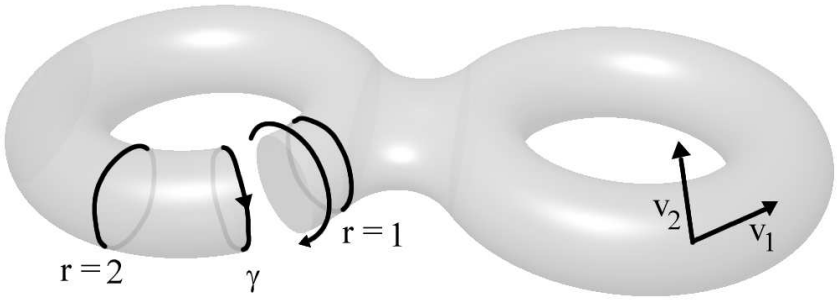
Surface automorphisms and the Nielsen realization problem

Maxim Hendriks



Doctoraalscriptie (Master's thesis)
Defended on January 15, 2007
Supervisors:
Martin Lübke (Universiteit Leiden)
Hansjörg Geiges (Universität zu Köln)

Mathematisch Instituut, Universiteit Leiden

Front illustration: a Dehn twist (see section 3.4)
Printed in 12pt Times New Roman
Available electronically from http://www.math.leidenuniv.nl/nieuw/en/theses/

Contents

superficial adj. *ˈsü-pə(r)-ˈfi-shəl* [from Latin *superficies*]
(1) of, or relating to, or located near a surface
(2) concerned only with the obvious or apparent: shallow

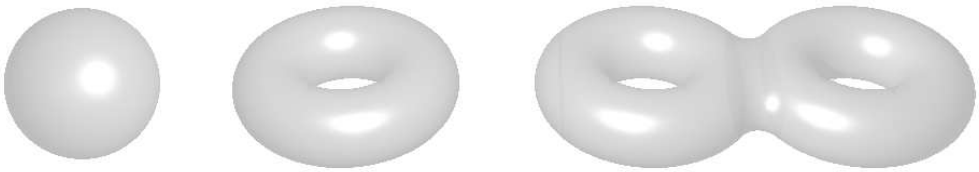
[Merriam-Webster]

1 Introduction

A mathematician is interested in many different abstract objects. With these objects often come maps between them, *morphisms* in category-theoretical terms. The invertible morphisms of an object to itself, its *automorphisms*, are interesting objects of study in themselves. Surfaces form one of the most intuitively accessible collections of mathematical objects. Since they are 2-dimensional, we may fully apply our imagination to them. In this thesis, we wish to study the automorphisms of a given surface. This can be done using only point-set topology, but other structure is often brought in, creating a mixture of point-set and differential topology as well as hyperbolic geometry. After introducing some fundamental concepts in section 1, we will study isotopies between automorphisms, mainly in the differentiable category, in section 2. In section 3 we will look at the function spaces of automorphisms and define the *mapping class group* of a surface (MCG for short), discussing some basic results about it. Section 4 gives a short overview of some hyperbolic geometry. Then in sections 5 and 6 we will present the (extended) *Nielsen realization problem*, asking how we can realize mapping classes by concrete automorphisms while respecting the group structure, and gather the results that have been obtained on this problem. Finally, section 7 will provide two representations of the mapping class group.

1.1 The basic objects

To begin with, we will not be working with all surfaces. For the sake of simplicity, we will restrict ourselves to *connected closed orientable surfaces*. It so happens that these are surprisingly easily classified up to homeomorphism. A nice presentation of this result can be found in Massey [28] or Munkres [34]. We can characterize all these surfaces by a single number, their *genus* $g = 1 - \frac{1}{2}\chi$, which tells us "how many holes they have". Here, χ is the Euler characteristic of the surface. We call the surface with g holes Σ_g. The first few are depicted below to get clear on the idea.



Σ_0 Σ_1 Σ_2

etc.

Σ_3

Figure 1. Connected closed orientable surfaces.

Furthermore, we have to decide on the category we want to work in. We could choose to use only the topology of the surface and view it as a topological manifold, or we could introduce a piecewise linear structure (we get a so-called PL-manifold) or a $\mathcal{C}^r$-differentiable structure for some $r \in \mathbb{N} \cup \{\infty\}$. For a given manifold the following implications hold:

$$\mathcal{C}^r\text{-differentiable} \Longrightarrow \mathrm{PL} \Longrightarrow \text{topological}.$$

These implications can not in general be reversed. In dimensions ≥ 4 there are topological manifolds which can not be given any PL-structure. Or a given topological manifold can be equipped with more than one possible PL or differentiable structure up to isomorphism in the PL or differentiable category respectively. Luckily, these problems do not arise for surfaces. It was proved early in the twentieth century that every surface possesses a unique PL and $\mathcal{C}^\infty$-differentiable structure, up to isomorphism. A consequence was also that a surface is triangulable and any two triangulations have a common refinement.

Though this structure is canonical, different categories have different sets of (auto)morphisms, so the choice of category is still important. According to what category of surface we choose, we distinguish between the collections of *homeo*morphisms $\mathrm{Homeo}(\Sigma_g)$, PL-homeomorphisms $\mathrm{PL}(\Sigma_g)$ and *diffeo*morphisms $\mathrm{Diff}^r(\Sigma_g)$ of a surface to itself, although in the last case I will mainly use $r = \infty$ without indication. In principle I will continue to use the word *automorphism* when I wish to remain indeterminate on the category under discussion.

A further possibility would be to look at a *complex structure* (also called *conformal structure*) on our surfaces. However, a surface admits many complex structures, as evinced already by the classical modular problem for the torus. The problem of looking at the auto-biholomorphism group of a given Riemann surface has a different flavor than in the above-mentioned categories, and we will not go into it. Some special kinds of conformal structure will be discussed in section 4, though.

Two important concepts that will star throughout our discussion are homotopy and isotopy. Given topological spaces X, Y, two maps $f, g : X \to Y$ are said to be *homotopic* if there is a map $F : X \times I \to Y$ such that $F(\cdot, 0) = f$ and $F(\cdot, 1) = g$. If moreover f and g are homeomorphisms onto their images, they will be called *isotopic* if there exists a homotopy $F : X \times I \to Y$ between them such that for every $t \in [0, 1]$, $F(\cdot, t) : X \to Y$ is also a homeomorphism onto its image. For differentiable maps f and g, we will speak of a differentiable isotopy if the map F is differentiable. The set of homotopy classes of morphisms from X to Y is often denoted by $[X, Y]$.

1.2 Applications

The theory of automorphisms of surfaces is important when studying 3-dimensional manifolds. An important tool in classifying these is a Heegaard splitting. This means separating the manifold into two handlebodies (the closed orientable surfaces imbedded in 3-space as shown above, together with the part of space bounded by it). It turns out that every closed orientable 3-manifold has a Heegaard splitting. So to classify these manifolds, it is useful to know how one can construct new 3-manifolds by glueing two handlebodies of the same genus together along their boundary, which is a closed orientable surface. This boils down to studying how you can map a surface to itself, which is what we are concerned with. In classifying surface bundles over a given manifold, the situation is similar: we need information on what automorphisms the fiber has. And in the field of string theory, surface automorphisms also seem to arise.

1.3 Algebraic information on surfaces

An algebraic object that is frequently used in the study of topological spaces is the *first homotopy group* or *fundamental group* of a surface, $\pi_1(\Sigma_g)$. There are also higher homotopy groups (i.e. $\pi_2, \pi_3, \pi_4, \ldots$) but we will have no need for them. To be precise, the fundamental group is defined using a base point as

$$\pi_1(\Sigma_g, p) := [(I, \partial I), (\Sigma_g, p)].$$

This is the group of homotopy classes of maps $(I, \partial I) \to (\Sigma_g, p)$, meaning such a map is from $I \to \Sigma_g$ and maps $\partial I = \{0, 1\}$ to p. The two groups $\pi_1(\Sigma_g, p_1)$ and $\pi_1(\Sigma_g, p_2)$ obtained by using two different base points are isomorphic. An isomorphism is given by conjugating a loop based at p_2 by a path from p_1 to p_2. Such an isomorphism is not canonical, however, since there is no canonical path from p_1 to p_2, not even up to homotopy. The fundamental group of a closed orientable surface can be calculated using a cut-and-paste diagram; this is a polygon together with its interior, of which the surface is a quotient space by identifying edges in pairs. The fundamental group can actually be used in this way to prove the classification of closed surfaces (see Munkres [34] chapter 12).

It turns out that $\pi_1(\Sigma_g, p)$ is generated by the homotopy classes of loops $\alpha_1, \beta_1, \ldots, \alpha_g, \beta_g$, as illustrated below for Σ_3, and that an explicit group presentation is given by

$$\pi_1(\Sigma_g, p) \cong \langle \alpha_1, \beta_1, \ldots, \alpha_g, \beta_g \mid [\alpha_1, \beta_1] \cdot [\alpha_2, \beta_2] \cdots [\alpha_g, \beta_g] = 1 \rangle.$$

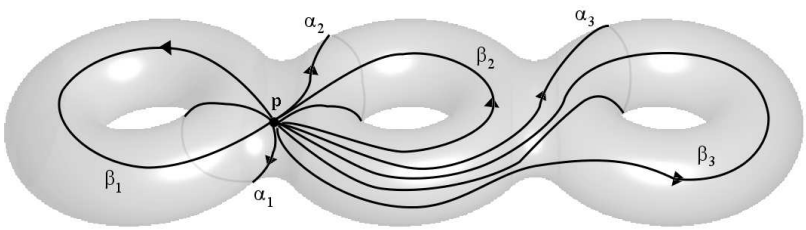


Figure 2. Loops generating $\pi_1(\Sigma_3, p)$.

Another useful set of algebraic objects for any topological space X is the collection of its *homology groups* $H_0(X), H_1(X), H_2(X), \ldots$. We will not explain the definition of a homology group; Bredon [5] gives a good introduction to homology theory. For a closed orientable surface Σ_g, $H_n(\Sigma_g) = 0$ for $n \geq 3$, and $H_2(\Sigma_g) \cong H_0(\Sigma_g) \cong \mathbb{Z}$. The most interesting homology group for Σ_g is the first homology group $H_1(\Sigma_g)$. This group is canonically isomorphic to the abelianized fundamental group (for any topological space). For Σ_g, we may regard the loops α_i and β_i generating $\pi_1(\Sigma_g, p)$ as 1-cycles representing homology classes. These homology classes therefore generate $H_1(\Sigma_g)$. Also, the relation in the above presentation of $\pi_1(\Sigma_g)$ becomes trivial when we allow the elements to commute (which is by definition the case if we work in the abelianized fundamental group). Thus we see that

$$H_1(\Sigma_g) \cong \mathbb{Z}^{2g}$$

and that $(\alpha_1, \beta_1, \ldots, \alpha_g, \beta_g)$ can function as a basis for $H_1(\Sigma_g)$, viewing this group as a $\mathbb{Z}$-module. This situation is pictured below, (ab)using the same letters as above, this time for homology classes.

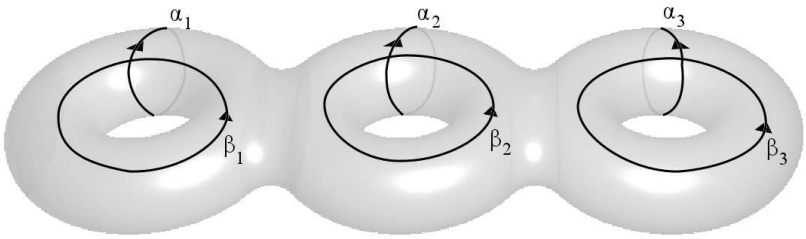


Figure 3. Homology classes generating $H_1(\Sigma_3)$.

1.4 Notation

We use the notations $H \subseteq G$, $H < G$ and $H \lhd G$ when H is a subset, subgroup and normal subgroup of G, respectively.

Given maps $f, g : X \to Y$, let us write $f \simeq_h g$ if they are homotopic and $f \simeq_i g$ if they are isotopic. If these maps are C^1-differentiable, $f \pitchfork g$ will mean f and g are transverse maps. This is to say that for all $x_1, x_2 \in X$ such that $f(x_1) = g(x_2) =: y$, we have the (not necessarily direct) sum

$$T_{x_1} f(T_{x_1} X) + T_{x_2} g(T_{x_2} X) = T_y(Y).$$

A closed curve will be a map $\gamma : S^1 \to \Sigma_g$ or a map $\gamma : I \to \Sigma_g$ with $\gamma(0) = \gamma(1)$. These definitions will be used interchangeably in the way which suits the application best. A (closed) curve will be called *simple* if the map $S^1 \to \Sigma_g$ is injective. If γ is a path from p to q (that is, $\gamma(0) = p$ and $\gamma(1) = q$) and δ is a path from q to r, the concatenation of these paths will be written as $\gamma * \delta$ and defined by

$$\gamma * \delta(t) := \begin{cases} \gamma(2t) & \text{for } 0 \leq t \leq \frac{1}{2} \\ \delta(2t-1) & \text{for } \frac{1}{2} \leq t \leq 1 \end{cases}.$$

For a path γ from p to q, the reverse path from q to p will be written as γ_{rev} and defined by

$$\gamma_{\text{rev}}(t) := \gamma(1-t).$$

We will denote the closed unit disc by D^2.

2 Curves on surfaces and isotopies between automorphisms

This section leads up to a basic result about the automorphisms of closed orientable surfaces: homotopic automorphisms are isotopic. In the process, we get a very tangible criterion to decide when two surface automorphisms are isotopic. Also, the road to this result showcases a few nice techniques in surface topology. I feel, therefore, it is a good starting point for our investigation. Most of the lemmata in this section can be found in Casson and Bleiler [7], Epstein [11] and Stillwell [40] in some form or other, though only Epstein works towards the same conclusion, and does so in the PL category. (I have not been able to find a source which does the proof consciously and correctly in the differentiable category!) Readers might like to compare the versions of several lemmata and their proofs using differential topology (here) as opposed to hyperbolic geometry (Casson and Bleiler).

Remark. Whether the main result holds for a larger collection of manifolds than just surfaces seems to be largely an open problem. In the literature, I could only find an analogous result for (3-dimensional) Seifert fibered space, and nothing for higher dimensions. As will be clear, the techniques used in the present section would not be applicable to higher dimensions.

For our treatment we will use the tools of transversality and tubular neighbourhoods. Thus we are actually obliged to work in the differentiable category. Without special mention, all maps in this section will thus be $\mathcal{C}^\infty$-differentiable. At the end this condition will be dispensed with to obtain the main result in the topological category. We start by introducing two seemingly unconnected notions for curves on surfaces, both illustrated in the figure below.

Definition 2.1 *The* minimal intersection number *of two simple closed curves γ_1, γ_2 on a surface is defined as $I_{\min}(\gamma_1, \gamma_2) := \min\{|\delta_1 \cap \delta_2| : \delta_1 \simeq_h \gamma_1, \delta_2 \simeq_h \gamma_2\}$. A set of curves $\{\gamma_1, \ldots, \gamma_n\}$ is said to* have minimal intersection *when $|\gamma_i \cap \gamma_j| = I_{\min}(\gamma_i, \gamma_j)$ for all $1 \leq i < j \leq n$. A* 2-gon between γ_1 and γ_2 *is a disc D embedded in the surface such that ∂D consists of two arcs that are part of γ_1 and γ_2, respectively.*

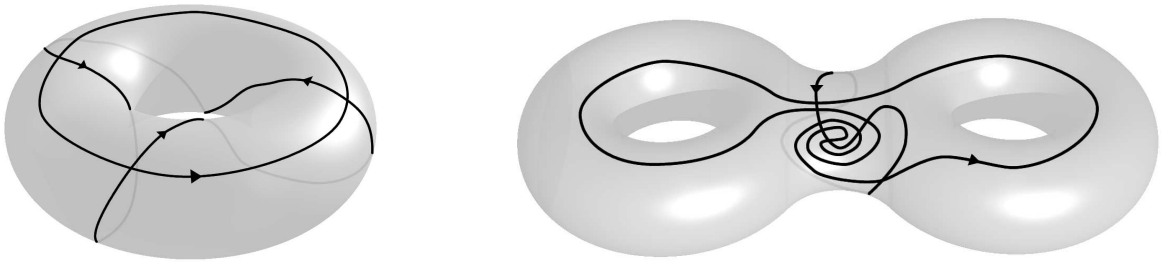


Figure 4. The two curves on Σ_1 shown to the left have minimal intersection, while the two curves on Σ_2 shown to the right clearly do not. We also see an embedded 2-gon there.

Surprisingly, these two notions are related.

Lemma 2.2 *Two smooth transverse simple closed curves γ and δ on any surface Σ have minimal intersection if and only if there is no embedded 2-gon D between them with* $\mathrm{int}(D) \cap (\gamma \cup \delta) = \emptyset$.

Remark. Transversality is essential here. First of all, non-transverse curves could coincide on an interval or touch but not cross, not bounding a disc in either situation. Second, a worse thing might occur: the curve δ could spiral towards a point p on γ, crossing γ an infinite number of times before reaching p, and then continue in a similar fashion by spiraling outwards, crossing γ an infinite number of times again. Any 2-gon between γ and δ would contain a smaller one, as can be seen from the image. In contrast, transverse curves have a discrete set of intersections points. Because the curves are closed, their number must be finite.

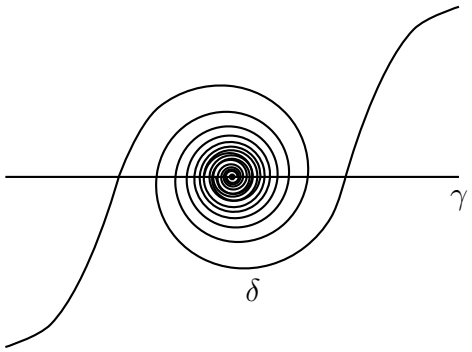


Figure 5. An infi nite spiraling family of intersections.

Proof. ($\Longrightarrow$) Suppose there was an embedded 2-gon of the required kind between the curves γ and δ. Then it is clear we could construct a homotopy for one of the curves which would dissolve these intersections, as shown below. Therefore the curves would not have minimal intersection.

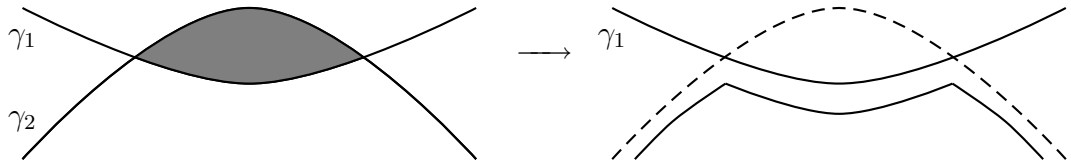


Figure 6. Dissolving the grey 2-gon shown on the left between the curves γ and δ.

($\Longleftarrow$) We take the nice contrapositive proof from Hass & Scott [16, lemma 3.1]. Consider the universal cover $\pi : \tilde{\Sigma} \to \Sigma$. We know that $\tilde{\Sigma}$ is either S^2 or $\mathbb{R}^2$ (see Epstein's corollary 1.8). We prove there are lifts $\tilde{\gamma}, \tilde{\delta} : \mathbb{R} \to \tilde{\Sigma}$ of γ, δ between which there is a 2-gon. By the Jordan curve theorem, it is enough to prove these lifts have at least two intersection points. When $\tilde{\Sigma} \cong S^2$, this fact is clear because there are lifts that intersect, and these must intersect an even number of times, since any lift to S^2 of a simple closed curve yields a periodic curve. When $\tilde{\Sigma} \cong \mathbb{R}^2$, the reasoning is the same if one of the lifts is periodic. Otherwise, view γ as a curve from $I \to \Sigma$ with $\gamma(0) = \gamma(1) \notin \gamma \cap \delta$ and lift it to $\tilde{\gamma} : I \to \tilde{\Sigma}$. This lift meets every fiber over $\gamma \cap \delta$ exactly once, so $|\gamma \cap \delta| = |\tilde{\gamma} \cap \pi^{-1}(\delta)|$. Suppose $\tilde{\gamma}$ intersects every component of $\pi^{-1}(\delta)$ only once.

Any loop $\gamma' \simeq_h \gamma$ is homotopic to a loop γ'' through $\gamma(0)$ with $|\gamma'' \cap \delta| = |\gamma' \cap \delta|$. And γ'' lifts to a curve from $\tilde{\gamma}(0)$ to $\tilde{\gamma}(1)$. Because all components of $\pi^{-1}(\delta)$ separate $\mathbb{R}^2$, this lift must intersect each component that $\tilde{\gamma}$ intersects. But then $|\gamma' \cap \delta| = |\gamma'' \cap \delta| \geq |\gamma \cap \delta|$, contradicting our assumption that $|\gamma \cap \delta|$ was not minimal.

