_2 = \gamma^n$.

Next, we will describe the δ-neighborhood of a geodesic line s, namely, the set

$$N_\delta(s) = \left\{ z \in \mathbb{H}^2 \ : \ d_{\mathbb{H}^2}(z, s) < \delta \right\}, \qquad \delta > 0.$$

Let us begin with the following example.

Example 3.7. Let Y be the positive y-axis, then the set $N_\delta(Y)$ is bounded by two straight lines from 0 to ∞ making the same angle θ with Y at 0, namely

$$N_\delta(Y) = \{ x + \mathbf{i}y \ : \ y > |x| \cot(\theta) \}, \quad \text{where} \quad \tanh\left(\frac{\delta}{2}\right) = \tan\left(\frac{\theta}{2}\right).$$

Indeed, let $z \in N_\delta(Y)$, then $d_{\mathbb{H}^2}(z, Y) < \delta$. Since for each $0 < \lambda \in \mathbb{R}$, the hyperbolic isometry f_λ leaves Y invariant, we have

$$d_{\mathbb{H}^2}(\lambda z, Y) = d_{\mathbb{H}^2}(\lambda z, f_\lambda(Y)) = d_{\mathbb{H}^2}(z, Y) < \delta\,.$$

Therefore, if $z \in N_\delta(Y)$, the whole straight line $L_z = \{ \lambda z \ : \ 0 < \lambda \in \mathbb{R} \} \subset N_\delta(Y)$. To see that $N_\delta(Y)$ is symmetric about Y, it suffices to observe that

$$d_{\mathbb{H}^2}(x + \mathbf{i}y, Y) = d_{\mathbb{H}^2}(-x + \mathbf{i}y, Y)\,, \qquad x \in \mathbb{R}, y > 0\,.$$

This shows that $N_\delta(Y)$ is a sector at the origin.

To prove the formula for θ, let $z_0 = e^{\mathbf{i}\left(\frac{\pi}{2} - \theta\right)}$ be a point on the boundary of $N_\delta(Y)$. By the formula $(\star)$ of hyperbolic distance in terms of cross ratio above, we have

$$\delta = d_{\mathbb{H}^2}(\mathbf{i}, z_0) = \ln(\mathbf{i}, z_0, 1, -1) = \ln \frac{\mathbf{i} - 1}{\mathbf{i} + 1} \cdot \frac{z_0 + 1}{z_0 - 1} = \ln\left(\frac{-1 + \mathbf{i}e^{\mathbf{i}\theta}}{\mathbf{i} - e^{\mathbf{i}\theta}}\right).$$

In other words, $e^\delta = \dfrac{-1 + \mathbf{i}e^{\mathbf{i}\theta}}{\mathbf{i} - e^{\mathbf{i}\theta}}$. Consequently,

$$\tanh\left(\frac{\delta}{2}\right) = \frac{e^\delta - 1}{e^\delta - 1} = \frac{\dfrac{-1 + \mathbf{i}e^{\mathbf{i}\theta}}{\mathbf{i} - e^{\mathbf{i}\theta}} - 1}{\dfrac{-1 + \mathbf{i}e^{\mathbf{i}\theta}}{\mathbf{i} - e^{\mathbf{i}\theta}} + 1} = \tan\left(\frac{\theta}{2}\right).$$

Using an isometry in $\mathrm{Isom}^+(\mathbb{H}^2)$, we obtain the δ-neighborhood of a general geodesic line in $\mathbb{H}^2$ (Figure 3.3 below).

Corollary 3.8. *Let s be a geodesic line in $\mathbb{H}^2$ ending at $\{ z_0, z_\infty \}$ in $\overline{\mathbb{R}}$. Then for each $\delta > 0$, $\partial N_\delta(s)$ consists of two circular arcs passing through z_0 and z_∞ so that the angle between each of the arcs and s is equal to θ where*

$$\tanh\left(\frac{\delta}{2}\right) = \tan\left(\frac{\theta}{2}\right) .$$

In particular,

$$\overline{N_\delta(s)} \cap \partial\overline{\mathbb{H}^2} = \overline{s} \cap \partial\overline{\mathbb{H}^2} .$$

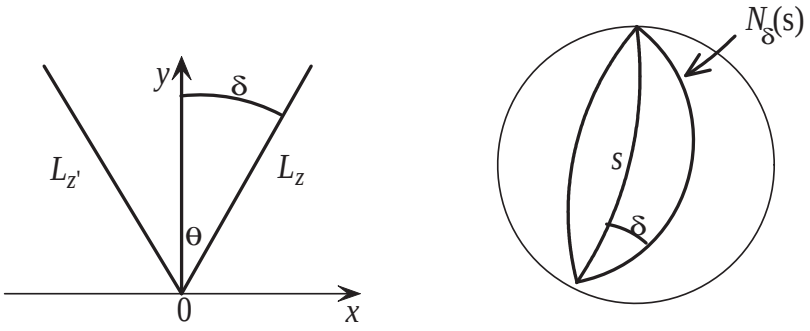

Figure 3.3

3.2. **Hyperbolic Geometry on Surfaces.** A *hyperbolic surface* Σ is a smooth surface together with a complete Riemannian metric of constant curvature -1. If the surface Σ is closed, we may equivalently define a hyperbolic structure as a collection of smooth charts $\{(U_i, \phi_i) \mid \phi_i(U_i) \subset \mathbb{H}^2\}$ so that $\Sigma = \bigcup_i U_i$ and each transition function $\phi_i \circ \phi_j^{-1}$ on $\phi_j(U_i \cap U_j)$ is a restriction of an isometry in $\mathrm{Isom}^+(\mathbb{H}^2)$. See [BP] for more details.

Proposition 3.9. *Every closed orientable surface $\Sigma_{g,0}$ with $g \geqslant 2$ has a hyperbolic structure.*

Proof. The surface $\Sigma_{g,0}$ is the quotient space of a regular Euclidean $4g$-gon by identifying pairs of opposite sides by Euclidean translations. If the sides are labeled cyclically as $a_1, \ldots, a_{2g}$, the algebraic relation on the boundary is given by

$$a_1 a_2 \cdots a_{2g}\, a_1^{-1} a_2^{-1} \cdots a_{2g}^{-1} .$$

Note that all the $4g$ vertices of the polygon are identified.

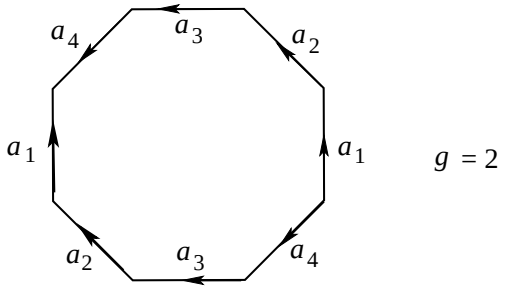

Figure 3.4

To construct a hyperbolic structure on $\Sigma_{g,0}$, consider the set of all regular hyperbolic $4g$-gons in $\mathbb{H}^2$. Such a polygon is determined up to isometry by its edge length, denoted by t. Let $\alpha(t)$ be its inner angle. Then we have

$$\lim_{t\to+\infty} \alpha(t) = 0 \qquad \text{and} \qquad \lim_{t\to 0} \alpha(t) = \frac{(4g-2)\pi}{4g}$$

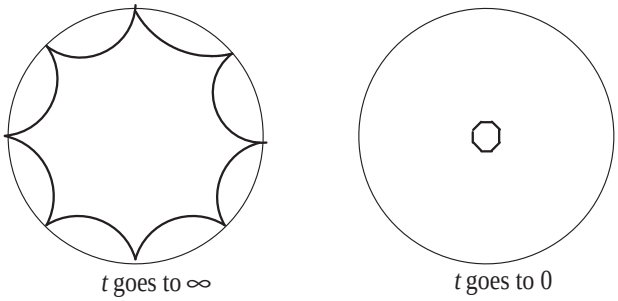

Figure 3.5

Indeed, the first equation comes from the fact that when $t \to +\infty$, the adjacent edges of the polygon are tangent to each other at the circle of infinity. The second equation comes from the fact that when $t \to 0$, the polygon becomes asymptotically Euclidean, and $\frac{(4g-2)\pi}{4g}$ is the inner angle of the regular Euclidean $4g$-gon. Here we have used the fact that the notion of angles in the disk model of the hyperbolic plane coincides with that of the Euclidean space.

Since $\alpha(t)$ is a continuous function of t according to the cosine law, there exists a t_0 such that $\alpha(t_0) = \frac{2\pi}{4g}$. Take this regular hyperbolic $4g$-gon with inner angle $\frac{2\pi}{4g}$ and identify opposite sides by hyperbolic isometries. Then the quotient space $\Sigma_{g,0}$ has the induced hyperbolic structure. This is due to the Poincaré Polyhedron

Theorem, see Maskit, [Ma], or Beardon, [Be]. Note that the key condition that the sum of angles at all the vertices in the polygon is 2π is satisfied by the choice of $2\pi/4g$. □

By Proposition 3.9, we can identify the universal cover of the closed surface $\Sigma_{g,0}$ with the hyperbolic plane $\mathbb{H}^2$. Let

$$\Theta : \mathbb{H}^2 \to \Sigma_g = \Sigma_{g,0}\,.$$

be the covering projection. Let $p_0 \in \Sigma_g$ and choose $z_0 \in \mathbb{H}^2$ such that $\Theta(z_0) = p_0$. Then the fundamental group $\pi_1(\Sigma_g, p_0)$ can be identified with the deck transformation group G of the universal cover.

