Chapter One

Curves, Surfaces, and Hyperbolic Geometry

A linear transformation of a vector space is determined by, and is best understood by, its action on vectors. In analogy with this, we shall see that an element of the mapping class group of a surface S is determined by, and is best understood by, its action on homotopy classes of simple closed curves in S. We therefore begin our study of the mapping class group by obtaining a good understanding of simple closed curves on surfaces.

Simple closed curves can most easily be studied via their geodesic representatives, and so we begin with the fact that every surface may be endowed with a constant-curvature Riemannian metric, and we study the relation between curves, the fundamental group, and geodesics. We then introduce the geometric intersection number, which we think of as an "inner product" for simple closed curves. A second fundamental tool is the change of coordinates principle, which is analogous to understanding change of basis in a vector space. After explaining these tools, we conclude this chapter with a discussion of some foundational technical issues in the theory of surface topology, such as homeomorphism versus diffeomorphism, and homotopy versus isotopy.

1.1 SURFACES AND HYPERBOLIC GEOMETRY

We begin by recalling some basic results about surfaces and hyperbolic geometry that we will use throughout the book. This is meant to be a brief review; see [208] or [119] for a more thorough discussion.

1.1.1 SURFACES

A *surface* is a 2-dimensional manifold. The following fundamental result about surfaces, often attributed to Möbius, was known in the mid-nineteenth century in the case of surfaces that admit a triangulation. Radò later proved, however, that every compact surface admits a triangulation. For proofs of both theorems, see, e.g., [204].

THEOREM 1.1 (Classification of surfaces) *Any closed, connected, orientable surface is homeomorphic to the connect sum of a 2-dimensional*

sphere with $g \geq 0$ tori. Any compact, connected, orientable surface is obtained from a closed surface by removing $b \geq 0$ open disks with disjoint closures. The set of homeomorphism types of compact surfaces is in bijective correspondence with the set $\{(g, b) : g, b \geq 0\}$.

The g in Theorem 1.1 is the *genus* of the surface; the b is the number of *boundary components*. One way to obtain a noncompact surface from a compact surface S is to remove n points from the interior of S; in this case, we say that the resulting surface has n *punctures*.

Unless otherwise specified, when we say "surface" in this book, we will mean a compact, connected, oriented surface that is possibly punctured (of course, after we puncture a compact surface, it ceases to be compact). We can therefore specify our surfaces by the triple (g, b, n). We will denote by $S_{g,n}$ a surface of genus g with n punctures and empty boundary; such a surface is homeomorphic to the interior of a compact surface with n boundary components. Also, for a closed surface of genus g, we will abbreviate $S_{g,0}$ as S_g. We will denote by ∂S the (possibly disconnected) boundary of S.

Recall that the *Euler characteristic* of a surface S is

$$\chi(S) = 2 - 2g - (b + n).$$

It is a fact that $\chi(S)$ is also equal to the alternating sum of the Betti numbers of S. Since $\chi(S)$ is an invariant of the homeomorphism class of S, it follows that a surface S is determined up to homeomorphism by any three of the four numbers g, b, n, and $\chi(S)$.

Occasionally, it will be convenient for us to think of punctures as *marked points*. That is, instead of deleting the points, we can make them distinguished. Marked points and punctures carry the same topological information, so we can go back and forth between punctures and marked points as is convenient. On the other hand, all surfaces will be assumed to be without marked points unless explicitly stated otherwise.

If $\chi(S) \leq 0$ and $\partial S = \emptyset$, then the universal cover $\widetilde{S}$ is homeomorphic to $\mathbb{R}^2$ (see, e.g., [199, Section 1.4]). We will see that, when $\chi(S) < 0$, we can take advantage of a hyperbolic structure on $\widetilde{S}$.

1.1.2 THE HYPERBOLIC PLANE

Let $\mathbb{H}^2$ denote the hyperbolic plane. One model for $\mathbb{H}^2$ is the *upper half-plane model*, namely, the subset of $\mathbb{C}$ with positive imaginary part ($y > 0$), endowed with the Riemannian metric

$$ds^2 = \frac{dx^2 + dy^2}{y^2},$$

where $dx^2 + dy^2$ denotes the Euclidean metric on $\mathbb{C}$. In this model the geodesics are semicircles and half-lines perpendicular to the real axis.

It is a fact from Riemannian geometry that any complete, simply connected Riemannian 2-manifold with constant sectional curvature -1 is isometric to $\mathbb{H}^2$.

For the *Poincaré disk model* of $\mathbb{H}^2$, we take the open unit disk in $\mathbb{C}$ with the Riemannian metric

$$ds^2 = 4\frac{dx^2 + dy^2}{(1 - r^2)^2}.$$

In this model the geodesics are circles and lines perpendicular to the unit circle in $\mathbb{C}$ (intersected with the open unit disk).

Any Möbius transformation from the upper half-plane to the unit disk is an isometry between the upper half-plane model for $\mathbb{H}^2$ and the Poincaré disk model of $\mathbb{H}^2$. The group of orientation-preserving isometries of $\mathbb{H}^2$ is (in either model) the group of Möbius transformations taking $\mathbb{H}^2$ to itself. This group, denoted $\mathrm{Isom}^+(\mathbb{H}^2)$, is isomorphic to $\mathrm{PSL}(2, \mathbb{R})$. In the upper half-plane model, this isomorphism is given by the following map:

$$\pm \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \left(z \mapsto \frac{az + b}{cz + d} \right).$$

The boundary of the hyperbolic plane. One of the central objects in the study of hyperbolic geometry is the *boundary at infinity* of $\mathbb{H}^2$, denoted by $\partial\mathbb{H}^2$. A point of $\partial\mathbb{H}^2$ is an equivalence class $[\gamma]$ of unit-speed geodesic rays where two rays $\gamma_1, \gamma_2 : [0, \infty) \to \mathbb{H}^2$ are equivalent if they stay a bounded distance from each other; that is, there exists $D > 0$ so that

$$d_{\mathbb{H}^2}(\gamma_1(t), \gamma_2(t)) \leq D \text{ for all } t \geq 0.$$

Actually, if γ_1 and γ_2 are equivalent, then they can be given unit-speed parameterizations so that

$$\lim_{t \to \infty} d_{\mathbb{H}^2}(\gamma_1(t), \gamma_2(t)) = 0.$$

We denote the union $\mathbb{H}^2 \cup \partial\mathbb{H}^2$ by $\overline{\mathbb{H}^2}$. The set $\overline{\mathbb{H}^2}$ is topologized via the following basis. We take the usual open sets of $\mathbb{H}^2$ plus one open set U_P for each open half-plane P in $\mathbb{H}^2$. A point of $\mathbb{H}^2$ lies in U_P if it lies in P, and a point of $\partial\mathbb{H}^2$ lies in U_P if every representative ray $\gamma(t)$ eventually lies in P, i.e., if there exists $T \geq 0$ so that $\gamma(t) \in P$ for all $t \geq T$.

In this topology $\partial\mathbb{H}^2$ is homeomorphic to S^1, and the union $\overline{\mathbb{H}^2}$ is homeomorphic to the closed unit disk. The space $\overline{\mathbb{H}^2}$ is a compactification of $\mathbb{H}^2$

and is called *the* compactification of $\mathbb{H}^2$. In the Poincaré disk model of $\mathbb{H}^2$, the boundary $\partial\mathbb{H}^2$ corresponds to the unit circle in $\mathbb{C}$, and $\overline{\mathbb{H}^2}$ is identified with the closed unit disk in $\mathbb{C}$.

Any isometry $f \in \mathrm{Isom}(\mathbb{H}^2)$ takes geodesic rays to geodesic rays, clearly preserving equivalence classes. Also, f takes half-planes to half-planes. It follows that f extends uniquely to a map $\overline{f} : \overline{\mathbb{H}^2} \to \overline{\mathbb{H}^2}$. As any pair of distinct points in $\partial\mathbb{H}^2$ are the endpoints of a unique geodesic in $\mathbb{H}^2$, it follows that $\overline{f}$ maps distinct points to distinct points. It is easy to check that in fact $\overline{f}$ is a homeomorphism.