So we have proved there is an embedded 2-gon between some lifts $\tilde{\gamma}, \tilde{\delta} : \mathbb{R} \to \tilde{\Sigma}$. If its interior still contains points of, say, $\pi^{-1}(\gamma)$, then some component $\tilde{\gamma}_2$ of $\pi^{-1}(\gamma)$ must enter and exit it through $\tilde{\delta}$ — it can not intersect $\tilde{\gamma}$, since γ is simple. So then our old 2-gon contains a smaller one, between $\tilde{\gamma}_2$ and $\tilde{\delta}$. This procedure of finding smaller 2-gons can be continued only a finite number of times, since all components of $\pi^{-1}(\gamma)$ and $\pi^{-1}(\delta)$ are pairwise transverse. So there is a 2-gon B between some components $\tilde{\gamma}$ and $\tilde{\delta}$ with interior disjoint from $\pi^{-1}(\gamma) \cup \pi^{-1}(\delta)$.

We now show that B homeomorphically projects to an embedded 2-gon with interior disjoint from $\gamma \cup \delta$. Call the two vertices of our 2-gon (the points of $\tilde{\gamma} \cap \tilde{\delta} \cap B$) x and y. Suppose we have a non-trivial element $g \in \pi_1(\Sigma) \cong \mathrm{Deck}(\tilde{\Sigma}/\Sigma)$ for which $g \cdot B \cap B \neq \emptyset$. Since $\mathrm{int}(B) \cap (\tilde{\gamma} \cup \tilde{\delta}) = \emptyset$, g must map some point of ∂B to ∂B. Because the set $\{p \in \partial B | g \cdot p \in B\}$ is closed, g must map a point of $\{x, y\}$ to this same set. Deck transformations act freely, so we must have, say, $g \cdot x = y$. It follows that g fixes $\tilde{\gamma}$ and $\tilde{\delta}$ setwise, because no other components of a $\pi^{-1}(\gamma)$ and $\pi^{-1}(\delta)$ contain either x or y. Thus γ and δ both represent g as an element of $\pi_1(\Sigma)$ and intersect in a single point. On an orientable surface this is a contradiction because the mod 2 self-intersection number $I_2(\gamma, \gamma)$ is zero (see Guillemin & Pollack [15]). On a non-orientable surface, this is contrary to assumption, because the curves would have minimal intersection, since $I_{\min}(\gamma, \delta) \geq 1$ if $I_2(\gamma, \delta) = 1$. We conclude that $g \cdot B \cap B = \emptyset$ and that $\pi|_{B_n}$ is a homeomorphism onto its image. □

We can use this lemma to prove that we can disentangle two curves, so that they have minimal intersection, by means of an ambient isotopy.

Lemma 2.3 *For any two smooth simple closed curves γ_1 and γ_2 on a surface Σ there is an isotopy $J : \Sigma \times I \to \Sigma$ such that $J(\cdot, 0) = \mathrm{id}_\Sigma$ and $J(\cdot, 1) \circ \gamma_1$ has minimal intersection with γ_2.*

Proof. To start with, there exists an isotopy $J_0 : S^1 \times I \to \Sigma$ such that $J_0(\cdot, 0) = \gamma_1$ and $(J_0(\cdot, 0) \circ \gamma_1) \pitchfork \gamma_2$ (by Bredon [5], chapter II, corollary 15.6), and this may be extended to an ambient isotopy $J_1 : \Sigma \times I \to \Sigma$ such that $J_1(\cdot, 0) = \mathrm{id}_\Sigma$ and $(J_1(\cdot, 1) \circ \gamma_1) \pitchfork \gamma_2$ by the isotopy extension theorem (see Hirsch [17], chapter 8). Now suppose that the curves do not have minimal intersection. From the previous lemma, we now know that there is an embedded 2-gon between γ_1 and γ_2 with interior disjoint from these curves. It is obvious that we can find a neighbourhood of this 2-gon which looks like the left figure in the previous proof. We may then dissolve this 2-gon by means of an isotopy J_2 which is the identity outside this neighbourhood and 'pushes γ_2 across γ_1', obtaining the situation shown in the right figure above. In this way, we reduce the number of points in $\gamma_1 \cap \gamma_2$ by two. By compactness and transversality, the two curves intersect each other only a finite number of times, so this procedure can be iterated until minimal intersection is attained by the sequence of isotopies $J_1, J_2, \ldots, J_k$. □

Definition 2.4 *A closed curve on a surface is called* essential *if it is not nullhomotopic.*

The following general lemma of surface topology comes up regularly.

Lemma 2.5 *Two disjoint homotopic essential simple closed curves on a compact orientable surface Σ bound an annulus.*

Proof. We subdivide our curves and consider them as 1-chains A_0 and B_0. These are homologous, because they are homotopic (by assumption). So there is a 2-chain N_0 with $\partial N_0 = A_0 - B_0$. We may choose N_0 so that it includes any point on the surface at most once in its image by subdivision. This 2-chain forms a finite triangulation of part of the surface, which we view as a compact orientable simplicial complex with the 1-simplices of A_0 and B_0 as its two boundary components.

We now determine the Euler characteristic $\chi(N_0)$. To this end we construct a sequence of simplicial complexes $N_0, N_1, N_2, \ldots$ with corresponding boundary sequences $A_0, A_1, A_2, \ldots$ and $B_0, B_1, B_2, \ldots$ by steps of three types, applied successively in no specific order. (I) Whenever N_k contains a 2-simplex with exactly one of its sides belonging to $A_k - B_k$, we cut away this 2-simplex and named side to obtain N_{k+1}. We add the other two sides and their common point to A_k to form A_{k+1}. (II) If N_k contains a 2-simplex with exactly two of its sides belonging to $A_k - B_k$, we cut away this simplex, named sides and their common point to obtain N_{k+1}. Then we add the third side to A_k to obtain A_{k+1}. (III) And if N_k contains a 2-simplex with three of its sides in $A_k - B_k$, we cut away this simplex and all its sides from N_k, and from A_k we delete the sides and the points that do not occur in other 1-simplices of A_k. We do not alter B_k, that is we set $B_{k+1} := B_k$. The procedure is illustrated below.

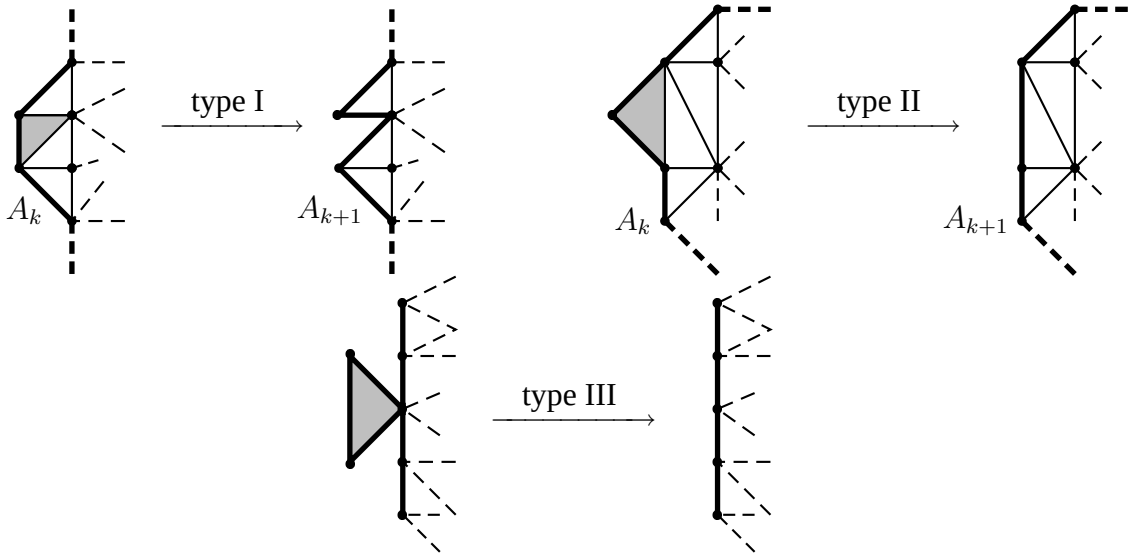


Figure 7. Changing the simplicial complex N_k to N_{k+1} by one of three kinds of alterations.

It is easy to check (for all three types of steps) that $\chi(N_{k+1}) = \chi(N_k)$ and that A_{k+1} stays a non-trivial 1-cycle. As long as there are 2-simplices in N_k, we can continue this process, because

there will be a 2-simplex with at least one side in $A_k - B_k$. As there are finitely many simplices in N_0, after a finite number of steps we will be left with a complex without 2-simplices, which must be equal to $B_0 \cong S^1$. This shows that $\chi(N_0) = \chi(S^1) = 0$.

The classification of compact orientable surfaces (see Massey [28]) tells us a surface with two boundary components and Euler characteristic zero is an annulus. □

From this lemma it now follows that we can freely manipulate (pairs of) essential simple closed curves within their homotopy classes, using ambient isotopies. In particular, homotopic essential simple closed curves on a surfaces are isotopic.

Lemma 2.6 *On an orientable surface Σ, two smooth essential simple closed curves γ, δ are homotopic if and only if they are ambient isotopic. We can generalize this to pairs of simple closed curves under additional conditions. If (γ_1, γ_2) and (δ_1, δ_2) are ordered pairs of essential simple closed curves with minimal intersection, $\gamma_1 \not\simeq_h \gamma_2$, $\gamma_1 \not\simeq \gamma_{2rev}$, $\delta_1 \not\simeq_h \delta_2$, $\delta_2 \not\simeq \delta_{2rev}$ and $\gamma_i \simeq_h \delta_i$ for $i = 1, 2$, then there is an ambient isotopy which moves one pair to the other.*

Proof. If two curves are ambient isotopic, they are certainly homotopic. The same holds for pairs of curves which already fullfill the side conditions.

Now for the non-trivial direction. We first tackle the problem for single curves. Notice that $I_{\min}(\gamma, \delta) = 0$: there is a homotopy $H : S^1 \times I \to \Sigma$ that moves γ to δ. After performing this homotopy, we can make γ disjoint from δ by a small homotopy inside a tubular neighbourhood of δ.[1] By lemma 2.3, we can now make the curves disjoint by an ambient isotopy. Assume this has been done. We invoke lemma 2.5 to conclude that γ and δ are the boundary curves of an annulus $S^1 \times I$.

Giving $S^1 \times I$ a product orientation enables us to say that γ and δ both wind around the annulus either to the right or to the left. Suppose they differed in direction. Then, because they are homotopic, $\gamma \simeq_h \gamma_{\text{rev}}$, whence $[\gamma]^2 = 1 \in \pi_1(\Sigma_g)$. But the fundamental group of an orientable surface does not contain elements of order 2, or of any finite order, for that matter, see Epstein [11] lemma 4.3. Therefore, γ and δ wind around the annulus in the same direction. We can thus construct an ambient isotopy between the two curves using collars on both sides of the annulus. This is geometrically obvious and proves our theorem for single curves.

Now for the generalization to pairs of curves. We have just proved we can move γ_1 to δ_1 by an ambient isotopy, so we start by doing that. Next we wish to push across 2-gons to ensure $\gamma_2 \cap \delta_2 = \emptyset$, as in lemma 2.3. However, we want to keep γ_1 fixed as a set, since it is already in place. Because γ_1 has minimal intersection with γ_2 and δ_2, whenever it intersects some 2-gon between γ_2 and δ_2, it crosses this 2-gon from γ_2 to δ_2 (or the other way around, and this may happen several times). Therefore we can indeed keep it fixed as set by 'pushing carefully'. We are now in a situation where $\gamma_1 = \delta_1$ and $\gamma_2 \cap \delta_2 = \emptyset$. This is shown in the figure on the left:

[1] Such a neighbourhood exists by the Tubular Neighbourhood Theorem, see Bredon [5] chapter II, theorem 11.14. Notice that we use the orientability of the surface here, for on a non-orientable surface, the normal bundle of δ in Σ might not have a nowhere zero section. Indeed, in that case we would have $I_{\min}(\gamma, \delta) = 1$.

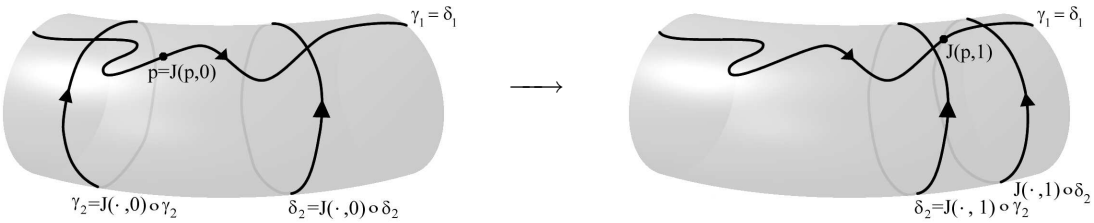


Figure 8. An isotopy across an annulus, keeping a set of disjoint curves across it fi xed.

We know the pair (γ_2, δ_2) bounds an annulus. Furthermore, both γ_2 and δ_2 have minimal intersection with $\gamma_1 = \delta_1$. Since γ_1 is not homotopic to γ_2 or its reverse, it is not contained in the annulus. This implies that if γ_1 intersects the annulus between γ_2 and δ_2 at all, it does so by crossing it a finite number of times, over disjoint paths. We can therefore move γ_2 across this annulus to δ_2 with an ambient isotopy that keeps γ_1 fixed as a set, and only moves points in a neighbourhood of the annulus. The result is shown above on the right. With this sequence of isotopies we achieve our goal. □

Remark. The preceding lemma can not be generalized straightforwardly to n-tuples of curves for $n \geq 3$. One could suspect the following holds. Given are two n-tuples $(\gamma_1, \ldots, \gamma_n)$ and $(\delta_1, \ldots, \delta_n)$ of essential, pairwise non-homotopic, pairwise minimally intersecting simple closed curves on a compact orientable surface Σ. If $\gamma_i \simeq_h \delta_i$ for $i = 1, \ldots, n$, then there is an ambient isotopy $J : \Sigma \times I \to \Sigma$ with $J(\cdot, 0) = \mathrm{id}_\Sigma$ and $J(\gamma_i(S^1), 1) = \delta_i(S^1)$ for $i = 1, \ldots, n$. But this is not true. It is true an ambient isotopy could be constructed which freed every γ_i from its corresponding δ_i by pushing carefully across 2-gons. But we could encounter an insurmountable obstacle, as shown in the following picture:

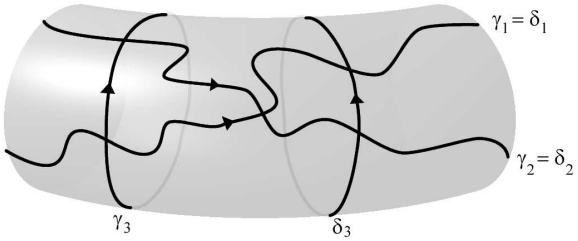


Figure 9. It is impossible in general to construct an ambient isotopy positioning three curves at will. In this particular instance this can be concluded from the fact that the only intersection point between γ_1 and γ_2 lies to the right of γ_3, but to the left of δ_3 on our orientable surface.

In view of this problem, if we want to generalize the lemma, we have to impose a restriction preventing these kinds of intersections. For this we introduce the following concept.

Definition 2.7 *A* closed essential 1-submanifold *of a compact orientable surface is a finite union of disjoint essential simple closed curves* $\{\gamma_1, \ldots, \gamma_n\}$ *such that* $\gamma_i \not\simeq_h \gamma_j, (\gamma_j)_{rev}$ *for* $i \neq j$*. Two closed essential* 1*-submanifolds* $N_1 = \{\gamma_1, \ldots, \gamma_n\}$ *and* $N_2 = \{\delta_1, \ldots, \delta_n\}$ *are called homotopic if for every* γ_i *there is a unique* δ_j *homotopic to it.*

If we restrict ourselves to closed essential 1-submanifolds, the following result is almost trivial.

Lemma 2.8 *If two closed essential* 1*-submanifolds* N *and* M *of a compact orientable surface* Σ *are homotopic, there is an ambient isotopy that moves one to the other. This can be generalized to pairs* (N_1, N_2) *and* (M_1, M_2) *of transverse closed essential* 1*-manifolds. Suppose* $N_i \simeq_h M_i$*, all curves of* N_1 *(*M_1*) have minimal intersection with all curves of* N_2 *(*M_2*) and no curve of* N_1 *(*M_1*) is homotopic to a curve of* N_2 *(*M_2*) or its reverse. Then there is an ambient isotopy moving one pair to the other.*

Proof. For the single submanifold case, we simply compose isotopies for each component, keeping all previously moved components γ_i of N fixed while working on one. This is possible because these components all have a tubular neighbourhood disjoint from any γ_j that has not been put in place yet. (We could cut these from the surface in every step, actually making this into an induction argument.)

In the case of pairs of submanifolds, we first move N_1 into position as in the last paragraph. Then N_2 can then also be dealt with componentwise, using the same subtlety as in lemma 2.6. We do not encounter the problematic situation mentioned above, because all the curves of N_2 are disjoint. This completes the demonstration. □

We need two more ingredients for our main result.

Theorem 2.9 *Between any two diffeomorphisms* h_1 *and* h_2 *of the disc* D^2 *that either both preserve or both reverse orientation, there is a differentiable isotopy.*

In the topological category, the proof is relatively straightforward. It is enough to prove that, given a homeomorphism h which preserves orientation, we can construct an isotopy to the identity. First, we use an isotopy to make sure that $h(0) = 0$ by extending an isotopy of $\{h(0)\} \times I \to D^2$ moving $h(0)$ back to 0 along a path, to an isotopy $J_1 : D^2 \times I \to D^2$. Second, we may rotate the whole disc around 0 by an isotopy J_2, so we assume that $h(0,0) = 0$ and $h(1,0) = (1,0)$. Now the action of h on S^1 may be described by the angle function $\theta(x) = \arg(h(x)) - \arg(x)$, where we choose $\arg$ in $[0, 2\pi)$. This θ is a continuous function $S^1 \to \mathbb{R}$ because of the assumption $h(1,0) = (1,0)$. We therefore use the isotopy

$$J_3(p,t) := \begin{cases} \rho_{-t\theta(p/||p||)}(p) & \text{if } p \neq (0,0) \\ (0,0) & \text{if } p = (0,0) \end{cases}$$

on the whole disc, where ρ_ϕ denotes rotation around the origin by an angle ϕ. We may thus also assume that $h|_{S^1}$ is the identity. In our fourth and last step, the famous Alexander trick gives us an isotopy J_4 between our map h and the identity:

$$J_3(p,t) := \begin{cases} t \cdot h\left(\frac{p}{t}\right) & \text{for } ||p|| < t \\ p & \text{for } ||p|| \geq t \end{cases}$$

This proof can be generalized to any dimension. The Alexander trick, however, is not adaptable to the differentiable category. The proof of the differentiable version is highly non-trivial, even

in dimension 2, and can be found in either a famous paper by Smale [39] or one of Munkres [31]. The higher-dimensional differentiable version was proven by Cerf [8] for $n \geq 6$.

Definition 2.10 *A set of curves on a closed orientable surface is said to* bind the surface *or* fill the surface *if the complement of the curves is a disjoint union of open discs.*

For $g \geq 1$ there exist two transverse closed essential 1-submanifolds N_1 and N_2 with minimal intersection that bind Σ_g. An example that is easy to verify is the pair $N_1 = \{\gamma_1, \ldots, \gamma_{2g-1}\}$, $N_2 = \{\delta\}$ shown below for Σ_3. One can actually find two essential *curves* which have minimal intersection and bind the surface. But in my opinion the greater generality of the lemma on 1-submanifolds is more enlightening than using two cleverly constructed curves, whose supposed minimal intersection would be far from obvious.

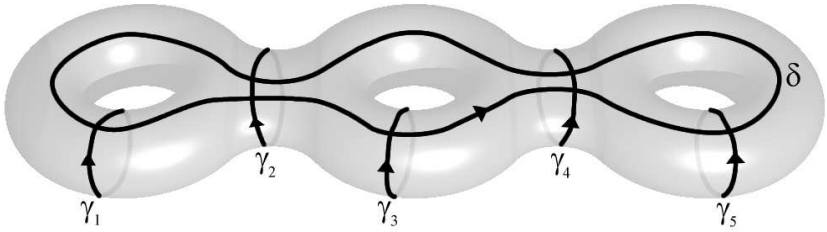


Figure 10. Two closed essential 1-submanifolds that bind Σ_3.

At last we are in a position to prove our main theorem, at least in the differentiable category.

Theorem 2.11 *Two diffeomorphisms of a closed orientable surface Σ_g are (topologically) homotopic if and only if they are differentiably isotopic.*

Proof. We prove the non-trivial implication. Let h_1 and h_2 be the diffeomorphisms. Supposing that $g \geq 1$, we use a pair (N_1, N_2) of transverse closed essential 1-submanifolds binding the surface. Our diffeomorphisms are by assumption homotopic, so $h_1(N_i) \simeq_h h_2(N_i)$. Lemma 2.8 therefore implies there is an ambient isotopy $J_1 : \Sigma_g \times I \to \Sigma_g$ moving the former pair into the latter. The two diffeomorphisms $\tilde{h}_1 : p \mapsto J_1(h_1(p), 1)$ and h_2 agree on $N_1 \cup N_2$. Moreover, the complement of $\tilde{h}_1(N_1 \cup N_2) = h_2(N_1 \cup N_2)$ consists of disjoint discs. The closure of these open discs are thus closed discs on whose boundary $\tilde{h}_1$ and h_2 agree. So lemma 2.9 tells us that these maps restricted to such a disc are (differentiably) isotopic. We can glue together the isotopies for the separate discs to form an isotopy $J_2 : \Sigma_g \times I \to \Sigma_g$ which 'adjusts' all their interiors. The composition

$$J(p,t) := \begin{cases} J_1(p, 2t) & \text{for } 0 \leq t \leq \frac{1}{2} \\ J_2(p, 2t-1) & \text{for } \frac{1}{2} \leq t \leq 1 \end{cases}$$

is then our sought after isotopy.

