MAPPING CLASS GROUPS AND CURVE GRAPHS

IAN BIRINGER

Abstract. These are informal notes for a Spring 2025 course at Boston College on the mapping class groups of surfaces and their curve graphs.

For the mapping class groups part, the main reference is Farb-Margalit [14], although the exposition and level of detail is different and I present some material from different perspectives. For instance, the discussion of the Birman exact sequence is initially done in a more elementary way, and more details/context about the homotopy exact sequence proof are given, and pseudo-Anosov maps are introduced first through half-translation surfaces. For the curve graph, we give basic examples (including the Farey graph), talk a bit about Gromov hyperbolicity, prove the curve graph is infinite diameter and hyperbolic (via bicorn curves), and then talk informally about various properties of the mapping class group that are similar to properties of hyperbolic groups. We prove the mapping class group is finitely generated and briefly discuss subsurface projection and the Masur-Minsky distance formula. If you're reading this, let me know if you find any errors!

1. Mapping class groups of manifolds

Let M be a manifold, which we basically always assume is connected and orientable, unless otherwise specified. Consider the group

$$\mathrm{Homeo}(M) := \{ \text{ homeomorphisms } M \longrightarrow M\},$$

under composition. This is a *huge* group. It's uncountable as long as $\dim M \geq 1$, and is very complicated algebraically. For instance:

Open Question 1.1. *Is it true that if* $\dim M \geq 2$*, and* Γ *is any finitely generated torsion free group, there's an injective homomorphism* $\Gamma \hookrightarrow \mathrm{Homeo}(M)$*?*

It's necessary to include something like the torsion-free assumption, since for any given M and prime p there's some k such that $(\mathbb{Z}/p\mathbb{Z})^k$ doesn't embed in $\mathrm{Homeo}(M)$, see [33]. See also Fisher [17] for a survey of related topics.

We endow $\mathrm{Homeo}(M)$ with the compact-open topology, which turns it into a topological group. If we give M a (Riemannian, say) metric d_M, the compact-open topology of $\mathrm{Homeo}(M)$ is the same as that induced by the sup metric

$$d(f,g) = \sup_{x \in M} d_M(f(x), g(x)).$$

Let $\mathrm{Homeo}_0(M) \subset \mathrm{Homeo}(M)$ be the path component of the identity.

Claim 1.2. *If* G *is a topological group and* $G_0 \subset G$ *is the (path) component of the identity, then* G_0 *is a normal subgroup of* G*.*

Proof. The group operations are continuous, and $G_0 \times G_0$ is path connected, so the same is true of its image under the multiplication map $m : G \times G \longrightarrow G$. Since the image contains id, it's contained in G_0, and hence G_0 is closed under multiplication.

Inversion is similar. For normality, fix $g \in G$ and perform a similar argument using the (continuous) conjugation map $x \mapsto gxg^{-1}$. $\square$

Let $\mathrm{Homeo}_+(M)$ be the group of orientation preserving (o.p.) homeomorphisms of M. We then define the following two groups:

$$\mathrm{Map}(M) := \mathrm{Homeo}_+(M)/\mathrm{Homeo}_0(M),$$
$$\mathrm{Map}^{\pm}(M) := \mathrm{Homeo}(M)/\mathrm{Homeo}_0(M)$$

These groups are the *mapping class group* of M and the *extended mapping class group* of M, respectively. I apologize for the notation, realizing you might prefer to use $\mathrm{Map}(M)$ to refer to the second quotient. However, in the setting of surfaces (where we'll mostly be working), it's typical to define the 'mapping class group' only using orientation preserving homeomorphisms. Note that $\mathrm{Map}(M)$ is an index 2 normal subgroup in $\mathrm{Map}^{\pm}(M)$, and that in the two quotients above, two homeomorphisms are identified exactly when they're isotopic.

Example 1.3. *In zero dimensions,* $\mathrm{Map}^{\pm}(\{p_1, \dots, p_n\}) \cong S_n$, *since the isotopy relation is trivial so we are just looking at bijections of an n-point set.*

Example 1.4. $\mathrm{Map}^{\pm}(\mathbb{R}) \cong \mathbb{Z}/2\mathbb{Z}$, *and similarly for S^1. To see this, note that since there are orientation reversing homeomorphisms of $\mathbb{R}$, the map*

$$\mathrm{Homeo}(\mathbb{R}) \longrightarrow \mathbb{Z}/2\mathbb{Z}$$

given by orientation is surjective, so we just have to check that its kernel $\mathrm{Homeo}^+(\mathbb{R})$ *is equal to* $\mathrm{Homeo}_0(\mathbb{R})$. *But a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is in* $\mathrm{Homeo}^+(\mathbb{R})$ *if and only if it's continuous, increasing and* $\lim_{x\to\pm\infty} f(x) = \pm\infty$. *And if f is, then so is*

$$f_t := tf + (1-t)id$$

for all $t \in [0,1]$; for instance, if $x < y$ then

$$f_t(x) = tf(x) + (1-t)x < tf(y) + (1-t)y = f_t(y).$$

So, (f_t) is an isotopy from f to id.

Here's one indication that mapping class groups are more well behaved from a group theoretic standpoint than homeomorphism groups.

Theorem 1.5. *If M is compact,* $\mathrm{Map}^{\pm}(M)$ *is countable.*

To prove this, we need two lemmas.

Lemma 1.6. *Given M compact, there's some $\epsilon = \epsilon(M, d_M)$ such that if $d(f,g) < \epsilon$, then f, g are isotopic.*

This follows for instance from the main theorem of [13], which we'll perhaps talk about later in the course. However, it's perhaps more believable than the theorem. Note also that Lemma 1.6 implies that path components in $\mathrm{Homeo}(M)$ are open, and hence that path components are the same as connected components.

Lemma 1.7. *If M is compact, then* $\mathrm{Homeo}(M)$ *is second countable.*

Proof. Given a finite subset $X \subset M$, a finite covering $\mathcal{B}$ by open sets, and a function $D : X \longrightarrow \mathcal{B}$, you can consider the neighborhood

$$\mathcal{N}(X, \mathcal{B}, D) := \{f \in \mathrm{Homeo}(M) \mid f(x) \in D(x)\ \forall x \in X\}.$$

This is a countable basis for $\mathrm{Homeo}(M)$. $\square$

We can now prove the theorem.

Proof of Theorem 1.5. Say $\mathrm{Map}^{\pm}(M)$ is uncountable. Then there's an uncountable set $A \subset \mathrm{Homeo}(M)$ such that no two elements of A are pairwise isotopic. So, by the first lemma we have $d(f, g) \geq \epsilon$ for all $f, g \in A$. But this contradicts second countability, since there should be a basis element contained in the $\epsilon/2$-neighborhood of every $f \in A$, and all these have to be distinct. $\square$

2. Mapping class groups of finite type surfaces

Let's now move up a dimension. A *finite type surface* is a surface S obtained by taking a compact surface, possibly with boundary, and removing a finite set of points. Finite type (orientable, which we'll always assume tacitly from now on) surfaces are classified by their genus, the number of boundary components, and the number of points removed, which we call *punctures.* Sometimes we may write surfaces as $S_{g,b,p}$, where the subscripts are these three numbers, but we'll try to remind the reader which are which frequently.

Claim 2.1. *A surface S has finite type $\iff$ $\pi_1 S$ is finitely generated.*

Proof. $\implies$ is easy using classification, as you can just write down generating sets, or even without, since you can get a finite generating set from the 1-skeleton of a finite triangulation, say. For $\impliedby$, find a finite embedded graph $G \subset S$ that π_1-surjects, and thicken it to a π_1-surjective compact subsurface $X \subset S$ with boundary. Using Van Kampen and the π_1-surjectivity, you can show that for every component $Y \subset S \setminus int(X)$, the intersection $\partial Y \cap X$ is a π_1-surjective boundary component of Y, and hence Y is either a disc or an annulus (possibly half-open). From there you can deduce that S is finite type. $\square$

Remark 2.2. *It follows that every surface (say, without boundary) and finitely generated fundamental group is the interior of a compact surface with boundary. In general, a manifold is* tame *if it's the interior of a compact manifold. It turns out there are non-tame, simply connected* 3*-manifolds (see the 'Whitehead manifold', which is a nested union of solid tori, each knotted in the next), and similar examples exist in higher dimensions.*

However, it turns out that a weaker statement is true for 3*-manifolds: any* 3*-manifold with finitely generated π_1 has a 'compact core', which is a compact submanifold that induces an isomorphism on π_1; this is the Scott Core Theorem* [50]*. The proof is a much more complicated version of that of the claim above. One algebraic consequence here is that all finitely generated fundamental groups of surfaces and* 3*-manifolds are automatically finitely presented. In general, a group G is called* coherent *if all finitely generated subgroups of it are finitely presented. Since covers of surfaces/*3*-manifolds are surfaces/*3*-manifolds, it follows that $\pi_1 M$ is coherent whenever M is a surface or* 3*-manifold. Similarly, free groups are coherent since subgroups of free groups are free. In contrast, $F_2 \times F_2$ is not coherent, since the kernel of the map $F_2 \times F_2 \longrightarrow \mathbb{Z}$ taking all four generators to $1 \in \mathbb{Z}$ is finitely generated but not finitely presented, see Example 9.22 in* [63]*. The group $SL_n\mathbb{Z}$ is coherent for $n = 2$ because it has a finite index free subgroup, and for $n \geq 4$ it's incoherent since you can embed $F_2 \times F_2$ in it. It's an open question of Serre whether $SL_3\mathbb{Z}$ is coherent.*

For finite type surfaces, it's useful to have a definition of a mapping class group that treats punctures and boundary components differently. So, we define

$$\mathrm{Homeo}(S, \partial S) := \{ \text{ homeomorphisms } f : S \longrightarrow S \text{ such that } f|_{\partial S} = id\}.$$

We let $\mathrm{Homeo}_0(S, \partial S)$ be its identity path component, and we let

$$\mathrm{Map}(S) := \mathrm{Homeo}^+(S, \partial S)/\mathrm{Homeo}_0(S, \partial S),$$

which in other words is the group of orientation preserving homeomorphisms of S that are the identity on the boundary, mod isotopies that are the identity on the boundary. Let's compute some examples.

Example 2.3. $\mathrm{Map}(D^2) = 1$. *This is called the* Alexander trick. *Say you have a homeomorphism $f : D^2 \longrightarrow D^2$ that's the identity on ∂D^2. Define*

$$F(x, t) := \begin{cases} (1-t)f(\frac{x}{1-t}) & 0 \leq |x| < 1-t \\ x & 1-t \leq |x| \leq 1. \end{cases}$$

Then $F(x, 0) = f(x)$, $F(x, 1) = x$, and this is an isotopy.

To identify more complicated mapping class groups, we need to better understand how to manipulate curves on surfaces. We say that two simple closed curves α, β on S are in *minimal position* if they intersect transversely and t$|\alpha \cap \beta|$ is at most $|\alpha' \cap \beta|$ for any α', β' homotopic to α, β. Here, the minimal $|\alpha' \cap \beta'|$ over all curves homotopic to α, β is sometimes just called the *intersection number* $i(\alpha, \beta)$, so two curves are in minimal position if they realize their intersection number.

Lemma 2.4 (The bigon criterion)**.** *Transverse curves $\alpha, \beta \subset S$ are in minimal position if only if they do not bound an embedded bigon, i.e. a topological disk embedded in S whose boundary is the concatenation of an arc of α and an arc of β.*

You can find a proof in Farb-Margalit [14]. The proof given uses hyperbolic geometry to interpret the minimal intersection number of curves homotopic to α, β in terms of the behavior at infinity of the preimages of α, β in the universal cover.

The following theorems were both proved by Baer in the 1920's [1, 2].

Theorem 2.5. *If $\alpha, \beta : S^1 \longrightarrow S$ are two homotopic simple closed curves on a surface S, then there is an ambient isotopy $f_t : S \longrightarrow S$ (rel ∂S) such that $f_0 = id$ and $f_1 \circ \alpha = \beta$.*

Here, 'ambient' refers to the fact that we're isotoping the whole surface, not just the simple closed curve.

Proof. Given two homotopic curves α, β, isotope α so that it is transverse to β. Since α, β are homotopic, they can be made disjoint after a homotopy, so if α, β intersect then there is a bigon. Isotope α through the bigon, decreasing its intersection number with β. Continue, until α is disjoint from β.

Suppose first that α, β are homotopically trivial in S. Then α, β bound discs on S. You can then isotope α to β by first isotoping α through the disc it bounds to a small metric disc (with respect to some Riemannian), then push this disc along a path in S to a similar small disc inside the disc bounded by β, and then expand it to give the disc bounded by β.

Now suppose that α, β are homotopically essential. Picking a basepoint and based homotopic based curves homotopic to α, β, let $\pi : X \longrightarrow S$ be the cover corresponding to the associated infinite cyclic subgroup of $\pi_1 S$, and let $\tilde{\alpha}, \tilde{\beta}$ be the pre-images of α, β under π. Then X is homotopic to an open annulus and the components of $\tilde{\alpha}, \tilde{\beta}$ are either essential simple closed curves on X or properly embedded arcs. Moreover, there is at least one no component of each that is an essential simple closed curve. It follows that there is a component $X \setminus (\tilde{\alpha} \cup \tilde{\beta})$ whose closure is an annulus A bounded by one component of $\tilde{\alpha}$ and one component of $\tilde{\beta}$. The restriction $\pi|_A$ is injective on ∂A, and hence is injective on A as well: indeed, if we let $B \subset A$ be the subset of points $b \in A$ such that $|(\pi|_A)^{-1} \circ \pi(b)| \geq 2$, then B is closed, and is also open since if $b \in B$ and $b' \in B$ satisfies $\pi(b) = \pi(b')$, then neither b nor b' lie on ∂A, so we can pick a small neighborhood of their common image that is evenly covered by neighborhoods of b, b' in A. So, A embeds as an annulus in S bounded by α, β.

We can now isotope α to β through the embedded annulus □

There are also versions of the above results for arcs rather than curves. Namely, if S has non-empty boundary, an *arc* in S is an embedding

$$\gamma : I \longrightarrow S, \quad \gamma^{-1}(\partial S) = \partial I$$

and we say that two arcs are *homotopic* if there is a homotopy from one to the other that fixes ∂I. Two arcs α, β are in *minimal position* if $|\alpha \cap \beta|$ is minimal among all homotopic arcs, and the *bigon criterion* says that two transverse arcs are in minimal position if and only if they do not bound a bigon. You can then prove that any two homotopic arcs on S are isotopic. There's also a bigon criterion for non-simple objects: if a (non-simple) closed curve that self-intersects transversely doesn't realize the minimal self-intersection number in its homotopy class, it must bound a bigon or a monogon. See for instance §2 of [55], although the proof of the bigon criterion in Farb-Margalit [14] modifies to cover this case. Similarly, geodesic arcs and non-simple closed curves always realize the minimal (self-) intersection number in their homotopy classes.

Theorem 2.6. *If $f, g : S \longrightarrow S$ are orientation preserving homeomorphisms that are homotopic rel ∂S, then f, g are isotopic rel ∂.*

Note a reflection of $int(D^2)$, or a reflection of $int(S^1 \times [0,1])$ in the second coordinate, is homotopic to the identity, but not isotopic to the identity. The proof involves an inductive application of the previous theorem. Very roughly, given f homotopic to the identity, pick an essential simple closed curve α on S and first use the previous theorem to isotope f so that $f|_\alpha = id$. Continue, cutting along simple closed curves and arcs, until we cut S up into discs, and then we use the Alexander trick to isotope f to the identity on the discs.

It is also worth noting that the Theorem 2.6 fails in higher dimensions, see e.g. Friedman-Witt [19] for an example in 3 dimensions.

Example 2.7. $\mathrm{Map}(S^2) = 1$*, since given $f \in \mathrm{Homeo}(S^2)$ we can take a simple closed curve $\alpha \subset S^2$, and using Theorem 2.5 we can find an isotopy g_t that takes $f(\alpha)$ to α, so that then $g_t \circ f$ is an isotopy from f to a map fixing α, and then we can apply the Alexander trick on each complementary disk.*

Example 2.8. *If $A = S^1 \times [0,1]$, then $\mathrm{Map}(A) \cong \mathbb{Z}$, and is generated by the isotopy class of a* Dehn twist*, defined by:*

$$T : A \longrightarrow A, \quad T(x,t) = (e^{2\pi t i}x, t). \tag{1}$$

In other words, we want to show that $n \mapsto [T^n] \in \mathrm{Map}(A)$ is an isomorphism. Fix a point $p \in S^1$ and let γ be the arc $p \times [0,1]$. Then $T^n(\gamma)$ wraps around the annulus n times, so is not homotopic to γ rel ∂ unless $n = 0$, implying the map is injective. For surjectivity, take a homeomorphism $f : A \longrightarrow A$ that's the identity on the boundary. Then $f(\gamma)$ is homotopic to $T^n(\gamma)$ for some n. It suffices to show $g = T^{-n} \circ f$ is isotopic to the identity. By Theorem 2.5 we can isotope g so that it fixes γ pointwise, and then use the Alexander trick to isotope f to the identity on the disc you get by cutting A along γ.

In general, if S is an oriented surface and $\alpha \subset S$ is a simple closed curve, we can let $A \subset S$ be an annular neighborhood of S (which is a one-sided neighborhood if α is a boundary component of S), parametrized so that the product orientation on $S^1 \times [0,1]$ agrees with the orientation on S. The (positive) *Dehn twist* $T_\alpha : S \longrightarrow S$ is the identity on $S \setminus int(A)$, and is defined on A just as in (1). The isotopy class of T_α is well defined by the (unoriented) curve α and the orientation of S. To distinguish T_α from its inverse, note that if γ is an arc that cross the annulus A, then if we walk along $T_\alpha(\gamma)$ toward α, we twist around α 'to the right'.

Proposition 2.9. *If α is a homotopically essential simple closed curve in $int(S)$ that doesn't bound a puncture, then the Dehn twist T_α has infinite order in $\mathrm{Map}(S)$.*

The proof is essentially the same as for the annulus, except that instead of a transverse arc you may have to use a transverse simple closed curve. See Prop 3.1 of Farb-Margalit [14] for details.

3. Relationships between mapping class groups

Let $\bar{S}$ be a compact surface with r boundary components, and let S be its interior. We have the following relationship between the two mapping class groups:

Theorem 3.1. *If $\bar{S}$ is not an annulus or disc, there is a short exact sequence*

$$1 \longrightarrow \mathbb{Z}^r \longrightarrow \mathrm{Map}(\bar{S}) \longrightarrow \mathrm{Map}(S) \longrightarrow 1,$$

where $\mathbb{Z}^r$ is generated by Dehn twists around the boundary components.

The main point of the proof is to show that if $f : \bar{S} \longrightarrow \bar{S}$ is the identity on $\partial\bar{S}$ and is isotopic to the identity (not rel ∂), then you can compose it with boundary twists so that it's isotopic to the identity rel ∂. This uses an argument similar to surjectivity in Example 2.8. See Theorem 3.19 in Farb Margalit [14]. When $\bar{S}$ is a disk, the Dehn twist around the boundary is trivial in $\mathrm{Map}(\bar{S})$, and when $\bar{S}$ is an annulus, the Dehn twists around its boundary components represent the same element in $\mathrm{Map}(S)$, and we have $\mathrm{Map}(S) \cong \mathbb{Z}$ as in Example 2.8.