Classification of isometries of $\mathbb{H}^2$. We can use the above setup to classify nontrivial elements of $\mathrm{Isom}^+(\mathbb{H}^2)$. Suppose we are given an arbitrary nontrivial element $f \in \mathrm{Isom}^+(\mathbb{H}^2)$. Since $\overline{f}$ is a self-homeomorphism of a closed disk, the Brouwer fixed point theorem gives that $\overline{f}$ has a fixed point in $\overline{\mathbb{H}^2}$. By considering the number of fixed points of $\overline{f}$ in $\overline{\mathbb{H}^2}$, we obtain a classification of isometries of $\mathbb{H}^2$ as follows.

Elliptic. If $\overline{f}$ fixes a point $p \in \mathbb{H}^2$, then f is called *elliptic*, and it is a rotation about p. Elliptic isometries have no fixed points on $\partial\mathbb{H}^2$. They correspond to elements of $\mathrm{PSL}(2,\mathbb{R})$ whose trace has absolute value less than 2.

Parabolic. If $\overline{f}$ has exactly one fixed point in $\partial\mathbb{H}^2$, then f is called *parabolic*. In the upper half-plane model, f is conjugate in $\mathrm{Isom}^+(\mathbb{H}^2)$ to $z \mapsto z \pm 1$. Parabolic isometries correspond to those nonidentity elements of $\mathrm{PSL}(2,\mathbb{R})$ with trace ± 2.

Hyperbolic. If $\overline{f}$ has two fixed points in $\partial\mathbb{H}^2$, then f is called *hyperbolic* or *loxodromic*. In this case, there is an f-invariant geodesic *axis* γ; that is, an f-invariant geodesic in $\mathbb{H}^2$ on which f acts by translation. On $\partial\mathbb{H}^2$ the fixed points act like a source and a sink, respectively. Hyperbolic isometries correspond to elements of $\mathrm{PSL}(2,\mathbb{R})$ whose trace has absolute value greater than 2.

It follows from the above classification that if $\overline{f}$ has at least three fixed points in $\overline{\mathbb{H}^2}$, then f is the identity.

Also, since commuting elements of $\mathrm{Isom}^+(\mathbb{H}^2)$ must preserve each other's fixed sets in $\overline{\mathbb{H}^2}$, we see that two nontrivial elements of $\mathrm{Isom}^+(\mathbb{H}^2)$ commute if and only if they have the same fixed points in $\overline{\mathbb{H}^2}$.

1.1.3 HYPERBOLIC SURFACES

The following theorem gives a link between the topology of surfaces and their geometry. It will be used throughout the book to convert topological

problems to geometric ones, which have more structure and so are often easier to solve.

We say that a surface S *admits a hyperbolic metric* if there exists a complete, finite-area Riemannian metric on S of constant curvature -1 where the boundary of S (if nonempty) is totally geodesic (this means that the geodesics in ∂S are geodesics in S). Similarly, we say that S *admits a Euclidean metric*, or *flat metric* if there is a complete, finite-area Riemannian metric on S with constant curvature 0 and totally geodesic boundary.

If S has empty boundary and has a hyperbolic metric, then its universal cover $\widetilde{S}$ is a simply connected Riemannian 2-manifold of constant curvature -1. It follows that $\widetilde{S}$ is isometric to $\mathbb{H}^2$, and so S is isometric to the quotient of $\mathbb{H}^2$ by a free, properly discontinuous isometric action of $\pi_1(S)$. If S has nonempty boundary and has a hyperbolic metric, then $\widetilde{S}$ is isometric to a totally geodesic subspace of $\mathbb{H}^2$. Similarly, if S has a Euclidean metric, then $\widetilde{S}$ is isometric to a totally geodesic subspace of the Euclidean plane $\mathbb{E}^2$.

THEOREM 1.2 *Let S be any surface (perhaps with punctures or boundary). If $\chi(S) < 0$, then S admits a hyperbolic metric. If $\chi(S) = 0$, then S admits a Euclidean metric.*

A surface endowed with a fixed hyperbolic metric will be called a *hyperbolic surface*. A surface with a Euclidean metric will be called a *Euclidean surface* or *flat surface*.

Note that Theorem 1.2 is consistent with the Gauss–Bonnet theorem which, in the case of a compact surface S with totally geodesic boundary, states that the integral of the curvature over S is equal to $2\pi\chi(S)$.

One way to get a hyperbolic metric on a closed surface S_g is to construct a free, properly discontinuous isometric action of $\pi_1(S_g)$ on $\mathbb{H}^2$ (as above, this requires $g \geq 2$). By covering space theory and the classification of surfaces, the quotient will be homeomorphic to S_g. Since the action was by isometries, this quotient comes equipped with a hyperbolic metric. Another way to get a hyperbolic metric on S_g, for $g \geq 2$, is to take a geodesic $4g$-gon in $\mathbb{H}^2$ with interior angle sum 2π and identify opposite sides (such a $4g$-gon always exists; see Section 10.4 below). The result is a surface of genus g with a hyperbolic metric and, according to Theorem 1.2, its universal cover is $\mathbb{H}^2$.

We remark that while the torus T^2 admits a Euclidean metric, the once-punctured torus $S_{1,1}$ admits a hyperbolic metric.

Loops in hyperbolic surfaces. Let S be a hyperbolic surface. A *neighborhood of a puncture* is a closed subset of S homeomorphic to a once-punctured disk. Also, by a *free homotopy* of loops in S we simply mean an

unbased homotopy. If a nontrivial element of $\pi_1(S)$ is represented by a loop that can be freely homotoped into the neighborhood of a puncture, then it follows that the loop can be made arbitrarily short; otherwise, we would find an embedded annulus whose length is infinite (by completeness) and where the length of each circular cross section is bounded from below, giving infinite area. The deck transformation corresponding to such an element of $\pi_1(S)$ is a parabolic isometry of the universal cover $\mathbb{H}^2$. This makes sense because for any parabolic isometry of $\mathbb{H}^2$, there is no positive lower bound to the distance between a point in $\mathbb{H}^2$ and its image. All other nontrivial elements of $\pi_1(S)$ correspond to hyperbolic isometries of $\mathbb{H}^2$ and hence have associated axes in $\mathbb{H}^2$.

We have the following fact, which will be used several times throughout this book:

> *If S admits a hyperbolic metric, then the centralizer of any nontrivial element of $\pi_1(S)$ is cyclic. In particular, $\pi_1(S)$ has a trivial center.*

To prove this we identify $\pi_1(S)$ with the deck transformation group of S for some covering map $\mathbb{H}^2 \to S$. Whenever two nontrivial isometries of $\mathbb{H}^2$ commute, it follows from the classification of isometries of $\mathbb{H}^2$ that they have the same fixed points in $\partial\mathbb{H}^2$. So if $\alpha \in \pi_1(S)$ is centralized by β, it follows that α and β have the same fixed points in $\partial\mathbb{H}^2$. By the discreteness of the action of $\pi_1(S)$, we would then have that the centralizer of α in $\pi_1(S)$ is infinite cyclic. If $\pi_1(S)$ had nontrivial center, it would then follow that $\pi_1(S) \approx \mathbb{Z}$. But then S would necessarily have infinite volume, a contradiction.

1.2 SIMPLE CLOSED CURVES

Our study of simple closed curves in a surface S begins with the study of all closed curves in S and the usefulness of geometry in understanding them.

1.2.1 CLOSED CURVES AND GEODESICS

By a *closed curve* in a surface S we will mean a continuous map $S^1 \to S$. We will usually identify a closed curve with its image in S. A closed curve is called *essential* if it is not homotopic to a point, a puncture, or a boundary component.

Closed curves and fundamental groups. Given an oriented closed curve $\alpha \in S$, we can identify α with an element of $\pi_1(S)$ by choosing a path from

the basepoint for $\pi_1(S)$ to some point on α. The resulting element of $\pi_1(S)$ is well defined only up to conjugacy. By a slight abuse of notation we will denote this element of $\pi_1(S)$ by α as well.