The case of $S^2 = \Sigma_0$ has to be treated separately, because on S^2 there are no essential curves. So we look at the image $h_1(\gamma)$ of some simple closed curve γ. We can not use lemma 2.6, but $h_1(\gamma)$ can be made disjoint from $h_2(\gamma)$ by an ambient isotopy, because of lemma 2.3. (Or we could simply say that the complement of $h_2(\gamma)$ is open and not empty, so $h_1(\gamma)$ can be moved

to some open disc in this complement.) By the Schönflies Theorem (see Bredon [5], chapter IV, theorem 19.11) the curves $h_1(\gamma)$ and $h_2(\gamma)$ bound an annulus. We can therefore move $h_1(\gamma)$ to either $h_2(\gamma)$ or $h_2(\gamma)_{\text{rev}}$ by an ambient isotopy. Assume this has been done. Lemma 2.9 assures that we can now adjust h_1 by an ambient isotopy on the two remaining discs which form $S^2 - h_1(\gamma)$ so that h_1 equals either h_2 or $a \circ h_2$, where a is the antipodal map. But the latter possibility would imply that $h_2 \cong_h a \circ h_2$, which is impossible (see for example Birman [4], Theorem 4.4). $\square$

Remark. We can generalize theorem 2.11 to a (connected) compact orientable surface Σ that is not the disc D^2 or the annulus $S^1 \times I$. We know from the classification of compact orientable surfaces that $\Sigma \cong \Sigma_g - \{\text{int}(D_1), \ldots, \text{int}(D_n)\}$, where $D_i \cong D^2$ is a disc embedded in the surface. A diffeomorphism of Σ permutes the boundary circles $B_1, \ldots, B_n$. Any B_i is essential in Σ, otherwise it would bound a disc and we would be dealing with D^2. If $i \neq j$, then $B_i \not\simeq_h B_j$ in Σ, otherwise the two essential curves B_i and B_j would bound an annulus by lemma 2.5 and we would have $\Sigma \cong S^1 \times I$. Two homotopic diffeomorphisms h_1, h_2 must therefore permute the boundary components in the same way. The isotopies between them on each component may be extended to an isotopy of Σ by the isotopy extension theorem, so we may assume $h_1 = h_2$ on $\partial\Sigma$.

We now choose a set of curves binding Σ_g that avoid $B_1, \ldots, B_n$. Looking back at the lemmata used to prove theorem 2.11, we notice that we can perform them while keeping $D_1, \ldots, D_n$, and thus $\partial\Sigma$, fixed. Only theorem 2.9 can not be applied straight away, because now we are dealing with some $D^2 - \{\text{int}(D_{i_1}), \ldots, \text{int}(D_{i_k})\}$. However, choosing a set of cuts, we may divide this surface up into discs. These cuts are moved by h_1 and h_2 to new cuts with the same endpoints. using a differentiable isotopy, we may move the one set to the other, thus again reducing the problem to isotopies on discs.

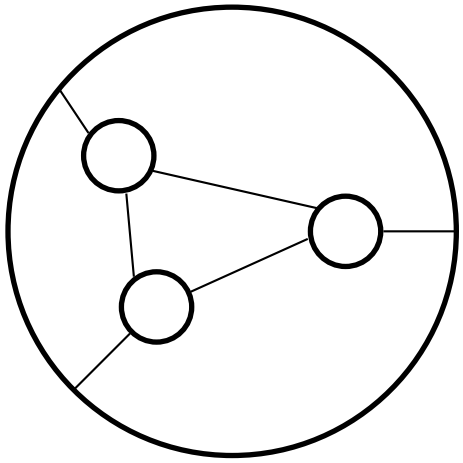


Figure 11. Cuts dividing up $D^2 - \{\text{int}(D_{i_1}), \ldots, \text{int}(D_{i_k})\}$.

The most important knowledge gained from the proof of theorem 2.11 is that the isotopy class of an automorphism is determined by what is does to a set of curves that bind the surface. Together

with the next result, this is foundational to the definition of the mapping class group we will encounter in the next section.

The extension of the main theorem to the topological category is tricky. A direct proof in this category does not seem to exist, which is not unreasonable, considering some of the awful things a general homeomorphism might do. The obvious strategy is therefore to use a result about more well-behaved automorphisms.

Epstein [11] proves that every surface homeomorphism is isotopic to a PL-homeomorphism. He also proves that two homotopic PL-homeomorphisms are PL-isotopic. This establishes our result independent of the category.

Theorem 2.12 *Given two automorphisms of a closed orientable surface Σ_g in either the topological, PL or Diff^r-category. They are isotopic in this category if and only if they are topologically homotopic.* □

We desire an extra result, which will be important in the next section, namely:

Theorem 2.13 *Every homeomorphism of a surface Σ is isotopic to a C^∞-diffeomorphism.*

Proof. We may assume by Epstein's theorem that the homeomorphism h is piecewise-linear. We metrize Σ and denote the distance function by d. By a result of Munkres [32, implied by theorem 6.3] for any continuous function $\varepsilon : \Sigma \to \mathbb{R}_+$ there is a diffeomorphism k such that

$$d(h(x), k(x)) < \varepsilon(x).$$

We now smoothly triangulate the surface by the simplicial complex S. Let $\{\sigma_i^0\}_{i\in I_0}$, $\{\sigma_i^1\}_{i\in I_1}$, $\{\sigma_i^2\}_{i\in I_2}$ denote the cells, the upper index indicating the dimension of the simplices. We choose ε in the following way:

$$\varepsilon(p) := \frac{1}{2} \cdot \inf \{\text{length}(\gamma) : \gamma \text{ an essential closed curve through } p\} .$$

Fix an ε-approximation k. To construct an isotopy from h to k, first, we move a 0-simplex $h(\sigma_i^0)$ to $k(\sigma_i^0)$ by an ambient isotopy on the ball $B_{\varepsilon(h(\sigma_i^0))}(h(\sigma_i^0))$ keeping the boundary fixed. These isotopies may then be combined to an isotopy of Σ by using the identity outside these discs. Assuming this has been done, we proceed to do the same for the 1-simplices $h(\sigma_i^1)$, keeping the 0-simplices fixed. This is possible because $h(\sigma_i^1)$ and $k(\sigma_i^1)$ are homotopic relative to their endpoints after our first isotopy, due to our choice of ε. After this step, we may do the same for the 2-simplices, using the Alexander trick. □

3 The structure of $\mathrm{Homeo}(\Sigma_g)$ and the MCG

3.1 Function spaces

The collections $\mathrm{Homeo}(\Sigma_g)$, $\mathrm{PL}(\Sigma_g)$ and $\mathrm{Diff}^r(\Sigma_g)$ of automorphisms of a surface Σ_g are all subsets of the set $\mathcal{C}(\Sigma_g, \Sigma_g)$ of continuous maps from Σ_g to itself. For any topological spaces X, Y, we may topologize the sets $\mathcal{C}(X, Y)$ and $\mathrm{Homeo}(X)$. These or any of their subspaces are called function spaces. Since we are dealing with spaces that might also be endowed with a PL-structure or differential structure, we consider them as nested spaces, although below we will define a topology on $\mathrm{Diff}^r(\Sigma_g)$ that is different from the subspace topology:

$$\mathrm{Diff}^r(\Sigma_g) \subset \mathrm{Homeo}(\Sigma_g) \subset \mathcal{C}(\Sigma_g, \Sigma_g) \quad \text{or} \quad \mathrm{PL}(\Sigma_g) \subset \mathrm{Homeo}(\Sigma_g) \subset \mathcal{C}(\Sigma_g, \Sigma_g).$$

But the automorphisms exhibit more structure. They naturally form a group under composition of maps, with id_{Σ_g} functioning as the unit element. For the topologies used in our discussion, they are topological groups with this multiplication. There are quite a few topologies on function spaces, and for a proper discussion of a few well-known ones, I refer the reader to Munkres [34] and Hirsch [17].

The topology that is of greatest interest to us is the compact-open topology. For any two topological spaces X and Y the compact-open topology is defined on $\mathcal{C}(X, Y)$. It is generated by the subbasis of all sets

$$B(K, V) := \{f \in \mathcal{C}(X, Y) \,|f(K) \subseteq V\},$$

where $K \subseteq X$ is compact and $V \subseteq Y$ is open. The fact that Σ_g is compact and metrizable simplifies things immensely. If we put a metric on Σ_g, we can use theorems 46.7 and 46.8 from Munkres [34], which claim:

Theorem 3.1 *Let X and Y be topological spaces. If X is compact and Y is a metric space, then the uniform topology, the topology of compact convergence and the compact-open topology on $\mathcal{C}(X, Y)$ coincide. Moreover, they are induced by the sup-metric*

$$d_{\mathcal{C}(X,Y)}(f, g) := \sup_{x \in X} d_Y(f(x), g(x)).$$

Note that the topology on $\mathcal{C}(\Sigma_g, \Sigma_g)$ does not depend on the metric chosen on Σ_g.

The most useful feature of the compact-open topology is the following. With this topology, a (continuous) path γ from f_1 to f_2 in $\mathcal{C}(\Sigma_g, \Sigma_g)$ corresponds to a homotopy $H : \Sigma_g \times I \to \Sigma_g$ between f_1 and f_2 given by $H(p, t) := \gamma(t)(p)$. If γ lies in $\mathrm{Homeo}(\Sigma_g)$, then the corresponding homotopy is actually an isotopy.

Remark. Contrast this with a coarser topology, such as the topology of pointwise convergence. With this topology, a continuous family of Dehn twists (to be defined in section 3.4) that stays fixed outside an annulus whose width shrinks to zero, connects two non-isotopic homeomorphisms. This is something we do not want to happen.

If the surface has been given a $\mathcal{C}^r$-differentiable structure ($r \in \mathbb{N} \cup \{\infty\}$), we may also use the $\mathcal{C}^r$-topology on $\mathrm{Diff}^r(\Sigma_g)$. Actually, this topology comes in two flavors, namely the weak topology $\mathcal{C}^r_W$ and the strong topology $\mathcal{C}^r_S$ (also called fine topology or Whitney topology). We define them on $\mathcal{C}^r(\Sigma_g, \Sigma_g)$. The $\mathcal{C}^r_W$ topology is generated by the following subbasis. Let $f \in \mathcal{C}^r(\Sigma_g, \Sigma_g)$ be a differentiable map, (ϕ, U) and (ψ, V) be charts on Σ_g and $K \subset U$ be a compact set such that $f(K) \subset V$. For $\varepsilon > 0$ we define the *weak subbasis element* $\mathcal{N}^r_W(f; (\phi, U), (\psi, V), K, \varepsilon)$ to be

$$\left\{ g \in \mathcal{C}^r(\Sigma_g, \Sigma_g) : \sup_{x \in \phi(K)} \left\| \frac{\partial^{|\mu|}(\psi f \phi^{-1})}{\partial x_\mu}(x) - \frac{\partial^{|\mu|}(\psi g \phi^{-1})}{\partial x_\mu}(x) \right\| < \varepsilon \text{ for } |\mu| \leq r \right\},$$

where $\partial^{|\mu|}/\partial x_\mu$ signifies the partial derivative with multi-index $\mu = (i_1, \ldots, i_s)$ of length $|\mu| = s$.

For the strong topology $\mathcal{C}^r_S$ we require a locally finite atlas $\Phi = \{(U_i, \phi_i)\}_{i \in I}$, a family $K = \{K_i\}_{i \in I}$ of compact subsets with $K_i \subset U_i$, a family of charts $\Psi = \{(V_i, \psi_i)\}_{i \in I}$ such that $f(K_i) \subset V_i$ and a family of positive numbers $\varepsilon = \{\varepsilon_i\}_{i \in I}$. This time, we give a basis, using the *strong basis element* $\mathcal{N}^r_S(f; \Phi, \Psi, K, \varepsilon)$, defined by

$$\left\{ g \in \mathcal{C}^r(\Sigma_g, \Sigma_g) : \sup_{x \in \phi_i(K_i)} \left\| \frac{\partial^{|\mu|}(\psi_i f \phi_i^{-1})}{\partial x_\mu}(x) - \frac{\partial^{|\mu|}(\psi_i g \phi_i^{-1})}{\partial x_\mu}(x) \right\| < \varepsilon_i \text{ for } |\mu| \leq r \text{ and } i \in I \right\}.$$

In our case, both the domain and range space of the function space are compact. One can easily see that the weak and strong topology then coincide. For complete details on these topologies, see Hirsch [17] and Munkres [33].

The $\mathcal{C}^r$-topology is much finer than the compact-open topology restricted to $\mathrm{Diff}^r(\Sigma_g)$. To construct a path inside $\mathrm{Diff}^r(\Sigma_g)$ between two diffeomorphisms in the $\mathcal{C}^r$-topology, we need more than a continuous family of $\mathcal{C}^r$-diffeomorphisms, a simple but necessary observation.

Example. We can take the $\mathcal{C}^\infty$-diffeomorphism $s : D^2 \to D^2$ given by

$$s(x, y) = \begin{pmatrix} \cos b(\frac{x^2+y^2}{2}) & -\sin b(\frac{x^2+y^2}{2}) \\ \sin b(\frac{x^2+y^2}{2}) & \cos b(\frac{x^2+y^2}{2}) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

where $b : \mathbb{R} \to \mathbb{R}$ is a $\mathcal{C}^\infty$ function with $b(x) = 0$ for $x \geq 1$ and $b(x) = \pi/2$ for $x \leq 0$. This is a kind of swirl:

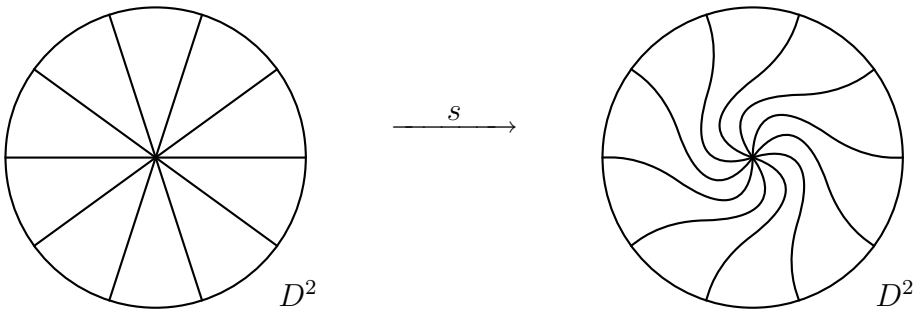


Figure 12. A swirl map on a small disc. A few radii and their images are drawn.

Now consider a point p on a metrized Σ_g, and some disc $D^2_u(p)$ of radius u. For every $t \in [0,u]$ we can construct the $\mathcal{C}^\infty$-diffeomorphism s_t by applying our swirl s to $D^2_t(p)$ and keeping points outside this disc fixed. Cleary, in the compact-open topology, this family of maps forms a continuous path from id_{Σ_g} to s_u. However, in the $\mathcal{C}^r$-topology (for any $r \geq 1$) this is not so, because for the first derivatives we see:

$$d_p(s_t) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \text{ for } t > 0, \text{ while } d_p(s_0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This can not be remedied by adjusting the expansion speed of the disc. It suffers from the same defect as the Alexander trick in this respect. The $\mathcal{C}^r$-topology is the unique topology such that a path in $\mathrm{Diff}^r(\Sigma_g)$ is continuous if and only if the isotopy corresponding to it is $\mathcal{C}^r$-differentiable. Remember that a differentiable isotopy from h_1 to h_2 is a $\mathcal{C}^r$-differentiable map $J : \Sigma_g \times (-\varepsilon, 1+\varepsilon) \to \Sigma_g$ such that $J(\cdot,t) = h_1$ for $t \leq 0$ and $J(\cdot,t) = h_2$ for $t \geq 1$; we require more than just for the maps $J(\cdot,t)$ to be diffeomorphisms.

The above example might presently suggest that this $\mathcal{C}^r$-topology is too fine for the study of isotopy classes of maps in the topological category. However, we have seen in section 2 that any two diffeomorphisms that act in the same way on the homotopy classes of curves on the surface can be joined by a differentiable isotopy, so the $\mathcal{C}^r$-topology does not separate $\mathrm{Diff}^r(\Sigma_g)$ into too many components. We will formulate this more precisely below.

3.2 The mapping class group

With either the compact-open topology or an even finer one, all function spaces under discussion become infinite-dimensional spaces (in terms of the Lebesgue covering dimension). This says in effect that there are so many ways to slightly alter a given automorphism, that we can not even begin to describe the possible alterations succinctly, by any finite number of parameters that is. In order to obtain useful information on the group structure of automorphisms, we therefore want to study homeomorphisms up to isotopy/homotopy. Thus we look at the path components of these spaces, the path component of the identity being the prime example.

Lemma 3.2 *In a topological group G the path component G_0 of the identity is a normal subgroup.*

Proof. To prove that $G_0 < G$, we remark that if $g \in G_0$, then there is a path γ from 1 to g. Obviously $g^{-1}\gamma$, defined by $(g^{-1}\gamma)(t) := g^{-1}\gamma(t)$ is a path from g^{-1} to 1, so that $g^{-1} \in G_0$ as well. If $g_1, g_2 \in G_0$ with paths γ_1, γ_2 from 1 to g_1, g_2 respectively, the product path $\gamma_1\gamma_2$ defined by $(\gamma_1\gamma_2)(t) := \gamma_1(t)\gamma_2(t)$ is a path from 1 to g_1g_2, and so $g_1g_2 \in G_0$. This proves that G_0 is a subgroup of G.

To show that $G_0 \lhd G$, we take $g \in G$ and $h \in G_0$. There is a path γ from 1 to h, and it is obvious that the path $g\gamma g^{-1}$ defined by $(g\gamma g^{-1})(t) := g\gamma(t)g^{-1}$ is a path from 1 to ghg^{-1}. So $ghg^{-1} \in G_0$, which is precisely what we need to prove. □

The path components of a topological group G are thus the cosets of G_0. Also, the quotient group G/G_0 is well-defined. For our specific objects of study, we therefore define the *mapping*

class group of Σ_g to be

$$\mathrm{MCG}(\Sigma_g) := \mathrm{Homeo}(\Sigma_g)/\,\mathrm{Homeo}_0(\Sigma_g).$$

We could hope for some extra topological structure on $\mathrm{MCG}(\Sigma_g)$ coming from $\mathrm{Homeo}(\Sigma_g)$, but alas, the induced quotient topology is discrete.

Lemma 3.3 $\mathrm{Homeo}_0(\Sigma_g)$ *is an open subgroup of* $\mathrm{Homeo}(\Sigma_g)$ *in the compact-open topology.*

Proof. It is enough to show that every point of $\mathrm{Homeo}_0(\Sigma_g)$ has an open neighbourhood contained in $\mathrm{Homeo}_0(\Sigma_g)$. Without loss of generality, we may prove this for id_{Σ_g} only, since we are working in a topological group. This means we have to show that there exists an $\varepsilon_0 > 0$ such that $d(h, 1) < \varepsilon_0 \Longrightarrow h \simeq_i 1$.

Choose a finite set of curves that bind the surface such as in theorem 2.11 and any suitable Riemannian metric, for example one induced from $\mathbb{R}^3$ by some embedding. For any of the binding curves there is an ε_0 such that all its ε_0-neighbourhoods are annuli. Because there is a finite number of curves, this ε may be chosen so that it satisfies this condition for all curves simultaneously. Therefore, as $d(x, h(x)) < \varepsilon$ for all x, h fixes the homotopy classes of these curves, and by the same reasoning as in theorem 2.11, we conclude $h \simeq_i 1$. □

Remark. The mapping class group can be defined for all topological spaces, and I do not know if it may not sometimes inherit a non-trivial topology. The above lemma can be generalized to compact manifolds by using a triangulation. We have seen this technique applied to surfaces in theorem 2.13.