When we work with surfaces with punctures, it's often useful to regard the punctures as 'marked points' on the surface, instead of absent points. If S is a compact surface and $P = \{p_1, \dots, p_n\}$ is a set of marked points on S, we can define

$$\mathrm{Homeo}(S,P) := \{ \text{ homeos } f : S \longrightarrow S \mid f(P) = P, f(S \setminus P) = f(S \setminus P) \}$$

and then define the mapping class group with marked points as

$$\mathrm{Map}(S,P) = \mathrm{Homeo}^+(S,P)/\mathrm{Homeo}_0(S,P).$$

There is then a canonical isomorphism

$$\mathrm{Map}(S,P) \longrightarrow \mathrm{Map}(S \setminus P)$$

defined by restricting a homeomorphism of (S,P) to $S \setminus P$. We also define the *pure homeomorphism group* of (S,P) to be the group

$$\mathrm{PHomeo}(S,P) := \{f \in \mathrm{Homeo}(S,P) \mid f(p_i) = p_i \ \forall p \in P\},$$

and the *pure mapping class group* of (S,P) to be the quotient

$$\mathrm{PMap}(S,P) := \mathrm{PHomeo}^+(S,P)/\mathrm{PHomeo}_0(S,P).$$

Note that there is a short exact sequence

$$1 \longrightarrow \mathrm{PMap}(S,P) \longrightarrow \mathrm{Map}(S,P) \longrightarrow S_P \longrightarrow 1,$$

where S_P is the symmetric group on P, i.e. the set of bijections from P to P.

Remark 3.2. *The exact sequence above splits exactly when $|P| \leq 3$. Indeed, if $n := |P| \leq 3$ then one can draw a picture of (S,P) embedded in $\mathbb{R}^3$ that is symmetric under the action of S_n. But if $n \geq 4$, one can use the fact that any finite subgroup F of $\mathrm{Map}(S)$ that preserves a boundary component has to be cyclic. (To prove this, use the Nielsen Realization theorem, which we'll cover later, to realize F as a group of isometries with respect to some hyperbolic metric with geodesic boundary, then map $F \longrightarrow S^1$ by recording the amount of rotation on the boundary component. This is an injective homomorphism since any isometry that's the identity on a boundary component is the identity.) Since the stabilizer of a point in S_n isn't cyclic if $n \geq 4$, the SES above doesn't split if $n \geq 4$.*

Example 3.3. *Let S be a sphere with three marked points $P = \{p_1, p_2, p_3\}$. Then $\mathrm{PMap}(S,P)$ is trivial. To see this, let's define a* simple arc *on (S,P) to be an embedding $\gamma : I \longrightarrow S$ with $\gamma(\partial I) \subset P$ and $\gamma(intI) \subset S \setminus P$. If $i \neq j$, let k be the remaining index in $\{1,2,3\}$ and let α_k be simple arc on S that has endpoints at p_i and p_j. We can choose $\alpha_1, \alpha_2, \alpha_3$ to be disjoint. You can check that up to homotopy rel P, each α_k is the unique simple arc from p_i to p_j. So, given $f \in \mathrm{PHomeo}(S,P)$, we can isotope f rel P to a homeomorphism fixing all three arcs, and then isotope it to the identity using the Alexander trick on each complementary triangle. Hence,*

$$\mathrm{Map}(S,P) \cong \mathrm{Map}(S \setminus P) \cong S_3,$$

the symmetric group on three letters.

Say now that S is a surface, possibly with boundary, and we fix a point $p \in S$. How does $\mathrm{Map}(S)$ compare to $\mathrm{Map}(S,p)$?

Theorem 3.4 (The Birman exact sequence, [5])**.** *There is an exact sequence*

$$\pi_1(S,p) \xrightarrow{\mathcal{P}} \mathrm{Map}(S,p) \xrightarrow{\mathcal{F}} \mathrm{Map}(S) \longrightarrow 1,$$

and if S is not a torus or open annulus, the map $\mathcal{P}$ is injective.

This beautiful theorem was proved by Joan Birman in her 1969 thesis! When S is a torus or open annulus it turns out that the map $\mathcal{P}$ is trivial, which we will discuss below. It is not necessary to assume that S has finite type in the statement above. Also, note that the theorem is trivial if $\chi(S) > 0$, i.e. if $S = S^2$ or D^2, since then all the groups and maps are trivial. So in the following, we will basically always assume that $\chi(S) < 0$.

Let's first describe the maps $\mathcal{P}, \mathcal{F}$ above. First is the *forgetful map*

$$\mathcal{F} : \mathrm{Map}(S,p) \longrightarrow \mathrm{Map}(S)$$

that takes in a homeomorphism fixing p and forgets that it fixes p. In other words, this is the map induced by the inclusion $\mathrm{Homeo}(S,P) \hookrightarrow \mathrm{Homeo}(S)$. The map

$$\mathcal{P} : \pi_1(S,p) \longrightarrow \mathrm{Map}(S,p)$$

is called the *point pushing map*, and is a little more complicated. Given $[\gamma] \in \pi_1(S,p)$, regarded as a map $\gamma : [0,1] \longrightarrow S$, take any isotopy $f_t : S \longrightarrow S$ such that $f_0 = id$ and $f_t(p) = \gamma(t)$, and then define the point pushing map via

$$\mathcal{P}([\gamma]) = [f_1] \in \mathrm{Map}(S,p).$$

One way to visualize such an isotopy is to image the surface is made of sand and to take your finger and push the point p around the loop γ. You can also write an example of a possible f_1 as a composition of explicit maps defined in neighborhood of embedded segments of γ.

To see that $\mathcal{P}$ is well-defined, we first prove the following lemma.

Lemma 3.5. *Say $[\gamma] \in \pi_1(S,p)$ and (f_t) is an isotopy as above. Then*

$$(f_1)_* : \pi_1(S,p) \longrightarrow \pi_1(S,p)$$

acts via $(f_1)_*([\alpha]) = [\gamma^{-1}][\alpha][\gamma]$, i.e. via conjugation by γ^{-1}.

Proof. Given $[\alpha] \in \pi_1(S,p)$, for each $s \in [0,1]$ let $\gamma_s = \gamma|_{[0,s]}$ and let

$$\alpha_s = \gamma_s \cdot (f_s \circ \alpha) \cdot \gamma_s^{-1}$$

and notice that (α_s) is a homotopy from $\alpha_0 = \alpha$ to $\alpha_1 = \gamma \cdot f_1 \circ \alpha \cdot \gamma^{-1}$. □

Then $\mathcal{P}$ is well-defined because any two maps on (S,p) that induce the same map on $\pi_1(S,p)$ are homotopic rel p, and therefore isotopic rel p. (At least when $S \neq S^2$, you get that two such maps are homotopic using that S is a $K(\pi,1)$.)

Example 3.6 (Point pushing along a simple loop)**.** *Let $\gamma : S^1 \longrightarrow S$ be an essential (oriented) simple loop on S that passes through p. Orientation preservingly parametrize an annular neighborhood $A \supset \gamma$ as $S^1 \times [-1,1]$, where $\gamma(t) = (t,0)$. Define an isotopy $f_s : S \longrightarrow S$ supported in A via the formula*

$$f_s(z,r) = (e^{2\pi \cdot s \cdot (1-|r|)i} z, r).$$

Then (f_s) rotates p around the circle γ, so $\mathcal{P}([\gamma]) = [f_1]$. But we also see that f_1 is the composition of a twist around $S^1 \times -1$ and the inverse of a twist around $S^1 \times 1$. So in other words, to point push around a simple closed curve γ, we can look in the direction of the orientation of γ and push off γ to the left creating a curve α, and push off to the right making a curve β, and then

$$\mathcal{P}([\gamma]) = T_\alpha \circ T_\beta^{-1}.$$

Proof of the Birman Exact Sequence. The map $\mathcal{P}$ is a homomorphism since if γ, δ are loops based at p and $(f_t), (g_t)$ are corresponding isotopies as above, then we can make an isotopy for the concatenation $\gamma \cdot \delta$ by letting $h_t = f_{2t}$ when $t \in [0,1/2]$, and $h_t = g_{2t-1} \circ f_1$ when $t \in [1/2,1]$, and then $h_1 = g_1 \circ f_1$. The image of $\mathcal{P}$ is contained in the kernel of $\mathcal{F}$, since f_1 above is always isotopic to the identity on S. And if $[f] \in \mathrm{Map}(S,p)$ is in the kernel of $\mathcal{F}$, there is an isotopy f_t with $f_0 = id$ and $f_1 = f$, and we can let $\gamma(t) = f_t(p)$, in which case we have $[f] = \mathcal{P}([\gamma])$.

Let's now show that $\mathcal{P}$ is injective if S is not a torus or annulus, i.e. if $\chi(S) \neq 0$. If $\chi(S) < 0$, the group $\pi_1(S,p)$ has trivial center, so the conjugation action of γ is nontrivial, implying that f_1 is not homotopic to the identity rel p by Lemma 3.5. If $\chi(S) > 0$ then $\pi_1(S,p)$ is trivial, so there's nothing to prove. □

3.1. **The BES from a homotopy long exact sequence.** You can also see the Birman exact sequence as part of the long exact sequence on homotopy group associated to a certain fiber bundle. Let's assume $\partial S = \emptyset$ for simplicity. The crucial point is that there is a fiber bundle

$$\mathrm{Homeo}^+(S,p) \longrightarrow \mathrm{Homeo}^+(S) \xrightarrow{\mathcal{E}} S, \tag{2}$$

where $\mathcal{E}$ is the evaluation map $\mathcal{E}(f) = f(p)$. To see that this is a fiber bundle, take an open set $U \subset S$ that is homeomorphic to a disc. Given $u \in U$, there is a choice of $\phi_u \in \mathrm{Homeo}_+(S)$ such that $\phi_u(p) = u$ and where ϕ_u varies continuously with u. (Informally, identify $U \cong D^2$ and define homeomorphism rel ∂ of D^2 taking $0 \in D^2$ to $u \in D^2$ by pushing 0 out radially the desired distance and direction. Or write a messy coordinate expression. Then precompose with a homeomorphism of S taking p to $0 \in D^2 \cong U$.) Then the map

$$U \times \mathrm{Homeo}^+(S,p) \longrightarrow \mathrm{Homeo}^+(S), \quad (u,f) \mapsto \phi_u \circ f$$

is a homeomorphism onto $\mathcal{E}^{-1}(u)$, with inverse map $\psi \mapsto (\psi(p), \phi^{-1}_{\psi(p)} \circ \psi)$, which gives the total space the structure of a local product, as desired. There is a long exact sequence of homotopy groups for any fiber bundle, and the relevant part is

$$\pi_1(\mathrm{Homeo}^+(S), id) \to \pi_1(S,p) \to \pi_0(\mathrm{Homeo}_+(S,p)) \to \pi_0(\mathrm{Homeo}^+(S)) \to \pi_0(S).$$

Here, the map $\pi_1(S,p) \longrightarrow \pi_0(\mathrm{Homeo}_+(S,p)) \cong \mathrm{Map}(S,p)$ is exactly $\mathcal{P}$, since we get it by lifting a loop in the base space to a path in the total space starting at the identity and seeing where it ends up. Note that from this perspective, well-definedness of $\mathcal{P}$ follows from applications of the homotopy lifting property of fiber bundles rather than (same action on π_1) $\implies$ homotopy $\implies$ isotopy. The next map is $\mathcal{F}$, since it's induced by the projection from $\mathrm{Homeo}^+(S,p) \longrightarrow \mathrm{Homeo}^+(S)$. Since $\pi_0(S)$ is trivial, this gives us the desired exact sequence.

When $\chi(S) < 0$, injectivity of $\mathcal{P}$ also follows from the fact that $\pi_1 \mathrm{Homeo}^+(S) = 1$, which is a consequence of the following more general theorem[1].

Theorem 3.7 (mostly Hamstrom [22] and Yagasaki [65])**.** *If S is not a sphere, a torus or an open disk or annulus, then* $\mathrm{Homeo}_0(S)$ *is contractible.*

For finite type S, Mary-Elizabeth Hamstrom [22] showed in 1966 that aside from the exceptions above, $\mathrm{Homeo}_0(S)$ has trivial homotopy groups. Tatsuhiko Yagasaki [65] then showed in 2000 how to promote her result to apply to infinite surfaces. It turns out that $\mathrm{Homeo}_0(S)$ is always an ANR (absolute neighborhood retract), c.f. Yagasaki [64], and this implies that they have the homotopy type of CW complexes, by Milnor [39], so Whitehead's Theorem implies that trivial homotopy groups implies contractibility. It seems to be unclear whether $\mathrm{Homeo}(M)$ is an ANR in higher dimensions

So, how do we get intuition for the theorem above? As a simple example, $\mathrm{Homeo}(D^2, \partial D^2)$ is contractible by the Alexander trick, which gives an explicit contraction to the identity. In general, Hamstrom's argument is quite complicated, but the general flavor of the proof is an inductive argument on the complexity of the surface that is similar to the proof of Theorem 2.6. In that proof, we fix a curve $c \subset S$ and a homeomorphism f and used the bigon criterion to first isotope $f(c)$ to a curve disjoint from c, then isotoped it through an annulus to c. Extending this

[1]Note that $\mathrm{Homeo}^+(S)$ is usually disconnected, so when we talk about its fundamental group, we're really talking about the identity component $\mathrm{Homeo}_0(S)$.

isotopy to the whole surface, we can assume after isotope that $f(c) = c$, and then reduce the problem to the case of the surface obtained by cutting along c, eventually reducing all the way down to the Alexander trick. Hamstrom's argument roughly involves[2] doing the same thing to families of homeomorphisms that you get as images of maps from spheres into $\mathrm{Homeo}_0(M)$, but it requires a lot of care to make the isotopies vary continuously as you vary the homeomorphism.

What about the exceptional surfaces? For a torus and open annulus, you can pick a point $p \in S$, identify $\mathrm{Homeo}_0(S,p) \cong \mathrm{Homeo}_0(S \setminus p)$, which is contractible by Theorem 3.7, and use the fibration from (2) to get that the evaluation map $\mathcal{E} : \mathrm{Homeo}_0(S) \longrightarrow S$ induces an isomorphism on π_k for all k, and hence is a homotopy equivalence, since as noted above, $\mathrm{Homeo}_0(S)$ is an ANR. So for instance, if $A = S^1 \times [0,1]$ then $\pi_1(\mathrm{Homeo}_0(A))$ is generated by the loop of rotations

$$R_s : A \longrightarrow A, \quad R_s(z,t) = (e^{2\pi s i} z, t).$$

For $\mathbb{R}^2$ and the sphere S^2, the inclusions

$$SO(2) \hookrightarrow \mathrm{Homeo}_0(\mathbb{R}^2), \quad SO(3) \hookrightarrow \mathrm{Homeo}_0(S^2)$$

are homotopy equivalences by a theorem of Kneser [31], see also Friberg [18] for an elementary proof. I believe it's unknown whether similar theorems hold in higher dimensions. You can also ask whether analogues of Theorem 3.7 are true in the differentiable and PL categories. For PL homeomorphisms, this is a theorem of Peter Scott [51]. For diffeomorphisms, it's a theorem of Earle and Eels [11], see also Earle-McMullen [12] for a nice proof using quasiconformal mappings and the Douady-Earle extension theorem.

Also, the assertion that $\mathrm{Diff}(S^3) \sim O(4)$ is the *Smale conjecture*, proven in 1983 by Hatcher [24]. (In 3-dimensions, it's equivalent to consider $\mathrm{Homeo}(S^3)$, see the appendix to Hatcher's paper for references.) It turns out that the analogous statement $\mathrm{Diff}(S^n) \sim O(n+1)$ is false in higher dimensions, where in dimension 4 this is a recent (currently unpublished, as of Jan 2025) theorem of Watanabe [61]. See also Hatcher's survey [23] for more information on these topics.

3.2. **Braid groups.** Let S be a surface (possibly with ∂), and let

$$C^{ord}(S,n) := \{(x_1, \dots, x_n) \in S^n \mid x_i \neq x_j \ \forall i \neq j\}$$

be the *configuration space* of ordered n-tuples in *int*(S), and let

$$C(S,n) := C^{ord}(S,n)/S_n$$

be the quotient by S_n, so that $C(S,n)$ is identified with the set of n-element subsets of S. Note that S_n acts freely on $C^{ord}(S,n)$, so the projection

$$\pi : C^{ord}(S,n) \longrightarrow C(S,n)$$

is a covering map. The *surface braid group* associated to the pair (S,n) is

$$B(S,n) := \pi_1 C(S,n).$$

If $p = (p_1, \dots, p_n) \in C^{ord}(S,n)$, every loop in $C(S,n)$ based at $\pi(p)$ lifts to an arc

$$\gamma = (\gamma_1, \dots, \gamma_n) : [0,1] \longrightarrow C^{ord}(S,n)$$

[2]At least, this is my read of what's going on after scanning her paper.

such that $\gamma(0) = p$ and $\pi(\gamma(1)) = p$. We can visualize this path as the union of the n disjoint *strands*

$$\sigma_i := \{(\gamma_i(t), t) \mid t \in [0,1]\} \subset S \times [0,1],$$

where here each σ_i is a path from $(p_i, 0)$ to $(p_j, 1)$ for some j. We call the union of these strands a *braid*, since the strands can wind around each other, and $B(S,n)$ is then the set of all braids up to (level-preserving) isotopy[3]. The special case $B(n) := B(D^2, n)$ is just called the *braid group*; try to draw some examples of braids in this case. See also Chapter 9 of [15] for more details.

When $n = 1$, we have $B(S,n) = \pi_1 S$. There's a generalization to higher n of the Birman exact sequence that relates surface braid groups and the associated mapping class groups, which we will state as follows.

Theorem 3.8 (Birman exact sequence for surface braids)**.** *Given $n \geq 1$ and a subset $\{p_1, \ldots, p_n\} \in int(S)$ we have a short exact sequence*

$$B(S,n) \xrightarrow{\mathcal{P}} \mathrm{Map}(S, \{p_1, \ldots, p_n\}) \xrightarrow{\mathcal{F}} \mathrm{Map}(S) \longrightarrow 1,$$

and if $S \neq S^2, T^2, \mathbb{R}^2, S^1 \times \mathbb{R}$, then $\mathcal{P}$ is injective.

For the proof, you can essentially just repeat the argument in §3.1. Namely,

$$\mathrm{Homeo}^+(S) \longrightarrow C(S,n), \quad f \longmapsto \{f(p_1), \ldots, f(p_n)\}$$

is a fiber bundle with fiber $\mathrm{Map}(S, \{p_1, \ldots, p_n\})$, and the associated long exact sequence on homotopy groups gives the short exact sequence above. Just as before, $\mathcal{P}$ is the 'point pushing map' that is the end result of an isotopy that pushes points along the given loop in $C(S,n)$. To prove injectivity of $\mathcal{P}$, one can either use a direct, elementary argument as in Birman's original paper [5], exploiting as before the fact that $\pi_1 S$ is center-free, or one can get injectivity straight from the long exact sequence using that $\pi_1 \mathrm{Homeo}_0(S, \partial S) = 1$, see Theorem 3.7.

In particular, since $\mathrm{Map}(D^2) = 1$, we get:

Corollary 3.9. *For each n-point subset $\{p_1, \ldots, p_n\} \subset int(D^2)$, the map*

$$\mathcal{P} : B(n) \longrightarrow \mathrm{Map}(S, \{p_1, \ldots, p_n\})$$

is an isomorphism.

4. The torus

Let $T^2 = \mathbb{R}^2/\mathbb{Z}^2$. What are some examples of elements of $\mathrm{Map}(T^2)$? Well, say $A \in SL_2\mathbb{Z}$. Then there's an orientation preserving linear homeomorphism

$$A : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad A(v) = Av,$$

and if $w \in \mathbb{Z}^2$ we have $A(v+w) - A(v) = A(w) \in \mathbb{Z}^2$, so A descends to an orientation preserving homeomorphism $f_A : T^2 \longrightarrow T^2$, giving a map

$$SL_2\mathbb{Z} \longrightarrow \mathrm{Map}(T^2), \quad A \mapsto f_A \tag{3}$$

Theorem 4.1. *The map* (3) *is an isomorphism.*

[3] Here, we mean an isotopy $f_t : S \times [0,1] \longrightarrow S \times [0,1]$ that is constant on the second factor. In fact, one can show that you get the same group whether or not you stipulate that the isotopies are level preserving.

To prove this, we need to understand curves on T^2 better. An element $(p,q) \in \mathbb{Z}^2$ is called *primitive* if p,q are coprime. Equivalently, (p,q) is primitive if it's not a proper multiple of any element of $\mathbb{Z}^2$, or if it's part of a basis for $\mathbb{Z}^2$.