There is a bijective correspondence:

$$\left\{ \begin{array}{c} \text{Nontrivial} \\ \text{conjugacy classes} \\ \text{in } \pi_1(S) \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{Nontrivial free} \\ \text{homotopy classes of oriented} \\ \text{closed curves in } S \end{array} \right\}$$

An element g of a group G is *primitive* if there does not exist any $h \in G$ so that $g = h^k$, where $|k| > 1$. The property of being primitive is a conjugacy class invariant. In particular, it makes sense to say that a closed curve in a surface is primitive.

A closed curve in S is a *multiple* if it is a map $S^1 \to S$ that factors through the map $S^1 \xrightarrow{\times n} S^1$ for $n > 1$. In other words, a curve is a multiple if it "runs around" another curve multiple times. If a closed curve in S is a multiple, then no element of the corresponding conjugacy class in $\pi_1(S)$ is primitive.

Let $p : \widetilde{S} \to S$ be any covering space. By a *lift* of a closed curve α to $\widetilde{S}$ we will always mean the image of a lift $\mathbb{R} \to \widetilde{S}$ of the map $\alpha \circ \pi$, where $\pi : \mathbb{R} \to S^1$ is the usual covering map. For example, if S is a surface with $\chi(S) \leq 0$, then a lift of an essential simple closed curve in S to the universal cover is a copy of $\mathbb{R}$. Note that a lift is different from a path lift, which is typically a proper subset of a lift.

Now suppose that $\widetilde{S}$ is the universal cover and α is a simple closed curve in S that is not a nontrivial multiple of another closed curve. In this case, the lifts of α to $\widetilde{S}$ are in natural bijection with the cosets in $\pi_1(S)$ of the infinite cyclic subgroup $\langle \alpha \rangle$. (Any nontrivial multiple of α has the same set of lifts as α but more cosets.) The group $\pi_1(S)$ acts on the set of lifts of α by deck transformations, and this action agrees with the usual left action of $\pi_1(S)$ on the cosets of $\langle \alpha \rangle$. The stabilizer of the lift corresponding to the coset $\gamma\langle \alpha \rangle$ is the cyclic group $\langle \gamma\alpha\gamma^{-1} \rangle$.

When S admits a hyperbolic metric and α is a primitive element of $\pi_1(S)$, we have a bijective correspondence:

$$\left\{ \begin{array}{c} \text{Elements of the conjugacy} \\ \text{class of } \alpha \text{ in } \pi_1(S) \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{Lifts to } \widetilde{S} \text{ of the} \\ \text{closed curve } \alpha \end{array} \right\}$$

More precisely, the lift of the curve α given by the coset $\gamma\langle \alpha \rangle$ corresponds to the element $\gamma\alpha\gamma^{-1}$ of the conjugacy class $[\alpha]$. That this is a bijective correspondence is a consequence of the fact that, for a hyperbolic surface S, the centralizer of any element of $\pi_1(S)$ is cyclic.

If α is any multiple, then we still have a bijective correspondence between elements of the conjugacy class of α and the lifts of α. However, if α is not primitive and not a multiple, then there are more lifts of α than there are conjugates. Indeed, if $\alpha = \beta^k$, where $k > 1$, then $\beta\langle\alpha\rangle \neq \langle\alpha\rangle$ while $\beta\alpha\beta^{-1} = \alpha$.

Note that the above correspondence does not hold for the torus T^2. This is so because each closed curve has infinitely many lifts, while each element of $\pi_1(T^2) \approx \mathbb{Z}^2$ is its own conjugacy class. Of course, $\pi_1(T^2)$ is its own center, and so the centralizer of each element is the whole group.

Geodesic representatives. A priori the combinatorial topology of closed curves on surfaces has nothing to do with geometry. It was already realized in the nineteenth century, however, that the mere existence of constant-curvature Riemannian metrics on surfaces has strong implications for the topology of the surface and of simple closed curves in it. For example, it is easy to prove that any closed curve α on a flat torus is homotopic to a geodesic: one simply lifts α to $\mathbb{R}^2$ and performs a straight-line homotopy. Note that the corresponding geodesic is not unique.

For compact hyperbolic surfaces we have a similar picture, and in fact the free homotopy class of any closed curve contains a unique geodesic. The existence is indeed true for any compact Riemannian manifold. Here we give a more hands-on proof of existence and uniqueness for any hyperbolic surface.

Proposition 1.3 *Let S be a hyperbolic surface. If α is a closed curve in S that is not homotopic into a neighborhood of a puncture, then α is homotopic to a unique geodesic closed curve γ.*

Proof. Choose a lift $\widetilde{\alpha}$ of α to $\mathbb{H}^2$. As above, $\widetilde{\alpha}$ is stabilized by some element of the conjugacy class of $\pi_1(S)$ corresponding to α; let ϕ be the corresponding isometry of $\mathbb{H}^2$. By the assumption on α, we have that ϕ is a hyperbolic isometry and so has an axis of translation A; see Figure 1.1.

Consider the projection of A to S and let γ_0 be a geodesic closed curve that travels around this projection once. Any equivariant homotopy from $\widetilde{\alpha}$ to A projects to a homotopy between α and a multiple of γ_0, which is the desired γ. One way to get such a homotopy is to simply take the homotopy that moves each point of $\widetilde{\alpha}$ along a geodesic segment to its closest-point projection in A. This completes the proof of the existence of γ. Note that we do not need to worry that the resulting parameterization of γ is geodesic since any two parameterizations of the same closed curve are homotopic as parameterized maps.

To prove uniqueness, suppose we are given a homotopy $S^1 \times I \to S$

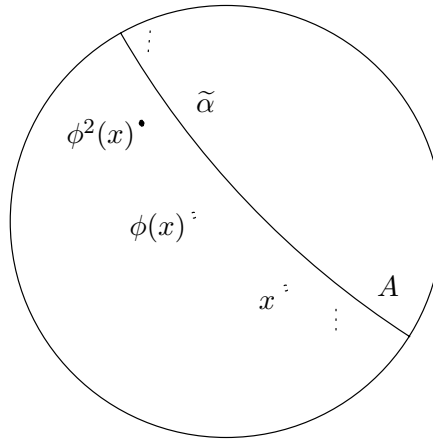

Figure 1.1 A lift $\widetilde{\alpha}$ of a closed curve α and the axis A for the corresponding isometry ϕ.

from α to a multiple γ' of some simple closed geodesic γ_0'. By compactness of $S^1 \times I$, there exists a constant $C \geq 0$ such that no point of α is moved a distance greater than C by the homotopy. In the universal cover $\mathbb{H}^2$, the homotopy lifts to a homotopy from the lift $\widetilde{\alpha}$ of α to a geodesic lift $\widetilde{\gamma}_0'$ of γ_0', and points of $\widetilde{\alpha}$ are moved a distance at most C. It follows that the endpoints of $\widetilde{\alpha}$ in $\partial\mathbb{H}^2$ are the same as those of $\widetilde{\gamma}_0'$. Since a geodesic in $\mathbb{H}^2$ is uniquely determined by its endpoints in $\partial\mathbb{H}^2$, this proves that the geodesic closed curve γ_0' is the same as γ_0 up to sign. The closed curve γ' is then specified by which multiple of γ_0 it is. But different multiples of γ_0 correspond to conjugacy classes in $\mathrm{Isom}^+(\mathbb{H}^2)$ that have different translation lengths and/or translation directions. Conjugacy classes with differing translation lengths are distinct, and so distinct multiples of γ_0 do not lie in the same free homotopy class. □

It follows from Proposition 1.3 that for a compact hyperbolic surface we have a bijective correspondence:

$$\left\{ \begin{array}{c} \text{Conjugacy classes} \\ \text{in } \pi_1(S) \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{Oriented geodesic} \\ \text{closed curves in } S \end{array} \right\}$$

1.2.2 Simple Closed Curves

A closed curve in S is *simple* if it is embedded, that is, if the map $S^1 \to S$ is injective. Among the reasons for the particular importance of simple closed curves is that we can easily classify them up to homeomorphism

of S (see Section 1.3), we can cut along them (see Section 1.3), and we can twist along them (see Section 3.1). As mentioned above, we will study homeomorphisms of surfaces via their actions on simple closed curves.

Any closed curve α can be approximated by a smooth closed curve, and a close enough approximation α' of α is homotopic to α. What is more, if α is simple, then α' can be chosen to be simple. Smooth curves are advantageous for many reasons. For instance, smoothness allows us to employ the notion of transversality (general position). When convenient, we will assume that our curves are smooth, sometimes without mention.