In one of the differentiable categories, denote it by $\mathrm{Diff}^\infty(\Sigma_g)$, we could attempt the same construction and define

$$\mathrm{MCG}_d(\Sigma_g) := \mathrm{Diff}(\Sigma_g)/\,\mathrm{Diff}_0(\Sigma_g).$$

using the $\mathcal{C}^r$-topology on $\mathrm{Diff}(\Sigma_g)$. Do we end up with a different group? By theorem 2.11, if $h \in \mathrm{Homeo}_0(\Sigma_g)$ is differentiable, then it is differentiably isotopic to id_{Σ_g}, so $\mathrm{Diff}_0(\Sigma_g) = \mathrm{Homeo}_0(\Sigma_g) \cap \mathrm{Diff}(\Sigma_g)$. Consider the natural projection

$$\pi|_{\mathrm{Diff}(\Sigma_g)} : \mathrm{Diff}(\Sigma_g) \to \mathrm{Homeo}(\Sigma_g)/\,\mathrm{Homeo}_0(\Sigma_g).$$

If $h_1, h_2 \in \mathrm{Diff}(\Sigma_g)$ are in the same coset of $\mathrm{Diff}_0(\Sigma_g)$, then $h_2 = h_1 h$ for some $h \in \mathrm{Diff}_0(\Sigma_g) < \mathrm{Homeo}_0(\Sigma_g)$, so $\pi|_{\mathrm{Diff}(\Sigma_g)}(h_2) = \pi|_{\mathrm{Diff}(\Sigma_g)}(h_1 h) = \pi|_{\mathrm{Diff}(\Sigma_g)}(h_1)$. Thus the projection factors through to a group homomorphism

$$\bar{\pi} : \mathrm{Diff}(\Sigma_g)/\,\mathrm{Diff}_0(\Sigma_g) \longrightarrow \mathrm{Homeo}(\Sigma_g)/\,\mathrm{Homeo}_0(\Sigma_g).$$

Suppose that $\bar{\pi}(h \cdot \mathrm{Diff}_0(\Sigma_g)) = 0$. This means $h \in \mathrm{Homeo}_0(\Sigma_g)$, which implies $h \in \mathrm{Diff}(\Sigma_g) \cap \mathrm{Homeo}_0(\Sigma_g) = \mathrm{Diff}_0(\Sigma_g)$. So $\bar{\pi}$ is injective. Because of theorem 2.13, $\pi|_{\mathrm{Diff}(\Sigma_g)}$ is surjective, so $\bar{\pi}$ is surjective as well. This shows that

$$\mathrm{MCG}_d(\Sigma_g) \cong \mathrm{MCG}(\Sigma_g).$$

The analogous result holds in the PL category as well. Note that for higher dimensional manifolds this would not be true and we would have to distinguish carefully between $\mathrm{MCG}(M)$, $\mathrm{MCG}_{\mathrm{PL}}(M)$ and $\mathrm{MCG}_d(M)$, if either of the latter two is even defined.

Returning to surfaces, we remark that just like in the topological category, we do not inherit any topological structure on $\mathrm{MCG}_d(\Sigma_g)$.

Lemma 3.4 *The group* $\mathrm{Diff}_0^r(\Sigma_g)$ *is an open subgroup of* $\mathrm{Diff}^r(\Sigma_g)$ *in the* $\mathcal{C}^r$*-topology.*

Proof. As in the preceding lemma, it is enough to construct a neighbourhood of the identity that is contained in $\mathrm{Diff}_0^r(\Sigma_g)$. We construct this neighbourhood as the intersection $U \cap \mathrm{Diff}^r(\Sigma_g)$, where U is a neighbourhood of id_{Σ_g} in $\mathrm{Homeo}(\Sigma_g)$ (with the compact-open topology), which is contained in $\mathrm{Homeo}_0(\Sigma_g)$. Such a U exists by the previous lemma. Because the $\mathcal{C}_S^r$-topology is finer than the compact-open topology, this is an open neighbourhood of id_{Σ_g} in $\mathrm{Diff}(\Sigma_g)$ with the $\mathcal{C}_S^r$-topology. But by theorem 2.12, any $h \in U \cap \mathrm{Diff}^r(\Sigma_g)$ must then be differentiably isotopic to id_{Σ_g}, implying that this neighbourhood is contained in $\mathrm{Diff}_0^r(\Sigma_g)$ $\square$

A very common subgroup of the mapping class group is the group $\mathrm{MCG}_+(\Sigma_g) < \mathrm{MCG}(\Sigma_g)$ of orientation-preserving automorphism isotopy classes of Σ_g. Some authors even completely restrict their attention to this subgroup and call *that* $\mathrm{MCG}(\Sigma_g)$. We will not adhere to this practice.

The structure of $\mathrm{MCG}(\Sigma_g)$ can be derived from that of $\mathrm{MCG}_+(\Sigma_g)$ by a semi-direct product. Every Σ_g can be imbedded in $\mathbb{R}^3$ in a such that there is a mirroring symmetry (see the figure below). This gives an orientation-reversing involution (that is, an automorphism of order 2) σ for Σ_g. The subgroup $\{1, \sigma\} < \mathrm{MCG}(\Sigma_g)$ acts on $\mathrm{MCG}_+(\Sigma_g)$ by conjugation and $\{1, \sigma\} \cap \mathrm{MCG}_+(\Sigma_g) = \{1\}$. Moreover, $\mathrm{MCG}_+(\Sigma_g) \lhd \mathrm{MCG}(\Sigma_g)$ since for an orientation preserving automorphism h and any automorphism g, ghg^{-1} is orientation preserving. From the above we conclude that $\mathrm{MCG}(\Sigma_g) \cong \mathrm{MCG}_+(\Sigma_g) \rtimes \mathbb{Z}_2$, with multiplication $(g_1, h_1) \cdot (g_2, h_2) = (g_1 c_{h_1}(g_2), h_1 h_2)$ for $g_1, g_2 \in \mathrm{MCG}_+(\Sigma_g)$ and $h_1, h_2 \in \{1, \sigma\}$, where c_{h_1} is the conjugation action with the element h_1.

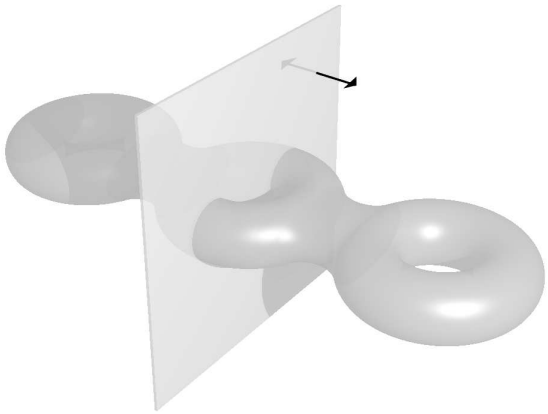


Figure 13. An orientation-reversing involution of Σ_g.

3.3 The MCG of the sphere and the torus

The simplest surfaces are $S^2 = \Sigma_0$ and $T^2 = \Sigma_1$. For these, the mapping class group is still tractable. A presentation can be given without a lot of effort. The case of S^2 follows from a little algebraic topology combined with our previous reasoning.

Theorem 3.5 *A homeomorphism $h : S^2 \to S^2$ is isotopic either to id or the antipodal map a. These maps themselves are not isotopic. Therefore, $\mathrm{MCG}(S^2) = \{\mathrm{id}, a\} \cong \mathbb{Z}_2$. and $\mathrm{MCG}_+(S^2) = \{1\}$.*

We give two proofs.

Proof. As in the proof of 2.11, look at a simple closed curve γ and $h \circ \gamma$. By an isotopy we can make sure that the latter is moved to either γ or γ_{rev}. By adjusting the remaining discs of S^2, we have an isotopy to either id or a. These maps are not isotopic since their degree differs: $\deg(\mathrm{id}) = 1$ while $\deg(a) = -1$. (See Bredon [5], chapter IV corollary 6.12.) □

Proof. Note that $S^2 = \partial D^3$. We can extend the homeomorphism h to one of D^3 by defining:

$$\tilde{h}(x) = \begin{cases} h\left(\frac{x}{||x||}\right)||x|| & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}.$$

From the Alexander lemma (accomplishing the same as lemma 2.9 in the topological category) we know that h is isotopic to either the identity or $x \mapsto -x$, and that these maps are not isotopic, for the latter is orientation reversing and the former orientation preserving. Since any isotopy of h can be extended in the same way as the map itself to D^3 and any isotopy of D^3 restricted to S^2 is an isotopy of S^2, the result follows. □

The torus is a little more complicated. We use its universal cover to discover its mapping class group. A variation on this proof can be found in Stillwell [40].

Theorem 3.6 $\mathrm{MCG}(T^2) \cong \mathrm{GL}_2(\mathbb{Z})$ *and* $\mathrm{MCG}_+(T^2) \cong \mathrm{SL}_2(\mathbb{Z})$.

Proof. The curves α_1 and β_1 defined in section 1.3 actually bind the surface in the case of T^2. As seen in theorem 2.11, this means that the isotopy class of an automorphism $h : T^2 \to T^2$ is uniquely determined by the homotopy classes $[h \circ \alpha_1]$ and $[h \circ \beta_1]$. Consider the universal covering $\pi : \mathbb{R}^2 \to T^2$ and choose coordinates such that $\tilde{\alpha}_1(t) = (t, 0)$ and $\tilde{\beta}_1(t) = (0, t)$ are lifts of α_1 and β_1, respectively. For any automorphism h there is a unique lift $\tilde{h} : \mathbb{R}^2 \to \mathbb{R}^2$ of h fixing $(0, 0)$. The above-mentioned homotopy classes are uniquely determined by the endpoints of $\tilde{h} \circ \tilde{\alpha}_1$ and $\tilde{h} \circ \tilde{\beta}_1$, both lying in $\mathbb{Z}^2$. We take the unique linear map $L : \mathbb{R}^2 \to \mathbb{R}^2$ with $L((1, 0)) = \tilde{h} \circ \tilde{\alpha}_1(1)$ and $L((0, 1)) = \tilde{h} \circ \tilde{\beta}_1(1)$. As h is invertible, so is $\tilde{h}$, and therefore L. It follows that L maps $\mathbb{Z}^2$ to itself, implying $L \in \mathrm{GL}_2(\mathbb{Z})$. Since elements of $\mathrm{GL}_2(\mathbb{Z})$ commute with $\mathrm{Deck}_\pi(\mathbb{R}^2)$, L induces an automorphism of T^2 (isotopic to h). But in fact, every element of $\mathrm{GL}_2(\mathbb{Z})$ projects down to T^2, we have a group isomorphism from $\mathrm{MCG}(T^2)$ to $\mathrm{GL}_2(\mathbb{Z})$. □

The dynamics of (isotopy classes of) toral automorphisms gives rise to a further classification. We have three different types of orientation preserving mapping classes. Fixing a correspondence between $\mathrm{MCG}_+(T^2)$ and $\mathrm{SL}_2(\mathbb{Z})$, the 'linear' element A_h of each mapping class $[h]$ is one of three kinds, determined by the trace of the A_h. The behaviour is best understood by looking at the action of the corresponding element of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{R}^2$. Of course, the type does not depend on the particular correspondence chosen.

1. $|\operatorname{tr}(A_h)| < 2$: this means A_h is *periodic*. Moreover, $A_h^{12} = 1$;
2. $|\operatorname{tr}(A_h)| = 2$: implies that there is a simple closed curve which is fixed setwise (but the map is not periodic). A_h is called *reducible*.
3. $|\operatorname{tr}(A_h)| > 2$: in this case there is an irrational number $\lambda > 0$ and two transverse dense immersions of $\mathbb{R}$ in T^2 such that A_h is expanding by a factor λ along one of these and contracting by a factor λ along the other. We call A_h *Anosov*.

In contrast to these first few cases, the mapping class group of higher genus surfaces is much more complex, and has been an active area of research in the past 40 years. A lot of progress has been made, of which I can present only a small part. But first we look at an interesting class of automorphisms.

3.4 Dehn twists

Consider an oriented surface Σ_g. We describe a family of automorphisms called Dehn twists. They can be defined in $\mathrm{Diff}(\Sigma_g)$, but we will not bother and just use homeomorphisms, focussing on the essentials.

Definition 3.7 *Let γ be a simple closed curve on Σ_g, and consider a tubular neighbourhood N of γ. Take an orientation preserving homeomorphism $i : A \to N$, where A is the annulus plane parametrized in polar coordinates by*

$$\{(r, \theta) : 1 \leq r \leq 2, \theta \in \mathbb{R}/2\pi\mathbb{Z}\}.$$

and is given the standard orientation. Then the automorphism $D_{\gamma,N,i}$ is defined by

$$D_\gamma(p) := \begin{cases} i \circ D \circ i^{-1} & \text{if } p \in N \\ p & \text{if } p \notin N \end{cases},$$

where $D : A \xrightarrow{\sim} A$ is given by

$$(r, \theta) \mapsto (r, \theta + 2\pi(r - 1)).$$

This map is called a left handed Dehn twist around γ. *A right handed Dehn twist around γ is the inverse of a left handed twist.*

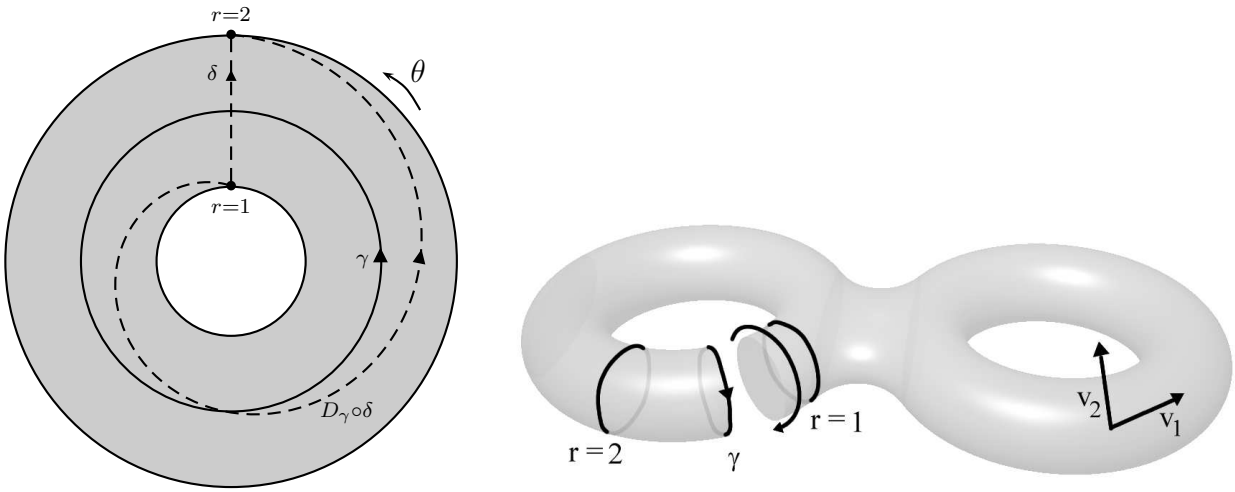


Figure 14. A left handed Dehn twist about a simple closed curve γ. On the left the action on the annulus is illustrated in the plane. An arc δ running straight across the annulus is mapped to a spiral that winds around the annulus once (dashed lines). The Dehn twist is most readily visualized as in the illustration on the right. We cut open the surface along γ, take the loose ends in both hands, twist one end a full turn to the left (with respect to the orientation indicated by (v_1, v_2)) and glue the surface back together along γ.

Our description suggests that N and i are not very important. And indeed this is so for our purposes, as we are interested in isotopy classes.

Lemma 3.8 *The isotopy class of a Dehn twist $D_{\gamma,N,i}$ does not depend on the choice of N and i, and homotopic closed curves give isotopic twists.*

Proof. First, suppose we have D_{γ,N,i_1} and D_{γ,N,i_2}. It is sufficient to prove that $i_1 \circ D \circ i_1^{-1} \simeq_i i_2 \circ D \circ i_2^{-1}$, keeping the boundaries of N fixed, or equivalently that $D \circ (i_2^{-1} \circ i_1) \circ D^{-1} \simeq_i (i_2^{-1} \circ i_1)$ as maps from $A \to A$, keeping ∂A fixed. An isotopy is given in polar coordinates by $J(r, \theta, t) := E \circ (i_2^{-1} \circ i_1) \circ E^{-1}(r, \theta, t)$, where $E : A \xrightarrow{\sim} A$ is defined to be

$$E(r, \theta, t) := \begin{cases} (r, \theta + 2\pi(r-1)/t) & \text{if } r \leq 1 + t \\ (r, \theta) & \text{if } r > 1 + t \end{cases}.$$

Now suppose we have two tubular neighbourhoods N_1, N_2 of γ. There is a tubular neighbourhood $N_3 \subseteq N_1 \cap N_2$ and both N_1 and N_2 can be deformed to N_3 by an ambient isotopy, implying there is also an isotopy $J : \Sigma_g \times I \to \Sigma_g$ such that $J(\cdot, 0) = \mathrm{id}$ and $J(N_1, 1) = N_2$. It follows that $K : \Sigma_g \times I \to \Sigma_g$ given by $K(p, t) := D_{\gamma, J(N_1,t)}(p)$ is an isotopy from D_{γ,N_1} to D_{γ,N_2}. (We may omit mention of the specific maps i here, by the previous step.)

Lastly, if $\gamma_1 \simeq_h \gamma_2$, there is an ambient isotopy $J : \Sigma_g \times I \to \Sigma_g$ with $J(\cdot, 0) = \mathrm{id}_{\Sigma_g}$ and $J(\gamma, 1) = \gamma_2$ by lemma 2.6. Having chosen a tubular neighbourhood N of γ_1 we may choose $J(N, 1)$ as a tubular neighbourhood for γ_2. We now see that $D_{\gamma_2, J(N,1)} = J(\cdot, 1) D_{\gamma_1,N} J(\cdot, 1)^{-1}$ and that the map $K(p, t) := J(p, t) D_{\gamma_1,N}(p) J(p, t)^{-1}$ is an isotopy between these, demonstrating our last claim. □

Remark. Because of the previous lemma, we will write a Dehn twist as D_γ. Sometimes, we will also loosely refer to the isotopy class of some D_γ as a Dehn twist. Actually, $[D_\gamma]$ also does not depend on the direction of γ, that is, $[D_\gamma] = [D_{\gamma_{\text{rev}}}]$. This is immediate, because we may still choose the same N and i.

Dehn twist are intuitively accessible examples of non-trivial automorphisms. Although he introduced them in the 1920s, Max Dehn published about them for the first time only in 1938 (see [9]). He proved that these twists are very powerful: they actually generate the whole group $\mathrm{MCG}_+(\Sigma_g)$ of orientation preserving homeomorphisms of a surface. In his paper, he constructed a concrete set of $2g(g-1)$ twists that accomplish this. In the 1960s, Raymond Lickorish revived interest in Dehn twists and found a generating set with only $3g-1$ twists (see Lickorish [24] and [25], and also Birman [4] and Ivanov [20] for simplified proofs). For this reason, Dehn twists are also referred to as Lickorish twists by some, but we will stick to using the inventor's name. The number of twist generators was cut down to $2g+1$ by Humphries [19], who also proved this is the minimum number of Dehn twists which can generate $\mathrm{MCG}_+(\Sigma_g)$.

Theorem 3.9 (Dehn, Lickorish, Humphries) *Let $\alpha_1, \ldots, \alpha_g$, $\beta_1, \ldots, \beta_g$ and $\gamma_1, \ldots, \gamma_{g-1}$ be the $3g-1$ curves on Σ_g ($g \geq 1$) shown in the figure below. These generate $\mathrm{MCG}_+(\Sigma_g)$, and in fact $\alpha_3, \ldots, \alpha_g$ can be left out to obtain a set of $2g+1$ generators. This is the minimum number of Dehn twists that can generate $\mathrm{MCG}_+(\Sigma_g)$.*

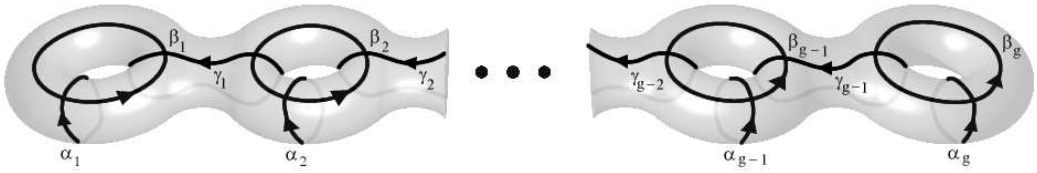


Figure 15. Lickorish' set of twist generators for Σ_g.

3.5 Presentations of the MCG

After the work of Lickorish, attempts were made to find other generating sets of $\mathrm{MCG}_+(\Sigma_g)$. One of the most recent results of these efforts has been the article [6] by Brendle and Farb. They have obtained the following.

Theorem 3.10 (Brendle, Farb) *For every $g \geq 1$, the group $\mathrm{MCG}_+(\Sigma_g)$ is generated by 3 torsion elements (i.e. elements of finite order).*

Theorem 3.11 (Brendle, Farb) *For $g \geq 3$, the group $\mathrm{MCG}_+(\Sigma_g)$ is generated by 6 involutions (i.e. elements of order 2).*

A generator set is one thing, a presentation something else. For $\mathrm{MCG}_+(\Sigma_2)$, one was found by Birman in 1973, see [4]. After groundbreaking work by Hatcher and Thurston, the first explicit presentations for $\mathrm{MCG}_+(\Sigma_g)$ ($g \geq 3$) was found in the early 1980s by Harer and Wajnryb. An overview of this work can be found in the expository article [42] by Wajnryb.

All the presentations mentioned use some set of Dehn twists as generators, although not the one mentioned in the previous subsection.