Lemma 4.2. *An element $(p,q) \in \mathbb{Z}^2$ can be represented by a* simple *closed curve if and only if (p,q) is primitive.*

Proof. The Euclidean geodesic from 0 to (p,q) projects to a closed curve in the given homotopy class, which is the line of slope p/q on the torus. Since the slope is constant, it's simple if and only if it's not a proper power of another loop, which is the case if and only if (p,q) is primitive.

Moreover, if $n \in \mathbb{N}$ then (np,nq) can be represented by a self-transverse, self-intersecting loop on T^2 with no bigons (take a curve that goes around the (p,q)-curve n times and perturb to be self-transverse), so the bigon criterion says that it's impossible to homotope this curve to be simple. $\square$

Lemma 4.3. *If α, β are simple closed, oriented curves on T^2 that are in minimal position and are in the homology classes $v,w \in \mathbb{Z}^2$ (thought of as column vectors) respectively, then $|\alpha \cap \beta|$ is the absolute value of the algebraic intersection number, which is the determinant*

$$D = \det \begin{pmatrix} v & w \end{pmatrix}.$$

Proof Sketch. It's true if $v = (1,0)$, since a (r,s)-curve intersects a $(1,0)$-curve s times, counting orientation. (Note it's sufficient for this to draw any two representatives in minimal position, since both algebraic intersection number and $i(\alpha,\beta)$ are invariant under homotopy for minimal position curves.) Given an arbitrary pair v,w, there's an element $A \in SL_2\mathbb{Z}$ that takes v to $(1,0)$. But as f_A preserves orientation, it preserves algebraic intersection numbers, and also

$$\det \begin{pmatrix} v & w \end{pmatrix} = \det A \cdot \det \begin{pmatrix} v & w \end{pmatrix} = \det \begin{pmatrix} Av & Aw \end{pmatrix},$$

so the result follows. $\square$

Any orientation preserving homeomorphim $f : T^2 \longrightarrow T^2$ acts preserving algebraic intersection number, so it follows that

$$f_* : H_1(T^2, \mathbb{Z}) \longrightarrow H_1(T^2, \mathbb{Z})$$

satisfies $\det(v\ w) = \det(f_*(v)\ f_*(w))$, implying $\det f_* = 1$, so the action on homology is via an element of $SL_2\mathbb{Z}$. This gives a map

$$\mathrm{Map}(T^2) \longrightarrow \mathrm{SL}_2\mathbb{Z}, \quad f \mapsto f_*,$$

which we claim is the inverse to the map (3).

First, it's clear that $(f_A)_* = A$, since $A : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ takes a path from 0 to v to a path from 0 to Av. Next, we want to show that if $f : T^2 \longrightarrow T^2$ is orientation preserving and $f_* = A$, the f is isotopic to A. But since T^2 is a $K(\pi,1)$, any two maps that induce the same action on π_1 are homotopic (here you can actually just use the straight line homotopy between lifts of the two homeos to $\mathbb{R}^2$), and hence are isotopic by Theorem 2.6.

Elements of $SL_2\mathbb{R}$ can be classified according to their trace.

Definition 4.4. We say that $A \in SL_2\mathbb{R}$ is *elliptic* if $|\mathrm{tr}A| < 2$, is *parabolic* if $|\mathrm{tr}A| = 2$ and $A \neq \pm I$, and is *hyperbolic* if $|\mathrm{tr}A| > 2$.

Any $A \in SL_2\mathbb{R}$ has two complex conjugate eigenvalues $\lambda, \frac{1}{\lambda}$, with $\frac{1}{\lambda} = \bar{\lambda}$. So, either the eigenvalues are both real, or $|\lambda| = 1$. Since $\lambda + \frac{1}{\lambda} = \mathrm{tr}A$, if both are real, then $|\mathrm{tr}A| \geq 2$, while otherwise, we have $|\lambda| = 1$ and $\lambda \neq \pm 1$, in which case $|\mathrm{tr}A| < 2$. So, for $A \in SL_2\mathbb{R}$ it follows that

(1) A is elliptic if it only if it has two non-real conjugate eigenvalues $e^{i\theta}, e^{-i\theta}$,
(2) A is parabolic if and only if it has a single eigenvalue, which is ± 1,
(3) A is hyperbolic if only if it has two real eigenvalues $\lambda, \frac{1}{\lambda}$, say with $|\lambda| > 1$,

Moreover, one can show that in the three cases, A is conjugate in $SL_2\mathbb{R}$ to

(1) the rotation matrix $R_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \in SO(2)$,
(2) the matrices $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$,
(3) the matrices $\pm \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$, where $\lambda > 0$,

So, what do elements of these types look like in $SL_2\mathbb{Z}$?

Lemma 4.5. *An element $A \in SL_2\mathbb{Z}$ is elliptic if and only if it has finite order.*

Proof. In the list of conjugacy representatives above, only the rotation matrices can ever be finite order, so finite order implies elliptic.

Assume A is elliptic. Then A is conjugate to a rotation matrix $R_\theta \in SL_2\mathbb{Z}$, but $SO(2) \cap SL_2\mathbb{Z}$ is finite since the former is compact and the latter discrete, so R_θ has finite order and hence A does too. $\square$

Examples of elliptic elements in $SL_2\mathbb{Z}$ include the matrices

$$R_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix},$$

which have order 4 and 6, respectively. Here, $R_{\pi/2}$ acts as a $\pi/2$ rotation on $\mathbb{R}^2$, preserving $\mathbb{Z}^2$. There is some $A \in SL_2\mathbb{R}$ such that $ABA^{-1} = R_{\pi/3}$, a rotation by $\pi/3$. Conjugating the translation group $\mathbb{Z}^2 \subset \mathbb{R}^2$ by the linear action of A gives the symmetry group of the regular hexagonal lattice, say centered at the origin, so one way to understand B is to view it as the obvious order 6 symmetry of this lattice. In fact, every elliptic in $SL_2\mathbb{Z}$ is conjugate in $SL_2\mathbb{Z}$ to a power of one of the two examples above.

Examples of parabolic elements of $SL_2\mathbb{Z}$ are the matrices in (2) above and their powers. The most often cited example of a hyperbolic matrix in $SL_2\mathbb{Z}$ is

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

which has eigenvalues ϕ^2 and ϕ^{-2}, where ϕ is the golden ratio. Some eigenvectors are $(\phi, 1)$ and $(-1/\phi, 1)$, respectively. (Use that $1 + \phi = \phi^2$.) You can conjugate all these examples by elements of $SL_2\mathbb{Z}$ to get loads of other examples.

So, how does the trace classification manifest when we view elements of $SL_2\mathbb{Z}$ as elements of the mapping class group $\mathrm{Map}(T^2)$? Elliptic elements are just the finite order elements of $\mathrm{Map}(T^2)$, and they can all be realized as finite order linear homeomorphisms. As a particular example, if $A = -I$ then $f_A \in \mathrm{Map}(T^2)$ is called the *hyperelliptic involution.* It reverses the orientation of every simple closed curve, and with respect to the standard pic of the torus in $\mathbb{R}^3$, it can be visualized by skewering the torus (through 4 points) and rotating 180 degrees.

Lemma 4.6. *Assume $A \in SL_2\mathbb{Z}$ is parabolic. Then A has a real ± 1-eigenspace, generated by a primitive vector $(p,q) \in \mathbb{Z}^2$, and $f_A : T^2 \longrightarrow T^2$ is (isotopic to) a power of a Dehn twist around the (p,q)-curve on T^2, composed with the hyperelliptic involution if* $\mathrm{tr}A = -2$.

Proof. The matrix $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ acts as a Dehn twist around the $(1,0)$-curve (composed with the hyperelliptic involution if it's a negative) and $(1,0)$ generates the relevant eigenspace. Taking powers and conjugating, we get the lemma for an arbitrary parabolic A. □

So, what if $A \in SL_2\mathbb{Z}$ is hyperbolic? Then it has two 1-dimensional real eigenspaces, E_+ and E_-, corresponding to the eigenvalues λ with $|\lambda| > 1$, and $1/\lambda$, respectively. Note that E_+, E_- are both lines of irrational slope: if E_+ had rational slope, say, then it would pass through an element $v \in \mathbb{Z}^2$, but then $A^{-n}(v) \to 0$ even though it lies in $\mathbb{Z}^2$ for all n, a contradiction. Let $\mathcal{F}_+, \mathcal{F}_-$ be the 1-dimensional foliations of $\mathbb{R}^2$ consisting of lines parallel to E_+, E_-, respectively. The linear action of A on $\mathbb{R}^2$ preserves these two foliations, stretching the leaves of $\mathcal{F}_+$ by a factor of λ and contracting the leaves of $\mathcal{F}_-$ by $1/\lambda$. Rigorously, if $p \in \mathbb{R}^2$ and v_+, v_- are tangent vectors at p that are tangent to $\mathcal{F}_+, \mathcal{F}_-$, then

$$|dA(v_+)| = |\lambda||v_+|, \quad |dA(v_-)| = \frac{1}{|\lambda|}|v_-|, \tag{4}$$

where $|\cdot|$ is the Euclidean norm. The foliations $\mathcal{F}_+, \mathcal{F}_-$ are called the *stable and unstable* foliations of A and/or f_A: for instance, $\mathcal{F}_+$ is 'stable' since any line in $\mathbb{R}^2$ that's not a leaf of $\mathcal{F}_+$ converges to a leaf of $\mathcal{F}_+$ under iteration by A.

The foliations $\mathcal{F}_+, \mathcal{F}_-$ descend to 'stable and unstable' foliations of T^2 by lines of constant slope, which we also denote by $\mathcal{F}_+, \mathcal{F}_-$, abusively. Note that in the quotient, the leaves of both foliations are still topologically lines rather than circles, because their slopes are irrational. The map $f_A : T^2 \longrightarrow T^2$ preserves both foliations, stretching $\mathcal{F}_+$ and contracting $\mathcal{F}_-$ as in (4), with respect to the Euclidean metric on T^2. In general, diffeomorphisms that preserve two transverse foliations, stretching one and contracting the other, are called *Anosov diffeomorphisms.*

5. Pseudo-Anosov maps

In this section we'll develop a theory of *pseudo-Anosov homeomorphisms* of surfaces S, generalizing Anosov homeomorphism of the torus. We'll define these maps using two perspectives. First, we generalize the Euclidean structure on T^2 to certain singular Euclidian metrics on S, and define pseudo-Anosov maps as maps that are represented by hyperbolic matrices in isometric local coordinates. Second, we define pseudo-Anosov maps to be homeomorphisms that preserve two transverse singular foliations on S, stretching one and contracting the other.

5.1. **Translation and half-translation surfaces.** The following is a very brief discussion, and we refer the reader to [34], for instance, for more details. The basic idea that we would like to generalize Anosov maps on the torus to higher complexity surfaces S using a certain Euclidean structure on S. Since Gauss-Bonnet says that any surface S with a locally Euclidean metric has $\chi(S) = 0$, one uses instead structures that are locally Euclidean except for a finite set a singularities.

Say that $P \subset \mathbb{R}^2$ is a region in the plane bounded by finitely many edges, which are Euclidean line segments, rays or lines, and that we can glue the edges of P via Euclidean translations that take the side of the edge facing P to the side of the image edge not facing P. (For example, P could be a regular $4g$-gon, with the usual gluing of opposite sides that gives a genus g surface.) The quotient surface S is called a *translation surface*. Any finite type topological surface can be realized in many ways as a translation surface; such a realization is called a *translation structure* on S. Note that if S is allowed to be noncompact, then one must take P to have infinite area, e.g. an annulus is a gluing of an infinite strip.

Since the edges of P are glued by isometries, the Euclidean metric on P descends to a metric on the quotient surface S. This metric is locally Euclidean except at a finite set $X \subset S$ of *singular points*. Here, a singular point obtained when the sum of the interior angles of P at a set of identified vertices does not sum to 2π; the total angle sum is called the *cone angle* at the singular point. Note that if $\chi(S) \neq 0$ then the Gauss-Bonnet Theorem says that S must have singular points.

The cone angle at each singular point x is a multiple of 2π. Indeed, for every directed edge e of P, let's say that the 'angle' of e is the angle it makes with the horizontal, which is well-defined mod 2π. Translations preserve angle, so glued directed edges have the same angle. If $v_1, \ldots, v_n$ are the vertices that project to x, and each v_i is the initial vertex of a directed edge e_i such that e_i is glued to e_{i+1}, interpreted cyclically, then one can show inductively that mod 2π, the sum of the interior angles of P at $v_1, \ldots, v_i$ is the difference between the angles of e_1, e_{i+1}. Setting $i = n$, the total interior angle sum is zero mod 2π.

If $X \subset S$ is the set of singular points, the nonsingular set $S \setminus X$ has an atlas of charts whose transition maps are Euclidean translations. Here, the charts are created by patching together precompositions of the quotient map $P \longrightarrow S$ by (compositions of) the edge-gluing translations.

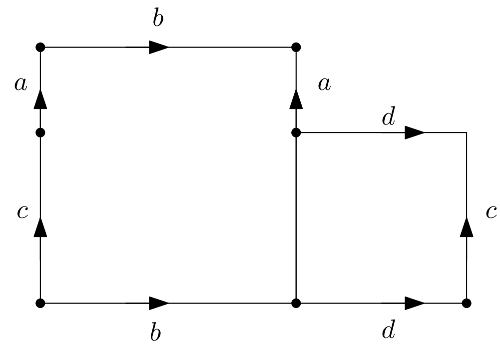
Remark 5.1 (Monodromy, and an intrinsic definition)**.** *Suppose $X \subset S$ is the singular set and $\gamma \subset S \setminus X$ is a loop. The* monodromy *around γ is defined as follows. Realize γ as a concatenation $\gamma = \gamma_1 \cdots \gamma_n$ segments each of which is contained in the codomain of a chart $\phi_i : U_i \longrightarrow V_i$. Pre-composing our charts with translations, we can assume that the common endpoint of γ_i and γ_{i+1} is the image of the same point under ϕ_i and ϕ_{i+1}. The monodromy around γ is then defined to be the transition map $\phi_n^{-1} \circ \phi_1$. For example, if P is an infinite strip and S is the annulus obtained by gluing its side via a translation, then the monodromy around an embedded loop in S is that translation or its inverse, depending on the orientation of the loop. However, if γ is a small embedded loop that bounds a disc in S with a single singular point x, then the monodromy around γ is trivial, i.e. the identity map; indeed, one can build the charts above so that they all take a fixed vertex of P to x, and then the monodromy will fix that vertex.*

Using this language one could define a translation structure on S intrinsically to be a finite subset $X \subset S$ and an atlas on $S \setminus X$ with transition maps that are Euclidean translations, such that the monodromy around every point of X is trivial. Given such a structure, you can cut S up into a topological disc with geodesics, and then lift the disk to $\mathbb{R}^2$ to get a region P that glues up to S as above.

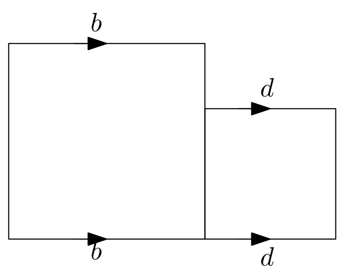
If S is a translation surface, a homeomorphism $f : S \longrightarrow S$ is *affine* if it preserves the singular set and whenever $p \in S$ is nonsingular, there are charts around $p, f(p)$

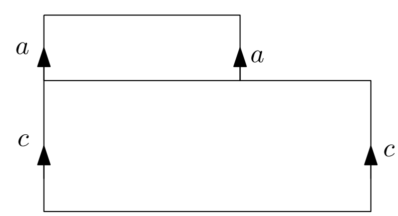
in which f is represented by the action of a matrix $A \in SL_2\mathbb{R}$, referred to as the *derivative* of f. Note that the matrix A is well-defined and is independent of p.

Example 5.2. *Let P be the polygon that's the union of two squares, one with side lengths $\phi = (1+\sqrt{5})/2$, and one with side lengths 1. Glue sides via translations to form a surface S, as indicated below.*



We can divide S into two 'vertical annuli' and into two 'horizontal annuli'.





The product of the Dehn twists in the vertical annuli, and the product of twists in the horizontal annuli, are isotopic to affine maps with derivative

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1+\phi \\ 0 & 1 \end{pmatrix},$$

respectively. The group generated by these matrices is huge — it turns out to be a lattice in $SL_2\mathbb{R}$ — and all the matrices therein are derivatives of affine maps.

Remark 5.3. *If S is a translation surface, its* Veech group *is the group of derivatives of affine maps of S, and S is called a* Veech surface *if its Veech group is a lattice in $SL_2\mathbb{R}$. Veech surfaces S have some fantastic dynamical properties: for instance, the Veech Dichotomy* [58] *says that if S is Veech, then for each θ the angle-θ flow F_θ on S is either periodic, or 'uniquely ergodic', meaning that it mixes S up so well that up to scale the Euclidean area measure on S is the only Borel measure invariant under F_θ.*

More generally, a *half-translation surface* is obtained if we replace translations above by maps of the form $z \mapsto \pm z + w,\ w \in \mathbb{R}^2$. Most of the above discussion is unchanged, but now the edge gluings only preserve angles mod π, and cone angles at singular points can be arbitrary multiples of π. Affine maps on half-translation surfaces are defined as before, but the derivative of an affine map is only well defined up to sign, i.e. as an element of $PSL_2\mathbb{R}$.

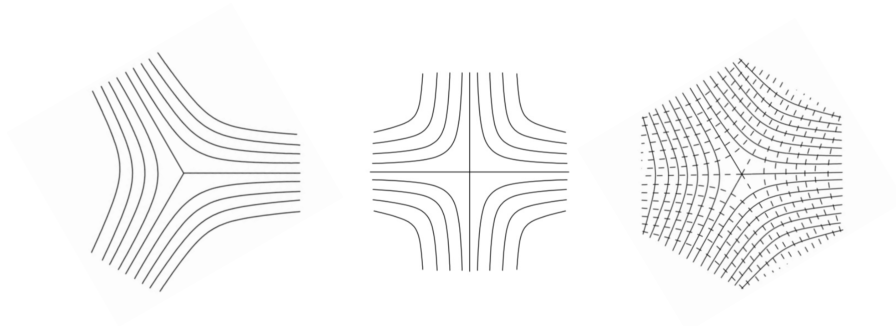


FIGURE 1. k-pronged singular foliations of D^2, for $k = 3, 4$, and then two k-pronged singular foliations that intersect transversely.

Definition 5.4 (pseudo-Anosov, geometric definition)**.** A map $f : S \longrightarrow S$ is called *pseudo-Anosov* if there is a half-translation structure on S such that f is affine with derivative a hyperbolic element of $PSL_2\mathbb{R}$.

For example, we get many pseudo-Anosov maps on a genus 2 surface by looking at affine maps of the L-shaped translation surface above. Of course, the action of a hyperbolic element of $SL_2\mathbb{Z}$ on the torus $\mathbb{R}^2/\mathbb{Z}^2$ is also (pseudo)-Anosov under this definition. More generally, suppose that S is a *square-tiled translation surface*, meaning a gluing via translations of a polygon that's a union of squares with integer vertices in $\mathbb{R}^2$. Then there's a natural quotient map

$$\pi : S \longrightarrow T^2,$$

obtained by reducing mod $\mathbb{Z}^2$, and this is a branched cover, branched over $0 \in T^2$. If $A \in SL_2\mathbb{Z}$, then $f_A : S \longrightarrow S$ lifts to a homeomorphism of S exactly when the action of f_A on $\pi_1(T^2 \setminus 0)$ fixes the subgroup $\pi_1(S \setminus \pi^{-1}(0)) \subset \pi_1(T^2 \setminus 0)$ up to conjugacy. There are finitely many subgroups of $\pi_1(T^2 \setminus 0)$ with a given index, and $SL_2\mathbb{Z}$ permutes these, so the subset $H \leq SL_2\mathbb{Z}$ consisting of all A such that f_A lifts has finite index. In particular, any A has a power A^n such that f_{A^n} lifts, and if A is hyperbolic type, then any lift $\tilde{f}_{A^n} : S \longrightarrow S$ is pseudo-Anosov.