Simple closed curves are also natural to study because they represent primitive elements of $\pi_1(S)$.

Proposition 1.4 *Let α be a simple closed curve in a surface S. If α is not null homotopic, then each element of the corresponding conjugacy class in $\pi_1(S)$ is primitive.*

Proof. We give the proof for the case when S is hyperbolic. Fix a covering map $\mathbb{H}^2 \to S$ and let $\phi \in \mathrm{Isom}^+(\mathbb{H}^2)$ be the hyperbolic isometry corresponding to some element of the conjugacy class of α. The primitivity of the elements of the conjugacy class of α is equivalent to the primitivity of ϕ in the deck transformation group.

Assume that $\phi = \psi^n$, where ψ is another element of the deck transformation group and $n \in \mathbb{Z}$. In any group, powers of the same element commute, and so ϕ commutes with ψ. Thus ϕ and ψ have the same set of fixed points in $\partial\mathbb{H}^2$.

Let $\widetilde{\alpha}$ be the lift of the closed curve α that has the same endpoints in $\partial\mathbb{H}^2$ as the axis for ϕ. We claim that $\psi(\widetilde{\alpha}) = \widetilde{\alpha}$. We know that $\psi(\widetilde{\alpha})$ is some lift of α. Since α is simple, all of its lifts are disjoint and no two lifts of α have the same endpoints in $\partial\mathbb{H}^2$. Thus $\psi(\widetilde{\alpha})$ and $\widetilde{\alpha}$ are disjoint and have distinct endpoints. Now, we know that $\psi^{n-1}(\psi(\widetilde{\alpha})) = \phi(\widetilde{\alpha}) = \widetilde{\alpha}$. Since the fixed points in $\partial\mathbb{H}^2$ of ψ^{n-1} are the same as the endpoints of $\widetilde{\alpha}$, the only way $\psi^{n-1}(\psi(\widetilde{\alpha}))$ can have the same endpoints at infinity as $\widetilde{\alpha}$ is if $\psi(\widetilde{\alpha})$ does. This is to say that $\psi(\widetilde{\alpha}) = \widetilde{\alpha}$, and the claim is proven.

Thus the restriction of ψ to $\widetilde{\alpha}$ is a translation. As $\phi = \psi^n$, the closed curve α travels n times around the closed curve in S given by $\widetilde{\alpha}/\langle\psi\rangle$. Since α is simple, we have $n = \pm 1$, which is what we wanted to show. □

Simple closed curves in the torus. We can classify the set of homotopy classes of simple closed curves in the torus T^2 as follows. Let $\mathbb{R}^2 \to T^2$ be the usual covering map, where the deck transformation group is generated by the translations by $(1, 0)$ and $(0, 1)$. We know that $\pi_1(T^2) \approx \mathbb{Z}^2$, and if

we base $\pi_1(T^2)$ at the image of the origin, one way to get a representative for (p, q) as a loop in T^2 is to take the straight line from $(0, 0)$ to (p, q) in $\mathbb{R}^2$ and project it to T^2.

Let γ be any oriented simple closed curve in T^2. Up to homotopy, we can assume that γ passes through the image in T^2 of $(0, 0)$ in $\mathbb{R}^2$. Any path lifting of γ to $\mathbb{R}^2$ based at the origin terminates at some integral point (p, q). There is then a homotopy from γ to the standard straight-line representative of $(p, q) \in \pi_1(T^2)$; indeed, the straight-line homotopy from the lift of γ to the straight line through $(0, 0)$ and (p, q) is equivariant with respect to the group of deck transformations and thus descends to the desired homotopy.

Now, if a closed curve in T^2 is simple, then its straight-line representative is simple. Thus we have the following fact.

Proposition 1.5 *The nontrivial homotopy classes of oriented simple closed curves in T^2 are in bijective correspondence with the set of primitive elements of $\pi_1(T^2) \approx \mathbb{Z}^2$.*

An element (p, q) of $\mathbb{Z}^2$ is primitive if and only if $(p, q) = (0, \pm 1)$, $(p, q) = (\pm 1, 0)$, or $\gcd(p, q) = 1$.

We can classify homotopy classes of essential simple closed curves in other surfaces. For example, in S^2, $S_{0,1}$, $S_{0,2}$, and $S_{0,3}$, there are no essential simple closed curves. The homotopy classes of simple closed curves in $S_{1,1}$ are in bijective correspondence with those in T^2. In Section 2.2 below, we will show that there is a natural bijection between the homotopy classes of essential simple closed curves in $S_{0,4}$ and the homotopy classes in T^2.

Closed geodesics. For hyperbolic surfaces geodesics are the natural representatives of each free homotopy class in the following sense.

Proposition 1.6 *Let S be a hyperbolic surface. Let α be a closed curve in S not homotopic into a neighborhood of a puncture. Let γ be the unique geodesic in the free homotopy class of α guaranteed by Proposition 1.3. If α is simple, then γ is simple.*

Proof. We begin by applying the following fact.

> *A closed curve β in a hyperbolic surface S is simple if and only if the following properties hold:*
>
> 1. *Each lift of β to $\mathbb{H}^2$ is simple.*
> 2. *No two lifts of β intersect.*

3. β is not a nontrivial multiple of another closed curve.

Thus if α is simple, then no two of its lifts to $\mathbb{H}^2$ intersect. It follows that for any two such lifts, their endpoints are not linked in $\partial\mathbb{H}^2$. But each lift of γ shares both endpoints with some lift of α. Thus no two lifts of γ have endpoints that are linked in $\partial\mathbb{H}^2$. Since these lifts are geodesics, it follows that they do not intersect. Further, by Proposition 1.4, any element of $\pi_1(S)$ corresponding to α is primitive. The same is then true for γ, and so γ cannot be a multiple. Since geodesics in $\mathbb{H}^2$ are always simple, we conclude that γ is simple. □

1.2.3 Intersection Numbers

There are two natural ways to count the number of intersection points between two simple closed curves in a surface: signed and unsigned. These correspond to the algebraic intersection number and geometric intersection number, respectively.

Let α and β be a pair of transverse, oriented, simple closed curves in S. Recall that the *algebraic intersection number* $\hat{i}(\alpha, \beta)$ is defined as the sum of the indices of the intersection points of α and β, where an intersection point is of index $+1$ when the orientation of the intersection agrees with the orientation of S and is -1 otherwise. Recall that $\hat{i}(\alpha, \beta)$ depends only on the homology classes of α and β. In particular, it makes sense to write $\hat{i}(a, b)$ for a and b, the free homotopy classes (or homology classes) of closed curves α and β.

The most naive way to count intersections between homotopy classes of closed curves is to simply count the minimal number of unsigned intersections. This idea is encoded in the concept of geometric intersection number. The *geometric intersection number* between free homotopy classes a and b of simple closed curves in a surface S is defined to be the minimal number of intersection points between a representative curve in the class a and a representative curve in the class b:

$$i(a, b) = \min\{|\alpha \cap \beta| : \alpha \in a,\ \beta \in b\}.$$

We sometimes employ a slight abuse of notation by writing $i(\alpha, \beta)$ for the intersection number between the homotopy classes of simple closed curves α and β.

We note that geometric intersection number is symmetric, while algebraic intersection number is skew-symmetric: $i(a, b) = i(b, a)$, while $\hat{i}(a, b) = -\hat{i}(b, a)$. While algebraic intersection number is well defined on homology classes, geometric intersection number is well defined only on free homotopy classes. Geometric intersection number is a useful invariant

but, as we will see, it is more difficult to compute than algebraic intersection number.

Observe that $i(a, a) = 0$ for any homotopy class of simple closed curves a. If α separates S into two components, then for any β we have $\hat{i}(\alpha, \beta) = 0$ and $i(\alpha, \beta)$ is even. In general, i and $\hat{i}$ have the same parity.

