CHAPTER 3

Representations of 3-manifold groups

Peter B. Shalen

University of Illinois at Chicago

Contents

April 23, 1999 – Draft version; Typeset by LaTeX

Supported by an NSF grant and a U. of I. University Scholar award

Handbook of Geometric Topology – Style files
Created by F.D. Mesman and W.j. Maas

The phrase "representations of 3-manifold groups" is likely to suggest different things to different people. When I was asked to write a chapter for this handbook on this subject, I thought it would be wise to focus on something I knew about, and I therefore decided to concentrate on the interaction between two kinds of representations of fundamental groups of 3-manifolds: representations by 2×2 matrices, and representations by automorphisms of trees. Representations of the first kind are related to hyperbolic structures on 3-manifolds, while those of the second kind are related to surfaces in 3-manifolds. The interaction between these two kinds of representations therefore provides a link between what are probably the two most useful kinds of structures on 3-manifolds, whence the utility of this theory in studying group actions, Dehn surgery, surfaces in knot exteriors, and degeneration of hyperbolic structures.

The core of this chapter is an attempt to present—with the necessary background—some of the content of a series of joint papers [17], [18], [19], —by Marc Culler and myself, and of the first chapter of our paper [15] with Cameron Gordon and John Luecke. I have been able to give only the briefest hints about the remarkable developments in the area that have been made by Steven Boyer and Xingru Zhang in [5], [6] and [7], by Daryl Cooper and Darren Long and their co-authors in [11], [13], and other papers—see [14] for a survey—and by Nathan Dunfield in [22] and [23]. I have been even briefer about my closely related joint work with John Morgan in [42], [43] and [44], and the subsequent breakthrough by Rips and the developments in geometry group theory that it has led to. To do justice to this material would have made the chapter twice as long, but of course the point of this enterprise is to try to inspire you to read all these papers. By and large the material that I have discussed in detail is prerequisite for the material that I have touched on lightly, and naturally it's material that I know well.

The chapter is meant to be organized rather as if it were the notes for a course, and I've tried to keep the informal tone of a course or lecture series, as I did in my survey articles [57] and [58]. As the main topic of the chapter involves tying together ideas from several areas of mathematics other than topology, I have tried to provide reasonable introductions to the relevant ideas from these other areas. In some case this has meant not just introducing statements or even just proofs of relevant theorems, but trying to provide some real context by showing how these ideas are used within the areas from which they're borrowed. That's why you'll find little sections about topics like Ihara's theorem on discrete subgroups of $\mathrm{SL}(2, \mathbf{Q}_p)$ and Lagrange's theorem that every positive integer is a sum of four squares.

The different sections are meant to be read in order, but in case you don't like being kept in suspense, I'll give a very quick outline of what's going to happen. (Of course I'll have to use some terms that may not mean much until you've read the relevant sections, but sometimes just seeing key words can be of value.) In Section 1 I'll present some generalities about group actions, particularly actions of fundamental groups, and I show how surfaces and hyperbolic structures lead naturally to actions on trees and representations in $\mathrm{SL}(2, \mathbf{C})$. In Section 2 I'll present a construction that goes the opposite way—starting with an action of π_1 of a 3-manifold on a tree, you can get a surface in the manifold. This idea is basically due to Stallings, and has lots of direct applications to 3-manifold theory, of which I'll do some samples.

Section 3 will be basically a mini-course on the tree for SL_2, which is a special case of the Bruhat-Tits building of an algebraic group that was given an elegant self-contained treatment by Serre. Like Serre, in his longer course [55], I'll assume only the most elementary algebra. This section provides the germ of the connection between matrix representations and representations by tree automorphisms. Before Culler and I wrote [17], this material had already been applied to 3-manifold theory via Bass's GL_2 subgroup theorem [3]. The applications in [17] and the papers depending on it differ from this first application in that they involve considering entire families of representations, which form algebraic varieties. In Section 4 I'll do a little rudimentary algebraic geometry, referring you to standard texts for some of the harder results, and I'll introduce varieties of representations of groups, and the varieties of characters that are closely related to them. I'll say a bit about what these things look like for the case of fundamental groups of hyperbolic 3-manifolds.

In Section 5 I'll present the basic theory that Culler and I developed in [17], tying together the material from Sections 3 and 4. In the last subsection of Section 5, and in Section 6, I'll do a first application to topology—the existence of an essential separating surface in the complement of a nontrivial knot, first conjectured by Neuwirth. A second application, in Section 7, is a proof of the Smith Conjecture about periodic tame homeomorphisms of S^3 with 1-dimensional fixed point set. (I'll talk a bit about the history of this when the time comes.) In Chapter 8 I'll introduce some machinery that's needed for the applications to Dehn surgery that I'll talk about in Chapter 9, and for studying more refined questions related to the Neuwirth Conjecture in Section 10. In Section 11 I'll give a hint about how the techniques are related to geometric questions about degenerations of hyperbolic structures.

I'm very grateful to Marc Culler for helping talk me through some difficult spots in the writing. (Of course, if Marc hadn't done all this joint work with me I wouldn't have anything to write about.) I'm very grateful to Benson Farb and Bob Daverman for reading the entire manuscript and making a huge number of helpful comments. I'd also like to thank Jeremy Teitelbaum for trying to make sure I didn't say anything too silly in the passages about p-adic numbers.

Although this chapter is something very different from what I had in mind when I sat down to write it, I hope it may prove useful.

1. Some basic concepts and examples

1.1. Representations and actions

Recall that an *action* of a group Γ on a set X is a function $\cdot : \Gamma \times X \to X$ that satisfies the identities $1 \cdot x = x$ and $(\gamma\delta) \cdot x = \gamma \cdot (\delta \cdot x)$. There is a lot of structure associated in an elementary way with an action, such as the partition of X into *orbits:* two elements x and y are in the same orbit if and only if $\gamma \cdot x = y$ for some

$\gamma \in \Gamma$. The *stabilizer* of an element $x \in X$, often denoted Γ_x, is the subgroup of Γ consisting of all $\gamma \in \Gamma$ such that $\gamma \cdot x = x$. For any $x \in X$, the map $\Gamma \to X$ defined by $\gamma \to \gamma \cdot x$ induces a bijection between the set of cosets of the form $\gamma\Gamma_x$, for $\gamma \in \Gamma$, and the orbit $\Gamma \cdot x$ of x. A set $S \subset X$ is *invariant* under the action if it is a union of orbits, i.e. if $\gamma \cdot x \in S$ whenever $x \in S$ and $\gamma \in \Gamma$. An element of X is said to be *fixed* by Γ if $\{x\}$ is invariant, i.e. is an entire orbit. The action is *free* if the stabilizer of every element of X is the trivial subgroup of Γ.

There's a natural bijective correspondence between actions of Γ on X and *representations* of Γ in the symmetric group $\mathcal{S}(X)$, i.e. homomorphisms $\rho : \Gamma \to \mathcal{S}(X)$. If $\cdot$ is an action, the corresponding representation sends γ to the element $s \mapsto \gamma \cdot s$ of $\mathcal{S}(X)$. I'll generally be talking about sets X that have some extra structure, and focusing on actions which preserve this structure in the sense that the corresponding representations take values in the automorphism group $\mathrm{Aut}(X)$. Thus X may be a vector space over some field K, in which case $\mathrm{Aut}(X)$ is the group of linear automorphisms; in terms of the action this means that the identities $\gamma\cdot(x+y) = \gamma\cdot x+\gamma\cdot y$ and $\gamma \cdot (ax) = a\gamma \cdot x$ hold when $x, y \in X$ and $a \in K$. So we can talk about linear actions and linear representations, and the difference between the two is purely notational. Similarly we can talk about topological actions, or actions by homeomorphisms, of Γ on a topological space X; simplicial actions on a simplicial complex; and so forth.

A representation $\Gamma \to \mathrm{Aut}(X)$ is termed *faithful* if it is an injective homomorphism. An action is said to be *effective* if it corresponds to a faithful representation. Thus Γ acts effectively on X if and only if for every $\gamma \in \Gamma - \{1\}$ there is an $x \in X$ with $\gamma \cdot x \neq x$.

Suppose that we are given actions of a group Γ on two sets X and Y. A map of sets $f : X \to Y$ is said to be Γ-*equivariant* if we have $f(\gamma \cdot x) = \gamma \cdot f(y)$ for every $x \in X$. For any group Γ, there is a category in which the objects are sets equipped with actions of Γ and the morphisms are Γ-equivariant maps. By giving the sets, actions and maps extra structure, one can define many natural subcategories. For example, one can require the sets to be vector spaces, or topological spaces, or simplicial complexes, and require both actions and maps to be linear, or continuous, or simplicial. There are natural terms to designate isomorphisms in these categories: Γ-equivariant linear isomorphisms, Γ-equivariant homeomorphisms, Γ-equivariant simplicial isomorphisms and so on.

Of course, we think of two actions as being "the same" if we have an isomorphism (in the appropriate category) between the sets in question which is equivariant with respect to the given actions. I'll express this by saying that the actions—or the corresponding representations—are *equivalent.* (This is the classical term in the case of linear representations). In more direct terms, $\rho : \Gamma \to \mathrm{Aut}(X)$ and $\rho' : \Gamma \to \mathrm{Aut}(Y)$ are equivalent if and only if there is an isomorphism $\phi : X \to Y$ such that $\phi \circ \rho = \rho' \circ \phi$.

By an *invariant* of a representation one means any sort of datum associated with the representation which depends only on its equivalence class. I'll illustrate the idea by talking briefly about invariants of linear representations, which, besides being a very classical thing to look at, will be especially important in this chapter.

The most obvious invariant of a linear representation $\rho : \Gamma \to \mathrm{Aut}(V)$, where V

is a vector space over some given field, is the dimension of V, which is often called the dimension of the representation. It's just about obvious that any n-dimensional representation is equivalent to a representation in $\mathrm{Aut}(\mathbf{C}^n) = \mathrm{GL}_n(\mathbf{C})$, and that two representations $\rho, \rho' : \Gamma \to \mathrm{GL}_n(\mathbf{C})$ are equivalent if and only if $\rho' = i_A \circ \rho$ for some matrix $A \in \mathrm{GL}_n(\mathbf{C})$; here I am denoting by i_A the inner automorphism $X \mapsto AXA^{-1}$ of $\mathrm{GL}_n(\mathbf{C})$. In particular, the property of being *unimodular,* i.e. of sending the entire group Γ into $\mathrm{SL}_n(\mathbf{C})$, depends only on the equivalence class of a representation $\rho : \Gamma \to \mathrm{GL}_n(\mathbf{C})$.

Let me now specialize to the case of unimodular representations of dimension 2; this is the case that will be important in this chapter, and, conveniently, the case that I am competent to discuss. By far the most important invariant of such a representation is its *character.* The character of $\rho : \Gamma \to \mathrm{GL}_n(\mathbf{C})$ is the function $\chi : \Gamma \to \mathbf{C}$ defined by $\chi(\gamma) = \mathrm{trace}\,\rho(\gamma)$. It is very nearly true that the character is a *complete* invariant, which would mean that 2-dimensional unimodular representations with the same character were always equivalent. To see that this is not quite true, note, for example, any homomorphism of Γ into the group of matrices of the form $\begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ has the same character as the *trivial representation* that sends the entire group to the identity matrix. It turns out that the only bad examples of this sort involve *reducible* representations. A representation of Γ in $\mathrm{SL}(2, \mathbf{C})$ is said to be reducible if it is equivalent to a representation by upper triangular matrices; in terms of the corresponding action of Γ on $\mathbf{C}^2$, this means that some 1-dimensional subspace is invariant under the action. You will find a proof of the following elementary result in [17]:

Proposition 1.1.1. *Let Γ be any finitely generated group. If two unimodular representations ρ and ρ' of Γ in $\mathrm{SL}(2, \mathbf{C})$ have the same character, then either ρ and ρ' are equivalent or they are both reducible.*

There is presumably a similar result about $\mathrm{GL}_n(\mathbf{C})$, but one probably has to be more careful about the statement.

1.2. A word about base points

This chapter is about representations (or actions) of fundamental groups of connected 3-manifolds. (Incidentally, when I say "manifold" I always mean "manifold with (possibly empty) boundary." This seems to be the modern convention among topologists. A manifold is *closed* if it is compact and has empty boundary.)

In dealing with fundamental groups it is always necessary to decide what to do about base points. In many situations I will be talking about *equivalence classes* of representations (or actions) of the fundamental group of a connected manifold. The point I would like to make here is that in these situations, there is a strong sense in which the choice of a base point is irrelevant. To see why this is so, let's consider two points x and w in the connected manifold M. Any path α from x to w defines an isomorphism $I_\alpha : \pi_1(X, w) \to \pi_1(X, x)$: for any loop γ based at w we have $I_\alpha([\gamma]) = [\alpha * \gamma * \bar{\alpha}]$. If α and α' are two paths from x to w, the composition $I_{\alpha'} \circ I_\alpha^{-1}$ is the inner automorphism $i_{[\alpha' * \alpha]}$ of $\pi_1(X, x)$. So although the isomorphism between $\pi_1(X, w)$ and $\pi_1(X, x)$ is not canonical, it is canonical modulo composition with inner automorphisms.

Now if we are given an action of $\pi_1(M, x)$ on a set X, and if $\rho : \pi_1(M, x) \rightarrow \mathcal{S}(X)$ is the corresponding representation, then for any path α from x to w we have a representation $\rho \circ I_\alpha : \pi_1(X, w) \rightarrow \mathcal{S}(X)$. If in place of α we consider another path α', we have $I_{\alpha'} = i_g I_\alpha$ for some $g \in \pi_1(M)$, and hence $\rho \circ I_{\alpha'} = i_{\rho(g)} \circ \rho \circ I_\alpha$. Since $\rho \circ I_{\alpha'}$ and $\rho \circ I_\alpha$ differ by post-composition with an inner automorphism, they are equivalent representations. So a representation of $\pi_1(M, x)$ defines an equivalence class of representations of $\pi_1(M, w)$. It's equally easy to see that this equivalence class depends only on the equivalence class of the given representation of $\pi_1(M, w)$, and that we get in this way an absolutely canonical bijection between equivalence classes of representations of $\pi_1(M, x)$ and equivalence classes of representations of $\pi_1(M, w)$. So one can talk without ambiguity about *equivalence classes of representations or actions of* $\pi_1(M)$, without specifying a base point.

There will be various other kinds of situations in this chapter where base points will be suppressed for a very similar reason. I will give a hint here and there to remind you of what is going on.

1.3. The universal covering

If M is a connected manifold of any dimension n, and x is a base point in M, the most basic example of an action of $\pi_1(M, x)$ is the usual action by homeomorphisms on the universal covering space $\tilde{M}$. The uniqueness theorem for the universal covering tells us that this action is well-defined up to equivalence. (Here the underlying category is that of topological spaces. Thus two actions on spaces are equivalent if and only if there is an equivariant homeomorphism between the spaces.) Following the convention I've just explained, I will say—without mentioning a base point—that the action of $\pi_1(M)$ on $\tilde{M}$ is canonically defined up to equivalence. If M is given a triangulation, then $\tilde{M}$ inherits a triangulation, and we can then interpret the action of $\pi_1(M)$ as a simplicial action, defined up to equivalence in the simplicial category.

Giving other kinds of structure in the connected manifold M often leads to new actions of $\pi_1(M)$ that are induced by its action on $\tilde{M}$.

1.4. The tree associated with a hypersurface

As one nice example, suppose that we are given a hypersurface in M, i.e. a codimension-1 submanifold F of M, not necessarily connected. In this section I will be assuming that F admits a *bicollaring,* i.e. a homeomorphism h of $F \times [-1, 1]$

onto a neighborhood of F in M such that $F(x,0) = x$ for every $x \in F$ and $h(F \times [-1,1]) \cap \partial M = h(\partial F \times [-1,1])$. We can use a bicollaring h to define a partition of the space M into disjoint subsets. The subsets are of two types: the components of $M - h(F \times (-1,1))$, and the sets of the form $F_i \times \{t\}$, where F_i is a component of F and $t \in (0,1)$. We can regard the sets in this partition as forming a topological space with the quotient topology, and you will see easily that it is a graph, i.e. a 1-dimensional CW complex, which has one edge for every component of F and one vertex for every component of $M - F$. Note that h defines an identification of each edge of $\mathcal{G}$ with the interval $[-1,1]$.

By construction there is a natural map $r : M \to \mathcal{G}$, but it is also very easy to construct a map $i : \mathcal{G} \to M$ such that the composition $r \circ i$ maps each edge and each vertex of $\mathcal{G}$ into itself. In particular $r \circ i$ is homotopic to the identity map of $\mathcal{G}$, from which it follows that $\mathcal{G}$ is connected and that $\pi_1(\mathcal{G})$ is isomorphic to a retract (hence both a subgroup and a quotient) of $\pi_1(M)$. The graph $\mathcal{G}$ is often called the *dual graph* of F in M. (Since regular neighborhood theory tells us that the bicollaring h is unique up to ambient isotopy, the graph is well-defined up to simplicial isomorphism, and even the map r is well-defined in a sense that's easy to work out.)

Now consider the universal covering $(\tilde{M}, p)$ of M. Given the bicollaring h of F, it's a simple exercise in covering space theory to show that $\tilde{F} = p^{-1}(F)$ has a unique bicollaring $\tilde{h}$ in M such that $p(\tilde{h}(x,t)) = h(p(x),t)$ for all $x \in \tilde{F}, t \in [0,1]$. Let T denote the dual graph of $\tilde{F}$ in $\tilde{M}$ defined in terms of this induced bicollaring $\tilde{h}$. Then T is simply connected since $\pi_1(T)$ is a retract of $\pi_1(\tilde{M}) = \{1\}$; that is, T is a *tree.* Now the sets that make up the partition defining T are the components of the sets of the form $p^{-1}(A)$, where A ranges over the sets in the partition defining the dual graph $\mathcal{G}$. So the partition defining T is invariant under the action of $\pi_1(M)$, in the sense that each element of $\pi_1(M)$ maps each set in the partition onto a possibly different set in the partition. Hence the action of $\pi_1(M)$ on $\tilde{M}$ induces an action on T. In fact, this induced action is the unique action that makes the quotient map $\tilde{M} \to T$ equivariant.

Since T is the dual graph of $\mathcal{G}$ defined by the bicollaring $\tilde{h}$, each closed edge e of T comes equipped with an identification with $[-1,1]$, i.e. a homeomorphism $\eta_e : [-1,1] \to e$. Since T is simply connected, and therefore has no multiple edges, the identification of the edges of T with linear intervals give T the structure of a simplicial complex. Because of the precise way in which $\tilde{h}$ is induced from the bicollaring h of M, it's easy to check that for each element $\gamma \in \pi_1(M)$ and each edge e of T, and each $t \in [-1,1]$, we have $\gamma \cdot \eta_e(t) = \eta_{\gamma \cdot e}(t)$. This implies that the action of $\pi_1(M)$ on T is simplicial, but it also shows a little more—namely, that if an element $\gamma \in \pi_1(M)$ leaves an edge e of T invariant, then it actually fixes the edge pointwise. This property is expressed by saying that $\pi_1(M)$ acts on T *without inversions.* A simplicial automorphism of a tree is called an *inversion* if it leaves some edge invariant but interchanges its endpoints.

Because the bicollaring h is unique up to ambient isotopy, the tree T is well-defined up to simplicial equivalence once the hypersurface F is given.

It's easy to describe the stabilizers of the vertices and edges of T under the action of $\pi_1(M)$. Each vertex s of T corresponds to a component $\tilde{K}$ of $\tilde{M} - \tilde{h}((-1,1))$, and it follows from the construction that the stabilizer $\pi_1(M)_s$ of s coincides with the stabilizer $\pi_1(M)_{\tilde{K}}$ of $\tilde{K}$. But $\tilde{K}$ is a component of $p^{-1}(K_j)$ for some component K_j of $M - h((-1,1))$, and from covering space theory one knows that the stabilizers of the various components of $p^{-1}(K_j)$ are precisely the conjugates of $\mathrm{im}(\pi_1(K_j) \to \pi_1(M))$ in $\pi_1(M)$. (It follows from the kind of thing I talked about in Subsection 1.2 that the subgroup $\mathrm{im}(\pi_1(K_j) \to \pi_1(M))$ of $\pi_1(M)$ is well-defined up to conjugacy, so it makes sense to talk about "conjugates of $\mathrm{im}(\pi_1(K_j) \to \pi_1(M))$ in $\pi_1(M)$.") Furthermore, we have $\mathrm{im}(\pi_1(K_j) \to \pi_1(M)) = \mathrm{im}(\pi_1(C_j) \to \pi_1(M))$, where C_j is the component of $M - F$ containing K_j. So we see that the stabilizers of the vertices of T are precisely the conjugates of the subgroups $\mathrm{im}(\pi_1(C_j) \to \pi_1(M))$ in $\pi_1(M)$, where C_j ranges over the components of $M - F$. Similarly, the stabilizers of the edges of T are precisely the conjugates of the subgroups $\mathrm{im}(\pi_1(F_i) \to \pi_1(M))$ in $\pi_1(M)$, where F_i ranges over the components of F.

The picture of the action is especially nice in the case where the inclusion homomorphism $\pi_1(F_i) \to \pi_1(M)$ is injective for every component F_i of F. For one thing, the stabilizer of an edge of T is then actually isomorphic to F_i, the isomorphism being canonical up to composition with an inner automorphism. We have a similarly nice situation with regard to the vertex stabilizers. This is because the injectivity of the inclusion homomorphisms $\pi_1(F_i) \to \pi_1(M)$ implies that for every component C_j of $M - F$, the inclusion homomorphism $\pi_1(C_j) \to \pi_1(M)$ is injective. You can see this very neatly in terms of the above discussion of $\tilde{M}$: the injectivity of the homomorphisms $\pi_1(C_j) \to \pi_1(M)$ is equivalent to the assertion that the components of $\tilde{M} - \tilde{F}$ are simply connected. But the injectivity of the $\pi_1(F_i) \to \pi_1(M)$ implies that the components of $\tilde{F}$ are simply connected, and since $\tilde{M}$ is simply connected as well, the simple connectivity of the components of $\tilde{M} - \tilde{F}$ follows from van Kampen's theorem. Now, from the injectivity of the $\pi_1(C_j) \to \pi_1(M)$ and the general discussion above, we see that the stabilizers in $\pi_1(M)$ of the vertices of T are isomorphic to the groups $\pi_1(C_j)$ for components C_j of $M - F$. Again the isomorphisms are canonical up to composition with inner automorphisms.

Having shown how a bicollared hypersurface gives rise to an action on a tree, I should say a word about why the condition that a hypersurface be bicollared is a natural one. The obvious necessary conditions for a hypersurface F of M to be bicollared are that F be *properly embedded* and *two-sided*. To say that F is properly embedded in M means that F is a closed subset of M and that $F \cap \partial M = \partial F$. To say that a properly embedded hypersurface F is two-sided (or "locally separating") means that the complement of F relative to any sufficiently small neighborhood of F is disconnected. (For simplicity, or out of habit, I will be talking mostly about orientable manifolds in this chapter. If M is orientable then a properly embedded hypersurface $F \subset M$ is two-sided if and only if it is orientable—which means, if you like, that each component is orientable.)

Conversely, any halfway-reasonable properly embedded orientable hypersurface in M is bicollared: for example, if M comes with a smooth or piecewise-linear structure then any smooth or PL hypersurface in M which is orientable or properly embedded is bicollared. (If you don't make some assumption about the surface then you can run into weird examples like the Alexander horned sphere [39].) For technical reasons it is not always convenient to be working with a smooth or PL structure, so I'll often just make it a hypothesis that the surfaces I talk about are bicollared.

1.5. The three-dimensional case: essential surfaces

The construction I described in the last subsection is especially useful in the case $n = 3$. Before making a few simple comments about this case I need to make some basic definitions and remarks about surfaces in 3-manifolds which are important for everything I'll be talking about in this chapter.

A 3-manifold M is said to be *irreducible* if M is connected and every bicollared 2-sphere in M is the boundary of a 3-ball contained in M. A bicollared surface $F \subset M$ is said to be *boundary-parallel* if F is the frontier of a set $P \subset M$ such that the pair (P, F) is homeomorphic to $(F \times [0,1], F \times \{1\})$. (The *frontier* of a set is the intersection of its closure with the closure of its complement; I am reserving the term *boundary* for the intrinsic sense, as in "manifold with boundary." Note that in the definition I've just given, the homeomorphism of P onto $F \times [0,1]$ has to map $P \cap \partial M$ onto $(\partial F \times [0,1] \cup (F \times \{0\}).)$

Definition 1.5.1. For most of this chapter I'll be using the following definition. A surface F in a compact, irreducible, orientable 3-manifold is said to be *essential* if it has the following properties:

(i) F is bicollared;

(ii) the inclusion homomorphism $\pi_1(F_i) \to \pi_1(M)$ is injective for every component F_i of F;

(iii) no component of F is a 2-sphere;

(iv) no component of F is boundary-parallel; and

(v) F is nonempty.

Condition 1.5.1(ii) has a beautiful geometric interpretation. One situation in which the condition obviously *fails* to hold is the one in which there is a disk $D \subset M$ such that (a) $D \cap F = \partial D$, and (b) the simple closed ∂D is homotopically nontrivial in F—which by elementary surface topology (see [26], Theorem I.7) is the same as saying that it doesn't bound a disk in F. In this case, ∂D clearly defines (up to conjugation and inversion) a nontrivial element of $\ker(\pi_1(F_i) \to \pi_1(M))$, where F_i is the component of F containing ∂D. Now, conversely, it is a fundamental principle in 3-manifold theory ([31], proof of Lemma 6.1), which is easily deduced from two results due to Papakyriakopoulos, the Dehn Lemma and the Loop Theorem, that if a bicollared surface F fails to satisfy Condition 1.5.1(ii), then there is a disk D satisfying (a) and (b).

Property (a) of the disk D says it can be thought of as a properly embedded disk in the manifold M' obtained by splitting M along F. The proof of Papakyriaokoulos's theorem is usually done in the PL category, and gives the additional conclusion that (c) D is bicollared in M'. A disk satisfying (a), (b) and (c) is often called a *compressing disk* for F. Thus, for a bicollared surface F, Condition 1.5.1(ii) is equivalent to the condition that F there is no compressing disk for F.

The properties that I have included in the definition of an "essential surface" are very similar to those that people typically include in the definition of an "incompressible surface." However, Haken's term "incompressible" has been used in so many ways in recent years that there are now almost as many definitions as there are 3-manifold topologists, and—to make matters worse—people get emotional about the issue of what the term should mean. That's why I am avoiding it in this chapter.

According to the discussion at the end of Subsection 1.4, Condition 1.5.1(i) implies that the stabilizer of each edge of T is isomorphic to the fundamental group of some component F_i of F, and that the stabilizer in $\pi_1(M)$ of each vertex of T is isomorphic to the fundamental group of some components C_j of $M - F$; and that these isomorphisms are canonical up to composition with inner automorphisms.

Conditions 1.5.1(ii)—(iv) in the definition of an essential surface also give nice information about the action associated to the surface. A (simplicial) action (without inversions) of a group Γ on a tree T is said to be *trivial* if there is a vertex of T which is fixed by the entire group Γ.

Proposition 1.5.2. *Let F be an essential surface in a compact, connected, orientable, irreducible 3-manifold M. Then the action on a tree associated to F is nontrivial.*

Proof. I'll call the tree T. Assume that the action is trivial, so that the stabilizer of some vertex of T is all of $\pi_1(M)$. This translates into saying that for some component C of $M - F$, the injection $\pi_1(C) \to \pi_1(M)$ is an isomorphism (as in "isomorphism onto").

By Condition 1.5.1(iv) we have $F \neq \emptyset$. Let F_0 be a component of F. Since $F_0 \cap C = \emptyset$, there is some component C_0 of $M - F_0$ such that the inclusion homomorphism $\pi_1(C_0) \to \pi_1(M)$ is surjective. This gives a contradiction right off the bat if F_0 doesn't separate M, i.e. if C_0 is the only component of $M - F_0$, because then even $H_1(M - F_0) \to H_1(M)$ fails to be surjective. Suppose now that $M - F_0$ has a second component C_1. Consider the dual tree T' to the connected essential surface F_0. If e is any edge of T', one endpoint s_0 of e, corresponding to a component of $p^{-1}(C_0)$, is stabilized by all of $\pi_1(M)$. The other endpoint s_1 of e has stabilizer $\pi_1(M)_{s_1} = \pi_1(M)_{s_1} \cap \pi_1(M)_{s_0} = \pi_1(M)_e$. This means that the inclusion homomorphism $\pi_1(F) \to \pi_1(\overline{C_1})$ is an isomorphism.

We also know that the $\overline{C_i}$ are irreducible; otherwise, since M is irreducible, F_0 would be contained in a ball, and would not be essential. Now the main point is to apply a theorem due to Stallings, for which the best reference is Brown and Crowell's paper [8] in which a stronger theorem is proved: Stallings's theorem says that if K is a compact, connected, orientable, irreducible 3-manifold and $F \subset K$ is a connected 2-manifold such that $\pi_1(F) \to \pi_1(K)$ is an isomorphism, then either K is a ball and $F = \partial K$, or the pair (K, F) is homeomorphic to $(F \times [0,1], F \times \{1\})$.

We can apply this with F_0 and $\overline{C_1}$ in place of F and K, and we get a contradiction to either 1.5.1(ii) or 1.5.1(iii).

$\square$

So an essential surface in M gives rise to a nontrivial action (defined up to equivalence) of $\pi_1(M)$ on a tree. In Section 2 we'll see how to go the other way—to start with an action of $\pi_1(M)$ on a tree and use it to construct, in a noncanonical but ultimately very useful way, an essential surface in M.

1.6. The action associated to a hyperbolic structure

Another kind of structure in a connected manifold M (of dimension $n \geqslant 2$) that leads to an action of the fundamental group is a (complete) hyperbolic structure. I will refer you to Bonahon's chapter in this volume for an introduction to hyperbolic geometry, which plays an important role in most of the topics I will be covering in this chapter. If we are given a hyperbolic structure on M, we can identify the universal covering of $\tilde{M}$ with the hyperbolic space $\mathbf{H}^n$ by some isometry. The natural action of $\pi_1(M)$ then becomes an action by isometries on $\mathbf{H}^n$—or, in slightly different language, a representation ρ_0 of $\pi_1(M)$ in the isometry group $\mathrm{Isom}(\mathbf{H}^n)$. The representation that we get in this way is well-defined up to equivalence once we have specified a hyperbolic structure on M. The representation is readily seen to be faithful, and to be *discrete* in the sense that $\rho_0(\pi_1(M))$ is a discrete subgroup of the topological group $\mathrm{Isom}(\mathbf{H}^n)$. If M is orientable then ρ_0 takes values in the group $\mathrm{Isom}^+(\mathbf{H}^n)$ of orientation-preserving isometries, but it is still well-defined only up to conjugation in $\mathrm{Isom}(\mathbf{H}^n)$. Thus in general there may be two inequivalent representations in $\mathrm{Isom}^+(\mathbf{H}^n)$ associated to a given hyperbolic structure on M, and they will differ by conjugation by an orientation-reversing involution $J \in \mathrm{Isom}^+(\mathbf{H}^n)$, a reflection about a hyperplane.

In the three-dimensional case, $\mathrm{Isom}(\mathbf{H}^3)$ may be identified isomorphically with $\mathrm{PSL}(2,\mathbf{C})$. The identification is canonical modulo inner automorphisms. An element of $\mathrm{PSL}(2,\mathbf{C})$ is a coset modulo $\pm I$ of a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. I'll denote this coset by $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$. We may choose the orientation-reversing involution J so that the group-theoretical conjugation $A \to A^J$ is realized by complex conjugation of matrices:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \longmapsto \begin{bmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{bmatrix}.$$

Thus the two equivalence classes of representations in $\mathrm{PSL}(2,\mathbf{C})$ associated with a given hyperbolic structure differ from each other by a complex conjugation.

Sometimes it's more convenient to work with representations in $\mathrm{SL}(2,\mathbf{C})$ instead of $\mathrm{PSL}(2,\mathbf{C})$. For this purpose, it is useful to have the following result of Thurston's:

Proposition 1.6.1. *Let M be a (connected) orientable hyperbolic 3-manifold, and let $\rho_0 : \pi_1(M) \to \mathrm{PSL}(2, \mathbf{C})$ be a representation associated to the hyperbolic structure. (So ρ_0 belongs to one of the two equivalence classes discussed above.) Then there is a lift of ρ_0 to $\mathrm{SL}(2, \mathbf{C})$, i.e. a representation $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ such that $p\tilde{\rho}_0 = \rho_0$, where $p : \mathrm{SL}(2, \mathbf{C}) \to \mathrm{PSL}(2, \mathbf{C})$ is the quotient projection.*

The representation $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ is even less canonical than the representation ρ_0. Whereas there are in general two choices for ρ_0 in terms of a given hyperbolic structure on M, it is a simple exercise, given Proposition 1.6.1, to show that when $\pi_1(M)$ is finitely generated, the number of lifts of a given ρ_0 is $|H_1(M; \mathbf{Z}_2)|$. However, a lift of ρ_0, being a linear representation, is a pretty down-to-earth kind of object, and one from which the hyperbolic structure itself can be recovered. For these reasons, having Proposition 1.6.1 is often very convenient in applications, as we shall see.

2. Actions of 3-manifold groups on trees

The ideas in this section are mostly due to Stallings, who developed them in a series of papers beginning with [61], and presented some of them in his book [62]. I'll be presenting them from a point of view which is fairly close to the one that was used in [44] and, a little later, in [15]. This point of view is influenced by Serre's book [55].

In this section, M will denote a compact, orientable, irreducible 3-manifold. In Subsection 1.4 I described how an essential surface in M gives rise to a nontrivial (simplicial) action, without inversions, of $\pi_1(M)$ on a tree. The construction of the action from the surface is canonical up to equivalence. It is far from being true that every nontrivial action without inversions of $\pi_1(M)$ on a tree arises from this construction. Indeed, I pointed out that under the action associated with an essential surface F, the stabilizer of each edge or vertex of the tree is isomorphic to the fundamental group of a component of F or of $M - F$. By contrast, I will point out in Subsection 2.3 below that for many reasonable choices of M there are very simple and natural examples of nontrivial actions without inversions of $\pi_1(M)$ on trees for which the edge and vertex stabilizers are not even finitely generated!

Nevertheless, it turns out that with every nontrivial action without inversions of $\pi_1(M)$ on a tree one can "associate," in an interesting way, an essential surface in M. I've used quotation marks here because the construction of the surface from the action is far from being canonical, as it depends on many choices. Furthermore, one cannot in general reconstruct an action from a surface "associated" to it. Nevertheless, such a surface contains important information about the action. In this section I will describe the construction of a surface associated to an action, talk a little about the extent to which this construction behaves like an inverse to the construction of 1.4, and give some important applications. These are only first applications, though, because everything I will be talking about in the rest of the chapter depends on the construction I'll describe here.

2.1. Constructing an equivariant map

Suppose, then, that we're given a simplicial action $\cdot$ of $\pi_1(M)$ on a tree T. (Eventually it will matter that the action is nontrivial and without inversions; I'll point out where these hypotheses come up.) The first step in constructing an essential surface associated with the action is to construct a (continuous) $\pi_1(M)$-equivariant map $\tilde{f} : \tilde{M} \to T$.

Let's fix a triangulation for M, and give $\tilde{M}$ the induced triangulation. The strategy for constructing $\tilde{f}$ is to construct (continuous) maps $\tilde{f}^{(i)}$ from the i-skeleta $\tilde{M}^{(i)}$ of $\tilde{M}$ to T for $i = 0, 1, 2, 3$; each $\tilde{f}^{(i)}$ will be $\pi_1(M)$-equivariant, and $\tilde{f}^{(i)}$ will extend $\tilde{f}_{i-1}$ for $i = 1, 2, 3$. Of course we'll define $\tilde{f}$ to be $\tilde{f}_3$.

To construct $\tilde{f}^{(0)}$ we first pick a *complete system of orbit representatives* for the action of $\pi_1(M)$ on $M^{(0)}$, i.e. a set $S^{(0)} \subset M^{(0)}$ such every orbit for the action of $\pi_1(M)$ on $M^{(0)}$ meets $S^{(0)}$ in exactly one point. Now, using the fact that $\pi_1(M)$ acts freely on M—and hence on $M^{(0)}$—we can see that if $h^{(0)}$ is any map whatever from $S^{(0)}$ to the vertex set $T^{(0)}$ of T then $h^{(0)}$ has one and only one extension $\tilde{f}^{(0)} : \tilde{M}^{(0)}$ which is $\pi_1(M)$-equivariant. Uniqueness is clear, since any such map must in particular satisfy

$$\tilde{f}^{(0)}(\gamma \cdot s) = \gamma \cdot h^{(0)}(s) \text{ for all } s \in S^{(0)} \text{ and } \gamma \in \pi_1(M). \tag{2.1.1}$$

To get existence we notice that (2.1.1) makes sense as a definition of $\tilde{f}^{(0)}$ because every vertex in $M^{(0)}$ can be expressed *in exactly one way* in the form $\gamma \cdot s$ with $\gamma \in \pi_1(M)$. The point here is that if $\gamma \cdot s = \gamma' \cdot s$ then s is a fixed point of $\gamma^{-1}\gamma'$, and since the action is free we must have $\gamma = \gamma'$. Now that we know the definition 2.1.1 makes sense, there is no problem checking that the map $\tilde{f}^{(0)}$ is equivariant.

It's significant that we could have started with *any* map $h : S^{(0)} \to T^{(0)}$ for this step. (Actually at this stage we don't even need h to map $S^{(0)}$ into $T^{(0)}$, but that will be nice to know later.) This illustrates how far our construction is from being canonical. The flexibility in the definition of $\tilde{f}$ turns out to be very useful; I'll return to this issue later in this section.

Now suppose that $\tilde{f}^{(i)}$ has been constructed for a given i with $0 \leqslant i \leqslant 2$. Let us pick a complete system of orbit representatives $S^{(i+1)}$ for the action of $\pi_1(M)$ on the set of all $i+1$-simplices of $\tilde{M}$. For any simplex $\sigma \in S^{(i+1)}$ we have a map $\tilde{f}^{(i)}|\partial\sigma : \partial\sigma \to T$. Since T is contractible, this map can be extended to a map $\tilde{f}_\sigma : \sigma \to T$. Now there is a unique continuous, $\pi_1(M)$-equivariant map $\tilde{f}^{(i+1)} : M^{(i+1)} \to T$ which restricts to h_σ on each $\sigma \in S^{(i+1)}$. Indeed, such a map must obviously be given by

$$\tilde{f}^{(i+1)}(\gamma \cdot x) = \gamma \cdot h_\tau(x) \text{ for all } \tau \in S^{(0)}, x \in \tau \text{ and } \gamma \in \pi_1(M).$$

We can show, almost exactly as in the construction of $\tilde{f}^{(0)}$, that the map given by this formula is well-defined and equivariant. Continuity is then easy.