Theorem 3.12 (Wajnryb) *The group* $\mathrm{MCG}_+(\Sigma_g)$ *(*$g \geq 2$*) admits the finite presentation generated by elements* $\{b_2, b_1, a_1, e_1, a_2, e_2, \ldots, a_{g-1}, e_{g-1}, a_g\}$ *representing the Dehn twists around the curves in the following figure*

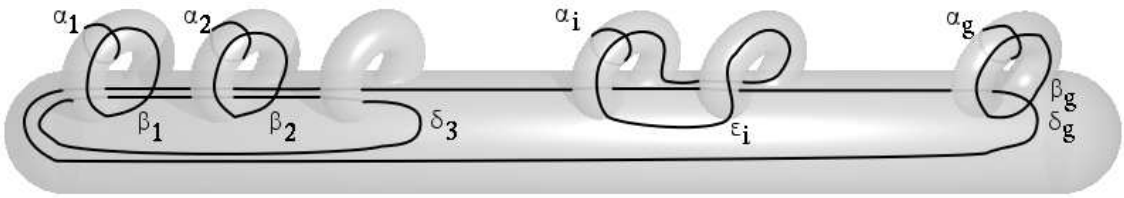


Figure 16. Wajnryb's set of twist generators for Σ_g.

with the following relations:

$$\begin{aligned}
xyx &= yxy \quad \textit{for consecutive elements } x, y \textit{ in the list} \\
xy &= yx \quad \textit{for non-consecutive elements in the list} \\
b_2a_2b_2 &= a_2b_2a_2 \\
b_2b_1 &= b_1b_2 \\
(b_1a_1e_1a_2)^5 &= b_2a_2e_1a_1b_1^2a_1e_1a_2b_2 \\
d_3a_1a_2a_3 &= d_{1,2}d_{1,3}d_{2,3} \\
d_g &\textit{ commutes with } b_1a_1e_1a_2\cdots a_{g-1}e_{g-1}a_ga_ge_{g-1}a_{g-1}\cdots e_1a_1b_1
\end{aligned}$$

where

$$\begin{aligned}
d_{1,2} &= (a_2e_1a_1b_1)^{-1}b_2(a_2e_1a_1b_1) \\
d_{1,3} &= t_2d_{1,2}t_2^{-1} \\
d_{2,3} &= t_1d_{1,3}t_1^{-1} \\
d_2 &= d_{1,2} \\
d_i &= (b_2a_2e_1b_1^{-1}t_2t_3\cdots t_{i-1})d_{i-1}(b_2a_2e_1b_1^{-1}t_2t_3\cdots t_{i-1})^{-1} \quad \textit{for } i = 3, 4, \ldots, g \\
t_i &= e_ia_ia_{i+1}e_i \quad \textit{for } i = 1, 2, \ldots, g-1
\end{aligned}$$

A special case of this presentation is Birman's presentation for $\mathrm{MCG}_+(\Sigma_2)$.

Theorem 3.13 *The group* $\mathrm{MCG}_+(\Sigma_2)$ *admits the presentation with generators* $g_1, \ldots, g_5$ *and*

relations

$$\begin{aligned} g_i g_j &= g_j g_i \qquad (|i-j| \geq 2, 1 \leq i,j \leq 5) \\ g_i g_{i+1} g_i &= g_{i+1} g_i g_{i+1} \qquad (1 \leq i \leq 4) \\ (g_1 g_2 g_3 g_4 g_5)^6 &= 1 \\ (g_1 g_2 g_3 g_4 g_5^2 g_4 g_3 g_2 g_1)^2 &= 1 \\ [g_1 g_2 g_3 g_4 g_5^2 g_4 g_3 g_2 g_1, g_i] &= 1 \qquad (1 \leq i \leq 5) \end{aligned}$$

where $g_1 = [D_{\alpha_1}]$, $g_2 = [D_{\beta_1}]$, $g_3 = [D_{\gamma_1}]$, $g_4 = [D_{\beta_2}]$ *and* $g_5 = [D_{\alpha_2}]$. *See Birman [4], theorem 4.8.*

The element $g_1 g_2 g_3 g_4 g_5^2 g_4 g_3 g_2 g_1$ is isotopic to a rotation over π around a longitudinal symmetry axis in the standard embedding we have seen before of Σ_2:

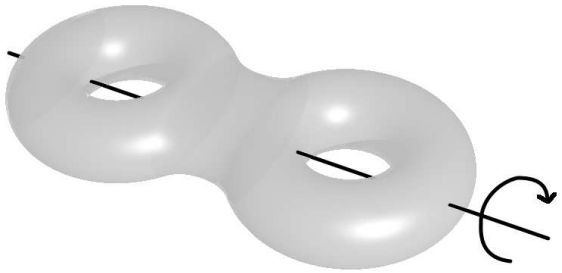


Figure 17. The hyperelliptic involution of Σ_2.

This mapping class has order 2 and is called the *hyperelliptic involution*. A special property that sets it apart is that it commutes with all g_i. Because these generate $\mathrm{MCG}_+(\Sigma_2)$, the hyperelliptic involution commutes with the whole group. We remark that higher genus surfaces have trivial center.

4 Geometric structure

4.1 Enter hyperbolic geometry

One of the main tools that is used in studying surfaces and their automorphisms is hyperbolic geometry. To apply this tool we put extra structure on a (topological) surface. Their are several possible approaches. One is to introduce a Riemannian metric of constant curvature. The other is to define a geometric structure, which is the route we will take. This does not eliminate the need for Riemannian geometry, but we tidily sweep it under the carpet. All of this is to be found in Benedetti & Petronio [3]. We take the Poincaré disc model as a definition.

Definition 4.1 Hyperbolic n-space $\mathbb{H}^n$ *is* $\mathrm{int}(D^n) \subset \mathbb{R}^n$ *as a smooth manifold. For* $v_x, w_x \in T_x\mathbb{H}^n$ *we define the inner product*

$$\langle v_x, w_x \rangle := \frac{4}{(1 - \sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n v_i w_i,$$

making $\mathbb{H}^n$ *into a smooth Riemannian manifold.*

With this metric, the distance between two points $x, y \in \mathbb{H}^n$ can be calculated to be

$$d(x, y) = \mathrm{arccosh}\left(1 + \frac{2||x - y||^2}{(1 - ||x||^2)(1 - ||y||)^2}\right).$$

Fixing y, we see that as x approaches the boundary S^n of $\mathbb{H}^n$, $d(x, y) \to \infty$. Contrary to Euclidean appearances, the boundary is infinitely far away, and the embedding in $\mathbb{R}^n$ by which we defined hyperbolic n-space is certainly no isometric embedding! Geodesics in $\mathbb{H}^n$ are segments of Euclidean circles orthogonal to S^n or of lines through the origin. Angles between curves may be measured as between Euclidean curves, as can be gleaned directly from the definition of the metric. In other words, the embedding of $\mathbb{H}^n$ in $\mathbb{R}^n$ is conformal. The sphere S^{n-1} is variously called the 'sphere at infinity', 'sphere of directions' or 'boundary' of $\mathbb{H}^n$ and denoted by $\partial\mathbb{H}^n$. Although it does not belong to $\mathbb{H}^n$ properly, it is extremely useful in hyperbolic geometry. We will encounter it again in section 7.

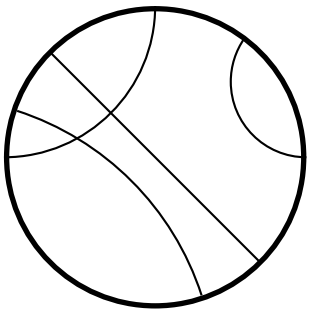


Figure 18. The hyperbolic plane $\mathbb{H}^2$ with a few lines drawn in it.

Definition 4.2 *A* local isometry $f : M \to N$ *between smooth Riemannian manifolds is a smooth map such that for all* $p \in M$ *and tangent vectors* $v, w \in T_p$*:*

$$\langle Tf(v), Tf(w)\rangle_{f(p)} = \langle v, w\rangle_p.$$

An isometry *is a local isometry that is bijective.*

Local isometries are exactly the (locally) distance-preserving maps from one Riemannian manifold to another. It is easy to see an isometry is automatically a diffeomorphism. We denote the group of isometries from a Riemannian manifold to itself by $\mathrm{Isom}(M)$. For example, for $\mathbb{R}^2$ with the usual metric, these are the affine maps $x \mapsto Ax + b$, where $A \in O(2)$. We now introduce a general concept of structure on a manifold, which we will then apply to the hyperbolic plane.

Definition 4.3 *Given is some manifold* X^n *with a group of homeomorphisms* G *acting on it. A* local G-map *is a map* $\xi : V \to W$ *between open sets of* X *such that every* $x \in V$ *has a neighbourhood on which* ξ *is the restriction of an element of* G*. An* (X, G)-structure on a manifold M^n *is a maximal atlas* $\{(U_i, \phi_i)\}_{i\in I}$ *on* M*, where* $\phi_i : U_i \xrightarrow{\sim} U_i' \subseteq X$ *is a homeomorphism between open sets, such that for any two charts* ϕ_i, ϕ_j*, the map* $\phi_j \circ \phi_i^{-1} : \phi_i(U_i \cap U_j) \to \phi_j(U_i \cap U_j)$ *is a local* G*-map.*

An example is a differentiable structure, which can be thought of as an $(\mathbb{R}^n, \mathrm{Diff}(\mathbb{R}^n))$-structure. We are interested in *geometric* structures, meaning we use some Riemannian manifold and its isometries. In particular, a *spherical* structure is an $(S^n, \mathrm{Isom}(S^n))$-structure, a *Euclidean* or *flat* structure is an $(\mathbb{R}^n, \mathrm{Isom}(\mathbb{R}^n))$-structure and a *hyperbolic* structure is an $(\mathbb{H}^n, \mathrm{Isom}(\mathbb{H}^n))$-structure. These terms are often used as adjectives for some manifold. We will for example speak of a hyperbolic surface $(\Sigma_g, \mathcal{H})$, meaning a surface Σ_g with a hyperbolic structure $\mathcal{H}$.

It turns out that a closed (orientable) surface can be fitted with exactly one of the three latter structures. The sphere S^2 is spherical (no wonder), T^2 is flat, and Σ_g is hyperbolic for $g \geq 2$.[2] This should immediately raise our interest in these hyperbolic structures if we want to study surfaces. Any complete connected (X, G)-manifold M is isometric to $X/\pi_1(M)$, where $\pi_1(M) \leq G$ operates freely and properly discontinuously on X. The simplest example of this is $T^2 \cong \mathbb{R}^2/\mathbb{Z}^2$, where $\pi_1(T^2) \cong \mathbb{Z}_2$ acts by translations on $\mathbb{R}^2$. More interesting still, any surface Σ_g for $g \geq 2$, equipped with a hyperbolic structure, is isometric to $\mathbb{H}^2/\Gamma$ for some $\Gamma \cong \pi_1(\Sigma_g)$, $\Gamma < \mathrm{Isom}(\mathbb{H}^2)$.

The orientation preserving isometries of $\mathbb{H}^2$ fall into three categories.

1. an *elliptic* isometry has a fixed point in $\mathbb{H}^n$. It is best described as a (hyperbolic) rotation around this point;

[2] In terms of Riemannian geometry, the three geometric structures we defined exhaust the possibilities for a Riemannian metric with constant Gauss curvature. A hyperbolic surface has Gauss curvature -1, a flat surface 0 and a spherical surface $+1$. Any closed surface can be fitted with a constant curvature metric and then scaled to one of these types. In three dimensions, the situation is much more difficult, and only in the 1970s has William Thurston been able to shed light on it.

2. a *parabolic* isometry has one fixed point on $\partial\mathbb{H}^n$. It moves points around on so-called horospheres through this fixed point. In our model, a horosphere is a euclidean sphere touching $\partial\mathbb{H}^n$;

3. a *hyperbolic* isometry has two fixed points on $\partial\mathbb{H}^n$. From a euclidean perspective, one of these, p, is repelling, the other, q, attracting. Points are moved towards q along curves which are equidistant from the (unique) geodesic with end points p and q, which is called the axis of the isometry.

All the covering transformations of a hyperbolic surface $(\Sigma_g, \mathcal{H}) \cong \mathbb{H}^2/\Gamma$ are in fact isometries of hyperbolic type. This in fact already shows that a torus can not support a hyperbolic structure, because its fundamental group is $\mathbb{Z}^2$, and two isometries of hyperbolic type along different axes do not commute.

4.2 Thurston's classification of surface automorphisms

We have seen that every mapping class of the torus contains a canonical element, given by a matrix in $\mathrm{GL}_2(\mathbb{Z})$. This was explained in subsection 3.3. In terms of their dynamics, such automorphisms fall into three distinct classes.

Thurston solved the problem whether the same can be done for higher genus. The hyperbolic nature of these surfaces makes it much harder, but it turns out there are still three types of behaviour. Thurston's work was widely circulated as a preprint before being published in concise form in [41]. A more introductory text to this material is Casson & Bleiler [7]. In Fathi et al. [12], a detailed exposition can be found.

Theorem 4.4 *Let Σ_g be a surface of genus at least two and let $h : \Sigma_g \to \Sigma_g$ be a homeomorphism. Then the isotopy class of h contains a homeomorphism k satisfying one of the following:*

1. *k is* periodic *and it is an isometry of $(\Sigma_g, \mathcal{H})$ for some hyperbolic structure $\mathcal{H}$;*

2. *k is* pseudo-Anosov*: it leaves a pair of transverse measured foliations with finitely many singular points on Σ_g invariant. (This means that points in the same leaf get mapped to the same leaf.) On one of them it is expanding by an irrational factor λ, on the other it is contracting by this factor;*

3. *k is* reducible*: there is an essential closed 1-submanifold on Σ_g that is left invariant and a power of k fixes this submanifold pointwise. The complement of this submanifold is a surface with finitely many components $C_1, \ldots, C_n$, for which holds: if $k^i(C_j) = C_j$, then $k^i|_{C_j}$ is of type 1 or 2.*

For example, a Dehn twist D_γ is reducible. The cases are not completely mutually exclusive. A periodic homeomorphism may fix some essential closed 1-submanifold. Pseudo-Anosov behaviour and periodicity preclude each other.

5 The Nielsen realization problem

We now discuss a famous problem on surface automorphisms. The problem is easy enough to state, but we first give a historical introduction.

An automorphism $h : \Sigma_g \to \Sigma_g$ almost induces an automorphism of the fundamental group of the surface. The variability resides in the fact that the fundamental group is defined using a base point p, and h need not keep this fixed. In general h induces the group isomorphism

$$\begin{aligned} h_* : \pi_1(\Sigma_g, p) &\to \pi_1(\Sigma_g, h(p)), \\ [\zeta] &\mapsto [h \circ \zeta] \end{aligned}$$

but the latter group is not canonically isomorphic to the former. For a path γ from $h(p)$ to p there is the path isomorphism

$$\begin{aligned} \gamma_* : \pi_1(\Sigma_g, h(p)) &\to \pi_1(\Sigma_g, p). \\ [\zeta] &\mapsto [\gamma_{\text{rev}} * \zeta * \gamma] \end{aligned}$$

Note that a path isomorphism γ_* is an inner automorphism of $\pi_1(\Sigma_g, p)$ if γ is a loop from p to p, for then $\gamma_* : [\zeta] \mapsto [\gamma_{\text{rev}} * \zeta * \gamma] = [\gamma]^{-1} \cdot [\zeta] \cdot [\gamma]$. From this it follows that for two different choices of paths γ_1, γ_2 from $h(p)$ to p the path isomorphisms γ_{1*} and γ_{2*} differ by an inner automorphism of $\pi_1(\Sigma_g, p)$, since $\gamma_{2*} = (\gamma_{1_{\text{rev}}} * \gamma_2)_* \gamma_{1*}$ and $\gamma_{1_{\text{rev}}} * \gamma_2$ is a loop from p to p. Now since $\mathrm{Inn}(\pi_1(\Sigma_g, p)) \lhd \mathrm{Aut}(\pi_1(\Sigma_g, p))$, the quotient group

$$\mathrm{Out}(\pi_1(\Sigma_g, p)) := \mathrm{Aut}(\pi_1(\Sigma_g, p)) / \mathrm{Inn}(\pi_1(\Sigma_g, p)),$$

called the *outer automorphism group* of the fundamental group, is defined. And because of the foregoing remarks, h induces a well-defined element of this group by setting

$$\begin{aligned} \nu : \mathrm{Homeo}(\Sigma_g) &\to \mathrm{Out}(\pi_1(\Sigma_g, p)) \\ h &\mapsto \gamma_* \circ h_* \mod \mathrm{Inn}(\pi_1(\Sigma_g, p)) \end{aligned}$$

for some path γ from $h(p)$ to p. Indeed, ν is a group homomorphism. Given two homeomorphisms $h_1, h_2 : \Sigma_g \to \Sigma_g$, we choose paths γ_i from $h_i(p)$ to p for $i = 1, 2$ and δ from $h_2 \circ h_1(p)$ to $h_1(p)$. Then

$$\begin{aligned} \nu(h_2 \circ h_1) &= (\delta * \gamma_1)_* \circ (h_2 \circ h_1)_* \\ &= \gamma_{1*} \circ \delta_* \circ h_{2*} \circ h_{1*} \\ &= (\gamma_{1*} \circ \delta_* \circ h_{2*} \circ \gamma_{1\text{rev}*}) \circ (\gamma_{1*} \circ h_{1*}) \\ &= (((h_2 \circ \gamma_{1_{\text{rev}}}) * \delta * \gamma_1)_* \circ h_{2*}) \circ (\gamma_{1*} \circ h_{1*}) \\ &= (\gamma_{2*} \circ h_{2*}) \circ (\gamma_{1*} \circ h_{1*}) \\ &= \nu(h_2)\nu(h_1) \end{aligned}$$

where we used $h_{2*} \circ \gamma_{1\text{rev}*}([\zeta]) = [h_2 \circ (\gamma_1 * \zeta * \gamma_{1_{\text{rev}}})] = [(h_2 \circ \gamma_1) * h_2 \circ \zeta * (h_2 \circ \gamma_{1_{\text{rev}}})] = [(h_2 \circ \gamma_{1_{\text{rev}}})_{\text{rev}} * h_2 \circ \zeta * (h_2 \circ \gamma_{1_{\text{rev}}})] = (h_2 \circ \gamma_{1_{\text{rev}}})_* \circ h_{2*}([\zeta])$ in the fourth step. The diagram below might assist in the visualization of this computation.

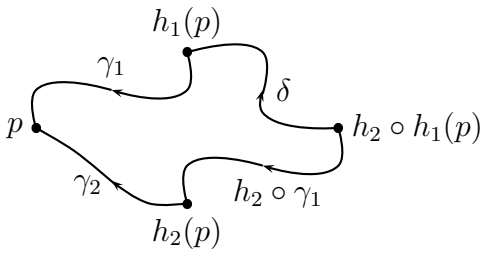


Figure 19. A visual aid in understanding that ν is a homomorphism.

The name ν is given in honour of Jakob Nielsen. In the 1920s, he became interested in the converse problem. Is every element of $\mathrm{Out}(\pi_1(\Sigma_g))$ induced by a concrete element of $\mathrm{Homeo}(\Sigma_g)$? He was able to prove this is indeed so in [35]. He then investigated the more general question of what subgroups of $\mathrm{Out}(\pi_1(\Sigma_g, p))$ can be represented by subgroups of $\mathrm{Homeo}(\Sigma_g)$. This turns out to be very difficult and is our main concern in the rest of this text. We want to specify more precisely where the problem resides.

To get clear on this we consider automorphisms $h_1, h_2 : \Sigma_g \to \Sigma_g$ that are homotopic, say by $H : \Sigma_g \times I \to \Sigma_g$. We can define the path δ from $h_1(p)$ to $h_2(p)$ by $\delta(t) := H(p, t)$. For any loop ζ from p to p, the loop $h_2 \circ \zeta$ is homotopic to $\delta_{\text{rev}} * (h_1 \circ \zeta) * \delta$ because $h_1 \simeq_h h_2$. Put otherwise, $h_{2*} = \delta_* \circ h_{1*}$. From this it follows that for paths γ_i from $h_i(p)$ to p:

$$\gamma_{2*} \circ h_{2*} = \gamma_{2*} \circ \delta_* \circ h_{1*} = (\delta * \gamma_2)_* \circ h_{1*} = \gamma_1 \circ h_{1*} \mod \mathrm{Inn}(\pi_1(\Sigma_g, p)).$$

So in fact ν factors through to $\mathrm{MCG}(\Sigma_g)$ and induces a map

$$\begin{aligned} \bar{\nu} : \mathrm{MCG}(\Sigma_g) &\to \mathrm{Out}(\pi_1(\Sigma_g, p)) \\ [f] &\mapsto \gamma_* \circ f_* \mod \mathrm{Inn}(\pi_1(\Sigma_g, p)) \end{aligned}$$

For $\bar{\nu}$ the following strong result emerged from results of Baer [1] and [2], Dehn, and Nielsen. The first complete proof was written up by Mangler [26].

Theorem 5.1 (Baer-Dehn-Nielsen) $\bar{\nu} : \mathrm{MCG}(\Sigma_g) \xrightarrow{\sim} \mathrm{Out}(\pi_1(\Sigma_g))$ *for all* $g \geq 1$.