Intersection numbers on the torus. As noted above, the nontrivial free homotopy classes of oriented simple closed curves in T^2 are in bijective correspondence with primitive elements of $\mathbb{Z}^2$. For two such homotopy classes (p, q) and (p', q'), we have

$$\hat{i}((p, q), (p', q')) = pq' - p'q$$

and

$$i((p, q), (p', q')) = |pq' - p'q|.$$

To verify these formulas, one should first check the case where $(p, q) = (1, 0)$ (exercise). For the general case, we note that if (p, q) represents an essential oriented simple closed curve, that is, if it is primitive, then there is a matrix $A \in \mathrm{SL}(2, \mathbb{Z})$ with $A((p, q)) = (1, 0)$. Since A is a linear, orientation-preserving homeomorphism of $\mathbb{R}^2$ preserving $\mathbb{Z}^2$, it induces an orientation-preserving homeomorphism of the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ whose action on $\pi_1(T^2) \approx \mathbb{Z}^2$ is given by A. Since orientation-preserving homeomorphisms preserve both algebraic and geometric intersection numbers, the general case of each formula follows.

Minimal position. In practice, one computes the geometric intersection number between two homotopy classes a and b by finding representatives α and β that realize the minimal intersection in their homotopy classes, so that $i(a, b) = |\alpha \cap \beta|$. When this is the case, we say that α and β are in *minimal position*.

Two basic questions now arise.

1. Given two simple closed curves α and β, how can we tell if they are in minimal position?

2. Given two simple closed curves α and β, how do we find homotopic simple closed curves that are in minimal position?

While the first question is a priori a minimization problem over an infinite-dimensional space, we will see that the question can be reduced to a finite check—the bigon criterion given below. For the second question, we will see

that geodesic representatives of simple closed curves are always in minimal position.

1.2.4 THE BIGON CRITERION

We say that two transverse simple closed curves α and β in a surface S form a *bigon* if there is an embedded disk in S (the bigon) whose boundary is the union of an arc of α and an arc of β intersecting in exactly two points; see Figure 1.2.

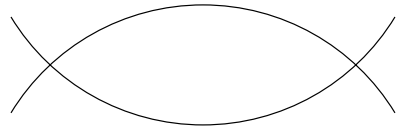

Figure 1.2 A bigon.

The following proposition gives a simple, combinatorial condition for deciding whether or not two simple closed curves are in minimal position. It therefore gives a method for determining the geometric intersection number of two simple closed curves.

Proposition 1.7 (The bigon criterion) *Two transverse simple closed curves in a surface S are in minimal position if and only if they do not form a bigon.*

One immediate and useful consequence of the bigon criterion is the following:

> *Any two transverse simple closed curves that intersect exactly once are in minimal position.*

Before proving Proposition 1.7, we need a lemma.

Lemma 1.8 *If transverse simple closed curves α and β in a surface S do not form any bigons, then in the universal cover of S, any pair of lifts $\widetilde{\alpha}$ and $\widetilde{\beta}$ of α and β intersect in at most one point.*

Proof. Assume $\chi(S) \leq 0$, so the universal cover $\widetilde{S}$ is homeomorphic to $\mathbb{R}^2$ (the case of $\chi(S) > 0$ is an exercise). Let $p\colon \widetilde{S} \to S$ be the covering map.

Suppose the lifts $\widetilde{\alpha}$ and $\widetilde{\beta}$ of α and β intersect in at least two points. It follows that there is an embedded disk D_0 in $\widetilde{S}$ bounded by one subarc of $\widetilde{\alpha}$ and one subarc of $\widetilde{\beta}$.

By compactness and transversality, the intersection $(p^{-1}(\alpha) \cup p^{-1}(\beta)) \cap D_0$ is a finite graph if we think of the intersection points as vertices. Thus there is an *innermost disk*, that is, an embedded disk D in $\widetilde{S}$ bounded by one arc of $p^{-1}(\alpha)$ and one arc of $p^{-1}(\beta)$ and with no arcs of $p^{-1}(\alpha)$ or $p^{-1}(\beta)$ passing through the interior of the D (see Figure 1.3). Denote the two vertices of D by v_1 and v_2, and the two edges of D by $\widetilde{\alpha}_1$ and $\widetilde{\beta}_1$.

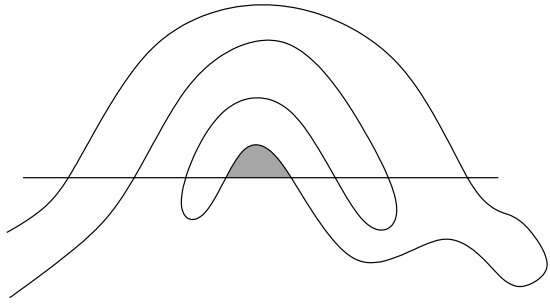

Figure 1.3 An innermost disk between two lifts.

We first claim that the restriction of p to ∂D is an embedding. The points v_1 and v_2 certainly map to distinct points in S since $\widetilde{\alpha}$ and $\widetilde{\beta}$ intersect with opposite orientations at these points. If a point of $\widetilde{\alpha}_1$ and a point of $\widetilde{\beta}_1$ have the same image in S, then both points would be an intersection of $p^{-1}(\alpha)$ with $p^{-1}(\beta)$, violating the assumption that D is innermost. If two points of $\widetilde{\alpha}_1$ (or two points of $\widetilde{\beta}_1$) map to the same point in S, then there is a lift of $p(v_1)$ between these two points, also contradicting the assumption that D is innermost.

We can now argue that D projects to an embedded disk in S. Indeed, if x and y in D project to the same point in S, then $x = \phi(y)$ for some deck transformation ϕ. Since ∂D embeds under the covering map, $\phi(\partial D) \cap \partial D$ is either empty or all of ∂D (in the case that ϕ is the identity). By the Jordan curve theorem, we then see that either $\phi(D)$ or $\phi^{-1}(D)$ must be contained in D. Now, by the Brouwer fixed point theorem, ϕ has a fixed point, which is a contradiction unless ϕ is the identity. □

We give two proofs of the bigon criterion. One proof uses hyperbolic geometry, and one proof uses only topology. We give both proofs since each of the techniques will be important later in this book.

First proof of Proposition 1.7. First suppose that two curves α and β form a bigon. It should be intuitive that there is a homotopy of α that reduces its intersection with β by 2, but here we provide a formal proof. We can

choose a small closed neighborhood of this bigon that is homeomorphic to a disk, and so the intersection of $\alpha \cup \beta$ with this disk looks like Figure 1.2. More precisely, the intersection of $\alpha \cup \beta$ with this closed disk consists of one subarc α' of α and one subarc β' of β intersecting in precisely two points. Since the disk is simply connected and since the endpoints of α' lie on the same side of β', we may modify α by a homotopy in the closed disk so that, inside this disk, α and β are disjoint. This implies that the original curves were not in minimal position.

For the other direction, we treat only the case $\chi(S) < 0$. The case $\chi(S) = 0$ is similar, and the case $\chi(S) > 0$ is easy. Assume that simple closed curves α and β form no bigons. Let $\widetilde{\alpha}$ and $\widetilde{\beta}$ be nondisjoint lifts of α and β. By Lemma 1.8, $\widetilde{\alpha}$ intersects $\widetilde{\beta}$ in exactly one point x.

It cannot be that the axes of the hyperbolic isometries corresponding to $\widetilde{\alpha}$ and $\widetilde{\beta}$ share exactly one endpoint at $\partial\mathbb{H}^2$ because this would violate the discreteness of the action of $\pi_1(S)$ on $\mathbb{H}^2$; indeed, in this case the commutator of these isometries is parabolic and the conjugates of this parabolic isometry by either of the original hyperbolic isometries have arbitrarily small translation length. Further, these axes cannot share two endpoints on $\partial\mathbb{H}^2$, for then the corresponding hyperbolic isometries would have the same axis, and so they would have to have a common power ϕ (otherwise the action of $\pi_1(S)$ on this axis would be nondiscrete). But then $\phi^n(x)$ would be an intersection point between $\widetilde{\alpha}$ and $\widetilde{\beta}$ for each n.