It's easy to adapt this construction so as to guarantee that the map $\tilde{f}$ is simplicial with respect to some $\pi_1(M)$-invariant triangulation of $\tilde{M}$ (and a given triangulation of T). All that we have to notice is that the in the induction step, the map $\tilde{f}^{(i)}|\partial\sigma$ can be extended to a *simplicial* map $\tilde{f}_\sigma : \sigma \to T$; this follows from the simplicial approximation theorem for pairs (see [59], Chapter 3, Section 4, Theorem 8 and Lemma 1).

2.2. Constructing a surface

Now, fixing a $\pi_1(M)$-invariant triangulation of $\tilde{M}$ in which $\tilde{f}$ is simplicial, let's consider a point x of T which is not a vertex. Consider the subset $P = \tilde{f}^{-1}(x)$ of $\tilde{M}$, and for any i-simplex σ of $\tilde{M}$ consider the subset $P \cap \sigma$ of σ. If $\tilde{f}$ does not map σ onto the edge e of T containing x then $P \cap \sigma = \emptyset$. (This is always the case if $i = 0$.) If $\tilde{f}$ does map σ onto e, then since no vertex of σ is mapped to x, the set $P \cap \sigma$ is an $(i-1)$-cell properly embedded in σ: if we think of the simplex σ as being embedded in an affine space, $P \cap \sigma$ is the intersection of σ with a hyperplane missing the vertices of σ. Now since P meets every simplex of $\tilde{M}$ either in the empty set or in a properly embedded codimension-1 cell, it's easy to see that P is a 2-manifold, properly embedded in $\tilde{M}$.

(The most interesting point is checking that P is locally Euclidean at any point z where it meets a 1-simplex τ of $\tilde{M}$. The 2-simplices incident to τ look like the pages of a cyclic book with τ as binding. The set P meets each page in a 1-cell which has an endpoint at z and is otherwise disjoint from τ. Since P meets each 3-simplex in a 2-cell, we can recover its whole intersection with the open star of τ by connecting every two successive 1-cells with a 2-cell, giving an open 2-disk.)

The basic idea for associating surfaces in M with actions of $\pi_1(M)$ on trees is now easy to explain. Let's denote by E the set of all midpoints of edges of the tree T. Since, by the discussion above, $\tilde{f}^{-1}(x)$ is a properly embedded 2-manifold in $\tilde{M}$ for each $x \in E$, the set $\tilde{F} = \tilde{f}^{-1}(E)$ is a properly embedded 2-manifold in $\tilde{M}$. On the other hand, since E is clearly $\pi_1(M)$-invariant and $\tilde{f}$ is $\pi_1(M)$-equivariant, $\tilde{F}$ is invariant under the action of $\pi_1(M)$ on $\tilde{M}$. So $\tilde{F}$ is the inverse image, under the covering projection, of some properly embedded 2-manifold in $F \subset M$.

In Subsection 2.4 I will show how to modify the map $\tilde{f}$ so that the surface $F \subset M$ that it defines is essential. Before doing this I need to deal with a small technical point, and make some definitions and remarks.

The technical point involves a slightly stronger version of the condition that $\tilde{F}$ be a 2-manifold, which makes life much simpler when we are worrying about such things as making F essential. Let $E \subset T$ be a discrete set containing no vertices of T. A continuous map $\tilde{f} : \tilde{M} \to T$ is said to be *transverse to* X if each point $z \in \tilde{f}^{-1}(X)$ has a neighborhood U with a homeomorphism $h : U \to V \times (0,1)$, for some open subset V of $\mathbf{R}^2$, such that $\tilde{f}|U = j \circ q \circ h$, where $q : (0,1) \times (0,1) \to (0,1)$ is the projection to the second factor, and j is some homeomorphism of $(0,1)$ onto an open interval in some edge of T. It's immediate that if $\tilde{f}$ is transverse to X then $\tilde{f}^{-1}(X)$ is a bicollared 2-manifold in M. On the other hand, if you examine the proof that I gave above that, for the equivariant map $\tilde{f}$ that I constructed, $\tilde{f}^{-1}(E)$ is a properly embedded 2-manifold, you will have no trouble obtaining the stronger conclusion that this map $\tilde{f}$ is transverse to E.

It follows from the discussion I gave above that if $\tilde{f} : \tilde{M} \to T$ is a $\pi_1(M)$-equivariant map transverse to E, then $\tilde{f}^{-1}(E)$ is the inverse image under the covering projection of a properly embedded surface F in M. This surface is well-defined once we have fixed the map $\tilde{f}$. I'll say that a bicollared surface $F \subset M$ is *dual to* the given action of $\pi_1(M)$ on T if it arises via this construction from some $\pi_1(M)$-equivariant map transverse to E.

2.3. Remarks on dual surfaces

When the fundamental group of a compact, orientable, irreducible 3-manifold M acts simplicially on a tree T, properties of the action are reflected in the behavior of surfaces dual to the action. The following statement covers a lot of applications:

2.3.1. *If F is a dual surface to an action of $\pi_1(M)$ on a tree T, then:*
(i) for each component C_i of $M - F$, the subgroup $\mathrm{im}(\pi_1(C_i) \to \pi_1(M))$ *of $\pi_1(M)$ is contained in the stabilizer of some vertex of T; and*
(ii) for each component F_j of F, the subgroup $\mathrm{im}(\pi_1(F_j) \to \pi_1(M))$ *of $\pi_1(M)$ is contained in the stabilizer of some edge of T.*

(Of course, as I mentioned in Subsection 1.2, the subgroup $\mathrm{im}(\pi_1(C_i) \to \pi_1(M))$ is defined only up to conjugacy in $\pi_1(M)$, but statement (i) makes sense because the conjugate of a vertex stabilizer is still the stabilizer of a (probably different) vertex. Likewise for (ii).)

Properties 2.3.1(i) and (ii) are just about immediate from the definition of a dual surface. If F is defined by an equivariant map $\tilde{f} : \tilde{M} \to T$, transvserse to the set E of midpoints of edges of T, then for any component C_i of $M - F$, the subgroup $\Gamma_i = \mathrm{im}(\pi_1(C_i) \to \pi_1(M))$ is the stabilizer of a component $\tilde{C}$ of $M - \tilde{F}$, where $F = f^{-1}(E)$. (Varying Γ_i within its conjugacy class just replaces $\tilde{C}$ by another component of $M - \tilde{F}$.) Now $\tilde{F}$ maps $\tilde{C}$ into a component S of $T - T^0$, where T^0 denotes the set of vertices of T. The equivariance of $\tilde{F}$ implies that Γ_0 stabilizes S. But S is just the open star, relative to the first barycentric subdvision of T, of a vertex s of T; and since the action of $\pi_1(M)$ is simplicial, the stabilizer of S is the same as the stabilizer of s. This proves (i), and (ii) is even easier. □

By the way, if you feel that the inclusions of subgroups given by 2.3.1 really ought to be equalities, you should have a look at the end of this section.

If we are in the really stupid case where the dual surface F is *empty,* then the only component of $M - F$ is the whole manifold M, and it follows from 2.3.1(ii) that $\pi_1(M)$ fixes a vertex of T. Remember that we express this by saying that the action of $\pi_1(M)$ on T is trivial. So:

2.3.2. *If $\pi_1(M)$ acts nontrivially on T then any surface dual to the action is nonempty.*

Sometimes it's useful to think of a dual surface $F \subset M$ in terms of a map of M into a graph, i.e. a 1-dimensional CW-complex. In fact, since Γ is assumed to act on T without inversions, the orbit space $\mathcal{G} = T/\Gamma$ has the structure of a graph in a natural way, the open 1-cells being the homeomorphic images of edges of T. (The presence of inversions would make certain edges get folded in two.) Now if $\tilde{f} : \tilde{M} \to T$ is a $\pi_1(M)$-equivariant map transverse to E, there is a unique map $f : M \to \mathcal{G}$ such that the diagram

$$\begin{tikzcd}
\tilde{M} \arrow[r, "\tilde{f}"] \arrow[d] & T \arrow[d] \\
M \arrow[r, "f"] & \mathcal{G}
\end{tikzcd}$$

commutes. We clearly have $F = f^{-1}(\bar{E})$, where $\bar{E}$ is the image of E in $\mathcal{G}$; you may think of E as the set of midpoints of edges of $\mathcal{G}$. If there were a simple characterization of maps $f : M \to \mathcal{G}$ that are induced by equivariant maps from $\tilde{M}$ to T, this would give a simplified definition of dual surfaces, but the fact is that characterizing such maps is pretty messy. Still, this way of looking at a dual surface is sometimes useful. For one thing, since f is itself transverse to $\bar{E}$ (in essentially the same sense as before), and since each point of $\bar{E}$ is two-sided in $\mathcal{G}$, it's an easy exercise to conclude that each component of F is two-sided in M. So:

2.3.3. *If $\pi_1(M)$ acts on T without inversions then the components of any dual surface are two-sided.*

(Since we've assumed M to be orientable, 2.3.3 is the same as saying that F is orientable, but 2.3.3 would be true even without this assumption.)

By and large, the dual surfaces that I will be working with in this chapter will be piecewise linear with respect to some triangulation of the ambient manifold. As I mentioned at the end of Section 1.4, a two-sided, properly embedded PL hypersurface in a PL manifold is always bicollared. However, one can show without ever mentioning a PL structure on M, or putting any additional restrictions on the surface, that:

2.3.4. *Any surface that's dual to an action of $\pi_1(M)$ on a tree is bicollared.*

The point is that the condition that $f : M \to \mathcal{G}$ is transverse to $\bar{E}$ immediately implies that the surface $F = f^{-1}(\bar{E}) = \tilde{f}^{-1}(E)$ is *locally flat:* this simply means that every point $x \in \tilde{F}$ has a neighborhood V in $\tilde{M}$ which can be mapped homeomorphically onto $\mathbf{R}^3$ in such a way that $\tilde{F} \cap V$ is mapped onto $\mathbf{R}^2$. And it is a theorem of Morton Brown's [9] that a locally flat, two-sided, properly embedded hypersurface in an n-manifold is always bicollared.

Now let's turn to the question of how the notion of dual surface is related to the construction described in Subsection 1.4 that associates a tree with a surface. If F is a bicollared surface in M, and if T is the tree, on which $\pi_1(M)$ acts without inversions, which is given by the construction of 1.4, then since T is by definition a quotient of the universal covering $\tilde{M}$, we have a natural map $\tilde{M} \to T$. It's straightforward to check that this map is equivariant and transverse to the midpoints of edges of T, and that F is the dual surface it defines.

The situation is more subtle if we begin with an action of $\pi_1(M)$ without inversions on a tree T, choose a dual surface F to the action, and compare the given action of $\pi_1(M)$ on T with its action on the tree T' which is associated with F by the construction of Subsection 1.4. This construction does make sense here, since we just saw that F is bicollared. It's natural to wonder whether T' and T are equivariantly isomorphic. This would mean that the inclusions given by 2.3.1 were equalities.

But this is false. In fact, there are perfectly reasonable examples in which the fundamental group of a compact, irreducible, orientable 3-manifold M acts on a tree T in such a way that the stabilizers of all the edges and vertices of T in $\pi_1(M)$ are infinitely generated. This means that we can never have equality in 2.3.1(i) or (ii), regardless of which dual surface we choose.

We can see this phenomenon by using an especially simple example of a tree: the real line $\mathbf{R}$, triangulated with a vertex at each integer point. The action of the group $\mathbf{Z}$ on $\mathbf{R}$ by translations is a simplicial action. Now if Γ is any group and $\phi : \Gamma \to \mathbf{Z}$ is a homomorphism, pulling back the standard action of $\mathbf{Z}$ on $\mathbf{R}$ via ϕ gives an action of Γ on $\mathbf{R}$, which is nontrivial if and only if the homomorphism ϕ is nontrivial. The stabilizer of any edge or vertex is simply the kernel of ϕ. Now Stallings's fibration theorem [60], [31] asserts that if M is a compact, irreducible, orientable 3-manifold, and $\phi : \pi_1(M) \to \mathbf{Z}$ is a homomorphism, then $K = \ker \phi$ can be finitely generated only if M is a locally trivial fiber bundle over S^1 and ϕ is the homomorphism induced by the bundle map. In this case K is just the image of the fundamental group of a fiber F, and it's easy to see that F is just a dual surface; this gives a class of examples where we do have $T' = T$.

In general, however, there is no reason why a homomorphism $\pi_1(M) \to \mathbf{Z}$ should be realized by a bundle map to S^1. Among 3-manifolds which have positive first betti number, and whose fundamental groups therefore admit homomorphisms onto $\mathbf{Z}$, there are certainly plenty of examples—my unjustified intuition says they are a majority—which cannot be fibered over S^1 at all. You can get a sense of the issues involved by looking at [63].

Still, it is not hard to understand the relationship between the trees T and T' in general. By fiddling a little you should be able to show that with a little care in choosing the bicollaring of the dual surface F which is used in constructing the tree T', we can guarantee that the equivariant map $\tilde{f} : \tilde{M} \to T$, which defined F, factors through a $\pi_1(M)$-equivariant map $T' \to T$. Although I won't be using this equivariant map in this chapter, it certainly gives a nice picture. For example, you can read off the inclusions 2.3.1(i) and (ii) from the existence of this map, since equivariance implies that every edge or vertex stabilizer of T' is contained in an edge or vertex stabilizer of T. This is presumably the way that mathematical aristocrats—to borrow a phrase of Raoul Bott's—think about 2.3.1. Personally I have the soul of a petty bourgeois. Otherwise I would never have finished this chapter by the deadline.

2.4. Making a dual surface essential

Now I'll indicate how, given a compact, irreducible 3-manifold M and an action of $\pi_1(M)$ on a tree T, you can construct an essential surface in M that's dual to the given action. We saw in the last section that there is some bicollared surface dual to the action. What I'll give here is a construction that can be carried out whenever a given surface F, dual to the action, is not essential. This construction replaces F by a new surface F', also dual to the action. I will then point out that F' is always "simpler" than F in a suitable sense that I'll make precise; this will imply that by repeating the construction a finite number of times we always arrive at an essential surface.

The most important part of the definition of an essential surface is the π_1-injectivity condition 1.5.1(i). The heart of the matter is therefore to give a construction for simplifying F in the case where 1.5.1(i) fails to hold; by the discussion following Definition 1.5.1, this is the case in which there is a compressing disk D for F. In this situation there is a natural operation, called a *compression*, by which you can replace F by a new surface F': you remove from F an annular neighborhood A of the simple closed curve ∂D, and to the resulting surface you attach two parallel copies of D, say D_1 and D_2, whose boundaries are the two components of ∂A. (Defining "parallel" here involves part (c) of the definition of a compressing disk given in Subsection 1.5, and it's easy to show that F' is bicollared.) A little later I will point out a precise and useful sense in which F' is "simpler" than F. Right now let me point out why the surface F' is dual to the action of $\pi_1(M)$ on T.

Since F is dual to the action, there is a $\pi_1(M)$-equivariant map $\tilde{f} : \tilde{M} \to T$, transverse to E, such that $f^{-1}(E) = p^{-1}(F)$, where $p : \tilde{M} \to M$ is the covering projection. We are required to find another $\pi_1(M)$-equivariant map $\tilde{f}' : \tilde{M} \to T$, transverse to E, such that $(f')^{-1}(E) = p^{-1}(F)$. Let's choose a nice neighborhood B of D in M, so that B is homeomorphic to a ball and meets F in the annulus A, which is properly embedded in B. Then B is the union of a ball X^+ and a solid torus X^-, where $X^+ \cap X^- = A$; we may take the disks D_1 and D_2 to be contained in X^+, and properly embedded in X^-. (Formal proofs of things like this can be given by using regular neighborhood theory. See [32].)

Now let $\tilde{B}$ be a component of $p^{-1}(B)$, so that p maps $\tilde{B}$ homeomorphically onto B, and let $\tilde{A}$, $\tilde{D}_1$, $\tilde{D}_2$, $\tilde{X}^+$ and $\tilde{X}^-$ denote the inverse images in $\tilde{B}$ of A, D_1, D_2, X^+ and X^-. Since $\tilde{f}$ is transverse to E and $\tilde{A} = \tilde{B} \cap f^{-1}(E)$, and since T is a tree, $\tilde{f}$ must map X^+ and X^- to the closures of different components of $T - E$, say Y^+ and Y^-. Let h denote the map from $\partial\tilde{B} \cup D_1 \cup D_2$ to T which agrees with $\tilde{f}$ on $\partial\tilde{B}$ and maps D_1 and D_2 to the point $p(A)$ of E. Then h can be extended to a map $g : \tilde{B} \to T$ such that $g^{-1}(E) = D_1 \cup D_2$. To see this, note that D_1 and D_2 divide B into three balls. If C is any of these balls, h maps ∂C to the closure of either Y^+ or Y^-; this makes it possible to extend $h|\partial C$ to a map from C to the closure or Y^+

or Y^- in such a way that the interior of C is mapped to either Y^+ or Y^-.

Now g admits a unique extension to a $\pi_1(M)$-equivariant map $\tilde{f}'_B : p^{-1}(B) \to T$, since each component of $p^{-1}(B)$ is the image of $\tilde{B}$ under a unique element of $\pi_1(M)$. If we define $\tilde{f}'$ to agree with f'_B on its domain and with $\tilde{f}$ on the rest of $\tilde{M}$, it is clear that $\tilde{f}'$ is continuous and $\pi_1(M)$-equivariant, and that $(\tilde{f}')^{-1}(E) = p^{-1}(F')$. You should convince yourself that with just a little more care in the construction of the extension g of h, we can guarantee that $\tilde{f}'$ is actually transverse to E. This shows that the surface F' is dual to the action.

(One comment that's worth making about the map $\tilde{f}'$ that's been constructed here is that even if $\tilde{f}$ were simplicial with respect to some subdivision of $\tilde{M}$ and the given triangulation of T—as was the case with the map $\tilde{f}$ that I constructed in Subsection 2.1—it's almost certainly necessary to subdivide T before $\tilde{f}'$ can be made simplicial.)

In the other cases where F fails to be essential, the construction of F' is significantly easier. We saw in Subsection 2.3 that F can't fail to satisfy Condition 1.5.1(iv). If F satisfies Condition (i) of 1.5.1 but does not satisfy both Conditions (ii) and (iii), it has a component F_0 which is either a 2-sphere or a boundary-parallel surface. If F_0 is a 2-sphere, then since M is irreducible, F_0 bounds a ball B_0; we may assume F_0 to be chosen so that B_0 doesn't contain any other 2-sphere component of F. Since F already satisfies 1.5.1(i), this implies that B_0 contains no component of F whatever.

It's not hard to see that $F' = F - F_0$ is again a dual surface. For example, in the case where F_0 is boundary-parallel, there's a submanifold M' which is a deformation-retract of M, such that $M' \cap F = F'$. If we think of a deformation-retraction from M to M' as a map $\rho : M \to M$, then ρ is covered by a $\pi_1(M)$-equivariant map $\tilde{\rho} : \tilde{M} \to \tilde{M}$. If $\tilde{f} : \tilde{M} \to T$ is a $\pi_1(M)$-equivariant map, transverse to E, with $\tilde{f}^{-1}(E) = p^{-1}(F)$, then $\tilde{f}' = \tilde{f} \circ \tilde{\rho}$ is also $\pi_1(M)$-equivariant map and transverse to E, and we have $\tilde{f}^{-1}(E) = p^{-1}(F')$. I'll let you work out the case where F_0 is a 2-sphere.

Now we need to address the sense in which F' is "simpler" than F. For any compact 2-manifold F, let me define the *complexity* of F to be the nonnegative integer

$$c(F) = \sum (2 - \chi(F_i))^2, \tag{2.4.1}$$

where F_i ranges over the components of F and χ denotes the Euler characteristic. Since a compact, connected 2-manifold has Euler characteristic at most 2, the expressions whose squares appear in the sum (2.4.1) are always nonnegative.

The reason this complexity is useful is that the operations I've described above, which replace a nonessential dual surface by a new dual surface, usually decrease the complexity. Specifically, this is always true of the first operation I described, the compression, which can be carried out when the given dual surface F fails to satisfy Condition (i) of Definition 1.5.1. To see this, notice that if F' is obtained by a compression from F, we have $F' = (F - \operatorname{int} A) \cup D_1 \cup D_2$, where A is an annulus in F whose core curve doesn't bound a disk in F, and D_1 and D_2 are disks with $(D_1 \cup D_2) \cap F = \partial D_1 \cup \partial D_2 = \partial A$. If F_0 denotes the component of F containing A, then F' is obtained from F by replacing F_0 by the 2-manifold $F_0' = (F_0 - \operatorname{int} A) \cup D_1 \cup D_2$, which has either one or two components. Now when we form the union of two surfaces that meet along a collection of common boundary curves, the Euler characteristic of the union is the sum of the Euler characteristics of the two pieces; hence

$$\chi(F_0') = (\chi(F_0) - \chi(A)) + (\chi(D_1) + \chi(D_2)) = \chi(F_0) + 2.$$

Now there are two cases. If F_0' is connected, the effect of our operation on the sum (2.4.1) is to replace the term $2 - \chi(F_0))^2$ by the term the term $2 - \chi(F_0'))^2$, where $\chi(F_0') = \chi(F_0) + 2 \leqslant 2$, so it's clear that $c(F') < c(F)$.

If F_0' has two components, say F_α' and F_β', then neither F_α' nor F_β' is a sphere; this is because the core curve of A did not bound a disk in F. Hence if we set $a = 2 - \chi(F_\alpha')$ and $b = 2 - \chi(F_\beta')$, then a and b are both strictly positive. Now we have

$$2 - \chi(F_0) = 4 - \chi(F_0') = 4 - (\chi(F_\alpha') + \chi(F_\beta') - 2) = a + b.$$

So the effect of our operation on the sum (2.4.1) is to replace the term $(a + b)^2$ by the term $a^2 + b^2$. Since a and b are strictly positive, we again have $c(F') < c(F)$.

The other operations on F that I described, for the cases where Condition (ii) or (iii) of Definition 1.5.1 fails, amount to discarding a component. This cannot increase the complexity of F, although it may keep it the same if the component in question is a sphere. It's now pretty clear how to get an essential surface which is dual to the given action. We choose a surface F whose complexity is minimal among all dual surfaces; and among those dual surfaces having minimal complexity, we choose F so as to have the smallest possible number of components. If F were not essential, one of our operations would produce either a dual surface of strictly smaller complexity, or one having the same complexity but fewer components, and in either case we'd have a contradiction.

2.5. Applications, I: Nonseparating surfaces

The simplest consequence of the constructions described in this section is that if M is a compact, orientable, irreducible 3-manifold such that $\pi_1(M)$ admits a nontrivial action on a tree, then M contains an essential surface. Conversely according to Proposition 1.5.2, if M contains an essential surface then $\pi_1(M)$ admits a nontrivial action on a tree.

In Subsection 2.3 I already mentioned an especially simple way of constructing an action of $\pi_1(M)$ on a tree: pull back the action of $\mathbf{Z}$ on $\mathbf{R}$ by translations, via some homomorphism from $\pi_1(M)$ to $\mathbf{Z}$. Such homomorphisms are in canonical bijective correspondence with elements of $H^1(M; \mathbf{Z})$. So the general construction described in this section allows one to associate (noncanonically) an essential surface F with any nontrivial element $c \in H^1(M; \mathbf{Z})$. It's a good exercise to show that the image in $H_1(M, \partial M; \mathbf{Z})$ of the fundamental class $[F]$ of F is the Poincaré dual of c. (To make this precise, we need to orient the components of F so that $[F]$ will be well defined; choosing the right orientations is part of the exercise. Here's a hint: the equivariant map $\tilde{f} : \tilde{M} \to \mathbf{R}$ from which F is constructed induces a map from M to S^1.)

Since $[F] \neq 0$, there is at least one component F_0 of F such that $[F_0] \neq 0$, which is tantamount to saying that F does not separate M. In particular:

Proposition 2.5.1. *If M is a compact, orientable, irreducible 3-manifold with positive first betti number, then M contains a nonseparating essential surface.*

This particular result is fundamental in 3-manifold theory. A *Haken manifold* is defined to be a compact, orientable, irreducible 3-manifold which is either homeomorphic to a ball or contains an essential surface. If M is a compact, orientable, irreducible 3-manifold with nonempty boundary, then either ∂M has a 2-sphere component, in which the irreducibility of M implies that M is a ball, or some component of ∂M has genus > 0, in which case an elementary application of Poincaré duality (see for example the proof of Lemma 4.9 in [31]) shows that M has positive first betti number, so that M contains an incompressible surface according to Proposition 2.5.1. So we may state the

Corollary 2.5.2. *Every compact, orientable, irreducible 3-manifold with nonempty boundary is a Haken manifold.*

Now suppose that M is a Haken manifold not homeomorphic to a ball. If we choose an essential surface $F \subset M$, then splitting M along F gives a new—possibly disconnected—manifold M', each component of which has nonempty boundary, and it's not hard to conclude from the irreducibility of M and the essentiality of F that each component of M' is irreducible. So each component of M' is a Haken manifold. If some component of M' is not a ball, it follows that there is an essential surface F' in some component of M', and we can split M' along F' to get a manifold M'', each component of which is a Haken manifold. It was first shown by Haken—see [31], pp. 140–142 for a simplified proof—that if the surfaces F' are chosen with a little care then this process must terminate; that is, there is some $n \geqslant 0$ such that every component of $M^{(n)}$ is a ball. The finite sequence $M, M', \dots M^{(n)}$ is called a *hierarchy* for the manifold M.

Some of the deepest results in 3-manifold theory are theorems about Haken manifolds that are proved by induction on the length of a hierarchy. To prove such a theorem we need only prove that it is true for a given manifold M whenever it is true for each component of the manifold M' obtained by splitting M along an essential surface. The first results of this type were obtained by Waldhausen; for the special case of a closed Haken manifold M his results imply that every irreducible 3-manifold M_1 with $\pi_1(M_1) \cong \pi_1(M)$ is homeomorphic to M, that every outer automorphism of $\pi_1(M)$ is induced by a self-homeomorphism of M, and that self-homeorphisms of M which induce the same outer automorphism are isotopic.

Another particularly famous example is Thurston's theorem characterizing those Haken manifolds which admit hyperbolic metrics of finite volume.

Corollary 2.5.2 applies in particular when M is the exterior of a tame knot in a closed, orientable 3-manifold Σ. By a *knot* in M I mean a subset K homeomorphic to S^1; to say that M is tame means that it has a tubular neighborhood—that is, a neighborhood V which can be mapped homeomorphically to $S^1 \times D^2$ in such a way that K is mapped onto $S^1 \times \{0\}$. (Every smooth or piecewise linear knot is automatically tame.) The *exterior* of K is the closure of $\Sigma - V$. If Σ is a homology 3-sphere (for example S^3), there is a stronger version of Corollary 2.5.2. The proof of the corollary gave a nonseparating essential surface F in M, but I claim we can take F to have a connected boundary. This is an easy consequence of a relative version of the main construction of this section; since I'll need the relative version elsewhere in the chapter, I will spell out the conclusion:

Proposition 2.5.3. *Let M be a compact, orientable, irreducible* 3*-manifold, and suppose that we are given a nontrivial action of $\pi_1(M)$ on a tree T. Let $B_1, \dots, B_k \subset \partial M$ be disjoint (compact) subpolyhedra, and consider the action of each $\pi_1(B_i)$ on T obtained by pulling back the action of $\pi_1(M)$ via the inclusion homomorphism $\pi_1(B_i) \to \pi_1(M)$. (These actions are well-defined up to equivalence; see Subsection 1.2.) Suppose that for $i = 1, \dots, k$ we are given a $\pi_1(B_i)$-equivariant map $\tilde{g}_i : \tilde{B}_i \to T$, where $\tilde{B}_i$ denotes the universal covering space of B, and suppose that each $\tilde{g}_i$ is transverse to the set E of all midpoints of edges of T, so that $\tilde{g}_i^{-1}(E)$ is the pre-image of a unique closed* 1*-manifold $C_i \subset B$. Then there is a dual surface F to the action of $\pi_1(M)$ on T such that $B_i \cap \partial F \subset C_i$ for $i = 1, \dots, k$.*

After a few preliminaries, proving this is just a matter of carrying out the constructions described in Subsections 2.1, 2.2 and 2.4 with a little care. Working out the details is a perfect exercise in understanding these constructions. One uses the equivariance property of $\tilde{g}_i$ to show that $\tilde{g}_i$ factors through a map $\tilde{g}'_i : \tilde{B}_i/N_i$ to T, where N_i is the kernel of the inclusion homomorphism $\pi_1(B_i) \to \pi_1(M)$. The quotient surface $\tilde{B}_i/N_i$ can be identified with a boundary component of the universal cover $\tilde{M}$ of M. Let's define $\tilde{g}' : (\tilde{B}_1/N_1) \cup \dots (\tilde{B}_k/N_k) \to T$ to be the map which restricts to $\tilde{g}'_i$ on each $\tilde{B}_i/N_i$. There is a unique extension of $\tilde{g}$ to a $\pi_1(M)$-equivariant map $\tilde{f}_\partial : p^{-1}(B) \to T$, where $p : \tilde{M} \to M$ is the covering projection and $B = B_1 \cup \cdots B_k$.

If one begins with a triangulation of M in which B is a subcomplex, then in carrying out the inductive construction given in Subsection 2.1 for the $\pi_1(M)$-equivariant map $\tilde{f}$, simplicial with respect to some $\pi_1(M)$-equivariant subdivision of $\tilde{M}$ and the given triangulation of T, one can make the choices of extensions $\tilde{f}_\sigma$ in such a way that for each simplex $\sigma \subset p^{-1}(B)$ we have $\tilde{f}_\sigma = \tilde{f}_\partial|\sigma$. According to the discussion in Subsections 2.2 and 2.3, $\tilde{f}$ is transverse to $E \subset T$. The dual surface defined by $\tilde{f}$—let's call it F_0—has the property that $B_i \cap F_0 = C_i$ for $i = 1, \dots, k$. The arguments of Subsection 2.4 show that by a finite sequence of modifications we can replace F_0 by an essential surface dual to the action. If you examine the effect of these operations on the boundary, you will find that $\partial F \subset \partial F_0$, so that $B_i \cap \partial F \subset C_i$ for $i = 1, \dots, k$, as asserted in Proposition 2.5.3.

Now, as I was saying, Proposition 2.5.3 can be used to show that if M is the exterior of a knot K in a homology 3-sphere Σ, then there is an essential nonseparating surface $F \subset M$ with connected boundary. To see this, first notice that $H_1(M; \mathbf{Z})$, which is the abelianization of $\pi_1(M)$, is infinite cyclic and generated by the meridian class, as is easily deduced from the Mayer-Vietoris theorem. Hence, up to sign, there's a unique homomorphism $\phi : \pi_1(M) \to \mathbf{Z}$, and ϕ maps the conjugacy class λ defined by the longitude of K to 0, and maps the conjugacy class μ defined by the meridian of K to a generator of $\mathbf{Z}$. (See Boyer's chapter for terminology concerning meridians and longitudes.) If we identify ∂M with $S^1 \times S^1$ so that $S^1 \times \{\text{point}\}$ is a meridian and $\{\text{point}\} \times S^1$ is a longitude, then the universal cover of ∂M becomes identified with $\mathbf{R} \times \mathbf{R}$. The hypotheses of Proposition 2.5.3 now hold if we set $T = \mathbf{R}$ and $B = \partial M$, let $\pi_1(M)$ act on $\mathbf{R}$ by the pullback via ϕ of the action of $\mathbf{Z}$ on $\mathbf{R}$ by translations, and define $\tilde{g} : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ to be the projection to the first factor. The 1-manifold C described in the statement of Proposition 2.5.3 has the form $\{\text{point}\} \times S^1$ and is therefore a longitude. Hence Proposition gives a dual surface F whose boundary is either a longitude or the empty set. But if $\partial F = \emptyset$ then the equivariant map defining F maps the boundary of $\widetilde{M}$ into the complement of the set E of midpoints of edges of $\mathbf{R}$, hence into an interval $[n - \frac{1}{2}, n + \frac{1}{2}]$; by equivariance, the vertex n of $\mathbf{R}$ is fixed by $\pi_1(\partial M)$. This is absoid, since μ acts on $\mathbf{R}$ by a unit translation.

So we do indeed have an essential surface $F \subset M$ whose boundary is connected, and is in fact a longitude. It's easy to see that in the tubular neighborhood V of K whose interior was removed from Σ to get M, there is an annulus A whose boundary curves are the knot K and a longitude, which we can take to be ∂F. So $F^+ = F \cup A$ is an embedded (but not properly embedded) compact, orientable surface in Σ with boundary F. Such a surface is called a Seifert surface. While Seifert surfaces are great for making plastic models, properly embedded surfaces in knot exteriors are generally more useful for theoretical work, as you will quickly discover if you try to write down the properties of F^+ that translate the condition that F is essential. Of course, F is so closely related to F^+ that one quickly slips into the habit of calling F a Seifert surface.

One immediate consequence of the existence of an essential Seifert surface in a knot K is that if K is nontrivial then the *group* of K, defined to be $\pi_1(\Sigma - K) \cong \pi_1(M)$, has a nonabelian free subgroup. (To say that K is nontrivial means that is doesn't bound a disk in Σ. So a Seifert surface, which by definition is orientable and has a connected boundary, must have genus > 0, so that $\pi_1(F)$, which injects into $\pi_1(M)$ by condition (i) of Definition 1.5.1, is a nonabelian free group.) This is a simple illustration of how the study of essential surfaces gives information about the structure of a knot group. I'll give a fancier illustration of this in Subsection 5.6 and Section 6.

2.6. Applications, II: Free products and stuff

It's routine to show that if an irreducible 3-manifold M contains an essential (bicollared) disk D then $\pi_1(M)$ is either a nontrivial free product or an infinite cyclic group. If D separates M then the irreducibility of M can be used to show that neither component of $M - D$ is simply connected, so that M is a nontrivial free product by van Kampen's theorem. If D doesn't separate M, van Kampen's theorem exhibits $\pi_1(M)$ as a free product of $\mathbf{Z}$ with $\pi_1(M - D)$, which may or may not be trivial; in either case we get the desired conclusion.

Conversely, if $\pi_1(M)$ is either a nontrivial free product or an infinite cyclic group, then M contains an essential disk. To prove this from the point of view of this section, the main thing you have to notice is that a group Γ which is either infinite cyclic or a free product admits a nontrivial action on a tree T in which the stabilizer of each edge is trivial. If Γ is infinite cyclic we can take $T = \mathbf{R}$. If Γ is a nontrivial free product, one can either construct the tree from the algebra of a free product—a good exercise—or describe it topologically as follows. Think of Γ as the fundamental group of a space X which is the union of two disjoint nonsimply connected spaces A and B and a topological arc which meets each of A and B in a single point. If $\tilde{X}$ is the universal cover of X and $p : \tilde{X} \to X$ is the covering projection, we can obtain T as a quotient of $\tilde{X}$ by identifying each component of $p^{-1}(A \cup B)$ to a point. The usual action of $\pi_1(X)$ on $\tilde{X}$ induces an action on T with the required property.

Now if $\pi_1(M)$ acts on a tree with trivial edge stabilizers, then according to the results of Subsections 2.2 and 2.4, there is an essential dual surface F to the action; and according to 2.3.1(ii), the fundamental group of each component of F is contained in the stabilizer of an edge of T and is therefore trivial. As my definition of essential surface rules out 2-sphere components, the components of F must be essential disks.

There is a useful generalization of the notion of a free product. Suppose we are given groups A, B and C, and injective homomorphisms $i : C \to A$ and $j : C \to B$. The *free product of A and B amalgamated over C*, denoted rather vaguely by $A \star_C B$, is defined to be quotient of the free product $A \star B$ obtained by adding the relations $i(c) = j(c)$ for all $c \in C$. It is a theorem, of which you will find an elegant account in [37], that the natural homomorphisms from A and B (and hence from C) to $A \star_C B$ are injective. One identifies A, B, and C with subgroups of $A \star_C B$, and one says that the amalgamated free product $A \star_C B$ is *nontrivial* if A and B are proper subgroups of $A \star_C B$.

Of course this comes up naturally in topology: if Z is, say, a connected bicollared hypersurface in an n-manifold M such that $M - Z$ has two connected components X and Y, and if the inclusion homomorphism from $\pi_1(Z)$ to $\pi_1(M)$ happens to be injective, then van Kampen's theorem exhibits $\pi_1(M)$ as the amalgamated free product $\pi_1(X) \star_{\pi_1(Z)} \pi_1(Y)$. (Here we should choose a base point in Z and take the homomorphisms i and j to be induced by inclusion.)

Generalizing the fact about free products which I talked about earlier in this section, one can show that any amalgamated free product $A \star_C B$ acts on a tree in such a way that the edge stabilizers are precisely the conjugates of C in $A \star_C B$, and the vertex stabilizers are precisely the conjugates of A and B. You can do this topologically by an argument very close to the one I gave for free products, or you can work it out algebraically from the normal form for elements of an amalgamated free product given in [37]. You can also find a proof in [55], about which I'll be saying more shortly. It's clear from the definitions that the action on the tree is nontrivial if and only if $A \star_C B$ is a nontrivial amalgamated free product.

If F is a separating, connected, essential surface in a closed, orientable, irreducible 3-manifold then $\pi_1(M)$ is a nontrivial free product with amalgamation. This amounts to saying that neither component of $M - F$ can carry $\pi_1(M)$, a fact which you can extract from the proof of 1.5.2 above, or deduce from the statement.

In the converse direction, if $\pi_1(M)$ is a nontrivial free product with amalgamation, then since $\pi_1(M)$ admits a nontrivial action an a tree, M must contain an essential surface. This is because $\pi_1(M)$ admits a nontrivial action on a tree, and you can apply the very first sentence of Subsection 2.5. However, the surface we get this way need not separate M. (Note that the dual surface to the action on the tree could fail to be connected.)

It's clear from these observations, together with the ones I made in Subsection 2.5, that a compact, orientable, irreducible 3-manifold is a Haken manifold if and only if *either* the first betti number of M is strictly positive or $\pi_1(M)$ is a nontrivial free product with amalgamation. In [67], Waldhausen attributed this result to D.B.A. Epstein.

Feustel [27] showed that if $\pi_1(M) \cong A \star_C B$, where C is isomorphic to the fundamental group of a closed, orientable surface of positive genus, then there is a separating, closed, essential surface $F \subset M$ such that the inclusion homomorphisms map $\pi_1(F)$ onto C and map the fundamental groups of the components of $M - F$ onto A and B.

Of course there is no need to stop at free products with amalgamation. If Z is any bicollared hypersurface in an n-manifold M such that the inclusion homomorphism from $\pi_1(Z_i)$ is injective for each component Z_i of Z, one can use van Kampen's theorem to compute $\pi_1(M)$ from the fundamental groups of the components of Z and of $M - Z$. The appropriate structure was described elegantly in the work of Bass and Serre presented in [55], who introduced the notion of the "fundamental group of a graph of groups" as the relevant generalization of an amalgamated free product. As you would expect, the fundamental group of a graph of groups has a canonical action on a tree. What is more surprising is the converse: every action of an arbitrary group Γ on a tree arises in an essentially unique way from an isomorphism of Γ with the fundamental group of a graph of groups. There is also a topological approach to this theory; see [53].