5.2. **The topological perspective.** Here, we describe pseudo-Anosovs as homeomorphisms of S that preserve two transverse 'foliations', stretching one and contracting the other. By Poincaré-Hopf, a surface S admits a 1-dimensional foliation exactly when $\chi(S) = 0$, so if we want a theory that applies more generally, we need to allow singularities of the types drawn in Figure 1, which we call *k-pronged singular foliations of the open disc D^2*. Namely:

Definition 5.5. A *singular foliation* $\mathcal{F}$ of a finite type surface S consists of a finite set X of *singular points* of S, together with a decomposition of S into 1-dimensional subsets called *leaves*, such that

- the decomposition of $\mathcal{F}$ into leaves restricts to that given by an actual foliation on the complement $S \setminus X$ of the singular set,
- every singular point lies in a chart $U \longrightarrow D^2$ taking leaf intersections to the leaves of the k-pronged singular foliation, for some $k \geq 3$,
- every puncture of S is contained in a chart $U \longrightarrow D^2 \setminus 0$ taking leaf intersections to the leaves of the k-pronged singular foliation, for some $k \geq 1$.

Here, if p is a singular point or puncture, then k is called the *order* of p in $\mathcal{F}$.

If $\mathcal{F}$ is a singular foliation on S, a *transverse measure* on $\mathcal{F}$ is a map

$$\mu : \{ \text{ arcs } \alpha : [0,1] \longrightarrow S \setminus (\text{singular points}) \text{ that are transverse to } \mathcal{F}\} \longrightarrow \mathbb{R}_{>0}$$

that is additive under reparametrization and concatenation, and where $\mu(\alpha_s)$ is constant under all homotopies $\alpha_s : [0,1] \longrightarrow S,\ s \in [0,1]$ 'rel $\mathcal{F}$', meaning that for fixed t, the points $\alpha_s(t)$ all lie on the same leaf of $\mathcal{F}$. We call the pair $(\mathcal{F}, \mu)$ a *measured foliation* on S. Two singular foliations $\mathcal{F}_\pm$ on S are *transverse* if their leaves are transverse near all nonsingular points of S.

Equivalently, $\mathcal{F}_\pm$ are transverse if all points on S have the same orders with respect to $\mathcal{F}_\pm$, and every point lies in the domain of a chart $\phi : U \longrightarrow D_k$ taking to two transverse k-pronged singular foliations of D^2, as pictured on the right of Figure 1 for $k = 3$. As an example, if $\theta \neq \theta' \mod \pi$, the angle θ and θ' foliations of a translation surface are transverse.

Definition 5.6 (pseudo-Anosov, topological definiton)**.** A homeomorphism

$$f : S \longrightarrow S$$

is called *pseudo-Anosov* if there are two transverse measured foliations $(\mathcal{F}_\pm, \mu_\pm)$, called the *stable and unstable foliations* of f, and a number $\lambda > 0$, called the *dilatation* of f, such that $f(\mathcal{F}_\pm) = \mathcal{F}_\pm$ and

$$f_*\mu_+ = \lambda\mu_+, \quad f_*\mu_- = 1/\lambda \cdot \mu_-.$$

Here, $f_*\mu$ referes to the pushforward transverse measure, where

$$f_*\mu(\alpha) := \mu(f^{-1}(\alpha)).$$

Let's try to reconcile this notion with the earlier one. Suppose first that S is a half-translation surface and $f : S \longrightarrow S$ is affine with derivative a hyperbolic element $A \in PSL_2\mathbb{R}$. Then the eigenfoliations $\mathcal{F}_\pm$ parallel to the eigenspaces $E_\pm$ of A are preserved by f. For a transverse measure, take a smooth arc α transverse to $\mathcal{F}_+$, say, and define

$$\mu_+(\alpha) = \int_0^1 |\pi_-(\alpha'(t))|\, dt$$

where π_- is the projection onto the second factor in $\mathbb{R}^2 = E_+ \oplus E_-$ and $|\cdot|$ is the Euclidean norm. Now, in local coordinates f^{-1} is represented by A^{-1}, so if $v \in E_-$ then $|df^{-1}(v)| = \lambda|v|$, where $\lambda > 1$ is the eigenvalue of E_+. So we have

$$|\pi_-(f^{-1} \circ \alpha'(t))| = \lambda|\pi_-(\alpha'(t))| \implies f_*\mu = \lambda\mu,$$

and the analysis for μ_- is the same.

Conversely, suppose that f preserves two singular foliations $\mathcal{F}_\pm$ and scales transverse measures via $f_*\mu_\pm = \lambda^{\pm 1}\mu_\pm$. We want to build a half-translation structure on S. To do this, let X be the common singular set of $\mathcal{F}_\pm$, take a point $p \in S \setminus X$, let $U \subset S \setminus X$ be a small ball around p, and define a chart

$$\phi : U \longrightarrow V \subset \mathbb{R}^2$$

by taking $q \in U$, choosing a path α_q in U from p to q that is transverse to both foliations, and defining $\phi(q)$ so that the absolute values of its coordinates are $(\mu_-(\alpha), \mu_+(\alpha))$. To define the signs of the coordinates of $\phi(q)$, you pick local orientations on $\mathcal{F}_\pm$ near p so that the oriented intersection numbers $i(\ell_-, \ell_+)$ are positive when $\ell_\pm$ are leaves of $\mathcal{F}_\pm$, and then define the sign of the first coordinate of

$\phi(q)$ to be positive if α_q intersects the leaves of $\mathcal{F}_-$ positively, and similarly for the second coordinate. This ϕ will be a well-defined homeomorphism onto its image, and it takes $\mathcal{F}_\pm$ to the horizontal and vertical foliations of V. Transition maps between charts will be Euclidean isometries that preserve the horizontal/vertical foliations, so maps of the form $z \mapsto \pm z + w$. In these coordinates, f will be affine, represented by the diagonal matrix with entries $\pm\lambda, \pm 1/\lambda$.

5.3. Properties of pseudo-Anosov maps.

Fact 5.7. *The invariant foliations $\mathcal{F}_\pm$ have no closed leaves, and no 'saddle connections', i.e. leaf segments with two singular endpoints and interior disjoint from the singular set.*

Proof. There are only finitely many saddle connections in a singular foliation on a finite type surface, so if there is a saddle connection c in $\mathcal{F}_+$, say, there is some n such that $f^n(c) = c$. But $\mu_-(f^n(c)) = \lambda^{-n}\mu_-(c) \neq \mu_-(c)$, a contradiction. The argument for closed leaves is somewhat similar. Briefly, the closed leaves in $\mathcal{F}_+$ lie in a union of finitely many foliated annuli, and μ_- is constant on the leaves in each annulus. So, μ_- takes on only finitely many values on closed leaves, but it is supposed to be multiplied by λ when we apply f, so there are no closed leaves. $\square$

Proposition 5.8. *Suppose $f : S \longrightarrow S$ is pseudo-Anosov. If γ is a simple closed curve on S, then all the iterates $f^n(\gamma)$ are homotopically distinct.*

To prove this, we need a few facts about singular Euclidean structures. So, suppose S is a half translation surface. A *geodesic* on S is a path that locally minimizes length, with respect to the singular Euclidean metric on S. A *saddle connection*[4] on S is a geodesic segment that intersects the singular locus of S exactly at its endpoints. Every closed geodesic on S either lies in the nonsingular part of S or is a concatenation of saddle connections.

Fact 5.9. *If two closed geodesics α, β on a half-translation surface are homotopic to disjoint simple closed curves, then α, β cannot intersect transversely. Moreover, if α, β are homotopic, then they are either equal or they bound a annulus with embedded interior, no interior singular points and all interior angles equal to π.*

Nontransverse intersections of geodesics can only happen at singular points.

Proof. This follows from Gauss-Bonnet for Euclidean surfaces with singularities:

$$\sum_i (\pi - \epsilon_i) + \sum_j (2\pi - \alpha_j) = 2\pi\chi, \tag{5}$$

where the ϵ_i are the interior angles at corners of the boundary, and α_j are the cone angles at singular points.

If α, β are geodesics that are homotopic to disjoint simple closed curves, and they intersect transversely, there are arcs of α, β that bound a (possibly degenerate) bigon B. (Note: as α, β can intersect nontransversely at singular points, we're really applying the bigon criterion to slight perturbations of α, β that intersect transversely. When we perturb back the bigon so it is bounded by arcs of α, β, it may degenerate, since the two arcs may run through a common singular point. However, Gauss-Bonnet still applies, via a continuity argument.) The bigon B has

[4] With respect to the terminology in Fact 5.7, if $\mathcal{F}$ is the angle-θ foliation of S, then the saddle connections of $\mathcal{F}$ are exactly the saddle connections of S that are leaf segments of $\mathcal{F}$.

only two interior angles that can be less than π, so the first sum in the Gauss-Bonnet formula is less than 2π, while the second sum is less than zero, and $2\pi\chi = 2\pi$, a contradiction.

Now assume that α, β are homotopic, and not equal. By the previous paragraph, we can assume that all intersections are nontransverse, in which case α, β bound a degenerate annulus A. Applying (5) to A, the right side is zero and both sums are nonpositive, so both sums are zero, implying all boundary angles are π and there are no interior cone points. Moreover, since all boundary angles are π, the annulus has embedded interior, since if some spanning arc is collapsed to a point, the two geodesics α, β have to be equal. $\square$

It follows that the set of slopes of segments occurring in a geodesic is constant in its homotopy class.

Proof of Proposition 5.8. Realize f as a hyperbolic affine map of a half-translation structure on S, say where the derivative is diagonal with entries $\lambda, 1/\lambda$, $\lambda > 1$. Given a closed curve c on S, let $s(c)$ be the smallest slope that appears in a segment of a geodesic representative of c on S. By Fact 5.7, $s(c) > 0$. So, the numbers $s(f^n(c)) = \lambda^{2n} s(c)$ are all distinct. $\square$

Remark 5.10. *Above, we've assumed that $\partial S = \emptyset$ for simplicity. If S has boundary, one can define pseudo-Anosov just as in the topological definition, but one has to be a bit more careful about what the foliations look like near the boundary. See e.g. Exposé 11 in* [16].

6. The Nielsen-Thurston Classification

Let S be a finite type surface without boundary. The following theorem of Nielsen and W. Thurston[5] classifies elements of the mapping class group $\mathrm{Map}(S)$ in a way that generalizes the trace classification of elements of $SL_2\mathbb{Z} \cong \mathrm{Map}(T^2)$.

Nielsen-Thurston Classification. *Every homeomorphism of S is isotopic to a homeomorphism $f : S \longrightarrow S$ that is either*

(1) periodic*, i.e. $f^n = Id$ for some n,*
(2) pseudo-Anosov, or
(3) reducible*, i.e. there is some multicurve $C \subset S$ such that $f(C) = C$.*

Here, a *multicurve* is a union of disjoint, essential, nonperipheral, non-pairwise-isotopic simple closed curves on S.

We often call an element of $\mathrm{Map}(S)$ periodic, pseudo-Anosov or reducible if it admits a representative of one of that type. Note that a homeomorphism of S can be both periodic and reducible: for instance, draw a genus $g+1$ surface as a circular necklace with g links in the chain. Then rotation by $2\pi/g$ is finite order and maps a pants decomposition of S to itself, permuting the components. A pseudo-Anosov map cannot be periodic or reducible, by Proposition 5.8.

A Dehn twist is an example of a reducible homeomorphism that is not periodic, even after an isotopy. A periodic map that is not reducible (even after isotopy) is

[5]Thurston came up with the modern statement of the theorem and proved it, but he and most other mathematicians were unaware that much earlier, Nielsen had written a series of papers on this topic that contained essentially all the necessary ideas.

the map $r : S \longrightarrow S$ defined by representing a genus g surface S as the opposite side gluing of a $4g$-gon P, and then rotating P by one click, i.e. by $\pi/(2g)$. Indeed:

Claim 6.1. *There is no multicurve that is fixed up to isotopy by r.*

Proof. Realize P as a regular hyperbolic polygon with $\pi/(2g)$-angles. Then S inherits a hyperbolic metric with respect to which r is an isometry. The quotient of S by $\langle r \rangle$ is a sphere Q with three cone points c, v, m, which are the projections of the center of P, all the vertices of P, and the midpoint of all the edges of P.

Suppose r preserves a multicurve up to isotopy. Then it leaves invariant the geodesic realization C of that multicurve. Thhis C can't pass through the center of P, or through a vertex, since r rotates around those points by an angle less than π. There are no simple closed curves on Q that do not pass through the cone points, so each component $c \subset C$ must intersect the set of edge midpoints of P. It follows that r^{2g} restricts to a reflection on c, with 2 fixed points that are distinct edge midpoints p, q, antipodal on c. There is some power r^i such that $r^i(p) = q$, and so $r^i(c) = c$, with r^i exchanging the antipodal points p, q. So, r^{2g} and r^i act on c as distinct order 2 elements, implying the stabilizer of c can't be cyclic, a contradiction. □

The NT classification implies the following characterization of pseudo-Anosovs.

Proposition 6.2. *A homeomorphism $f : S \longrightarrow S$ is isotopic to a pseudo-Anosov map if and only if for every simple closed curve γ on S, all the iterates $f^n(\gamma)$ are homotopically distinct.*

Proof. We prove the 'if' direction, as the other direction is Proposition 5.8. If f is not isotopic to a pseudo-Anosov map, then it is either finite order or reducible. In the first case, $f^n = id$ for some n, so we're done, and in the second we can take γ to be a component of the invariant multicurve. □

If S is a finite type surface with boundary, the classification of homeomorphism is slightly different: any homeomorphism of S fixing ∂S is isotopic rel ∂S to a homeomorphism that is the composition of a periodic, pseudo-Anosov (on $int(S)$), or reducible homeomorphism with powers of Dehn twists around the boundary components. This is a consequence of the statement above and Theorem 3.1.

Why is case (3) called 'reducible'? Take a separating simple closed curve c on S, and let f be a homeomorphism of S that fixes c and restricts to the identity on one component of $S \setminus c$ and a pseudo-Anosov homeomorphism on the other. Then f 'reduces' into two maps of the other types on subsurfaces.

Definition 6.3. A homeomorphism $g : S \longrightarrow S$ is *pure* if there we can write $S = \cup S_i$ as a union of essential subsurfaces-with-boundary with disjoint interiors, such that g leaves each S_i invariant and acts on each as either the identity, a pseudo-Anosov, or a Dehn twist if S_i is an annulus.

Using the NT classification, one can prove that *any homeomorphism f has a pure power*. Namely, one takes a maximal reducing system C, passes to a power so that f preserves each component of C and each complementary component. Applying the NT classification, we can pass to a power so on each complementary component f is either pseudo-Anosov, the identity, or some composition of the two with powers of boundary twists. Then we add in annuli around each component of C where one sees a twist power.

6.1. **Thurston's theorem on composing multitwists.** Above, we saw how to construct pseudo-Anosov maps via branched covers and via explicit compositions of affine maps on specific translation surfaces. Here is a construction that is often often easier to apply in practice. Suppose A, β are multicurves in S, in minimal position. Then α, β *fill* S if every component of $S \setminus (\alpha \cup \beta)$ is a disk.

Theorem 6.4 (W. Thurston)**.** *If S is a finite type surface and α, β are multi curves that fill S, then there is a real number $\mu = \mu(\alpha, \beta) > 0$ such that the representation*

$$\rho : \langle T_\alpha, T_\beta \rangle \longrightarrow PSL_2\mathbb{R}, \quad \rho(T_\alpha) = \begin{pmatrix} 1 & \mu \\ 0 & 1 \end{pmatrix} \quad \rho(T_\beta) = \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix}$$

has the following properties:

(1) an element $f \in \langle T_\alpha, T_\beta \rangle$ is periodic, a multitwist power, or a pseudo-Anosov according to whether $\rho(f)$ is elliptic, parabolic or hyperbolic,
(2) if f is pseudo-Anosov, the dilatation of f is the larger eigenvalue of $\rho(f)$.

Moreover, when α, β are simple closed curves that fill S, we have $\mu = i(\alpha, \beta)$.

Proof. Let's sketch the proof when α, β are simple closed curves. Regard the union $\alpha \cup \beta$ as the 1-skeleton of a decomposition of S into topological polygons. All vertices have degree 4, so dual to this there is a decomposition of S into quadrilaterals, glued edge to edge. Each square is crossed by one arc of α and one arc of β. Equip S with a half-translation structure in which each square is identified with the unit square $[0,1]^2 \subset \mathbb{R}^2$, such that the intersecting arc of α is horizontal and that of β is vertical. Then T_α, T_β can be represented as affine maps with derivatives

$$\begin{pmatrix} 1 & \mu \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix}$$

respectively. For T_α, notice that all the squares in our dual decomposition of S can be listed as $S_1, \ldots, S_\mu$ in such a way that the right side of S_i is glued to the left side of S_{i+1}, cyclically. Performing just these gluings, we get an annulus A of height 1 and circumference μ that further glues up to give S, with α the core curve of A. We can then represent the Dehn twist in this annulus by the shear map

$$\begin{pmatrix} 1 & \mu \\ 0 & 1 \end{pmatrix}.$$

The description of an affine map representing T_β is similar. The representation ρ then just takes f to the derivative of the affine representation of f that one gets by composing the affine maps above. Note that the affine representatives of T_α, T_β fix all the singular points of S, with respect to the half-translation structure.

If $\rho(f)$ is elliptic, then $\rho(f)^n = id$ for some n, and since f fixes all the singular points on S, we have that f^n is the identity in a neighborhood of the singular locus, implying that $f^n = id$ globally, so f is periodic.

If $\rho(f)$ is parabolic, the 1-dimensional eigenspace of $\rho(f)$ gives a singular foliation of S that is preserved by f, and along which f acts with derivative 1. Since f fixes the singular locus, and a power of f leaves invariant the singular leaves of the foliation, and hence acts trivially on them, the singular leaves cannot nontrivially accumulate in S, as if they did there would be an open subset of S on which a power of f acts as the identity, contradicting that the derivative is parabolic. So, the singular leaves of f are all closed, and divide S into a union of surfaces admitting non-singular foliations, which must then all be annuli. It follows that f is a power of a multitwist in these annuli.

If $\rho(f)$ is hyperbolic, then f is a pseudo-Anosov by the first definition in the previous section, with dilatation the largest eigenvalue of $\rho(f)$. □

In the multitwist case, the proof is similar. We still get a dual decomposition of S into rectangles, but in order to represent T_α, T_β simultaneously as affine maps, we have to be careful in choosing the widths and heights of the rectangles. This requires the Perron Frobenius theorem. See pg 394 of Farb-Margalit [14].

Corollary 6.5. *If α, β are multicurves that fill S, then $T_\alpha \circ T_\beta^{-1}$ is pseudo-Anosov.*

Proof. The composition $T_\alpha T_\beta^{-1}$ goes to

$$\begin{pmatrix} 1 & \mu \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \mu & 1 \end{pmatrix} = \begin{pmatrix} 1+\mu^2 & \mu \\ \mu & 1 \end{pmatrix},$$

which has trace bigger than 2. □

It's also worth noting that if α, β are simple closed curves that fill S and intersect more than twice, you can just use $T_\alpha \circ T_\beta$ above instead. However, in general you need the inverse. For example, note that

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix},$$

which has trace 1, so the product of twists around two curves on the torus that intersect once actually has finite order.

A lot of the time, nearly all elements in the group $\langle T_\alpha, T_\beta \rangle$ are pseudo-Anosov. For instance, we have the following, which implies that if α, β are simple closed curves that fill and intersect more than twice, then all elements of $\langle T_\alpha, T_\beta \rangle$ are pseudo-Anosov except those that are conjugate to powers of T_α, T_β.

Claim 6.6. *If $\mu > 2$, then the subgroup*

$$\Gamma = \left\langle \begin{pmatrix} 1 & \mu \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix} \right\rangle < PSL_2\mathbb{R}$$

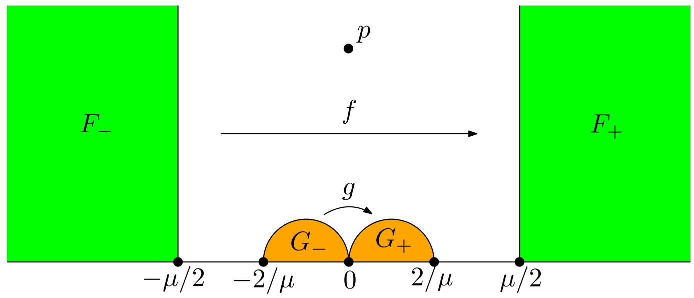
is freely generated by the given elements, is discrete, and the only non-hyperbolic elements in Γ are conjugates of powers of the generators, or in the case that $\mu = 2$, conjugates of powers of their product.