We conclude that any lift of α intersects any lift of β at most once and that any such lifts have distinct endpoints on $\partial\mathbb{H}^2$. But we can now see that there is no homotopy that reduces intersection. Indeed, if $\widetilde{\alpha}$ is a particular lift of α, then each fundamental domain of $\widetilde{\alpha}$ intersects the set of lifts of β in $|\alpha \cap \beta|$ points. Now, any homotopy of β changes this π_1-equivariant picture in an equivariant way, so since the lifts of α and β are already intersecting minimally in $\mathbb{H}^2$, there is no homotopy that reduces intersection. □

Second proof of Proposition 1.7. We give a different proof that two curves not in minimal position must form a bigon. Let α and β be two simple closed curves in S that are not in minimal position and let $H : S^1 \times [0,1] \to S$ be a homotopy of α that reduces intersection with β (this is possible by the definition of minimal position). We may assume without loss of generality that α and β are transverse and that H is transverse to β (in particular, all maps are assumed to be smooth). Thus the preimage $H^{-1}(\beta)$ in the annulus $S^1 \times [0,1]$ is a 1-submanifold.

There are various possibilities for a connected component of $H^{-1}(\beta)$: it could be a closed curve, an arc connecting distinct boundary components, or an arc connecting one boundary component to itself. Since H reduces the

intersection of α with β, there must be at least one component δ connecting $S^1 \times \{0\}$ to itself. Together with an arc δ' in $S^1 \times \{0\}$, the arc δ bounds a disk Δ in $S^1 \times [0,1]$. Now, $H(\delta \cup \delta')$ is a closed curve in S that lies in $\alpha \cup \beta$. This closed curve is null homotopic—indeed, $H(\Delta)$ is the null homotopy. It follows that $H(\delta \cup \delta')$ lifts to a closed curve in the universal cover $\widetilde{S}$; what is more, this lift has one arc in a lift of α and one arc in a lift of β. Thus these lifts intersect twice, and so Lemma 1.8 implies that α and β form a bigon. □

Geodesics are in minimal position. Note that if two geodesic segments on a hyperbolic surface S together bounded a bigon, then, since the bigon is simply connected, one could lift this bigon to the universal cover $\mathbb{H}^2$ of S. But this would contradict the fact that the geodesic between any two points of $\mathbb{H}^2$ is unique. Hence by Proposition 1.7 we have the following.

Corollary 1.9 *Distinct simple closed geodesics in a hyperbolic surface are in minimal position.*

The bigon criterion gives an algorithmic answer to the question of how to find representatives in minimal position: given any pair of transverse simple closed curves, we can remove bigons one by one until none remain and the resulting curves are in minimal position. Corollary 1.9, together with Proposition 1.3, gives a qualitative answer to the question.

Multicurves. A *multicurve* in S is the union of a finite collection of disjoint simple closed curves in S. The notion of intersection number extends directly to multicurves. A slight variation of the proof of the bigon criterion (Proposition 1.7) gives a version of the bigon criterion for multicurves: two multicurves are in minimal position if and only if no two component curves form a bigon.

Proposition 1.3 and Corollary 1.9 together have the consequence that, given any number of distinct homotopy classes of essential simple closed curves in S, we can choose a single representative from each class (e.g. the geodesic) so that each pair of curves is in minimal position.

1.2.5 HOMOTOPY VERSUS ISOTOPY FOR SIMPLE CLOSED CURVES

Two simple closed curves α and β are *isotopic* if there is a homotopy

$$H : S^1 \times [0,1] \to S$$

from α to β with the property that the closed curve $H(S^1 \times \{t\})$ is simple for each $t \in [0,1]$.

In our study of mapping class groups, it will often be convenient to think about isotopy classes of simple closed curves instead of homotopy classes. One way to explain this is as follows. If $H : S^1 \times I \to S$ is an isotopy of simple closed curves, then the pair $(S, H(S^1 \times \{t\}))$ "looks the same" for all t (cf. Section 1.3).

When we appeal to algebraic topology for the existence of a homotopy, the result is in general not an isotopy. We therefore want a method for converting homotopies to isotopies whenever possible.

We already know $i(a, b)$ is realized by geodesic representatives of a and b. Thus, in order to apply the above results on geometric intersection numbers to isotopy classes of curves, it suffices to prove the following fact originally due to Baer.

Proposition 1.10 *Let α and β be two essential simple closed curves in a surface S. Then α is isotopic to β if and only if α is homotopic to β.*

Proof. One direction is vacuous since an isotopy is a homotopy. So suppose that α is homotopic to β. We immediately have that $i(\alpha, \beta) = 0$. By performing an isotopy of α, we may assume that α is transverse to β. If α and β are not disjoint, then by the bigon criterion they form a bigon. A bigon prescribes an isotopy that reduces intersection. Thus we may remove bigons one by one by isotopy until α and β are disjoint.

In the remainder of the proof, we assume $\chi(S) < 0$; the case $\chi(S) = 0$ is similar, and the case $\chi(S) > 0$ is easy. Choose lifts $\widetilde{\alpha}$ and $\widetilde{\beta}$ of α and β that have the same endpoints in $\partial\mathbb{H}^2$. There is a hyperbolic isometry ϕ that leaves $\widetilde{\alpha}$ and $\widetilde{\beta}$ invariant and acts by translation on these lifts. As $\widetilde{\alpha}$ and $\widetilde{\beta}$ are disjoint, we may consider the region R between them. The quotient $R' = R/\langle\phi\rangle$ is an annulus; indeed, it is a surface with two boundary components with an infinite cyclic fundamental group. A priori, the image R'' of R in S is a further quotient of R'. However, since the covering map $R' \to R''$ is single-sheeted on the boundary, it follows that $R' \approx R''$. The annulus R'' between α and β gives the desired isotopy. □

1.2.6 Extension of Isotopies

An isotopy of a surface S is a homotopy $H : S \times I \to S$ so that, for each $t \in [0, 1]$, the map $H(S, t) : S \times \{t\} \to S$ is a homeomorphism. Given an isotopy between two simple closed curves in S, it will often be useful to promote this to an isotopy of S, which we call an *ambient isotopy* of S.

Proposition 1.11 *Let S be any surface. If $F : S^1 \times I \to S$ is a smooth isotopy of simple closed curves, then there is an isotopy $H : S \times I \to S$ so that $H|_{S\times 0}$ is the identity and $H|_{F(S^1\times 0)\times I} = F$.*

Proposition 1.11 is a standard fact from differential topology. Suppose that the two curves are disjoint. To construct the isotopy, one starts by finding a smooth vector field that is supported on a neighborhood of the closed annulus between the two curves and that carries one curve to the other. One then obtains the isotopy of the surface S by extending this vector field to S and then integrating it. For details of this argument see, e.g., [95, Chapter 8, Theorem 1.3].

1.2.7 ARCS

In studying surfaces via their simple closed curves, we will often be forced to think about arcs. For instance, many of our inductive arguments involve cutting a surface along some simple closed curve in order to obtain a "simpler" surface. Simple closed curves in the original surface either become simple closed curves or collections of arcs in the cut surface. Much of the discussion about curves carries over to arcs, so here we take a moment to highlight the necessary modifications.

We first pin down the definition of an arc. This is one place where marked points are more convenient than punctures. So assume S is a compact surface, possibly with boundary and possibly with finitely many marked points in the interior. Denote the set of marked points by $\mathcal{P}$.

A *proper arc* in S is a map $\alpha : [0,1] \to S$ such that $\alpha^{-1}(\mathcal{P} \cup \partial S) = \{0,1\}$. As with curves, we usually identify an arc with its image; in particular, this makes an arc an unoriented object. The arc α is *simple* if it is an embedding on its interior. The homotopy class of a proper arc is taken to be the homotopy class within the class of proper arcs. Thus points on ∂S cannot move off the boundary during the homotopy; all arcs would be homotopic to a point otherwise. But there is still a choice to be made: a homotopy (or isotopy) of an arc is said to be *relative to the boundary* if its endpoints stay fixed throughout the homotopy. An arc in a surface S is *essential* if it is neither homotopic into a boundary component of S nor a marked point of S.

The bigon criterion (Proposition 1.7) holds for arcs, except with one extra subtlety illustrated in Figure 1.4. If we are considering isotopies relative to the boundary, then the arcs in the figure are in minimal position, but if we are considering general isotopies, then the half-bigon shows that they are not in minimal position.