You can think of the material in this chapter as being vaguely analogous to the Bass-Serre theory. If we start out, not with an abstract group but with the fundamental group of a (compact, irreducible, orientable) 3-manifold M, then we have seen how construct, from an action of $\pi_1(M)$ on a tree, not just a group-theoretical identification of $\pi_1(M)$ with the fundamental group of a graph of groups, but an essential surface in M itself, from which such an identification can be constructed. This construction is less canonical than the one given by the Bass-Serre theory, and the relationship between the dual surface and the action is less direct than the connection between a general group action and the associated graph of groups. It's nevertheless a useful construction. I've given a few hints in this section and the last one about why it's useful. To exploit it further one needs to combine it with ideas from algebra and geometry, which I'll be presenting in the next three sections.

3. The tree for SL_2

In the last two sections I illustrated how group actions on trees come up in 3-manifold theory. Another subject in which such actions come up naturally is the study of groups of 2×2 matrices with entries in a field: there is a natural way of constructing actions of such groups on trees, and this provides a beautiful and powerful way of analyzing the algebraic structure of these groups. In this section I will be giving a brief introduction to the ideas involved, from a purely algebraic point of view. In Section 5 I will explain the surprising interaction of these ideas from algebra with the topological theory of 3-manifolds.

The classic work on the tree for SL_2 is Serre's book [55]. The construction of the trees in question is a very special case of a construction due to Bruhat and Tits [10]. My aims here are to give a quick, self-contained account of enough of the material to allow you to read the rest of this chapter, and to inspire you to read [55] and perhaps [10].

3.1. Valuations

The starting point for this theory is the notion of a valuation. Valuations are objects that come up naturally in both number theory and complex analysis—so naturally, in fact, that anyone who has thought about the most elementary aspect of either subject has really worked with valuations, whether consciously or not. To illustrate the number-theoretic aspect of the idea, consider a prime number p. Given any integer $a \neq 0$, let us denote by $v(a) = v_p(a) \geqslant 0$ the exponent of p in the prime factorization of $|a|$. The fundamental theorem of arithmetic asserts that the surjective map $v : \mathbf{Z} - \{0\} \to \mathbf{N}$ (where $\mathbf{N}$ denotes the set of nonnegative integers) is well-defined. The reason the theorem is so powerful is that once we know v is well-defined, it follows immediately that it behaves well under addition and multiplication: we have

$$v(ab) = v(a) + v(b) \qquad \text{for all } a, b \neq 0, \tag{3.1.1}$$

and

$$v(a+b) \geqslant \min(v(a), v(b)) \text{ for all } a, b \text{ such that } a, b \text{ and } a+b \text{ are all nonzero.} \tag{3.1.2}$$

(The inequality (3.1.2) just says that if a power of p divides both a and b then it divides $a + b$.)

Now it is an elementary exercise to show that if a map v of an integral domain R onto $\mathbf{N}$ satisfies (3.1.1) and (3.1.2), and if K denotes the field of fractions of R, then v extends uniquely to a map $\bar{v}$ of K onto $\mathbf{Z}$ satisfying the identities (3.1.1) and (3.1.2). (Any such $\bar{v}$ must satisfy $\bar{v}(a/b) = v(a) - v(b)$ for $a, b \in R - \{0\}$, and you can check that this formula gives a well-defined extension with the required properties. Surjectivity is clear.) Now if K is a field, a *valuation* of K is defined to be a surjection $v : K - \{0\} \to \mathbf{Z}$ that satisfies (3.1.1) and (3.1.2). (In the general theory of such things, these are called discrete, rank-1 valuations, but as they are the only kind of valuation I will be talking about for most of this chapter, I will just call them valuations for now. I will have occasion to mention more general valuations in Section 11.)

3.2. The p-adic valuation

So the map v_p of $\mathbf{Z}$ extends to a valuation of $\mathbf{Q}$. I'll denote the extension by v_p as well; it's called the p-adic valuation. If $a/b \in \mathbf{Q}$ is a fraction that has been written in lowest terms, and if p appears with exponent $r > 0$ in the factorization of a, we have $v_p(a/b) = r$. We have $v_p(a/b) = -r$ if p appears with exponent $r > 0$ in the factorization of b; and if p divides neither a nor b we have $v(a/b) = 0$.

As an illustration of how basic this is in number theory, consider the theorem, which was beyond the reach of the ancient Greeks even for $k = 2$, that the k-th root of a positive integer n is either an integer or an irrational. The point is that if the k-th root—let's call it α—is rational, then for every prime p we have $v_p(n) = kv_p(\alpha)$; thus $k|v_p(n)$ for every p, and it follows from the definition of the v_p that n is a k-th power of an integer.

3.3. Fields of meromorphic functions

To illustrate the relevance of valuations to complex analysis, consider a meromorphic function f which is defined on an open set $u \subset \mathbf{C}$ and is not identically zero. For any point $z_0 \in u$ we can write $f(z)$, for z in some neighborhood of z_0, in a unique way as a laurent series

$$f(z) = \sum_{n=r}^{\infty} a_n (z - z_0)^n,$$

where r is an integer, not necessarily positive, and $a_r \neq 0$. I'll call r the *order* of f at z_0. (If $r > 0$ we may say that f has a zero of order r, and if $r < 0$ that it has a pole of order $|r|$.) The meromorphic functions on u form a field K under pointwise addition and multiplication, and it's pretty clear that the function $v : K - \{0\} \to \mathbf{Z}$ that assigns to each meromorphic function its order at z_0 is a valuation.

When v is a valuation of a field K it's convenient to extend v to a function, still denoted v, defined on all of K and taking values in a set $\mathbf{Z} \cup \{\infty\}$, where ∞ is a new element that we adjoin to $\mathbf{Z}$, by setting $v(0) = \infty$. If we do this then the identities (3.1.1) and (3.1.2) hold for all $a, b \in K$, provided that we interpret $+$, $\geqslant$ and min in the obvious ways on the set $\mathbf{Z} \cup \{\infty\}$.

3.4. The valuation ring

A valuation v of a field K gives a good deal of nice structure. It's immediate from the definitions that the elements $x \in K$ such that $v(x) \geqslant 0$ form a sub-ring (with unity) of K. (I'm following the convention here according to which $v(0)$ is defined to be ∞ and therefore to be $\geqslant 0$.) This ring is called the *valuation ring* associated to v, and I'll denote it $\mathcal{O}_v$. Note that a nonzero element $x \in K$ satisfies $v(x) = 0$ if and only if x and x^{-1} are both in $\mathcal{O}_v$. Thus the elements $x \in K$ with $v(x) = 0$ comprise the (multiplicative) group of units $\mathcal{O}_v^*$ of the ring $\mathcal{O}_v$.

The ideals in the ring $\mathcal{O}_v$ are of a very simple form. to see this, let us fix an element $\pi \in K$ with $v(\pi) = 1$. Such an element is called a *uniformizer.* Let $\mathcal{I}$ be any ideal in $\mathcal{O}_v$, and let's set $n = \min_{x \in \mathcal{I}} v(n)$. If we fix an $x_0 \in \mathcal{I}$ with $v(x_0) = n$, then $\pi^n x_0^{-1} = n - n = 0$, so that $\pi^n \in x_0 \mathcal{O}_v^* \subset \mathcal{I}$. Conversely, for any $x \in \mathcal{I}$ we have $v(x/\pi^{-n}) = v(x) - n \geqslant 0$, so that $\pi^n | x$ in $\mathcal{O}_v$. This shows that $\mathcal{I}$ is the principal ideal generated by π^n. So $\mathcal{O}_x$ is a principal ideal domain, and the only ideals are

$$(1) \supset (\pi) \supset \ldots \subset (\pi^n) \subset \ldots,$$

all of them linearly ordered by inclusion. In particular, $\mathcal{M}_v = (\pi)$, which consists of all elements $x \in K$ with $v(x) > 0$, is the unique maximal (proper) ideal of $\mathcal{O}_v$. note that $\mathcal{O}_v - \mathcal{M}_v = \mathcal{O}_v^*$. Since $\mathcal{M}_v$ is a maximal ideal, the ring $k_v = \mathcal{O}_v / \mathcal{M}_v$ is a field, called the *residue field* of v.

Let's see what all this looks like in the examples of Subsections 3.2 and 3.3. In the example of 3.2, the valuation ring is the ring $\mathbf{Z}_{(p)}$ consisting of all rational numbers which, when written in lowest terms, have denominators not divisible by the given prime p. In this example, p is itself a uniformizer, and the maximal ideal $p\mathbf{Z}_{(p)}$ of $\mathbf{Z}_{(p)}$ consists of all rational numbers which, when written in lowest terms, have numerators divisible by p. the residue field is easily seen to be isomorphic to $\mathbf{Z}/p\mathbf{Z}$; in fact, the usual homomorphism $\mathbf{Z} \to \mathbf{Z}/p\mathbf{Z}$ extends to a homomorphism $\mathbf{Z}_{(p)} \to \mathbf{Z}/p\mathbf{Z}$ with kernel $p\mathbf{Z}_{(p)}$.

In the example of Subsection 3.3, the valuation ring is the ring $\mathcal{O}_{z_0}$ consisting of all meromorphic functions on u which do not have poles at z_0. The function $z - z_0$ is a uniformizer, and the maximal ideal $\mathcal{M}_{z_0}$ of $\mathcal{O}_{z_0}$ consists of all meromorphic functions on u that have zeros at z_0. The homomorphism $f \mapsto f(z_0)$ from $\mathcal{O}_{z_0}$ to $\mathbf{C}$ is surjective, since it maps the field of constant functions isomorphically to $\mathbf{Z}$, and its kernel is obviously $\mathcal{M}_{z_0}$. Hence the residue field k_v is canonically isomorphic to $\mathbf{C}$ in this example.

3.5. The p-adics

One way of thinking about a valuation v of a field K—and this is valuable general knowledge, although it won't be essential for the applications to 3-manifolds—is that it defines a "nonarchimedean absolute value" on K. Let's choose any real constant $c > 1$, and let's set $|x| = c^{-v(x)}$ for every $x \in K - \{0\}$ and $|0| = 0$. Then we have $|x| = 0$ if and only if $x = 0$, and the properties (3.1.1) and (3.1.2) of v translate into the identities

$$|ab| = |a| \cdot |b|,$$

and

$$|a + b| \leqslant \max(|a|, |b|)$$

which hold for all $a, b \in K$. From this it is routine to deduce that K becomes a metric space if we define the distance between x and y to be $|x - y|$, and that the field operations are continuous in terms of the topology defined by this metric. The closed unit ball about 0 in K is obviously just the valuation ring $\mathcal{O}_v$. It's also routine to show that the field operations extend uniquely to the completion $\hat{K}$ of the metric space K, which thereby itself becomes a field, and that v extends uniquely to a valuation $\hat{v}$ of $\hat{K}$. It is not hard to show that the residue field of $\hat{v}$ is naturally isomorphic to the residue field of v.

The "p-adic distance" $d_p(x, y) = |x - y|_p$ defined by the valuation v_p of the field $\mathbf{Q}$ is an especially nice example. The definition of the distance depends on the choice of the constant c. For deep number-theoretic reasons it is customary to take the constant c to be p in this case. I'll use this choice of constant so as make my notation standard, but the choice of c does not affect anything I'll be talking about here. One nice feature of this example is that the unit ball $\mathcal{O}_v = \mathbf{Z}_{(p)}$ contains $\mathbf{Z}$ as a dense subset. To see this, note that if we're given any element of $\mathbf{Z}_{(p)}$, say a/b where $a, b \in \mathbf{Z}$ and p does not divide b, we have $p^n x + by = 1$ for some integers x and y, so that

$$|\frac{a}{b} - ay|_p = |\frac{axp^n}{b}| \leqslant p^{-n},$$

from which the assertion follows.

The field obtained by completing of $\mathbf{Q}$ with respect to the distance function d_p is called the field of p-adic numbers and is denoted $\mathbf{Q}_p$. The valuation ring of $\mathbf{Q}_p$, which is the closed unit ball about 0, is denoted $\mathbf{Z}_p$ and is called the ring of p-adic integers. Since $\mathbf{Z}$ is dense in $\mathbf{Z}_{(p)}$, it's easy to deduce that $\mathbf{Z}$ is dense in $\mathbf{Z}_p$ as well. So we can think of $\mathbf{Z}_p$ as the completion of $\mathbf{Z}$ with respect to the p-adic distance.

Now $\mathbf{Z}$ is a totally bounded metric space with respect to the p-adic distance. This is because for any integer $n > 0$ we can write $\mathbf{Z}$ as the union of p^n congruence classes modulo p^n, and each of these classes has diameter p^{-n}. so the completion $\mathbf{Z}_p$ of $\mathbf{Z}$ is compact. From this it's easy to deduce that $\mathbf{Q}_p$ is locally compact.

As a field with an absolute value, which is locally compact with respect to the topology defined by the absolute value, $\mathbf{Q}_p$ has many formal properties in common with the field $\mathbf{R}$ of real numbers. From a formal point of view it is interesting to ask questions about p-adic numbers that are analogous to familiar questions about real numbers. Actually such questions are of far more than formal interest, because of the role of the p-adic numbers in number theory. A famous example is the Hasse-Minkowski principle, which addresses the question of when a diophantine equation in n variables

$$\sum a_{ij} x_i x_j = 0,$$

where a_{ij} are integers for $i, j \in \{1, \ldots, n\}$, has a nontrivial solution, i.e. whether there are integers $x_1, \ldots, x_n$, not all 0, that satisfy the equation. We can obviously replace both occurrences of the word "integers" here by "rational numbers," and we can assume without loss of generality that the matrix (a_{ij}) is symmetric; the question is then one about nontrivial zeros of a quadratic form in $\mathbf{Q}$. The Hasse-Minkowski principle says that such a form has a nontrivial zero in $\mathbf{Q}$ if (and only if) it has a nontrivial zero in $\mathbf{R}$ and also in $\mathbf{Q}_p$ for every prime p. Saying the form has a nontrivial zero over $\mathbf{R}$ is the same as saying that it's indefinite—i.e. that it's neither positive definite nor negative definite—and this information can be read off from the signs of some minors of the matrix. The glorious part is that one can show, for example, that if $n \geqslant 5$ then the form has a nontrivial zero in $\mathbf{Q}_p$ for every p; so we get the elegant result that the above equation has a nontrivial integer solution whenever $n \geqslant 5$ and the (symmetrized) matrix of the form is indefinite.

The Hasse-Minkowski theorem is a special case of a "local-global" principle which says you can do certain kinds of things in $\mathbf{Q}$ if you can do them in $\mathbf{R}$ and in $\mathbf{Q}_p$ for every p. Of course the principle doesn't always work. Now you know as much about this as I do; to learn more, look at [54].

In Subsection 3.10 I'll give an actual proof of a p-adic analogue of a very familiar theorem involving real numbers.

3.6. Defining the tree for SL_2

Let K be any field. I'll show how to associate with any valuation v of K a tree $T = T_v$ on which $\mathrm{GL}(2, K)$ acts in a natural way. (As I've said, the tree is a special case of an object discovered by Bruhat and Tits. The description of it that I'll give is due to Serre.) Let's consider the standard 2-dimensional vector space $V = K^2$ over K. In particular we may regard V as a module over the valuation ring $\mathcal{O} = \mathcal{O}_v$. We define a *lattice* in V to be an $\mathcal{O}$-submodule of V which is finitely generated and spans V as a vector space over K. Since $\mathcal{O}$ is a principal ideal domain, any finitely generated $\mathcal{O}$-submodule of V is a free $\mathcal{O}$-module of some rank $\leqslant 2$. If the rank is < 2, the submodule cannot span V as a vector space. So any lattice is of rank 2: as far as their isomorphism type is concerned, all lattices look just like the standard lattice $\mathcal{O}^2 \subset K^2$.

In fact we can say slightly more. If Λ_0 and Λ_1 are lattices, and if $\{e_i, f_i\}$ is a basis of Λ_i as a free $\mathcal{O}$-module, then each $\{e_i, f_i\}$ is also a basis of V as a vector space over K. Hence there is a linear automorphism of V, which we can think of as an element A of $\mathrm{GL}(2, K)$, mapping $\{e_0, f_0\}$ onto $\{e_1, f_1\}$. In particular, A carries the lattice Λ_0 onto Λ_1.

When Λ_0 and Λ_1 are lattices, the element of $\mathrm{GL}(2, K)$ carrying Λ_0 onto Λ_1 is far from being unique, but it does define an invariant quantity according to the following result.

Lemma 3.6.1. *Let Λ_0 and Λ_1 be lattices, and let A and B be two linear automorphisms of $\mathrm{GL}(2, K)$ that carry Λ_0 onto Λ_1. Then $v(\det A) = v(\det B)$.*

Proof. Set $C = B^{-1}A$, so that $C(\Lambda_0) = \Lambda_0$. Hence if $\{e, f\}$ is a basis for Λ_0 as an $\mathcal{O}$-module, the matrix expressing C in terms of the basis $\{e, f\}$ has entries in $\mathcal{O}$, so that $\det C \in \mathcal{O}$. Since C^{-1} also leaves Λ_0 invariant, we have $(\det C)^{-1} \in \mathcal{O}$ as well. Thus $\det C$ is a unit in $\mathcal{O}$ and hence $v(\det C) = 0$. The conclusion of the lemma now follows from (3.1.1) and the multiplicativity of determinants. □

In view of this lemma we can associate an integer $\delta(\Lambda_0, \Lambda_1)$ with any ordered pair of lattices (Λ_0, Λ_1) by setting $\delta(\Lambda_0, \Lambda_1) = v(\det A)$, where A is an arbitrary linear automorphism of V mapping λ_0 onto Λ_1. From the multiplicativity of the determinant we see that

$$\delta(\Lambda_0, \Lambda_2) = \delta(\Lambda_0, \Lambda_1) + \delta(\Lambda_1, \Lambda_2) \tag{3.6.2}$$

for any lattices $\Lambda_0, \Lambda_1, \Lambda_2$. Since the identity has determinant 1, we also have

$$\delta(\Lambda, \Lambda) = 0 \tag{3.6.3}$$

for every lattice Λ.

If the lattices Λ_0 and Λ_1 satisfy $\Lambda_1 \subset \Lambda_0$, and if A is a linear transformation which maps Λ_0 onto Λ_1, then since $A(\Lambda_0) \subset \Lambda_0$, the argument used to prove Lemma 3.6.1 shows that in a suitable basis A has entries in $\mathcal{O}$ and therefore that $\det A \in \mathcal{O}$, i.e. $v(\det A) \geqslant 0$. So:

$$\text{If } \Lambda_1 \subset \Lambda_0 \text{ then } \delta(\Lambda_0, \Lambda_1) \geqslant 0. \tag{3.6.4}$$

Note also that if B is a linear transformation of V, then for any lattices Λ_0 and Λ_1 we have

$$\delta(B(\Lambda_0), B(\Lambda_1)) = \delta(\Lambda_0, \Lambda_1), \tag{3.6.5}$$

since if A is a linear transformation mapping Λ_0 onto Λ_1 then BAB^{-1} maps $A(\Lambda_0)$ onto $A(\Lambda_1)$.

We may think of $\delta(\cdot,\cdot)$ as an "algebraic distance" between lattices. What is more directly related to the construction of the tree T_v is a "geometric distance" between lattices, or more precisely between *homothety classes* of lattices. Two lattices $\Lambda, \Lambda' \subset V$ are said to be *(homothety)-equivalent,* or to represent the same homothety class, if there is a nonzero element α of K such that $\Lambda' = \alpha\Lambda$. You can see immediately that this really is an equivalence relation. It will turn out that our tree T_v is defined in such a way that its vertices are in bijective correspondence with homothety classes of lattices. The distance function that I will define will turn out to give the number of edges you have to follow to get from one vertex to another.

Because the homothety classes of lattices are going to be vertices, I will often use the letter s for a homothety class when I'm thinking of it as an object in its own right. (It's convenient here to follow Serre, who wrote in French: s means a vertex (*sommet*), whereas v is the valuation.) On the other hand, I will sometimes write $[\Lambda]$ for the homothety class of a lattice Λ that has already been named.

In order to define the geometric distance between homothety classes, we need two lemmas, of which the first is almost trivial.

Lemma 3.6.6. *If Λ_0 and Λ_1 are lattices, then Λ_1 is equivalent to a lattice Λ_1' such that $\Lambda_1 \subset \Lambda_0$.*

Proof. For $i = 0, 1$, let $\{e_i, f_i\}$ be a basis for Λ_i as an $\mathcal{O}$-module. Then $\{e_i, f_i\}$ is also a basis for V as a vector space, so we can write $e_1 = \alpha e_0 + \beta f_0$ for some $\alpha, \beta \in K$. Since $e_1 \neq 0$ we have $m_0 = -\min(v(\alpha), v(\beta)) \in \mathbf{Z}$. For any $m \geqslant m_0$ we have $v(\pi^m\alpha) = m + v(\alpha) \geqslant 0$, so that $\pi^m\alpha \in \mathcal{O}$, and likewise $\pi^m\beta \in \mathcal{O}$. Hence $\pi^m e_1 \in \Lambda_0$ for $m \geqslant m_0$. Similarly, we see that $\pi^m f_1 \in \Lambda_0$ for m sufficiently large. So we can find an m for which $\pi^m e_1$ and $\pi^m f_1$ both belong to Λ_0; hence $\pi^m\Lambda_1$, which is equivalent to Λ_1, is contained in Λ_0. □

Rappel 3.6.7. In the proof of the next lemma I'll be using a basic result on finitely generated modules over a principal ideal domain: *if L_0 is a free module of finite rank over a p.i.d. R, and L_1 is a submodule of L_0, then there exist a basis $\{e_1, \ldots, e_n\}$ for L_0 and elements $\alpha_1, \ldots, \alpha_n$ of R such that L_1 is generated by $\alpha_1 e_1, \ldots, \alpha_n e_n$.* (This result underlies one proof of the structure theorem for finitely generated modules over R: any finitely generated submodule M can obviously be written in the form L_0/L_1, where L_0 is a finitely generated free module and L_1 is a submodule of L_0; the above result then shows that M is a finite direct sum of cyclic modules.)

Lemma 3.6.8. *If Λ_0 and Λ_1 are lattices, then there is a unique lattice Λ_1' equivalent to Λ_1 such that $\Lambda_1 \subset \Lambda_0$ and Λ_0/Λ_1 is isomorphic as an $\mathcal{O}$-module to $\mathcal{O}/\beta\mathcal{O}$ for some nonzero element β of $\mathcal{O}$.*

Proof. By Lemma 3.6.6, we may assume that Λ_1 is already contained in Λ_0. By the result I mentioned before the proof, in Rappel 3.6.7, Λ_0 has a basis $\{e, f\}$ such that Λ_1 is generated by $\{\alpha e, \gamma f\}$ for some $\alpha, \gamma \in \mathcal{O}$. After possibly reversing the roles of e and f we may assume that $v(\alpha) \leqslant v(\gamma)$; setting $\beta = \gamma\alpha^{-1}$ we conclude that $v(\beta) \geqslant 0$, so that $\beta \in \mathcal{O}$. The lattice $\Lambda_1' = \alpha^{-1}\Lambda_1$, which is equivalent to Λ, has the basis $\{e, \beta f\}$. It follows that $\Lambda_1' \subset \Lambda_0$ and that Λ_0/Λ_1' is isomorphic as an $\mathcal{O}$-module to $\mathcal{O}/\beta\mathcal{O}$. This proves the existence assertion.

Now suppose there is a second lattice Λ_1'' equivalent to Λ_1 such that $\Lambda_1'' \subset \Lambda_0$ and such that Λ_0/Λ_1'' is isomorphic as an $\mathcal{O}$-module to $O/\delta\mathcal{O}$ for some $\delta \in \mathcal{O}$. We may write $\Lambda_1'' = \zeta\Lambda_1'$ for some $\zeta \in F$. Since Λ_1' contains the basis element e of $\mathcal{O}$, we have $\zeta e \in \Lambda_1'' \subset \Lambda_0$, which implies that $\zeta \in \mathcal{O}$. Since Λ_1'' is generated by ζe and $\zeta\beta f$, we have $\Lambda_0/\Lambda_1'' \cong \mathcal{O}/\zeta\mathcal{O} \oplus \mathcal{O}/\zeta\beta\mathcal{O}$. But the uniqueness part of the structure theorem for modules over a p.i.d. implies that $\mathcal{O}/\zeta\mathcal{O} \oplus \mathcal{O}/\zeta\beta\mathcal{O}$ can't be cyclic unless ζ is a unit in $\mathcal{O}$, in which case $\Lambda_1'' = \Lambda_1'$. This proves the uniqueness assertion, and completes the proof of Lemma 3.6.8. □

Given lattices Λ_1 and Λ_0, I will say that Λ_1 is *snugly embedded in* Λ_0 if Λ_0 and Λ_1 are related in the way described in the conclusion of Lemma 3.6.8, that is, if $\Lambda_1 \subset \Lambda_0$ and Λ_0/Λ_1 is a cyclic $\mathcal{O}$-module. Now if s_0 and s_1 are homothety classes of lattices, I'll define the "geometric distance" $d(s_0, s_1)$ to be the integer $\delta(\Lambda_0, \Lambda_1)$, where the Λ_i are representatives of the s_i such that Λ_1 is snugly embedded in Λ_0. According to Lemma 3.6.8, such representatives Λ_0 and Λ_1 exist, and Λ_1 is uniquely determined once Λ_0 has been chosen. To show that $d(s_0, s_1)$ is independent of the choice of Λ_0, note that if Λ_0 and Λ_0' are representatives of s_0, so that $\Lambda_0' = \alpha\Lambda_0$ for some nonzero element α of F, and if Λ_1 is a representative of s_1 that's snugly embedded in Λ_0, then $\Lambda_1' = \alpha\Lambda_1$ represents s_1 and is snugly embedded in Λ_0', and by applying (3.6.5) to the linear transformation $x \mapsto \alpha x$ we find that

$$\delta(\Lambda_0', \Lambda_1') = \delta(\alpha\Lambda_0, \alpha\Lambda_1) = \delta(\Lambda_0, \Lambda_1).$$

It follows from (3.6.4) that $d(s_0, s_1) \geqslant 0$ for any two homothety classes of lattices s_0 and s_1.

To understand the definition of d better, let's consider an arbitrary lattice Λ_0 and a lattice Λ_1 that's snugly embedded in Λ_0. From the proof of Lemma 3.6.8 (or, more precisely, the existence part of the proof and the uniqueness part of the statement) we see that Λ_0 has a basis $\{e, f\}$ such that e and βf generate Λ_1 for some $\beta \in F - \{0\}$. The linear transformation of V whose matrix in the basis $\{e, f\}$ is $\begin{pmatrix} 1 & 0 \\ 0 & \beta \end{pmatrix}$ has determinant β and maps Λ_0 onto Λ_1. Hence

$$d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1) = v(\beta). \tag{3.6.9}$$

It's worth noticing that if π is a uniformizer in $\mathcal{O}_v$ then in the above discussion we may take β to be a nonnegative power of π, since every nonzero element of $\mathcal{O}_v$ is a nonnegative power of π with a unit. If $\beta = \pi^n$ then $d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1) = n$.

It's useful to generalize the description of the distance that I just gave. If Λ_0 and Λ_1 are lattices with $\Lambda_1 \subset \Lambda_0$, then by Rappel 3.6.7 Λ_0 and Λ_1 have bases of the form $\{e, f\}$ and $\{\alpha e, \gamma f\}$ for some $\alpha, \gamma \in F - \{0\}$. Here, using the matrix $\begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}$, we find that

$$\delta(\Lambda_1, \Lambda_0) = v(\alpha\gamma) = v(\alpha) + v(\gamma).$$

This is used in the proof of the following lemma, which is the main step in the proof that d is a distance function.

Lemma 3.6.10. *If $\Lambda_0 \supset \Lambda_1$ are lattices, we have*

$$d([\Lambda_0], [\Lambda_1]) \leqslant \delta(\Lambda_0, \Lambda_1)$$

and

$$d([\Lambda_0], [\Lambda_1]) \equiv \delta(\Lambda_0, \Lambda_1) \pmod 2.$$

Furthermore, we have $d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1)$ if and only if Λ_1 is snugly embedded in Λ_0.

Proof. This is a lot like the existence part of the proof of Lemma 3.6.8. We fix bases $\{e, f\}$ and $\{\alpha e, \gamma f\}$ for Λ_0 and Λ_1, where $\alpha, \gamma \in \mathcal{O}$. As in the proof of 3.6.8, we may assume that $v(\alpha) \leqslant v(\gamma)$; and as in that proof it follows that $\beta = \gamma\alpha^{-1} \in \mathcal{O}$, and that $\Lambda_1' = \alpha^{-1}\Lambda_1$, which is equivalent to Λ_1, has the basis $\{e, \beta f\}$ and is therefore snugly embedded in Λ_0. By the remarks before the statement of the lemma we're proving now, we find that $d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1') = v(\beta)$, and that $\delta(\Lambda_1, \Lambda_0) = v(\alpha) + v(\gamma) = v(\beta) + 2v(\alpha)$, so that

$$\delta(\Lambda_0, \Lambda_1) = d([\Lambda_0], [\lambda_1]) + 2v(\gamma),$$

from which the first two assertions follow. If $\delta(\Lambda_0, \Lambda_1) = d([\Lambda_0], [\lambda_1])$ then $v(\gamma) = 0$; thus $\gamma \in \mathcal{O}^*$, and it follows that $\{e, \gamma f\}$ is a basis for Λ_1, hence that Λ_1 is snugly embedded in Λ_0. The converse follows from the definition of d. □

Let's denote by $T^{(0)}$ the set of all homothety classes of lattices, so that d is a nonnegative integer-valued function on $T^{(0)} \times T^{(0)}$.

Lemma 3.6.11. *$(T^{(0)}, d)$ is a metric space.*

Proof. Since any lattice is obviously snugly embedded in itself, it follows from (3.6.3) that $d(s, s) = 0$ for any s. Conversely, if $d(s_0, s_1) = 0$, and if we represent the s_i by lattices Λ_i where Λ_1 is snugly embedded in Λ_0, then Λ_0 and Λ_1 have bases $\{e, f\}$ and $\{e, \beta f\}$ for some $\beta \in \mathcal{O}$ with $v(\beta) = d([\Lambda_0], [\Lambda_1]) = 0$; hence $\beta \in \mathcal{O}^*$, from which it follows that $\Lambda_1 = \Lambda_0$ and hence $s_0 = s_1$.

To prove symmetry we consider two arbitrary elements s_0, s_1 of $T^{(0)}$, which we represent by lattices Λ_0 and Λ_1, where Λ_1 is snugly embedded in Λ_0. Again we choose a basis $\{e, f\}$ of Λ_0 such that e and $f' = \beta f$ form a basis of Λ_1. So $d(s_0, s_1) = \delta(\Lambda_0, \Lambda_1) = v(\beta)$. On the other hand, $\Lambda_0' = \beta\Lambda_0$ also represents s_0, and it has the basis $\{f', \beta e\}$; hence Λ_0' is snugly embedded in Λ_1, and $d(s_1, s_0) = \delta(\Lambda_1, \Lambda_0') = v(\beta)$.

To prove the triangle inequality we consider arbitrary elements v_0, v_1, v_2 of $T^{(0)}$. If Λ_0 is any lattice representing s_0 then successive applications of Lemma 3.6.8 give lattices Λ_1 and Λ_2 representing s_1 and s_2, with Λ_{i+1} snugly embedded in Λ_i for $i = 0, 1$. Now $\Lambda_2 \subset \Lambda_0$, and by Lemma 3.6.10 and (3.6.2) we have

$$d(s_0, s_2) \leqslant \delta(\Lambda_0, \Lambda_2) = \delta(\Lambda_0, \Lambda_1) + \delta(\Lambda_0, \Lambda_2) = d(s_0, s_1) + d(s_1, s_2),$$

where the last equality follows directly from the definition of d. □

Before I can move on to the definition of the tree T_v, I need to establish a couple of other properties of the metric space $T^{(0)}$. Like the last result, they are applications of Lemma 3.6.10.

Lemma 3.6.12. *For any $s_0, s_1, s_2 \in T^{(0)}$ we have*

$$d(s_0, s_2) \equiv d(s_0, s_1) + d(s_1, s_2) \pmod 2.$$

Proof. By successive applications of Lemma 3.6.6, we can represent the s_i by lattices Λ_i with $\Lambda_2 \subset \Lambda_1 \subset \Lambda_0$. By (3.6.2) and the second assertion of Lemma 3.6.10, we find that

$$d(s_0, s_2) \equiv \delta(\Lambda_0, \Lambda_2) = \delta(\Lambda_0, \Lambda_1) + \delta(\Lambda_1, \Lambda_2) \equiv d(s_0, s_1) + d(s_1, s_2) \pmod 2.$$

□

Lemma 3.6.13. *Let s_0 and s_1 be elements of $T^{(0)}$, set $n = d(s_0, s_1)$, and let p and q be nonnegative integers with $p + q = n$. Then there is a unique element s of $T^{(0)}$ such that $d(s_0, s) = p$ and $d(s, s_1) = q$.*

Proof. Let's represent the s_i by lattices Λ_i, with Λ_1 snugly embedded in Λ_0. There are bases $\{e, f\}$ and $\{e, \beta f\}$ of Λ_0 and Λ_1, and we may take $\beta = \pi^n$ where π is a uniformizer and $n = d(s_0, s_1)$. If we now define Λ to be the lattice generated by e and $\pi^p f$, it is clear that Λ is snugly embedded in Λ_0 and that Λ_1 is snugly embedded in Λ_1. By (3.6.9), the element $s = [\Lambda]$ of $T^{(0)}$ has the required properties.

Now suppose that some $s' \in T^{(0)}$ satisfies $d(s_0, s) = p$ and $d(s, s_1) = q$. We wish to prove that $s' = s$. By successive applications of Lemma 3.6.8, there exist a representative Λ' of s which is snugly embedded in Λ_0, and a representative Λ_1' of s_1 which is snugly embedded in Λ. Using the definition of d and (3.6.2), we find that

$$\begin{aligned} d([\Lambda_0], [\Lambda_1']) = n = p + q = d(s_0, s) + d(s', s_1) \\ = \delta(\Lambda_0, \Lambda') + \delta(\Lambda, \Lambda_1') = \delta(\Lambda_0, \Lambda_1'), \end{aligned}$$

which by Lemma 3.6.10 implies that Λ_1' is snugly embedded in Λ_0. It now follows from the uniqueness assertion of Lemma 3.6.8 that $\Lambda_1' = \Lambda_1$. Thus we have $\Lambda_1 \subset \Lambda \subset \Lambda_0$.

The lattices that contain Λ_1 and are contained in Λ_0 are in bijective correspondence with the submodules of $\Lambda_0/\Lambda_1 \cong \mathcal{O}/\pi^n\mathcal{O}$. Since every ideal in $\mathcal{O}$ is generated by a power of π, the only submodules of Λ_0/Λ_1 are those generated by π^k for $0 \leqslant k \leqslant n$. Hence every lattice that contains Λ_1 and is contained in Λ_0 is generated by e and $\pi^k\beta$ for some $k \leqslant n$. In particular Λ' has this form for some k. But then by (3.6.9) we have $k = d(s_0, s) = p$, so $\Lambda' = \Lambda$ and hence $s' = s$. Lemma 3.6.13 is now proved. □

Let's define an abstract simplicial 1-complex $T = T_v$ as follows. The set of vertices of T is $T^{(0)}$. A 1-simplex is an unordered pair (s, s') of vertices such that $d(s, s') = 1$. I'll use the same name T (or T_v) to refer to the geometric realization of this complex; and as is usual in such situations, it will be either be clear from the context which I mean, or it won't matter.

Theorem 3.6.14. *The* 1*-complex* T *is* 1*-connected, i. e. it is a tree.*

Proof. Let s and s' be any two elements of $T^{(0)}$. Set $k = d(s, s')$. By successive applications of the existence assertion of Lemma 3.6.13, we find elements $s = s_0, s_1, \ldots, s_k = s'$ of $T^{(0)}$ such that $d(s_{i-1}, s_i) = 1$ for $i = 1, \ldots, k$. This defines an edge path between the vertices s, s' of T and shows that T is connected.

To show that T is simply connected we must show that for every *reduced* edge path $s_0, \ldots, s_n$ of length $n > 0$ in T we have $s_n \neq s_0$. To say that the edge path is reduced means that in addition to having $d(s_{i-1}, s_i) = 1$ for $i = 1, \ldots, n$, we have $s_{i-1} \neq s_{i+1}$ whenever $0 < i < n$. What I'll prove, by induction on n, is that if $s_0, \ldots, s_n$ is any reduced edge path of length $n > 0$ then $d(s_0, s_n) = n$. For $n = 1$ this is trivial. Now suppose that $s_0, \ldots, s_{n+1}$ is a reduced edge path of length $n+1$, where $n > 0$, and assume that the assertion is true for shorter paths, so that $d(s_0, s_n) = n$ and $d(s_0, s_{n-1}) = n-1$. Since $d(s_n, s_{n+1}) = 1$, the triangle inequality gives $n - 1 \leqslant d(s_0, s_{n+1}) \leqslant n + 1$. By Lemma 3.6.12, we have

$$d(s_0, s_{n+1}) \equiv d(s_0, s_n) + d(s_n, s_{n+1}) = n + 1 \qquad (\text{mod } 2),$$

so we can't have $d(s_0, s_{n+1}) = n$. It remains to rule out the possibility that $d(s_0, s_{n+1}) = n - 1$. Assume that this does hold. Then we have

$$d(s_0, s_{n-1}) = d(s_0, s_{n+1}) = n - 1,$$

$$d(s_{n-1}, s_n) = d(s_{n+1}, s_n) = 1,$$

and

$$d(s_0, s_n) = n.$$

Invoking the uniqueness assertion of Lemma 3.6.13, taking $p = n - 1$, $q = 1$, and letting s_n play the role of s_1, we conclude that $s_{n+1} = s_{n-1}$. But this contradicts the assumption that $s_0, \dots, s_{n+1}$ is a reduced edge path. The proof of Theorem 3.6.14 is now complete.

3.7. The action

Now it's very easy to bring $\mathrm{GL}(2, K)$ into the picture. We can think of an element B of $\mathrm{GL}(2, K)$ as a linear automorphism of $V = K^2$. As such, B maps any lattice onto a lattice, and it obviously maps equivalent lattices to equivalent lattices. So there is a natural action of $\mathrm{GL}(2, K)$ on the set $T^{(0)}$ of homothety classes of lattices. It's also clear that if Λ_1 is snugly embedded in Λ_0 then $B \cdot \Lambda_1$ is snugly embedded in $B \cdot \Lambda_0$, and by (3.6.5) we have

$$d([B \cdot \Lambda_0, B \cdot \Lambda_1]) = \delta(B \cdot \Lambda_0, B \cdot \Lambda_1) = \delta(\Lambda_0, \Lambda_1) = d([\Lambda_0, \Lambda_1]),$$

so that $\mathrm{GL}(2, K)$ acts by isometries on $T^{(0)}$. In particular, each element of $\mathrm{GL}(2, K)$ carries 1-simplices onto 1-simplices, so that we have a natural action of $\mathrm{GL}(2, K)$ on the tree T.