Proof Sketch. Consider the action of Γ on the upper half plane, via fractional linear transformations. Writing the two generators as f, g, respectively, you can check that f acts as a shift to the right by μ, sending the complement $\mathbb{H}^2 \setminus F_-$ to F_+, while g fixes 0 and sends $\mathbb{H}^2 \setminus G_-$ to G_+.

To see that Γ is free, take some reduced word in f, g, like $w = g^2 f^{-3} g^{-5} f$. Where does w send the point p in the picture? Well, we have

$$p \xmapsto{f} \in F_+ \xmapsto{g^{-5}} \in G_- \xmapsto{f^{-3}} \in F_- \xmapsto{g^2} \in G_+,$$

so as $w(p) \in G_+$, in particular $w(p) \neq p$, so $w \neq 1$. A similar technique works in general, to prove both freeness of Γ, and with slightly more care, discreteness. This type of argument is called 'ping pong', since you're supposed to imagine the point p bouncing back and forth between the different sets above.



Assuming $\mu > 2$ for simplicity, one can also use ping pong to prove that the only nonhyperbolic elements of Γ are conjugate to powers of the generators. Indeed, take a cyclically reduced word $w \in \Gamma$. Then the initial and terminal letters of w are not inverses: say they're f and g for simplicity. If $n \in \mathbb{N}$, then w^n is a reduced word starting in f and ending in g, so $w^n(p) \in G_+$. Inverting, $w^{-n}(p) \in F_-$. So, w has to be hyperbolic with attracting fixed point in ∂G_+ and repelling fixed point in F_-. The only way this argument can break is if w starts and ends in f, f or g, g, say, in which case it could be that w is parabolic fixing the unique point of intersection of $\partial F_-, \partial F_+$ or $\partial G_-, \partial G_+$, i.e. w has the same fixed point as either f or g, in which case it's a power of f or g. If $\mu = 2$ a similar issue arises since G_+ intersects F_+ at infinity, and similarly for $-$. □

6.2. **Geometrization of mapping tori.** Let S be a closed surface with genus at least 2 and let $f : S \longrightarrow S$ be homeomorphism. The *mapping torus* of f is

$$M_f := S \times [0,1]/(x,0) \sim (f(x),1),$$

which is a closed 3-manifold. We then have:

Theorem 6.7 (Thurston [56])**.** *The mapping torus M_f admits a hyperbolic metric if and only if f is pseudo-Anosov.*

The 'only if' direction is a consequence of Nielsen-Thurston plus a bit of hyperbolic geometry. Namely, we use:

Lemma 6.8. *If M is a closed hyperbolic 3-manifold, then every abelian subgroup of $\pi_1 M$ is cyclic.*

Proof. Regard $M = \Gamma \backslash \mathbb{H}^3$, with $\pi_1 M \cong \Gamma$. Then every nontrivial element of Γ has hyperbolic type, and one can check that two hyperbolic elements commute only if they have the same fixed points, and the stabilizer of a geodesic in $\mathbb{H}^3$ is discrete only when it is cyclic. □

Above, the fundamental group of the mapping torus M_f splits as

$$1 \longrightarrow \pi_1 S \longrightarrow \pi_1 M_f \longrightarrow \mathbb{Z}.$$

Pick some $g \in \pi_1 M_f$ projecting to $1 \in \mathbb{Z}$. Then g acts by conjugation on $\pi_1 S$ via the map f_*, which is well-defined up to conjugacy. If f is reducible or periodic, there's some element $\gamma \in \pi_1 S$ such that $f_*^n(\gamma)$ is conjugate to γ, and then if we conjugate g appropriately by an element of $\pi_1 S$, we'll get a new g such that $g\gamma g^{-1} = \gamma$, in which case $\langle g, \gamma \rangle \cong \mathbb{Z}^2$.

The other direction, however, is much harder. Assuming f is pseudo-Anosov, Thurston iterates f to construct a sequence of hyperbolic metrics on $S \times \mathbb{R}$ that converge in some sense to a limit structure that turns out to be the infinite cyclic cover of a hyperbolic structure on M_f.

Theorem 6.7 is a special case of the following theorem, which was conjectured by Thurston in the 80's and proved by Perelman [44, 45, 46] in 2003.

Theorem 6.9 (The Hyperbolization Theorem)**.** *A closed* 3*-manifold* M *admits a hyperbolic metric if and only if*

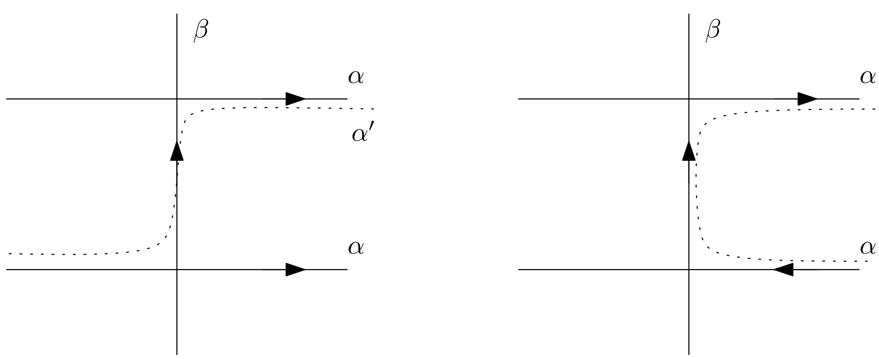
- M *is 'irreducible', i.e. every* 2*-sphere in* M *bounds a ball,*
- $\pi_1 M$ *is infinite, and has no* $\mathbb{Z}^2$ *subgroups.*

7. The curve graph and applications

Let S be a finite type surface that is *nonsporadic*, by which we mean that S is not a sphere with ≤ 4 punctures, or a torus with ≤ 1 puncture. The *curve graph* of S, written $C(S)$, is the graph whose vertices are isotopy classes of essential, nonperipheral simple closed curves on S, where edges connect classes with disjoint representatives. Note that every vertex of $C(S)$ has infinite degree and the mapping class group $\mathrm{Map}(S)$ acts on $C(S)$ by graph automorphisms.

Lemma 7.1. $C(S)$ *is connected.*

Proof. We'll do the proof when S is closed. Take two curves α, β and assume they intersect. If they intersect only once, they're both disjoint from the boundary of their regular neighborhood, which is essential since S is not a torus. So, assume they intersect at least 2 times. Focusing attention on two consecutive intersections of β with α, we have one of the following two pictures, depending on whether the signs of those intersections agree or not.



In each case, we can surger α to give an essential curve α' that intersects β fewer times and is disjoint from α. On the left, α' is the dotted curve, while on the right, α' is either of the two dotted curves. Iterating, we get a path from α to β. □

Since $C(S)$ is connected, we can equip it with a path metric where each edge has length 1. Doing the argument above a bit more carefully proves the inequality

$$d(\alpha, \beta) \leq 2\log_2(i(\alpha, \beta)) + 2.$$

To explain the log, note that in each of the two pictures, there are two possible choices for α', namely the dotted curve and its reflection over β. If we choose the one that intersects β fewer times, then at every step in the process we cut the number of intersections with β at least in half, so the total number of steps needed to get to β is logarithmic. Note that there's no bound in the other direction, since curves at distance 2 can intersect arbitrarily many times.

Remark 7.2. *Sometimes, it's useful to equip $C(S)$ with the structure of a simplicial complex, where $v_1, \dots, v_{k+1}$ span a k-simplex if they can all be realized disjointly, but for the most part we'll just consider $C(S)$ as a graph here.*

On sporadic surfaces, any two disjoint, nonperipheral, essential simple closed curves are isotopic, so the curve graph as defined above would be totally disconnected. However, at least on T^2, the once-punctured torus, and the 4-punctured sphere, one gets an interesting graph $C(S)$ by redefining the edge relation to be 'minimal intersection' instead of 'disjoint'.

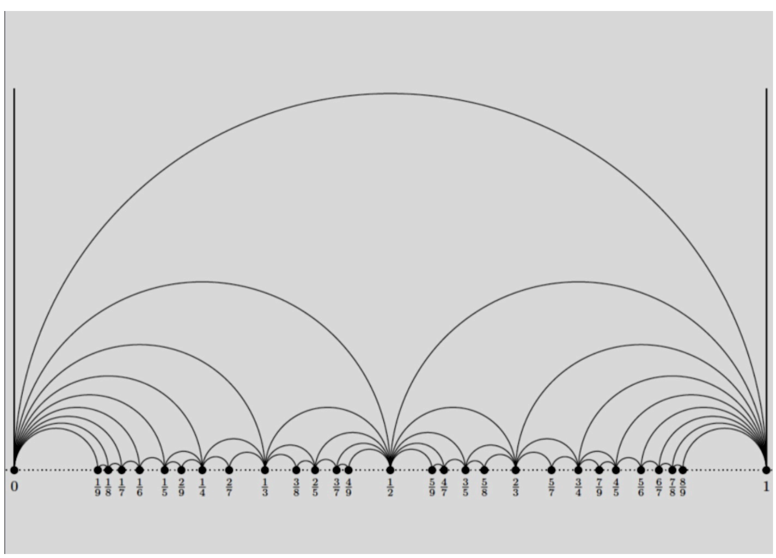
Example 7.3 (The Farey graph)**.** *The curve graph of the torus T^2 is called the Farey graph, denoted $\mathcal{F}$. Its vertices are isotopy classes of (unoriented) simple closed curves on T^2. We can identify the vertices of $\mathcal{F}$ with the extended rational numbers via*

$$[(p,q)] \in H_1(T^2, \mathbb{Z}) \longmapsto p/q \in \mathbb{Q} \cup \infty.$$

Edges connect vertices $p/q, r/s$ that intersect once, so where

$$1 = |det \begin{pmatrix} p & r \\ q & s \end{pmatrix}| = |ps - rq|.$$

It's convenient to picture the Farey graph by considering its vertex set $\mathbb{Q} \cup \infty$ as a subset of $\mathbb{R} \cup \infty$, which we identify with the boundary of the upper half plane $\mathbb{H}^2$, and drawing the edges of $\mathcal{F}$ as hyperbolic geodesics. For instance, here's a picture of a piece of the Farey graph (credit to Jan Karabaš).



The mapping class group $\mathrm{Map}(T^2) \cong SL(2, \mathbb{Z})$ *acts on its curve graph, and in the above picture, the action is the restriction of the action of $SL(2, \mathbb{Z})$ on $\mathbb{H}^2 \cup \partial_\infty \mathbb{H}^2$ by fractional linear transformations. This action is simply transitive on the set of (co-)oriented edges of $\mathcal{F}$. From this, you can deduce some useful properties:*

(1) None of the edges of $\mathcal{F}$ intersect. *Indeed, it suffices to check this when one of the edges is* $[0,\infty]$*. If the edge from* p/q *to* r/s *crosses* $[0,\infty]$*, then we can assume* r *is negative and* p,q,s *are positive, but then*

$$ps - rq \geq 2,$$

so p/q *and* r/s *can't be Farey neighbors.*

(2) Every component of $\mathbb{H}^2 \setminus \mathcal{F}$ is an ideal triangle. *It suffices to show this is true for the two components adjacent to* $[0,\infty]$*, but that's clear from the picture above.*

Each vertex α *of* $\mathcal{F}$ *has infinite valence, and its Farey neighbors all differ by Dehn twists around* α*. By (1), every edge of* $\mathcal{F}$ *separates* $\mathbb{H}^2 \cup \partial\mathbb{H}^2$*. You can use this, and the symmetry of* $\mathcal{F}$*, to prove for instance that* $\mathcal{F}$ *has infinite diameter. Namely, if you start with an edge* e_1 *of* $\mathcal{F}$*, and look to a specified side of that edge, you can always find another edge* e_2 *on that side that's disjoint from the first, even at the endpoints. Repeating, you get a sequence of edges* $e_1, e_2, \ldots$ *cutting off a nested sequence of half-planes* $P_1 \supset P_2 \supset \cdots$*. If you pick a vertex* v *outside* P_1 *and a vertex* w *inside* P_n*, then* $d(v,w) \geq n$*, since any path from* v *to* w *has to pass through a vertex of each* e_i*.*

There's some interesting number theory going on in the geometry of the Farey graph. For instance, it turns out that geodesics in the Farey graph are related to continued fraction expansions. See these notes of C. Series [52] *or Chapter 19 of Schwartz's book* [49]*.*

It's also sometimes useful to consider other related graphs, like the *arc graph* $\mathcal{A}(S)$ or the *arc and curve graph* $\mathcal{AC}(S)$, where now vertices are isotopy classes of properly embedded essential simple arcs in S, or either arcs or curves in S, respectively, and edges connect arcs or curves that intersect minimally. If S is allowed to have boundary, we require all isotopies to fix the boundary. Note that for finite type surfaces without boundary, there are only countably many isotopy classes of arcs and curves, while for surfaces with boundary, there are always uncountably many arcs up to isotopy, since there are uncountably many possible endpoints.

Example 7.4. *Suppose* A *is a compact annulus. The vertex set of the arc complex* $\mathcal{A}(A)$ *consists of all arcs from one boundary component to the other, mod isotopies rel endpoints. So as mentioned above,* $\mathcal{A}(A)$ *has uncountably many vertices. However,* $\mathcal{A}(A)$ *is quasi-isometric to* $\mathbb{R}$*. Indeed, identify*

$$A = \mathbb{R} \times [0,1]/\mathbb{Z},$$

where $\mathbb{Z}$ *acts by integer translations on the first factor. If* $\gamma \in \mathcal{A}(A)$*, there's a unique lift* $\tilde{\gamma}$ *that has an 'initial' endpoint in* $[0,1) \times [0,1]$*, and we define* $e(\gamma)$ *to be the first coordinate of the other ('terminal') endpoint of* $\tilde{\gamma}$*. Then*

$$|d(\gamma,\delta) - |e(\gamma) - e(\delta)|| \leq 2,$$

which you can prove using the following observations.

- *If* γ,δ *are disjoint then* $|e(\gamma) - e(\delta)| < 1$*. Hence, moving one step along a path in* $\mathcal{A}(A)$ *changes* $e(\cdot)$ *by at most* 1*, so* $d(\gamma,\delta) \geq |e(\gamma) - e(\delta)|$*.*
- *Given* γ,δ*, say with* $e(\gamma) \leq e(\delta)$*, you can construct a sequence of arcs* $\gamma = \gamma_0, \ldots, \gamma_n$*, where* $n = \lfloor e(\delta) - e(\gamma) \rfloor$*, by fixing the initial endpoint while incrementing* $e(\gamma_i)$ *until it is less than* 1 *from* $e(\delta)$*. After perturbing the*

endpoints of the γ_i's, each is disjoint from the next, so we have a path in $\mathcal{A}(A)$. You can check that $d(\gamma_n, \delta) \le 2$.

Here's a first truly nontrivial result about curve graphs. It was first proved by Masur-Minsky in [37] using different methods. I'm not sure where the proof below appears in the literature, but as far as I know it's the simplest one.

Proposition 7.5. *If S is a finite type surface, and not a sphere with at most 3 punctures, the curve graph $C(S)$ has infinite diameter.*

To get some intuition for the proof, first consider the following statement. *Suppose $A \subset \mathbb{R}^2$ and for all $a \in A$, the set A is invariant under a rotation*

$$R_a : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

around a by an angle at least $\pi/3$. Then either A is a point or $diam(A) = \infty$. To prove this, assume not and take two (distinct) points $a, b \in A$ at maximal distance. Since $d(a, R_b(a)) > d(a, b)$, a contradiction.

Proof. We'll do the proof in the nonsporadic case. First, let's note that the diameter of $C(S)$ is at least 3, i.e. that there are two curves that together fill the surface S. For instance, if P is a pants decomposition for S and γ is a simple closed curve that intersects each component of P, then $\gamma, T_P^2(\gamma)$ fill the surface. To see this, note that you can draw each component of P on the union $\gamma \cup T_P^2(\gamma)$, so any simple closed curve disjoint from both $\gamma, T_P^2(\gamma)$ has to lie in one of the complementary pants, and hence be one of the pants curves, but those all intersect γ.

Now suppose that $C(S)$ has finite diameter, and let α, β be curves at maximal distance. We want to show that for large n, we have $d(\alpha, T_\beta^n(\alpha)) > d(\alpha, \beta)$, a contradiction. So, a high power of T_β is playing the role of the rotation above. To measure angles, the appropriate replacement for the 'tangent space at b' is the arc complex of an annular neighborhood A of β.

Equip S with a hyperbolic metric and think of vertices of $C(S)$ as their geodesic representatives. Let A be a small, regular metric neighborhood of β. If γ is a closed geodesic on S that intersects β transversely, define the *projection* $\pi_A(\gamma) \in \mathcal{A}(A)$ to be any component of $\gamma \cap A$ that goes from one boundary component to the other. Note that there is a choice involved in defining each $\pi(\gamma)$, but the choice can only change the image by distance 1.

We claim that for n large, any geodesic in $C(S)$ from α to $T_\beta^n(\alpha)$ passes within 1 from β, so that we have

$$d(\alpha, T^n(\alpha)) \ge d(\alpha, \beta) + d(T^n(\alpha), \beta) - 2 = 2d(\alpha, \beta) - 2 > d(\alpha, \beta)$$

where the last inequality is because we know $d(\alpha, \beta) > 3$. Since α intersects β transversely, if n is large, $\pi_A(T_\beta^n(\alpha))$ spins many times around the annulus, so is very far from $\pi_A(\alpha)$ in $\mathcal{A}(A)$, say much farther than $diam(C(S))$. If a geodesic from α to $T_\beta^n(\alpha)$ does not pass within 1 of β, all its vertices project nontrivially to arcs in $\mathcal{A}(A)$, giving a path between $\pi_A(T_\beta^n(\alpha))$ and $\pi_A(\alpha)$, which contradicts that those vertices are far apart in the arc graph. ∎

8. Gromov hyperbolicity

Let X be a metric space. A *geodesic* in X is a path whose length is the distance between its endpoints. We say that X is 'geodesic' if every pair of points is connected by at least one geodesic. A *geodesic triangle* in X is a triple of geodesics

α, β, γ such that the terminal endpoint of each is the initial endpoint of the next, cyclically.

Definition 8.1 (Hyperbolicity)**.** Let $\delta \geq 0$. A geodesic metric space X is called *δ-hyperbolic* if each side of a geodesic triangle is contained in the δ-neighborhood of the union of the other two. More generally, we call X *(Gromov) hyperbolic* if it is δ-hyperbolic for some δ.

The above is often summarized by saying that triangles in X are *δ-thin.*

The prototypical example is hyperbolic n-space $\mathbb{H}^n$. To see this, note that any hyperbolic triangle is contained in a hyperbolic plane, and applying an isometry we can assume that the triangle has vertices $-1, 1, \infty$ in the upper half plane model, with $[-1, 1]$ the edge in question. The farthest point on this edge from the other two edges is i, which is certainly at most 1 from those edges, so $\mathbb{H}^n$ is 1-hyperbolic. If you're interested in the minimal δ, you can also just note that i is closest to $1 + \sqrt{2}i$ on the $[1, \infty]$ edge, and the distance between these two points is

$$\tanh^{-1}(1/\sqrt{2}) \approx .881375.$$

Next, every simplicial tree is 0-hyperbolic. Indeed, any path in a tree contains in its image the unique embedded path with the same endpoints, so if α, β, γ is a geodesic triangle, then $\alpha \cup \beta \supset \gamma$.

Euclidean space $\mathbb{R}^n$ is *not* Gromov hyperbolic: a huge triangle with angles bounded away from zero will require a huge δ.