For each $\gamma \in G$, let α be the geodesic segment from z_0 to $\gamma(z_0)$. Then $a_\gamma = \Theta(\alpha)$ is a loop based at p_0 in Σ_g. The map $\Phi : G \to \pi_1(\Sigma_g, p_0)$ having $\Phi(\gamma) = [a_\gamma]$ is an isomorphism between the groups. For any other point z in $\mathbb{H}^2$, let s be a path from z to $\gamma(z)$ in $\mathbb{H}$. Then the quotient of s is a loop $\Theta(s)$ in the surface Σ_g freely homotopic to a_γ. It is known that the free homotopy class of a loop in Σ_g corresponds to a conjugacy class of an element in $\pi_1(\Sigma_g, p_0)$.

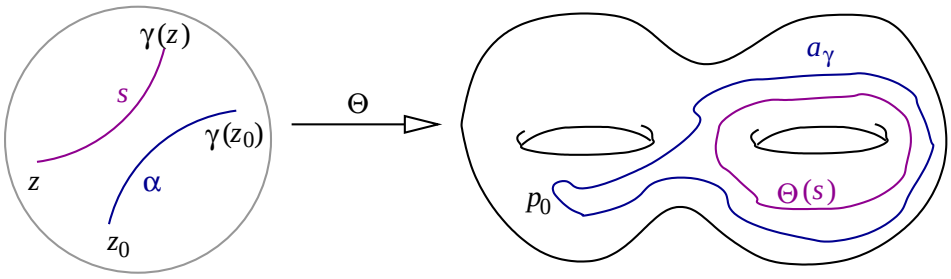

Figure 3.6

For this reason, we will take $G = \pi_1(\Sigma_g, p_0)$ and consider an element of it as either an isometry on $\mathbb{H}^2$ or a homotopy class of a loop at p_0 in Σ_g. For $z \in \mathbb{H}^2$, the point $\Theta(z) \in \Sigma_g$ is also denoted by $[z]$. In the rest of the paper, we will always equip Σ_g with a fixed hyperbolic structure. Since the action of G on $\mathbb{H}^2$ is free, there is no elliptic element in G.

Recall that the *injectivity radius* $\epsilon(q)$ at $q \in \Sigma_g$ is the supremum over all positive real numbers $\delta > 0$ such that the δ-disk at q is isometric to a δ-disk in $\mathbb{H}^2$. The *injectivity radius* $\epsilon(\Sigma_g)$ of Σ_g is defined by

$$\epsilon(\Sigma_g) :\stackrel{\text{def}}{=\!=} \inf_{q \in \Sigma_g} \epsilon(q)\,.$$

Lemma 3.10. *Under the above identification $\Sigma_g = \mathbb{H}^2/G$ where $G = \pi_1(\Sigma_g, p_0)$, we have the following:*

(1) *Each $\gamma \in G - \{\,\mathrm{id}\,\}$ is an isometry of hyperbolic type.*
(2) *For any $\gamma \in G - \{\,\mathrm{id}\,\}$, the projection of $\mathrm{Axis}(\gamma)$ onto Σ_g is the unique closed geodesic in the conjugacy class of $\gamma \in \pi_1(\Sigma_g, p_0)$.*

Proof. Let ϵ be the injective radius of Σ_g. Since Σ_g is compact, $\epsilon(\Sigma_g) > 0$. Then, by definition, any loop in Σ_g of length less than $\epsilon/2$ is null homotopic.

For statement (1), suppose that there is a parabolic element $\gamma \in G \setminus \{\,\mathrm{id}\,\}$. By the classification of isometries, one may assume that $\gamma(z) = z + a$ for $z \in \overline{\mathbb{R}}$, where $a \in \mathbb{R}$, after taking conjugation. By a direct calculation, $d_{\mathbb{H}^2}(z, z+a) = \frac{|a|}{\mathrm{Im}(z)}$. So, there exists a point $z \in \mathbb{H}^2$, with sufficiently large $\mathrm{Im}(z)$, such that

$$d_{\mathbb{H}^2}(z, \gamma(z)) < \frac{\epsilon}{2}\,.$$

This shows that the projection of the geodesic segment from z to $\gamma(z)$ is a loop ρ in Σ_g of length at most $\epsilon/2$ based at the point $[z]$. By the construction, ρ is freely homotopic to a representative of γ, and so γ must be trivial, which contradicts that $\gamma \neq \mathrm{id}$. Since γ cannot be an elliptic element, the first statement is established.

For statement (2), let $a = \Theta(\mathrm{Axis}(\gamma))$, then it is a geodesic in the conjugacy class of γ. Let b be another closed geodesic freely homotopic to a and $F : \mathbb{S}^1 \times [0,1] \to \Sigma_g$ be a smooth homotopy between a and b. Define

$$\widetilde{F} : \mathbb{R} \times [0,1] \to \Sigma_g \quad \text{by} \quad \widetilde{F}(s,t) = F(e^{\mathbf{i}s}, t)\,.$$

Then $\widetilde{F}$ can be lifted to $F^* : \mathbb{R} \times [0,1] \to \mathbb{H}$ so that $F^*(\mathbb{R}, 0)$ and $F^*(\mathbb{R}, 1)$ are two geodesic lines satisfying $\Theta \circ F^*(\mathbb{R}, 0) = a$ and $\Theta \circ F^*(\mathbb{R}, 1) = b$ respectively.

Since F is smooth, there is $c > 0$ such that the hyperbolic length of any curve $F(e^{is}, t)$, $t \in [0,1]$, is bounded by c for each choice of s. This shows that $d_{\mathbb{H}}(F^*(s,0), F^*(s,1)) \leqslant c$ for $s \in \mathbb{R}$. Thus, these two geodesic lines, $F^*(\mathbb{R}, 0)$ and $F^*(\mathbb{R}, 1)$, are within bounded distance. Thus, they must coincide. Hence a=b and uniqueness is established. □

Let $G = \pi_1(\Sigma_g, p_0)$ and $\gamma \in G - \{\,\mathrm{id}\,\}$, we will call the geodesic $\Theta(\mathrm{Axis}(\gamma)) \subset \Sigma_g$ the *geodesic representative* of γ. Note that if $\tau \in G$, then $\tau\gamma\tau^{-1}$ and γ have the same geodesic representative since $\mathrm{Axis}(\tau\gamma\tau^{-1}) = \tau(\mathrm{Axis}(\gamma))$ and $\Theta = \Theta \circ \tau$.

Recall that a loop $\rho : \mathbb{S}^1 \to \Sigma_g$ is called simple if ρ is an embedding.

Definition 3.11. An element $\gamma \in G - \{\,\mathrm{id}\,\}$ is call *simple* if its geodesic representative is an embedded circle in Σ_g. Two elements $\gamma_1, \gamma_2 \in G - \{\,\mathrm{id}\,\}$ are called *disjoint* if their geodesic representatives are disjoint. Two elements $\gamma_1, \gamma_2 \in G - \{\,\mathrm{id}\,\}$ are called *linked* if $\mathrm{Axis}(\gamma_1) \cap \mathrm{Axis}(\gamma_2) \neq \emptyset$ and $\mathrm{Axis}(\gamma_1) \neq \mathrm{Axis}(\gamma_2)$).

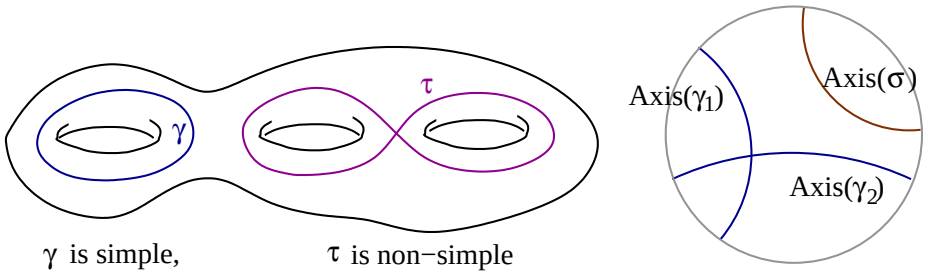

Figure 3.7

For $\gamma_1, \gamma_2 \in G - \{\mathrm{id}\}$ such that γ_1 is not conjugate to γ_2, the geodesic intersection number $I(\gamma_1, \gamma_2)$ is the number of intersection points between the closed geodesic representatives of γ_1 and γ_2.

Lemma 3.12. *Let the deck transformation group G be identified with $\pi_1(\Sigma_g, p_0)$ as before. Then*

(1) *An element* $\gamma \in G - \{\,\mathrm{id}\,\}$ *is simple if and only if*

$$g(\mathrm{Axis}(\gamma)) \cap \mathrm{Axis}(\gamma) = \emptyset \quad \textit{for all} \quad g \in G - \{\gamma^n \mid n \in \mathbb{Z}\}\,.$$

(2) *If there exists a simple loop* ρ *freely homotopic to a representative in* $\gamma \in G - \{\,\mathrm{id}\,\}$*, then* γ *is simple.*

(3) *Two elements* $\gamma_1, \gamma_2 \in G - \{\,\mathrm{id}\,\}$ *are disjoint if and only if*

$$G(\mathrm{Axis}(\gamma_1)) \cap G(\mathrm{Axis}(\gamma_2)) = \emptyset\,.$$

(4) *Suppose* $\gamma_1, \gamma_2 \in G - \{\,\mathrm{id}\,\}$ *are simple such that* $\mathrm{Axis}(\gamma_1) \cap \mathrm{Axis}(\gamma_2) \neq \emptyset$*. Then* $I(\gamma_1, \gamma_2) = 1$ *if and only if for all* $\sigma \in G - \{\gamma_2^n \mid n \in \mathbb{Z}\}$*,*

$$\sigma(\mathrm{Axis}(\gamma_1)) \cap \mathrm{Axis}(\gamma_2) = \emptyset\,.$$

Proof. To see (1), suppose that there is $g \in G - \{\gamma^n \mid n \in \mathbb{Z}\}$ and there exists $z \in g(\mathrm{Axis}(\gamma)) \cap \mathrm{Axis}(\gamma) \neq \emptyset$. Since $g \neq \gamma^n$,

$$\mathrm{Axis}(g\gamma g^{-1}) = g(\mathrm{Axis}(\gamma)) \neq \mathrm{Axis}(\gamma)\,.$$

Therefore, the fixed point sets of $g\gamma g^{-1}$ and γ separate each other in $\overline{\mathbb{R}}$ and so the curves $g(\mathrm{Axis}(\gamma))$ and $\mathrm{Axis}(\gamma)$ intersect transversally in $\mathbb{H}^2$. These two curves are projected to the same geodesic, $\Theta(\mathrm{Axis}(\gamma)) = \Theta(g\,\mathrm{Axis}(\gamma)) \subset \Sigma_g$. Then, $\Theta(z) \in \Sigma_g$ is a transversal self intersection of it; equivalently, γ is not simple.