In this chapter I will mostly be using the action of $\mathrm{SL}(2, K)$ on T that comes from restricting the action of $\mathrm{GL}(2, K)$. There are a couple of points to be made about this action of $\mathrm{SL}(2, K)$. First of all, $\mathrm{SL}(2, K)$ acts on T without inversions. This is because if s is any vertex of T, and Λ is any lattice representing s, then for any $B \in \mathrm{SL}(2, K)$ we have $d(s, B \cdot s) \equiv \delta(\Lambda, B \cdot \Lambda) = 0 \pmod 2$ by Lemma 3.6.10 and the definition of δ; in particular we always have $d(s, B \cdot s) \neq 1$, so B can't act as an inversion.

The second point to be made about the action of $\mathrm{SL}(2, K)$ is that the stabilizers of vertices have a very simple description. Let s be any vertex, let B be an element of the stabilizer $\mathrm{SL}(2, K)_s$, and let Λ be a lattice representing s. Then $B \cdot s$ is homothety-equivalent to s, so $B \cdot s = \alpha s$ for some $\alpha \in K - \{0\}$. Since $\delta(\cdot, \cdot)$ is well-defined, we have $2v(\alpha) = \delta(\Lambda, B \cdot \Lambda) = v(\det B) = v(1) = 0$. Hence α is a unit in $\mathcal{O}$, so that $B \cdot \Lambda = \Lambda$. Conversely, if $B \in \mathrm{SL}(2, K)$ leaves Λ invariant, it is obvious that $B \cdot s = s$. Thus $\mathrm{SL}(2, K)_s$ is the stabilizer of Λ. Now the stabilizer of the standard lattice $\mathcal{O}^2$ is the group $\mathrm{SL}(2, \mathcal{O})$ (consisting of all 2×2 matrices of determinant 1 with entries in $\mathcal{O}$). If A is an element of $\mathrm{GL}(2, K)$ such that $A \cdot \mathcal{O}^2 = \Lambda$, we have $\mathrm{SL}(2, K)_s = \mathrm{SL}(2, \mathcal{O})^A$, where exponentiation denotes conjugation. What I've shown is that the stabilizers in $\mathrm{SL}(2, K)$ of the vertices of T are just the conjugates of $\mathrm{SL}(2, \mathcal{O})$ in $\mathrm{GL}(2, K)$.

3.8. Getting to know the tree, I: the link of a vertex

A good starting point for understanding what the tree $T = T_v$ looks like is describing the link of a vertex. Let $s_0 = [\Lambda_0] \in T^{(0)}$ be given. The link of s_0 consists of all elements s of $T^{(0)}$ such that $d(s_0, s) = 1$. Any such s is represented by a unique lattice Λ which is snugly embedded in Λ_0; thus if π denotes a uniformizer in $\mathcal{O}_v$, the lattice Λ_0 has a basis $\{e, f\}$ (depending on s) such that $\{e, \pi f\}$ is a basis of Λ_0. So we have

$$\pi\Lambda_0 \subset \Lambda \subset \Lambda_0. \tag{3.8.1}$$

Now the lattices Λ that satisfy (3.8.1) are in bijective correspondence with submodules of the quotient module $V = \Lambda_0/\pi\Lambda_0$, which we can think of as a 2-dimensional vector space over the residue field $k = k_v = \mathcal{O}_v/\pi\mathcal{O}_v$ of v. If Λ is in fact generated by e and πf for some basis $\{e, f\}$ of λ_0, then the corresponding subspace of V is clearly 1-dimensional; and the converse is easy to deduce from Rappel 3.6.7. So we get a canonical bijection between vertices in the link of s_0 and 1-dimensional subpaces of V.

The set of 1-dimensional subspaces of a 2-dimensional vector space V over k is, by definition, a 1*-dimensional projective space*, or *projective line*, over k. I'll return to projective spaces in Section 5.2. For now let me just observe that if V has a basis we can identify the corresponding projective line with the disjoint union of k with a single element denoted ∞: the subspace spanned by the vector whose coordinates in the basis are a and b is identified with $a/b \in k$ if $b \neq 0$, and with ∞ if $b = 0$.

In the case where k is a finite field with q elements—for example when v is the p-adic valuation of $\mathbf{Q}$ or $\mathbf{Q}_p$ and $q = p$—the set $k \cup \{\infty\}$ has $q + 1$ elements, and hence each vertex of T has valence $q+1$. A good exercise in understanding the tree is to take $q = 2$, so that $k \cong \mathbf{Z}/2\mathbf{Z}$ and T is a trivalent tree, and, starting with an arbitrary vertex, to describe some nearby vertices. Suppose we denote by $[e, f]$ the vertex represented by the lattice generated by a given basis $\{e, f\}$ of F^2. If we write a given vertex in the form $[e, f]$, the vertices in its link are $[e, 2f]$, $[2e, f]$ and $[2e, e + f]$. (The latter vertex could equally well have been written as $[e + f, 2f]$.) Now we can find the three vertices in the link of, say, $[2e, e + f]$ by substituting $2e$ and $e + f$ for e and f in the expression for the vertices in the link of $[e, f]$; doing this directly gives $[2e, 2e + 2f]$, $[4e, f]$ and $[4e, 3e + f]$. However, $[2e, 2e + 2f]$ is simply the original vertex $[e, f]$ under a different name, which makes sense because we already know that $[e, f]$ and $[2e, e + f]$ are joined by an edge. So the two new vertices in the link of $[2e, e+f]$ are $[4e, f]$ and $[4e, 3e+f]$, which you may prefer to rename $[4e, -e + f]$. You can continue in this way and see various interesting new vertices appear at small distances from $[e, f]$.

Since the action of $\mathrm{SL}(2, K)$ on T is simplicial, it restricts to an action of the stabilizer $\mathrm{SL}(2, K)_s$ of any vertex s on the link of s. Up to equivalence, what we are looking at here is an action of $\mathrm{SL}(2, \mathcal{O})$ on the standard 1-dimensional projective space $k\mathrm{P}^1$ over the residue field k. Once we've said that, it's pretty clear what this action should be (again up to equivalence): the quotient homomorphism $\mathcal{O} \to k$ gives rise to a natural homomorphism $q : \mathrm{SL}(2, \mathcal{O}) \to \mathrm{SL}(2, k)$, and $\mathrm{SL}(2, k)$ acts in a natural way on $k\mathrm{P}^1$ because a linear transformation of the vector space k^2 permutes the 1-dimensional subspaces of k^2. (If you identify $k\mathrm{P}^1$ with $k \cup \{\infty\}$ then

$\mathrm{SL}(2,K)$ acts on $k\mathrm{P}^1$ by linear fractional transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \to \frac{az+b}{cz+d} \qquad \text{for } z \in k \cup \{\infty\}$$

as you can check by an easy calculation.) The obvious action of $\mathrm{SL}(2,\mathcal{O})$ on $k\mathrm{P}^1$ is obtained by pulling back this action of $\mathrm{SL}(2,K)$ via q. As is usual in such situations, proving that this natural action really is equivalent to the one obtained by restricting the action of $\mathrm{SL}(2,K)$ on T is just an exercise in keeping track of the definitions.

One consequence of this description of the action of $\mathrm{SL}(2,K)_s$ on the link of s is a description of the stabilizer in $\mathrm{SL}(2,K)$ of an edge of T. If e is an edge with endpoints s_0 and s_1, we can think of the stabilizer $\mathrm{SL}(2,K)_e$ of e as the stabilizer of s_1 within the group $\mathrm{SL}(2,K)_{s_0}$. But under the standard action of $\mathrm{SL}(2,K)$ on $k\mathrm{P}^1$, the stabilizer of a point of $k\mathrm{P}^1$ is conjugate to the group Δ of upper triangular matrices in $\mathrm{SL}(2,K)$. It follows that $\mathrm{SL}(2,K)_e$ is conjugate in $\mathrm{GL}(2,K)$ to $q^{-1}(\Delta) \subset \mathrm{SL}(2,\mathcal{O})$. The latter group can be described directly as consisting of all matrices of the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $c \in \pi\mathcal{O}$ and $a, b, d \in \mathcal{O}$.

This in turn has a neat consequence concerning the commutator subgroup $[\mathrm{SL}(2,K)_e, \mathrm{SL}(2,K)_e]$ of an edge stabilizer $\mathrm{SL}(2,K)_e$, which will come up a couple of times in the applications that I'll talk about later in this chapter. We have $q([\mathrm{SL}(2,K)_e, \mathrm{SL}(2,K)_e]) \subset [\Delta,\Delta]$, and the group $[\Delta,\Delta]$ is made up entirely of upper triangular matrices over k that have 1's on the diagonal. In particular, if e is an edge of T_v we have

$$\operatorname{trace} A \equiv 2 \qquad (\mathrm{mod}\ 2) \qquad \text{for every } A \in [\mathrm{SL}(2,K)_e, \mathrm{SL}(2,K)_e]. \tag{3.8.2}$$

3.9. Getting to know the tree, II: axes

Now that we have a good picture of the stabilizer of a vertex and how it acts on the link of the vertex, let's ask the opposite question: what can we say about the action of an element A of $\mathrm{SL}(2,K)$ that fixes no vertex of T? For simplicity, let's assume that A is diagonalizable over K, so that for some basis $\{e,f\}$ of K^2 we have $A(e) = \alpha e$ and $A(f) = \alpha^{-1} f$. The assumption that A fixes no vertex says that α and α^{-1} do not both belong to $\mathcal{O}$; by symmetry we may assume $\alpha^{-1} \notin \mathcal{O}$, so that $l = v(\alpha) > 0$. Using the notation of Subsection 3.8, let us set $s_n = [e, \pi^n f] \in T^{(0)}$ for every $n \in \mathbf{Z}$. We have $d(s_m, s_n) = |m-n|$ for all $m,n \in \mathbf{Z}$. It follows that there is an edge a_n joining s_n to s_{n+1} for each n, and that the s_n and a_n form a subcomplex L of T which, up to simplicial isomorphism, looks like the real line triangulated so that the integers are the vertices. The definition of the action of $\mathrm{SL}(2,K)$ on T implies that $A \cdot s_n = s_{n+2l}$ for every $n \in \mathbf{Z}$. (In fact, A maps the lattice generated by e and $\pi^n f$ onto the one generated by $\pi^{-l}e$ and $\pi^{n+l}f$.) So the "simplicial line" L is invariant under A, and A acts on L as a translation.

It's a neat combinatorial exercise to prove that if T is any tree and γ is any simplicial automorphism of T which is not an inversion and fixes no vertex of T, then there is a unique subcomplex L of T which is simplicially isomorphic to a line and is invariant under γ; furthermore, γ always acts on L as a translation. You will find this worked out in [55], or in a more general version in [42]. The line L is called the *axis* of γ. So what I've done here is to describe the axis in T of a diagonalizable element of $\mathrm{SL}(2, K)$ (when it exists, i.e. when the element has no fixed point in T).

3.10. Application: Ihara's theorem

To illustrate what can be done with the tree for SL_2, I will give Serre's elegant proof of a result due to Ihara which is the p-adic analogue of a simple fact about discrete subgroups of $\mathrm{SL}(2, \mathbf{R})$. Suppose that $\Gamma \subset \mathrm{SL}(2, \mathbf{R})$ is a discrete, torsion-free group. Then Γ maps injectively to $\mathrm{PSL}(2, \mathbf{R})$, the group of orientation-preserving isometries of the hyperbolic plane $\mathbf{H}^2$; the image is still discrete and of course torsion-free. From this one deduces that the action of Γ on $\mathbf{H}^2$ is free. This is because the stabilizer Γ_z of each point $z \in H^2$ is a discrete subgroup of the compact group $\mathrm{SL}(2, \mathbf{R})_z \cong \mathrm{SO}_2$, and is therefore finite, hence trivial since Γ is torsion-free. The discreteness of Γ also implies that Γ acts properly discontinuously on $\mathbf{H}^2$, so the quotient $\mathbf{H}^2/\Gamma$ is an orientable hyperbolic surface F having $\mathbf{H}^2$ as its universal covering space and Γ as its group of deck transformations. Hence $\Gamma \cong \pi_1(F)$. This shows that every discrete, torsion-free subgroup of $\mathrm{SL}(2, \mathbf{R})$ either is a free group or is isomorphic to the fundamental group of a closed orientable surface of genus $\geqslant 2$.

Ihara's theorem says that in the p-adic world the corresponding result is even simpler: every discrete, torsion-free subgroup of $\mathrm{SL}(2, \mathbf{Q}_p)$ is free! (Here the term "discrete" is to be interpreted in terms of the topology on $\mathrm{SL}(2, \mathbf{Q}_p)$ defined in the obvious way from the metric topology of $\mathbf{Q}_p$.) The proof closely parallels the one for the real case, but it uses the tree T for $\mathrm{SL}(2, \mathbf{Q}_p)$ in place of $\mathbf{H}^2$. First we show Γ acts freely on the set of vertices of T: this is formally identical to the proof in the real case that Γ acts freely on $\mathbf{H}^2$, once one knows that the stabilizer $\mathrm{SL}(2, \mathbf{Q}_p)_s$ is compact for every vertex; but by what we saw in Subsection 3.7, $\mathrm{SL}(2, \mathbf{Q}_p)_s$ is conjugate in $\mathrm{GL}(2, \mathbf{Q}_p)$ to $\mathrm{SL}(2, \mathbf{Z}_p)$; and as we saw in Subsection 3.5 that $\mathbf{Z}_p$ is compact, the group $\mathrm{SL}(2, \mathbf{Z}_p)$ is clearly compact as well. Now that we know that Γ acts freely on the vertices of T, it follows that it acts freely on the whole geometric simplicial complex T because, according to 3.7, there are no inversions. So the quotient T/Γ is a graph G having T as its universal covering space and Γ as its group of deck transformations. (Note that since the action is simplicial this time, proper discontinuity is not even an issue.) Hence $\Gamma \cong \pi_1(G)$, and it follows that Γ is free.

4. Varieties of representations and varieties of characters

I talked in Subsection 1.6 about the nearly canonical representation of the fundamental group of a finite-volume hyperbolic 3-manifold M in $\mathrm{PSL}(2, \mathbf{C})$ or $\mathrm{SL}(2, \mathbf{C})$. It turns out that when M has cusps, this representation can be "deformed" through infinite families of inequivalent representations which can be studied with the techniques of algebraic geometry. The punch line, later in the chapter, is going to be that deforming representations "off to infinity" produces actions of $\pi_1(M)$ on trees, which are defined using the construction of Section 3, and which in turn can be used to define incompressible surfaces in M using the constructions of Section 2.

In this section I'll try to provide a rough introduction to the needed foundational ideas from algebraic geometry, as well as presenting the more specialized material involving representations and hyperbolic manifolds. Although I won't be able to make this section as self-contained as Section 3, I'll try to give a hint of what the required material is about, and to provide references, where necessary, to sources where proofs are presented in an accessible form.

I've decided to present this theory from the point of view taken in [17], [18], [15], and [19], involving $\mathrm{SL}(2, \mathbf{C})$-representations. It has been shown, for example in [6] and [12], that stronger information can sometimes be obtained using $\mathrm{PSL}(2, \mathbf{C})$-representations; but this requires taking a less elementary point of view, and you may have your hands full already.

4.1. The variety of representations

Let Γ be any finitely generated group. We are interested in studying the set $R(\Gamma)$ of all representations of Γ in $\mathrm{SL}(2, \mathbf{C})$. Suppose we fix a finite system of generators of Γ, say $(g_1, \ldots, g_n)$. Then a representation $\rho : \Gamma \to \mathrm{SL}(2, \mathbf{C})$ is uniquely determined by specifying the n-tuple $(\rho_(g_1), \cdots, \rho(g_n))$. Here each $\rho(g_i)$ is a matrix $\begin{pmatrix} w_i & g_i \\ y_i & z_i \end{pmatrix} \in \mathrm{SL}(2, \mathbf{C})$, so we may think of ρ as being determined by the $4n$-tuple $(w_1, x_1, y_1, z_1, \ldots, w_n, x_n, y_n, z_n)$ of complex numbers. This gives a bijective correspondence $\rho \leftrightarrow (\rho(g_1), \cdots, \rho(g_n))$ between $R(\Gamma)$ and some subset of the complex affine space $\mathbf{C}^{4n}$. It will be useful to think of $R(\Gamma)$ as being identified with this subset of $\mathbf{C}^{4n}$ via this correspondence. (Of course this identification depends on choosing a system of generators $(g_1, \ldots, g_n)$ for Γ. I'll return to this issue in Subsection 4.3.)

Now suppose that $(r_j)_{j \in J}$ is a system of defining relators for Γ; here the index set J may be finite or infinite, and each r_j is a word in the generators $g_1, \ldots, g_n$. If $X_1, \ldots, X_n$ are 2×2 matrices, we denote by $r_j(X_1, \ldots, X_n)$ the matrix that's obtained from by substituting X_i for g_i in the word r_j for $i = 1, \ldots, n$. Then a $4n$-tuple $(w_1, \ldots, z_n)$ belongs to the set $R(\Gamma)$ if and only if we have

$$w_i z_i - x_i y_i = 1 \tag{4.1.1}$$

for $i = 1, \ldots, n$, and

$$r_j \left(\begin{pmatrix} w_1 & x_1 \\ y_1 & z_1 \end{pmatrix}, \ldots, \begin{pmatrix} w_n & x_n \\ y_n & z_n \end{pmatrix} \right) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.1.2}$$

for each $j \in J$. For each i the equation (4.1.1) is a polynomial equation in the coordinates of $\mathbf{C}^{4n}$. For each j we can rewrite the matrix equation (4.1.2) as a system of four polynomial equations in the coordinates. To do this, we first rewrite each occurrence of an inverse matrix $\begin{pmatrix} w_i & x_i \\ y_i & z_i \end{pmatrix}^{-1}$ as $\begin{pmatrix} z_i & -x_i \\ -y_i & w_i \end{pmatrix}$ (which is equal to $\begin{pmatrix} w_i & x_i \\ y_i & z_i \end{pmatrix}^{-1}$ in the presence of equations (4.1.1). Then we multiply out the left hand side of (4.1.2) and set each of the four matrix entries of the resulting product equal to the corresponding matrix entry on the right hand side.

This shows that the set $R(\Gamma) \subset \mathbf{C}^{4n}$ is the solution set to some system of polynomial equations in the coordinates of $\mathbf{C}^{4n}$. In general, a subset of an affine space $\mathbf{C}^N$ is called a *(complex affine) algebraic set* if it's the set of zeros of some system of polynomial equations in the coordinates. The set of defining equations that we have exhibited for $R(\Gamma)$ may be infinite, if Γ is not finitely presented; however, one of the first things that one proves in algebraic geometry—a corollary to the Hilbert basis theorem—is that any subset of $\mathbf{C}^N$ which is defined by a possibly infinite system of polynomial equations can actually be defined by some finite subsystem.

4.2. A little algebraic geometry

This is a good place to review a few basic concepts and results concerning algebraic sets in an affine space $\mathbf{C}^N$, where N is a natural number. (The proofs of these facts can be found in any introductory text on algebraic geometry. One book that I have found congenial is [45].) An algebraic set is said to be *reducible* if it can be expressed as the union of two proper algebraic subsets. Irreducible affine algebraic sets are often called *affine varieties.*

Another corollary of the Hilbert basis theorem states that any algebraic set $V \subset \mathbf{C}^N$ is a finite union of irreducible algebraic sets $V_1 \cup \ldots \cup V_k$. Once this has been established, it's obvious that we can choose the V_i so that $V_i \not\subset V_j$ whenever i and j are distinct indices $\leqslant k$. With this restriction, it isn't hard to show that the decomposition $V = V_1 \cup \ldots \cup V_k$ is unique apart from the order of the terms. The V_i are called the *irreducible components* of V. Unlike the connected components of a topological space, the irreducible components of an algebraic set V are not necessarily disjoint from one another. (The algebraic subset of $\mathbf{C}^2$ defined by the equation $zw = 0$, where z and w denote the coordinates, has the coordinate axes $z = 0$ and $w = 0$ as its irreducible components.)

If $V \subset \mathbf{C}^N$ is any algebraic set, the *coordinate ring* $\mathbf{C}[V]$ of V is defined, most concretely, to be the ring of all functions on V which are restrictions of functions on $\mathbf{C}^N$ defined by polynomials in the coordinates. Since $\mathbf{C}[V]$ contains the constant functions we may think of it as an algebra over $\mathbf{C}$, and as such it is generated by the restrictions to V of the coordinate functions on $\mathbf{C}^N$; in particular it is a finitely generated $\mathbf{C}$-algebra. If V is irreducible, it's a simple exercise in using the definitions to show that $\mathbf{C}[V]$ is an integral domain. In this case one denotes by $\mathbf{C}(V)$ the field of fractions of the integral domain $\mathbf{C}[V]$.

If V is an affine variety, any element of $\mathbf{C}(V)$ may be written in the form f/g, where $f, g \in \mathbf{C}(V)$ and $g \neq 0$ (i.e. the function g does not vanish identically on V). The points of V where g takes the value 0 form a proper algebraic subset of V. It is a general fact that any proper algebraic subset of an affine variety V is made up of irreducible components having lower dimension than V, and has a dense complement of V. Thus g is nonzero on an open dense subset U of V. The given element of $\mathbf{C}(V)$ defines a function on U whose value at a point $x = (z_1, \ldots, z_N)$ is $f(z_1, \ldots, z_N)/g(z_1, \ldots, z_N)$.

This leads to an alternative description of the elements of $\mathbf{C}(V)$ as equivalence classes of genuine functions. Each of the functions in question is required to have a domain which is the complement a proper algebraic subset of V, and on this domain it is required to be defined by rational functions in the coordinates of the affine space containing V. Two such functions are equivalent if they agree on the intersection of their domains.

For this reason, the elements of $\mathbf{C}(V)$ are called *rational functions* on V, and $\mathbf{C}(V)$ is referred to as the *function field* of V.

A *polynomial map* between algebraic sets $V \subset \mathbf{C}^M$ and $W \subset \mathbf{C}^N$ is a map $F : V \to W$ which is defined by polynomials in the ambient coordinates. More precisely, F is a polynomial map if there are elements $f_1, \ldots, f_N$ of $\mathbf{C}[V]$ such that $F(x) = (f_1(x), \ldots, f_N(x))$ for every $x \in V$. A priori, of course, if $f_1, \ldots, f_N$ are elements of $\mathbf{C}[V]$ then $(f_1(x), \ldots, f_N(x))$ is only a point of $\mathbf{C}^N$; to say that it lies in W for every $x \in V$ says that the f_i satisfy certain algebraic relations.

It's often useful to think of affine algebraic sets as forming a category, with polynomial maps playing the role of morphisms. In particular we have a natural notion of *isomorphism* of affine algebraic sets. From this point of view, the coordinate ring behaves like (you should excuse the expression) a contravariant functor: if $F : V \to W$ is a polynomial map, then for every $g \in \mathbf{C}[W]$, the function $F \circ g$ belongs to $\mathbf{C}[V]$. (We're just composing polynomials to get another polynomial.) The map $g \mapsto F \circ g$ is a homomorphism of $\mathbf{C}$-algebras from $\mathbf{C}[W]$ to $\mathbf{C}[V]$. It's obvious that if F is surjective—or more generally if it maps V onto a dense subset of W—then the associated homomorphism $\mathbf{C}[W] \to \mathbf{C}[V]$ is injective. So if V and W are irreducible, there is an induced homomorphism (necessarily injective!) of fields of fractions, from $\mathbf{C}(W)$ to $\mathbf{C}(V)$. So when we have fixed a polynomial map of V onto W, where V and W are irreducible, we can think of the field $\mathbf{C}(V)$ as an extension of $\mathbf{C}(W)$.

4.3. More on varieties of representations

Let's return to the study of the set of representations $R(\Gamma)$ of a finitely generated group Γ. I have pointed out that if we fix a system of generators $(g_1, \dots, g_n)$ of Γ defines a bijection, say η, of $R(\Gamma)$ onto an affine algebraic set in $\mathbf{C}^{4n}$. By definition we have $\eta(\rho) = (\rho(g_1), \dots, \rho(g_n))$. Now suppose that $(h_1, \dots, h_m)$ is a second system of generators of Γ, and let $\theta : \rho \mapsto (\rho(g_1), \dots, \rho(g_n))$ denote the corresponding bijection to an algebraic set in $\mathbf{C}^{4m}$. The composition $\theta \circ \eta^{-1}$ is a bijection between algebraic sets. Let's write $h_i = W_i(g_1, \dots, g_n)$ for $i = 1, \dots, m$, where each W_i is a word in n letters; and let's identify a point $(w_1, x_1, y_n, z_n, \dots, w_n, x_n, y_n, z_n)$ of $\mathbf{C}^{4m}$ with an n-tuple $\left(\begin{pmatrix} w_1 & g_1 \\ y_1 & z_1 \end{pmatrix}, \dots, \begin{pmatrix} w_n & g_n \\ y_n & z_n \end{pmatrix}\right)$ of 2×2 matrices, and likewise for $\mathbf{C}^m$. Then the composite bijection $\theta \circ \eta^{-1}$ maps an n-tuple of matrices $(X_1, \dots, X_n)$ to the m-tuple

$$(W_i(X_1, \dots, X_n))_{i=1}^m.$$

Because matrix multiplication and inversion involve only multiplying and adding entries and changing signs, it follows that $\theta \circ \eta^{-1}$ is a polynomial map. The same argument shows that $\eta \circ \theta^{-1}$ is a polynomial map. So the natural bijection between the two algebraic sets incarnating $R(\Gamma)$ is an isomorphism of algebraic sets; this means that the structure of an algebraic set that we have given to $R(\Gamma)$ is really a completely natural one.

I'll ordinarily be identifying $R(\Gamma)$ with an actual algebraic set in an affine space by fixing some set of generators, and the remark I just made says that nothing algebro-geometric about $R(\Gamma)$ really depends on the set of generators. I find it very reassuring to know that, although I don't know if I'll actually use it anywhere in this chapter. On the other hand, there is another remark I need to make about the algebraic set $R(\Gamma)$ which is absolutely fundamental for the mathematics I'll be talking about.

Suppose R_0 is an algebraic subset of $R(\Gamma)$, for example $R(\Gamma)$ itself or an irreducible component. Suppose we fix an element $\gamma \in \Gamma$. Then every $\rho \in R_0$ defines a matrix $\rho(\gamma) \in \mathrm{SL}(2, \mathbf{C})$. Since we are thinking of γ as being fixed, the entries of $\rho(\gamma)$ are determined by the element ρ of R_0; so we can write

$$\rho(\gamma) = \begin{pmatrix} a_\gamma(\rho) & b_\gamma(\rho) \\ c_\gamma(\rho) & d_\gamma(\rho) \end{pmatrix}, \tag{4.3.1}$$

where a, b, c, d are complex-valued functions on R_0 determined by the element γ. Since each $\rho \in R_0$ is a representation, we have $\rho(\gamma\delta) = \rho(\gamma)\rho(\delta)$, i.e.

$$\begin{pmatrix} a_{\gamma\delta}(\rho) & b_{\gamma\delta}(\rho) \\ c_{\gamma\delta}(\rho) & d_{\gamma\delta}(\rho) \end{pmatrix} = \begin{pmatrix} a_\gamma(\rho) & b_\gamma(\rho) \\ c_\gamma(\rho) & d_\gamma(\rho) \end{pmatrix} \begin{pmatrix} a_\delta(\rho) & b_\delta(\rho) \\ c_\delta(\rho) & d_\delta(\rho) \end{pmatrix} \tag{4.3.2}$$

for all $\rho \in R_0$ and $\gamma, \delta \in \Gamma$.

Now if as usual we think of $R(\Gamma)$ as being a concrete set in an affine space by fixing a set of generators for Γ, and if the element γ happens to be a generator, then it is immediate from (4.3.1) that $a_\gamma, b_\gamma, c_\gamma, d_\gamma$ are just the restrictions to R_0 of the four coordinate functions corresponding to that generator. So they belong to the coordinate ring $\mathbf{C}[R_0]$. For an arbitrary element γ, if we write out γ as a word in the generators and repeatedly apply (4.3.2), then again because of the polynomial nature of matrix multiplication and inversion, we conclude that $a_\gamma, b_\gamma, c_\gamma, d_\gamma \in \mathbf{C}[R_0]$ for every $\gamma \in \Gamma$. Notice also that for any $\gamma \in \Gamma$ and any $\rho \in R_0$ we have $a_\gamma(\rho)d_\gamma(\rho) - b_\gamma(\rho)c_\gamma(\rho) = \det \rho(\gamma) = 1$; so $a_\gamma d_\gamma - b_\gamma c_\gamma = 1$ for every $\gamma \in \Gamma$. This means that the matrix $\begin{pmatrix} a_\gamma & b_\gamma \\ c_\gamma & d_\gamma \end{pmatrix}$ is an element of $\mathrm{SL}(2, \mathbf{C}[R_0])$ for every γ. Finally, it follows from (4.3.2)that the map $\mathcal{P} : \Gamma \to \mathrm{SL}(2, \mathbf{C}[R_0])$ defined by

$$\mathcal{P}(\gamma) = \begin{pmatrix} a_\gamma & b_\gamma \\ c_\gamma & d_\gamma \end{pmatrix} \tag{4.3.3}$$

is a homomorphism, i.e. a representation of Γ in $\mathrm{SL}(2, \mathbf{C}[R_0])$. In particular, if R_0 is irreducible and if K denotes the field $\mathbf{C}(R_0)$, we may regard $\mathcal{P}$ as a representation in $\mathrm{SL}(2, K)$.

In [17] Culler and I named $\mathcal{P}$ the *tautological representation*, and topologists working in this area have generally used this term, although a very similar object is sometimes referred to by algebraists as a *universal representation.*

A central theme in this chapter, which first appeared in my joint paper [17] with Culler, is that the representation $\mathcal{P}$ can be used, via the theory that I presented in Section 3, to define actions of Γ on trees. But before I can explain how this works, I need to introduce a little more machinery.

4.4. Varieties of characters

From the beginning of the chapter I've been stressing the theme that one is interested in representations primarily *up to equivalence.* However, the elements of the set $R(\Gamma)$ are representations in $\mathrm{SL}(2, \mathbf{C})$, not equivalence classes of representations, and for some purposes this is a defect. In this subsection I'll show how to parametrize the *characters* of representations of a finitely generated group Γ by points of an affine algebraic set $X(\Gamma)$, much as the representations themselves are parametrized in the way I described in Subsection 4.1. According to Proposition 1.1.1, characters almost classify representations up to equivalence, so the points of $X(\Gamma)$ are very nearly in bijective correspondence with equivalence classes of representations.

We can think of the group $\mathrm{SL}(2, \mathbf{C})$ as acting on $R(\Gamma)$ by conjugation: for any $A \in \mathrm{SL}(2, \mathbf{C})$ and for any representation $\rho \in R(\Gamma)$ we can define $A \cdot \rho = i_A \circ \rho$, where i_A is the inner automorphism $X \mapsto AXA^{-1}$. The equivalence classes of representations are the orbits of this action. Furthermore, the action is algebraic in the sense that the map $(A, \rho) \mapsto A \cdot \rho$ is a polynomial map from $\mathrm{SL}(2, \mathbf{C}) \times R(\Gamma)$ to $R(\Gamma)$, as you can easily check by arguments like the ones I gave in Subsection 4.1.

The fancy-delancey point of view about the algebraic set $X(\Gamma)$ is that it is the quotient of $R(\Gamma)$ by the action of $\mathrm{SL}(2, \mathbf{C})$, *in the category of algebraic sets.* (Because inequivalent representations can sometimes have the same character, it is *not* the quotient in the category of "sets, period.") There is a general theory of quotients under group actions in algebraic geometry, called geometric invariant theory, which certainly subsumes the material I'll be covering in this section. The point of view I'll be presenting here is closely modeled on the point of view that Culler and I used in [17]. I will be adopting this point of view partly because I don't know geometric invariant theory, and partly because I want to show how elementary the material is. At one point I will refer you to [17] for a result that we proved using the "Burnside Lemma," but that is also quite elementary algebra.

To begin with, let's define a function $I_\gamma : R(\Gamma) \to \mathbf{C}$ for each $\gamma \in \Gamma$), by setting $I_\gamma(\rho) = \operatorname{trace} \rho(\gamma)$ for every representation $\rho \in R(\Gamma)$. Using the notation of Subsection 4.3, with $R_0 = R(\Gamma)$, we deduce from (4.3.1) that $I_\gamma(\rho) = a_\gamma(\rho) + d_\gamma(\rho)$; comparing this with (4.3.3), we conclude that

$$I_\gamma = \operatorname{trace} \mathcal{P}(\gamma) \tag{4.4.1}$$

for every $\gamma \in \Gamma$. In particular, I_γ is an element of the coordinate ring $\mathbf{C}[R(\Gamma)]$ for every $\gamma \in \Gamma$.

I'll define the *trace ring* $T(\Gamma)$ to be the sub-ring of $\mathbf{C}[R(\Gamma)]$ generated by all the functions I_γ for $\gamma \in \Gamma$. (By definition the elements of $T(\Gamma)$ are functions that can be expressed as integer polynomials in the I_γ.) The following elementary result provides a finite set of generators for $T(\Gamma)$ as a ring.

Proposition 4.4.2. *Suppose that a group Γ is generated by elements $\gamma_1, \ldots, \gamma_n$. Then the trace ring $T(\Gamma)$ is generated by the elements I_V, where V ranges over all elements of the form $\gamma_{i_1} \ldots \gamma_{i_k}$ with $1 \leqslant k \leqslant n$ and $1 \leqslant i_1 < \ldots < i_k \leqslant n$. (Note that this set of generators of $T(\Gamma)$ has $2^n - 1$ elements.)*

Proof. The proof is based on the identity

$$\operatorname{trace} AB + \operatorname{trace} AB^{-1} = (\operatorname{trace} A)(\operatorname{trace} B), \tag{4.4.3}$$

which holds for all $A, B \in \mathrm{SL}(2, \mathbf{C})$. This identity has a beautiful proof which I learned from Troels Jorgensen. The characteristic polynomial of A is $X^2 - (\operatorname{trace} A) + 1$, so the Cayley-Hamilton theorem says that $A^2 - (\operatorname{trace} A)A + I = 0$, i.e. $A + A^{-1} = (\operatorname{trace} A)I$. Now multiply both sides on the right by B and take traces to get (4.4.3).

We can interpret (4.4.3) in terms of the functions I_γ as saying that for any $\gamma, \gamma' \in \Gamma$ we have

$$I_{\gamma'\gamma} + I_{\gamma'\gamma^{-1}} = I_\gamma I_{\gamma'}; \tag{4.4.4}$$

in fact, if we evaluate both sides of (4.4.4) at a point ρ of $R(\Gamma)$ we get (4.4.3) with $A = \rho(\gamma')$ and $B = \rho(\gamma)$.

Let's denote by T_0 the sub-ring of $T(\Gamma)$ generated by elements of the special form described in the statement of the proposition. Using (4.4.4) we can prove by induction on the length of a word W in the generators $\gamma_1, \dots, \gamma_n$ that $I_W \in T_0$. This will give the conclusion. You can think of the induction as starting at length 0, where the assertion is trivial because I_1 (where 1 means the identity element of Γ) is the constant function 2. Now consider a word W of length $n > 0$, and assume the assertion is true for words of length $< n$. We can assume W is a reduced, since otherwise we can replace it by a shorter word representing the same element of Γ.

Suppose that W' is a word obtained from W by inverting a single letter somewhere in W: that is, W has the form $X\gamma_i^\epsilon Y$ as a word, for some $i \leqslant n$ and $\epsilon = \pm 1$, and $W' = X\gamma_i^{-\epsilon}Y$. (When I say that W has the form $X\gamma_i^\epsilon Y$ *as a word,* the juxtaposition of X, γ_i^ϵ, Y represents concatenation of words and not merely multiplication in the group. In particular, $n = \operatorname{length} W = \operatorname{length} X + \operatorname{length} Y + 1$. Note that the word W' need not be reduced.) I claim that $I_W \in T_0$ if and only if $I_{W'} \in T_0$. This is because we can rewrite I_W and $I_{W'}$ as $I_{YX\gamma_i^\epsilon}$ and $I_{YX\gamma_i^{-\epsilon}}$ in view of the familiar identity $\operatorname{trace} AB = \operatorname{trace} BA$, and then by (4.4.4) we find that

$$I_W + I_{W'} = I_{YX} I_{\gamma^\epsilon}.$$

Since $I_{XY} \in T_0$ by the induction hypothesis, and since $I_{\gamma^\epsilon} = I_\gamma$ is one of the generators of T_0, the claim follows.

Next I claim that we can interchange two successive letters in W without affecting the membership of I_W in T_0; that is, if $W = X\gamma_i^\epsilon\gamma_j^\zeta Y$ as a word, and $W' = X\gamma_j^\zeta\gamma_i^\epsilon Y$, then I_W belongs to T_0 if and only if $I_{W'}$ does. This is because the same argument used to prove my last claim shows that $I_W \in T_0$ if and only if $I_{X(\gamma_i^\epsilon\gamma_j^\zeta)^{-1}Y} = I_{X\gamma_j^{-\zeta}\gamma_i^{-\epsilon}Y}$ belongs to T_0, so this claim now follows from the last one.

Now by repeatedly interchanging successive letters we can replace W by a word which either fails to be reduced or has the form $\gamma_1^{k_1} \dots \gamma_n^{k_n}$ for some $k_1, \dots, k_n \in \mathbf{Z}$. If we assume W to have the latter form then after possibly inverting certain letters we can assume the k_i to be nonnegative. If some k_i is $\geqslant 2$, we can invert a single letter and get a nonreduced word. So we can assume W already has the form $\gamma_1^{k_1} \dots \gamma_n^{k_n}$ with each k_i equal to either 0 or 1. But in this case I_W is by definition a generator of T_0. This proves the proposition. □

Now, given a finitely generated group Γ, let's fix a set of generators $\gamma_1, \dots, \gamma_n$ for Γ. Setting $N = 2^n - 1$, let's index the words of the form $\gamma_{i_1} \dots \gamma_{i_k}$, with $1 \leqslant k \leqslant n$ and $1 \leqslant i_1 < \dots < i_k \leqslant n$, in some order as $V_1, \dots, V_N$. We define a map $t : R(\Gamma) \to \mathbf{C}^N$ by $t(\rho) = (I_{V_1}(\rho), \dots, I_{V_n}(\rho))$. If two points ρ, ρ' in $R(\Gamma)$ have the same image under t, i.e. if $I_{V_i}(\rho) = I_{V_i}(\rho')$, then it follows from Proposition 4.4.2 that $I_\gamma(\rho) = I_\gamma(\rho')$ for every $\gamma \in \Gamma$. By the definition of the I_γ this means that $\operatorname{trace} \rho(\gamma) = \operatorname{trace} \rho'(\gamma)$ for every γ, i.e. that ρ and ρ have the same character.

Conversely, if ρ and ρ' have the same character then in particular $t(\rho) = t(\rho')$. So the points of $t(R(\Gamma))$ are in natural bijective correspondence with the characters of representations of Γ in $\mathrm{SL}(2, \mathbf{C})$, and the map t sends each representation to the point corresponding to its character. From now on I will identify $t(R(\Gamma))$ with the set of characters of representations of Γ, just as I identified the set of representations itself with a subset of C^{4n} in Subsection 4.1.