One reason for the definition above is that it captures many of the important geometric features of $\mathbb{H}^n$ and trees, but is preserved under quasi-isometries. Namely, if X, Y are metric spaces, a map $f : X \longrightarrow Y$ is a (K, C)*-quasi-isometry* if we have

$$\frac{1}{K} d(x, y) - C \leq d(f(x), f(y)) \leq K d(x, y) + C \tag{6}$$

and if every point in Y is at most C away from some point of $f(X)$. Note that quasi-isometries need not be continuous: it's only the large scale metric data that matters. We say X, Y are *quasi-isometric* if there's a quasi-isometry from X to Y. You can check that

Proposition 8.2. *If X is Gromov hyperbolic and Y is quasi-isometric to X, then Y is Gromov hyperbolic.*

We'll prove this in a minute. However, here's why it's useful. Given a group G generated by a finite set S, the *Cayley graph* is the graph $Cay(G, S)$ whose vertex set is G, and where each g is connected to gs by an edge, for all $s \in S \cup S^{-1}$. The distance between two vertices of $Cay(G, S)$ is given by the 'word metric'

$$d_S(g, h) = |gh^{-1}|_S, \quad |x|_S = \min\{n \mid x = s_1 s_2 \cdots s_n,\ s_i \in S \cup S^{-1}\}.$$

Now, a group can have different finite generating sets S, T, giving different Cayley graphs and different word metrics d_S, d_T. But if $M = \max_{t \in T} |t|_S$, then

$$|x|_T \leq M |x|_S,$$

as any word in $T \cup T^{-1}$ can be rewritten as a word in $S \cup S^{-1}$ by writing each element of $T \cup T^{-1}$ out in the S generators. Increasing M so the same statement holds with S, T switched, we get that

$$\frac{1}{M} d_S(g, h) \leq d_T(g, h) \leq M d_S(g, h).$$

So, if we define a map $i : Cay(G,S) \longrightarrow Cay(G,T)$ that's the identity on vertices and takes each edge to the image of one of its vertices (say), then i is a quasi-isometry. Hence, G has a well-defined 'quasi-isometry class'.

Definition 8.3. A finitely generated group G is *(Gromov) hyperbolic* if some/any of its Cayley graphs are hyperbolic.

Free groups are hyperbolic, since they have Cayley graphs that are trees. We also have the following proposition:

Proposition 8.4 (Milnor-Schwartz)**.** *If G is a finitely generated group acting properly discontinuously and cocompactly on a geodesic metric space X, and $x \in X$, then the orbit map $G \longrightarrow X,\ g \mapsto g(x)$ is a quasi-isometry.*

You can find a proof of this in any geometric group theory text. It follows then from Propositions 8.2 and 8.4 that fundamental groups of closed hyperbolic n-manifolds are hyperbolic groups.

To prove the Proposition 8.2, we need a couple lemmas.

Lemma 8.5 (Detours are expensive)**.** *Suppose that X is Gromov hyperbolic, γ is a path from p to q in X, and $[p,q]$ is a geodesic from p to q. Then $\forall x \in [p,q]$,*

$$d(x,\gamma) \leq \delta \lfloor \log_2 \mathrm{length}(c) \rfloor + 1.$$

So, if you want to travel from p to q and stay D away from some point on $[p,q]$ in the process, you have to use a path whose length is exponential in D. By contrast, in $\mathbb{R}^2$, you need only a linear length in D, which you can see by considering a path consisting of two sides of an equilateral triangle. See Figure 2 for a proof.

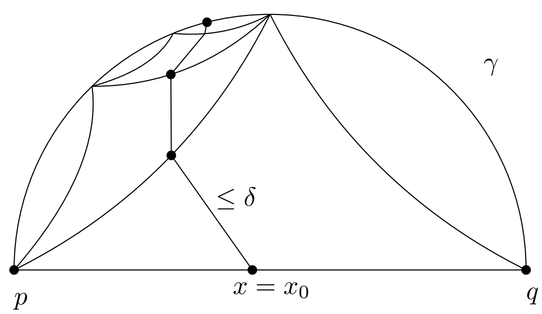


FIGURE 2. The proof of Lemma 8.5. Divide c in half by length and use the midpoint to draw a geodesic triangle. Then $x = x_0$ is within δ of some point x_1 on one of the other two sides. The repeat, dividing the corresponding half of c again in half and making a new geodesic triangle. In at most $\log_2 \mathrm{length}(c)$ steps we arrive at a subsegment of c with length at most 1, in which case we're done.

A (K,C)-*quasi-geodesic* is a map $\gamma : I \longrightarrow X$, where $I \subset \mathbb{R}$ is an interval, such that (6) holds for γ. For instance, the composition of a geodesic and a quasi-isometry is a quasi-geodesic.

Lemma 8.6 (Quasigeodesic stability, aka the 'Morse Lemma')**.** *Given K, C, δ, there's some D such that if X is δ-hyperbolic and γ is a (K,C)-quasigeodesic in X with the same endpoints as a geodesic $[p,q]$, then we have $d_{Haus}(\gamma,[p,q]) \leq D$.*

Here, d_{Haus} is Hausdorff distance, so the lemma says that each point of γ is within R of $[p, q]$, and vice versa. Note that this lemma fails for $\mathbb{R}^2$, as the two non-hypotenuse sides of a right triangle together form a $(2, 0)$-quasigeodesic. Indeed, taking points p, q on the two sides at distances a, b from the right angle,

$$\frac{a+b}{2} \leq \max\{a, b\} \leq d(p, q) = \sqrt{a^2+b^2} \leq a+b.$$

However, the right angle can be arbitrarily far from the hypotenuse.

Proof. First, we can replace γ with a piecewise linear approximation, parametrized by arc length, up to changing the image by a bouded Hausdorff distance, and changing the quasi-geodesic constants by a controlled amount.

Let D be the maximal distance from a point in $[p, q]$ to γ. We claim that D is bounded above in terms of K, C, δ. To see this, suppose $y \in [p, q]$ and $d(y, \gamma) = D$. Pick x, z on either side of y at distance D along $[p, q]$, and pick points a, c on γ at distance $\leq D$ from x, z, and let $\gamma' \subset \gamma$ be the segment with endpoints a, c. Then since γ' is an arc-length parametrized quasi-geodesic whose endpoints are at most $4D$ apart, length(γ') is linear in D. However, the concatenation of γ' with $[x, a]$ and $[c, z]$ is a path from x to z with length linear in D stays D away from the point y. In light of the previous lemma, D is bounded.

We now claim that γ lies in a bounded neighborhood of γ. Take the D from the previous paragraph, and decompose γ as a concatenation $\gamma_1 \cdot \gamma_2 \cdot \gamma_3$, where γ_2 is a maximal subpath that lies outside the D-neighborhood of $[p, q]$. Since every point of $[p, q]$ is within D of some point on γ, the subsets $U_1, U_3 \subset [p, q]$ consisting of points within γ_1, γ_3 are both closed and union to $[p, q]$, and hence they intersect, so we can take some point $x \in [p, q]$ that's within D of points $y_1 \in \gamma_1, y_3 \in \gamma_3$. But then the length of γ_2 is at most that of the subpath of γ from y_1 to y_3, which is bounded since $d(y_1, y_3) \leq 2D$. Hence, γ_2 can only stray a bounded distance from $[p, q]$. $\square$

As a consequence of the above, one can replace geodesics by quasi-geodesics in the definition of Gromov hyperbolicity. Namely, a metric space X is hyperbolic if and only if for every K, C, there's some $\delta > 0$ such that any (K, C)-quasi-geodesic triangle in X is δ-thin. This implies Proposition 8.2, quasi-isometry invariance of Gromov hyperbolicity. It also allows one to extend the definition of hyperbolicity to non-geodesic metric spaces, allowing us to talk about hyperbolicity of a group just in terms of the word metric rather than passing through the Cayley graph.

9. Hyperbolicity of the curve graph

In this section we prove hyperbolicity of the Farey graph, and then the curve graphs of nonsporadic surfaces.

Claim 9.1. *The Farey graph $\mathcal{F}$ is hyperbolic.*

There are many ways to prove this, see e.g. Minsky's lecture notes [41] for one example, where he shows that $\mathcal{F}$ is $\frac{3}{2}$-thin. The essential ingredient in all proofs, though, is that edges of the Farey graph separate. In our proof we'll sacrifice the particular δ for a more conceptual proof, using the fact that the dual graph the Farey tessellation is a 3-valent tree.

Proof. Let $\mathcal{T}$ be the dual tree to the Farey tessellation, and let $\hat{\mathcal{T}}$ be the graph obtained from $\mathcal{T}$ by adding in all the Farey vertices (but not the edges), and connecting each Farey vertex v to all Farey triangles that have v as a vertex. Combinatorially, if we think of $\mathcal{T}$ together with its embedding in the plane, then it comes with a collection of *horocycles*, biinfinite paths where all turns are in the same direction, and $\hat{\mathcal{T}}$ is obtained from $\mathcal{T}$ by coning off all the horocycle.

The Farey graph $\mathcal{F}$ is quasiisometric to $\hat{\mathcal{T}}$, via the identity map on the vertices of $\mathcal{F}$. It's coarse Lipschitz since if v, w are Farey neighbors, there's a path of length 2 from one to the other in $\hat{\mathcal{T}}$. For the other inequality, given any path in $\hat{\mathcal{T}}$ we can replace each dual vertex with an adjacent Farey vertex, noting that adjacent dual vertices give Farey vertices at distance at most 2 in $\mathcal{F}$.

To show that $\hat{\mathcal{T}}$ is hyperbolic, note that if γ is a geodesic in $\hat{\mathcal{T}}$ then we can push it into $\mathcal{T}$ (say with vertices in $\mathcal{T}$) by replacing each visit to a Farey vertex with a segment of the corresponding horocycle. The result is an embedded path $\bar{\gamma}$ in $\mathcal{T}$ (i.e. a geodesic) that is at Hausdorff distance 1 from γ in $\hat{\mathcal{T}}$. So if we have a geodesic triangle in $\hat{\mathcal{T}}$, push all three paths into $\mathcal{T}$, note that the union of two of these covers the third, and conclude that the original triangle was 2-thin, say. $\square$

The proof above is pretty specific to the Farey graph. However, we also have

Theorem 9.2. *If S is a nonsporadic finite type surface, $C(S)$ is hyperbolic.*

This was shown by Masur-Minsky [37] in 2001. We'll prove the theorem when S is closed, following the much simpler 2013 proof by Sisto-Przytycki [47], which builds on earlier work of Hensel-Przytycki-Webb [26]. The following (including pics) is all direct from Sisto's blog, at `https://alexsisto.wordpress.com/2013/09/20/an-even-shorter-proof-that-curve-graphs-are-hyperbolic/`.

We require the following fact. If X is a metric space and $\delta > 0$, a *δ-path* is a sequence $p_0, \ldots, p_n$ with $d(p_i, p_{i+1}) < \delta$ for all i, and $A \subset X$ is δ-connected if any two points in A can be joined by a δ-path in A.

Lemma 9.3 (Masur-Schleimer [36])**.** *Let X be a metric graph, let $\delta > 0$, and suppose that for each $x, y \in X$ we have a subset $A(x, y) \subset X$ such that*

(1) $A(x, y)$ is δ-connected,
(2) if $d(x, y) \leq 1$, then $diam A(x, y) \leq \delta$,
(3) for all x, y, z, the subset $A(x, z)$ is contained in the δ-neighborhood of the union of $A(x, y)$ and $A(y, z)$.

Then X is δ'-hyperbolic for some δ' depending only on δ.

See Bowditch [7] for a 1-page proof. This is sometimes called the 'guessing geodesics lemma' since it turns out after the fact that any geodesic from x to y lies at bounded Hausdorff distance from $A(x, y)$.

Proof of Theorem 9.2. Given $a, b \in C(S)$, let $A(a, b)$ be the set of simple closed curves composed a, b, and all simple closed curves consisting of an arc of a concat an arc of b. Elements of $A(a, b)$ are called *bicorn curves*.

We claim that $A(a, b)$ is 2-connected. First, $A(a, b)$ comes with a partial order, where $c < c'$ if the b-arc of c is contained in that of c'. (So a bicorn if bigger if it has more b in it.) If $c \in A(a, b) \setminus b$, there's always some bicorn d that's bigger and is at most 2 away in $C(S)$, since if $c = a' \cup b'$ then you can extend b' until it hits $int(a')$

again to create a new arc b'', which gives a bicorn d that intersects c at most once. So, starting at any bicorn curve, there's a path of bicorns to b, implying the claim.

It's clear that if a, b are disjoint, then $A(a, b) = \{a, b\}$, so has diameter at most 1. It remains to verify the thin triangle condition. So, say we have a, b, c and a bicorn $x \in A(a, b)$. We claim there's a bicorn $y \in A(a, c) \cup A(b, c)$ that's at most 2 away from x in $C(S)$. Well, put all curves in minimal position and consider 3 consecutive intersections of c with $x = a' \cup b'$. Two of these intersections are on a', say, and we can let $a'' \subset a'$ be the subarc they bound. Then a'' and an arc $c'' \subset c$ form a bicorn $y = a'' \cup c''$, and $i(x, y) \leq 2$, so $d_{C(S)}(x, y) \leq 2$. □

There's a classification of isometries for Gromov hyperbolic spaces, similar to the classification of isometries of $\mathbb{H}^n$. Namely, an isometry $f : X \longrightarrow X$ is *elliptic* if some/every orbit $O_x := \{f^n(x) \mid n \in \mathbb{N}\}$ of f is bounded, and *hyperbolic* if some/every orbit O_x is a quasi-geodesic. Note that orbits of isometries always satisfy

$$d(f^n(\gamma), f^m(\gamma)) \leq K|m - n|, \quad K = d(f(\gamma), \gamma)$$

by the triangle inequality, so the point here is that for hyperbolic type isometries we have a similar lower bound. The remaining isometries are called *parabolic*: they act in ways similar to parabolic isometries of $\mathbb{H}^n$, e.g. they fix a single point in the 'Gromov boundary' of X, which is the single accumulation point of every orbit O_x. See e.g. Ghys-De la Harpe [20] for the more classical case that X is a proper metric space (which $C(S)$ isn't), and Hamann [21] for the proofs in the general case.

The mapping class group $\mathrm{Map}(S)$ acts by isometries on $C(S)$, and you can see the Nielsen-Thurston classification of elements of $\mathrm{Map}(S)$ in terms of their actions on If $f \in \mathrm{Map}(S)$ is reducible or periodic, then f acts elliptically on $C(S)$.

Proposition 9.4. *Pseudo-Anosov mapping classes act hyperbolically on $C(S)$.*

Like hyperbolicity of $C(S)$, this was first shown by Masur-Minsky in [37], and in fact this was their proof that $C(S)$ has infinite diameter. We follow the argument given in Przytycki-Sisto [47]. But first we need the following lemma.

Lemma 9.5. *If $f : S \longrightarrow S$ is pseudo-Anosov, with invariant foliations $\mathcal{F}_\pm$, then every half-leaf of $\mathcal{F}_\pm$ is dense in S.*

Proof sketch, see [15, Cor 14.15] *for details.* This is a sharpening of Fact 5.7, which says that the foliations $\mathcal{F}_\pm$ do not contain closed leaves or saddle connections. To prove it, say for $\mathcal{F}_+$, equip S with a half-translation structure such that $\mathcal{F}_+, \mathcal{F}_-$ are the horizontal and vertical foliations, respectively. Fixing a small vertical arc τ on S, it suffices to show that every half-leaf of $\mathcal{F}_+$ intersects τ. We'll do this using the dynamics of the 'first return map' to τ. Namely, co-orient τ and for each point $p \in \tau$, flow along the leaf of $\mathcal{F}_+$ containing p in the direction of the co-orientation, stopping at the first time when either

(1) you hit a singularity of $\mathcal{F}_+$ or return to an endpoint of τ, or
(2) you return to the interior of τ.

Subdivide τ into arcs $\tau = \cup_i \tau_i$ by cutting along the finitely many points p where (1) occurs. Flowing each τ_i along $\mathcal{F}_+$ eventually comes back to τ (otherwise you get an infinite area strip in S, which is compact), so traces out a rectangle R_i in S. The union $\cup_i R_i$ is all of S, since otherwise it would have horizontal boundary that's a union of closed curves, contradicting Fact 5.7. So, every half leaf in S starts in some R_i and therefore intersects τ. □

As a corollary, suppose f is pseudo-Anosov and S is equipped with a half-translation structure on which f is affine. Then *for every* $\epsilon > 0$ *there's some* $\delta > 0$ *such that if* α *is an arc that makes an angle at least* ϵ *with the horizontal, and has length at least* ϵ*, while* β *is an arc that makes an angle less than* δ *with the horizontal and has length at least* $\frac{1}{\delta}$*, then* α, β *intersect.* This is the property of pseudo-Anosovs we'll use below. You can prove it with a compactness argument, and of course it also works with horizontal/vertical reversed.

Proof of Proposition 9.4. The basic idea is to construct a K-Lipschitz map $\beta : C(S) \longrightarrow \mathbb{R}$ that takes the action of f on $C(S)$ to the shift $x \mapsto x+1$ on $\mathbb{R}$. Then for $\alpha \in C(S)$, we'll have

$$Kd(f^n(\alpha), f^m(\alpha)) \geq |\beta(f^n(\alpha)) - \beta(f^m(\alpha))| = K|n-m|,$$

giving the lower bound necessary to show that orbits are quasigeodesic.

Equip S with a half-translation structure X_0 with respect to which f is affine, say with diagonal derivative, with diagonal entries $\lambda, 1/\lambda$, $\lambda > 1$. For each t, let X_t be the half-translation structure on S obtained by post-composing the charts of X_0 with the diagonal matrix with entries λ^t, λ^{-t}. Then f pushes forward X_t to X_{t+1}. The family $\{X_t \mid t \in \mathbb{R}\}$ will play the role of the $\mathbb{R}$ in the previous paragraph. Note that all the X_t have the same singular points and the same geodesics, in particular the same saddle connections. Note that the italicized statement in the paragraph before the proof on X_t with constants independent of t, since X_t, X_{t+1} are isometric and $\mathbb{R}/\mathbb{Z}$ is compact.

To define the Lipschitz projection, let l_t denote length on X_t. For each saddle connection c, let $\beta(c)$ be the unique time t such that $l_t(c)$ is minimal, or equivalently, the unique time such that c makes a $\pi/4$-angle with the horizontal on X_t. Note that as t increases starting at $\beta(c)$, the geodesic c becomes more horizontal, and its length increases, while if t decreases it becomes longer and more vertical.

Every essential simple closed curve γ on S is homotopic to a unique concatenation of saddle connections. Define

$$\beta : C(S) \longrightarrow \mathbb{R}, \quad \beta(\gamma) := \frac{1}{n}\sum_{i=1}^{n} \beta(c_i), \text{ if } \gamma = c_1 \cdots c_n.$$

Note that $\beta(f(\gamma)) = \beta(\gamma) + 1$, so β takes the f-action on $C(S)$ to the shift as desired. It remains to prove that β is K-lipschitz. For this, it suffices to show that if γ, γ' are disjoint curves, then $|\beta(\gamma) - \beta(\gamma)| \leq K$. (Really, we'll show the contrapositive of this.) We use the following observations.

(1) There's a universal lower bound for the length of any saddle connection in any X_t. To prove this, note that the minimal length of a saddle connection on X_t is nonzero, varies continuously with t, and is invariant under $t \mapsto t+1$.
(2) If $\beta(\gamma) \leq t$, then some saddle connection c of γ makes an angle at most $\pi/4$ with the *horizontal*, and similarly with $\geq$, *vertical*.
(3) Given $\epsilon > 0$, there's some K such that if $\beta(\gamma') \geq t + K$, then γ' has a saddle connection c' that makes an angle at most ϵ with the vertical in X_t, and has length at least $\frac{1}{\epsilon}$. To prove this, take some saddle connection c' of γ' with $\beta(c') \geq t+K$, apply (2) to say it makes an angle at most $\pi/4$ with the vertical on X_{t+K}, and that then as you decrease the parameter from t to $t+K$, the segment c' becomes much more vertical and much longer than the lower bound in (1).