Conversely, suppose that the quotient $\Theta(\mathrm{Axis}(\gamma))$ intersects itself at $q \in \Sigma_g$. Let $w \in \mathbb{H}^2$ such that $\Theta(w) = q$. Consider the lifting of $\Theta(\mathrm{Axis}(\gamma))$ in a neighborhood of w. Since Θ is a local isometry, there are two distinct geodesic lines $\tilde{s_i}$, $i = 1, 2$ intersecting at w, which are projected to $\Theta(\mathrm{Axis}(\gamma))$. Thus, there are $\sigma_i \in G$ such that $\tilde{s_i} = \sigma_i(\mathrm{Axis}(\gamma))$. Assume that $\tilde{s_1} \neq \mathrm{Axis}(\gamma)$. Then $\sigma_1 \notin \{\,\gamma^n \mid n \in \mathbb{Z}\,\}$. It follows that the orbit of $\mathrm{Axis}(\gamma)$ under the action of G is not pairwise disjoint. This completes the proof of (1).

For statement (2), suppose otherwise that γ is not simple. By (1), there are $\gamma_1, \gamma_2 \in G$ such that $\gamma_1\gamma_2^{-1} \notin \{\,\gamma^n \mid n \in \mathbb{Z}\,\}$ and

$$\gamma_1(\mathrm{Axis}(\gamma)) \cap \gamma_2(\mathrm{Axis}(\gamma)) \neq \emptyset\,.$$

Let ρ be a simple loop freely homotopic to $\Theta(\mathrm{Axis}(\gamma))$ and $\tilde{\rho}$ be a lifting of ρ. Since $\tilde{\rho}$ is compact, it is of bounded distance to $\mathrm{Axis}(\gamma)$. Further let $\tilde{\rho}_i = \gamma_i(\tilde{\rho})$ for $i = 1, 2$. We have $\tilde{\rho}_i \subset N_c(\gamma_i(\mathrm{Axis}(\gamma)))$, $i = 1, 2$, for some constant $c > 0$. It follows that $\tilde{\rho}_1 \cap \tilde{\rho}_2 \neq \emptyset$. In addition, $\Theta(\tilde{\rho}_i) = \rho$ for $i = 1, 2$; so ρ is not simple, which contradicts the assumption.

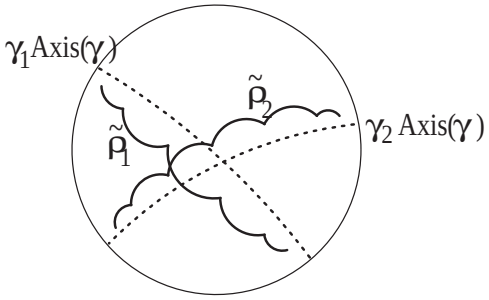

Figure 3.8

To see (3), note that γ_1, γ_2 are disjoint if and only if $\Theta(\mathrm{Axis}(\gamma_1)) \cap \Theta(\mathrm{Axis}(\gamma_2)$ is empty. The latter condition is the same as that $G(\mathrm{Axis}(\gamma_1)) \cap G(\mathrm{Axis}(\gamma_2)) = \emptyset$.

Finally, for (4), let $z_1 \in \mathrm{Axis}(\gamma_1) \cap \mathrm{Axis}(\gamma_2) \neq \emptyset$. Then $[z_1] = \Theta(z_1) \in \Sigma_g$ is an intersection point of the geodesic representatives of γ_1 and γ_2. For $i = 1, 2$, $\Theta^{-1}\{[z_1]\} \cap \mathrm{Axis}(\gamma_i) = G(z_1) \cap \mathrm{Axis}(\gamma_i)$ where $G(z_1)$ is the G-orbit of z_1. Since both γ_i are simple, by (1), $G(z_1) \cap \mathrm{Axis}(\gamma_i) = \{\gamma_i^n(z_1) \mid n \in \mathbb{Z}\}$ for $i = 1, 2$.

Suppose that $I(\gamma_1, \gamma_2) = 1$, that is, the geodesic representatives of γ_1 and γ_2 intersect only at the point $[z_1]$. If $w \in G(\mathrm{Axis}(\gamma_1)) \cap \mathrm{Axis}(\gamma_2)$, then $\Theta(w) = [z_1]$ and so $w \in G(z_1)$. It follows that $w \in G(z_1) \cap \mathrm{Axis}(\gamma_2)$. From the discussion in the previous paragraph, $w = \gamma_2^n(z_1)$ for some $n \in \mathbb{Z}$.

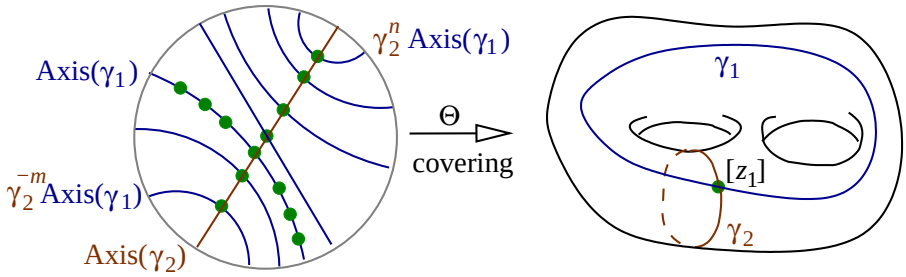

Figure 3.9

Hence, if $I(\gamma_1, \gamma_2) = 1$, then $G(\text{Axis}(\gamma_1)) \cap \text{Axis}(\gamma_2) \subset \{\gamma_2^n(z_1) \mid n \in \mathbb{Z}\}$. This shows that for $\sigma \in G - \{\gamma_2^n \mid n \in \mathbb{Z}\}$

$$\sigma(\text{Axis}(\gamma_1)) \cap \text{Axis}(\gamma_2) = \emptyset.$$

Conversely, suppose that there exists $\sigma \in G - \{\gamma_2^n \mid n \in \mathbb{Z}\}$ and $z_2 \in \mathbb{H}^2$ such that

$$z_2 \in \sigma(\text{Axis}(\gamma_1)) \cap \text{Axis}(\gamma_2)\,.$$

We will show that $I(\gamma_1, \gamma_2) \geqslant 2$ by proving that $[z_1] \neq [z_2]$. Assume otherwise that $[z_1] = [z_2]$, then $z_2 = g(z_1)$ for some $g \in G$.

Now, both z_1 and z_2 are in $\text{Axis}(\gamma_2)$. Thus, $z_2 = g(z_1) \in G(z_1) \cap \text{Axis}(\gamma)$. Since γ_2 is simple, by (1), we have $z_2 = \gamma_2^m(z_1)$ for some $m \in \mathbb{Z}$.

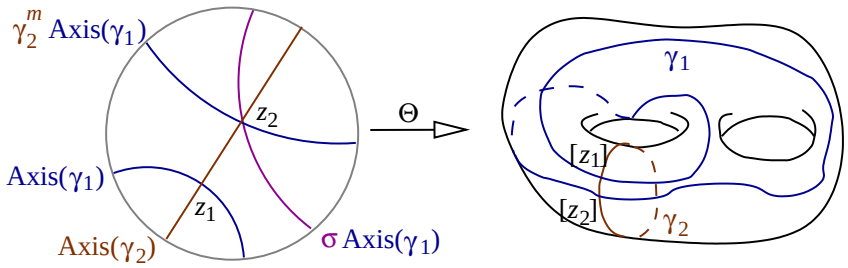

Figure 3.10

On the other hand, it has been assumed that $z_2 \in \sigma(\text{Axis}(\gamma_1))$ where $\sigma \notin \{\gamma_2^n \mid n \in \mathbb{Z}\}$. Thus, $\sigma(\text{Axis}(\gamma_1)) \neq \gamma_2^m(\text{Axis}(\gamma_1))$. However, they are projected to the same geodesic representative $\Theta(\text{Axis}(\gamma_1))$ with a self intersection point $[z_2]$. This contradicts that γ_1 is simple. □

4. Quasi-Isometry and Large Scale Geometry

In this section, we study large scale geometry, in particular, the properties of a quasi-isometry on the hyperbolic plane. The main result of this section is the result of Milnor that relates the isometric actions of the fundamental group with the quasi-isometric geometry of the universal cover. These notions constitute the key idea of Farb-Margalit's proof of Dehn-Nielsen theorem, [FM].

Definition 4.1. Suppose (X, d_X), (Y, d_Y) are two metric spaces and $F : X \to Y$ is a map which may not be continuous. We say F is a quasi-isometry (or more precisely K-quasi-isometry) if there exists a constant $K > 0$ such that

(1) for all $x_1, x_2 \in X$,

$$\frac{1}{K} d_X(x_1, x_2) - K \leqslant d_Y(f(x_1), f(x_2)) \leqslant K d_X(x_1, x_2) + K\,; \qquad \text{and}$$

(2) $Y = N_K(f(X)) \doteq \{\, y \in Y \mid d_Y(y, f(x)) \leqslant K \text{ for some } x \in X \,\}$.