Whereas it was essentially obvious that $R(\Gamma) \subset \mathbf{C}^{4n}$ was an algebraic set, the corresponding fact for characters requires more work. I will refer you to [17] for a proof, using the "Burnside Lemma," that $t(R(\Gamma)) \subset \mathbf{C}^N$ is an algebraic set. For a still more elementary proof of this, see [29]. From this point I will denote the algebraic set $t(R(\Gamma))$ by $X(\Gamma)$.

Since t maps $R(\Gamma)$ onto $X(\Gamma)$, we have a natural injective homomorphism $J : \mathbf{C}[X(\Gamma)] \to \mathbf{C}[R(\Gamma)]$ by Subsection 4.2. The algebra $\mathbf{C}[R(\Gamma)]$ is generated by the restrictions of the coordinate functions in $\mathbf{C}^N$. The homomorphism J carries the i-th coordinate function to its composition with t, which by the definition of t is just I_{W_i}. So the ring $J(\mathbf{C}[R(\Gamma)])$ is generated by the I_{W_i}. According to Proposition 4.4.2 it follows that $J(\mathbf{C}[R(\Gamma)])$ coincides with the sub-algebra $\mathbf{C}T[R(\Gamma)]$ generated by the functions I_γ for $\gamma \in \Gamma$. In particular, each I_γ is in the image of J, that is, it is obtained from a polynomial function on $X(\Gamma)$ by composition with t.

I'll generally just identify $C[R(\Gamma)]$ with its image under J. This means that each function f on $X(\Gamma)$ is identified with $J(f) = f \circ t$. As a special case, the function on $X(\Gamma)$ from which I_γ is obtained by composition will also be denoted I_γ. In this language we can say that the functions I_γ generate the algebra $\mathbf{C}[X(\Gamma)]$.

4.5. The irreducible component of a discrete faithful character

Having introduced the formalism of the character variety, I can now be precise about the ideas I was waving my hands about in the introduction to this section. Let N be an orientable hyperbolic 3-manifold of finite volume. According to Proposition 1.6.1, a (discrete, faithful) representation $\rho_0 : \pi_1(N) \to \mathrm{PSL}(2, \mathbf{C})$ associated to the hyperbolic structure of N admits a lift $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$. Thinking of $\tilde{\rho}_0$ as a point of $R(\pi_1(N))$, we get an associated point $\chi_0 = t(\tilde{\rho}_0) \in X(\pi_1(N))$. It should be clear at this point that when I talked in the introduction to the section about ways of "deforming $\tilde{\rho}_0$ through inequivalent representations," I was referring to the study of the irreducible component(s) of $X(\pi_1(N))$ containing χ_0.

When N is closed, everything that can be said about the subject is contained in two theorems [65], [66] due to Weil. Whenever N has finite volume, whether or not it is closed, the main result of [66], the "local rigidity theorem," implies that any discrete faithful representation $\tilde{\rho} : \pi_1(N) \to \mathrm{SL}(2, \mathbf{C})$ sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N))$ is equivalent to $\tilde{\rho}_0$ as a representation in $\mathrm{SL}(2, \mathbf{C})$. This is forerunner of the "strong rigidity theorem" later proved by Mostow, and can be easily deduced from it; Mostow's theorem asserts in this context that any two discrete faithful representations of $\pi_1(N)$ in $\mathrm{Isom}(\mathbf{H}^3)$ are equivalent. (See Bonahon's chapter in this volume.)

In the case where N is closed, the results of [65] imply that any representation $\tilde{\rho} : \pi_1(N) \to \mathrm{SL}(2, \mathbf{C})$ sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N)$ is still discrete and faithful. So Weil's results, taken together, say in this particular context that an entire neighborhood of $\tilde{\rho}_0$ in $R(\pi_1(N))$ is contained in a single equivalence class of representations, and hence maps to a point in $X(\pi_1(M))$. This suggests that the image χ_0 of $\tilde{\rho}_0$ should be an isolated point in $X(\pi_1(N))$, and this can in fact be deduced from Weil's results with a little fiddling.

An (irreducible) affine variety is always connected according to [45], GIVE PRECISE REFERENCE, so the isolated point χ must constitute a 0-dimensional irreducible component of $X(\pi_1(N))$, which is of course the only irreducible component containing χ_0. In the informal language of the introduction to the section, $\tilde{\rho}_0$ cannot be deformed through inequivalent representations when N is closed.

The correct generalization of this to the case of a finite-volume manifold which may have cusps is essentially due to Thurston. In one version, it states that if N is an orientable, finite-volume, hyperbolic 3-manifold with n cusps, and if $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ is defined as above, then any irreducible component X_0 of $X(\pi_1(N))$ containing $\chi_0 = t(\rho_0)$ has dimension n. (This is not the strongest known version; more about this at the end of this subsection.)

Thurston's proof of this is divided into two parts. In the first step, which is elementary, ingenious, and essentially algebraic, one shows that any component of $X(\pi_1(N))$ which contains χ_0 must have dimension at least n. The idea is to write down a presentation of $\pi_1(N)$ from which one can deduce that $X = X(\pi_1(N))$, which is realized concretely as an algebraic set in some affine space $\mathbf{C}^r$, can be defined "in a neighborhood of χ_0" by $r - n$ equations. (More precisely, this means that there is a polynomial map $f : \mathbf{C}^r \to \mathbf{C}^{r-n}$ such that $f^{-1}(0) \cap U = X \cap U$ for some neighborhood U of χ_0 in $\mathbf{C}^r$.) This implies the required lower bound on dimension by virtue of general facts about algebraic sets. The only facts used in describing X locally by the right kind of system of equations are that N is homeomorphic to the interior of a compact manifold M whose boundary consists of tori $B_0, \ldots, B_{n-1}$, and that the representation $\tilde{\rho}_0$ is irreducible and maps each of the subgroups $\mathrm{im}(\pi_1(B_i) \to \pi_1(M))$ isomorphically to a group of parabolic elements in $\mathrm{SL}(2, \mathbf{C})$.

You will find an account of this part of the argument in [17] (proof of Proposition 3.2.1).

I want to say a little more about the second part of Thurston's argument because it gives additional information which will be important in this chapter. This part uses hyperbolic geometry, and the key step is an adaptation of the main theorem of [65] to the case of a finite-volume hyperbolic manifold with cusps. In terms of the notation that I just introduced, the relevant result is that if a representation $\tilde{\rho} : \pi_1(N) \to \mathrm{SL}(2, \mathbf{C})$ is sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N)$, *and* if ρ shares with $\tilde{\rho}_0$ the property that it maps each of the groups $P_i = \mathrm{im}(\pi_1(B_i) \to \pi_1(M))$ (defined up to conjugacy, and often called *peripheral subgroups*) onto a group of parabolic elements in $\mathrm{SL}(2, \mathbf{C})$, then ρ is still discrete and faithful. By [66], ρ is then equivalent to ρ_0, if it is close enough to it.

Like the corresponding statement in the closed case, this one is easily translated into a statement about the character variety. Let X^* denote the algebraic subset of $X(\pi_1(N))$ obtained by adjoining the additional equations $I_\gamma^2 = 4$, for all conjugacy classes represented by elements of the subgroups P_i, to the defining equations for $X(\pi_1(N))$. So X^* consists of all characters of representations that send all the peripheral subgroups onto groups of parabolic elements. The translation of the adapted version of Weil's theorem is that χ_0 is an isolated point of X^*.

Now suppose that for each $i \leqslant n$ we fix a nontrivial element γ_i of P_i. (You should think of the elements $\gamma_1, \ldots, \gamma_n$ as being defined up to conjugacy, as the peripheral subgroups are.) It is elementary to see that in a neighborhood of χ_0, the algebraic set X^* is defined by adding the equations $I_{\gamma_i}^2 = 4$, for $i = 0, \ldots, n$, to the defining equations for X. (The main point, at least intuitively, is that if $\rho \in R(\pi_1(N))$ is a representation sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N))$, such that $I_{\gamma_1}, \ldots, I_{\gamma_n}$ vanish at $t(\rho)$, then each $\rho(\gamma_i)$ is a nontrivial element of $\mathrm{SL}(2, \mathbf{C})$ with trace ± 2. Any other element γ of P_i commutes with $\rho(\gamma_i)$ and must therefore also have trace ± 2, so that I_γ vanishes at $t(\rho)$. Again, translating this into the required statement requires a bit of fiddling.)

Now it's a basic fact about complex affine varieties that if a variety X has dimension d then any irreducible component of a subset of X defined by n additional equations has dimension at least $d - n$. (See [56], p. 60, Theorem 7.) If a component X_0 containing χ_0 had dimension $d > n$, then since X_* is defined in the neighborhood of χ_0 by the n additional equations $I_{\gamma_i} = 0$, any component of X_* containing χ_0 would have dimension $d - n > 0$. This is a contradiction since χ_0 is isolated in X^*.

For details, see [18].

This argument gives more information than the statement about dimension that I gave at the outset. Let's summarize what it shows:

Theorem 4.5.1. *Let N be an orientable hyperbolic 3-manifold of finite volume, and n denote the number of its cusps. Let $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ be a lift of a representation $\rho_0 : \pi_1(N) \to \mathrm{PSL}(2, \mathbf{C})$ associated to the hyperbolic structure of N, set $\chi_0 = t(\tilde{\rho}_0) \in X(\pi_1(N))$, and let X_0 be an irreducible component of $X(\pi_1(N))$ containing χ_0. Then $\dim X_0 = n$. Furthermore, if $B_0, \ldots, B_{n-1}$ are the boundary components of a compact core of N, if γ_i is an element whose conjugacy class is carried by B_i for $i = 0, \ldots, n-1$, then χ_0 is an isolated point of the algebraic subset*

$$X^* = \{\chi \in X_0 : I_{\gamma_1}^2 = \cdots = I_{\gamma_n}^2 = 4\}$$

of X_0. □

This is the theorem I will be quoting in later sections. I should point out that stronger statements are generally believed to be true: that χ_0 is a smooth point of $X(\pi_1(N))$—so that in particular there is only one irreducible component of $X(\pi_1(N))$ containing χ_0—and that in a neighborhood of χ_0, the functions $I_{\gamma_1}, \ldots, I_{\gamma_i}$ form a system of local coordinates for X_0. This last statement could be used to simplify very slightly some of the arguments that I'll be giving later on.

However, as this goes to press, it is not clear to me whether references for these stronger statements are available.

I'll conclude this subsection with a simplified statement of a special case of Theorem 4.5.1 which will come up in several applications. If $n = 1$, so that X_0 is a curve, and if we set $\gamma = \gamma_1$, saying that χ_0 is an isolated point of $X^* = X_0 \cap I_\gamma^{-1}\{\pm 2\}$ boils down to saying that the polynomial function I_γ is nonconstant on X_0. So we may state:

Corollary 4.5.2. *Let N be an orientable one-cusped hyperbolic 3-manifold of finite volume. Then there is a 1-dimensional irreducible component X_0 of $X(\pi_1(N))$, containing the character of the lift of a representation associated to the hyperbolic structure of N, such that if γ is any nontrivial element of $\pi_1(N)$ carried by the boundary of the compact core of N, the function I_γ is nonconstant on X_0.* □

5. Ideal points and trees

In this section I'll be talking about a construction that was first introduced by Culler and me in [17]. It gives a way of associating actions of a finitely generated group Γ on trees with "ideal points" (see Subsection 5.2 below) of a curve in the character space $X(\Gamma)$. This construction turns out to have various applications to 3-manifold theory, because $\mathrm{SL}(2, \mathbf{C})$-representations of π_1 of a connected 3-manifold are related to hyperbolic structures on the manifold (Subsection 1.6), whereas actions of π_1 on trees are related to essential surfaces (Subsection 1.5 and Section 2). All the subsequent sections of the chapter will depend in some way on the construction I'll be describing here.

The construction depends on a little more background material from algebraic geometry than was used in Section 4.

5.1. Some more algebraic geometry

A point P of an algebraic set V is said to be *smooth of dimension d*, where $0 \leqslant d \leqslant N$, if P has a neighborhood U in $\mathbf{C}^N$ with the property that $V \cap U = Z \cap U$, where Z is the solution set of a system of $N - d$ polynomial equations

$$f_1(z_1, \ldots, z_N) = \ldots = f_{N-d}(z_1, \ldots, z_N) = 0,$$

the f_i being polynomials in the coordinates $z_1, \ldots, z_N$, such that the $N \times (N - d)$-matrix of partial derivatives

$$(\frac{\partial f_j}{\partial z_i})_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant N-d}$$

at the point P is of rank $N - d$.

I pointed out in Subsection 4.2 that the irreducible components of an affine algebraic set are not in general disjoint from one another. However, it is not hard to show that a point lying in the intersection of two distinct components is not smooth in V, so a smooth point of V does lie in a unique irreducible component.

If V is an affine variety in $\mathbf{C}^N$, there is a unique natural number $d \leqslant N$, called the dimension of V, such that V has a dense subset consisting of smooth points of dimension d. If x is any smooth point of V, we can apply the complex implicit function theorem–which looks formally just like the real implicit function from advanced calculus, and can be deduced from it—to parametrize the points in some neighborhood of x by d complex coordinates. This gives the set of smooth points of d the structure of a complex submanifold of dimension d in $\mathbf{C}^N$.

An affine variety V of dimension 1 is called, naturally enough, an affine curve. In the 1-dimensional case, the existence of a local complex coordinate near a smooth point of V has especially nice consequences, for example for the study of the function field $\mathbf{C}(V)$. Using a local coordinate we can identify a neigborhood U of a smooth point $P \in V$ with a domain in $\mathbf{C}$ in such a way that $P = 0$. If we write a given element of $\mathbf{C}(V)$ in the form f/g, where $f, g \in \mathbf{C}[V]$, then the restrictions of $f, g \in \mathbf{C}[V]$ to U are holomorphic functions f_U and g_U. The quotient f_U/g_U is a meromorphic function, and takes a well-defined value in the Riemann sphere $\mathbf{C} \cup \infty$ at every point in its domain. More explicitly, if $f_U(z) = z^m F(z)$ and $g_U(z) = t^n G(z)$, where F and G are holomorphic and are nonzero at 0, then

$$\frac{f_U(z)}{g_U(z)} = z^{m-n}\frac{F(z)}{G(z)},$$

where F and G are holomorphic and nonzero. It's easy enough to show that the meromorphic function f_U/g_U is well-defined, i.e. does not depend on the way we wrote the given element of $\mathbf{C}(V)$ as a quotient. Its value at 0, which we may think of as the value of f/g at P, is of course $F(0)/G(0)$ if $m \geqslant n$ and ∞ if $m > n$.

In Subsection 4.2 I pointed out that an element of $\mathbf{C}(V)$ defines a natural complex-valued function on an open dense subset of V. What we are seeing here is a partial improvement of this: an element of $\mathbf{C}(V)$ defines a natural function on the set of all smooth points of V, although this function now takes values in $\mathbf{C} \cup \{\infty\}$. In any case, the value we have assigned to f/g at P is indeed the value in any reasonable sense; for example, it is the limit of $f(z)/g(z)$ as z approaches P through $U - \{P\}$.

The number $m - n$ that appeared in the discussion above is the *order* of f_U/g_U in the sense of Subsection 3.3. Now I pointed out in 3.3 that there is a valuation of the field of meromorphic functions on U which assigns to each function its order at 0. For the same reason, there is a valuation of the function field $\mathbf{C}(V)$ that assigns to each element f/g of $\mathbf{C}(V)$ the order at P of the corresponding function f_U/g_U on U.

For any point P of an arbitrary affine variety V, it is tempting to try to assign a value at P to any element of $\mathbf{C}(V)$, but this is not always possible. For example, if $V = \mathbf{C}^2$ (an affine space which we may think of as an algebraic subset of itself, defined by the empty set of equations), and if z and w denote the coordinates on $\mathbf{C}^2$, then the element z/w of $\mathbf{C}(V)$ defines a function on the complement of the line L defined by $w = 0$; furthermore, this function can be extended continuously to a map $\mathbf{C}^2 - \{(0,0)\} \to \mathbf{C} \cup \{\infty\}$ by giving it the value ∞ at every point of $L - \{(0,0)\}$. However, the function cannot be extended continuously to $(0,0)$, because it can take arbitrary limiting values through a sequence in $\mathbf{C}^2 - \{0,0\}$ which approaches $(0,0)$; for example, along the complement of $\{(0,0)\}$ in a line $z = \lambda w$, its value is identically equal to λ.

This problematic behavior can also occur at nonsmooth points of affine curves. For example, the two-variable polynomial $z^3 + w^3 + zw$ is easily seen to be irreducible, from which it follows by general principles that its zeros form an irreducible algebraic set of codimension 1, hence a curve, $V \subset \mathbf{C}^2$. Since the polynomial defining V is closely approximated by zw near $(0,0)$, one can show that the intersection of V with a suitable neighborhood of $(0,0)$ is made up of two "branches:" these are complex analytic 1-manifolds that are tangent to the coordinate axes $z = 0$ and $w = 0$. Consider the element z/w of $\mathbf{C}(V)$, where z and w now denote the generators of $\mathbf{C}[V]$ obtained by restricting the coordinate functions to V. The "rational function" z/w defines a genuine function on $V - \{(0,0)\}$, but this function approaches 0 as the argument approaches $(0,0)$ through the branch tangent to $z = 0$, and approaches ∞ as the argument approaches $(0,0)$ through the branch tangent to $w = 0$.

5.2. Projective varieties

Affine varieties are often awkward to work with because they are noncompact. One can typically learn much more about an affine variety V by studying a *projective completion* of V.

Recall that for a positive integer N, the complex projective N-space $\mathbf{CP}^N$ is the quotient of the $\mathbf{C}^{N+1} - \{(0,\dots,0)\}$ under the equivalence relation $\sim$ in which $(Z_0,\dots,Z_N) \sim (W_0,\dots,W_N)$ if and only if there is a complex number $\alpha \neq 0$ such that $W_i = \alpha Z_i$ for $i = 0,\dots,N$. I'll denote the equivalence class of $(Z_0,\dots,Z_N)$ by $[Z_0,\dots,Z_N]$. One says that $Z_0,\dots,Z_N$ are *homogeneous coordinates* for $[Z_0,\dots,Z_N]$.

If f is a homogeneous complex polynomial of degree $d \geqslant 0$ in $N+1$ indeterminates, then for any point $(Z_0,\dots,Z_N)$ of $\mathbf{C}^{N+1}$ and any $\alpha \in \mathbf{C}$ we have $f(\alpha Z_0,\dots,\alpha Z_N) = \alpha^d f(Z_0,\dots,Z_N)$. This means that although f does not have a well-defined value at a given point of $\mathbf{CP}^N$, it does have a well-defined set of zeros. A *projective algebraic set* in $\mathbf{CP}^N$ is the set of common zeros of a collection of homogeneous polynomials—of various degrees—in $N+1$ indeterminates. Some of the basic properties of affine algebraic sets have straightforward analogues for the projective case. Thus any projective variety can actually be defined as the zero set of a *finite* collection of homogeneous polynomials, and can be represented as a finite union of projective *varieties*, i.e. irreducible projective algebraic sets. (The definition of reducibility in the projective setting looks formally just like the affine definition.)

Let $H_0 \subset \mathbf{CP}^N$ denote the locus of zeros of the coordinate function Z_0, which we can think of as a first-degree homogeneous polynomial. The map $J_0 : \mathbf{C}^N \to \mathbf{CP}^N$ defined by $J_0(z_1, \ldots, z_N) = [1, z_1, \ldots, z_N]$ is a diffeomorphism of $\mathbf{C}^N$ onto the open dense subset $\mathbf{CP}^N - H_0$ of $\mathbf{CP}^N$, with inverse given by

$$[Z_0, Z_1, \ldots, Z_N] \mapsto (\frac{Z_1}{Z_0}, \ldots, \frac{Z_N}{Z_0}).$$

If V is any affine algebraic set in $\mathbf{C}^N$ then the closure $\overline{J_0(V)}$ is a projective algebraic set in $\mathbf{CP}^N$. If V is irreducible, so is $\overline{J_0(V)}$. (Sometimes we can find defining equations for $\overline{J_0(V)}$ by "homogenizing" equations for V: thus if $n = 2$ and V is defined by $z_1^2 + z_2^3 = 1$, we can define $\overline{J_0(V)}$ by $Z_1^2 + Z_2^3 = Z_0^3$. However, this will not always work. To give a trivial example, if we define $\emptyset \subset \mathbf{C}^2$ by the equations $z_1 = 1, z_1 = 2$, homogenizing gives the equations $Z_1 = Z_0, Z_1 = 2Z_0$. The solution set of the latter system consists of the point $[1, 1, 2]$, whereas $\overline{J_0(\emptyset)} = \emptyset$. What is always true is that if we homogenize an arbitrary system of equations defining V then $\overline{J_0(V)}$ is a union of irreducible components of the locus of zeros of the resulting homogeneous system.)

Because $\mathbf{CP}^N$ is obviously compact, projective varieties are always compact; this makes them more tractable objects than affine varieties for some purposes. If $V \subset \mathbf{C}^N$ is an affine variety, we can think of $\overline{J_0(V)}$ as a compactification of V, from which V can be "recovered" since $J_0(V) = \overline{J_0(V)} \cap J_0(C^N)$. This is often a useful way of getting information about an affine variety.

On the other hand, we can also use affine varieties to study projective ones. If we're looking at a projective variety $W \subset \mathbf{CP}^N$, we can assume that W is not contained in any of the "hyperplanes" $H_i \subset \mathbf{CP}^N$ defined by $Z_i = 0$, for $i{-}0, \ldots, N$, because otherwise we could think of W as a variety in a lower-dimensional projective space. Now if $J_0 : \mathbf{C}^N \to \mathbf{CP}^N - H_0$ is defined as above, and if set not contained in H_0, then $V_0 = V \cap (\mathbf{CP}^N - H_0)$ is "identified" via J_0 with the set $J_0^{-1}(W)$; it's not hard to show that $J_0^{-1}(W)$ is an affine variety in $\mathbf{C}^N$. But for $i = 0, \ldots, n-1$, we can do exactly the same construction using H_i and a similarly defined map $J_i : \mathbf{C}^N \to \mathbf{CP}^N - H_i$ in place of H_0 and J_0, to get an "identification" of the set $V_i = V \cap (\mathbf{CP}^N - H_i)$ with an affine variety. We have $V = \cup_{i=0}^N V_0$, and we can think of the V_i as domains of an "atlas of affine coordinate charts;" in terms of these, a projective variety is something that looks "locally" like an affine variety, in much the way that a differentiable manifold looks locally like $\mathbf{R}^n$.

The "transition maps" are easily understood in the setting of a projective variety. To simplify the notation a bit, let's consider a point P lying in the intersection of the chart domains V_0 and V_1. Suppose that P, regarded as a point of the affine variety V_0, has affine coordinates $z_1, z_2 \ldots, z_N$. Then $1, z_1, z_2, \ldots, z_N$ are homogeneous coordinates for P as a point of $V \subset \mathbf{CP}^N$. It follows that if we regard P as a point of V_1, its affine coordinates are $1/z_1, z_2/z_1, \ldots, z_N/z_1$. The same calculation shows that for any i and j, the "transition maps" relating affine coordinates in V_i to those in V_j are defined by rational functions in the coordinates.

In view of the description of the function field as a set of equivalence classes of functions, which I gave in Subsection 4.2, it now follows that there is a natural isomorphic identification of all the function fields $\mathbf{C}(V_0), \ldots, \mathbf{C}(V_N)$ with one another. So it makes sense to talk about the *function field* $\mathbf{C}(W)$ of a projective variety W.

This is a first example of how one can turn "local" definitions involving affine varieties into "global" ones involving projective varieties, in close analogy with the theory of differentiable manifolds. There are of course many other examples. A point of a projective variety is *smooth* if it is identified with a smooth point of an affine variety under one of the affine charts; for a point lying in more than one chart domain, this is independent of the choice of chart, as a calculation with the transition maps shows. Likewise, a projective variety has a *dimension* which is equal to the dimension of each of its affine pieces.

The results about smooth points of affine curves that I discussed in Subsection 5.1 are readily translated into the projective context: if W is a projective curve, i.e. a projective variety of dimension 1, and if P is a smooth point of W, then every element of $\mathbf{C}(W)$ has a well-defined value at P, this value being an element of $\mathbf{C} \cup \{\infty\}$; and furthermore, P gives rise to a valuation of $\mathbf{C}(W)$ in a natural way.

5.3. Canonical completions

In Subsection 5.2 I talked about *the* completion of an affine algebraic set, because I was thinking of such sets concretely as subsets of particular affine spaces. As with so many other kinds of mathematical objects, it is often useful to have the flexibility that comes from thinking of affine algebraic sets as being "defined up to isomorphism." In this context it is no longer permissible to speak about *the* completion of an affine algebraic set, because isomorphic affine algebraic sets may have nonisomorphic projective completions. Actually, if you're paying close attention you'll have noticed that I haven't defined isomorphism of projective varieties, but there are examples where projective completions of isomorphic affine varieties are not even homeomorphic.

Let me point out an especially trivial example of this phenomenon. The union of two "parallel lines" in $\mathbf{C}^2$, say $z_1 = 0$ and $z_1 = 1$, is isomorphic as an affine algebraic set to the union of two "skew lines" in $\mathbf{C}^3$, say $z_1 = z_2 = 0$ and $z_2 = z_3 = 1$. (It's a good exercise to write down the polynomial maps between $\mathbf{C}^2$ and $\mathbf{C}^3$ that restrict to an isomorphism and its inverse.) On the other hand, the completion of the first set in $\mathbf{CP}^2$ is the union of the "projective lines" $Z_1 = 0$ and $Z_1 = Z_0$, which are topological 2-spheres meeting in the point $(0, 0, 1)$, whereas the completion of the second set is the union of the projective lines $Z_1 = Z_2 = 0$ and $Z_2 = Z_3 = Z_0$, which are again topological 2-spheres but are disjoint (since a point of $\mathbf{CP}^3$ cannot have all its homogeneous coordinates equal to 0). Note that in this example the completion of the second set is smooth, whereas the completion of the first set is not.

From this point on I will be using the phrase "completion of V," where V is an algebraic set in some $\mathbf{C}^N$, to mean any projective variety of the form $\overline{J_0(V')}$, where

V' is an arbitrary affine algebraic set in some $\mathbf{C}^{N'}$ isomorphic to V and $J_0 : \mathbf{C}^{N'} \to \mathbf{CP}^{N'}$ is the standard embedding defined in Subsection 5.2. In situations where a particular completion $\hat{V}$ of V has been fixed, I will regard V as being identified in the obvious way with a subset of $\hat{V}$. In this situation I will often refer to the points of $\hat{V} - V$ as *ideal points* of the completion $\hat{V}$; in contrast the points of $V \subset \hat{V}$ may be referred to as *ordinary points*.

The example I just described shows that a 1-dimensional algebraic set may have one completion in which the ideal points are smooth, and another which fails to have this property. It can be shown that every affine curve has a completion in which all the ideal points are smooth. This is in fact a fairly direct consequence of one of the basic results in the theory of algebraic curves, which allows one to "resolve the singularity" at a nonsmooth point of a projective algebraic curve, replacing it by a finite number of smooth points. In more precise terms, if x is a singular point of a projective curve C, there exist a projective curve $\tilde{C}$ and a well-defined map $J : \tilde{C} \to C$, which is rational in local affine coordinates near every point, such that $J^{-1}(y)$ is a single point for every $y \neq x$, and $J^{-1}(x)$ consists of a finite number of smooth points. In [45] and [28] you will find a total of three quite different proofs of this result, all very enlightening. Now if $\hat{V}$ is any completion of an affine curve V, one can resolve all those singularities of $\hat{V}$ which occur at ideal points, and it is not hard to show that the resulting projective curve is still a completion of V.

Although I will not be defining the notion of isomorphism of projective varieties in this chapter, because I won't really need it, I ought to mention that up to isomorphism there is only one completion of an affine curve in which all the ideal points are smooth. Once one has studied the definitions, the proof that such a completion is canonical in this sense is a simple application of the fact, which I talked about above, that if W is a projective curve, every element of $\mathbf{C}(W)$ has a well-defined value in $\mathbf{C} \cup \{\infty\}$ at every smooth point of W.

5.4. Associating an action on a tree with an ideal point

Let Γ be a finitely generated group, and let C be a curve contained in $X(\Gamma)$, i.e. an irreducible 1-dimensional subvariety of $X(\Gamma)$. By Subsection 5.3, there is a projective completion $\hat{X}$ of X such that every ideal point of $\hat{X}$ is smooth. I will show how every ideal point x of $\hat{X}$ gives rise to a nontrivial action of Γ on a tree.

First of all, by the very way I defined $\mathbf{C}(\hat{X})$ in Subsection 5.2, there is a natural isomorphism of $\mathbf{C}(\hat{X})$ with $\mathbf{C}(X)$. From now on I'll write $F = \mathbf{C}(X) = \mathbf{C}(\hat{X})$. Any ideal point x of $\hat{X}$, because it is a smooth point, determines a valuation v_x of $F = \mathbf{C}(\hat{X})$ by the construction described in Subsections 5.1 and 5.2.

On the other hand, there is an irreducible subvariety R_C of $R(\Gamma)$ such that $t(R_C) = C$. (The point here is that since C is a subvariety of $X(\Gamma) = t(R(\Gamma))$, it is true for very general algebro-geometric reasons that C is the closure of $t(R_C)$ for some subvariety R_C of $R(\Gamma)$. The argument I alluded to in Subsection 4.4, based on the "Burnside Lemma," which shows that $t(R(\Gamma)) \subset \mathbf{C}^N$ is an algebraic set, also shows that $t(R_C)$ is an algebraic set, hence closed, hence equal to C. I'll have to refer you to [17] for details, but that's the philosophy.) As in Subsection 4.2, we can regard $K = \mathbf{C}(R_C)$ as an extension of the field $F = \mathbf{C}(X)$. We invoke the following extension theorem for valuations:

Theorem 5.4.1. *Let K be a finitely generated extension of a field F and let $v : F^* \to \mathbf{Z}$ be a valuation of F. Then there exist an integer $d > 0$ and a valuation $w : K^* \to \mathbf{Z}$ such that $w|F^* = dv$.*

This result is pretty well-known and elementary. The best reference I can give you is to my joint paper [2] with Roger Alperin, where we state it as Lemma 1.1 and give a proof that's self-contained except for a reference to Bourbaki's Commutative Algebra. If you read the proof and look up the reference, you'll know about as much valuation theory as you need for this subject. The result is an extension theorem in the sense that the function $\frac{1}{d}w$ is an extension of v to K^*, and is a valuation in a very slightly more general sense than the one I have defined here: it takes its values in the infinite cyclic group $\frac{1}{d}\mathbf{Z}$ rather than in $\mathbf{Z}$.

We can apply this theorem in our situation because K, being the function field of a (finite-dimensional) variety R_C, is finitely generated as an extension of $\mathbf{C}$; in fact, if C lives in an affine space $\mathbf{C}^N$ then the restrictions of the coordinate functions to R_C generate K over $\mathbf{C}$. So in particular K is finitely generated as an extension of F. The theorem gives a valuation w of F such that $w|F^* = dv_x$ for some $d > 0$. With the valuation w, as in Section 3, we can associate a tree $T = T_w$ on which $\mathrm{SL}(2, K)$ acts in a natural way. On the other hand, by Section 4 we have a tautological representation $\mathcal{P} : \Gamma \to \mathrm{SL}(2, K)$, and we can pull back the action of $\mathrm{SL}(2, K)$ on T via $\mathcal{P}$ to get an action of Γ on T. According to Subsection 3.7, Γ acts without inversions on T.

For any $\gamma \in \Gamma$, the function I_γ is an element of $\mathbf{C}(X)$. Since the ideal point x is smooth, it follows from what I said in Subsection 5.2 that I_γ has a well-defined value $I_\gamma(x) \in \mathbf{C} \cup \infty$ at x. The most important property of the action of Γ on the tree T is:

Property 5.4.2. *For any element $\gamma \in \Gamma$ the following statements are equivalent:*

(i) $I_\gamma(x) \in \mathbf{C}$, i.e. I_γ does not have a pole at x;

and

(ii) Some vertex of T is fixed by γ.

To prove this, first recall that by (4.4.1) the element I_γ of $F \subset K$ is the trace of $\mathcal{P}(\gamma) \in \mathrm{SL}(2, K)$. We have

$$\begin{aligned} I_\gamma(x) \in \mathbf{C} \iff v(I_\gamma) \geqslant 0 \iff w(\operatorname{trace} \mathcal{P}(\gamma)) \geqslant 0 \\ \iff \operatorname{trace} \mathcal{P}(\gamma) \in \mathcal{O}, \end{aligned} \tag{5.4.3}$$

where $\mathcal{O} \subset K$ is the valuation ring defined by the valuation w. Now if (ii) holds, i.e. if $\mathcal{P}(\gamma) \in \mathrm{SL}(2, K)$ fixes a vertex of T, then by Subsection 3.7 the element $\mathcal{P}(\gamma)$

lies in a conjugate, within $\mathrm{GL}(2, K)$, of $\mathrm{SL}(2, \mathcal{O})$. In particular, $\operatorname{trace} \mathcal{P}(\gamma) \in \mathcal{O}$, so that (i) holds by virtue of (5.4.3). To prove the converse we need the *rational canonical form* of a matrix in $\mathrm{SL}(2, K)$, which in this case is a pretty trivial matter. Suppose that (i) holds, so that $\operatorname{trace} \mathcal{P}(\gamma) \in \mathcal{O}$. If we are in the degenerate case where $\mathcal{P}(\gamma) = \pm I$ then γ acts trivially on the whole tree T. In the nondegenerate case we can choose a vector $e \in K^2$ such that e and its image f under the linear transformation $A = \mathcal{P}(\gamma)$ of K^2 are linearly independent. In the basis $\{e, f\}$, the linear matrix of the linear transformation A has the form

$$B = \begin{pmatrix} 0 & 1 \\ c & d \end{pmatrix},$$

so that A is conjugate to B in $\mathrm{GL}(2, K)$. In particular we have $c = -\det A = -1$ and $d = \operatorname{trace} A \in \mathcal{O}$, so that $B \in \mathrm{SL}(2, \mathcal{O})$. Thus $A = \mathcal{P}(\gamma)$ lies in a conjugate, within $\mathrm{GL}(2, K)$, of $\mathrm{SL}(2, \mathcal{O})$, and by Subsection 3.7, γ fixes a vertex of T.

An important consequence of Property 5.4.2 is the following property of the action:

Property 5.4.4. *The action of* Γ *on* T *is nontrivial. (*Recall from Subsection 1.5 that this means that no vertex of T is fixed by the entire group Γ.*)*

To prove this, note that if the action were trivial, then by Property 5.4.2 we would have $I_\gamma(x) \in \mathbf{C}$ for every element γ of Γ. In terms of the concrete description of the character variety that I gave in Subsection 4.4, each of the coordinate functions of $C \subset X(\Gamma)$ is of the form I_γ for some $\gamma \in \Gamma$. But some coordinate function must take the value ∞ at x since x is an ideal point. This contradiction completes the proof of Property 5.4.4.

5.5. More about the action

In some of the applications that I discuss in this chapter I will need an apparently more general version of Property 5.4.2: a finitely generated subgroup Γ_1 of Γ fixes a vertex of T if and only if I_γ takes a finite value at x for every $\gamma \in \Gamma$. My favorite way to prove this is to notice that it follows immediately from Property 5.4.2 itself and the following result:

Proposition 5.5.1. *Suppose that a finitely generated group* Γ *acts without inversions on a tree* T*, in such a way that each element* $\gamma \in \Gamma$ *fixes some vertex* v_γ *of* T*. Then there is a single vertex* v *of* T *which is fixed by the entire group* Γ*: thus* $\gamma \cdot v = v$ *for every* $\gamma \in \Gamma$*.*

You may find a proof of this in [55], but it's much more fun to do it as an exercise.
□

Property 5.4.2 also admits a generalization in a different direction. This version involves the notion of the *length function* associated to an action of a group Γ on a tree T. The length function l associated to an action $\cdot$ (without inversions) is defined by $l(\gamma) = \min_s d(s, \gamma \cdot s)$, where s ranges over the vertices of T. I mentioned in Subsection 3.8 that any $\gamma \in \Gamma$ either has a fixed point in T—in which case $l(\gamma) = 0$—or has a unique invariant line (an "axis") on which it acts by a translation; in this case it is not hard to show that $l(\gamma)$ is the (integer) distance through which γ translates vertices on its axis.

The second generalization of 5.4.2, involving the length functions l associated to the action of Γ on T, is that for any $\gamma \in \Gamma$, the length $l(\gamma)$ is equal to twice the order of the pole of I_γ at the ideal point x. (Here "the order of the pole" is taken to mean 0 if I_γ does not have a pole at x, so that 5.4.2 indeed appears as a special case.) This is also an excellent exercise. The only reference I know for it is [42], where it is proved in a much more general form.

One consequence of this generalization of 5.4.2 is that the length function defined by the action of Γ on $T = T_w$ is canonically associated to the ideal point x, i.e. does not depend on the choice of the extension w.

The length function associated to an action on a tree plays the same role as the character associated to a representation in $\mathrm{SL}(2, \mathbf{C})$, and there is an analogue of Proposition 1.1.1. To understand the statement, first note that according to Proposition 5.5.1, an action of a finitely generated group Γ on a tree is trivial if and only if the associated length function is 0. It's not hard to show that if Γ acts nontrivially on T then there is a unique minimal Γ-invariant subtree of T. (It can be described as the union of the axes of all the elements of Γ that do not have fixed vertices.) It's easy to see that if two actions of Γ on trees have minimal invariant subtrees that are equivariantly simplicially isomorphic, then they give rise to the same length function. The analogue of Proposition 1.1.1 gives a converse that is valid except for certain "degenerate" actions that are analogous to reducible representations.

An action of Γ on a tree is termed *abelian* if its length function has the form $l(\gamma) = |h(\gamma)|$ where h is a homomorphism from Γ to $\mathbf{Z}$; any function of this form does arise from an action of Γ on the tree $\mathbf{R}$, triangulated so that its vertex set is $\mathbf{Z}$. The following result is a special case of results proved in [16] and in [1]:

Proposition 5.5.2. *Let Γ be a finitely generated group. If two nonabelian actions of Γ on trees T and T' define the same length function, then the minimal Γ-invariant subtrees of T and T' are Γ-equivariantly simplicially isomorphic.*

In particular, for the case of the action of Γ on the tree T associated to an ideal point by the construction I described above, the restriction of the action to the minimal Γ-invariant subtree of T is something canonically defined by the given ideal point—except in the degenerate case where the action is abelian. I won't be using this fact anywhere in the rest of this chapter, but as Golde and Tevye said, it's nice to know.

I'll conclude this section by pointing out one more property of the action of Γ on the tree T associated to an ideal point x, which is just the translation of (3.8.2) into this context. (Check it.) This property and its variants come up a lot in applications to 3-manifolds.