This result implies that the problem of representing subgroups of $\mathrm{Out}(\pi_1(\Sigma_g, p))$ — that is, finding a set of automorphisms that induce the subgroup — is equivalent to representing subgroups of $\mathrm{MCG}(\Sigma_g)$. This representation problem is called the

Generalized Nielsen realization problem. Which subgroups $H \leq \mathrm{MCG}(\Sigma_g)$ admit a representation in $\mathrm{Homeo}(\Sigma_g)$? That is, for which H can we find a map $\sigma : H \to \mathrm{Homeo}(\Sigma_g)$ such that $\pi_0 \circ \sigma = 1$, where $\pi_0 : \mathrm{Homeo}(\Sigma_g) \to \mathrm{MCG}(\Sigma_g)$ is the canonical projection. Rephrased in the language of homological algebra: does the short exact sequence

$$0 \longrightarrow \mathrm{Homeo}_0(\Sigma_g) \longrightarrow \mathrm{Homeo}(\Sigma_g) \xrightarrow[\pi_0]{} \mathrm{MCG}(\Sigma_g) \longrightarrow 0$$

$$\sigma$$

split, and if not, by what subgroups $H' < \mathrm{Homeo}(\Sigma_g)$ and $H < \mathrm{MCG}(\Sigma_g)$ may we replace these terms so that it does?

We may replace $\mathrm{Homeo}(\Sigma_g)$ by $\mathrm{PL}(\Sigma_g)$ or $\mathrm{Diff}(\Sigma_g)$ to make it even harder, or even demand that our representation preserves a geometric structure. Note that we are not allowed to use more than one automorphism in each isotopy class. The adjective 'generalized' is used because Nielsen concentrated on finite subgroups, probably for simplicity. The difficulty already mentioned above thus resides in finding concrete automorphisms which together have the demanded group structure of their isotopy classes. We say the problem is solvable for $H \leq \mathrm{MCG}(\Sigma_g)$ if a representation exists. If the problem is solvable for H, then the same goes for all $H' \leq H$.

6 Partial solutions to the Nielsen realization problem

We now give a summary of the results that have been obtained on the Nielsen realization problem. Because of the variety and difficulty of techniques used, I can only sketch some proofs, and I do not claim to understand all details of every paper cited. But let us start with some positive results.

6.1 Positive results

The sphere. For the sphere the Nielsen realization problem is no problem at all. The whole of $\mathrm{MCG}(S^2)$ can be represented by the concrete group $\{\mathrm{id}, a\}$ where $a : x \mapsto -x$ is the antipodal map (in the standard embedding $S^2 \to \mathbb{R}^3$).

The torus. We recall that $\mathrm{MCG}(T^2) \cong \mathrm{GL}_2(\mathbb{Z})$. As a matter of fact, the Nielsen realization problem is solvable for the whole of $\mathrm{GL}_2(\mathbb{Z})$. We know that this group acts on the universal cover $\mathbb{R}^2$ of T^2, and since each element respects the equivalence classes under the identification $(x, y) \sim (x + m, y + n)$ $(m, n \in \mathbb{Z})$, this action descends to an action on the torus.

The fact that T^2 has a Euclidean structure plays an important role. We were able to construct a realization using linear maps. In terms of differential geometry, these are geodesic maps. For $g \geq 2$, can we find a 'natural' action of $\mathrm{MCG}(\Sigma_g)$ on $\mathbb{H}^2 \cong \tilde{\Sigma}_g$ descending to an action on Σ_g? If we interpret natural here as being by geodesic maps, we will not succeed. A geodesic map of $\mathbb{H}^2$ must already be an isometry. And although a surface of genus at least 2 can carry infinitely many hyperbolic structures, the isometry group of any closed hyperbolic surface is finite (see Kobayashi [23] theorem III.2.2 or Zieschang [43] theorem 15.21). So there is no nice realization of an infinite subgroup of mapping classes by geodesic maps, like on the torus. For example, a Dehn twist can not be realized in this way on a hyperbolic surface.

Cyclic subgroups. Cyclic subgroups are the easiest subgroups we can try to represent in any mapping class group. If $H \leq \mathrm{MCG}(\Sigma_g)$ is infinite cyclic, then any automorphism in a class $[g]$ that generates H has infinite order. They all solve the problem. For a finite cyclic subgroup $H \cong \mathbb{Z}/n\mathbb{Z}$ the problem boils down to the question: if a homeomorphism h satisfies $h^n \simeq_i 1$, is there a homeomorphism $\tilde{h} \simeq_i h$ for which $\tilde{h}^n = 1$? Nielsen proved that this is indeed the case for orientation preserving homeomorphisms in [37], using hyperbolic isometries. Thus we arrive at the result:

Theorem 6.1 *Cyclic subgroups of* $\mathrm{MCG}_+(\Sigma_g)$ *can be represented in* $\mathrm{Diff}_+(\Sigma_g)$.

Finite solvable groups. Fenchel extended Nielsen's previously obtained result to finite solvable subgroups of $\mathrm{MCG}_+(\Sigma_g)$, also using hyperbolic geometry, see [13]. His results are explained in chapter 3 of Zieschang [43].

Group extensions of $\pi_1(\Sigma_g)$ admitting a splitting. Eckmann and Müller [10] proved that the problem is solvable for an even greater number of finite groups of mapping classes, but their criterium is a little more difficult. The key theorem is of a group-theoretic nature. They prove

that the groups $\pi_1(\Sigma_g)$ are characterized algebraically by two properties: they have to be PD^2-groups, meaning they have to obey a property reminiscent of Poincaré-duality, and they have to split over a finitely generated subgroup. They then prove a theorem on virtual PD^2-group (i.e. groups containing a PD^2-group as a subgroup of finite index), which allows them to conclude the following.

Theorem 6.2 *Let G be a finite effective group extension of $\pi_1(\Sigma_g)$ ($g \geq 2$) which splits over a finitely generated subgroup. Then G can be realized as a group of isometries acting cocompactly, freely, and properly discontinuously on $\mathbb{H}^2$.*

Here a finite group extension of a group A is a group B such that $A \lhd B$ and $[B : A]$ is finite. Such an extension induces a conjugation action of B upon A by setting, for any $b \in B$, $c_b : a \mapsto bab^{-1}$ for all $a \in A$. The extension is called effective if this conjugation action is the identity on A only when $b \in A$. We apply the theorem as follows. Given a finite $H < \mathrm{MCG}(\Sigma_g) \cong \mathrm{Out}(\pi_1(\Sigma_g))$ for some $g \geq 2$, we can construct a finite effective extension $G \rhd \pi_1(\Sigma_g)$ such that $\xi : G/\pi_1(\Sigma_g) \to \mathrm{Out}(\pi_1(\Sigma_g))$ given by $\xi([g]) = [x \mapsto gxg^{-1}]$ maps $G/\pi_1(\Sigma_g)$ isomorphically onto H. If G splits over a finitely generated subgroup, we can then realize it by isometries of $\mathbb{H}^2$ according to their theorem. But since $\pi_1(\Sigma_g) \lhd G$, this action descends to an action of $H \cong G/\pi_1(\Sigma_g)$ on $\Sigma_g \cong \mathbb{H}^2/\pi_1(\Sigma_g)$.

Whether the group G splits is a technical issue. For instance, this is ensured if the first Betti number $\beta_1(G)$ is positive. We need not even know what the Betti number of a group is; it can be computed with the aid of the character $\chi(h) = \mathrm{tr}(\bar{\mu}(h))$ of the representation in first homology that we will encounter in section 7. According to Eckmann and Müller:

$$\beta_1(G) = \frac{1}{|H|} \sum_{h \in H} \chi(h).$$

Remark. Any subgroup $H \cong \mathbb{Z}^n$ generated by a collection of Dehn twists $\{[D_{\gamma_1}], \ldots, [D_{\gamma_n}]\}$ with $I_{\min}(\gamma_i, \gamma_j) = 0$ for all i, j is also representable just by taking any representatives of $[D_{\gamma_i}]$ and their products. The same goes for any free (non-abelian) subgroup. The essential problem is finding automorphisms having suitable relations in the group theoretical sense.

6.2 Finite groups in general

Steven Kerckhoff finally proved in [21] that the Nielsen realization problem is solvable for *all* finite subgroups of $\mathrm{MCG}(\Sigma_g)$, thus also overcoming the restriction to $\mathrm{MCG}_+(\Sigma_g)$. Just as the previous results on higher genus surfaces, his proof depends on hyperbolic geometry, but in a new way. Continuing to use hyperbolic geometry was no lucky guess. Kerckhoff and others had proved that if the Nielsen realization problem were solvable for a finite group of mapping classes, it had to be solvable for this finite group by using isometries with respect to a suitable hyperbolic metric. The question is therefore: how do we find the right hyperbolic metric. For this, Kerckhoff used a structure called the Teichmüller space.

Definition 6.3 *The* Teichmüller space $\mathcal{T}(\Sigma_g)$ *of a topological surface* Σ_g *is the quotient set of the collection* $\{(\Sigma_g, \mathcal{H}) : \mathcal{H}$ *is a hyperbolic structure on* $\Sigma_g\}$ *by the relationship*

$$(\Sigma_g, \mathcal{H}_1) \sim (\Sigma_g, \mathcal{H}_2) \quad :\Longleftrightarrow \quad \exists h : (\Sigma_g, \mathcal{H}_1) \xrightarrow{\sim} (\Sigma_g, \mathcal{H}_2), \quad h \simeq_i 1_{\Sigma_g}.$$

The Teichmüller space of Σ_g can be given a metric structure and thereby topologized. It turns out to be isometric (as a metric space) to $\mathbb{R}^{6g-6}$, see Benedetti & Petronio [3]. A mapping class of Σ_g also induces a homeomorphism on $\mathcal{T}(\Sigma_g)$, and Kerckhoff proved that a finite group of mapping classes must have a common fix point. This fix point gives the hyperbolic structure that solves the problem for the given subgroup of $\mathrm{MCG}(\Sigma_g)$.

Theorem 6.4 *Every finite subgroup* $H < \mathrm{MCG}(\Sigma_g)$ *can be represented in* $\mathrm{Diff}(\Sigma_g)$*. Moreover,* H *can be represented as a group of isometries with respect to some hyperbolic metric on* Σ_g*.*

Remark. We might ask what the finite subgroups of $\mathrm{MCG}(\Sigma_g)$ actually are. We have already seen that such a subgroup is the isometry group of some hyperbolic surface. By a theorem of Hurwitz (Zieschang [43] theorem 15.21, already cited above), such an isometry group can contain at most $84(g-1)$ elements and this bound is sharp. So this is the maximum order of any finite subgroup of $\mathrm{MCG}(\Sigma_g)$. Another notable result has been that every periodic element of $\mathrm{MCG}(\Sigma_g)$ has order at most $4g+2$, see Nielsen [36].

6.3 Groups with two ends / virtually cyclic groups

The Nielsen realization problem is also solvable for infinite groups with two ends. This is stated in the Kirby problem list [22] without reference, and I was not able to find the result in the literature. But first, although a very simple description of the mysterious sounding concept of "groups with two ends" will shortly be given, I nonetheless only discovered this after having ploughed through some literature, the results of which I do not wish to deprive the reader of.

The concept of an end of a group derives from the idea of an end of a topological space. This idea was developed by Freudenthal in the 1930s while working on compactifications. Some spaces (e.g. $\mathbb{R}$) can be separated by leaving out a compact subset, others (e.g. $\mathbb{R}^2$) can not. An end is more or less one of the components if we are allowed to cut away a large compact subset. To be exact:

Definition 6.5 *An* open end neighbourhood *of a non-compact topological space* X *is an open set* $U \subseteq X$ *containing, for some compact* $\emptyset \subset K \subseteq X$*, a component of* $X - K$*. An* end *of* X *is an equivalence class of nested sequences* $U_1 \supset U_2 \supset \cdots$ *of connected open end neighbourhoods such that* $\cap_{i=1}^{\infty} \bar{U}_i = \emptyset$*, where*

$$U_1 \supset U_2 \supset \cdots \quad \sim \quad V_1 \supset V_2 \supset \cdots$$

when for all U_i *there is a* V_j *such that* $U_i \subseteq V_j$ *and for all* V_i *there is a* U_j *such that* $V_i \subseteq U_j$*.*

We are interested in a numerical invariant, namely the number of ends a space has. For the intuition, $\mathbb{R}$ and a cylinder have two ends, $\mathbb{R}^n$ has one end for $n \geq 2$, and $\Sigma_g - \{p_1, \ldots, p_k\}$ has k ends. What we want to do is transport this concept to finitely generated groups and define for any such group its number of ends. To accomplish this, we give a definition of the number of ends of a locally finite graph. Rest assured this definition gives the same result gotten by suitably topologizing the graph and counting equivalence classes of ends as defined above.

Definition 6.6 *Given a connected graph X, where each vertex is incident to only finitely many edges, the* number of ends of X *is the supremum in $\mathbb{N} \cup \{\infty\}$ of the number of infinite components into which X can be separated by removing a finite number of edges.*

To get from groups to graphs, we introduce the Cayley graph of a group.

Definition 6.7 *The Cayley graph $C(G, T)$ of a group G with generator set T is the group whose vertices are the elements of G and whose edges are (g, gt) for any $g \in G$ and $t \in T \cup T^{-1}$. (For an order 2 element t one usually constructs two edges between g and $gt = gt^{-1}$.)*

If the group is finitely generated, the Cayley graph is locally finite. We topologize it using a metric. We identify each edge with $[0, 1]$ and define the distance between two points as the minimal length of an arc between them (this minimum exists). The Cayley graph of a group is connected, since any element $g \in G$ can be written as a word in $T \cup T^{-1}$, and this word is a recipe for a path from 1 to g. We therefore define *the number of ends $e(G)$ of a finitely generated group* G to be the number of ends of its Cayley graph $C(G, T)$ for some finite generating set T. The Cayley graph may depend on T, but its number of ends does not. Moreover, while a topological space and a graph can have any number of ends, there are only a limited number of possibilities for groups. In fact, $e(G) \in \{0, 1, 2, \infty\}$, see Hopf [18] for both of these last two claims.

Another result from [18] is that groups with two ends are exactly the groups that are *virtually* $\mathbb{Z}$ (also *virtually infinite cyclic*), meaning they have a finite index subgroup isomorphic to $\mathbb{Z}$. Still more is known, compare Scott & Wall [38, theorem 5.12]:

Theorem 6.8 *Let G be a finitely generated group. The following are equivalent.*

1. *$e(G) = 2$;*
2. *G has an infinite cyclic subgroup of finite index;*
3. *G has a finite normal subgroup H with quotient G/H either $\mathbb{Z}$ or $\mathbb{Z}_2 * \mathbb{Z}_2$;*
4. *Either G is a semi-direct product $H \rtimes \mathbb{Z}$ for some finite $H \lhd G$ or G is an amalgamated free product $G_1 *_H G_2$ of finite groups such that $[G_1 : H] = [G_2 : H] = 2$ and $H \lhd G$.*

How do we use this knowledge to realize the mapping classes of a group with two ends $G < \mathrm{MCG}(\Sigma_g)$? We could try to use the two possible specific forms of G as given in $4.$ of our theorem. In case G is an amalgamated free product $G_1 *_H G_2$, realizing it reduces to realizing both G_1 and G_2 such that the realizations coincide on $H = G_1 \cap G_2$. If we can do that, we can then realize a word in $G_1 *_H G_2$ by the product of the realizations of its letters (in G_1 and G_2),

and all group relations will be satisfied. However, I do not know how to solve the problem in this way.

There is a more successful attack on the problem that only makes use of the information that $\mathbb{Z} \lhd G$. Thurston [private communication] says that the following fact is true and can be found in [41]. Suppose we consider the classification of surface automorphisms by dynamical behaviour, as seen in section 4. If any automorphism h is in canonical form (periodic, reducible or pseudo-Anosov) and a mapping class g normalizes $\langle [h] \rangle$, which is to say that

$$g[h]g^{-1} \in \{[h], [h]^{-1}\},$$

then there is an automorphism $h_g \in g$ that normalizes $\langle h \rangle$ exactly:

$$h_g h h_g^{-1} \in \{h, h^{-1}\}.$$

What I actually need to finish the proof is the extra knowledge that this normalizing element is unique. Under this assumption, suppose we are given a group $G < \mathrm{MCG}(\Sigma_g)$ and a subgroup $\mathbb{Z} = \langle [h_0] \rangle < G$ where $\mathrm{ord}([h_0]) = \infty$. We may assume that $h_0 \in \mathrm{Homeo}(\Sigma_g)$ is a canonical representative of its mapping class. Because $\langle [h_0] \rangle$ is normal in G, conjugation with an element g acts as a group isomorphism of this subgroup. Therefore every element $g \in G$ obeys

$$g[h_0]g^{-1} = [h_0] \quad \text{or} \quad g[h_0]g^{-1} = [h_0]^{-1},$$

so we may choose the representative of $h_g \in g$ to be the unique element that normalizes h_0 exactly. This gives us a candidate for a realization. We must check whether the relations of G (besides these conjugations) hold for our realization. These relations can very generally be written as $g_1 g_2 = g_3$ for $g_1, g_2, g_3 \in G$. We prove that such a relation is realized correctly using the actions of the h_{g_i} on h_0. There are three possible combinations of these actions. For brevity we only tackle the case $h_{g_i} h_0 h_{g_i}^{-1} = h_0$ for $i = 1, 2, 3$, the others are analogous. We have $h_{g_1} h_{g_2} \simeq_i h_{g_3}$. Also,

$$h_0 = h_{g_1}(h_{g_2} h_0 h_{g_2}^{-1}) h_{g_1}^{-1} = (h_{g_1} h_{g_2}) h_0 (h_{g_1} h_{g_2})^{-1} \text{ and also } h_0 = h_{g_3} h_0 h_{g_3}^{-1}.$$

Because the exactly normalizing element is unique in its mapping class, we conclude that

$$h_{g_1} h_{g_2} = h_{g_3}.$$

So indeed we have a realization, on the assumption of a unique normalizing element.

6.4 Negative results

Cohomological obstructions. Morita gave the first negative result on the Nielsen realization problem in his complicated 1987 paper [29]. There he proved that for $g \geq 18$, there is no representation of $\mathrm{MCG}_+(\Sigma_g)$ in $\mathrm{Diff}_+(\Sigma_g)$. Afterwards, he improved his methods and got the same result — in fewer pages — for all $g \geq 5$ in 2001, see [30].

His proofs use oriented surface bundles. An oriented surface bundle is a principal fiber bundle $\pi : E \to M$ of oriented manifolds, with fiber Σ_g and an action $\phi : E \times \mathrm{Diff}_+(\Sigma_g) \to E$ which preserves fibers. The structure group of an oriented surface bundle is $\mathrm{Diff}_+(\Sigma_g)$. We will call such a bundle a Σ_g-bundle.

Definition 6.9 *A* characteristic class of a Σ_g-bundle *is a map α that yields, for some abelian group A and some $k \in \mathbb{Z}$, an element $\alpha(\pi) \in H^k(M, A)$ when we give a Σ_g-bundle $\pi : E \to M$ as input, such that for a bundle map between Σ_g-bundles*

$$\begin{array}{ccc} E_1 & \xrightarrow{\tilde{f}} & E_2 \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle \pi_2} \\ M_1 & \xrightarrow[f]{} & M_2 \end{array}$$

we have $\alpha(\pi_1) = f^*(\alpha(\pi_2))$.[3]

Now Morita comes up with a whole bunch of characteristic classes $\{e_i\}_{i\in\mathbb{N}}$ for a Σ_g-bundle, as follows. For a Σ_g-bundle $\pi : E^{n+2} \to M^n$, consider the subbundle of the tangent bundle of E called the vertical bundle $\xi : F \to E$ of π, defined as the kernel of $\pi_* : TE \to TM$. This is a 2-dimensional real vector bundle over E. Its Euler class is

$$\chi(\xi) := (\imath_E^F)^* D (\imath_E^F)_*([E]) \in H^2(E, \mathbb{Z}),$$

where $[E] \in H_{n+2}(E, \mathbb{Z})$ is the fundamental class and $D : H_{n+2}(F, \mathbb{Z}) \to H^2(F, \mathbb{Z})$ is the Poincaré-Lefschetz duality map. He then uses the Gysin homomorphism (integration along the fiber)

$$\mathrm{Gys}_\pi : H^{2(i+1)}(E, \mathbb{Z}) \to H^{2i}(M, \mathbb{Z})$$

and defines

$$e_i := \mathrm{Gys}(\chi(\xi)^{i+1}) \in H^{2i}(M, \mathbb{Z}).$$

The main part of his paper is devoted to proving that these classes are non-trivial. On the other hand, he proves that for the natural projection $p : \mathrm{Diff}_+(\Sigma_g) \to \mathrm{MCG}_+(\Sigma_g)$, the induced map $p^* : H^*(\mathrm{MCG}_+(\Sigma_g), \mathbb{Q}) \to H^*(\mathrm{Diff}_+(\Sigma_g), \mathbb{Q})$ on group cohomology annihilates e_i for $i \geq 3$, which is to say $p^*(e_i) = 0$. But this implies there can not be a map $s : \mathrm{MCG}_+(\Sigma_g) \to \mathrm{Diff}_+(\Sigma_g)$ for which $p \circ s = 1_{\mathrm{MCG}(\Sigma_g)}$, since this would entail $0 = s^* \circ p^*(e_i) = 1_{H^*(\mathrm{MCG}(\Sigma_g),\mathbb{Q})}(e_i) = e_i$ for $i \geq 3$, contradicting the non-triviality of these characteristic classes.