It follows directly from the definition that compositions of quasi-isometries are quasi-isometries. The followings are some examples of quasi-isometries.

Example 4.2. Let $X \subset Y$ be a subset such that

$$N_K(X) = \{\, y \in Y \mid d_Y(x, y) \leqslant K \text{ for some } x \in X \,\}$$

and $d_X(x_1, x_2) = d_Y(i(x_1), i(x_2))$ where $i : X \to Y$ is the inclusion map. Then the inclusion map i is a quasi-isometry. For specific examples, we may take $X = \mathbb{Z}$ and $Y = \mathbb{R}$, or more generally $X = \mathbb{Z}^n$ and $Y = \mathbb{R}^n$ in the standard Euclidean metrics.

Example 4.3. Suppose (M^n, g) is a compact Riemannian manifold with $p_0 \in M$. Moreover, suppose $Y = (\tilde{M}^n, \tilde{g})$ is the universal cover of M with $y_0 \in Y$ projected to p_0 so that the covering map is a local isometry. Let $X \subset Y$ be the orbit of a point $y \in Y$ under the isometric action of the deck transformation group $\pi_1(M, p_0)$. Then, by construction, $N_D(X) = Y$ where D is the diameter of M. Thus the inclusion map $i : X \to Y$ is a quasi-isometry. Note that X is a bijective image of the fundamental group $\pi_1(M, p_0)$ via the action map $\gamma \mapsto \gamma(y) \ : \ \pi_1(M, p_0) \to X$.

Our main interests are in the quasi-isometry classes of a finitely generated group G. Recall that a *symmetric generating set* S for G is a set which generates G and it satisfies that if $x \in S$, then $x^{-1} \in S$.

Definition 4.4. (Cayley Graph) Suppose G is a finitely generated group with a symmetric finite generating set S. The *Cayley graph* $\Gamma(G, S)$ is the graph of which the vertex set is G and two vertices $g_1, g_2 \in G$ are joined by an edge if $g_1^{-1} g_2 \in S$, i.e., $g_2 = g_1 s$ for some $s \in S$. A metric is given to $\Gamma(G, S)$ such that each edge has length one. Consequently, the distance between two vertices g_1, g_2 of $\Gamma(G, S)$ is the minimum length of all possible edge paths from g_1 to g_2.

Remark. For any edge between g_1, g_2 and any $g \in G$, a unique edge is determined by gg_1, gg_2. Thus, the left action of G on $\Gamma(G, S)$ is an isometric action.

Example 4.5. Let $G = \mathbb{Z}$, $S = \{ \pm 1 \}$, then $\Gamma(G, S)$ is isometric to $\mathbb{R}^1$ with the standard metric.

Example 4.6. Let $G = \mathbb{Z}^2$, $S = \{ \pm(1,0), \pm(0,1) \}$, then $\Gamma(G, S)$ is isometric to the standard lattice grid in the plane.

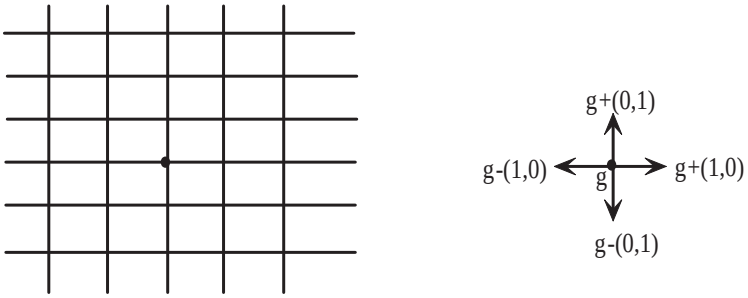

Figure 4.1

For a finitely generated group G with a symmetric generating set S, the *word length*, $|w|_S$ of an element $w \in G$ is the distance from w to the identity element id in the Cayley graph $\Gamma(G, S)$, namely,

$$|w|_S = \min \{ k \in \mathbb{N} \mid w = s_1 s_2 \cdots s_k, \text{ for some } s_1, \ldots, s_k \in S \} .$$

For instance, let $\mathbb{F}_2$ denote the free group on two generators x, y and let $S = \{ x, x^{-1}, y, y^{-1} \}$. Then we have $\left| x^3 y^{-1} x y^{-2} x^{-5} \right|_S = 12$ in $\Gamma(\mathbb{F}_2, S)$.

Example 4.7. If T is another symmetric generating set for the group G, then for any $w \in G$

$$|w|_S \leqslant K\,|w|_T\ , \tag{2}$$

where $K = \max\{\,|t|_S \mid t \in T\,\}$.

Indeed, suppose that $|w|_T = N$ and $w = t_1 \cdots t_N$ where $t_\ell \in T$. Now each t_ℓ can be in turns written in terms of S, namely,

$$t_\ell = s_{i_\ell,1} \cdots s_{i_\ell,n_\ell} \qquad \text{for } s_{i_\ell,j} \in S\,,$$

where $n_\ell = |t_\ell|_S \leqslant K$ for $\ell = 1, \ldots, N$. It follows that

$$w = s_{i_1,1} \cdots s_{i_1,n_1} \cdot s_{i_2,1} \cdots s_{i_2,n_2} \cdot \cdots\cdots \cdot s_{i_N,1} \cdots s_{i_N,n_N}, \qquad \text{where } s_{i_\ell,j} \in S.$$

Thus $|w|_S \leqslant \sum_{i=1}^{N} n_i \leqslant \sum_{i=1}^{N} K \leqslant K\,|w|_T$.

Lemma 4.8. *Given any two symmetric generating sets S and T of a group G, their Cayley graphs $\Gamma(G,S)$ and $\Gamma(G,T)$ are quasi-isometric.*

Proof. Denote d_S and d_T the metrics on $\Gamma(G,S)$ and $\Gamma(G,T)$ respectively. Note that from Example 4.7 above, the identity map from (G, d_S) to (G, d_T) is a quasi-isometry. Moreover, from Example 4.7, the inclusion $(G, d_T) \hookrightarrow \Gamma(G,T)$ is a quasi-isometry. Since a composition of quasi-isometries is still a quasi-isometry, it suffices to establish a quasi-isometry $\mathfrak{p} : \Gamma(G,S) \to (G, d_S)$ such that $\mathfrak{p}|_G$ is the identity map on G. Indeed, for $x \in \Gamma(G,S)$, simply take $\mathfrak{p}(x) = x$ if $x \in G$ and $\mathfrak{p}(x)$ to be any one of the two vertices of the edge containing x. Then $\mathfrak{p}$ is clearly a quasi-isometry. □

Due to Lemma 4.8, we see that the quasi-isometry class of the Cayley graph $\Gamma(G,S)$ is independent of the choice of the finite symmetric generating set. From now on we may discuss if a map $\phi : G \to G$ is a quasi-isometric map in the sense that G is given the metric with respect to any finite symmetric generating set.

Corollary 4.9. *If G is a finitely generated group and $\phi : G \to G$ is an isomorphism, then ϕ is a quasi-isometric map.*

Proof. Indeed, the map ϕ is a composition of $f : \Gamma(G,S) \to \Gamma(G,\phi(S))$ and $g : \Gamma(G,\phi(S)) \stackrel{\text{id}}{\to} \Gamma(G,S)$. Here f is a graph isomorphism and $g|_G = \text{id}_G$ is the identity map on G. By definition f is an isometry and by Lemma 4.8, g is a quasi-isometry. Thus the corollary follows. □

Theorem 4.10. (Milnor-Svarc) *Suppose (M,g) is a closed Riemannian manifold with the universal cover $(\tilde{M},\tilde{g})$ such that $G = \pi_1(M,p_0)$ acts isometrically on $\tilde{M}$ as the group of deck transformations. Then G is quasi-isometric to $\tilde{M}$ via the map sending γ to $\gamma(z)$ where $z \in \tilde{M}$ is a chosen point.*

Proof. Let Ω be a connected compact fundamental domain of the G action on $\tilde{M}$ and $K_1 = \text{diam}(\Omega)$. Define $S = \{\, \gamma \in G \mid \gamma(\Omega) \cap \Omega \neq \emptyset \,\}$. Note that S is a symmetric finite set since $\gamma(\Omega)\cap\Omega \neq \emptyset$ is equivalent to $\Omega \cap \gamma^{-1}(\Omega) \neq \emptyset$. Furthermore, it is known that G is generated by S. We choose the base point z to be in the interior of Ω and consider the map $P : G \to \tilde{M}$ given by

$$P(\gamma) = \gamma(z).$$

First of all, by definition of the diameter K_1, one has $N_{K_1}(P(G)) = \tilde{M}$.