Property 5.5.3. *If e is any edge of T_w and γ is any element of $[\Gamma_e, \Gamma_e]$, the commutator subgroup of the stabilizer of e in Γ, then $I_\gamma(x) = 2$.*

5.6. Separating surfaces in knot exteriors

In his classic treatise [47], Lee Neuwirth asked a number of questions about the structure of knot groups, i.e. fundamental groups of complements of nontrivial knots in S^3. (A knot is said to be *trivial* if it bounds a disk in S^3.) One of his questions was whether every knot group can be expressed as a nontrivial free product with amalgamation (see Subsection 2.6) in which the amalgamated subgroup is free. He proposed the idea of answering this question affirmatively by showing that if K is a nontrivial tame knot in S^3 then the exterior of K contains a separating essential surface. I will refer to this statement as the *weak Neuwirth Conjecture*, because in [46] Neuwirth formulated stronger topological and group-theoretic versions of his conjecture, some of which are still unproved. Of course you should compare the weak Neuwirth Conjecture with the elementary fact, which I talked about in Subsection 2.5, that the exterior of K always contains a nonseparating essential surface.

The weak Neuwirth Conjecture was proved in [18], in a much more general context than that of knot exteriors in S^3. It is included in the following result, which is proved in [18]:

Theorem 5.6.1 (Culler-Shalen)**.** *Let M be a compact, orientable, irreducible* 3*-manifold whose boundary is a torus. Suppose that $H_1(\partial M; \mathbf{Q}) \to H_1(M; \mathbf{Q})$ is surjective, but that M is not a solid torus. Then M contains a separating essential surface.*

(By the way, when M is the exterior of a tame knot in S^3, the surjectivity of $H_1(\partial M; \mathbf{Q}) \to H_1(M; \mathbf{Q})$ follows easily from the Mayer-Vietoris theorem. The irreducibility of M in this case follows from a classical result due to Alexander, Graeub and Moise, the so-called 3-dimensional PL Schönflies theorem (see for example [39], Section 17).)

I will discuss the stronger versions of Neuwirth's conjectures in Section 10. The essential point to be made here is that while Theorem 5.6.1 applies to irreducible tame knot exteriors in arbitrary rational homology 3-spheres, the stronger versions of the conjecture seem to depend on the hypothesis that the knot is in S^3, or at any rate in a 3-manifold of some more special sort.

Actually Theorem 5.6.1 is an essentially immediate consequence of a result, stated below as Theorem 5.6.2, which applies to irreducible tame knot exteriors in arbitrary closed, orientable, connected 3-manifolds; the statement of this theorem was not spelled out in [18]. To state it we need the notion of a *boundary slope*, which is discussed in more detail in Boyer's chapter in this volume. CHECK THIS. Briefly, if M is a compact, orientable, irreducible 3-manifold whose boundary is a torus, the boundary components of an essential surface F are all isotopic since they are disjoint, homotopically nontrivial simple closed curves on a torus. Their common isotopy class is called the *boundary slope* of F; for reasons that Boyer explains, the term *slope* is used to mean any isotopy class of nontrivial simple closed curves in ∂M. A slope is called a boundary slope of M if it is the boundary slope of some essential surface in M. A fundamental result of Hatcher's [30] implies that the set of boundary slopes of M is always finite.

Theorem 5.6.2. *Let M be a compact, orientable, irreducible 3-manifold whose boundary is a torus. Then either*

(i) *M is a solid torus, or*
(ii) *M contains an essential separating annulus, or*
(iii) *M contains an essential nonseparating torus, or*
(iv) *M has at least two boundary slopes.*

To see that this implies Theorem 5.6.1, notice that alternative (i) is ruled out by the hypothesis of Theorem 5.6.1, while alternative (ii) implies the conclusion of 5.6.1. The hypothesis that $H_1(\partial M;\mathbf{Q}) \to H_1(M;\mathbf{Q})$ is surjective rules out alternative (iii). This hypothesis also implies, by a simple homological argument, that there is at most one slope that can occur as the boundary slope of a nonseparating essential surface. So if alternative (iv) holds then M contains an essential separating surface.

In this subsection I will show how to apply the techniques that I've described to prove Theorem 5.6.2 in the special case where $N = \operatorname{int} M$ has a (finite-volume) hyperbolic structure. In this case one gets a stronger result, namely that alternative (iv) of Theorem 5.6.2 holds. I'll establish alternative (iv) in the following paraphrased form: if s is any slope, there is an essential surface which has nonempty boundary and has a boundary slope different from s.

Let s be represented by a simple closed curve c in ∂M and let γ be an element representing the conjugacy class determined by some orientation of c. We consider the curve $X_0 \subset X(\pi_1(M))$ given by Corollary 4.5.2. Since I_γ is nonconstant on X_0 it must have a pole on $\hat{X}_0$, which must occur at some ideal point x of X_0.

With the ideal point x we can associate a tree T_x and an action of $\pi_1(M)$ on T_x by the construction of Section 5. Furthermore, according to Section 2 there exists an essential surface $F \subset M$ which is dual to the action of $\pi_1(M)$ on T_x. To complete the proof it suffices to show that any such F has a nonempty boundary and that its boundary slope is different from s.

Assume that either F is closed, or that $\partial F \neq \emptyset$ and that the boundary slope of F is s. In either case, the simple closed curve c is isotopic to a simple closed curve in the complement of F. Thus the element $\gamma \in \pi_1(M)$ lies in a conjugate of $\operatorname{im}(\pi_1(C_i)) \to \pi_1(M)$ for some component C_i of $M - F$. According to 2.3.1(i), this implies that γ fixes some vertex of T_x. But by Property 5.4.2 this means that I_γ does not have a pole at x. Of course this contradicts our choice of x, and so the proof of the theorem is complete—in the special case where $\operatorname{int} M$ is hyperbolic.

6. The proof of the weak Neuwirth Conjecture

I'm going to present the main ideas in the proof of Theorem 5.6.2 in the general case, where we don't assume that int M has a hyperbolic structure. I will skip over a few technical algebraic details, which you can find in [18].

The starting point for the argument is Thurston's geometrization theorem [49], [50], which, together with the characteristic submanifold theorem ([35], [34]) guarantees that by splitting M along some disjoint system of essential tori $\{T_1, \ldots, T_k\}$ we can get a manifold M' such that for each component M'_i of M', either M'_i is a Sefiert fibered space or int M'_i is hyperbolic. We can assume that the component of M_0 of M' which contains the torus $B_0 = \partial M$ is not homeomorphic to $S^1 \times S^1 \times [0, 1]$, as otherwise we could replace $\{T_1, \ldots, T_k\}$ by a system of fewer tori with the same properties.

If M_0 is a Seifert fibered space (not homeomorphic to $S^1 \times S^1 \times [0, 1]$), it's a routine matter to check that either M_0 is homeomorphic to $D^2 \times [0, 1]$—in which case $M = M_0$ and alternative (i) of Theorem 5.6.2 holds—or M_0 contains an essential separating annulus A with $\partial A \subset B_0$. We can think of A as an essential annulus in M. This annulus will separate M (implying alternative (ii) of Theorem 5.6.2) unless there are two components T_1 and T_2 of ∂M_0 which lie in different components of $M_0 - A$ but lie in the same component of $M - M_0$; but if this happens then T_1 and T_2 are both nonseparating tori in M, so that alternative (iii) of Theorem 5.6.2 will hold. So the conclusion of the theorem holds whenever M_0 is a Seifert fibered space.

In the crucial case where int M_0 is hyperbolic, we generalize the argument of Subsection 5.6 to show that alternative (iv) of Theorem 5.6.2 holds. As I pointed out in Subsection 5.6, it suffices to show that if s is any slope, there is an essential surface in M which has nonempty boundary and has a boundary slope different from s. It turns out we can show more than this, namely that there is an essential surface $F \subset M_0$ with $\partial F \subset B_0$, and such that the common isotopy class of the components of ∂F is distinct from s.

In Subsection 5.6, in the case where int M was hyperbolic and had a single cusp, we used the 1-dimensional component of $\pi_1(M)$ given by Corollary 4.5.2. In the general case, if int M_0 has n cusps then ∂M_0 consists of n tori. We have labeled one of these B_0; let $B_1, \ldots, B_{n-1}$ denote the others. Theorem 4.5.1 gives an n-dimensional irreducible component X_0 of $X(\Gamma)$. I'll define a curve $Y_0 \subset X_0$ which plays the role that X_0 played in the one-cusp case. Specifically, I claim that X_0 contains a curve Y_0 such that

(i) for each $i = 1, \ldots, n-1$ and each element $\alpha \in \mathrm{im}(\pi_1(B_i) \to \pi_1(M))$, the function $I_\alpha | Y_0$ is identically equal to either 2 or -2, and

(ii) for each nontrivial element $\alpha \in \mathrm{im}(\pi_1(B_0) \to \pi_1(M))$, the function $I_\alpha | Y_0$ is nonconstant.

To construct Y_0 we first recall that X_0 is by definition an irreducible component of $X(\pi_1(M))$ containing the character χ_0 of the lift to $\mathrm{SL}(2, \mathbf{C})$ of a discrete faithful representation of $\pi_1(M)$ in $\mathrm{PSL}(2, \mathbf{C})$. Hence if we fix elements $\gamma_i \in \mathrm{im}(\pi_1(B_i) \to \pi_1(M))$ for $i = 1, \ldots, n-1$, we have $I_{\gamma_i}(\chi_0) = \chi_0(\gamma_i) = \pm 2$ for $1 \leqslant i \leqslant n-1$.

In particular, χ_0 lies in the algebraic subset Z of $X(\pi_1(N))$ obtained by adding the equations $I^2_{\gamma_i} = 4$ for $i = 1, \cdots, n-1$. Now by a general property of complex affine varieties that I already quoted in Section 4.5 (see [56], p. 60, Theorem 7), since X_0 has dimension n, and Z is defined by adding $n-1$ extra equations, each component of Z must have dimension at least 1. In particular there must be a curve Y_0 with $\chi_0 \in Y_0 \subset X_0$. To prove property (ii) for the curve Y_0, we use the fact that $I_\alpha(\chi_0) = \pm 2$ to conclude that if I_α were constant on Y_0, all the functions $I^2_\alpha, I^2_{\gamma_1}, \dots, I^2_{\gamma_{n-1}}$ would be identically equal to 4 on Y_0. Thus Y_0 would be contained in the algebraic subset X^* of $X(\pi_1(N))$ obtained by adding the equations $I^2_\alpha = I^2_{\gamma_1} = \ldots = I^2_{\gamma_{n-1}} = 4$. But this contradicts the last assertion of Theorem 4.5.1, according to which χ_0 is an isolated point of X^*.

To show that Y satisfies (i), one begins with the fact, which I already mentioned in Subsection 5.4 (the paragraph before the statement of the Extension Theorem for Valuations) that any (irreducible) curve in $X(\pi_1(M))$ is in fact the image under $t : R(\pi_1(M)) \to X(\pi_1(M))$ of a subvariety of $R(\pi_1(M))$. Having fixed a subvariety R_0 of $R(\pi_1(M))$ with $t(R_0) = Y_0$, one considers an index i with $1 \leqslant i < n$ and an element α of $\operatorname{im}(\pi_1(B_i) \to \pi_1(M))$. For any point $\rho \in R_0$ we have $\operatorname{trace}\rho(\gamma_i) = I_{\gamma_i}(t(\rho)) = \pm 2$, since $t(\rho) \in Y_0 \subset Z$. Thus $\rho(\gamma_i)$ is either $\pm I$ or a conjugate of $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. But $\rho(\alpha)$ commutes with $\rho(\gamma_i)$ since $\pi_1(B_i)$ is abelian; as the only matrices that commute with $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ are those of the form $\pm \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ for $\lambda \in \mathbf{C}$, we must have either $\rho(\gamma_i) = \pm I$ or $\operatorname{trace}\rho(\alpha) = \pm 2$.

In view of the irreducibility of R_0, there are two possibilities: either (a) $\rho(\alpha_i) = \pm I$ for *every* $\rho \in R_0$, or (b) $\operatorname{trace}\rho(\alpha) = \pm 2$ for every $\rho \in R_0$. Now (b) is exactly what we need, because it translates into the statement that the function I_α takes the value 2 or -2 at every point of Y_0; this implies the conclusion of (i) since Y_0 is irreducible. So we need only rule out (a). Well, if (a) holds then every point of X_0, in particular $\chi_0 = t(\rho_0)$, is the character of a nonfaithful representation. But since ρ_0 is irreducible, any representation with character χ_0 is equivalent to ρ_0, and since ρ_0 is faithful we have a contradiction.

Having constructed the curve Y_0 satisfying (i) and (ii), we proceed to construct the required surface F. We follow the same basic procedure as in Subsection 5.6, using Y_0 in place of X_0. Remember that we are given a slope s in B_0, and that we require that the essential surface $F \subset M_0$ have its boundary contained in B_0, and that the common isotopy class of the components of ∂F be distinct from s. We represent s by a simple closed curve c in ∂M, and we let γ be an element representing the conjugacy class determined by some orientation of c. By property (ii) of the curve Y_0, the function $I_\alpha | Y_0$ is nonconstant, and therefore has a pole at some ideal point x of $\tilde{Y}_0$. The ideal point x determines an action of $\pi_1(M)$ on a tree T. If F is any essential surface in M dual to this action, the same argument that we used in Subsection 5.6 shows that F must have boundary components contained in B_0, and that the common isotopy class of these boundary components cannot be s. The new twist is that we must pick the surface F in such a way that it has no boundary components in any component $B_i \neq B_0$ of ∂M. This is made possible by property (i) of Y_0; I'll indicate how this works.

First we translate property (i) into a property of the action of $\pi_1(M)$ on T. For $i = 1, \ldots, n-1$, set $\Gamma_i = \operatorname{im}(\pi_1(B_i) \to \pi_1(M))$. Of course, this subgroup is defined only up to conjugacy, but we fix a concrete representative Γ_i in the conjugacy class of subgroups. Property (ii) says that I_γ takes a finite value at the ideal point x for every $\gamma \in \Gamma_i$. Thus by Proposition 5.5.1, there is a vertex v_i of T which is fixed by Γ_i.

The property of the action that I have stated here is exactly what is needed to guarantee that we can choose the a surface F dual to the action in such a way that a F is disjoint from $B_1, \ldots, B_{n-1}$. (Remember that the surface dual to a given action is in general far from being canonical.) The transition from the property of the action to the property of the (suitably chosen) dual surface is contained in the following result, which is a corollary to Proposition 2.5.3.

Corollary 6.0.1. *Let M be a compact, orientable, irreducible 3-manifold, and let $B_1, \ldots, B_k$ be disjoint subpolyhedra of ∂M. Suppose that we are given an action of $\pi_1(M)$ on a tree T, without inversions, and suppose that for each $i \leqslant k$, the subgroup $\Gamma_i = \operatorname{im}(\pi_1(B_i) \to \pi_1(M))$ of $\pi_1(M)$ fixes a vertex v_i of T. (Of course the subgroups Γ_i are defined only up to conjugacy, but if a given subgroup fixes a vertex then any conjugate subgroup fixes a (possibly different) vertex; so the condition makes sense.) Then there is an essential surface $F \subset M$, dual to the action of $\pi_1(M)$ on T, such that $F \cap B_i = \emptyset$ for $i = 1, \ldots, k$.*

This is just the special case of Proposition 2.5.3 in which the map $\tilde{g}_i$, for $i = 1, \ldots, k$, is the constant map that sends $\tilde{B}_i$ to v_i; it has the required equivariance property because Γ_i fixes v_i. In the notation of 2.5.3 we have $C_i = \emptyset$ for $i = 1, \ldots, k$, so that the conclusion of 2.5.3 implies that $F \cap B_i = \emptyset$. □

In the case of the action associated to the ideal point x that we are considering, I have shown that the hypothesis of Corollary 6.0.1 holds if we set $k = n-1$ and define $B_1, \ldots B_{n-1}$ as above. So the corollary gives an essential dual surface F to the action which has the property that $F \cap B_i = \emptyset$ for $i = 1, \ldots, n-1$. This completes the proof of Theorem 5.6.2. □

7. The Smith Conjecture

It is a long-standing conjecture that a tame periodic homeomorphism h of S^3 is topologically linear, i.e. conjugate to the restriction to S^3 of an orthogonal transformation of $\mathbf{R}^4$. (To say that h is *tame* means that for every fixed point P of h, there exist an h-invariant neighborhood V of h and a homeomorphism j of V onto the unit ball $B^3 \subset \mathbf{R}^3$, such that jhj^{-1} is the restriction of a linear automorphism of $\mathbf{R}^3$. A periodic *diffeomorphism* is automatically tame.) The Smith Conjecture, which was first proved in 1978 (see [41]), is equivalent to the special case of this conjecture in which h is assumed to have period $n > 1$ (not much of a restriction), to preserve orientation, and to have nonempty fixed point set. Under these assumptions, it follows from classical theorems due to P. A. Smith that the fixed point set of h is homeomorphic to a 1-sphere, so that it may regarded as a knot in S^3; and the hypothesis that h is tame implies that the fixed point set is a tame knot.

It was long unknown whether this fixed point set could be a nontrivial knot. In fact, classical arguments may be used to show that the homeomorphism is topologically linear if and only if the fixed point set is unknotted. This connection with classical knot theory was one source of fascination with the Smith Conjecture. However, the proof that was eventually given did not use this particular reduction to knot theory, but another one which in a sense is more direct.

If h is a tame, periodic homeomorphism whose fixed point set is a knot, it's elementary to show that the orbit space $\Sigma = S^3/\langle h\rangle$ is a connected, closed, orientable 3-manifold; and, furthermore, that the orbit map $p : S^3 \to \Sigma$ is a branched covering map, whose branch set is a tame knot $K \subset \Sigma$. It's pretty clear that the branched covering S^3 of Σ is regular and that its covering group is the cyclic group $\langle h\rangle$. Furthermore, it's an entirely elementary exercise to prove that h is topologically linear if and only if the knot $K \subset \Sigma$ is trivial, in the sense that it bounds a disk in Σ. So the Smith Conjecture is in fact an immediate consequence of the following result.

Theorem 7.0.1. *Let K be a tame knot in a connected, closed, orientable* 3-*manifold Σ. Let n be an integer* > 1*, and let $\tilde{\Sigma}$ denote the n-fold cyclic regular branched covering of Σ, branched over K. Assume that $\tilde{\Sigma}$ is simply connected. Then the knot K is trivial.*

Of course this result is really stronger than the Smith Conjecture, because it is assumed only that $\tilde{\Sigma}$ is simply connected, not that it is diffeomorphic to S^3. For this reason the result was known for a time as the "Generalized Smith Conjecture," until a still more general version was announced by Thurston. . .

In this section I will sketch a proof of Theorem 7.0.1 which was first given in [17]. As I will point out below, it gives a rather stronger form of Theorem 7.0.1 than the form that was first proved in [41]. In any case, it is a good illustration of the use of the character variety in this subject.

7.1. Preliminary observations

A preliminary step in the proof of Theorem 7.0.1 is to reduce it to the special case where the exterior M of K is irreducible. This may be done by using the Kneser finiteness theorem—of which you will find an account in Bonahon's chapter in this volume, or in [31] (Theorem 3.15)—to decompose M as the connected sum of an irreducible manifold, itself the exterior of some tame knot K_0 in a closed orientable 3-manifold Σ_0, and a closed manifold. The n-fold cyclic branched cover $\tilde{\Sigma}_0$ of Σ branched over K_0 is then a connected summand of $\tilde{\Sigma}$, so if $\tilde{\Sigma}$ is simply connected, so is $\tilde{\Sigma}_0$. It's also a routine matter to check that K is trivial if and only if K_0 is.

There is a basically similar preliminary reduction to the case where K is a *prime* knot. A tame knot is said to be *prime* if its exterior contains no essential annulus whose boundary components are meridians. (I'm using the term "essential" as I did in Section 1, and since at this point I'm looking at a knot whose exterior is irreducible, the context is consistent with that of Section 1.) The reason for the term "prime" is that if the exterior of K does contain an essential annulus with meridian boundaries, we can decompose K as a *connected sum* of two tame knots K_1 and K_2 in closed orientable 3-manifolds Σ_1, Σ_2. (The connected sum is defined by removing from each Σ_i a ball B_i that meets K_i in an unknotted arc α_i, then gluing together $\Sigma_1 - \operatorname{int} B_1$ and $\Sigma_2 - \operatorname{int} B_2$ by some homeomorphism of their boundaries that matches $\partial\alpha_1$ with $\partial\alpha_2$; this gives a knot $K_1 \# K_2$ in the manifold $\Sigma_1 \# \Sigma_2$. For $\Sigma_1 = \Sigma_2 = S^3$, this formalizes the idea of tying two knots in succession in the same piece of string.) There is a prime decomposition of a tame knot analogous to the prime decomposition for an oriented 3-manifold described in Bonahon's chapter, and using this it is not hard to reduce the proof of Theorem 7.0.1 to the case where K, in addition to having an irreducible exterior, is prime.

Having made these reductions, we consider a prime tame knot, with irreducible exterior M, in a connected, closed, orientable 3-manifold Σ, and an integer $n > 1$. We let $\tilde{\Sigma}$ denote the n-fold cyclic regular branched covering of Σ, branched over K. Working contrapositively, we assume that K is a nontrivial knot, and we wish to show that $\pi_1(\Sigma)$ is a nontrivial group. We let μ denote the meridian μ of K, which we think of as an element of $\pi_1(M)$, represented by a simple closed curve in ∂M. When I need to refer to this simple closed curve itself, regarded as a subset of ∂M, I will denote it $|\mu|$. It is quite elementary to see that there is an exact sequence

$$1 \to \pi_1(\tilde{\Sigma}) \to |\pi_1(M) : \mu^n = 1| \to \mathbf{Z}/n\mathbf{Z} \to 0.$$

Here by $|\pi_1(M) : \mu^n = 1|$ I of course just mean the group obtained from $\pi_1(M)$ by adding the relation $\mu^n = 1$. (Nowadays this group is often referred to as an orbifold group.)

7.2. The character variety appears

Thus one way of showing that $\pi_1(\tilde{\Sigma})$ is nontrivial is to show that $|\pi_1(M) : \mu^n = 1|$ is not a cyclic group. Thus it certainly suffices to find a representation $\bar{\rho} : \pi_1(M) \to \mathrm{PSL}(2, \mathbf{C})$ with noncyclic image such that $\bar{\rho}(\mu)$ has order n. To carry this a step further, it suffices to find a representation $\rho : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ such that

(i) the image of $\rho(\pi_1(M))$ under the natural homomorphism
$\mathrm{SL}(2, \mathbf{C}) \to \mathrm{PSL}(2, \mathbf{C})$ is noncyclic,
and
(ii) $\rho(\mu)$ has order $2n$.

(I am using here the fact that $-I$ is the only element of order 2 is $\mathrm{SL}(2, \mathbf{C})$, so that any element of order $2n$ is $\mathrm{SL}(2, \mathbf{C})$ maps to an element of order n in $\mathrm{PSL}(2, \mathbf{C})$.)

Notice that I haven't claimed that there is always a representation ρ satisfying (i) and (ii). What I'll show, rather, is that in the crucial case where $\operatorname{int} M \cong \Sigma - K$ is hyperbolic (i. e. has a hyperbolic structure of finite volume), the attempt to find a representation ρ satisfying (i) and (ii) either succeeds or—in the case where it fails—leads to an alternative way of showing that $\pi_1(\tilde{\Sigma})$ is nontrivial. It will also turn out that when $\operatorname{int} M$ is not hyperbolic, this alternative method nearly always works—and the exceptional cases are easily handled.

Assume for the moment, then, that $\operatorname{int} M$ is hyperbolic. Corollary 4.5.2 gives a curve $X_0 \subset X(\pi_1(M))$. We are attempting to find a point $\chi \in X_0$ which is the character of a representation ρ satisfying (i) and (ii). For condition (i) the following lemma is of obvious relevance.

Lemma 7.2.1. *Let N be an orientable hyperbolic 3-manifold of finite volume with a single cusp, and let X_0 be a curve in $X(\pi_1(N))$ given by Corollary 4.5.2. Then for every point $\chi \in X_0$ there is a representation $\rho \in t^{-1}(\chi)$ such that the image of $\rho(\pi_1(N))$ under the natural homomorphism $\mathrm{SL}(2, \mathbf{C}) \to \mathrm{PSL}(2, \mathbf{C})$ is noncyclic.*

Sketch of proof. We have $X_0 = t(R_0)$, where R_0 is an irreducible component of $R(\pi_1(N))$ containing ρ_0, a lift of the discrete, faithful representation of $\pi_1(N)$ to $\mathrm{SL}(2, \mathbf{C})$. It's easy to see that ρ_0 is irreducible. It's also easy to see that the reducible representations of $\pi_1(N)$ form a closed algebraic subset of $R(\pi_1(N))$. (In fact, it's shown in [17] that a representation ρ is reducible if and only if $\operatorname{trace} \rho(\gamma) = 2$ for every element γ of the commutator subgroup of $\pi_1(N)$.) So there's an open, dense subset of R_0 consisting of irreducible representations.

If $\rho \in R_0$ is irreducible then $\rho(\pi_1(N))$ is nonabelian. This easily implies that the centralizer of $\rho(\pi_1(N))$ is $\{\pm I\}$. Hence under the action of $\mathrm{SL}(2, \mathbf{C})$ by conjugation on R_0, the orbit $\rho^{\mathrm{SL}(2, \mathbf{C})}$ of ρ—which for the irreducible representation coincides with the fiber $t^{-1}(t(\rho))$ of ρ—is homeomorphic to $\mathrm{PSL}(2, \mathbf{C})$ and thus has dimension 3. Since X_0 has dimension 1 and the generic fiber of $t|R_0 : R_0 \to X_0$ has dimension 3, it follows that the dimension of R_0 is 4. By a general property of algebraic maps between complex varieties, it now follows that *every* fiber of $t|R_0$ has dimension at least $\dim R_0 - \dim X_0 = 3$.

Suppose now that there is a point $\chi \in X_0$ such that for every representation $\rho \in t^{-1}(\chi)$, the image of $\rho(\pi_1(N))$ in $\mathrm{PSL}(2, \mathbf{C})$ is cyclic. If for some given $\rho \in t^{-1}(\chi)$ the group $\rho(\pi_1(N))$ is not itself cyclic, then $\rho(\pi_1(N))$ must contain $-I$ and the quotient $\rho(\pi_1(N))/\{\pm I\}$ must be cyclic. It follows that in this case we must have $\rho(\pi_1(N)) \cong (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/2\mathbf{Z})$ for some $n \geqslant 0$.

Let S denote the set of all homomorphisms of $\pi_1(N)$ onto groups of the form $(\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/\epsilon\mathbf{Z})$, where n ranges over the nonnegative integers and ϵ over $\{1, 2\}$. Since $\pi_1(N)$ is finitely generated, S is countable. We can write

$$t^{-1}(\chi) \subset \bigcup_{\phi \in S} A_\phi;$$

here $A_\phi \subset R(\pi_1(N))$ denotes the set of all homomorphisms of the form $\sigma \circ \phi$, where σ ranges over all faithful representations of $\phi(\pi_1(N)) = (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/\epsilon\mathbf{Z})$ in $\mathrm{SL}(2, \mathbf{C})$. Such a representation σ must send the generator of $\mathbf{Z}/\epsilon\mathbf{Z}$ to $-I$ if $\epsilon = 2$ (and to I if $\epsilon = 1$). Hence, if we are given $\phi \in S$ and we fix an element $\gamma \in \pi_1(N)$

such that $\phi(\gamma)$ is the standard generator of $(\mathbf{Z}/n\mathbf{Z}) \times \{0\}$, a representation $\rho \in A_\phi$ is uniquely determined by the element $\rho(\gamma)$. If $\rho \in A_\phi \cap t^{-1}(\chi)$ then the trace of $\rho(\gamma)$ must be equal to $\chi(\gamma)$. Since the set of matrices with a given trace is a 2-dimensional subvariety of $\mathrm{SL}(2, \mathbf{C})$, it follows that the set $A_\phi \cap t^{-1}(\chi) \subset R(\pi_1(N))$ has dimension at most 2 for any $\phi \in S$. Thus

$$t^{-1}(\chi) = \bigcup_{\phi \in S} A_\phi \cap t^{-1}(\chi)$$

is a countable union of sets of dimension at most 2. Since we showed that the complex algebraic set $t^{-1}(\chi)$ has dimension 3, we now have a contradiction.

7.3. The character variety argument completed

Having finished the proof of Lemma 7.2.1, we return to the main menu. (Hi, my name's Peter, I'm your waiter.) Remember we are trying to find a representation $\rho : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ satisfying (i) and (ii). In view of the lemma (applied to $N = \operatorname{int} M$), it's enough to find a point $\chi \in X_0$ such that every $\rho \in t^{-1}(x)$ satisfies (ii). The simple observation that gets us started in doing this is that if ω is any primitive $2n$-th root of unity in $\mathbf{C}$, then any matrix with trace $\omega + \omega^{-1}$ has order $2n$ in $\mathrm{SL}(2, \mathbf{C})$. This is because the matrix

$$\begin{pmatrix} \omega & 0 \\ 0 & \omega^{-1} \end{pmatrix}$$

has the right trace and the right order; and since $n > 1$ we have $\omega + \omega^{-1} \neq \pm 2$, so that any two matrices with trace $\omega + \omega^{-1}$ and determinant 1 are conjugate.

The upshot of all this is that if I_μ takes the value $\omega + \omega^{-1}$ at some point $\chi \in X_0$, then any representation $\rho \in t^{-1}(\chi)$ will satisfy (ii); by the lemma, some $\rho \in t^{-1}(\chi)$ will also satisfy (i). So ρ will give rise to a representation of $|\pi_1(M) : \mu^n = 1|$ in $\mathrm{PSL}(2, \mathbf{C})$ with a noncyclic image, and in particular it will follow that $|\pi_1(M) : \mu^n = 1|$ is noncyclic, and hence that $\pi_1(\tilde{\Sigma})$ is nontrivial.

Now by Corollary 4.5.2, the function I_μ is nonconstant on X_0. Furthermore, as there is a canonical isomorphism $\mathbf{C}(\hat{X}_0) \to \mathbf{C}(X_0)$, we can extend I_μ to a rational function $\hat{I}_0 : \hat{X}_0 \to \mathbf{C} \cup \{\infty\}$; since I_0 is nonconstant, $\hat{I}_0$ is surjective. Thus *either* I_μ takes the value $\omega + \omega^{-1}$ at some point of X_0, *or* $\hat{I}_\mu$ takes the value $\omega + \omega^{-1}$ at some point $x \in \hat{X}_0 - X_0$. In the latter case we can still show that $\pi_1(\tilde{\Sigma})$ is nontrivial, but by a quite different method.

The idea is, of course, to look at the tree T associated to the ideal point x. Since I_μ takes a finite value at x, it follows from 5.4.2 that μ fixes a vertex of T. Since μ generates the image of the fundamental group of $|\mu| \subset \partial M$ in $\pi_1(M)$, it follows from Corollary 6.0.1 that there is an essential surface $F \subset M$, dual to the action of $\pi_1(M)$ on T, such that $F \cap |\mu| = \emptyset$. (In applying the corollary we take $k = 1$

and $B_1 = |\mu|$.) Thus either F has boundary components which are all parallel to $|\mu|$—i.e. its boundary slope is the meridional slope—or it is closed.

At this point it is easy to reduce the proof of Theorem 7.0.1 to the proof of the following result, which can be proved by arguments due to Meeks-Yau, Gordon-Litherland and Thurston.

Theorem 7.3.1. *Let K be a prime tame knot in a closed, connected, orientable 3-manifold Σ, such that the exterior M of K is irreducible. Suppose that either M contains a closed essential surface, or the meridian of K is a boundary slope in M. Then for any $n > 1$, the n-fold cyclic branched covering space $\tilde{\Sigma}$ of Σ branched over K contains a closed, connected essential surface. (Here the term "essential" can be interpreted according to the definition I gave in Section 1, even though we don't know that $\tilde{M}$ is irreducible.)*

The point is that what I have shown above is that if $\operatorname{int} M$ is hyperbolic, and if we are in the case where the approach to the proof of Theorem 7.0.1 based on the character variety fails, the hypothesis of Theorem 7.3.1 must hold. But the conclusion of 7.3.1 certainly implies the conclusion of 7.0.1, since if $\tilde{F}$ and is essential in $\tilde{M}$, then according to the definition I gave in Section 1, $\tilde{F}$ has genus > 0, and $\pi_1(\tilde{\Sigma})$, which contains an isomorphic copy of $\pi_1(F)$, is therefore nontrivial. If $\operatorname{int} M$ is not hyperbolic we are still OK, because then by Thurston's geometrization theorem, either M contains an essential torus and we can still apply Theorem 7.3.1, or M is Seifert fibered, in which case the proof of Theorem 7.0.1 is an elementary exercise based on the classification of the Seifert fibered spaces.

7.4. The equivariant loop theorem

I will give only the briefest hint about the proof of Theorem 7.3.1, since the techniques don't have much to do with the subject of this chapter. If $F \subset M$ is an essential surface which is either closed or has meridional boundary slope, then the pre-image of F in $\tilde{\Sigma}$, say $\tilde{F}_0$, is a closed bicollared surface, possibly disconnected but definitely invariant under the action of $\mathbf{Z}/n\mathbf{Z}$ on the cyclic branched covering space $\tilde{\Sigma}$. Using the primality of K and the irreducibility of M, it's not hard to show that $\tilde{F}_0$ has a component of genus > 0. If this component is essential, we're happy. If not, the equivariant version of the Dehn-Lemma-Loop-Theorem due to Meeks and Yau [38] allows one to do compressions on F_0 in such a way that the resulting surface F_1 is still invariant under the $\mathbf{Z}/n\mathbf{Z}$ action. The primality and irreducibility again imply that F_1 must have a component of positive genus, and we can repeat the process until we see an essential component appearing. (By a finiteness argument like the one I described in Subsection 2.4, the process cannot continue indefinitely.)

The big ingredient here is the equivariant Dehn-Loop Theorem. This was first proved by Meeks and Yau using minimal surface techniques. Their proof, which required a lot of hard analysis, was later reinterpreted in a purely combinatorial setting by Edmonds [25], Dunwoody [24] and Jaco and Rubinstein [33].

7.5. The root-of-unity phenomenon

In the argument given in Subsection 7.3, the only ideal points that had to be considered were those where the function I_μ took finite values of the apparently special form $\omega+\omega^{-1}$, where ω is some root of unity. We saw that just by using the finiteness of I_μ at the ideal point we got topological information—that M contains a closed essential surface or that μ is a boundary slope—which was crucial for the proof. It is natural to wonder whether the fact that one obtains a very special kind of finite value, namely $\omega+\omega^{-1}$, where ω is a root of unity, provides additional restrictions on the situation.

From this point of view, the following result, proved in [11], is striking. Suppose we have a compact orientable manifold M bounded by a torus, with $\operatorname{int} M$ hyperbolic, and for simplicity suppose that M contains no closed essential surfaces. Suppose that for some nontrivial element α of $\operatorname{im}(\pi_1(\partial M)\to\pi_1(M))$, the function I_γ takes a finite value $c\in\mathbf{C}$ at some ideal point x of the curve given by Corollary 4.5.2. Then $c=\omega+\omega^{-1}$ for some root of unity ω.

As a hint about why this should be so, consider a dual surface F to an action on a tree T associated to x. Since by hypothesis F cannot be closed, the arguments of Subsection 7.3 show that α belongs to the conjugacy class in $\pi_1(M)$ defined by the boundary slope of α. Now consider for a moment the special case in which ∂F is connected. In this case, α is a product of commutators in $\operatorname{im}(\pi_1(F)\to\pi_1(M))$, which by 2.3.1(ii) is a subgroup of the stabilizer of an edge of T. By Property 5.5.3, it follows that $I_\gamma(x)=2$. This proves the assertion in this case, since we can take $\omega=1$.

The proof in the general case is a refinement of this argument. If F is a dual surface having the minimal number of boundary components among all surfaces dual to the action, and if some component of F has n boundary components, then ω can be shown to be an n-th root of unity.

Nathan Dunfield found a remarkable application of this result in his paper [23]. As I will barely have a chance to mention Dunfield's results in Sections 9 and 10, you will have to look at his paper to appreciate his application of the root-of-unity phenomenon.

7.6. Extensions of the theorem

If you examine the proof of Theorem 7.0.1 that I have sketched here, you will see that it really gives a proof of the following stronger result:

Theorem 7.6.1. *Let K be a nontrivial tame knot in a closed, connected, orientable 3-manifold Σ. Let n be an integer >1, and let $\tilde{\Sigma}$ denote the n-fold cyclic regular branched covering of Σ, branched over K. Then either (i) $\pi_1(\tilde{\Sigma})$ has a nontrivial representation in $\mathrm{PSL}(2,\mathbf{C})$, or (ii) $\tilde{\Sigma}$ contains an essential surface. Condition (i) can be replaced by the stronger condition that the "orbifold group" $|\pi_1(M):\mu^n=1|$, where M is the exterior of K and μ the meridian, has a representation in $\mathrm{PSL}(2,\mathbf{C})$ with noncyclic image.*

This "PSL$(2, \mathbf{C})$-version" of the Smith Conjecture is the result which I mentioned earlier as being stronger than the version of the Smith Conjecture proved in [41]. In [17] you will find a proof of an essentially more general result than Theorem 7.6.1, which applies to many noncyclic regular coverings; it is a "PSL$(2, \mathbf{C})$-version" of a generalization of the Smith Conjecture due to Davis and Morgan [21]. Thurston's orbifold theorem, which you may read about in Bonahon's chapter in this volume, is in turn much stronger than Theorem 7.6.1. (It is also much harder to prove!) In the cases where the "PSL$(2, \mathbf{C})$-versions" give nondegenerate representations in PSL$(2, \mathbf{C})$ of the fundamental group of the branched covering—or the "orbifold group" which contains the group of the branched covering as a finite-index subgroup—Thurston's result actually gives a geometric structure on the manifold whose existence implies the existence of such a representation. Furthermore, the geometric structure is invariant under the group of symmetries of the branched covering, and this accounts for the extension of the representation to the "orbifold group": see Bonahon's chapter in this volume. Thurston's theorem also applies in more general situations than the other results.

8. Degrees and Norms

The material in this section first appeared in [15] and was worked out by Marc Culler and myself.

In Subsection 5.6 and Sections 6 and 7, very simple properties of algebraic curves were used to prove nontrivial theorems about 3-manifolds. In Subsection 5.6 and Section 6 we used the simple fact that a nonconstant rational function f on a projective algebraic curve C always has at least one pole. In Section 7 we used the essentially equivalent fact that such a function f takes every value in $\mathbf{C} \cup \{\infty\}$ at least once. These facts were applied to the functions $\hat{I}_\gamma$ on a projective completion $\hat{X}_0$ of a curve X_0 given by Corollary 4.5.2. (Recall that $\hat{X}_0$ is a curve in $X(\pi_1(N))$, where N is a hyperbolic 3-manifold with one cusp and γ is a nontrivial peripheral element of $\pi_1(N)$).)