Corollary 1.9 (geodesics are in minimal position) and Proposition 1.3 (existence and uniqueness of geodesic representatives) work for arcs in surfaces with punctures and/or boundary. Here we switch back from marked points to punctures to take advantage of hyperbolic geometry. Proposition 1.10 (homotopy versus isotopy for curves) and Theorem 1.13 (extension of iso-

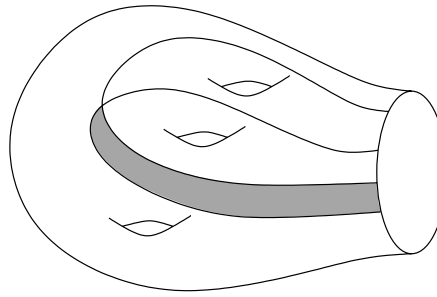

Figure 1.4 The shaded region is a half-bigon.

topies) also work for arcs.

1.3 THE CHANGE OF COORDINATES PRINCIPLE

We now describe a basic technique that is used quite frequently in the theory of mapping class groups, often without mention. We call this technique the *change of coordinates principle*. One example of this principle is that, in order to prove a topological statement about an arbitrary nonseparating simple closed curve, we can prove it for any specific simple closed curve. We will see below that this idea applies to any configuration of simple closed curves that is given by topological data.

1.3.1 CLASSIFICATION OF SIMPLE CLOSED CURVES

As a prelude to our explanation of the change of coordinates principle, we present a classification of simple closed curves in a surface.

We first need to introduce an essential concept. Given a simple closed curve α in a surface S, the surface obtained by *cutting* S along α is a compact surface S_α equipped with a homeomorphism h between two of its boundary components so that

1. the quotient $S_\alpha/(x \sim h(x))$ is homeomorphic to S, and
2. the image of these distinguished boundary components under this quotient map is α.

It also makes sense to cut a surface with boundary or marked points along a simple proper arc; the definition is analogous. Similarly, one can cut along a finite collection of curves and arcs. There are several distinct situations for cutting along a single arc, depending on whether the endpoints of the arc lie

on a boundary component or a puncture, for instance, and the cut surface is allowed to have marked points on its boundary.

We remark that the cutting procedure is one place where it is convenient to assume that all curves under consideration are smooth. Indeed, if γ is a smooth simple closed curve in a surface S, then the pair (S, γ) is locally diffeomorphic to $(\mathbb{R}^2, \mathbb{R})$, and one can immediately conclude that the surface obtained from S by cutting along γ is again a surface, now with two additional boundary components. Hence the classification of surfaces can be applied to the cut surface.

We say that a simple closed curve α in the surface S is *nonseparating* if the cut surface S_α is connected. We claim the following.

> *If α and β are any two nonseparating simple closed curves in a surface S, then there is a homeomorphism $\phi : S \to S$ with $\phi(\alpha) = \beta$.*

In other words, up to homeomorphism, there is only one nonseparating simple closed curve in S. This statement follows from the classification of surfaces, as follows. The cut surfaces S_α and S_β each have two boundary components corresponding to α and β, respectively. Since S_α and S_β have the same Euler characteristic, number of boundary components, and number of punctures, it follows that S_α is homeomorphic to S_β. We can choose a homeomorphism $S_\alpha \to S_\beta$ that respects the equivalence relations on the distinguished boundary components. Such a homeomorphism gives the desired homeomorphism of S taking α to β. If we want an orientation-preserving homeomorphism, we can ensure this by postcomposing by an orientation-reversing homeomorphism fixing β if necessary.

A simple closed curve β is *separating* in S if the cut surface S_β is not connected. Note that when S is closed, β is separating if and only if it is the boundary of some subsurface of S. This is equivalent to the vanishing of the homology class of β in $H_1(S, \mathbb{Z})$. By the "classification of disconnected surfaces," we see that there are finitely many separating simple closed curves in S up to homeomorphism.

The above arguments give the following general classification of simple closed curves on a surface:

> *There is an orientation-preserving homeomorphism of a surface taking one simple closed curve to another if and only if the corresponding cut surfaces (which may be disconnected) are homeomorphic.*

The existence of such a homeomorphism is clearly an equivalence relation. The equivalence class of a simple closed curve or a collection of simple

closed curves is called its *topological type*. For example, a separating simple closed curve in the closed surface S_g divides S_g into two disjoint subsurfaces of, say, genus k and $g-k$. The minimum of $\{k, g-k\}$ is called the *genus* of the separating simple closed curve. By the above, the genus of a curve determines and is determined by its topological type. Note that there are $\lfloor \frac{g}{2} \rfloor$ topological types of essential separating simple closed curves in a closed surface.

The uninitiated may have trouble visualizing separating simple closed curves that are not the obvious ones. We present a few in Figure 1.5, and we encourage the reader to draw even more complicated separating simple closed curves.

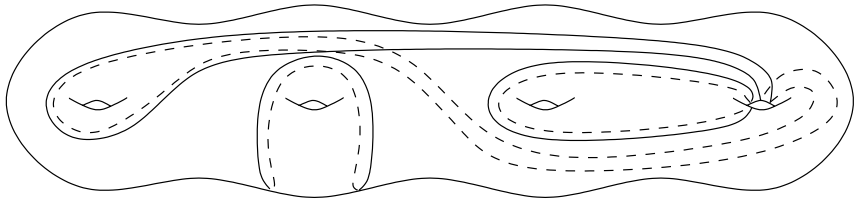

Figure 1.5 Some nonobvious separating simple closed curves.

1.3.2 THE CHANGE OF COORDINATES PRINCIPLE

The change of coordinates principle is a kind of change of basis for curves in a surface S. It roughly states that any two collections of simple closed curves in S with the same intersection pattern can be taken to each other via an orientation-preserving homeomorphism of S. In this way an arbitrary configuration can be transformed into a standard configuration. The classification of simple closed curves in surfaces given above is the simplest example.

We illustrate the principle with two sample questions. Suppose α is *any* nonseparating simple closed curve α on a surface S.

1. Is there a simple closed curve γ in S so that α and γ *fill* S, that is, α and γ are in minimal position and the complement of $\alpha \cup \gamma$ is a union of topological disks?

2. Is there a simple closed curve δ in S with $i(\alpha, \delta) = 0$? $i(\alpha, \delta) = 1$? $i(\alpha, \delta) = k$?

Even for the genus 2 surface S_2, it is not immediately obvious how to answer either question for the nonseparating simple closed curve α shown

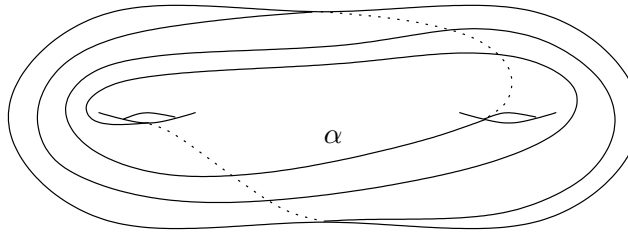

Figure 1.6 A simple closed curve on a genus 2 surface.

in Figure 1.6. However, we claim that Figure 1.7 gives proof that the answer to the first question is yes in this case, as we now show. The curves β and γ in Figure 1.7 fill the surface (check this!). By the classification of simple closed curves in a surface, there is a homeomorphism $\phi : S_2 \to S_2$ with $\phi(\beta) = \alpha$. Since filling is a topological property, it follows that $\phi(\gamma)$ is the curve we are looking for since it together with $\alpha = \phi(\beta)$ fills S_2.

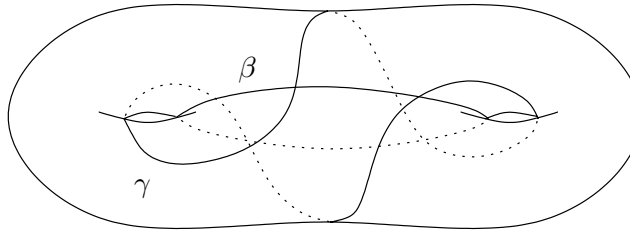

Figure 1.7 Two simple closed curves that fill a genus 2 surface.

We think of ϕ as changing coordinates so that the complicated curve α becomes the easy-to-see curve β. The second question can be answered similarly.