So, if γ, γ' satisfy $\beta(\gamma) = t$ and $\beta(\gamma') \geq t + K(\epsilon)$, then on X_t there's a saddle connection c of γ that's not too short and makes an angle at most $\pi/4$ with the horizontal, while γ' has a saddle connection that's very long and nearly vertical. Hence, by the last sentence in the second paragraph of the proof, the saddle connections c, c' have to intersect transversely, which by Fact 5.9 implies the same for γ, δ. Hence, if γ, δ are disjoint then $|\beta(\gamma) - \beta(\delta)| \leq K$. □

9.1. **The Gromov boundary.** Any δ-hyperbolic metric space X has a *Gromov boundary* ∂X, defined as follows. If $x, y, p \in X$, the *Gromov product* of x, y with respect to p is defined as

$$\langle x, y\rangle_p := \frac{1}{2}(d(x,p) + d(y,p) - d(x,y)).$$

As an exercise, you can show that there's some $D = D(\delta)$ such that $\langle x, y\rangle_p$ is within D of the distance from p to any geodesic $[x, y]$. Fixing $p \in X$, we then define

$$\partial X := \{ \text{ sequences } (x_n) \text{ in } X\}/\sim$$

where $(x_n) \sim (y_n)$ if $\langle x_n, y_n\rangle_p \to \infty$. Here, you can check that $\langle \cdot, \cdot\rangle_p$ extends via

$$\langle (x_n), (y_n)\rangle_p := \lim_{n\to\infty} \langle x_n, y_n\rangle_p$$

to a product on the Gromov boundary, and you can then put a topology on ∂X wherein ξ, ξ' are close if their Gromov product is large. Note that this definition agrees with the usual definition of $\partial\mathbb{H}^n$ as the boundary sphere, since after fixing $p \in \mathbb{H}^n$, points of $\mathbb{H}^n$ are near the same boundary point exactly when the geodesic between them is far from p. You can also check that if T is a tree, then ∂T is the usual Cantor set boundary.

To describe the boundary of $C(S)$, imagine we have a sequence (γ_n) that goes to infinity in $C(S)$. You can imagine that γ_n converges on S to the leaves of a foliation $\mathcal{F}$. Since (γ_n) goes to infinity in $C(S)$, for large n the curves γ_n are wrapping sort of everywhere around the surface S, and it turns out that the limiting foliation $\mathcal{F}$ is *minimal*, meaning all leaves are dense. (It's minimal in the sense that it doesn't restrict to any foliation of a proper subsurface.)

Theorem 9.6 (Klarreich [30])**.** *$\partial C(S)$ can be identiifed with the set of minimal, singular foliations on S that admit transverse measures, considered up to isotopies and Whitehead moves.*

See Klarreich's paper for definitions and more explanation. Whitehead moves involve merging two singularities of a foliation by collapsing a saddle connection joining them. The identification above is actually even homeomorphic, where the topology on the right is given by projective convergence of transverse measures.

It turns out that every hyperbolic type isometry $f : X \longrightarrow X$ of a hyperbolic metric space X has exactly two fixed points $\xi_\pm \in \partial X$, and if $x \in X$ then $f^n(x) \to \xi_\pm$ as $n \to \pm\infty$. If f is a pseudo-Anosov map of S, its two fixed points in $\partial C(S)$ are its stable and unstable foliations.

10. Free and abelian subgroups of Map(S)

The above indicates that perhaps Map(S) shares certain characteristics of a discrete group acting by isometries on $\mathbb{H}^n$. As motivation, let's recall the following basic fact.

Proposition 10.1. *Suppose $\Gamma \subset Isom(\mathbb{H}^n)$ is a discrete group. If $f, g \in \Gamma$ have hyperbolic type, then either*

- *the fixed point sets of f, g are disjoint, and sufficiently large powers f^n, g^m generate a free group, or*
- *$\langle f, g\rangle$ has a finite index cyclic subgroup with hyperbolic generator.*

Proof Sketch. If the fixed point sets are disjoint, you can show f^n, g^m generate a free group for large n, m using ping pong.

If f, g share a single fixed point, you can check that $\langle f, g\rangle$ is not discrete, a contradiction to the hypotheses. For instance, if f, g share their attracting fixed points and translate by $s, t > 0$ along their axes, respectively, fix $a, b \in \mathbb{N}$ and a point x on the axis of f, and note that as $n \to \infty$ we have

$$d(f^{-n} \circ g^{-b} \circ f^a \circ f^n(x), x) \to as - bt,$$

since f^n takes x close to the attracting fixed points, where the axes are nearly the same, and where $g^{-b} \circ f^a$ acts nearly by translation by $as - bt$. Moreover, since the axes of f, g are different, the translation distance on the left hand side above is never zero. But $as - bt$ can be chosen to be arbitrarily close to zero, so $\langle f, g\rangle$ isn't discrete.

If f, g have the same axis α, then you can fix $x \in \alpha$ and map

$$\langle f, g\rangle \longrightarrow \mathbb{R}, \quad h \mapsto d(h(x), x).$$

This map has finite kernel, and its image is a discrete subgroup of $\mathbb{R}$ (hence isomorphic to $\mathbb{Z}$), since otherwise we contradict discreteness of Γ. So we get a short exact sequence

$$1 \longrightarrow F \longrightarrow \langle f, g\rangle \longrightarrow \mathbb{Z} \longrightarrow 1,$$

where F is finite, so taking any element that projects to a generator of $\mathbb{Z}$ gives a finite index cyclic subgroup of $\langle f, g\rangle$. $\square$

The analogous result is true in the mapping class group.

Theorem 10.2. *Suppose $f, g \in \mathrm{Map}(S)$ are pseudo-Anosov. Then either*

- *the invariant foliations of f, g are all distinct, and sufficiently large powers f^n, g^m generate a free group, or*
- *$\langle f, g\rangle$ has a finite index cyclic subgroup with pseudo-Anosov generator.*

The fact that large powers of pseudo-Anosovs with different foliations generate a free group is originally due to McCarthy [38] and independently Ivanov (c.f. [28]). The rest of the content above you can find in [16, §9.5] and [28, Lemma 5.11].

As an application, one can prove:

Theorem 10.3 (McCarthy [38], Ivanov [28])**.** *Suppose that S is a finite type surface and $\Gamma \subset \mathrm{Map}(S)$. Then either Γ is virtually abelian, or it contains a nonabelian free group.*

Abelian subgroups of $\mathrm{Map}(S)$ are well understood. For instance, two pure elements f, g of $\mathrm{Map}(S)$ commute if and only if their active subsurfaces are all either equal or disjoint, and on any shared active subsurface f, g are either both Dehn twists, or are pseudo-Anosovs with the same invariant foliations. As another example, the maximal rank of a free abelian subgroup of $\mathrm{Map}(S)$ is $3g - 3$, realized for instance by products of twists along curves in a pants decomposition. See [6] and [28] for more details.

Theorem 10.3 is often referred to as the *Tits alternative* for mapping class groups, after the following theorem of Jacques Tits.

Theorem 10.4 (Tits [57])**.** *If k is a field of characteristic zero and $\Gamma \subset GL(n,k)$ is a subgroup, then either Γ is virtually solvable (i.e. it has a finite index solvable subgroup) or Γ contains a nonabelian free group.*

As a reminder, a group Γ is solvable if the *derived series*, defined by $\Gamma_0 = \Gamma$ and $\Gamma_i := [\Gamma_{i-1}, \Gamma_{i-1}]$, terminates at the trivial group after finitely many steps. In the mapping class group, every solvable subgroup is virtually abelian, by work of Birman-Lubotsky-McCarthy [6].

Theorem 10.3 is true as stated for groups Γ acting discretely on $\mathbb{H}^n$. Basically, the point is to show that either Γ has a global fixed point at infinity, in which case it is virtually abelian, or it contains a hyperbolic type element f. In the latter case, either Γ centralizes f, or you can find a conjugate g of f that's hyperbolic type with different fixed points, and then large powers of f, g generate a nonabelian free group. The same general outline works for $\mathrm{Map}(S)$, although not exactly, which is perhaps expected since $\mathrm{Map}(S)$ isn't actually hyperbolic. Instead, for $\Gamma \subset \mathrm{Map}(S)$ the analogue of having a global fixed point at infinity is that Γ is *reducible*, i.e. there's a nontrivial multicurve fixed by all of Γ. If Γ is not reducible, you show Γ has a pseudo-Anosov and follow the outline above. If Γ is reducible, then after passing to a finite index subgroup you can assume that it fixes all the complementary components of the reducing system, and then you can do an inductive argument, passing to associated subgroups of the mapping class groups of those components.

There's one other result in this vein I'd like to mention. Recall that if f, g are pseudo-Anosovs with different invariant foliations, then large powers of f, g generate a free group. For general $f, g \in \mathrm{Map}(S)$, you can show that certain large powers of f, g generate either an abelian group or a free group; for instance, you get a $\mathbb{Z}^2$ if f, g are supported on disjoint subsurfaces. There are analogous statements for finite collections $f_1, \ldots, f_n$ of mapping classes, where if the elements are sufficiently complicated then they generate the group with the 'expected' commutativity relations. One such statement is due to Koberda [32] and another to Clay-Leininger-Mangahas [10], which we now describe.

If $G = (V, E)$ is a finite graph, the *right angled Artin group (RAAG)* on G is

$$\Gamma(G) := \langle v \in V \mid [v, w], (v, w) \in E \rangle.$$

So, the generators are vertices of G, and they commute if there's an edge between them. If $f_1, \ldots, f_n$ are pure mapping classes, we can let $G = G(f_1, \ldots, f_n)$ be the graph with vertex set $f_1, \ldots, f_n$, and an edge between f_i and f_j if their supports can be realized disjointly, and we call $\Gamma(G)$ the *expected RAAG*.

Theorem 10.5 (Clay-Leininger-Mangahas [10])**.** *There is some $C > 0$ as follows. Suppose that $f_1, \ldots, f_n \in \mathrm{Map}(S)$ are realized on pairwise nonisotopic, connected subsurfaces $S_i \subset S$, and the translation distance $\tau_{C(S_i)}(f_i) \geq C$ for each i. If Γ is the expected RAAG, then the map $\Gamma \longrightarrow \mathrm{Map}(S)$ is injective, and is even a quasi-isometric embeddding.*

Here, $\tau_{C(S_i)}(f_i)$ is the translation distance on the curve graph $C(S_i)$, i.e. the minimum of $d(\gamma, f_i(\gamma))$ where $\gamma \in C(S_i)$. It's a quasi-isometric embedding with respect to any word metric on $\mathrm{Map}(S)$ coming from a finite generating set.

Remark 10.6. *If $G = (V, E)$ and $\Gamma = \Gamma(G)$ is the associated RAAG, the* extension graph G^e *is defined to have vertices of the form $\gamma v \gamma^{-1}$, where $v \in V, \gamma \in \Gamma$, and two vertices are defined to be adjacent if they commute in Γ. Then Γ acts on G^e by conjugation. Kim-Koberda* [29] *show that G^e plays the same role for Γ as $C(S)$ does for* $\mathrm{Map}(S)$*. In particular, G^e is infinite diameter and hyperbolic (it's actually quasi-isometric to a tree) and there's a sort of Nielsen-Thurston classification for elements depending on whether they act elliptically or hyperbolically on G^e.*

11. Finite generation of $\mathrm{Map}(S)$

One can use a variant of the curve graph to show mapping class groups are finitely generated.

Theorem 11.1 (c.f. [14, §4.1])**.** *If S is a finite type surface, then* $\mathrm{PMap}(S)$ *is generated by finitely many Dehn twists.*

Note that $\mathrm{PMap}(S) \subset \mathrm{Map}(S)$ is a finite index subgroup, so it follows from the above that $\mathrm{Map}(S)$ is also finitely generated. Indeed, we can just add to the finite generating set in the theorem a finite set of mapping classes that realize all permutations of the punctures.

For closed surfaces S, Humphries [27] gave a particularly simple set of $2g + 1$ Dehn twists generating $\mathrm{Map}(S)$, and showed that no smaller set of twists generates. We remark that any closed surface, $\mathrm{Map}(S)$ is 2-generated by work of Wajnryb [59], but the generators are not twists. Wajnryb also gave an explicit presentation for $\mathrm{Map}(S)$ using the Humphries generators [60].

Proof Sketch of Theorem 11.1, see e.g. [14] *for details.* First, let's assume that S has genus 0 with n punctures. Here, the result follows from the Birman exact sequence. Namely, if we fix one of the punctures, p, and let $\hat{S}$ be the surface obtained by filling in p, we get a short exact sequence

$$1 \longrightarrow \pi_1(\hat{S}, p) \xrightarrow{\mathcal{P}} \mathrm{PMap}(S) \longrightarrow \mathrm{PMap}(\hat{S}) \longrightarrow 1,$$

where $\mathcal{P}(\gamma)$ is a point push around γ. Now, $\pi_1(\hat{S}, p)$ is finitely generated by simple closed curves, and a point push around a simple closed curve is isotopic on S to the composition of two Dehn twists, so it follows that image of $\mathcal{P}$ is finitely generated by Dehn twists. An inductive argument then implies that $\mathrm{PMap}(S)$ is too.

Dealing with genus is a bit more complicated. Consider the graph $\mathcal{N}(S)$ whose vertices are nonseparating curves on S, and where edges connect curves that intersect exactly once. Then $\mathrm{PMap}(S)$ acts on $\mathcal{N}(S)$, transitively on directed edges. Indeed, if $[a, b]$ and $[c, d]$ are edges, note that cutting along $a \cup b$ and $c \cup d$ gives a surface with one less genus and one additional boundary components, which is divided into 4 arcs marked by the two curves. Classification of surfaces implies there's a homeomorphism from one picture to the other, which induces a homeomorphism of S taking $[a, b]$ to $[c, d]$. We claim:

(1) $\mathcal{N}(S)$ is connected.

This is pretty easy to prove when S is closed. In Figure 7.1, each possible picture comes with two possible surgeries, one of which is guaranteed to be nonseparating. So, given two nonseparating curves, you can find a path in $C(S)$ from one to the other that only includes nonseparating curves. But then for any two disjoint nonseparating curves, there's a third nonseparating curve that intersects both once. See e.g. Farb-Margalit [14] for details in the case with punctures.

Using (1), we prove that $\mathrm{PMap}(S)$ can be generated by finitely many twists using induction on the genus of S. The base case is the genus 0 case above. For the inductive case, assume it's true for surfaces with smaller genus. Then

(2) For each $a \in \mathcal{N}(S)$, the stabilizer $\Gamma_a \leq \mathrm{PMap}(S)$ is contained in a subgroup of $\mathrm{PMap}(S)$ generated by a finite collection of Dehn twists.

Why? Well, cutting along a gives surface S_a with smaller genus. Then Γ_a is generated by $\mathrm{PMap}(S_a), T_a$, and an element that switches the two sides of a. By induction, $\mathrm{PMap}(S_a)$ is finitely generated by twists, and T_a is a twist. To deal with the side-switching, pick some curve b intersecting a once. Then $(T_a T_b^{-1})^3$ fixes a and switches its two sides, since you can look at it just in the punctured torus neighborhood of $a \cup b$, in which T_a, T_b can be represented as upper/lower triangular elements of $SL_2\mathbb{Z}$ with 1's on the off-diagonal, and then

$$-\left(\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ -1 & 1\end{pmatrix}\right)^3 = -I,$$

which switches the two sides of any curve on the punctured torus. So, Γ_a is contained in the group $\langle \mathrm{PMap}(S_a), T_a, T_b\rangle$, which is finitely generated by twists.

Fix two adjacent vertices $a_0, a_1 \in \mathcal{N}(S)$, with stabilizers Γ_0, Γ_1. We claim that

$$\langle \Gamma_0, \Gamma_1 \rangle = \mathrm{PMap}(S),$$

which by (2) will finish the proof. To do this, pick $f \in \mathrm{PMap}(S)$, and a path

$$a_0, a_1, \ldots, a_n = f(a_0)$$

in $\mathcal{N}(S)$, and let Γ_i be the stabilizer of a_i. Note that for each i, there's an element of Γ_i taking a_{i-1} to a_{i+1}, by directed edge transitivity of the action of $\mathrm{PMap}(S)$.

We first claim that for each i, the stabilizer Γ_i of a_i lies in $\langle \Gamma_0, \Gamma_1 \rangle$. Indeed, if it's true up to i, then take some $h_i \in \Gamma_1$ such that $h_1(a_{i-1}) = a_{i+1}$, which we can do since $\mathrm{PMap}(S)$ acts transitively on directed edges. Then

$$\Gamma_{i+1} = h_i \Gamma_{i-1} h_1^{-1} \subset \langle \Gamma_0, \Gamma_1 \rangle.$$

Suppose that n is even. Then we can find $h_i \in \Gamma_i$ such that if

$$h = h_{n-1} h_{n-3} \cdots h_3 h_1,$$

then $h(a_0) = a_n$. From the previous paragraph, $h \in \langle \Gamma_0, \Gamma_1 \rangle$. But $h^{-1} f \in \Gamma_0$, so $f \in \Gamma_0, \Gamma_1\rangle$. If n is odd, do the same using even indices, finding h such that $h(a_1) = a_n$, and hence $h^{-1} f \in \Gamma_1$. $\square$

12. Subsurface projection

One of the main tools in an inductive study of $C(S)$ and $\mathrm{Map}(S)$ is *subsurface projection*, defined in [35] by Masur-Minsky.

Definition 12.1. Suppose that $X \subset S$ is an essential subsurface that is not an annulus. We define the *subsurface projection* map

$$\pi_X : C(S) \longrightarrow C(X) \cup \{\emptyset\}$$

as follows. If γ can be isotoped to be disjoint from X, we set $\pi_X(\gamma) = \emptyset$. Otherwise, let $\pi_X(\gamma)$ be any essential curve on X disjoint from γ. We set

$$d_X(\alpha, \beta) := d_{C(X)}(\pi_X(\alpha), \pi_X(\beta))$$

The 'any' above may be disconcerting, but we're really only interested in the large scale geometry of π_X and any two choices for $\pi_X(\gamma)$ as above will be ≤ 2 away from each other in $C(X)$. Sometimes, people pick $\pi_X(\gamma)$ to be some essential curve in X produced by surgering γ and ∂X.

As an example, note that if α is a simple closed curve on S intersecting a subsurface X transversely, and $f : S \longrightarrow S$ is a mapping class that is pseudo-Anosov on X, then $d_X(\alpha, f^n(\alpha)) \to \infty$.

For annular subsurfaces $A \subset S$, one needs a slightly different definition. The only interesting 'curve graph' associated to an annulus is its arc graph, which records how much arcs twist around the annulus. However, even if a curve γ in S twists a lot around the core curve of A, this twisting may happen outside of A, so the intersection with A may not record it. One way to fix this is to replace A by a compactification of the cover of S corresponding to A.

Definition 12.2. Suppose $A \subset S$ is an essential annulus. Let S_A be the cover of S corresponding to $\pi_1 A$. If we hyperbolize S, then the hyperbolic metric lifts to a metric on S_A, so we can identify $S_A = \mathbb{H}^2/\mathbb{Z}$, where $\mathbb{Z}$ acts hyperbolically. Then

$$\overline{S}_A = \mathbb{H}^2 \cup \partial\mathbb{H}^2/\mathbb{Z}$$

is a compactification of S_A. We set $C(A) := \mathcal{A}(\overline{S}_A)$, the arc complex of this compactified cover. And then we define

$$\pi_A : C(S) \longrightarrow C(A) \cup \{\emptyset\}$$

by taking curves that don't intersect A transversely to $\emptyset$, and taking any other geodesic γ to the closure in $\overline{S}_A$ of any component of the preimage of γ in S_A that runs from one end of S_A to the other.

Theorem 12.3 (Bounded Geodesic Image (BGI) [35])**.** *There's some $D = D(S)$ as follows. If $\alpha, \beta \in C(S)$, $X \subset S$ is a proper subsurface, and $d_X(\alpha, \beta) > D$, then any geodesic in $C(S)$ from α to β has to pass through a curve disjoint from X.*

For intuition, imagine a geodesic from α to β. If there's no curve disjoint from X on the geodesic, it has to project to a path from $\pi_X(\alpha)$ to $\pi_X(\beta)$, which are far apart, so it has to spend a bunch of time traversing this projection. However, if you just skip out of X at some point, your projections can just warp from $\pi_X(\alpha)$ to $\pi_X(\beta)$. The BGI says this is always the most efficient method if the projection distances are large enough.