Since the left action of G on both $(\Gamma(G,S), d_S)$ and $(\tilde{M},d)$ are isometries, it suffices to verify the quasi-isometry condition that there exists a constant K such that for each $\gamma \in G$,

$$\frac{1}{K} d_S(1,\gamma) - K \leqslant d(P(1),P(\gamma)) \leqslant K\, d_S(1,\gamma) + K\,, \tag{3}$$

Indeed, we have $d_S(\gamma_1,\gamma_2) = d_S(1,\gamma_1^{-1}\gamma_2)$ and

$$d(P(\gamma_1),P(\gamma_2)) = d(\gamma_1(z),\gamma_2(z)) = d(z,\gamma_1^{-1}\gamma_2(z)) = d(P(1),P(\gamma_1^{-1}\gamma_2))\,.$$

Now, we claim that the second inequality in (3) holds, namely,

$$d(P(1),P(\gamma)) \leqslant K_2\, d_S(1,\gamma) + K_2 \qquad \text{where } K_2 = \max(2K_1,1). \tag{3a}$$

Indeed, let $d_S(1,\gamma) = n$ and write $\gamma = \gamma_1\gamma_2\cdots\gamma_n$, where $\gamma_i \in S$. Then

$$\begin{aligned} d(P(1),P(\gamma)) &= d(z,\gamma(z)) = d(z,\gamma_1\gamma_2\cdots\gamma_n(z)) \\ &\leqslant \sum_{i=1}^{n-1} d(\gamma_1\cdots\gamma_{i-1}(z),\gamma_1\cdots\gamma_i(z)) \\ &\leqslant \sum_{i=1}^{n-1} d(z,\gamma_i(z)) \leqslant \sum_{i=1}^{n-1} 2K_1 \\ &\leqslant 2K_1 n = 2K_1 d_S(1,\gamma) \leqslant K_2 d_S(1,\gamma). \end{aligned}$$

Here we have used the fact that z and $\gamma_i(z)$ lie in $\Omega \cup \gamma_i(\Omega)$ with $\Omega \cap \gamma_i(\Omega) \neq \emptyset$, and thus

$$d(z,\gamma_i(z)) \leqslant \operatorname{diam}(\Omega \cup \gamma_i(\Omega)) \leqslant 2K_1.$$

To see the first inequality in (3), that

$$d_S(1,\gamma) \leqslant K\, d(P(1),P(\gamma)) + K^2\,, \tag{3b}$$

for some K, we let

$$\delta = \frac{1}{2}\inf_{\xi\in G}\{\, d(\Omega,\xi(\Omega)) \mid \Omega\cap\xi(\Omega) = \emptyset\,\} > 0\,.$$

By the definition of δ, if $\xi \in G$ satisfies that $d(\xi(z),z) \leqslant \delta$, then $\xi \in S$. Choose the shortest geodesic segment L in $\tilde{M}$ joining z to $\gamma(z)$. Write

$$d(\gamma(z),z) = n\delta + \delta' \quad \text{where} \quad 0 \leqslant \delta' < \delta, \qquad n \in \mathbb{Z}_{\geqslant 0}\,.$$

Let $z_1 = z$, $z_2,\ldots,z_n$, $z_{n+1} = \gamma(z)$ be points along the geodesic L so that the distance from z_i to z along L is $(i-1)\cdot\delta$, $i = 1,\ldots,n$. In particular,

$$\begin{aligned} d(z_i,z_{i+1}) &= \delta, \qquad i \leqslant n-1, \quad \text{and} \\ d(z_n,z_{n+1}) &= d(z_n,\gamma(z)) = \delta' < \delta\,. \end{aligned}$$

For each such z_i along L, $z_i \in h_i(\Omega)$ for some $h_i \in G$. Now $d(z_i,z_{i+1}) \leqslant \delta$ implies that

$$d(\Omega, h_i^{-1}h_{i+1}(\Omega)) = d(h_i(\Omega),h_{i+1}(\Omega)) \leqslant \delta\,.$$

Thus, we have $\gamma_i = h_i^{-1}h_{i+1} \in S$. As a consequence, one obtains

$$h_{i+1} = h_i\cdot\gamma_i = \gamma_1\gamma_2\cdots\gamma_i \qquad \text{for each } i = 1,\ldots,n.$$

Now, by the construction, both $z_{n+1} = \gamma(z) \in \gamma(\Omega)$ and $z_{n+1} \in h_{n+1}(\Omega)$. Thus $\gamma(\Omega) = h_{n+1}(\Omega)$, that is,

$$\gamma = h_{n+1} = \gamma_1\gamma_2\cdots\gamma_n \qquad \text{where } \gamma_i \in S.$$

In particular, $|\gamma|_S \leqslant n$ according to definition. In other words,

$$d_S(1,\gamma) = |\gamma|_S \leqslant n = \left(\frac{1}{\delta}\right)\delta n \leqslant \frac{1}{\delta}\sum_{i=1}^{n} d(z_i, z_{i+1}) \leqslant \frac{1}{\delta}d(z,\gamma(z))\,.$$

This shows that $\delta\, d_S(1,\gamma) \leqslant d(P(1), P(\gamma)$ and Inequality (3b) follows.

In summary, the condition (3) for P being a quasi-isometry is satisfied by taking $K = \max(\frac{1}{\delta}, 1, K_2)$. □

We now come to the main technical result for proving Dehn-Nielsen Theorem.

Proposition 4.11. *Suppose $\phi : \pi_1(\Sigma_g, p_0) \to \pi_1(\Sigma_g, p_0)$ is an isomorphism. If $\alpha, \beta \in \pi_1(\Sigma_g, p_0) - \{\,\mathrm{id}\,\}$ satisfy $\mathrm{Axis}(\alpha) \cap \mathrm{Axis}(\beta) = \emptyset$, then*

$$\mathrm{Axis}(\phi(\alpha)) \cap \mathrm{Axis}(\phi(\beta)) = \emptyset\,.$$

We will proceed by assuming that $\mathrm{Axis}(\phi(\alpha)) \cap \mathrm{Axis}(\phi(\beta)) \neq \emptyset$ and derive a contradiction. Before doing so, we need to consider a quasi-isometry on $\mathbb{H}^2$ and establish a lemma.

Let $G = \pi_1(\Sigma_g, p_0)$. By Corollary 4.9, and Theorem 4.10, we may replace the isomorphism $\phi : G \to G$ by a K-quasi-isometric map $F : \mathbb{H}^2 \to \mathbb{H}^2$ such that the following diagram commutes up to quasi-isometry where P is the evaluation map defined by $P(\gamma) = \gamma(z_0)$ for a fixed chosen point $z_0 \in \mathbb{H}^2$.

$$\begin{array}{ccc} G & \xrightarrow{\ \phi\ } & G \\ {\scriptstyle P}\downarrow & & \downarrow{\scriptstyle P} \\ \mathbb{H}^2 & \xrightarrow{\ F\ } & \mathbb{H}^2 \end{array}$$

Lemma 4.12. *If $F : \mathbb{H}^2 \to \mathbb{H}^2$ is a K-quasi-isometry, then there exists a constant $c = c(K)$ such that for any geodesic line L, $F(L) \subset N_c(L')$ for some geodesic line L'.*

We omit the details of the proof of the lemma. See Lemma 3.43, page 51, in Kapovich's book [Kap] for a proof.

Proof of Proposition 4.11. By Lemma 4.12, we see that the quasi-isometry F sends the two geodesic lines $A = \mathrm{Axis}(\alpha)$ and $B = \mathrm{Axis}(\beta)$ into $N_c(A')$ and $N_c(B')$, where A', B' are two geodesic lines. Moreover, if $p_n, q_n \in \partial\mathbb{H}^2$ are sequences converging to the end points of A, then A' has end points $\lim F(p_n)$ and $\lim F(q_n)$. Since $F \circ P = P \circ \phi$, by taking $p_n = \alpha^n(z_0)$ and $q_n = \alpha^{-n}(z_0)$, one can see that the end points of A' are the same as those of $\mathrm{Axis}(\phi(\alpha))$. Thus, $A' = \mathrm{Axis}(\phi(\alpha))$ and similarly, $B' = \mathrm{Axis}(\phi(\beta))$.

Recall that we start with the assumption that $\mathrm{Axis}(\phi(\alpha)) \cap \mathrm{Axis}(\phi(\beta)) \neq \emptyset$. Thus, as in Figure 4.2, we have $A' \cap B' \neq \emptyset$ and the end points of A' and B' are respectively denoted $\{a_1, a_2\}$ and $\{b_1, b_2\}$.

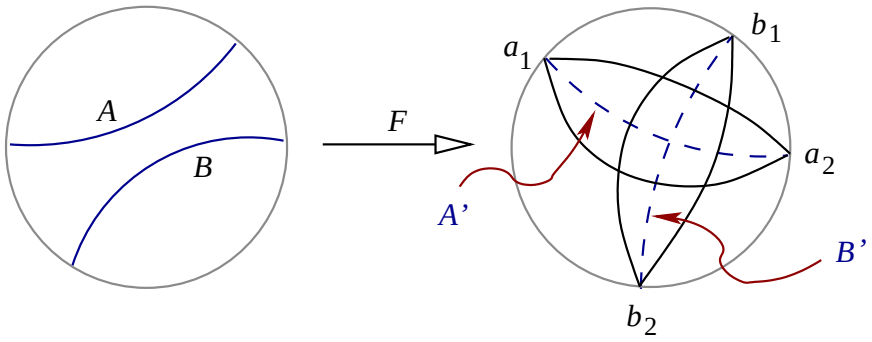

Figure 4.2

Choose two curves $\tilde{A} \subset N_R(A)$ and $\tilde{B} \subset N_R(B)$ of constant distance R to A and B respectively so that the distance from $\tilde{A}$ to $\tilde{B}$ is at least $2R$.

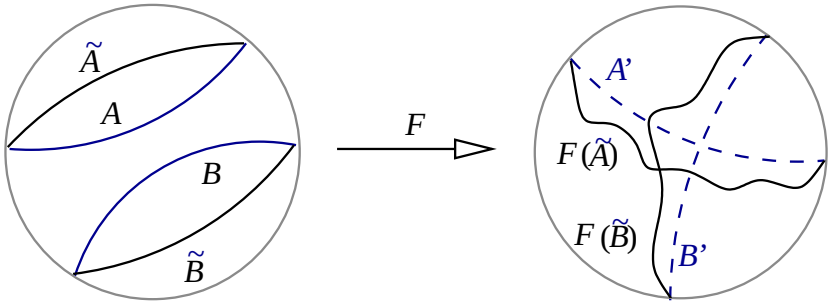

Figure 4.3

Since F is a quasi-isometry, by taking R large, we can make $\text{dist}(F(\tilde{A}), F(\tilde{B}))$ arbitrarily large.