1.3.3 Examples of the Change of Coordinates Principle

The change of coordinates principle applies to more general situations. We give several examples here. Most of the proofs are minor variations of the above arguments and so are left to the reader.

1. Pairs of simple closed curves that intersect once. Suppose that α_1 and β_1 form such a pair in a surface S. Let S_{α_1} be the surface obtained by cutting S along α_1. There are two boundary components of S_{α_1} corresponding to the two sides of α_1. The image of β_1 in S_{α_1} is a simple arc connecting these boundary components to each other. We can cut S_{α_1} along this arc to obtain

a surface $(S_{\alpha_1})_{\beta_1}$. The latter is a surface with one boundary component that is naturally subdivided into four arcs—two coming from α_1 and two coming from β_1. The equivalence relation coming from the definition of a cut surface identifies these arcs in order to recover the surface S with its curves α_1 and β_1.

If α_2 and β_2 are another such pair, there is an analogous cut surface $(S_{\alpha_2})_{\beta_2}$. By the classification of surfaces, $(S_{\alpha_2})_{\beta_2}$ is homeomorphic to $(S_{\alpha_1})_{\beta_1}$, and moreover there is a homeomorphism that preserves equivalence classes on the boundary. Any such homeomorphism descends to a homeomorphism of S taking the pair $\{\alpha_1, \beta_1\}$ to the pair $\{\alpha_2, \beta_2\}$.

2. Bounding pairs of a given genus. A *bounding pair* is a pair of disjoint, homologous, nonseparating simple closed curves in a closed surface. Figure 1.8 shows one example, but we again encourage the reader to find more complicated examples. The genus of a bounding pair in a closed surface is defined similarly to the genus of a separating simple closed curve.

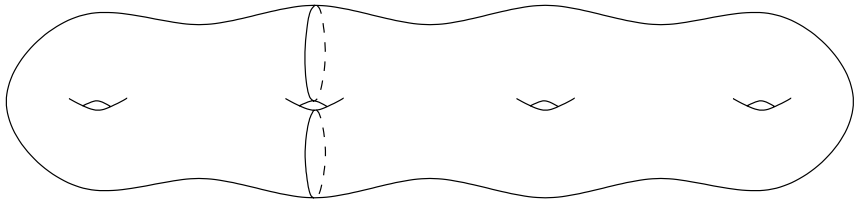

Figure 1.8 A genus 1 bounding pair.

3. Pairs (or k-tuples) of disjoint simple closed curves whose union does not separate.

4. Pairs of simple closed curves $\{\alpha, \beta\}$ with $i(\alpha, \beta) = |\alpha \cap \beta| = 2$ and $\hat{i}(\alpha, \beta) = 0$ and whose union does not separate.

5. Nonseparating simple proper arcs in a surface S that meet the same number of components of ∂S.

6. Chains of simple closed curves. A *chain* of simple closed curves in a surface S is a sequence $\alpha_1, \ldots, \alpha_k$ with the properties that $i(\alpha_i, \alpha_{i+1}) = 1$ for each i and $i(\alpha_i, \alpha_j) = 0$ whenever $|i - j| > 1$. A chain is *nonseparating* if the union of the curves does not separate the surface.

Any two nonseparating chains of simple closed curves with the same number of curves are topologically equivalent. This can be proved by induction. The starting point is the case of nonseparating simple closed curves, and the inductive step is example 5: cutting along the first few arcs, the next arc becomes a nonseparating arc on the cut surface. Note that example 1

is the case $k = 2$. One can also prove by induction that every chain in S_g of even length is nonseparating, and so such chains must be topologically equivalent.

We remark that the homeomorphism representing the change of coordinates in each of the six examples above can be taken to be orientation-preserving.

1.4 THREE FACTS ABOUT HOMEOMORPHISMS

In this subsection we collect three useful facts from surface topology. Each allows us to replace one kind of map with a better one: a homotopy of homeomorphisms can be improved to an isotopy; a homeomorphism of a surface can be promoted to a diffeomorphism; and $\mathrm{Homeo}_0(S)$ is contractible, so in particular any isotopy from the identity homeomorphism to itself is homotopic to the constant isotopy.

1.4.1 HOMOTOPY VERSUS ISOTOPY FOR HOMEOMORPHISMS

When are two homotopic homeomorphisms isotopic? Let us look at two of the simplest examples: the closed disk D^2 and the closed annulus A. On D, any orientation-reversing homeomorphism f induces a degree -1 map on $S^1 = \partial D^2$, and from this follows that f is not isotopic to the identity. However, the straight-line homotopy gives a homotopy between f and the identity. On $A = S^1 \times I$, the orientation-reversing map that fixes the S^1 factor and reflects the I factor is homotopic but not isotopic to the identity.

It turns out that these two examples are the only examples of homotopic homeomorphisms that are not isotopic. This was proved in the 1920s by Baer using Proposition 1.10 (see [8, 9] and also [56]).

THEOREM 1.12 *Let S be any compact surface and let f and g be homotopic homeomorphisms of S. Then f and g are isotopic unless they are one of the two examples described above (on $S = D^2$ and $S = A$). In particular, if f and g are orientation-preserving, then they are isotopic.*

In fact, a stronger, relative result holds: if two homeomorphisms are homotopic relative to ∂S, then they are isotopic relative to ∂S. Theorem 1.12 can be proven using ideas from the proof of Proposition 2.8.

Theorem 1.12 also holds when S has finitely many marked points. In that case, we need to expand our list of counterexamples to include a sphere with one or two marked points.

1.4.2 Homeomorphisms versus Diffeomorphisms

It is sometimes convenient to work with homeomorphisms and sometimes convenient to work with diffeomorphisms. For example, it is easier to construct the former, but we can apply differential topology to the latter. The following theorem will allow us to pass back and forth between homeomorphisms and diffeomorphisms of surfaces.

Theorem 1.13 *Let S be a compact surface. Then every homeomorphism of S is isotopic to a diffeomorphism of S.*

It is a general fact that any homeomorphism of a smooth manifold can be approximated arbitrarily well by a smooth map. By taking a close enough approximation, the resulting smooth map is homotopic to the original homeomorphism. However, this general fact, which is easy to prove, is much weaker than Theorem 1.13 because the resulting smooth map might not be smoothly invertible; indeed, it might not be invertible at all.

Theorem 1.13 was proven in the 1950s by Munkres [167, Theorem 6.3], Smale, and Whitehead [213, Corollary 1.18]. In part, this work was prompted by Milnor's discovery of the "exotic" (nondiffeomorphic) smooth structures on S^7.

Theorem 1.13 gives us a way to replace homeomorphisms with diffeomorphisms. We can also replace isotopies with smooth isotopies. In other words, if two diffeomorphisms are isotopic, then they are smoothly isotopic; see, for example, [30].

In this book, we will switch between the topological setting and the smooth setting as is convenient. For example, when defining a map of a surface to itself (either by equations or by pictures), it is often easier to write down a homeomorphism than a smooth map. On the other hand, when we need to appeal to transversality, extension of isotopy, and so on, we will need to assume we have a diffeomorphism.

One point to make is that we will actually be forced to consider self-maps of a surface that are not smooth; pseudo-Anosov homeomorphisms, which are central to the theory, are special maps of a surface that are never smooth (cf. Chapter 13).

1.4.3 Contractibility of Components of $\mathrm{Homeo}(S)$

The following theorem was proven by Hamstrom in a series of papers [77, 78, 79] in the 1960s. In the statement, $\mathrm{Homeo}_0(S)$ is the connected component of the identity in the space of homeomorphisms of a surface S.

THEOREM 1.14 *Let S be a compact surface, possibly minus a finite number of points from the interior. Assume that S is not homeomorphic to S^2, $\mathbb{R}^2$, D^2, T^2, the closed annulus, the once-punctured disk, or the once-punctured plane. Then the space $\mathrm{Homeo}_0(S)$ is contractible.*

The fact that $\mathrm{Homeo}_0(S)$ is simply connected is of course an immediate consequence of Theorem 1.14. This fact will be used, among other places, in Section 4.2 in the proof of the Birman exact sequence. There is a smooth version of Theorem 1.14; see [53] or [73].