This allows you to think of large powers of pseudo-Anosovs f or twists on a subsurface X coarsely as π-rotations around X. Indeed, we have

$$d_X(f^n(\alpha), \alpha) \to \infty$$

so for large n, any geodesic from α to $f^n(\alpha)$ passes through a curve disjoint from X, and the distance $d(\alpha, f^n(\alpha)) \sim 2d(\alpha, \partial X)$.

The BGI is one tool in Masur-Minsky's investigation of the geometry of $\mathrm{Map}(S)$. Here's more of their program. Equip $\mathrm{Map}(S)$ with a finite generating set S, and its corresponding word metric, which we denote by $d_S(\cdot, \cdot)$. Here we describe how to estimate d_S using distances in curve graphs of subsurfaces of S. As a first step, let's define a *(complete, clean) marking* μ of S to be a pants decomposition (called the 'base of the marking) and a set of 'transversals', by which we mean that for each pants curve γ, we pick a simple closed curve on S disjoint from all the other pants curves that intersects γ minimally, i.e. either once or twice depending on whether γ sits inside of a punctured torus or 4-punctured sphere component of the complement of the other pants curves.

Let $\mathcal{M}(S)$ be the locally finite graph whose vertices are markings of S, and where edges connect markings that differ by 'elementary moves', in which we either twist a transversal around its curve, or swap a curve and its transversal, and then surger the other transversals to make them disjoint from the new pants curve. See [35] for details. It turns out that $\mathcal{M}(S)$ is connected (try to prove it!) and the mapping class group $\mathrm{Map}(S)$ acts on $\mathcal{M}(S)$ by isometries, with finite quotient. In particular, by Milnor-Svarc, $\mathrm{Map}(S)$ and $\mathcal{M}(S)$ are quasi-isometric.

So, one can estimate distances in $\mathrm{Map}(S)$ using distances in $\mathcal{M}(S)$. But is that any more tractable? Masur-Minsky's idea is to show that the distance between two markings can be understood in terms of their 'projections' to curve graphs of subsurfaces, and since curve graphs are hyperbolic their geometry is more tractable.

Theorem 12.4 (The Distance Formula [35]). *There's some $K = K(S)$ as follows. If μ, μ' are markings of S, then*

$$d(\mu, \mu') \sim_K \sum_{X \subseteq S} [d_X(\mu, \mu')]_K$$

Here, $f \sim_K g$ means $f(x) \leq Kg(x) + K$ and vice versa, while

$$[x]_K = \begin{cases} 0 & x \leq K \\ x & x \geq K. \end{cases}$$

To define $d_X(\mu, \mu')$, take any base or transversal curves in μ, μ' that intersect X transversely (these always exist) and measure their d_X-distance.

Really, the picture that Masur-Minsky develop is as follows. Take μ, μ', and start by drawing a geodesic γ_S in $C(S)$ from μ to μ'. There are finitely many 'large projection subsurfaces' (LPSs) $X \subset S$, i.e. subsurfaces where $d_X(\mu, \mu') > K$. Partially order the large projection X's with respect to inclusion. For each maximal LPS X, there's some vertex v_X on γ_S that's disjoint from X. Imagine a geodesic γ_X in $C(X)$ from $\pi_X(\mu)$ to $\pi_X(\mu')$ as living in the link of v_X. Then look at an LPS that's maximal among those properly contained in X and continue inductively. The collection of curve graph geodesics in subsurfaces that you get is essentially what Masur-Minsky call a 'heirarchy', and there's a natural way to 'resolve' any heirarchy into a quasi-geodesic path in $\mathrm{Map}(S)$.

Similar distance formulas can be used to estimate the intersection number between two curves (Watanabe [62]), and the distances between two points in Teichmuller space with either the Teichmuller metric [48] or the Weil-Petersson metric (c.f. Theorem 2.6 in [9]). For instance, Watanabe showed that if $\alpha, \alpha' \in C(S)$ then

$$\log i(\alpha, \alpha) \sim_K \sum_{X \subseteq S} [d_X(\alpha, \alpha')]_K + \sum_{A \subset X} [\log d_A(\alpha, \alpha')]_K ,$$

where X ranges over non-annular subsurfaces and A over annular subsurfaces.

Distance formulas have also been found for various subgraphs of the curve graph. For example, say H is a handlebody with boundary S and $\mathcal{D}(H) \subset C(S)$ is the *disc graph*, whose vertices are 'meridians', i.e. curves that bound disks in H. In [36], Masur-Schleimer define a *witness*[6] for H to be a subsurface $X \subset S$ such that every

[6]They actually use the term 'hole' instead of witness in that paper, but the accepted terminology has since changed. In the distance formula for $\mathcal{D}(H)$, the surfaces X 'witness' the distance

meridian for H intersects X. classify all witnesses, and show that

$$d_{\mathcal{D}(H)}(\alpha,\beta) \sim_K \sum_{\text{witnesses } X\subset S} [d_X(\alpha,\beta)]_K.$$

Here's one application, from Masur-Minsky [35].

Theorem 12.5. *Fix a word metric $|\cdot|$ on* $\mathrm{Map}(S)$. *Then there's $K = K(S)$ as follows. If f, g are conjugate pseudo-Anosov elements of* $\mathrm{Map}(S)$*, there's an element $h \in \mathrm{Map}(S)$ such that $g = hfh^{-1}$ such that $|h| \le K(|f| + |g|)$.*

Tao [54] later proved the same statement for arbitrary pairs of conjugate mapping classes. You can use this, for instance, to give an exponential-time algorithm that solves the problem of when two mapping classes are conjugate. (If f, g are, you're guaranteed to find a conjugator by searching through the ball of radius $K(|f|+|g|)$.) The fact that the conjugacy problem is solvable (via some unnamed algorithm) was proved earlier by Hemion [25]. Basically, the point is that the conclusion of Theorem 12.5 is an exercise for pairs of conjugate elements of a hyperbolic group, and that the hyperbolicity of curve graphs $C(X), X \subset S$ is enough to make similar ideas work for $\mathrm{Map}(S)$.

Remark 12.6. *In* [3]*, Behrstock-Hagen-Sisto defined a notion of 'heirarchically hyperbolic space/group' (HHS/HHG) that generalizes the picture above, where the mapping class group can be described by a collection of hyperbolic spaces (the curve graphs $C(X)$, $X \subset S$), which are related by some sort of (subsurface) projection. Other examples of HHGs are $Out(F_n)$, RAAGS, and groups acting geometrically on CAT(0) cube complexes, and in each case there is a distance formula similar to Theorem 12.4. See e.g.* [53] *for a survey.*

12.1. **The Ending Lamination Theorem.** Minsky's motivation in developing the machinery in the last section was to solve Thuston's *Ending Lamination Conjecture*, which informally says that the geometry of a hyperbolic 3-manifold M with finitely generated fundamental group is determined up to isometry by certain geometric invariants of its topological ends. Using his work with Masur above, Minsky [40] and Brock-Canary-Minsky [8] showed that in fact one can combinatorially model the geometry of M using these ending invariants. We'll describe a bit of this theory briefly in one special case.

Let S be a closed orientable surface with genus at least 2, let $f : S \longrightarrow S$ be pseudo-Anosov and let M_f be its mapping torus. As noted before, Thurston [56] showed that M_f admits a hyperbolic metric. Mostow's Rigidity Theorem (c.f. [4]) says that this metric is unique up to isometry. But what does the metric look like? To make the question a bit easier, let's study instead the geometry of the cover

$$N \longrightarrow M_f$$

corresponding to the fundamental group of the surface S, so that $N \cong S \times \mathbb{R}$. Here, there are two ends to N and the two 'ending invariants' discussed above are essentially the two invariant foliations $\mathcal{F}_\pm$ associated to f. So, can you somehow read off the geometry of N from these foliations?

For instance, say γ is a simple closed curve on S and let $\ell_N(\gamma)$ be the length of the geodesic in N homotopic to $\gamma \times \{0\}$. Minsky showed:

in $\mathcal{D}(H)$. The original terminology was supposed to reflect that X behaves like a big hole in the ground that you might be forced to walk around for a long time to get from α to β.

Theorem 12.7 (Minsky, c.f. [42, 43, 40])**.** *Given $\epsilon > 0$ sufficiently small, there's a $K = K(\epsilon, S)$ as follows. If $\gamma \subset S$ is a simple closed curve, then $\ell_N(\gamma) < \epsilon$ if and only if there's some subsurface $X \subset S$ with $\gamma \subset \partial X$ such that*

$$d_X(\mathcal{F}_-, \mathcal{F}_+) \geq K.$$

This is a more precise version of the main theorems in [42, 43]; it follows from the more complicated statement of the *Short Curve Theorem* in [40]. Here, $d_X(\mathcal{F}_-, \mathcal{F}_+)$ is defined by taking the distance between the 'subsurface projections' of $\mathcal{F}_\pm$ in X, where we define the projection of $\mathcal{F}_\pm$ by isotoping X so that ∂X is transverse to $\mathcal{F}_\pm$, and then defining $\pi_X(\mathcal{F}_\pm)$ to be any vertex of $C(X)$ that is disjoint from some arc of $\ell_\pm \cap X$, where $\ell_\pm$ is a leaf of $\mathcal{F}_\pm$.

Really, what happens is that the geometry of M_f reflects a heirarchy of curve complex geodesics joining μ to $f(\mu)$, where μ is a marking. Subsurfaces $X \subset S$ where $d_X(\mu, f(\mu))$ is large correspond to long regions in M_f homeomorphic to $X \times I$, where the curves in $\partial X \times I$ are all very short in M_f.

References

1. Reinhold Baer, *Kurventypen auf flächen.*, (1927).
2. ———, *Isotopie von kurven auf orientierbaren, geschlossenen flächen und ihr zusammenhang mit der topologischen deformation der flächen.*, (1928).
3. Jason Behrstock, Mark Hagen, and Alessandro Sisto, *Hierarchically hyperbolic spaces, i: Curve complexes for cubical groups*, Geometry & Topology **21** (2017), no. 3, 1731–1804.
4. Riccardo Benedetti and Carlo Petronio, *Lectures on hyperbolic geometry*, Universitext, Springer-Verlag, Berlin, 1992. MR 1219310 (94e:57015)
5. Joan S Birman, *Mapping class groups and their relationship to braid groups*, Communications on Pure and Applied Mathematics **22** (1969), no. 2, 213–238.
6. Joan S Birman, Alex Lubotzky, and John McCarthy, *Abelian and solvable subgroups of the mapping class groups*, (1983).
7. Brian H Bowditch, *Tight geodesics in the curve complex*, Inventiones mathematicae **171** (2008), no. 2, 281–300.
8. Jeffrey Brock, Richard Canary, and Yair Minsky, *The classification of Kleinian surface groups 2: the ending lamination conjecture*, Preprint.
9. Jeffrey Brock, Howard Masur, and Yair Minsky, *Asymptotics of weil–petersson geodesics ii: bounded geometry and unbounded entropy*, Geometric and Functional Analysis **21** (2011), no. 4, 820–850.
10. Matt T Clay, Christopher J Leininger, and Johanna Mangahas, *The geometry of right-angled artin subgroups of mapping class groups*, Groups, Geometry, and Dynamics **6** (2012), no. 2, 249–278.
11. CJ Earle and J Eells, *The diffeomorphism group of a compact riemann surface*, (1967).
12. Clifford J Earle and Curt McMullen, *Quasiconformal isotopies*, Holomorphic Functions and Moduli I: Proceedings of a Workshop held March 13–19, 1986, Springer, 1988, pp. 143–154.
13. Robert D Edwards and Robion C Kirby, *Deformations of spaces of imbeddings*, Annals of Mathematics **93** (1971), no. 1, 63–88.
14. Benson Farb and Dan Margalit, *A primer on mapping class groups (pms-49)*, vol. 41, Princeton university press, 2011.
15. ———, *A primer on mapping class groups*, Princeton Mathematical Series, vol. 49, Princeton University Press, Princeton, NJ, 2012. MR 2850125
16. Albert Fathi, Valentin Poénaru, and François Laudenbach, *Thurston's work on surfaces (mn-48)*, (2021).
17. David Fisher, *Groups acting on manifolds: around the zimmer program*, Geometry, rigidity, and group actions **72157** (2011).
18. Bjorn Friberg, *A topological proof of a theorem of kneser*, Proceedings of the American Mathematical Society **39** (1973), no. 2, 421–426.
19. John L Friedman and Donald M Witt, *Homotopy is not isotopy for homeomorphisms of 3-manifolds*, Topology **25** (1986), no. 1, 35–44.

20. Etienne Ghys and Pierre da la Harpe, *Sur les groupes hyperboliques d?apres mikhael gromov*, vol. 83, Springer Science & Business Media, 2013.
21. Matthias Hamann, *Group actions on metric spaces: fixed points and free subgroups*, arXiv preprint arXiv:1301.6513 (2013).
22. Mary-Elizabeth Hamstrom, *Homotopy groups of the space of homeomorphisms on a 2-manifold*, Illinois Journal of Mathematics **10** (1966), no. 4, 563–573.
23. Allen Hatcher, *A 50-year view of diffeomorphism groups*, Slides from a talk at the 50th Cornell Topology Festival (2012).
24. Allen E Hatcher, *A proof of the smale conjecture, diff(s^3)?o(4)*, Annals of Mathematics **117** (1983), no. 3, 553–607.
25. Geoffrey Hemion, *On the classification of homeomorphisms of 2-manifolds and the classification of 3-manifolds*, (1979).
26. Sebastian Hensel, Piotr Przytycki, and Richard CH Webb, *Slim unicorns and uniform hyperbolicity for arc graphs and curve graphs*, arXiv preprint arXiv:1301.5577 (2013).
27. Stephen P Humphries, *Generators for the mapping class group*, Topology of Low-Dimensional Manifolds: Proceedings of the Second Sussex Conference, 1977, Springer, 2006, pp. 44–47.
28. Nikolai V. Ivanov, *Subgroups of Teichmüller modular groups*, Translations of Mathematical Monographs, vol. 115, American Mathematical Society, Providence, RI, 1992, Translated from the Russian by E. J. F. Primrose and revised by the author. MR MR1195787 (93k:57031)
29. Sang-Hyun Kim and Thomas Koberda, *The geometry of the curve graph of a right-angled artin group*, International Journal of Algebra and Computation **24** (2014), no. 02, 121–169.
30. Erica Klarreich, *The boundary at infinity of the curve complex and the relative teichmüller space*, Preprint (1999).
31. Hellmuth Kneser, *Die deformationssätze der einfach zusammenhängenden flächen*, Mathematische Zeitschrift **25** (1926), no. 1, 362–372.
32. Thomas Koberda, *Right-angled artin groups and a generalized isomorphism problem for finitely generated subgroups of mapping class groups*, Geometric and Functional Analysis **22** (2012), 1541–1590.
33. Larry N Mann and Jin-Chen Su, *Actions of elementary p-groups on manifolds*, Transactions of the American Mathematical Society **106** (1963), no. 1, 115–126.
34. Daniel Massart, *A short introduction to translation surfaces, veech surfaces, and teichmüller dynamics*, Surveys in geometry I, Springer, 2022, pp. 343–388.
35. H. A. Masur and Y. N. Minsky, *Geometry of the complex of curves. II. Hierarchical structure*, Geom. Funct. Anal. **10** (2000), no. 4, 902–974. MR MR1791145 (2001k:57020)
36. Howard Masur and Saul Schleimer, *The geometry of the disk complex*, Journal of the American Mathematical Society **26** (2013), no. 1, 1–62.
37. Howard A. Masur and Yair N. Minsky, *Geometry of the complex of curves. I. Hyperbolicity*, Invent. Math. **138** (1999), no. 1, 103–149. MR MR1714338 (2000i:57027)
38. John McCarthy, *A ?tits-alternative? for subgroups of surface mapping class groups*, Transactions of the American Mathematical Society **291** (1985), no. 2, 583–612.
39. John Milnor, *On spaces having the homotopy type of a cw-complex*, Transactions of the American Mathematical Society **90** (1959), no. 2, 272–280.
40. Yair Minsky, *The classification of Kleinian surface groups i: models and bounds*, Preprint.
41. Yair N Minsky, *A geometric approach to the complex of curves on a surface*, Topology and Teichmüller spaces (Katinkulta, 1995) (1996), 149–158.
42. Yair N. Minsky, *Kleinian groups and the complex of curves*, Geom. Topol. **4** (2000), 117–148 (electronic). MR 2182094 (2006k:57051)
43. ———, *Bounded geometry for Kleinian groups*, Invent. Math. **146** (2001), no. 1, 143–192. MR 1859020 (2004f:30032)
44. G. Perelman, *The entropy formula for the ricci flow and its geometric applications*, arXiv:math.GT/0405568.
45. ———, *Finite extinction time for the solutions to the ricci flow on certain three-manifolds*, arXiv:math.GT/0405568.
46. ———, *Ricci flow with surgery on three-manifolds*, arXiv:math.GT/0405568.
47. Piotr Przytycki and Alessandro Sisto, *A note on acylindrical hyperbolicity of mapping class groups*, Hyperbolic geometry and geometric group theory **73** (2017), 255–264.
48. Kasra Rafi, *A combinatorial model for the teichmüller metric*, GAFA Geometric And Functional Analysis **17** (2007), 936–959.

49. Richard Evan Schwartz, *Mostly surfaces*, vol. 60, American Mathematical Soc., 2011.
50. G. P. Scott, *Compact submanifolds of* 3*-manifolds*, J. London Math. Soc. (2) **7** (1973), 246–250. MR MR0326737 (48 #5080)
51. GP Scott, *The space of homeomorphisms of a 2-manifold*, Topology **9** (1970), no. 1, 97–109.
52. Caroline Series, *Continued fractions and hyperbolic geometry*, Loughborough LMS Summer School (2015).
53. Alessandro Sisto, *What is a hierarchically hyperbolic space?*, arXiv preprint arXiv:1707.00053 (2017).
54. Jing Tao, *Linearly bounded conjugator property for mapping class groups*, Geometric and Functional Analysis **23** (2013), no. 1, 415–466.
55. Dylan Paul Thurston, *Positive basis for surface skein algebras*, Proceedings of the National Academy of Sciences **111** (2014), no. 27, 9725–9732.
56. William P. Thurston, *Hyperbolic structures on* 3*-manifolds. II. Surface groups and* 3*-manifolds fibering over the circle*, Preprint (1998).
57. Jacques Tits, *Free subgroups in linear groups*, Journal of algebra **20** (1972), no. 2, 250–270.
58. William A Veech, *Teichmüller curves in moduli space, eisenstein series and an application to triangular billiards*, Inventiones mathematicae **97** (1989), 553–583.
59. Bronislaw Wajnryb, *Mapping class group of a surface is generated by two elements*, Topology **35** (1996), no. 2, 377–383.
60. Bronisław Wajnryb, *An elementary approach to the mapping class group of a surface*, Geometry & Topology **3** (1999), no. 1, 405–466.
61. Tadayuki Watanabe, *Some exotic nontrivial elements of the rational homotopy groups of Diff(S4)*, arXiv preprint arXiv:1812.02448 (2018).
62. Yohsuke Watanabe, *Intersection numbers in the curve complex via subsurface projections*, Journal of Topology and Analysis **9** (2017), no. 03, 419–439.
63. Daniel T Wise, *An invitation to coherent groups*, What?s next??the mathematical legacy of William P. Thurston **205** (2020), 326–414.
64. Tatsuhiko Yagasaki, *Homeomorphism groups of 2-manifolds and spaces of embeddings into 2-manifolds (unsolved problems and its progress in general?geometric topology)*, ?????????? **1107** (1999), 46–56.
65. ______, *Homotopy types of homeomorphism groups of noncompact 2-manifolds*, Topology and its Applications **108** (2000), no. 2, 123–136.