These simple facts about a nonconstant rational function f on a projective algebraic curve C can be regarded as consequences of the fact that f is a *branched covering map.* (Coincidentally the same concept came up, one dimension higher, in the last section.) This is an especially natural point of view in the case where C is smooth, and I'll discuss this case first. To see that f is a branched covering map means that there is a finite set $\Psi \subset \mathbf{C} \cup \{\infty\}$ such that $f|C - f^{-1}(\Psi) : C - f^{-1}(\Psi) \to (\mathbf{C} \cup \{\infty\}) - \Psi$ is a covering map. By compactness, it suffices to show that for each point $x \in C$ there exist a neighborhood U of x in C, and homeomorphic identifications of U and $F(V)$ with the unit disk in $\mathbf{C}$, under which $F|U$ becomes the map $z \mapsto z^n$ for some positive integer $n = n_x$. This in turn is true because f is nonconstant and is complex analytic in terms of local coordinates on C and $\mathbf{C} \cup \{\infty\}$. The integer n_x is the degree of the zero of the function $f - c$ if $c = f(x) \in \mathbf{C}$; if f has a pole at x then n_x is the order of the pole. For any $y \in \mathbf{C} \cup \{\infty\}$, we have $y \in \Psi$ if and only if $n_x > 1$ for some $x \in f^{-1}(y)$.

The fact that f is a branched covering map also implies that it has a well-defined *degree.* The degree may be defined as the degree of the covering map $f|C - f^{-1}(\Psi) : C - f^{-1}(\Psi) \to (\mathbf{C} \cup \{\infty\} - \Psi$. Thus for any $y \in (\mathbf{C} \cup \{\infty\}) - \Psi$, the number of points in $f^{-1}(y)$ is the degree of f. More generally, for any $y \in \mathbf{C} \cup \{\infty\}$, the degree of f is

$$\sum_{x \in f^{-1}(y)} n_x.$$

When y is 0 or ∞ this says that the degree counts the zeros or poles of f with multiplicity, the multiplicity of a zero or pole being its order.

In the general case, where C is not necessarily smooth, we can get a picture of the function f by resolving the singularities of C. In Section 5.3 I mentioned the process of resolving a singularity of a projective curve. If we apply this process successively at all the singular points of C we get a smooth curve, sometimes called the *normalization* of C and denoted C^ν, and a generically one-one map $\nu : C^\nu \to C$. This curve does not depend on any choices made in constructing it, but is canonically associated with C.

Now if f is a nonconstant rational function on C, and if we set $f^\nu = \nu \circ f : C^\nu \to \mathbf{C} \cup \{\infty\}$, we can define the degree of f to be the degree of f^ν. It is still the case that for any point y lying outside a suitable finite subset of $\mathbf{C} \cup \{\infty\}$, we have $\operatorname{Card} f^{-1}(y) = \deg f$.

If f is a nonconstant rational function on an *affine* curve C, and if $\hat{C}$ is a projective completion of C whose ideal points are smooth, then f extends to a rational function $\hat{f}$ on $\hat{C}$, and we can define the degree f to be the degree of $\hat{f}$. The interpretation as the generic number of points in a fiber still works.

The degree of f can also be defined from an algebraic point of view. The map f gives an identification of the function field of C with an extension of the function field of $\mathbf{C}$, which is a field of rational functions in one indeterminate. The degree of f is the degree of this extension. From this point of view there is no distinction between the smooth case and the singular case, or between the affine case and the projective case. For details, see [45].

It turns out that the study of the degree of the functions $\hat{I}_\gamma : X_0 \to \mathbf{C}$ has real applications to the study of 3-manifolds. This was first made clear by my joint work with Culler that appeared as Chapter I of [15], and was developed further in some remarkable papers by Boyer and Zhang [5], [6], [7]. I will begin the discussion of this degree in the present section. In Sections 9 and 10 I will give some topological applications. All the material in this section is extracted from [15].

8.1. Degrees of trace functions; defining the norm

Throughout this section I'll be talking about a hyperbolic 3-manifold N with one cusp; as in the earlier sections I'll let M denote its compact core, I'll choose a curve X_0 with the properties stated in Corollary 4.5.2, and I'll let $\hat{X}_0$ denote a projective completion of X_0 in which the ideal points are smooth. If we want to calculate the degree of $\hat{I}_\gamma$ for a given element $\gamma \in \operatorname{im}(\pi_1(\partial M \to \pi_1(M))$, it is in a sense simplest to do so by counting the poles of $\hat{I}_\gamma$, if only because these all occur at the finitely many ideal points of $\hat{X}_0$. If $x_1, \ldots, x_n$ are the ideal points, we can write

$$\deg I_\gamma = \sum_{i=1}^{n} P_{x_i}(I_\gamma), \tag{8.1.1}$$

where $P_{x_i}(f)$ denotes the order of the pole of a function f at x_i in the sense of Subsection 5.5: thus

$$P_{x_i}(I_\gamma) = \max(0, -v_i(I_\gamma)), \tag{8.1.2}$$

where v_i is the valuation of $\mathbf{C}(X_0)$ associated to the ideal point x_i.

In order to understand the nature of the right hand side of (8.1.2), we consider the tautological representation. Let $*$ be a base point in ∂M, let R_0 be a component of $R(\Gamma)$ that maps onto X_0 (see Subsection 5.4) and let $\mathcal{P} : \pi_1(M, *) \to \mathrm{SL}(2, \mathbf{C}(R_0))$ denote the tautological representation. This is relevant to understanding the terms in the sum (8.1.2), because by (4.4.1) we have $I_\gamma = \operatorname{trace} \mathcal{P}(\gamma)$. Since the subgroup $\Lambda = \operatorname{im}(\pi_1(\partial M, *) \to \pi_1(M, *))$ is abelian, its image under $\mathcal{P}$ is conjugate in $\mathrm{SL}(2, K)$, where K is some finite extension of $\mathbf{C}(R_0)$, either to a group of diagonal matrices or to a group of matrices of the form $\pm \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ with $\lambda \in K$. Actually the second alternative is impossible, because it would make $I_\gamma = \operatorname{trace} \mathcal{P}(\gamma)$ equal to $2 \in \mathbf{C}$, a constant function, for every $\gamma \in \Lambda$, whereas we know from Corollary 4.5.2 that these functions I_γ are all nonconstant. So there is a homomorphism η from Λ to K^*, the multiplicative group of the field K, such that

$$\eta(\gamma) = \begin{pmatrix} \eta(\gamma) & 0 \\ 0 & \eta(\gamma)^{-1} \end{pmatrix}$$

for every $\gamma \in \Lambda$. So for each $\gamma \in \Lambda$ we have

$$I_\gamma = \eta(\gamma) + \eta(\gamma)^{-1} \tag{8.1.3}$$

.

To calculate $v_i(I_\gamma)$ from (8.1.3), we first use the extension theorem for valuations, Theorem 5.4.1, to get a valuation w_i of K such that $w_i | \mathbf{C}(V) = d_i v_i$ for some positive integer d_i. Now it's an elementary exercise, starting from the definition of a valuation, to show that if w is a valuation of a field K and f is an element of K, then

$$\max(0, -w(f + f^{-1})) = |w(f)|.$$

Putting this together with (8.1.2) and (8.1.3), we get

$$P_{x_i}(I_\gamma) = \frac{1}{d_i}\max(0, -w_i(I_\gamma)) = \frac{1}{d_i}|w_i(\eta(\gamma))|.$$

To simplify the notation a little, let's set $\ell_i(\gamma) = \frac{1}{d_i}w_i(\eta(\gamma))$, so that $\ell_i : \Lambda \to \mathbf{Z}$ is a homomorphism of abelian groups for $i = 1, \ldots, n$. Then we have

$$P_{x_i}(I_\gamma) = |\ell_i(\gamma)|, \tag{8.1.4}$$

which, combined with (8.1.1), gives

$$\deg I_\gamma = \sum_{i=1}^{n} |\ell_i(\gamma)|. \tag{8.1.5}$$

So the integer-valued function on the rank-two free abelian group Λ that assigns to each $\gamma \in \Lambda$ the degree of I_γ is a function of a very special sort: it is a finite sum of functions, each of which is the absolute value of a homomorphism $\Lambda \to \mathbf{Z}$. To make this look more familiar, it is useful to think of Λ in a slightly different way. Remember that the inclusion homomorphism $\pi_1(\partial M, *) \to \pi_1(M)$ is injective, so that Λ is isomorphic in a canonical way to $\pi_1(\partial M, *)$; since $\pi_1(\partial M, *)$ is abelian, it is in turn canonically isomorphic to $H_1(\partial M, \mathbf{Z})$. Finally, the latter group can be identified with a lattice in the 2-dimensional vector space $V = H_1(\partial M, \mathbf{R})$. So we can identify Λ with this lattice by an isomorphism of groups. When we do this, each of the homomorphisms $\ell_i : \Lambda \to \mathbf{Z}$ can be extended to a linear form $V \to \mathbf{R}$, which I'll still denote ℓ_i. Then, copying the formula (8.1.5), we can define a function $\|\cdot\| : V \to [0, \infty) \subset \mathbf{R}$ by

$$\|u\| = \sum_{i=1}^{n} |\ell_i(\gamma)|, \tag{8.1.6}$$

so that $\|\gamma\| = \deg I_\gamma$ for every $\gamma \in \Lambda$. From the formula (8.1.6) we deduce immediately that

$$\|u_1 + u_2\| \leqslant \|u_1\| + \|u_2\| \tag{8.1.7}$$

for all $u_1, u_2 \in V$, and

$$\|ru\| = |r|\|u\| \tag{8.1.8}$$

for all $u \in V, r \in \mathbf{R}$.

If V is any vector space, a function $\|\cdot\| : V \to \mathbf{R}$ that satisfies (8.1.7) and (8.1.8) is called a *seminorm.* It's called a *norm* if it also satisfies

$$\|u\| > 0 \tag{8.1.9}$$

for every nonzero vector $u \in V$. Before giving a little context for these definitions, let me give the simple proof that the function $\| \cdot \|$ that I've defined on the 2-dimensional vector space $V = H_1(\partial M, \mathbf{R})$ satisfies (8.1.9) and is therefore actually a norm.

The key point is that (8.1.9) is true if $0 \neq u \in \Lambda$, because then $\|u\| = \deg I_u$, and since I_u is nonconstant according to Corollary 4.5.2, it has a strictly positive degree. Now we certainly can't have $\|u\| = 0$ for *every* $u \in V$, since $\|u\| > 0$ when $u \in \Lambda$; so in the expression (8.1.6) defining $\| \cdot \|$, the linear forms ℓ_i can't all be identically zero. After re-indexing we can assume that ℓ_1 is not identically zero. If for some vector $u_0 \neq 0$ we have $\|u_0\| = 0$, then in particular $\ell_1(u_0) = 0$, so u_0 spans the kernel of ℓ_1. But since, by construction, ℓ_1 maps Λ to $\mathbf{Z}$, the kernel of ℓ_1 is spanned by an element of Λ. So after multiplying u_0 by a nonzero constant we can assume $u_0 \in \Lambda$, and as $\|u_0\| = 0$ we now have a contradiction.

So far we have established the following properties of $\| \cdot \| : V \to \mathbf{R}$:

Property 8.1.10. *The function* $\| \cdot \| : V \to \mathbf{R}$ *is a norm;*

and

Property 8.1.11. *For each* $\gamma \in \Lambda$ *we have* $\|\gamma\| = \deg I_\gamma$.

The norm on $V = H_1(\partial M; \mathbf{R})$ that I have described is essentially the same as the one defined in Section 1 of [15]. Actually these two norms differ by a factor of 2; the reason for this will emerge.

I'll sometimes denote this norm $\| \cdot \|$ on $H_1(\partial M)$ by $\| \cdot \|_M$. If you want to justify this notation on strictly logical grounds, you will have to check that the norm depends only on M, and not on the choice of the lift ρ_0 of the discrete faithful representation to $\mathrm{SL}(2, \mathbf{C})$ that was used in defining the norm. It seems to me that this is easy enough to check, but if you don't want to go to the trouble you can just think of $\| \cdot \|_M$ as depending on a choice which is suppressed from the notation.

8.2. A word about norms

Before going on, I would like to give a brief discussion of the geometric meaning of norms on a finite-dimensional vector spaces. The most familiar example of a norm on $\mathbf{R}^n$ is of course the Euclidean norm $\| \cdot \|_E$, defined by $\|(x_1, \ldots, x_n)\|_E = \sqrt{x_1^2 + \ldots + x_n^2}$. If $\| \cdot \|$ is a norm on a vector space V one can define a metric d on V, generalizing the definition of the Euclidean metric, by setting $d(u, v) = \|u - v\|$. What is interesting about this metric in the finite-dimensional case is not the topology it defines, because it is an elementary fact, sometimes called the "equivalence of norms theorem," that any metric defined by a norm on $\mathbf{R}^n$ gives rise to the same topology as the Euclidean metric.

(In a somewhat sharper form, the equivalence of norms theorem states that for any norm $\| \cdot \|$ there are constants $C, C' > 0$ such that $\|u\| \leqslant C\|u\|_E$ and $\|u\|_E \leqslant$

$C\|u\|$ for every $u \in \mathbf{R}^n$. To prove the existence of C one writes $u = (x_1, \dots, x_n)$ in the standard basis as $\sum_{i=1}^n x_i e_i$, and uses (8.1.7) and (8.1.8) to conclude that

$$\|u\| \leqslant \sum_{i=1}^{n} |x_i| \|e_i\| \leqslant \|u\|_E \sum_{i=1}^{n} \|e_i\|,$$

so that we can take C to be $\sum_{i=1}^n \|e_i\|$. The existence of C implies in particular that $\|\cdot\|$ is continuous in the usual topology of $\mathbf{R}^n$. Since, by (8.1.9), $\|\cdot\|$ is strictly positive on the Euclidean unit sphere S^{n-1}, it now follows that $\|\cdot\|$ takes a minimum value $c > 0$ on S^{n-1}. The proof of the equivalence of norms theorem is now completed by setting $C' = \frac{1}{c}$ and invoking (8.1.8) again.)

What *is* interesting about a norm (or the associated metric) on a finite-dimensional vector space is the geometric structure to which it gives rise. Specifically, if $\|\cdot\|$ is a norm on a vector space V, then the ball B of radius 1 about the origin, consisting of all $u \in V$ such that $\|u\| \leqslant 1$, is a convex, compact subset of V having 0 as an interior point. (Convexity is immediate from properties (8.1.7) and (8.1.8). In proving that B is compact and that $0 \in \operatorname{int} B$, we may assume that $V = \mathbf{R}^n$; in this case the assertions follow from the equivalence of norms theorem.) It follows from (8.1.9) that B is also *balanced* in the sense that $-u \in B$ whenever $u \in B$. Conversely, if $B \subset V$ is a balanced, convex, compact set with $0 \in \operatorname{int} B$, it is clear that for each $u \in \mathbf{R}^n$ there is a nonnegative real number $r \in [0, \infty)$ such that $u = ru_0$ for some $u_0 \in B$. By compactness there is in fact a least such r, say $r = r_u$. It is a straightforward exercise to check that the function $\|\cdot\| : V \to [0, \infty)$ defined by $\|u\| = r_u$ is a norm, and that this construction is precisely the inverse of the construction that associates to each norm its unit ball. So when V is finite-dimensional we have a canonical bijection between norms on V and balanced, convex, compact sets whose interiors contain the origin. A norm is an appealing algebraic way of encoding the structure of a certain kind of geometric object.

Sometimes I'll find it convenient to look at the ball

$$B_r = \{u \in V : \|u\| \leqslant r\}$$

of radius r associated to a norm $\|\cdot\|$, where r is a positive number not necessarily equal to 1. The difference between B_r and B_1 is not a big deal, because (8.1.8) says that $B_r = rB_1 = \{ru : u \in B_1\}$. We can also think of B_r as the unit ball for $r\|\cdot\|$, which according to the definition is itself a norm on V.

Another useful fact, which is also easy to prove by the elementary methods I've been talking about, is that the unit sphere of a norm $\|\cdot\|$, i. e. the set of all $u \in V$ with $\|u\| = 1$, is precisely the boundary of its unit ball.

8.3. Further properties of the norm

Now I'll discuss some more properties of the norm $\|\cdot\| = \|\cdot\|_M$ on $V = H_1(\partial M)$, where M is the compact core of a one-cusped orientable finite-volume hyperbolic

3-manifold, and the associated balanced convex set $B \subset V$. The first thing to notice is that because the expression (8.1.6) that was used to define $\|\cdot\|$ is a finite sum of absolute values of linear forms, the unit sphere ∂B is a polygon, i.e. a finite union of line segments. This may seem almost obvious, but giving a careful proof of it leads to important information. To begin with, we note that of the linear forms ℓ_{x_i} that appear in the expression (8.1.6), it may happen that some are identically zero. After reindexing the ℓ_i, if need be, we may assume that there is a natural number $k \leqslant n$ such that

$$\|u\| = \sum_{i=1}^{k} |\ell_i(\gamma)| \tag{8.3.1}$$

for every $u \in V$, and ℓ_i is not identically zero for any $i \leqslant k$. (Actually we must have $k \geqslant 2$, as otherwise $\|\cdot\|$ wouldn't satisfy (8.1.9).) The kernel of ℓ_i, for each $i \leqslant n$, is a line L_i through the origin in the plane V. Each L_i divides the plane into two half-planes, H_i^+ and H_i^-, such that ℓ_i is $\geqslant 0$ on H_i^+ and $\leqslant 0$ on H_i^-.

The lines $L_1, \ldots, L_k$ divide the plane into $2k$ sectors. Let Σ denote any one of these sectors. For each $i \leqslant k$, the sector Σ is contained in either H_i^+ or H_i^-. Hence the i-th term $\|\ell_i\|$ in the sum (8.3.1) is identically equal on Σ either to the linear form ℓ_i or to the linear form $-\ell_i$. It follows that $\|\cdot\||\Sigma$ coincides with the restriction to Σ of a function λ_Σ which is a finite sum of linear forms, and is therefore itself a linear form. The intersection of Σ with the unit ball ∂B of $\|\cdot\|$ coincides with the intersection of Σ with the line on which λ_Σ is equal to 1. This is a line segment. It follows that ∂B is a polygon, as it is the union of the $2k$ line segments $\partial B \cap \Sigma$, where Σ ranges over the sectors.

However, this argument shows more. Since the line segment $\partial B \cap \Sigma$ is the intersection of Σ with a line not passing through 0, the endpoints of this segment must lie on the rays that make up the frontier of Σ. Each of these rays is contained in one of the lines L_i. So every vertex of ∂B lies on one of the L_i. What's neat is that the lines L_i have direct topological meaning in terms of the 3-manifold M. Since the definition of the ℓ_i involved ideal points of the curve $\hat{X}_0$, you will probably guess that the meaning of the L_i has something to do with essential surfaces; and you will not be wrong.

As I mentioned in Subsection 5.6, the term "slope" is used to indicate an unoriented isotopy class of simple closed curves in ∂M. These are in bijective correspondence with indivisible elements of $\pi_1(\partial M)$ modulo sign. In this section we are identifying $\pi_1(\partial M)$ with the lattice $\Lambda \subset V$. I'll talk about the "slope of an indivisible element of Λ," so that there are just two inidivisible elements with any given slope and they differ by sign. In 5.6 I also pointed out that there is a "boundary slope" associated with each bounded essential surface $F \subset M$. I'll call an element of Λ a *boundary class* if its slope is a boundary slope. Two indivisible elements of Λ which differ by sign span the same 1-dimensional subspace of V. So corresponding to each "slope" there is a line through the origin in the plane V. I'll call a line through the origin a *boundary line* if it corresponds to a boundary slope. According to the theorem of Hatcher's that I talked about in Subsection 5.6, only finitely many lines in V occur as boundary lines of bounded essential surfaces in M.

Proposition 8.3.2. *Each of the lines $L_1, \ldots, L_k$ is the boundary line of some bounded essential surface in M. So each vertex of the polygon ∂B lies on a boundary line.*

To prove this, we recall that for each $i \leqslant k$ the linear form ℓ_i restricts to a nontrivial homomorphism from Λ to $\mathbf{Z}$, so that $\ker \ell_i$ is a direct factor of Λ and is therefore spanned by some indivisible element γ_i, which is the homology class determined up to sign by some simple closed curve $C \subset \partial M$. What we have to prove is that there is a bounded essential surface $F \subset M$ whose boundary components are all isotopic in ∂M to C.

Remember that L_i is the kernel of the linear form ℓ_i, which is defined in terms of an ideal point x_i of X_0. According to Section 5, there is an action of $\pi_1(M)$ on a tree T_i associated to x_i. Since $\gamma_i \in \ker \ell_i$, we find from (8.1.4) that

$$P_{x_i}(I_{\gamma_i}) = |\ell_i(\gamma_i)| = 0.$$

In other words, I_{γ_i} does not have a pole at x_i. It therefore follows from Property 5.4.2 that γ_i fixes a vertex of T_i.

Now we argue as in Section 7. Let $|\gamma_i|$ denote a simple closed curve realizing the slope of γ_i. Since γ_i generates the image of the fundamental group of $|\gamma_i| \subset \partial M$ in $\pi_1(M)$, it follows from Corollary 6.0.1 that there is an essential surface $F \subset M$, dual to the action of M on T, such that $F \cap |\gamma_i| = \emptyset$. Thus either F has boundary components which are all parallel to $|\gamma_i|$—i. e. its boundary class is γ_i—or it is closed.

Of course the conclusion that γ_i is a boundary class is exactly what we want, because it says that L_i is a boundary line. So we need to rule out the possibility that F is closed. If we assume F is closed, then $\partial M \subset M - F$, so by 2.3.1(i) the subgroup Λ of $\pi_1(M)$ fixes a vertex of T_i. This assertion makes sense even though Λ, as a subgroup of $\pi_1(M)$, is defined only up to conjugation: if a given subgroup fixes some vertex of T, then any conjugate subgroup also fixes some—possibly different—vertex.) By Property 5.4.2 and (8.1.4) it then follows that

$$|\ell_i(\gamma)| = P_{x_i}(I_\gamma) = 0$$

for every $\gamma \in \Lambda$, i.e. that ℓ_i is identically zero. But this is false since $i \leqslant k$. The proof of Proposition 8.3.2 is now complete.

Everything I have said up to now about the norm has involved calculating the degrees of the functions I_γ in terms of poles. The degree of a function can also be calculated in terms of its zeros; in the next section I'll show how to get new information about the norm by studying the zeros of certain functions closely related to the I_γ, and comparing this with the information we already have.

9. Applications to Dehn surgery

9.1. The Cyclic Surgery Theorem

The first application of the norm $\|\cdot\|_M$ that I talked about in the last section was given in [15], where it was used to prove the Cyclic Surgery Theorem. I will refer you to Boyer's chapter in this volume for the motivation, but I will review the statement of the Cyclic Surgery Theorem here. If M is a compact, orientable 3-manifold whose boundary is a torus, and α is a slope (i. e. an isotopy class of simple closed curves on ∂M), I will denote by $M(\alpha)$ the manifold obtained from M by the Dehn filling with filling slope α. By definition, this means that $M(\alpha)$ is obtained from the disjoint union $M \amalg (S^1 \times D^2)$ by gluing the boundaries of M and $S^1 \times D^2$ via some homeomorphism which maps a simple closed curve representing the slope of α to a curve $\{*\} \times \partial D^2$ for some point $* \in S^1$.

Theorem 9.1.1. (Culler, Gordon, Luecke, Shalen [15]) *Let M be a compact, orientable 3-manifold whose boundary is a torus. Suppose that M is irreducible but is not a Seifert fibered space. Let α and β be indivisible elements of $\Lambda = \mathrm{im}(\pi_1(\partial M) \to \pi_1(M)$ such that $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ are cyclic. Then the geometric intersection number $\Delta(\alpha, \beta)$ is at most one.*

9.2. The "no-closed-surface" case

The proof of the Cyclic Surgery Theorem turns out to be a lot simpler in the special case in which we assume that M contains no closed essential surfaces. I will spend most of this section discussing the proof in this case first. In Subsection 9.5 I'll come back and say a little about the refinements that one has to make to handle the general case.

So for the rest of this subsection I will be assuming that M contains no closed essential surfaces.

In particular M contains no essential tori; and since the hypothesis of the theorem rules out the Seifert fibered case, it follows from Thurston's geometrization theorem that the interior of M—call it N—has a hyperbolic structure of finite volume. Since ∂M is one torus, N has one cusp. Let's fix a curve X_0 with the properties stated in Corollary 4.5.2. As in Section 8 I'll identify $\Lambda = \mathrm{im}(\pi_1(\partial M) \to \pi_1(M)$ with $H_1(\partial M; \mathbf{Z})$, which I'll think of as a lattice in the 2-dimensional vector space $V = H_1(\partial M; \mathbf{R})$. As in Section 8 we have a norm $\|\cdot\| = \|\cdot\|_M$ defined on V. A key step in the proof of the cyclic surgery theorem in this case turns out to be extracting information about the norm of an element α of Λ from the condition that $\pi_1(M(\alpha))$ is cyclic.

The following simple but arresting lemma, which was discovered by Marc Culler, gives the first indication that the character variety may be useful in studying classes $\alpha \in \Lambda$ for which $M(\alpha)$ is cyclic. It was the starting point for Culler's and my work on this subject.

Lemma 9.2.1. *Let α be an indivisible element of Λ such that $\pi_1(M(\alpha))$ is cyclic. Let χ be a point of X_0 such that $I_\alpha(\chi) = \pm 2$. Then $I_\gamma(\chi) = \pm 2$ for every $\gamma \in \Lambda$.*

Proof. By Lemma 7.2.1, applied to $N = \operatorname{int} M$, there is a representation $\rho \in t^{-1}(\chi)$ such that the image of $\rho(\pi_1(M))$ under the natural homomorphism $p : \mathrm{SL}(2, \mathbf{C}) \to \mathrm{PSL}(2, \mathbf{C})$ is noncyclic. If $\rho(\alpha) = \pm I$, then $p \circ \rho : \pi_1(M) \to \mathrm{PSL}(2, \mathbf{C})$ maps α to the identity element of $\mathrm{PSL}(2, \mathbf{C})$ and hence factors through a homomorphism of $|\pi_1(M) : \alpha = 1|$ onto the noncyclic group $p(\rho(\pi_1(M)))$. This is impossible, since $|\pi_1(M) : \alpha = 1| \cong \pi_1(M(\alpha))$ is noncyclic. So $\rho(\alpha) \neq \pm I$. On the o ther hand, we have $\operatorname{trace} \rho(\alpha) = I_\alpha(\chi) = \pm 2$. So after composing ρ with a suitable inner automorphism of $\mathrm{SL}(2, \mathbf{C})$ we can assume that $\rho(\alpha) = \pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. (Of course a conjugation doesn't change the character of ρ.)

Now for any $\gamma \in \Lambda$ the matrix $\rho(\gamma)$ commutes with $\rho(\alpha) = \pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, since Λ is abelian. So $\rho(\gamma)$ must have the form $\pm \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$. In particular $I_\gamma(\chi) = \operatorname{trace} \rho(\gamma) = \pm 2$. This proves the lemma. Is that cool, or what?

In thinking about the consequences of Lemma 9.2.1, it's nice to think in terms of the functions $f_\gamma = I_\gamma^2 - 4 : X_0 \to \mathbf{C}$ which are defined for all $\gamma \in \Lambda$. It's very easy to see that squaring a function doubles its degree, and that adding a constant doesn't change the degree, so in terms of the norm $\| \cdot \|_M$ defined in Section 8 we have

$$\deg f_\gamma = 2 \deg I_\gamma = 2\|\gamma\| \tag{9.2.2}$$

for every $\gamma \in \Lambda$. Now Lemma 9.2.1 says that if $\pi_1(M(\alpha))$ is cyclic, then at every point of X_0 where f_α takes the value 0, all the functions f_γ for $\gamma \in \Lambda$ take the value 0.

Recall from Corollary 4.5.2 that if γ is any nontrivial element of Λ then I_γ is nonconstant on X_0; hence so is f_γ. If we allow ourselves to think in very fuzzy terms for a moment, we can think of the degree of f_γ as counting the points where f_γ takes the value 0, and Lemma 9.2.1 says that f_γ, for an arbitrary $\gamma \in \Lambda$, takes the value 0 wherever f_α does; this suggests the

Fuzzy Idea 9.2.3. *Maybe we should expect to have $\deg f_\alpha \leqslant \deg f_\gamma$, and hence $\|\alpha\| \leqslant \|\gamma\|$, for every $\gamma \in \Lambda$, if $\pi_1(M(\alpha))$ is cyclic. In other words, when $\pi_1(M(\alpha))$ is cyclic, maybe the the norm of α should be minimal among all norms of nontrivial elements of the lattice Λ.*

However, we have not come even close to proving this. This is because calculating the degree of a function in terms of its zeros makes sense only when the domain is a smooth projective curve, and even in this case the orders of the zeros must be taken into account. The curve X_0 is not projective, it need not be smooth, and we have ignored orders.

Of these three issues, the last two turn out to be essentially technical. First of all, the proof of Lemma 9.2.1 can be souped up, by an argument involving the tautological representation and valuations of function fields, to show that if χ is a smooth point of X_0 where f_α has a zero of a certain order, then f_γ has a zero of at least the same order. To put this in symbols, let's write $Z_x(f)$ to mean the order of zero of a function f at a smooth point of a curve, and to mean 0 if f does not have a zero at x: thus $Z_x(f) = \max(0, v_x(f))$, where v_x is the valuation of the function field associated to the smooth point x. Then the souped-up version of Lemma 9.2.1 says that if α and γ are elements of Λ with α indivisible and $\pi_1(M(\alpha))$ cyclic, then $Z_\chi(f_\alpha) \leqslant Z_\chi(f_\gamma)$ for every smooth point χ of X_0.

There's also a version of this that works for nonsmooth points of X_0. Let's set $X_0^\nu = \nu^{-1}(X_0) \subset \hat{X}_0^\nu$. The following result is the ultimate version of Lemma 9.2.1.

Theorem 9.2.4. *Let α be an indivisible element of Λ such that $\pi_1(M(\alpha))$ is cyclic. Then for every point x of X_0^ν we have*

$$Z_x(f_\alpha^\nu) \leqslant Z_x(f_\gamma^\nu).$$

Again, the most important thing to say about the proof of this theorem is that it's a essentially an elaborate technical refinement of the simple proof of Lemma 9.2.1. As amusing as the algebra involved is, I will fight back the impulse to give any of the argument, but I will make one comment to help get you into the right mood for reading the details in [15]. Although on the face of it X_0^ν would appear to be only a subset of a projective curve, it turns out to have the structure of an affine curve, the *affine normalization* of X_0. From the algebraic point of view, the coordinate ring $\mathbf{C}[X^\nu]$ is the *integral closure* of $\mathbf{C}[X]$.

Theorem 9.2.4 deals with two of the issues I mentioned above in connection the proposed "proof" of the Fuzzy Idea 9.2.3. The remaining issue concerns ideal points, and is different in nature. In fact, it should be clear by now that anything related to ideal points of X_0 has something to do with the topology of M, specifically with the essential surfaces in M.

The natural context for all this is provided by the normalized projective completion $\hat{X}_0^\nu$. According to the discussion in the introduction to Section 8, we have

$$\deg f = \deg \hat{f}^\nu = \sum_{x \in \hat{X}_0^\nu} Z_x(\hat{f}^\nu). \tag{9.2.5}$$

Of course this last expression is really a finite sum since $Z_x(\hat{f}^\nu) = 0$ for all but finitely many points $x \in \hat{X}_0^\nu$.

Now let α be an indivisible element of Λ such that $\pi_1(M(\alpha))$ is cyclic. According to (9.2.5), the inequality proposed in 9.2.3 will hold if $Z_x(f_\alpha^\nu) \leqslant Z_x(f_\gamma^\nu)$ for every $x \in \hat{X}_0^\nu$; and according to Theorem 9.2.4, this is true whenever $x \in X_0^\nu$.

Now suppose that $x \in \hat{X}_0^\nu - X_0^\nu$. We would like to know that $Z_x(f_\alpha^\nu) \leqslant Z_x(f_\gamma^\nu)$ in this case. This is not even an issue unless $Z_x(f_\alpha^\nu) \neq 0$, so that $I_\alpha(\nu(x)) = \pm 2$. Suppose this happens, and let's look at the tree T associated to the ideal point

$\nu(x)$ of $\hat{X}_0$. Since I_α takes the finite value 2 at x it follows from 5 that α fixes a vertex of T. Let C be a simple closed curve in ∂M realizing the slope of α. Since α generates the image of the fundamental group of $C \subset \partial M$ in $\pi_1(M)$, it follows from Corollary 6.0.1 that there is an essential surface $F \subset M$, dual to the action of $\pi_1(M)$ on T, such that $F \cap C = \emptyset$. Thus either F has boundary components which are all parallel to C—i. e. its boundary slope is the slope of α—or it is closed.

We are excluding from the discussion the case where M contains closed essential surfaces, so we need only worry about the case where F is bounded and has α as a boundary class. In this case, where α is a boundary class of M, we need to resort to some serious topology. The relevant argument here is due to Gordon. If F is any essential surface in M having α as a boundary class, we can use F to construct a closed bicollared surface $F(\alpha)$ in the manifold $M(\alpha)$. After all, $M(\alpha)$ is constructed from M by attaching a solid torus K in which α is a meridian; since the components of ∂F are meridians in K, we can attach disjoint disks in K to the boundary components of F, and this gives the surface $F(\alpha)$. Let me suspend the conventions of this subsection for a moment so that I can state the following result with all relevant hypotheses so as to make its self-contained nature clear.

Theorem 9.2.6 (Gordon)**.** *Let M^3 be compact, connected, orientable and irreducible, with ∂M a torus. Suppose that M contains no closed essential surfaces. Let α be a boundary class for M. Among all essential surfaces in M having α as a boundary class, α, let us choose one, say F_0, whose number of boundary components is minimal. Then either (i) F has strictly positive genus and $F(\alpha)$ is an essential surface in $M(\alpha)$, or (ii) F has genus 0, so that $F(\alpha)$ is a 2-sphere; and either F is a fiber in a fibration of M over S^1, or $F(\alpha)$ decomposes $M(\alpha)$ as the connected sum of two nontrivial lens spaces.*

A lens space is *nontrivial* if it's not a 3-sphere. To say that $F(\alpha)$ decomposes $M(\alpha)$ as the connected sum of two nontrivial lens spaces means that the closure of each component of $M(\alpha) - F(\alpha)$ is homeomorphic to the manifold obtained from a nontrivial lens space by removing the interior of a 3-ball with bicollared boundary. Such a ball is unique up to ambient isotopy in $M(\alpha)$; see [32].

Note that if (i) holds then $\pi_1(M(\alpha))$ contains an isomorphic copy of $\pi_1(F(\alpha))$, where $F(\alpha)$ is a closed, orientable surface of positive genus; and that if (ii) holds, $\pi_1(M(\alpha))$ is a free product of two nontrivial cyclic groups. So:

Corollary 9.2.7. *Under the hypotheses of Theorem 9.2.6, $\pi_1(M(\alpha))$ is not cyclic.*

Resuming the conventions of this subsection, we can now go back to considering a class $\alpha \in \Lambda$ such that $\pi_1(M(\alpha))$ is cyclic. If I_α takes a finite value at some ideal point of X_0 then we have seen that, under the assumption that M contains no closed essential surfaces, α must be a boundary class; and this contradicts Corollary 9.2.7. So I_α must have a pole at every ideal point, which means that the inequality of Theorem 9.2.4 holds even when $x \in \hat{X}_0^\nu - X_0^\nu$, since in that case the left hand side is zero. So we can sum the inequality over all $x \in \hat{X}_0^\nu$, and according to (9.2.5) we get

$$\deg f_\alpha \leqslant \deg f_\gamma$$

whenever α and γ are elements of Λ such that α is indivisible and $\pi_1(M(\alpha))$ is cyclic. Thus the "fuzzy idea" 9.2.3 turns out to be entirely justified in the case where M contains no closed essential surfaces, and we may state it as a

Theorem 9.2.8. *If $\pi_1(M(\alpha))$ is cyclic then $\|\alpha\|$ is minimal among all norms of nonzero elements of Λ. (Let me emphasize that this theorem depends on the blanket assumption made in this section that M contains no closed essential surfaces.)*

In order to make the transition between Theorem 9.2.8 and the conclusion of the Cyclic Surgery Theorem, we set $m = \min_{0 \neq \gamma \in \Lambda} \|\gamma\|$, so that m is a natural number and any $\alpha \in \Lambda$ with $\pi_1(M(\alpha))$ cyclic has norm m. The argument is based on considering the ball B_m of radius m for the norm $\| \cdot \| = \| \cdot \|_M$. From what I said in Subsections 8.2 and 8.3 it follows that the unit ball B of $\| \cdot \|$ is a compact, convex, balanced set bounded by a polygon, and hence so is $B_m = mB$. We can paraphrase Theorem 9.2.8 by saying that $\operatorname{int} B_m$ contains no points of Λ except 0.

It happens that number-theorists have long been interested in properties of convex, balanced polygons whose interiors contain no points of a give lattice. In the next subsection I will state and prove a theorem on this subject, which is one of the simplest results in Minkowski's "geometry of numbers," and illustrate how powerful it is in number theory. Then, in Subsection 9.4, I will show how to use Minkowski's theorem to finish the proof of the Cyclic Surgery Theorem in the case where there is no closed essential surface.

9.3. A little geometry of numbers: a theorem of Minkowski's

Suppose that V is an n-dimensional vector space. The volume element (Lebesgue measure) on V, which we can think of as a function that assigns a nonnegative real number to every bounded, measurable set in V, is well-defined up to multiplication by a positive constant. A convex set is always measurable, so it has a well-defined volume once we have fixed a volume element. If Λ is a lattice in V, i.e. a discrete, cocompact subgroup of the additive group of V, then the quotient V/Λ, which is topologically an n-torus, inherits a volume element when we fix a volume element on V; the volume of the whole torus V/Λ is then called the covolume of Λ. Of course, multiplying the volume element by a positive constant has the effect of multiplying the covolume by the same constant. Hence the conclusion of the following theorem is independent of the choice of a volume element.

Theorem 9.3.1 (Minkowski)**.** *Let Λ be a lattice in an n-dimensional vector space V. Let B be a compact, convex, balanced subset of V such that* $\operatorname{int} B$ *contains no point of Λ except 0. Then*

$$\operatorname{vol} B \leqslant 2^n \operatorname{covol} \Lambda.$$

Proof. We may take the volume element to be normalized so that $\operatorname{covol}\Lambda = 1$. It follows from convexity that $\operatorname{int} B$ has the same volume as B. The volume element on V induces a volume element on $V/2\Lambda$ as well as on V/Λ. Since V/Λ has volume 1, the volume of $V/2\Lambda$ is 2^n. So it's enough to show that the natural map $p : V \to V/2\Lambda$ maps $\operatorname{int} B$ injectively.

Let x and y be points of B with $p(x) = p(y)$; then $x - y \in 2\Lambda$, i.e. $\frac{x-y}{2} \in \Lambda$. But since $\operatorname{int} B$ is convex and balanced, and contains x and y, we have $\frac{x-y}{2} \in B$. The hypothesis concerning B now implies that $\frac{x-y}{2} = 0$, so $x = y$. This is all there is to the proof of the theorem.

(This is logically equivalent to the proof you will find in books on the subject, but the use of the torus $V/2\Lambda$ makes it more intuitive. John Morgan pointed out this version of the argument to me several years ago.) □

I will illustrate the use of Minkowski's theorem in number theory by showing how it can be used to prove Lagrange's famous theorem that every positive integer is the sum of four squares. I've adapted this from [48]. One way of paraphrasing the statement that a given n is a sum $a^2 + b^2 + c^2 + d^2$ of four squares of integers is to say that n is the squared absolute value $|h|^2$ of a (Hamiltonian) quaternion $h = a + bi + cj + dk$ with integer "coordinates." Since we have $|hk| = |h||k|$ for any quaternions h and k, the property of being a sum of four squares is preserved under the formation of products; so it's enough to prove that every prime p is the sum of four squares. The case $p = 2$ is clear, so let's take p odd.