In the special case that $F|_{\tilde{A}}$ and $F|_{\tilde{B}}$ are continuous, we choose R such that $\text{dist}(F(\tilde{A}), F(\tilde{B})) \geq 1$. It is easy to conclude from continuity of F that both $\mathbb{H}^2 \setminus F(\tilde{A})$ and $\mathbb{H}^2 \setminus F(\tilde{B})$ are disconnected. According to the assumption that $\text{Axis}(A') \cap \text{Axis}(B') \neq \emptyset$, we have $F(\tilde{A}) \cap F(\tilde{B}) \neq \emptyset$ for any choice of R. This contradicts that $\text{dist}(F(\tilde{A}), F(\tilde{B})) \geq 1$.

In the general case, since $F(A) \subset N_c(A')$ and $F(B) \subset N_c(B')$ and F is a quasi-isometry, for any curve $\tilde{A}$ (respectively $\tilde{B}$) of constant distance to A (respectively B), there exists $0 < c' \in \mathbb{R}$ such that $F(\tilde{A}) \subset N_{c'}(A')$ and $F(\tilde{B}) \subset N_{c'}(B')$. Now by the assumption that $A' \cap B' \neq \emptyset$, it follows that for any choices of $\tilde{A}$ and $\tilde{B}$ there is a point z which is within a constant distance c^* to both $F(\tilde{A})$ and $F(\tilde{B})$ where c^* depends only on K. However, as we see from the above case, by taking R large, we can make $\text{dist}(F(\tilde{A}), F(\tilde{B})) > 2c^*$. This is a contradiction. □

5. Dehn-Nielsen Theorem

We follow the proof appeared in Farb-Margalit's book to prove Dehn-Nielsen Theorem in this section. The basic idea of the proof is the same as that of Dehn.

Given a group G, we use $\mathrm{Aut}(G)$ to denote the group of all self-isomorphisms of G. Furthermore, let

$$\mathrm{Inn}(G) = \big\{\phi \in \mathrm{Aut}(G) \mid \phi(x) = g^{-1}xg \text{ for all } x \in G\,, \text{ for some } g \in G\big\}$$

be the group of *inner automorphisms* of G. It is easy to check that $\mathrm{Inn}(G)$ is a normal subgroup of $\mathrm{Aut}(G)$. We define the *outer automorphism* group of G to be the quotient group

$$\mathrm{Out}(G) = \mathrm{Aut}(G)/\,\mathrm{Inn}(G).$$

For instance, $\mathrm{Out}\,(\mathbb{Z}^n) = \mathrm{GL}(n,\mathbb{Z})$.

Given a surface Σ with a base point $p_0 \in \Sigma$, a self-homeomorphism h on Σ induces an isomorphism $h_* : \pi_1(\Sigma, p_0) \to \pi_1(\Sigma, h(p_0))$. Let

$$\theta_\xi : \pi_1(\Sigma, h(p_0)) \to \pi_1(\Sigma, p_0)$$

be an isomorphism induced by a path ξ from $h(p_0)$ to p_0. Then $\theta_\xi \circ h_* \in \mathrm{Aut}(\pi_1(\Sigma, p_0))$. Moreover, different choices of paths ξ, η from $h(p_0)$ to p_0 will result in two automorphisms, $\theta_\xi \circ h_*$ and $\theta_\eta \circ h_*$. They are obviously related by an inner automorphism. Thus h_* represents a well defined element in $\mathrm{Out}(\pi_1(\Sigma, p_0))$, which we still denote h_* for simplicity. Furthermore, $(h_1 \circ h_2)_* = (h_1)_*(h_2)_*$ for self-homeomorphisms h_1, h_2 of Σ. In particular, by sending the homotopy class $[h]$ to h_*, we produce a group homomorphism

$$\Psi : \Gamma(\Sigma) = \mathrm{Homeo}(\Sigma)/\simeq \quad \longrightarrow \quad \mathrm{Out}(\pi_1(\Sigma, p_0)).$$

Dehn-Nielsen Theorem. *Suppose $\Sigma = \Sigma_{g,0}$ is a closed surface of genus $g \geqslant 1$, then Ψ is an isomorphism.*

5.1. **Injectivity of Ψ.** The proof is essentially the same as the proof for the torus given in Theorem 1.7. We sketch the argument as follows.

First take a collection of simple closed curves $a_1, \ldots, a_{2g}$ in Σ so that

(1) $\bigcap_{i=1}^{2g} a_i = \{\, p_0 \,\} = a_j \cap a_k$ for all pair j, k of indices; and
(2) $\Sigma - \bigcup_{i=1}^{2g} a_i$ is a topological disk.

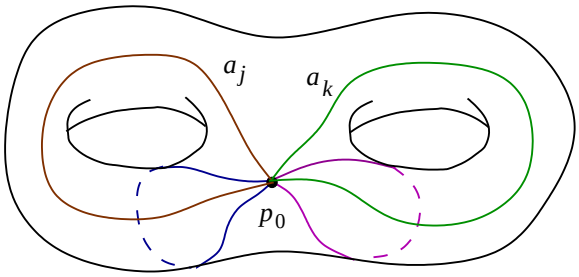

Figure 5.1

It is known from surface topology that $\pi_1(\Sigma, p_0)$ is generated by the homotopy classes $[a_1], \ldots, [a_{2g}]$.

Let $[h] \in \ker \Psi \subset \Gamma(\Sigma)$. That is, $h \in \mathrm{Homeo}(\Sigma)$ such that $h_* \in \mathrm{Inn}(\pi_1(\Sigma, p_0))$. Then, there is a loop b based at p_0 such that for each $i = 1, 2, \ldots, 2g$, the loop $h(a_i)$ is homotopic to ba_ib^{-1} relative p_0. We want to show that h is homotopic to the identity map. To construct the homotopy H from h to id_Σ, let us define H on $\Sigma \times \{0, 1\} \cup \bigcup_{i=1}^{2g} a_i \times I$ by,

$$H(x, 0) = h(x)\,, \quad x \in \Sigma$$

$$H(x, 1) = x\,, \quad x \in \Sigma$$

$$H(p_0, t) = b(t), \quad t \in [0, 1]$$

$$H|_{a_i \times [0,1]} = \text{the homotopy from } a_i \text{ to } h(a_i), i = 1, \ldots, 2g.$$

Now using the facts that $\pi_2(\Sigma, p_0) = 0$ and $(\Sigma \times [0, 1]) \setminus (\bigcup_{i=1}^{2g} a_i \times [0, 1])$ is a 3-cell with H defined on the boundary of the 3-cell, we conclude that H extends to $\Sigma \times [0, 1]$ and becomes a homotopy between h and id_Σ. □

5.2. **Surjectivity of Ψ.** Suppose $\phi : \pi_1(\Sigma, p_0) \to \pi_1(\Sigma, p_0)$ is an isomorphism. We will show that $\phi = h_*$ for some $h \in \mathrm{Homeo}(\Sigma)$.

First of all, we may assume that the genus $g \geqslant 2$ since the result for $g = 1$ was proved in Theorem 1.7. In this case, we will identify the universal cover of Σ with the hyperbolic plane $\mathbb{H}^2$ and $\pi_1(\Sigma, p_0)$ with the deck transformation group G acting isometrically on $\mathbb{H}^2$.

In the rest of the proof, we use $[s]$ to denote the free homotopy class of a loop s in Σ. We will also identify $[s]$ with the conjugacy class of an element in $\pi_1(\Sigma, p_0)$.

Lemma 5.1. *For the automorphism $\phi : \pi_1(\Sigma, p_0) \to \pi_1(\Sigma, p_0)$ and $\alpha, \beta \in \pi_1(\Sigma, p_0)$,*

(1) *if α is simple, then so is $\phi(\alpha)$*
(2) *if α, β are disjoint, then $\phi(\alpha)$ and $\phi(\beta)$ are disjoint*
(3) *if α, β are simple, and $I(\alpha, \beta) = 1$, then $I(\phi(\alpha), \phi(\beta)) = 1$.*

Proof. By Proposition 4.11, we identify G with the orbit $G(z) \subset \mathbb{H}^2$ and extend $\phi : G \to G$ to a quasi-isometry on $\mathbb{H}^2$, which is still denoted by $\phi : \mathbb{H}^2 \to \mathbb{H}^2$.

By choosing a hyperbolic metric on Σ, we can identify each conjugacy class in $\pi_1(\Sigma, p_0) - \{\,\mathrm{id}\,\}$ with its geodesic representative.

For (1), let α be simple, i.e., its geodesic representative s is simple. By definition, s is simple if and only if $\Theta^{-1}(s)$ is a disjoint union of geodesics, where $\Theta : \mathbb{H}^2 \to \Sigma$ is the universal covering projection.

Let us pick a $y \in G$ such that $\alpha = \{xyx^{-1} \mid x \in G\}$. Then

$$\Theta^{-1}(s) = \bigcup_{x \in G} x(\mathrm{Axis}(y))$$

Then s is simple if and only if $\mathrm{Axis}(x_1 y x_1^{-1}) \cap \mathrm{Axis}(x_2 y x_2^{-1}) = \emptyset$ for all pairs of $x_1, x_2 \in G$ with $x_1 x_2^{-1} \notin \{y^n \mid n \in \mathbb{Z}\}$. According to Proposition 4.11 that every quasi-isometry preserves the disjointness of axes, together with the fact that for each $z \in G$, $\mathrm{Axis}(\phi(z))$ and $\phi(\mathrm{Axis}(z))$ have the same end points in $\partial\mathbb{H}^2$, we have

$$\mathrm{Axis}(\phi(x_1)\phi(y)\phi(x_1)^{-1}) \cap \mathrm{Axis}(\phi(x_2)\phi(y)\phi(x_2)^{-1}) = \emptyset\,.$$

On the other hand, it follows from the surjectivity of $\phi : G \to G$ that

$$\phi(\alpha) = \{x\phi(y)x^{-1} \mid x \in G\} \qquad \text{and therefore}$$
$$\mathrm{Axis}(x_1\phi(y)x_1^{-1}) \cap \mathrm{Axis}(x_2\phi(y)x_2^{-1}) = \emptyset\,,$$

for each pair of $x_1, x_2 \in G$ with $x_1 x_2^{-1} \notin \{\phi(y)^n \mid n \in \mathbb{Z}\}$. Thus $\phi(\alpha)$ is simple.