Decompositional obstructions. In his recent paper [27] (of which only a preprint is available yet), Marković proves that for $g \geq 6$, $\mathrm{MCG}_+(\Sigma_g)$ can not be realized in $\mathrm{Homeo}_+(\Sigma_g)$. This is more general than Morita's result, which can not be generalized to the topological category;

[3] In category-theoretical terms a characteristic class is a natural transformation between the functor $b_{\mathrm{Diff}_+(\Sigma_g)} : \mathrm{Top} \to \mathrm{Set}$ which associates to a topological space X the set of all Σ_g-bundles over it and the functor $H^* : \mathrm{Top} \to \mathrm{Set}$ which associates to a topological space X the set of all its cohomology classes.

Thurston proved that no cohomological obstructions exist there. Marković employs a wholly different technique, using suitable decompositions of a surface. I will present a short sketch of his paper, which might function as a guidance when reading his work. Note that at some points, I use different notation which I think is clearer.

Definition 6.10 *An* upper semi-continuous decomposition *of a surface Σ is a partition $\mathbf{S}$ of Σ into closed, connected subsets, such that*

1. *no $S \in \mathbf{S}$ separates Σ;*

2. *for a sequence $(S_n)_{n=1}^{\infty}$ of elements of $\mathbf{S}$ with Hausdorff limit S_0, there is an $S \in \mathbf{S}$ such that $S_0 \subseteq S$.*

Here a sequence of closed and bounded subsets $(S_n)_{n=1}^{\infty}$ of a metric space X is said to have Hausdorff limit S_0 if this sequence converges to the closed and bounded set S_0 in the power set 2^X equipped with the Hausdorff metric. The second condition is equivalent to the condition

2.′ *for every $S \in \mathbf{S}$ and every open $U \supset S$, there is an open set V such that $U \supset V \supset S$ and V is a union of elements of $\mathbf{S}$.*

Definition 6.11 *A closed connected subset S of a surface Σ is called* acyclic *if there is an open disc U in Σ containing S, such that $U - S$ is homeomorphic to an annulus. (This is equivalent to the general definition that the homology $\tilde{H}_*(S) = 0$.)*

For an upper semi-continuous decomposition $\mathbf{S}$, we define $\mathbf{S}_{ac}$ to be the union of acyclic elements of $\mathbf{S}$. It turns out that every component of $\mathbf{S}_{ac}$ is a subsurface of Σ_g.

Definition 6.12 *Let $G \leq \mathrm{Homeo}_+(\Sigma_g)$ be a group of homeomorphisms. We call an upper semi-continuous decomposition of Σ_g* admissible for G *if*

1. *for all $f \in G$ and all $S \in \mathbf{S}$ we have $f(S) = S$;*

2. *for any $S \in \mathbf{S}$, every point in a frontier component of $\Sigma_g - S$ is a limit of points from $\Sigma_g - S$ belonging to acyclic elements of $\mathbf{S}$.*

We can define a partial order on the collection $\mathcal{S}(G)$ of all admissible decompositions for a given group $G \leq \mathrm{Homeo}_+(\Sigma_g)$ by setting $\mathbf{S}_\alpha \leq \mathbf{S}_\beta$ if every element of $\mathbf{S}_\alpha$ is contained in an element of $\mathbf{S}_\beta$. Marković proves that, given two admissible decompositions $\mathbf{S}_1$ and $\mathbf{S}_2$ of Σ_g for G, there is a common lower bound $\mathbf{S}_1 \cap \mathbf{S}_2 \in \mathcal{S}(G)$. The collection $\mathcal{S}(G)$ is thus a directed set. Also, any chain in this partial order has a lower bound. He uses these facts to prove:

Theorem 6.13 *For every $G \leq \mathrm{Homeo}(\Sigma_g)$ the partial order $(\mathcal{S}(G), \leq)$ has a smallest element.*

Proof. By Hausdorff's maximal principle, which is equivalent to the axiom of choice, any chain in $\mathcal{S}(G)$ can be extended to a maximal chain. So let us start with a chain of one random element of $\mathcal{S}(G)$ and extend it to a maximal chain. Now we take the lower bound $\mathbf{S}$ of this chain (which exists in $\mathcal{S}(G)$, as mentioned above) and show that this is our smallest element. Here the fact that $\mathcal{S}(G)$ is a directed set comes in. Taking any element $\mathbf{T}$, we can form $\mathbf{S} \cap \mathbf{T} \in \mathcal{S}(G)$, and $\mathbf{S} \cap \mathbf{T} \leq \mathbf{S}$. But since our chain was maximal, we must have $\mathbf{S} \cap \mathbf{T} = \mathbf{S}$, otherwise we could simply adjoin $\mathbf{S} \cap \mathbf{T}$ to our chain to get a longer one. But this implies $\mathbf{S} = \mathbf{S} \cap \mathbf{T} \leq \mathbf{T}$. Since $\mathbf{T}$ was chosen without any restriction, $\mathbf{S}$ is the smallest element of our partial order. □

Admissible decompositions are a concrete, yet refined tool to study realizations of groups of homeomorphisms. A random homeomorphism might not fix any small set of Σ_g setwise, so the minimal admissible decomposition would simply be the trivial one: $\mathbf{S} = \{\Sigma_g\}$. However, if the homeomorphisms obey certain relations, as the realizations of subgroups of $\mathrm{MCG}_+(\Sigma_g)$ must, a small admissible decomposition contains valuable information. A first hint at the connection is a theorem that Marković proves:

Theorem 6.14 *If* $\mathbf{S}$ *is an admissible decomposition for* $h : \Sigma_g \xrightarrow{\sim} \Sigma_g$*, such that* $\mathbf{S}_{ac} = \Sigma_g$*, then* $f \simeq_i 1$.

Marković considers a particular kind of homeomorphism of Σ_g. He starts with an Anosov homeomorphism A of T^2 which has a fixed point; one of the 'linear' elements gotten from $\mathrm{GL}_2(\mathbb{Z})$ as described in subsection 3.3 will do. Out of this, he creates a homeomorphism of a subsurface

$$T := T^2 - \mathrm{int}(D^2) \subset \Sigma_g$$

by a blow up procedure. He then extends this to a homeomorphism A_{Σ_g} of Σ_g. Now, using a smart covering space of Σ_g and a string of lemmata, he proves that any homeomorphism $h \simeq_i A_{\Sigma_g}$ admits a decomposition $\mathbf{S}(h)$ such that $\mathbf{S}_{ac}(h)$ contains a subsurface isotopic to T.

Figure 20. The subsurface T, on which A_{Σ_g} is the blow up of an Anosov map of the torus.

Moreover, given any homeomorphism k isotopic to a homeomorphism which is the identity on T, and such that k also commutes with h, suprisingly, $\mathbf{S}(h)$ is an admissible composition for k too. Assuming we have a realization

$$\sigma : \mathrm{MCG}_+(\Sigma_g) \to \mathrm{Homeo}(\Sigma_g)$$

of $\mathrm{MCG}_+(\Sigma_g)$, $[k]$ and $[h]$ commute, because there are representatives in their classes which are the identity on complementary subsets of Σ_g, and these representatives commute. This implies that $\sigma([k])$ and $\sigma([h])$ must also commute.

Next, Marković looks at Dehn twists. Denote the smallest admissible decomposition of a homeomorphism in the isotopy class of a twist D_α by $\mathbf{S}(D_\alpha)$. Then for any homeomorphism h:

$$\mathbf{S}(D_{h\circ\alpha}) = h(\mathbf{S}(D_\alpha)).$$

Also, $\mathbf{S}_{ac}(D_\alpha)$ contains only one component that is non-planar, and this is a subsurface of Σ_g with two ends, both homotopic to α. Denote its complement by B_α. If we have two loops α and β with $I_{\min}(\alpha,\beta) = 0$, then $[D_\alpha]$ and $[D_\beta]$ commute. It follows that their realizations must also commute, so $\sigma([D_\beta])(B_\alpha) = B_\alpha$.

Now we look at the surface Σ_g ($g \geq 3$), which can be represented as in the following figure:

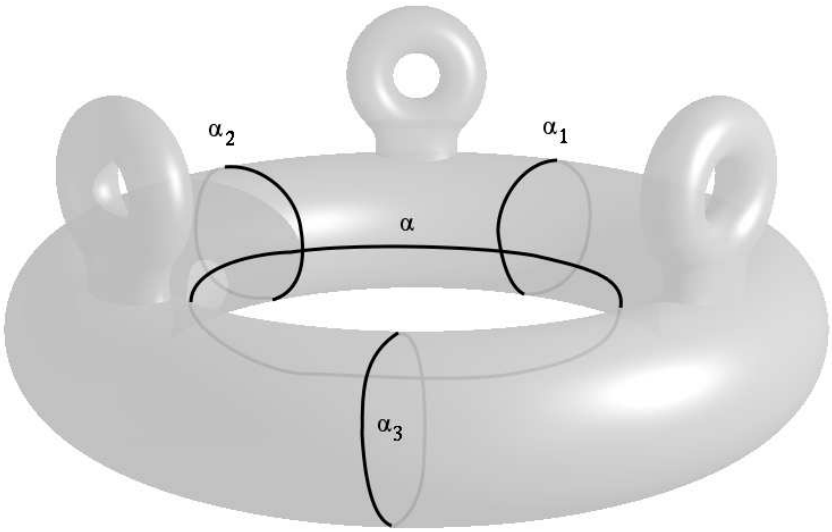


Figure 21. Marković picture of Σ_g as $g-1$ handles attached to a torus (for the case $g = 4$).

There is the obvious rotation ρ taking α_i to α_{i+1} ($i \in \mathbb{Z}_{g-1}$). Choosing a complex structure on Σ_g, we can conjugate $\sigma([\rho])$ to a conformal map by a homeomorphism isotopic to 1, according to a theorem of Hurwitz. (This uses that ρ has finite order.) Without loss of generality, we can therefore assume that $\sigma([\rho])$ is conformal, by conjugating the whole realization.

We choose a closed annulus A_α around α which contains B_α, and such that $\sigma([\rho])(A_\alpha) = A_\alpha$. Then we transfer the whole situation to the complex plane by a conformal map, sending A_α to an annulus $A(0,r,r')$ around 0 with radii $r' > r$. Now we analyze what the homeomorphisms $\sigma(D_{\alpha_i})$, restricted to A_α and conjugated to $A(0,r,r')$, do for $i = 1,\ldots,g-1$. Marković constructs a small closed set $E \supset \partial A(0,r,r') \cap B_{\alpha_1}$ such that the sets $R^i(E)$ are mutually disjoint, where R is the rotation of order $g-1$ of $A(0,r,r')$. And then he constructs a connected subset $H \subset E \cap \partial A(0,r,r')$, say in the circle of radius r, that has to contain points from at least $g-5$

sets $R^i(E)$. Considering the length of the set (interval) H, this leads to an impossibility when

$$2\pi r\frac{g-5}{g-1} \geq 2\pi r\frac{1}{g-1},$$

which is the case if $g \geq 6$.

7 Different representations of the MCG

Besides looking at realizations of a group of mapping classes as a group of concrete automorphisms, we can also consider wholly different realizations. We treat two of these. They might come in handy in future attempts to find the realizations we are after.

7.1 A matrix representation using $H_1(\Sigma_g)$

We devise a representation of $\mathrm{MCG}(\Sigma_g)$, being a realization by linear maps on some vector space, by means of the first homology group $H_1(\Sigma_g)$. This seems more natural than dawdling with the fundamental group, because it is in line with the free homotopy of curves that homotopic homeomorphims induce.

Any $h \in \mathrm{Homeo}(\Sigma_g)$ induces a chain map $h_\Delta : \Delta(\Sigma_g) \to \Delta(\Sigma_g)$ and hence a map $h_* \in \mathrm{Aut}(H_1(\Sigma_g))$. The map

$$\begin{aligned} \mu : \mathrm{Homeo}(\Sigma_g) &\to \mathrm{Aut}(H_1(\Sigma_g)) \\ h &\mapsto h_* \end{aligned}$$

is a homomorphism, since $H_1(\Sigma_g)$ is a functorial object over Σ_g. The homotopy theorem of homology theory is exactly the statement that if h_1 and h_2 are homotopic, then $(h_1)_* = (h_2)_*$. Thus μ factors to

$$\begin{aligned} \bar{\mu} : \mathrm{MCG}(\Sigma_g) &\to \mathrm{Aut}(H_1(\Sigma_g)). \\ [h] &\mapsto h_* \end{aligned}$$

Alas, it is not injective, so we have no faithful representation. For example, take the Dehn twist about any simple closed separating curve δ. (This means that $\Sigma_g - \delta$ has two components.) Many such curves exist that are not nullhomotopic, so the Dehn twist around one is not isotopic to the identity. However, the curve is nullhomologous, because a triangulation of one of the separated components yields a 2-chain of which γ is the boundary (in the homological sense). Since the chain group is commutative, every effect the map induced on it by the Dehn twist D_δ could have would be to map $c \mapsto c + k\delta$ for some $k \in \mathbb{Z}$ and any 1-chain c. However, this action is annulled in homology, so $(D_\delta)_* = 1_{H_1(\Sigma_g)}$.

To make this representation more explicit, we choose the basis $(\alpha_1, \beta_1, \ldots, \alpha_g, \beta_g)$ for $H_1(\Sigma_g)$ (see the introduction). The image $\bar{\mu}(\mathrm{MCG}_+(\Sigma_g))$ is generated by the images of the Dehn twists around α_i, β_i and γ_i as defined in section 3.4. With respect to our basis these twists map to the following matrices:

$$\bar{\mu}(D_{\alpha_i}) = \begin{pmatrix} 1 & 0 & & & & & & \\ 0 & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & 1 & 1 & & & \\ & & & 0 & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & 0 \\ & & & & & & 0 & 1 \end{pmatrix} \quad \bar{\mu}(D_{\beta_i}) = \begin{pmatrix} 1 & 0 & & & & & & \\ 0 & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & 1 & 0 & & & \\ & & & 1 & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & 0 \\ & & & & & & 0 & 1 \end{pmatrix}$$

$$\bar{\mu}(D_{\gamma_i}) = \begin{pmatrix} 1 & 0 & & & & & & & & \\ 0 & 1 & & & & & & & & \\ & & \ddots & & & & & & & \\ & & & 1 & 1 & 0 & 1 & & & \\ & & & 0 & 1 & 0 & 0 & & & \\ & & & 0 & -1 & 1 & -1 & & & \\ & & & 0 & 0 & 0 & 1 & & & \\ & & & & & & & \ddots & & \\ & & & & & & & & 1 & 0 \\ & & & & & & & & 0 & 1 \end{pmatrix}$$

A simple computation yields that for any one of the above matrices, call it A, we have

$$A^T \Omega A = \Omega, \text{ where } \Omega \text{ is the } 2g \times 2g \text{ block diagonal matrix with blocks } \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Thus the image $\bar{\mu}(\mathrm{MCG}_+(\Sigma_g))$ is contained in the group of integral symplectic matrices $\mathrm{Sp}_{2g}(\mathbb{Z})$. To generate the whole of $\mathrm{MCG}(\Sigma_g)$, we need to add one orientation-reversing automorphism, for example the involution σ shown at the end of section 3.2. It is easy to see that σ sends α_i to α_{g-i} and β_j to $-\beta_{g-j}$, so we have

$$\bar{\mu}(\sigma) = \begin{pmatrix} & & & & & & 1 & 0 \\ & & & & & & 0 & -1 \\ & & & & & \iddots & & \\ & & & 1 & 0 & & & \\ & & & 0 & -1 & & & \\ & & \iddots & & & & & \\ 1 & 0 & & & & & & \\ 0 & -1 & & & & & & \end{pmatrix}$$

which is an antisymplectic matrix, obeying

$$\bar{\mu}(\sigma)^T \Omega \bar{\mu}(\sigma) = -\Omega.$$

7.2 A realization in $\mathrm{Homeo}(\mathrm{ST}(\Sigma_g))$

We now treat a construction that is of interest to us because it enables us to construct a faithful realization of $\mathrm{MCG}(\Sigma_g)$ in $\mathrm{Homeo}(\mathrm{ST}(\Sigma_g))$. Here $\mathrm{ST}(\Sigma_g)$ is the unit tangent bundle (also called **S**pherical **T**angent bundle) of Σ_g, which is the subbundle of $T\Sigma_g$ of tangent vectors of length 1. To speak sensibly about the length of tangent vectors, we must be working on a surface with a Riemannian metric, which we will implicitly assume henceforth. The treatment is based on Gromov [14] and Casson & Bleiler [7], the general facts about hyperbolic geometry used can be found in Benedetti & Petronio [3].

To every point $p \in \mathbb{H}^n$ and tangent vector $v_p \in T_p(\mathbb{H}^n)$ corresponds a unique point $\partial(v_p) \in \partial\mathbb{H}^n$ (the boundary at infinity). This point is the intersection point of the geodesic ray defined by (p, v_p) with $S^{n-1} \subset \mathbb{R}^n$, confer section 4. Gromov attributes the following simple but useful remark to Cheeger. An orthonormal 2-frame $(v_{p,1}, v_{p,2})$ at a point $p \in \mathbb{H}^n$ gives us a triplet of points on $\partial\mathbb{H}^n$. We set

$$\begin{aligned} \mathrm{Ch} \;:\; \mathrm{St}_2(\mathbb{H}^n) &\longrightarrow \partial^3(\mathbb{H}^n) \\ (v_{p,1}, v_{p,2}) &\mapsto (\partial(-v_{p,1}), \partial(v_{p,1}), \partial(v_{p,2})). \end{aligned}$$

This is a differentiable map from the Stiefel manifold $\mathrm{St}_2(\mathbb{H}^n)$ of orthonormal 2-frames in $\mathbb{H}^n$ to $\partial^3(\mathbb{H}^n)$, which is the set of triples (x_1, x_2, x_3) with $x_1, x_2, x_3 \in \partial\mathbb{H}^n$ and $x_i \neq x_j$ for $i \neq j$. This map has a differentiable inverse: for a triple $(x_1, x_2, x_3) \in \partial^3(\mathbb{H}^n)$ there is a unique geodesic l between x_1 and x_2, and a unique perpendicular from x_3 onto l, which give us a pair $(v_{p,1}, v_{p,2})$ that map to (x_1, x_2, x_3) under Ch. Thus Ch is a diffeomorphism, the so-called *Cheeger diffeomorphism*.

Now any closed connected hyperbolic n-manifold M has a universal cover $\widetilde{M}$, which has a unique hyperbolic structure that makes the projection a local isometry. And a simply connected complete hyperbolic manifold is known to be homeomorphic to $\mathbb{H}^n$. So we have $M = \mathbb{H}^n/\Gamma$, where $\Gamma \cong \pi_1(M)$ is the group of deck transformations of the cover $\pi : \widetilde{M} \to M$, acting by isometries. We now invoke a general result whose proof can be found in [3], Propositions C.1.2 and C.1.8.[4]

Lemma 7.1 *Let $h : M^n \to M^n$ be a homeomorphism of a closed orientable hyperbolic manifold and $\tilde{h} : \mathbb{H}^n \to \mathbb{H}^n$ a lift to the universal cover. Then $\tilde{h}$ has a unique continuous extension to $\mathbb{H}^n \cup \partial\mathbb{H}^n$ and the restriction of this extension to $\partial\mathbb{H}^n$ is a homeomorphism of $\partial\mathbb{H}^n$.* □

The action of an element of Γ on $\mathbb{H}^2$ can be viewed as a special case of this, namely a lift of id_M. Thus this action extends to the boundary $\partial\mathbb{H}^n$ and hence to $\partial^3(\mathbb{H}^n)$ by defining $\gamma((x_1, x_2, x_3)) := (\gamma x_1, \gamma x_2, \gamma x_3)$ for any $\gamma \in \Gamma$ and $(x_1, x_2, x_3) \in \partial^3(\mathbb{H}^n)$. Our map Ch can now be defined on $\mathrm{St}_2(M)$ by taking a lift of an orthonormal frame to $\mathrm{St}_2(\mathbb{H}^n)$ with the covering map $T\pi$ (which is locally a bundle isomorphism of $\mathrm{St}_2(\mathbb{H}^n) \to \mathrm{St}_2(M)$) and then applying the Cheeger diffeomorphism. Since any two lifts differ by an element of Γ — which acts on $\mathrm{St}_2(\mathbb{H}^n)$ by the tangent

[4] The result is stated there for quasi-isometries, a class of maps to which the lifts of homeomorphisms of compact connected orientable hyperbolic manifolds belong.

maps of its elements — the image in $\partial^3(\mathbb{H}^n)$ is defined up to the action of Γ on this space and we get the Cheeger diffeomorphism

$$\mathrm{Ch} : \mathrm{St}_2(M) \xrightarrow{\sim} \partial^3(\mathbb{H}^n)/\Gamma.$$

The Cheeger diffeomorphism comes in handy when we have a homeomorphism $h : M \to M$. The homeomorphism lifts to the universal cover $\mathbb{H}^n$, as mentioned before, but not uniquely. A lift $\tilde{h}$ is determined by the image $\tilde{h}(p) \in \pi^{-1}(h(p))$ of a single point $p \in \mathbb{H}^n$. Because of the previous lemma, any lift induces a unique map on $\partial\mathbb{H}^n$, and two lifts differ by an action of Γ, which also extends to the boundary sphere. Thus h induces a unique (i.e. independent from the choice of lift) map $\hat{h}$ from $\partial^3(\mathbb{H}^n)/\Gamma \to \partial^3(\mathbb{H}^n)/\Gamma$. The Cheeger diffeomorphism implies that this is equivalent to an induced map

$$\hat{h} : \mathrm{St}_2(M) \xrightarrow{\sim} \mathrm{St}_2(M).$$

We can define the map $C : \mathrm{Homeo}(M) \to \mathrm{Homeo}(\mathrm{St}_2(M))$ by $h \mapsto \hat{h}$. This map is a group homomorphism. Given homeomorphisms $h_1, h_2 : M \to M$ with lifts to $\mathbb{H}^n$ chosen as $\tilde{h}_1$ and $\tilde{h}_2$, we can choose $\widetilde{h_2 h_1} := \tilde{h}_2\tilde{h}_1$ as a lift of h_2h_1. By lemma 7.1 the extensions to $\partial\mathbb{H}^n$ obey the same law, from which it is immediate that $\widehat{h_2h_1} = \hat{h}_1\hat{h}_2$. We will now prove that the homomorphism C only depends on the homotopy class of h.