Having made this reduction, we now forget about quaternions, and reinterpret our goal more naïvely as showing that p is the square of the Euclidean norm $\|v\|$ of a point $v = (a, b, c, d)$ in the standard lattice $\mathbf{Z}^4 \subset \mathbf{R}^4$. The approach is to construct a subgroup Λ of $\mathbf{Z}^4$ which has index p^2, and is therefore a lattice of covolume p^2 in terms of the Euclidean volume element, and such that for every $v \in \Lambda$ we have $\|v\|^2 \equiv 0 \pmod{p}$. If we have such a Λ, and if B denotes the ball of radius $\sqrt{2p}$ about the origin, then by elementary calculus we find that

$$\operatorname{vol} B = 2\pi^2 p^2 > 16p^2 = 2^4 \operatorname{covol}\Lambda,$$

so that by Minkowski's theorem, there is a nonzero element v of $\Lambda \cap \operatorname{int} B$. Then $0 < \|v\|^2 < 2p$, but the property we're assuming for Λ says that $\|v\|^2$ is divisible by p; so we must have $\|v\|^2 = p$, which is what we need.

The best way to find a Λ with the right properties is to think of $\mathbf{R}^4$ in yet a third way, as the complex vector space $\mathbf{C}^2$, in which case the lattice $\mathbf{Z}^4$ becomes $\mathcal{O}^2$, where $\mathcal{O} = \mathbf{Z}[i]$ is the ring of Gaussian integers, consisting of all complex numbers $a + ib$ with $a, b \in \mathbf{Z}$. Given any $s \in \mathcal{O}$, we can define a homomorphism of additive groups $H : \mathcal{O}^2 \to \mathcal{O}$ by $H_s(z, w) = z - sw$. Then $\Lambda = \Lambda_s = H_s^{-1}(p\mathcal{O})$ is a subgroup of index p^2 in $\mathcal{O}^2$, and for any $(z, w) \in \Lambda$ we have $|z|^2 \equiv |s|^2|w|^2 \pmod{p}$. Hence if we can choose $s \in \mathcal{O}$ so that $|s|^2 \equiv -1 \pmod{p}$, it will follow that $\|(z, w)\|^2 = |z|^2 + |w|^2 \equiv 0(\text{mod } p)$ for any $(z, w) \in \Lambda$, as required.

But the existence of such an s is easy. The homomorphism $x \to x^2$ from the multiplicative group $(\mathbf{Z}/p\mathbf{Z})^*$ to itself has kernel $\{\pm 1\}$ and hence has image of order $(p-1)/2$, so the set S of squares in $\mathbf{Z}/p\mathbf{Z}$ has cardinality $(p+1)/2$. It follows that the set $T \subset \mathbf{Z}/p\mathbf{Z}$ consisting of elements of the form $1 - x^2$ also has $(p+1)/2$ elements, and the paucity of pigeon holes forces S and T to intersect. So there exist integers u and v such that $-1 - u^2 \equiv v^2 \pmod{p}$, and $s = u + iv$ then satisfies $|s|^2 \equiv -1 \pmod{p}$.

9.4. The "no-closed-surface" case, concluded

To complete the proof of the Cyclic Surgery Theorem in the case where M contains no closed essential surfaces, we apply Minkowski's theorem to the set B_m in our 2-dimensional vector space $V = H_1(M, \mathbf{R})$. Let's normalize the volume element on V in such a way that our lattice $\Lambda = H_1(M, \mathbf{Z})$ has co-area 1. The relevance of the theorem comes from the observation that, under this normalization, if γ and γ' are two elements of Λ, then the area of the parallelogram with vertices $\pm\gamma, \pm\gamma'$ is just $2\Delta(\gamma, \gamma')$.

Suppose that α and β are indivisible elements of Λ with $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ cyclic. By Theorem 9.2.8 we have $\|\alpha\| = \|\beta\| = m$, so that $\alpha, \beta \in \partial B_m$. The parallelogram P with vertices $\pm\alpha, \pm\beta$ is therefore contained in B_m, so that

$$\Delta(\alpha, \beta) = \frac{1}{2}\operatorname{area} P \leqslant \frac{1}{2}\operatorname{area} B \leqslant 2, \tag{9.4.1}$$

where in the last step we have used Theorem 9.3.1 to conclude that $\operatorname{area} B \leqslant 4$.

To complete the proof of the Cyclic Surgery Theorem in this case we need only rule out the possibility that all the inequalities in (9.4.1) are equalities. If this happens, then in particular α and β are vertices of B_m. Now according to Proposition 8.3.2, each of the vertices of B, and hence of B_m, lies on a boundary line defined by some bounded essential surface in M. So in this situation we conclude that α and β are boundary classes. However, in the situation of this section, knowing that even *one* of the classes α and β is a boundary class is enough to give a contradiction. This follows from Corollary 9.2.7, since $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ have both been assumed cyclic, and M contains no essential surfaces.

And that is how the Cyclic Surgery Theorem is proved in the special case I've been talking about.

9.5. The case where there are closed surfaces

In the general case, the proof of the Cyclic Surgery Theorem breaks up into three cases. For two of the cases–the case in which either α or β is a boundary slope, and the case in which $\operatorname{int} M$ is not hyperbolic—I will refer you to Boyer's chapter in this volume. In the case that neither α nor β is a boundary slope, the proof is a refinement of the proof that I gave in the special case where M contains no closed essential surfaces. The main step is to prove the following result, which is mostly a refinement of the proof of Theorem 9.2.8.

Theorem 9.5.1. *Let M be a manifold with a single torus boundary such that* $\operatorname{int} M$ *has a hyperbolic structure of finite volume, and let $\|\cdot\|$ be the norm on $\Lambda = H_1(\partial M)$ defined in Section 8. Let α be an indivisible element of Λ which is not a boundary class, and suppose that $\pi_1(M(\alpha))$ is cyclic. Then $\|\alpha\|$ is minimal among all norms of nonzero elements of Λ.*

The basic strategy used in the proof is the same as in the proof of Theorem 9.2.8: one shows that for any point x of the curve $\hat{X}^\nu$ where $f(\alpha)$ has a zero, f_γ has a zero of at least the same order for each nontrivial element γ of Λ. When $x \in X^\nu \subset \hat{X}^\nu$ this is proved in exactly the same way as above. If x projects to an ideal point of $\hat{X}$, the above arguments give an action of $\pi_1(M)$ on a tree T under which α fixes a vertex. As before, we can associate an essential surface F with this action, and we can take it to be disjoint from a simple closed curve realizing the slope of α. If F had a nonempty boundary, its boundary slope would be α, and we would have a contradiction. The difficulty is that F may now be closed. Section I.6 of [15] is devoted to the proof that in this situation, if $Z_x(f_\alpha) > Z_x(f_\gamma)$, then we can replace a given closed surface F dual to the action of $\pi_1(M)$ on T by a new dual surface which has nonempty boundary, and has α as a boundary class. The starting point is the observation that since $\pi_1(M)$ is cyclic, there must be a compressing disk for F in M. See [15] for more.

9.6. Other applications to surgery

In [7], Boyer and Zhang proved the following analogue of the Cyclic Surgery Theorem:

Finite Surgery Theorem (Boyer-Zhang). *Let M be a compact, orientable 3-manifold whose boundary is a torus. Suppose (for simplicity) that* $\operatorname{int} M$ *has a hyperbolic metric of finite volume. Let α and β be indivisible elements of Λ such that $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ are finite. Then the geometric intersection number $\Delta(\alpha, \beta)$ is at most three. Furthermore, up to sign there are at most five indivisible elements α of Λ such that $\pi_1(M(\alpha))$ is finite.*

The bounds of three and five are both sharp. For more discussion of the context of the statement, see Boyer's chapter in this volume.

A key step in the proof of the Finite Surgery Theorem is the following surprising analogue of Theorem 9.5.1, which was proved in [5]; I will have to refer you to [5] for an account of the very remarkable new ideas that enter into the proof.

Theorem 9.6.1. *Let M be a manifold with a single torus boundary such that* $\operatorname{int} M$ *has a hyperbolic structure of finite volume, and let $\|\cdot\|$ be the norm on $\Lambda = H_1(\partial M)$*

defined in Section 8. Let α be an indivisible element of Λ which is not a boundary class, and suppose that $\pi_1(M(\alpha))$ is finite. Then

$$\|\alpha\| \leqslant \max\{2m, m+8\},$$

where

$$m = \min_{0 \neq \gamma \in \Lambda} \|\gamma\|.$$

Using Theorem 9.6.1 and the kinds of arguments that I've discussed earlier in this chapter, Boyer and Zhang were able to show that the bound of 3 on the geometric intersection number asserted in the Finite Surgery Theorem holds unless the unit ball for the norm $\|\cdot\| = \|\cdot\|_M$ is of a special type. The same argument works if one replaces the usual norm $\|\cdot\|_M$ by a slightly different norm $\|\cdot\|'_M$; the definition of $\|\cdot\|'_M$ is just like that of $\|\cdot\|_M$, except that in place of the usual curve X_0 one uses the possibly reducible curve X_1 obtained by saturating X_0 under the action of the Galois group of $\mathbf{C}$ over $\mathbf{Q}$. The completion of the proof of the Finite Surgery Theorem given in [7] essentially involves showing that polygons of these exceptional types do not arise as unit balls of norms $\|\cdot\|'_M$ for 3-manifolds M whose interiors have 1-cusped hyperbolic structures. In order to do this, Boyer and Zhang needed to interpret the polygons arising from 3-manifolds from a new point of view, based on the theory of the so-called A-polynomial, which was developed in [11] and [13], among other papers.

If M is a compact 3-manifold with torus boundary, there is a natural map r from the character variety $X(\pi_1(M))$ to $X(\pi_1(\partial M))$: the image under r of a character of $\pi_1(M)$ is its precomposition with the inclusion homomorphism $\pi_1(\partial M) \to \pi_1(M)$. Basically the A-polynomial—or the slight variant of it used in [5], which I'll call the A'-polynomial—gives information about the image $r(X_1)$ in the case where $\operatorname{int} M$ has a 1-cusped hyperbolic structure and X_1 is the Galois-saturated curve I've just described. It turns out that there is a curve $Y \subset \mathbf{C}^2$ which admits a canonical degree-two rational map to $r(X_1)$. The A'-polynomial of M is a canonically defined two-variable integer polynomial whose locus of zeros is Y. (The A-polynomial is defined similarly except that in place of X_1 one uses a curve X_2 that's possibly still bigger than X_1 in the sense that it may have still more irreducible components.)

Boyer and Zhang re-interpreted the unit ball of the norm $\|\cdot\|_M$ as a "geometric dual" to the so-called Newton polygon of the A'-polynomial, which is another convex plane polygon obtained from the curve A' by an algebro-geometric construction. Then they adapted to the A'-polynomial properties of the A-polynomial established in [11], and the more surprising properties established by Cooper and Long in [13], to deduce restrictions on the unit ball of $\|\cdot\|_M$ which rule out the exceptional polygons that arise in the proof of the Finite Surgery Theorem. For a survey of the relevant material on the A-polynomial, see Cooper and Long's paper [14].

In [23], Dunfield established a fundamental property of the map r, namely that $r|X_0 : X_0 \to r(X_0)$ is a birational map, which is to say that it has degree one. His ingenious proof, besides using the result from [11] about the exactness of Hodgson's volume form, uses deeper properties of hyperbolic manifolds than had previously been brought to bear on the theory that I've been describing in this chapter: Thurston's Dehn surgery theorem and, crucially, the volume rigidity theorem of Goldman, Gromov and Thurston. Dunfield gave elegant applications of his birationality theorem to Dehn surgery and to related topics. One of his results on Dehn surgery states that if a hyperbolic knot in S^3 is small, i.e. if its exterior contains no closed essential surface, and if it admits a nontrivial cyclic surgery, then it admits a nonintegral boundary slope. (For the terminology I'm using here, it's best to see Boyer's chapter. I'll talk about another of Dunfield's consequences of his birationality theorem in the next section.)

10. Boundary slopes and genera of essential surfaces in knot exteriors: the Neuwirth Conjecture revisited and the Poincaré Conjecture approached

According to Theorem 5.6.2, if a tame knot in any closed orientable 3-manifold Σ satisfies some mild conditions then it has at least two boundary slopes. In the case where $\pi_1(\Sigma)$ is cyclic (for example if $\Sigma = S^3$) one gets strictly stronger results. Some of these are best understood in terms of explicitly identifying slopes as elements of $\mathbf{Q} \cup \{\infty\}$, by the formalism explained in Boyer's chapter in this volume. Recall that if $\mu \in \Lambda = H_1(\partial M)$ denotes the meridian of M, and if we choose an element $\lambda \in \Lambda$ such that $\{\mu, \lambda\}$ is a basis of λ, then we get a bijection between slopes and elements of $\mathbf{Q} \cup \{\infty\}$ by letting the slope of $a\mu + b\lambda$ correspond to a/b. As a/b has at least as much of a right to be called a slope as the corresponding unoriented isotopy class, I will allow myself to blur the distinction between the two when λ has been chosen, or—as as in what we're about to see—when the choice doesn't matter.

Changing the choice of λ has the effect of subjecting all slopes to an integer translation, possibly followed by a change of sign. In particular, the absolute value of the difference between two slopes does not depend on the choice of λ. Indeed, a straightforward calculation shows that the difference between the slopes of two indivisible elements $\alpha, \beta \in \lambda$ is given in invariant form by the expression

$$\frac{\Delta(\alpha, \beta)}{\Delta(\alpha, \mu)\Delta(\beta, \mu),}$$

where Δ denotes the geometric intersection number of α and β as in Theorem 9.1.1.

The slope of the meridian is always ∞. When ∞ is not a boundary slope, the set of boundary slopes is a subset of $\mathbf{Q}$ which is finite by Hatcher's theorem [30] and is well-defined modulo integer translation and change of sign. In particular, the *diameter* of the set of boundary slopes—the difference between its greatest and least elements—is well defined.

10.1. A lower bound for the diameter of the boundary slopes

The following result is proved in [19].

Theorem 10.1.1. *Let K be a nontrivial tame knot in a homotopy 3-sphere. Suppose that ∞ is not a boundary slope of K. Then the diameter of the set of boundary slopes of K is $\geqslant 2$.*

An immediate corollary to this theorem is that if ∞ is not a boundary slope, there is always a boundary slope of absolute value at least 1. This is a small step in the direction of the conjecture that there is always a boundary slope which is a nonzero integer. You should think of this conjecture as a stronger form of the conjecture of Neuwirth's that I sketched the proof of in Chapter 6; in fact it is closely related to some of the stronger versions stated by Neuwirth himself.

Theorem 10.1.1 is deduced in [19], by a relatively routine argument, from the following more general result, Theorem 10.1.2. A *cable knot* is a knot that lies on the boundary of some tame solid torus and has intersection number at least 2 with a meridian disk for that solid torus.

Theorem 10.1.2. *Let Σ be a closed, connected, orientable 3-manifold such that $\pi_1(\Sigma)$ is cyclic. Let K be a nontrivial tame knot in Σ which is not a cable knot. Suppose that ∞ is not a boundary slope of K. Then the diameter of the set of boundary slopes of K is $\geqslant 2$.*

This result, like Theorem 5.6.2, is technically easier to prove in the case where $\Sigma - K$ is hyperbolic. The proof in the hyperbolic case, which I will give here, turns out to be a remarkably simple application of Theorem 9.5.1. The proof in the general case involves souping up the argument that works in the hyperbolic case in rather the same way that the arguments of Subsection 5.6 were souped up in Section 6, although the details—for which I will refer you to [19]—are more involved.

In proving Theorem 10.1.2 in the hyperbolic case, it's nice to think in terms of the invariant description of the difference of two slopes that I gave above. From this point of view, what we have to prove is that there exist two boundary classes α and β such that

$$\frac{\Delta(\alpha,\beta)}{\Delta(\alpha,\mu)\Delta(\beta,\mu)} \geqslant 2. \tag{10.1.3}$$

As in Section 9, let's set $m = \min_{0\neq\gamma\in\Lambda} \|\gamma\|_M$. If μ denotes the meridian of K (in the sense of Boyer's chapter, for example), then the "Dehn-filled" manifold $M(\mu)$ is just Σ, which has cyclic fundamental group, so that $\|\mu\| = m$ by Theorem 9.5.1. Thus if B_m denotes the ball of radius m for the norm $\|\cdot\| = \|\cdot\|_M$, we have $\mu \in \partial B_m$.

Now recall from Subsection 8.3 that the unit ball B of $\|\cdot\|$ is a compact, convex, balanced set bounded by a polygon, and that each vertex of the polygon lies on a boundary line; hence the same is true of $B_m = mB$. Let e denote an edge of the polygon ∂B_m containing μ. The endpoints of e are vertices α_0 and β_0 of ∂B_m, and are therefore positive multiples of boundary classes α and β. I'll complete the proof by showing that the inequality (10.1.3) holds with these choices of α and β.

The geometric intersection number $\Delta(\cdot,\cdot)$ is the absolute value of an alternating integer-valued bilinear pairing on the lattice $\Lambda = H_1(\partial M;\mathbf{Z})$. Let's extend the latter pairing to an alternating real-valued bilinear pairing on the vector space $V = H_1(\partial M;\mathbf{R})$; I'll write $\Delta(\cdot,\cdot)$ for the absolute value of this extended pairing as well. If we write $\alpha = a\|\alpha_0\|$ and $\beta = b\|\beta_0\|$, with $a, b > 0$, then we have $\Delta(\alpha,\beta) = ab\Delta(\alpha_0,\beta_0)$, whereas $\Delta(\alpha,\mu) = a\Delta(\alpha_0,\mu)$ and $\Delta(\beta,\mu) = b\Delta(\beta_0,\mu)$. Hence

$$\frac{\Delta(\alpha,\beta)}{\Delta(\alpha,\mu)\Delta(\beta,\mu)} = \frac{\Delta(\alpha_0,\beta_0)}{\Delta(\alpha_0,\mu)\Delta(\beta_0,\mu)}. \tag{10.1.4}$$

Since μ is on the segment e with endpoints α_0, β_0, we can write $\mu = t\alpha_0 + (1-t)\beta_0$ for some $t \in [0,1]$. So we have

$$\Delta(\alpha_0,\mu) = (1-t)\Delta(\alpha_0,\beta_0) \qquad \text{and} \qquad \Delta(\beta_0,\mu) = t\Delta(\alpha_0,\beta_0).$$

Combining this with the equality (10.1.4), we get

$$\frac{\Delta(\alpha,\beta)}{\Delta(\alpha,\mu)\Delta(\beta,\mu)} = \frac{1}{t(1-t)\Delta(\alpha_0,\beta_0)}. \tag{10.1.5}$$

As I pointed out in Section 9, it follows from Theorem 9.3.1 that B_m has area at most 4. Now we reason as in the proof of Theorem 9.1.1, but with α_0 and β_0 playing the roles of α and β in that argument. The parallelogram P with vertices $\pm\alpha, \pm\beta$ is therefore contained in B_m, and we have

$$\Delta(\alpha_0,\beta_0) = \frac{1}{2}\operatorname{area} P \leqslant \frac{1}{2}\operatorname{area} B_m \leqslant 2.$$

Combining this inequality with the equality (10.1.5) we find that

$$\frac{\Delta(\alpha,\beta)}{\Delta(\alpha,\mu)\Delta(\beta,\mu)} \geqslant \frac{1}{2t(1-t)}.$$

But the right-hand side of this last inequality is bounded below by 2, since the function $t(1-t)$ on $[0,1]$ takes its maximum at $t = 1/2$. So the inequality (10.1.3) is established, and Theorem 10.1.1 is established in the hyperbolic case.

10.2. Some related results

Nathan Dunfield has shown that Theorem 10.1.1 is sharp: there exists a hyperbolic knot K in a closed orientable 3-manifold Σ with $\pi_1(\Sigma) \cong \mathbf{Z}/10\mathbf{Z}$ such that the set of boundary slopes of K has diameter 2. Furthermore, the greatest and least slopes are half-integers. (Remember that the set of slopes is defined only modulo an integer translation and a change of sign, so that the greatest and least slopes are not well-defined. On the other hand, both the diameter of the set of boundary slopes, and the condition that the greatest and least boundary slopes are integers, is invariant.) This example is presented in [19].

In this example, $\pi_1(\Sigma)$ is cyclic of even order. By contrast, Dunfield has given an argument in [22] which shows that if K is a hyperbolic knot in a closed 3-manifold Σ such that $\pi_1(\Sigma)$ is cyclic of odd order, and if the diameter of the set of all boundary slopes is exactly 2, then the greatest and least slopes cannot be integers or half-integers. The proof involves the same ingredients as Dunfield's theorem about Dehn surgery which I mentioned in Subsection 9.6.

Although the restrictions on the set of boundary slopes that I have stated are the only ones known for a general knot in a manifold with cyclic π_1, one can get a little more information about the set of essential surfaces in the knot exterior by looking beyond the set of slopes. For example, the main theorem of my forthcoming paper [20] with Culler gives, as a special case, information about any nontrivial knot K in a homotopy 3-sphere (e.g. S^3). Suppose that the exterior of K contains only two incompressible surfaces (so that both are bounded, one is a spanning surface and one has boundary slope $\neq 0$). If the genus g of K is $\geqslant 2$ and if s denotes the non-zero boundary slope, then

$$\frac{g}{\log_2 g} \leqslant 24s^2.$$

It's interesting to compare this with the known examples of knots in S^3 with only two essential surfaces in their exteriors, which are the torus knots: for a type (p, q) torus knot, the genus is $(p-1)(q-1)/2$ and the nonzero boundary slope is pq. So for torus knots the genus is a little less than the slope, whereas the general theorem gives an upper bound for the genus which is slightly worse-than-quadratic in the slope. (The general form of the theorem of [20], which involves the notion of a "strict boundary slope," applies to certain hyperbolic knots as well, such as the figure-eight knot.) The method of proof involves considering the functions $I_\gamma - 2$ associated to *nonperipheral* elements γ, and comparing the orders of their poles with the orders of their zeros by using some of the facts pointed out in Section 5.

In addition to the connection with classical knot theory, these results are potentially related to the Poincaré Conjecture. One can think of them as characterizations of the trivial knot in a closed orientable 3-manifold Σ with cyclic π_1: for example, Theorem 10.1.1 says that a knot in Σ is trivial if and only if the diameter of the set of its boundary slopes is strictly less than 2. If one can characterize the trivial knot, one can try to show that an *arbitrary* closed orientable 3-manifold contains some knot satisfying the condition and having an irreducible exterior. It would then follow that any closed orientable 3-manifold with cyclic π_1 would contain a trivial knot with irreducible exterior, and would therefore be a lens space. (This would give the Poincaré Conjecture as a special case.)

11. R-trees, degenerations of hyperbolic structures, and other stories

In [64], Thurston gave a criterion for the set $\mathrm{AH}(M)$ of "homotopy hyperbolic structures" on a compact irreducible 3-manifold M (with boundary) to be compact. The simplest way to define $\mathrm{AH}(M)$ involves using the $\mathrm{PSL}(2, \mathbf{C})$-character variety $PX(\pi_1(M))$, which I haven't defined in this chapter: it's the subset of $PX(\pi_1(M))$ consisting of all characters of faithful representations whose images are discrete subgroups of $\mathrm{PSL}(2, \mathbf{C})$. The question of compactness is unaffected by replacing $\mathrm{PSL}(2, \mathbf{C})$ by $\mathrm{SL}(2, \mathbf{C})$: saying that $\mathrm{AH}(M)$ is compact is equivalent to saying that the set $D(\pi_1(M)) \subset X(\pi_1(M))$ consisting of all characters of discrete faithful representations in $\mathrm{SL}(2, \mathbf{C})$ is compact. Thurston's theorem, which played a key role in his original proof of his geometrization theorem for Haken manifolds, asserts that this set is compact if and only if M contains no essential disks or annuli. The "only if" part is easy, and in proving the converse it's easy to see that one can assume that M contains no connected essential surfaces with nonnegative Euler characteristic.

When Marc Culler and I were doing the work that led to [17], we noticed a connection between our methods and the statement of Thurston's result. In fact, we noticed a simple proof of the weaker statement that when M contains no essential surfaces with nonnegative Euler characteristic, the intersection of $D(\pi_1(M))$ with any curve $C \subset X(\pi_1(M))$ is compact. If the conclusion were false there would be a sequence (χ_i) of points of $C \cap D(\pi_1(M))$ approaching an ideal point x of C. The discreteness and faithfulness of the representations ρ_i defining the χ_i implies that under the action on a tree T associated to x by the construction of Section 5, the stabilizer Γ_e of any edge e of T is a "small" subgroup of $\pi_1(M)$ in the sense that it contains no nonabelian free subgroup. Briefly, this is because, if γ and δ generated a free subgroup of $[\Gamma_e, \Gamma_e]$, the functions I_γ and $I_{[\gamma,\delta]}$ would take the value 2 at x by Property 5.5.3. So $\operatorname{trace} \rho(\gamma)$, and $\operatorname{trace} \rho([\gamma, \delta])$ would be close to 2 for large i, and this would contradict Jørgensen's inequality [36] about discrete subgroups of $\mathrm{SL}(2, \mathbf{C})$.

Now let's associate an essential surface F with the action of $\pi_1(M)$ on T. Since the edge stabilizers are small subgrooups of $\pi_1(M)$, it follows from 2.3.1(ii) that each component of F has a small fundamental group, and hence has Euler characteristic $\geqslant 0$. This contradicts the hypothesis.

John Morgan and I were able to turn this into a proof of Thurston's compactness theorem. The task was to replace the curve C by a whole irreducible component of $X(\pi_1(M))$, and this required generalizing all the material in Sections 2, 3, and 5. In this theory, the discrete, rank-1 valuations that appear in Section 3 are replaced by more general valuations in which the "value" group $\mathbf{Z}$ is replaced by a general ordered abelian group; the simplicial trees that appear in Sections 2 3, are replaced by $\mathbf{R}$-trees, which can be thought of as metric spaces in which any two points are joined by a unique topological arc; and the essential surfaces that appear in Section 2 are replaced by essential measured laminations, which can be thought of as "irrational" counterparts of essential surfaces, and which typically look locally like the product of an open set in $\mathbf{R}^2$ with a Cantor set.

Formally the two main steps in this argument were to show that a suitable kind of sequence tending to infinity in $D(\pi_1(M))$ defines an action of $\pi_1(M)$ on an **R**-tree with small arc stabilizers, and that if a 3-manifold group admits such an action then it contains an essential connected surface of nonnegative Euler characteristic. In [40], Morgan generalized the first step to hyperbolic manifolds of arbitrary dimension. Morgan and I then formulated a general conjecture about groups that act on **R**-trees with small arc stabilizers; in view of the result of [40], this conjecture implied a high-dimensional analogue of Thurston's compactness theorem. As a by-product of our proofs, we also showed that if a finitely generated 3-manifold acts *freely* on an **R**-tree then it's a free product of free groups and surface groups. We conjectured that *any* group which acts freely on an **R**-tree is of this form.

This conjecture on free actions was proved by Rips. Using Rips's ideas, Bestvina-Feighn [4] and Rips-Sela independently proved a version of our conjecture on actions with small arc stabilizers which is strong enough to imply the high-dimensional version of Thurston's compactness theorem. In another direction, Paulin [51], [52] discovered a partial generalization of the results of [42] and [40] which permit applications to much more general kinds of objects than hyperbolic manifolds. This has given rise to an entire new area of geometric group theory in which methods involving actions on **R**-trees are applied to the study of outer automorphisms, decompositions of groups, and other questions. You can learn more about this—including aspects that I haven't even mentioned, such as the connection with Thurston's compactification of Teichmüller space—from my old survey articles [57] and [58], or from Bestvina's chapter in this volume.

References

[1] R. C. Alperin and H. Bass, "Length functions of group actions on Λ-trees." *Combinatorial Group Theory and Topology (Alta, Utah 1984)*,Ann. of Math. Studies, no. 111, Princeton Univ. Press, 1987, pp. 265–378.

[2] R. C. Alperin and P. B. Shalen, "Linear groups of finite cohomological dimension." *Invent. Math.* **66** (1982), 89–98.

[3] H. Bass, "Finitely generated subgroups of GL_2." *The Smith Conjecture. Papers presented at the Symposium held at Columbia University, New York, 1979.* Pure and Applied Mathematics, **112**. Academic Press, 1984, pp. 127–136.

[4] M. Bestvina and M. Feighn, "Stable actions of groups on real trees." *Invent. Math.* **121** (1995), 287–321.

[5] S. Boyer and X. Zhang, "Finite Dehn surgery on knots."*J. Amer. Math. Soc.* **9** (1996), 1005–1050.

[6] S. Boyer and X. Zhang, "On Culler-Shalen seminorms and Dehn filling." Preprint.

[7] S. Boyer and X. Zhang, "A proof of the finite filling conjecture." Preprint.

[8] E. M. Brown and R. H. Crowell, "Deformation retractions of 3-manifolds into their boundaries." *Ann. of Math.* **82** (1965), 445–458.

[9] M. Brown, "Locally flat imbeddings of topological manifolds." *Ann. of Math. (2)* **75** (1962), 331-341.

[10] F. Bruhat and J. Tits, "Groupes réductifs sur un corps local." *Inst. Hautes Etudes Sci. Publ. Math.* No. 41 (1972), 5–251.

[11] D. Cooper, M. Culler, H. Gillet, D. Long, and P. B. Shalen, "Plane curves associated to character varieties of 3-manifolds." *Invent. Math.* **118** (1994), 47–84.

[12] D. Cooper and D. Long, "An undetected slope in a knot manifold." *Topology '90 (Columbus, Ohio, 1990)*, Ohio State Res. Inst. Math. Publ. 1, de Gruyter, 1992, pp. 111–121.

[13] D. Cooper and D. Long, "The A-polynomial has ones in the corners." *Bull. London Math. Soc.* **29** (1997), 231–238.

[14] D. Cooper and D. Long, "Representation theory and the A-polynomial of a knot." Knot theory and its applications. *Chaos Solitons Fractals* **9** (1998), 749–763.

[15] M. Culler, C. Gordon, J. Luecke, and P. B. Shalen, "Dehn surgery on knots." *Ann. of Math.* **125** (1987), 237–300.

[16] M. Culler and J. W. Morgan, "Group actions on $\mathbf{R}$-trees." *Proc. London Math. Soc. (3)* **55** (1987), 571-604.

[17] M. Culler and P. B. Shalen, "Varieties of group representations and splittings of 3-manifolds." *Ann. of Math.* **117** (1983), 109–146.

[18] M. Culler and P. B. Shalen, "Bounded separating, incompressible surfaces in knot manifolds." *Invent. Math.* **75** (1984), 537–545.

[19] M. Culler and P. B. Shalen, "Boundary slopes of knots." *Comm. Math. Helv.*, to appear.

[20] M. Culler and P. B. Shalen, "On knots with only two essential surfaces." In preparation.

[21] M. Davis and J. W. Morgan, "Finite group actions on homotopy 3-spheres." *The Smith Conjecture. Papers presented at the Symposium held at Columbia University, New York, 1979.* Pure and Applied Mathematics, **112**. Academic Press, 1984, pp. 181–225

[22] N. Dunfield, "Examples of non-trivial roots of unity at ideal points of hyperbolic 3-manifolds." *Topology*, to appear.

[23] N. Dunfield, "Cyclic surgery, degrees of maps of character curves, and volume rigidity of hyperbolic manifolds." preprint.

[24] M. J. Dunwoody, "An equivariant sphere theorem." *Bull. London Math. Soc.* **17** (1985), 437–448.

[25] A. L. Edmonds, "A topological proof of the equivariant Dehn lemma." *Trans. Amer. Math. Soc.* **297** (1986), 605–615.

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

[27] C. D. Feustel, "A splitting theorem for closed orientable 3-manifolds." *Topology* **11** (1972), 151–158.

[28] W. Fulton, *Algebraic curves. An introduction to algebraic geometry.* Advanced Book Classics. Addison-Wesley, 1989. xxii + 226 pp. Series

[29] F. González-Acunã and J. M. Montesinos-Amilibia, "On the character variety of group representations in $\mathrm{SL}(2, \mathbf{C})$ and $\mathrm{PSL}(2, \mathbf{C})$." *Math. Z.* **214** (1993), 627–652.

[30] A. E. Hatcher, "On the boundary curves of incompressible surfaces." *Pacific J. Math.* **99** (1982), 373–377.

[31] J. Hempel, 3-*Manifolds.* Ann. of Math. Studies, no. 86, Princeton Univ. Press, 1976, 12 + 195 pp.

[32] J. F. P. Hudson, *Piecewise linear topology. University of Chicago Lecture Notes prepared with the assistance of J. L. Shaneson and J. Lees.* W. A. Benjamin, Inc., New York-Amsterdam 1969, ix + 282 pp.

[33] W. Jaco and J. H. Rubinstein, "PL equivariant surgery and invariant decompositions of 3-manifolds." *Adv. in Math.* **73** (1989), 149–191.

[34] W. Jaco and P. B. Shalen, "Seifert fibered spaces in 3-manifolds." *Mem. Amer. Math. Soc.* **21** (1979), no. 220, viii + 192 pp.

[35] K. Johannson, *Homotopy Equivalences of* 3-*Manifolds with Boundaries.* Lecture Notes in Mathematics, **761**. Springer, 1979, ii + 303 pp.

[36] T. Jorgensen, "On discrete groups of Möbius tranformations." *Amer. J. Math.* **98** (1976), 739–749.

[37] A. G. Kurosh, *The Theory of Groups.* Second English Edition, 2 volumes, Chelsea, 1960. Volume I, 272 pp. Volume II, 308 pp.

[38] W. H. Meeks III and S. T. Yau, "The equivariant loop theorem for three-dimensional manifolds and a review of the existence theorems for minimal surfaces." *The Smith Conjecture.*

Papers presented at the Symposium held at Columbia University, New York, 1979. Pure and Applied Mathematics, **112**. Academic Press, 1984, pp. 153–163.

[39] E. E. Moise, *Geometric Topology in Dimensions* 2 *and* 3. Graduate Texts in Mathematics, **47**. Springer, 1977, x + 262 pp.

[40] J. W. Morgan, "Groups actions on trees and the compactification of the space of classes of $\mathrm{SO}(n,1)$-representations." *Topology* **25** (1986), 1–33.

[41] J. Morgan and H. Bass, Eds., *The Smith Conjecture. Papers presented at the Symposium held at Columbia University, New York, 1979.* Pure and Applied Mathematics, **112**. Academic Press, 1984, xv + 243 pp.

[42] J. Morgan and P. B. Shalen, "Degenerations of hyperbolic structures, I." *Ann. of Math.* **120** (1984), 401–476.

[43] J. Morgan and P. B. Shalen, "Degenerations of hyperbolic structures, II." *Ann. of Math.* **127** (1988), 403–456.

[44] J. Morgan and P. B. Shalen, "Degenerations of hyperbolic structures, III." *Ann. of Math.* **127** (1988), 457–519.

[45] *Algebraic geometry. I. Complex projective varieties.* Grundlehren der Mathematischen Wissenschaften, no. 221. Springer, 1976, x + 186 pp.

[46] L. P. Neuwirth, "Interpolating manifolds for knots in S^3." *Topology* **2** (1963), 359–365.

[47] L. P. Neuwirth, *Knot Groups.* Ann. of Math. Studies, no. 56, Princeton Univ. Press, 1965, vi + 113 pp.

[48] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An Introduction to the Theory of Numbers.* Fifth Edition. John Wiley and Sons, 1991, xiv + 529 pp.

[49] J.-P. Otal, *Le Théorème d'Hyperbolisation pour les Variétés Fibrées de Dimension* 3, Astérisque **235**, Société Mathématique de France (1996).

[50] J.-P. Otal, *Thurston hyperbolization of Haken manifolds*, in *Surveys in Differential Geometry* (S.-T. Yau ed.), to appear (1997).

[51] F. Paulin, "Topologie de Gromov équivariante, structures hyperboliques et arbres réels. *Invent. Math.* **94** (1988), 53–80.

[52] F. Paulin, "Outer automorphisms of hyperbolic groups and small actions on **R**-trees." *Arboreal group theory (Berkeley, CA, 1988).* MSRI Publ. 19, Springer New York, 1991, pp. 331–343.

[53] P. Scott and C. T. C. Wall, "Topological methods in group theory." *Homological Group Theory (Proc. Sympos. Durham 1977)*, London Math. Soc. Lecture Note Series **37**. Cambridge Univ. Press, 1979, pp. 137–203.

[54] J.-P. Serre, *A Course in Arithmetic.* Translated from the French. Graduate Texts in Mathematics, No. 7. Springer-Verlag, 1973, viii+115 pp.

[55] J.-P. Serre, *Trees*. Translated from the French by John Stillwell. Springer-Verlag, 1980. ix+142 pp.

[56] I. R. Shafarevich, *Basic Algebraic Geometry*. Translated from the Russian by K. A. Hirsch. Springer-Verlag, 1977. xv+439 pp.

[57] P. B. Shalen, "Dendrology of groups." *Essays in Group Theory*, S. Gersten, Ed. M.S.R.I. Pub. **8**, Springer-Verlag 1987, pp. 265-320.

[58] P. Shalen, "Dendrology and its applications." *Group Theory from a Geometrical Viewpoint, ICTP, Trieste, Italy, 26 March–6 April, 1990*, E. Ghys, A. Haefliger, and A. Verjovsky, Eds. World Scientific Publishing, 1991, pp. 543–616.

[59] E. H. Spanier, *Algebraic Topology*. Springer, 19??, xvi + 528 pp.

[60] J. R. Stallings, "On fibering certain 3-manifolds." *Topology of* 3*-Manifolds and Related Topics (Proc. Univ. of Georgia Institute 1961).* Prentice-Hall, 1962, pp. 95–100.

[61] J. R. Stallings, "A topological proof of Grushko's theorem on free products." *Math. Z.* **90** (1965), 1–8.

[62] J. Stallings, *Group theory and three-dimensional manifolds. A James K. Whittemore Lecture in Mathematics given at Yale University, 1969.* Yale Mathematical Monographs **4**, Yale University Press, 1971, v+65 pp.

[63] W. P. Thurston, "A norm for the homology of 3-manifolds." *Mem. Amer. Math. Soc.* **59**

(1986), no. 339, 1–viii and 99–130.

[64] W. P. Thurston, "Hyperbolic structures on 3-manifolds. I. Deformation of acylindrical manifolds." *Ann. of Math. (2)* **124** (1986), 203–246.

[65] A. Weil, "On discrete subgroups of Lie groups. *Ann. of Math. (2)* **72** (1960), 369–384.

[66] A. Weil, "On discrete subgroups of Lie groups. II. *Ann. of Math. (2)* **75** (1962), 578–602.

[67] F. Waldhausen, "Gruppen mit Zentrum und 3-dimensionale Mannigfältigkeiten." *Topology* **6** (1967), 505–517.