The proof of (2) is similar. Namely, α, β are disjoint if and only if for any $x \in \alpha$ and $y \in \beta$, $\mathrm{Axis}(x) \cap \mathrm{Axis}(y) = \emptyset$. Again, the quasi-isometry ϕ preserves the disjointness of axes. Thus the result follows.

For statement (3), we can use Proposition 4.11 together with Lemma 3.12. □

Finally, we come to the proof of Dehn-Nielsen's theorem.

Let $\phi : \pi_1(\Sigma, p_0) \to \pi_1(\Sigma, p_0)$ be an automorphism. Choose oriented simple loops $\{ c_1, \ldots, c_{2g} \}$ as shown in Figure 5.2. Note that $\Sigma \setminus \cup_{i=1}^{2g} c_i$ is connected.

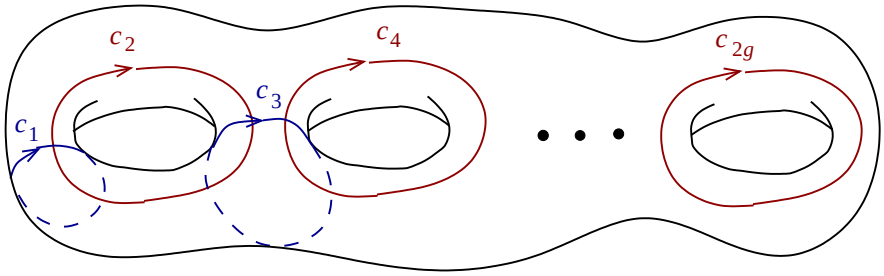

Figure 5.2

More precisely, if we denote $[c_i]$ the free homotopy class of the unoriented loop c_i for $i = 1, \ldots, 2g$, the sequence $\{ c_i \ : \ i = 1, \ldots, 2g \}$ satisfies

(1) $I([c_i], [c_{i+1}]) = 1$ for $i = 1, \ldots, 2g-1$;
(2) $I([c_i], [c_j]) = 0$ if $|i - j| \geqslant 2$ for $i, j = 1, \ldots, 2g$;
(3) the algebraic intersection sign at $c_i \cap c_{i+1}$ from c_i to c_{i+1} is 1 for each $i = 1, \ldots, 2g-1$.

By Lemma 5.1, $\phi([c_i])$ is simple for each i. Then, there is a new sequence $\{ d_1, \ldots, d_{2g} \}$ of simple loops such that $[d_i] = \phi([c_i])$. Moreover, by Lemma 5.1 again, $I(\phi([c_i]), \phi([c_{i+1}])) = 1$ and $I(\phi([c_i]), \phi([c_j])) = 0$ for $|i-j| \geqslant 2$. By taking d_i to be the geodesic representative of $[d_i]$, we may further assume that

$$|d_i \cap d_{i+1}| = 1 \text{ for } i = 1, \ldots, 2g-1 \text{ and } |d_i \cap d_j| = 0 \text{ for } |i-j| \geqslant 2 .$$

By the classification of surfaces, there exists a homeomorphism $h \in \mathrm{Homeo}(\Sigma)$ such that

$$h(c_i) = d_i \quad \text{for } i = 1, 2, \ldots, 2g\,,$$

and hence, $h_*([c_i]) = \phi([c_i])$ for all i. Indeed, since both c_1, d_1 are nonseparating, we may find $h_1 \in \mathrm{Homeo}(\Sigma)$ so that $h_1(c_1) = d_1$. Thus, we may assume $c_1 = d_1$ after composition with h_1. Next, since both c_2, d_2 are nonseparating in with $|c_1 \cap c_2| = |c_1 \cap d_2| = 1$, we are able to find $h_2 \in \mathrm{Homeo}(\Sigma)$ so that

$$h_2(c_1) = c_1 \quad \text{and} \quad h_2(c_2) = d_2.$$

Inductively, after taking composition, we find $h \in \mathrm{Homeo}(\Sigma)$ with the required property, $h(c_i) = d_i$, $i = 1, ..., 2g$. Note that in the process, each c_k is nonseparating in the surface obtained by cutting along $c_1, \ldots, c_{k-1}$.

We further claim that there is an involution τ of Σ such that

$$h_* = \phi \quad \text{or} \quad h_* = \tau_* \circ \phi \qquad \text{in the group } \mathrm{Out}(\pi_1(\Sigma, p_0)).$$

To see this, let us consider the automorphism $\phi \circ h_*^{-1} \in \mathrm{Aut}(\pi_1(\Sigma, p_0))$, which takes $[c_i]$ to $[c_i]$. Thus, we may simply assume that ϕ fixes each free homotopy class $[c_i]$ of the unoriented curve c_i. With this assumption, the goal is to show $\phi \in \mathrm{Inn}(G)$ or $\tau_* \circ \phi \in \mathrm{Inn}(G)$.

To this end, choose a set of generators $z_1, \ldots, z_{2g}$ for $\pi_1(\Sigma, p_0)$ as shown so that z_i and c_i are freely homotopic as oriented loops.

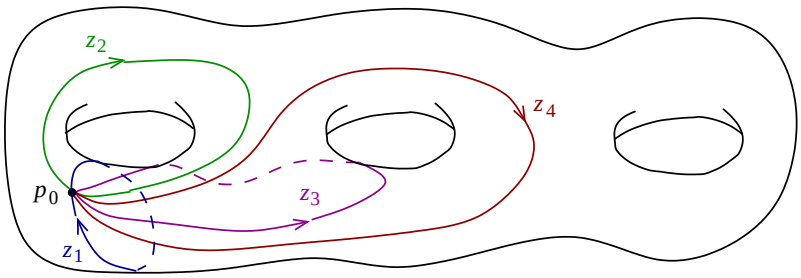

Figure 5.3

There are two involutions τ_1 and τ_2 of Σ leaving each c_i invariant. The first involution τ_1 is the hyper-elliptic involution which reverses orientation of each c_i.

The second involution τ_2 is the reflection of Σ about a "plane" which leaves the orientation of c_{2i+1} invariant and reverses the orientation of each c_{2i}.

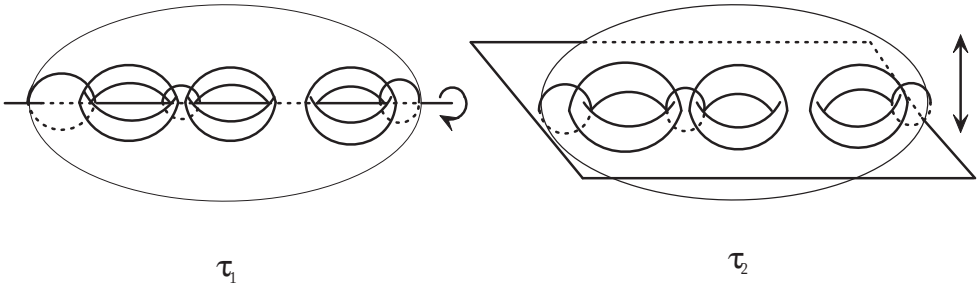

Figure 5.4

Since $\phi([c_1]) = [c_1]$, we have $\phi(z_1) = az_1a^{-1}$ or $\phi(z_1) = az_1^{-1}a^{-1}$ for some $a \in G$. By composing ϕ with the inner automorphism $x \mapsto a^{-1}xa$ and the involution $(\tau_1)_*$ if necessary, we may assume that $\phi(z_1) = z_1$.

Now, due to the fact that $\mathrm{Axis}(z_1) \cap \mathrm{Axis}(z_2) \neq \emptyset$, we conclude that

$$\mathrm{Axis}(z_1) \cap \mathrm{Axis}(\phi(z_2)) \neq \emptyset\,.$$

Similarly, $\phi([c_2]) = [c_2]$ implies that $\phi(z_2) = bz_2^{\pm 1}b^{-1}$ for some $b \in G$. By Lemma 3.12, statement (4), and $I([c_1], [c_2]) = 1$, we conclude

$$\phi(z_2) = z_1^n z_2^{\pm 1} z_1^{-n} \quad \text{for some } n \in \mathbb{Z}.$$

By composing ϕ with the inner automorphism $x \mapsto z_1^n x z_1^{-n}$ and the involutions $(\tau_1)_*$ and $(\tau_2)_*$ if necessary, we may assume that

$$\phi(z_2) = z_2.$$

We claim that $\phi(z_i) = z_i$ for $i \geqslant 3$.

Indeed, to see $\phi(z_3) = z_3$, we only need the above assumptions of $\phi(z_1) = z_1$ and $\phi(z_2) = z_2$ together with the fact that $\phi([c_3]) = [c_3]$.