Lemma 7.2 *Let $h_1, h_2 : M \to M$ be homotopic homeomorphisms of a closed orientable hyperbolic manifold. For any lift $\tilde{h}_1$ of h_1 there is a lift $\tilde{h}_2$ of h_2 to the universal cover $\mathbb{H}^n$ that extends to the same map on $\partial\mathbb{H}^n$.*

Proof. Let $H : M \times I \to M$ be a homotopy between h_1 and h_2. This homotopy can be lifted to the universal cover with one point prescribed. If we choose the lift $\tilde{H} : \mathbb{H}^n \times I \to \mathbb{H}^n$ so that $\tilde{H}(p, 0) = \tilde{h}_1(p)$, then $\tilde{H}(\cdot, 0) = \tilde{h}_1$ will hold, since lifts for h_1 are uniquely determined by the choice of one image point. The map $\tilde{H}(\cdot, 1)$ is a lift of h_2, and we now show this is the desired lift. As H is uniformly continuous with respect to the hyperbolic metric on M, the hyperbolic lengths of the paths $\tilde{H}(p, \cdot)$ are uniformly bounded. Therefore the Euclidean distance (in the Poincaré model of $\mathbb{H}^n$) between $\tilde{H}(p, 0)$ and $\tilde{H}(p, 1)$ tends to 0 as p tends to $\partial\mathbb{H}^n$. Hence the induced maps on $\partial\mathbb{H}^n$ are the same. □

The map $\hat{h}$ is by construction independent of the choice of lift of h. This means that for homotopic homeomorphisms h_1, h_2, we may freely choose the lifts as in the foregoing proof, whence the induced maps $\hat{h}_1$ and $\hat{h}_2$ are seen to be equal. That is, the homomorphism C factors through to

$$\bar{C} : \mathrm{MCG}(M) \to \mathrm{Homeo}(\mathrm{St}_2(M)).$$

In the special case of a closed orientable hyperbolic surface Σ_g ($g \geq 2$), the space $\mathrm{St}_2(\Sigma_g)$ has two components. If we orient Σ_g, we may distinguish these as positively and negatively oriented frames and accordingly write $\mathrm{St}_2(\Sigma_g) = \mathrm{St}_{2,+}(\Sigma_g) \amalg \mathrm{St}_{2,-}(\Sigma_g)$. For the sake of concreteness, imagine $\Sigma_g = \mathbb{H}^2/\Gamma$ with $\mathbb{H}^2$ the standard Poincaré disc, and regard right-hand frames to be positively oriented.

Likewise, the space $\partial^3(\mathbb{H}^2)$ splits as $\partial^3_+(\mathbb{H}^2) \amalg \partial^3_-(\mathbb{H}^2)$, where the two component names denote the spaces of triples $(x_1, x_2, x_3) \in \partial^3(\mathbb{H}^2)$ that lie in counterclockwise and clockwise order, respectively. Since our surface is orientable, the group Γ consists only of orientation preserving isometries, and so the homeomorphisms it induces on $\partial\mathbb{H}^2$ have the same property. The action of Γ on $\partial^3(\mathbb{H}^2)$ therefore preserves (counter)clockwiseness and the resulting quotient space $\partial^3(\mathbb{H}^2)/\Gamma$ splits as $\partial^3_+(\mathbb{H}^2)/\Gamma \amalg \partial^3_-(\mathbb{H}^2)/\Gamma$ and the Cheeger diffeomorphism maps $\mathrm{St}_{2,+}(\Sigma_g)$ to $\partial^3_+(\mathbb{H}^2)/\Gamma$ and $\mathrm{St}_{2,-}(\Sigma_g)$ to $\partial^3_-(\mathbb{H}^2)/\Gamma$.

In dimension 2 a tangent vector uniquely determines a positively oriented orthonormal 2-frame with this tangent vector as the first of the pair. In the same way, it uniquely determines a negatively oriented orthonormal 2-frame. This gives us canonical diffeomorphisms

$$\mathrm{Fr}_+ : \mathrm{ST}(\Sigma_g) \to \mathrm{St}_{2,+}(\Sigma_g)$$
$$\mathrm{Fr}_- : \mathrm{ST}(\Sigma_g) \to \mathrm{St}_{2,-}(\Sigma_g).$$

For a homeomorphism $h : \Sigma_g \to \Sigma_g$ we would like to construct an induced homeomorphism $\bar{h} : \mathrm{ST}(\Sigma_g) \to \mathrm{ST}(\Sigma_g)$, like we did above for $\mathrm{St}_2(M)$, with the same nice properties: independence under isotopy and preservation of the group structure. In order to do this, we would like to restrict the induced homeomorphism on $\mathrm{St}_2(\Sigma_g)$ to one of its components, but this can not always be done. The map h can be lifted to $\mathbb{H}^2$ and then extended to $\partial\mathbb{H}^2$ and thereby to $\partial^3(\mathbb{H}^2)$. But this action can only be restricted to $\partial^3_+(\mathbb{H}^2)$ or $\partial^3_-(\mathbb{H}^2)$ if $\tilde{h}$ is orientation preserving, which is the case exactly when h is. We must then also choose to which component we restrict. An orientation reversing $\tilde{h}$ leads to a swap of the components of $\partial^3(\mathbb{H}^2)$, giving two restricted maps back and forth. The figure below gives an overview of the situation.

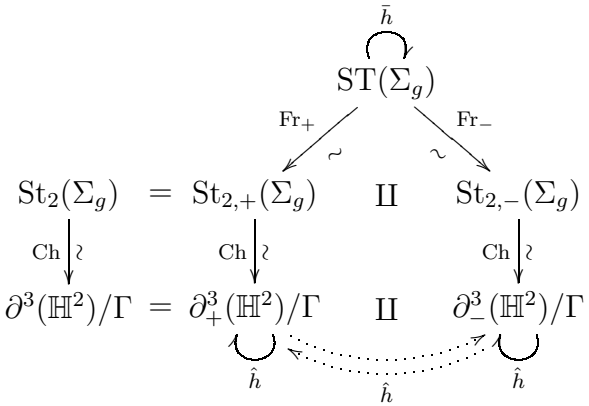


Figure 22. The maps $\hat{h}$ for orientation preserving h have been drawn with solid arrows in the figure, the ones for orientation reversing h are dashed. We want to induce a map $\bar{h}$ on $\mathrm{ST}(\Sigma_g)$.

Remark. It might have occurred to the reader that there is a simpler way to induce a homeomorphism from $\mathrm{ST}(\Sigma_g) \to \mathrm{ST}(\Sigma_g)$, namely the normalized differential

$$v_p \mapsto T_p(h)(v_p)/||T_p(h)(v_p)|| \text{ for } v_p \in \mathrm{ST}(\Sigma_g).$$

But of course this map depends on this precise mapping of directions and would not allow factorization through to $\mathrm{MCG}(\Sigma_g)$.

The problem of reconciling the induced maps $\hat{h}$ for orientation reversing h with the induced maps for orientation preserving h is not easily solved. Reconsidering the problem from the viewpoint of $\mathrm{St}_2(\Sigma_g)$, we might look at the restriction $\hat{h}|_{\mathrm{St}_{2,+}(\Sigma_g)} : \mathrm{St}_{2,+}(\Sigma_g) \to \mathrm{St}_{2,-}(\Sigma_g)$ of a map induced on the whole of $\mathrm{St}_2(\Sigma_g)$. We could try to compose it with some fixed homeomorphism, e.g. the canonical frame mirroring map $m : \mathrm{St}_{2,-}(\Sigma_g) \to \mathrm{St}_{2,+}(\Sigma_g)$ given by $(v_1, v_2) \mapsto (v_1, -v_2)$. Or we could 'square' the map by composing it with $\hat{h}|_{\mathrm{St}_{2,-}(\Sigma_g)}$ to obtain $(\hat{h} \circ \hat{h})|_{\mathrm{St}_{2,+}(\Sigma_g)}$. In both cases we end up with induced maps on $\mathrm{St}_{2,+}(\Sigma_g)$, but the group structure is not preserved.
The only sensible thing to do is to restrict our attention to orientation preserving homeomorphisms. (In Kirby's problem list [22], this is not stated clearly.) We still have to make the choice which component to use, giving two distinct induced maps. If we choose $\mathrm{St}_{2,+}(\Sigma_g)$, we can lift $\tilde{h}$ to

$$\bar{h} := \mathrm{Fr}_+^{-1} \circ \hat{h} \circ \mathrm{Fr}_+ : \mathrm{ST}(\Sigma_g) \xrightarrow{\sim} \mathrm{ST}(\Sigma_g).$$

Obviously, the induced maps $\bar{h}$ depend only on the isotopy class of h and obey the group law $\overline{h_2 h_1} = \bar{h}_2 \bar{h}_1$. Moreover, for different isotopy classes, the induced maps are different, proving that we have a faithful group realization. For suppose that $\hat{h}_1 = \hat{h}_2 : \partial^3_+(\mathbb{H}^2)/\Gamma \to \partial^3_+(\mathbb{H}^2)/\Gamma$. Let c be an oriented simple closed curve on Σ_g and $\tilde{c}$ an oriented lift of this curve to $\mathbb{H}^2$. If $\tilde{c}$ has endpoints $p_1, p_2 \in \partial\mathbb{H}^2$ (and is oriented from p_1 to p_2), then $\tilde{h}_1(p_i) = \gamma(\tilde{h}_2(p_i))$ for $i = 1, 2$ and a certain $\gamma \in \Gamma$, because h_1 and h_2 induce the same action on pairs (even triplets) of points on $\partial\mathbb{H}^2$ up to an action of Γ. This means that $\tilde{h}_1 \circ \tilde{c}$ and $\gamma \circ \tilde{h}_2 \circ \tilde{c}$ have the same ordered pair of endpoints, but these are lifts of $h_1 \circ c$ and $h_2 \circ c$, respectively. This implies that the latter two curves are homotopic, see Casson & Bleiler [7, lemma ???]. Remember our closed curve c was arbitrary. From the proof of theorem 2.11 it follows that $h_1 \simeq_i h_2$.

One more remark on orientations. The space $\partial^3(\mathbb{H}^2)$ is orientable, because it is a subspace of $\partial\mathbb{H}^2 \times \partial\mathbb{H}^2 \times \partial\mathbb{H}^2$ and this space inherits a canonical product orientation from $\partial\mathbb{H}^2$. Since Γ acts freely on $\partial^3(\mathbb{H}^2)$ by orientation preserving diffeomorphisms, $\mathrm{St}_2(\Sigma_g) \cong \partial^3(\mathbb{H}^2)/\Gamma$ is orientable as well, inheriting an orientation (induced locally, if you will, from the total space $\partial^3(\mathbb{H}^2)$ of this covering). And this means the same goes for $\mathrm{ST}(\Sigma_g)$, being homeomorphic to a component of $\mathrm{St}_2(\Sigma_g)$.

Now if $h : \Sigma_g \to \Sigma_g$ is orientation preserving, so is $\tilde{h} : \mathbb{H}^2 \cup \partial\mathbb{H}^2 \to \mathbb{H}^2 \cup \partial\mathbb{H}^2$, whence also $\tilde{h} : \partial^3(\mathbb{H}^2) \to \partial^3(\mathbb{H}^2)$ and $\tilde{h} : \partial^3_+(\mathbb{H}^2)/\Gamma \to \partial^3_+(\mathbb{H}^2)/\Gamma$. To induce $\bar{h} : \mathrm{ST}(\Sigma_g) \to \mathrm{ST}(\Sigma_g)$, we conjugate this last map with $\mathrm{Ch} \circ \mathrm{Fr}_+$. But this implies that $\bar{h}$ is also orientation preserving, since the degree $\deg$ of maps between connected orientable topological manifolds satisfies $\deg(g \circ f) = \deg(g)\deg(f)$, and for a homeomorphism h we have $\deg(h) \in \{-1, 1\}$, indicating whether the homeomorphism is orientation preserving ($\deg(h) = 1$) or orientation reversing ($\deg(h) = -1$). Summarizing, we have proved the following theorem.

Theorem 7.3 (Cheeger, Gromov) *An orientation preserving homeomorphism $h : \Sigma_g \to \Sigma_g$ of closed oriented hyperbolic surfaces induces an orientation preserving homeomorphism $\bar{h} : \mathrm{ST}(\Sigma_g) \to \mathrm{ST}(\Sigma_g)$. The map*

$$S : \mathrm{Homeo}_+(\Sigma_g) \to \mathrm{Homeo}_+(\mathrm{ST}(\Sigma_g))$$

given by $h \mapsto \bar{h}$ is a group homomorphism that factors through to the faithful realization

$$\bar{S} : \mathrm{MCG}_+(\Sigma_g) \to \mathrm{Homeo}_+(\mathrm{ST}(\Sigma_g)).$$

Bibliography

[1] Baer, R. (1927), Kurventypen auf Flächen, *Journal der reine und angewandte Mathematik* 156, pp. 231–246.

[2] Baer, R. (1928), Isotopie von Kurven auf orientierbaren, geschlossenen Flächen und ihr Zusammenhang mit der topologischen Deformation der Flächen, *Journal der reine und angewandte Mathematik* 159, pp. 101–111.

[3] Benedetti, R., Petronio, C. (1992), *Lectures on hyperbolic geometry*, Universitext, Springer-Verlag: Berlin etc.

[4] Birman, J.S. (1974), *Braids, links and mapping class groups*, Princeton University Press: Princeton.

[5] Bredon, G.E. (1993), *Topology and geometry*, Graduate Texts in Mathematics 139, Springer-Verlag: Berlin etc.

[6] Brendle, T.E., Farb, B. (2004), Every mapping class group is generated by 6 involutions, *Journal of Algebra* Vol. 278 Issue 1, pp. 187–198.

[7] Casson, A.J., Bleiler, S.A. (1988), *Automorphisms of surfaces after Nielsen and Thurston*, London Mathematical Society Student Texts 9, Cambridge University Press: Cambridge.

[8] Cerf, J. (1970), *La stratification naturelle des espaces de fonctions différentiables réelles et le théorème de la pseudo-isotopie*, Publications mathématiques de l'I.H.É.S., tome 39.

[9] Dehn, M. (1938), Die Gruppe der Abbildungsklassen, *Acta Mathematica* 69, pp. 135–206.

[10] Eckmann, B., Müller, H. (1982), Plane motion groups and virtual Poincaré duality of dimension two, *Inventiones Mathematicae* 69, pp. 293–310.

[11] Epstein, D.B.A. (1966), Curves on 2-manifolds and isotopies, *Acta Mathematica* 115, pp. 83–107.

[12] Fathi, A., Laudenbach, F., Poénaru, V. (1979), *Travaux de Thurston sur les surfaces* (séminaire Orsay), Astérisque 66–67, Société Mathematique de France: Paris.

[13] Fenchel, W. (1948), Estensioni di gruppi discontinui e transformazioni periodiche delle superficie, *Rendiconti della Accademia Nazionale dei Lincei* 5, pp. 326–329.

[14] Gromov, M. (2000), Three remarks on geodesic dynamics and fundamental groups, *L'Enseignement Mathematique* 46, pp. 391–402.

[15] Guillemin, V., Pollack, A. (1974), *Differential topology*, Prentice-Hall Inc.: Englewood Cliffs, New Jersey etc.

[16] Hass, J., Scott, P. (1985), *Intersections of curves on surfaces*, Israel Journal of Mathematics 51, pp. 90–120.

[17] Hirsch, M.W. (1976), *Differential topology*, Graduate Texts in Mathematics 33, Springer-Verlag: Berlin etc.

[18] Hopf, H. (1943), Enden offener Räume und unendliche diskontinuierliche Gruppen, *Commentarii Mathematici Helvetici* 16, pp. 81–100.

[19] Humphries, S.P. (1977), Generators for the mapping class group, in: *Topology of low-dimensional manifolds*, Lecture Notes in Mathematics 722, Springer-Verlag: Berlin etc.

[20] Ivanov, N.V. (2001), Mapping class groups, in: *The handbook of geometric topology*, edited by Sher, R.B. and Daverman, R.J., North-Holland: Amsterdam.

[21] Kerckhoff, S.P. (1983), The Nielsen realization problem, *Annals of Mathematics (2)* 117, pp. 235–265.

[22] Kirby, R. (1993), Problems in low-dimensional topology, in: *Geometric topology* (part 2), edited by Kazez, W.H., American Mathematical Society, Providence (RI). (An update is available via Kirby's webpage, http://math.berkeley.edu/∼kirby/.)

[23] Kobayashi, S. (1972), *Transformation groups in differential geometry*, Ergebnisse der Mathematik 70, Springer-verlag: Berlin etc.

[24] Lickorish, W.B.R. (1964), A finite set of generators for the homeotopy group of a 2-manifold, *Proceedings of the Cambridge Philosophical Society* 60, pp. 769–778.

[25] Lickorish, W.B.R. (1966), A finite set of generators for the homeotopy group of a 2-manifold (corrigendum), *Proceedings of the Cambridge Philosophical Society* 62, pp. 679–681.

[26] Mangler, W. (1939), Die Klassen von topologischen Abbildungen einer geschlossenen Fläche auf sich, *Mathematische Zeitschrift* 44, pp. 541–554.

[27] Marković, V. (2006), *Realization of the mapping class group by homeomorphisms* (preprint), see http://www.maths.warwick.ac.uk/∼markovic/.

[28] Massey, W.S. (1967), *Algebraic topology: an introduction*, Graduate Texts in Mathematics 56, Springer-Verlag: Berlin etc.

[29] Morita, S. (1987), Characteristic classes of surface bundles, *Inventiones Mathematicae* 90, pp 551–577.

[30] Morita, S. (2001), *Geometry of characteristic classes*, Translations of mathematical monographs 199, Iwanami Series in Modern Mathematics, American Mathematical Society: New York.

[31] Munkres, J.R. (1960), Differentiable isotopies on the 2-sphere, *Michigan Mathematical Journal* 7, pp. 193–197.

[32] Munkres, J.R. (1960), Obstructions to the smoothing of piecewise-differentiable homeomorphisms, *Annals of Mathematics* 72, pp. 521–554.

[33] Munkres, J.R. (1966), *Elementary differential topology*, Annals of mathematics studies 54, Princeton University Press: Princeton.

[34] Munkres, J.R. (2000), *Topology*, 2nd edition, Prentice Hall: New York.

[35] Nielsen, J. (1927), Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen, *Acta Mathematica* 50, pp. 189–358.

[36] Nielsen, J. (1936), Topologischer Beweis eines Satzes von Wiman, *Mathematische Tidsskrift* 1936 B, pp. 11–24.

[37] Nielsen, J. (1942), Abbildungsklassen endlicher Ordnung, *Acta Mathematica* 75, pp. 23–115.

[38] Scott, P., Wall, C.T.C. (1979), Topological methods in group theory, in: *Homological group theory*, Proceedings of a symposium, held at Durham in september 1977, London Mathematical Society Lecture Note Series 36, Cambridge University Press: Cambridge.

[39] Smale, S. (1959), Diffeomorphisms of the 2-sphere, *Proceedings of the American Mathematical Society* 10, pp. 621–626.

[40] Stillwell, J. (1995), *Classical topology and combinatorial group theory*, second edition, Graduate Texts in Mathematics 72, Springer-Verlag: Berlin etc.

[41] Thurston, W.P. (1988), On the geometry and dynamics of diffeomorphisms of surfaces, *Bulletin of the American Mathematical Society (new series)* 19, pp. 417–431.

[42] Wajnryb, B. (1999), An elementary approach to the mapping class group of a surface, *Geometry and Topology* 3, pp. 405–466.

[43] Zieschang, H. (1981), *Finite groups of mapping classes of surfaces*, Lecture Notes in Mathematics 875, Springer-Verlag: Berlin etc.

luctor, sed scripsi