Due to $I(\phi[c_3], \phi[c_2]) = I([c_3], [c_2]) = 1$, and $\mathrm{Axis}(z_2) \cap \mathrm{Axis}(z_3) \neq \emptyset$, we conclude that $\mathrm{Axis}(\phi(z_2)) \cap \mathrm{Axis}(\phi(z_3)) \neq \emptyset$ and so

$$\phi(z_3) = z_2^n z_3^{\pm 1} z_2^{-n} \qquad \text{for some } n \in \mathbb{Z}.$$

Suppose that $n \neq 0$ and then we will derive a contradiction. In such case, the self-composition ϕ^k of ϕ satisfies

$$\phi^k(z_3) = z_2^{nk} z_3^{\pm 1} z_2^{-nk} .$$

As a result, $\mathrm{Axis}(\phi^k(z_3)) = z_2^{nk}(\mathrm{Axis}(z_3))$ converges to the end points of $\mathrm{Axis}(z_2)$ in $\overline{\mathbb{H}^2}$ as $k \to \infty$ in the Hausdorff metric.

On the other hand, $\phi^k(z_1 z_2) = z_1 z_2$ is represented by a simple loop of which the geodesic representative intersects z_3 at one point. In particular

$$\mathrm{Axis}(z_3) \cap \mathrm{Axis}(z_1 z_2) \neq \emptyset .$$

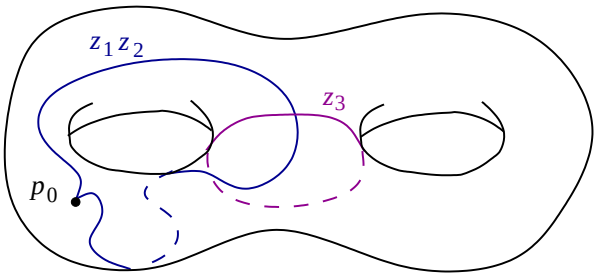

Figure 5.5

Thus, for all k, $\phi^k(\mathrm{Axis}(z_3)) \cap \phi^k(\mathrm{Axis}(z_1 z_2)) \neq \emptyset$. On the other hand,

$$\phi^k(\mathrm{Axis}(z_3)) = \mathrm{Axis}(\phi^k(z_3)) = z_2^{nk}(\mathrm{Axis}(z_3)) .$$

Therefore, $z_2^{nk}(\mathrm{Axis}(z_3)) \cap \mathrm{Axis}(z_1 z_2) \neq \emptyset$ for all k large. This is impossible since $\mathrm{Axis}(z_1 z_2)$ is disjoint from the end-points of $\mathrm{Axis}(z_2)$ in $\overline{\mathbb{H}^2}$.

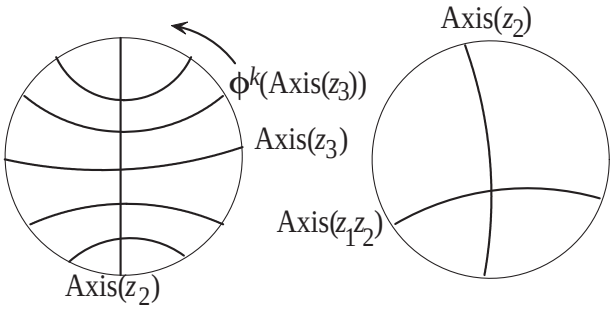

Figure 5.6

It follows that $\phi(z_3) = z_3^{\pm 1}$. We claim $\phi(z_3) = z_3^{-1}$ is also impossible. If otherwise, $\phi(z_3) = z_3^{-1}$, we derive a contradiction as follows. By Figure 5.7, we can see that $I(z_2 z_3, z_1^{-1} z_2) = 0$ and $I(z_2 z_3^{-1}, z_1^{-1} z_2) \neq 0$.

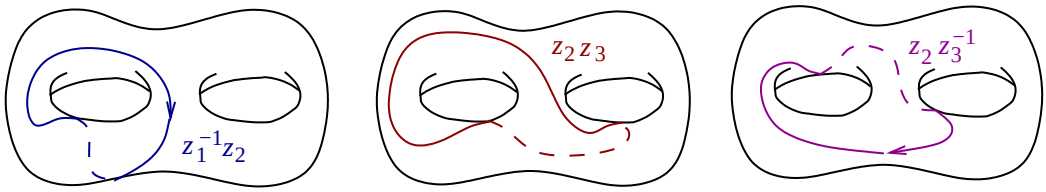

Figure 5.7

Now all of the classes $z_2 z_3$, $z_1^{-1} z_2$, $z_2 z_3^{-1}$ are simple as shown. Thus using $I(z_2 z_3, z_1^{-1} z_2) = 0$ and statement (2) of Lemma 5.1, we conclude that the conjugacy classes of $\phi(z_2 z_3) = z_2 z_3^{-1}$ and $\phi(z_1^{-1} z_2) = z_1^{-1} z_2$ are represented by disjoint simple loops. This contradicts that $I(z_2 z_3^{-1}, z_1^{-1} z_2) \neq 0$.

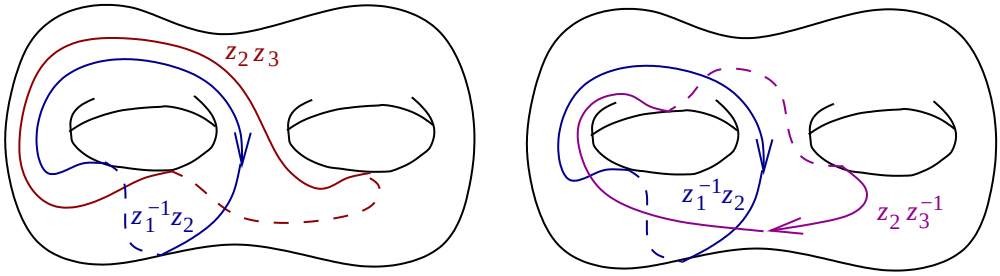

Figure 5.8

Thus we conclude that $\phi(z_3) = z_3$. Next, the facts that $\phi(z_2) = z_2$, $\phi(z_3) = z_3$ and $\phi([c_4]) = [c_4]$ lead to $I([z_2], [c_4]) = 0$ and $I([z_3], [c_4]) = 1$. Following exactly the same argument above and using these facts, we conclude that $\phi(z_4) = z_4$.

Inductively, we conclude that $\phi(z_i) = z_i$ for each $i = 1, \ldots, 2g$. Hence $\phi = \mathrm{id}_\Sigma$. The surjectivity of Ψ is thus established. □

References

[Au] D. Auroux, "A stable classification of Lefschetz fibration", *Geometry & Topology* 9 (2005), pp. 203–217.

[Be] A. Beardon, *Geometry of Discrete Groups.* GTM 91, Springer-Verlag, 1983.
[BH] C. Bödigheimer and R. Hain, *Mapping class group and moduli spaces of Riemann surfaces.* Contemporary Math. 150, American Math. Society, 1993.
[BV] X. Buff, J. Fehrehbach, P. Lochak, L. Schneps, and P. Vogel, *Moduli Spaces of Curves, Mapping Class Groups and Field Theory.* SMF/AMS Texts and Monographs, vol. 9, American Math. Society, 1999.
[CB] A. Casson and S. Bleiler. *Automorphisms of surfaces after Nielsen and Thurston.* London Math. Soc., Student Text 9, Cambridge Univ. Press, 1988.
[De] Max Dehn, *Papers on group theory and topology.* Translated and introduced by John Stillwell, Springer Verlag, 1987.
[Don] S. Donaldson, "Lefschetz fibrations in symplectic geometry", *Doc. Math. J. DMV.*, Extra Volume ICMII (1998), pp. 309-314.
[Ep] D. B. A. Epstein. "Curves on 2-manifolds and isotopies," *Acta Math.,* vol. 115 (1966), pp. 83–107.
[FM] B. Farb and D. Margalit, *A Primer on Mapping Class Groups.* Working Draft, available at http://www.math.uchicago.edu/∼margalit/mcg/mcgv31.pdf
[Gom] R. Gompf, "A new construction of symplectic manifolds", *Ann. of Math.* 142 (1995), pp. 527-595.
[HL] Richard Hain and Eduard Looijenga, *Mapping class groups and moduli spaces of Curves.* Proc. Sympoc. Pure Math. 62, Part 2, Amer. Math. Soc., 1997.
[HT] M. Handel and W. P. Thurston, "New proofs of some results of Nielsen", *Adv. in Math.* 56 (1985), no. 2, pp. 173–191.
[Kap] Michael Kapovich, *Hyperbolic manifolds and discrete groups.* Progress in Mathematics, 183. Birkhuser Boston, Inc., Boston, MA, 2001.
[Ma] Bernard Maskit, *Kleinian groups.* Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 287. Springer-Verlag, 1988.
[Ms] William Massey, *A basic course in Algebraic Topology.* GTM 127. Springer Verlag, 1991.
[Pa] A. Papadopoulos, editor, *Handbook of Teichmüller theory.* European Mathematical Society Publishing House, Zürich.
[Ra] John G. Ratcliff, *Foundations of hyperbolic manifolds.* Second edition, GTM 149. Springer-Verlag, New York, 2006. xii+779 pp.
[Ro] D. Rolfson, *Knots and links.* Publish or Perish, Houston, 1976.
[St] J. Stillwell, *Geometry of surfaces.* Springer-Verlag, 1992.
[St2] J. Stillwell, *Classical Topology and Combinatorial Group Theory.* Second edition. Springer-Verlag, 1993.
[Th] W. P. Thurston, "On the geometry and dynamics of diffeomorphisms of surfaces", *Bull. Amer. Math. Soc.* 19 (1988), pp. 417–431.
[Th2] W. P. Thurston, *The Geomety and Topology of 3-manifolds.* Mimeographed Notes, Mathematics Department, Princeton University, New Jersey, 1979.

Department of Mathematics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China

E-mail address: thomasau@cuhk.edu.hk

Department of Mathematics, Rutgers University, Piscataway, New Jersey, USA

E-mail address: fluo@math.rutgers.edu

Department of Mathematics, Rutgers University, Piscataway, New Jersey, USA

E-mail address: tianyang@math.rutgers.edu