CHAPTER 4

Free and surface groups

In this chapter we study scl in free groups, and some related groups. The methods are largely geometric and depend on realizing the groups in question as fundamental groups of particularly simple low-dimensional manifolds.

The first main theorem proved in this chapter is the Rationality Theorem (Theorem 4.24), which says that in a free group F, the unit ball of the scl norm on $B_1^H(F)$ is a rational polyhedron; i.e. scl is a piecewise linear rational function on finite dimensional rational subspaces of $B_1^H(F)$. It follows that scl takes on only rational values in free groups. The method of proof is direct: we show how to explicitly construct extremal surfaces bounding finite linear combinations of conjugacy classes. As a byproduct, we obtain a polynomial-time algorithm to calculate scl in free groups, which can be practically implemented, at least in some simple cases. This algorithm gives an interesting conjectural picture of the spectrum of scl on free groups, and perhaps some insight into the spectrum of scl on word-hyperbolic groups in general.

The polyhedrality of the unit ball of the scl norm is related to certain rigidity phenomena. Each nonzero element in $B_1^H(F)$ projectively intersects the boundary of the unit ball of the scl norm in the interior of some face. The smaller the codimension of this face, the smaller the space of quasimorphisms which are extremal for the given element. The situations displaying the most rigidity are therefore associated to faces of the unit ball of codimension one. It turns out that for a free group, such faces of codimension one exist, and have a geometric meaning. In § 4.2 we discuss the Rigidity Theorem (Theorem 4.78), which says that if F is a free group, associated to each isomorphism $F \to \pi_1(S)$ (up to conjugacy), where S is a compact oriented surface, there is a top dimensional face π_S of the unit ball of the scl norm on F, and the unique homogeneous quasimorphism ϕ_S dual to π_S (up to scale and elements of H^1) is the rotation quasimorphism associated to a hyperbolic structure on S.

Finally, in § 4.3, we discuss diagrammatic methods to study scl in free groups. In particular, we discuss a technique due to Duncan–Howie which uses left-invariant orders on one-relator groups to obtain sharp lower bounds on scl in free groups.

Some of the material in this chapter is developed more fully in the papers [**47, 43, 45, 46**].

4.1. The Rationality Theorem

The goal of this section is to prove the Rationality Theorem for free groups. Essentially, this theorem says that the unit ball in the scl norm on $B_1^H(F)$ is a rational polyhedron. Polyhedral norms occur in other contexts in low-dimensional

topology, and the best-known example is that of the Thurston norm on the 2-dimensional homology of a 3-manifold. We briefly discuss this example.

4.1.1. Thurston norm. Let M be a 3-manifold. Thurston [**196**] defined a pseudo-norm on $H_2(M, \partial M; \mathbb{R})$ as follows.

For each properly embedded surface S in M, define $\|S\|_T = -\chi^-(S)$. For each relative class $A \in H_2(M, \partial M; \mathbb{Z})$, define

$$\|A\|_T = \inf_S -\chi^-(S)$$

where the infimum is taken over all properly embedded surfaces S for which $[S]$ represents the class A. Thurston shows that this function satisfies the following two crucial properties:

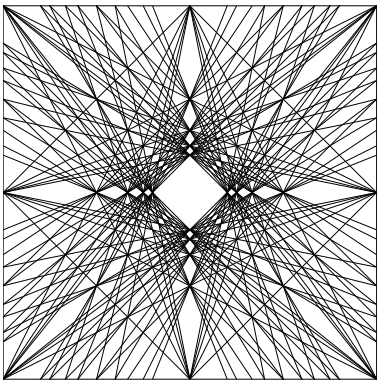
- it is linear on rays; that is, $\|nA\|_T = n\|A\|_T$ for any integral class A and any non-negative integer n
- it is subadditive; that is, $\|A+B\|_T \leq \|A\|_T + \|B\|_T$ for all integral classes A, B.

By the first property, $\|\cdot\|_T$ can be extended by linearity to all of $H_2(M, \partial M; \mathbb{Q})$. By the second property, it can be extended to a unique continuous function on $H_2(M, \partial M; \mathbb{R})$, which is linear on rays and subadditive. Such a function satisfies the axioms of a (pseudo)-norm, and is called the *Thurston norm* on homology. Note that this function is generally only a pseudo-norm; it takes the value 0 on the span of integral classes which can be represented by surfaces of non-negative Euler characteristic. If M is irreducible and atoroidal, $\|\cdot\|_T$ is a genuine norm.

By construction, $\|A\|_T \in \mathbb{Z}$ for all $A \in H_2(M, \partial M; \mathbb{Z})$. A norm on a finite dimensional vector space which takes integer values on integer vectors (with respect to some basis) can be characterized in a finite amount of data, as follows.

LEMMA 4.1. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ which takes integer values on the lattice $\mathbb{Z}^n$. Then the unit ball of $\|\cdot\|$ is a finite sided polyhedron whose faces are defined by integral linear equalities.*

PROOF. Let U be any open set in $\mathbb{R}^n$ containing 0. We claim that there are only *finitely many* integral linear functions ϕ on $\mathbb{R}^n$ such that the subspace $\phi \leq 1$ contains U. Let ϕ be such a linear function. Then there is a (unique) *integral* vector v_ϕ such that $\phi(w) = \langle v_\phi, w\rangle$ where $\langle\cdot,\cdot\rangle$ denotes the ordinary inner product on $\mathbb{R}^n$. Since U is open, there is some positive number ϵ such that the ball of radius ϵ in the (ordinary) L^1 norm is contained inside U. Hence if ϕ is as above, every co-ordinate of v_ϕ has absolute value at most $1/\epsilon$. On the other hand, since v_ϕ is integral, there are only *finitely many* functions ϕ with this property (the adjacent Figure shows all level sets $\phi = 1$ in 2 dimensions for $\epsilon = 1/5$). This proves the claim.



Let B denote the unit ball in the $\|\cdot\|$ norm. For the remainder of the proof we assume $n = 3$ (the general case is not significantly more complicated). For each integral basis $\{v_1, v_2, v_3\}$, there is a unique integral linear function on $\mathbb{R}^3$ that agrees with $\|\cdot\|$ on the elements of the basis. Pick some primitive integral vector v_1,

and then extend v_1 to an integral basis $\{v_1, v_2, v_3\}$. For each pair of integers i, j let $v_2^i = v_2 + iv_1$ and $v_3^{i,j} = v_3 + iv_1 + jv_2$, and let $\phi_{i,j}$ denote the integral linear function that agrees with $\|\cdot\|$ on the basis $\{v_1, v_2^i, v_3^{i,j}\}$. Fix a small open set U containing 0 as above, whose closure is contained in the interior of B. For each sufficiently large fixed j, the functions $\phi_{i,j}$ for i big compared to j satisfy $\phi_{i,j} \leq 1$ on U. By convexity of B and the discussion above, for each fixed j there is a ψ_j such that $\psi_j = \phi_{i,j}$ for all sufficiently large i (depending on j). The plane $\psi_j = 1$ intersects ∂B in two straight lines joining $v_1/\|v_1\|$ to each of $v_2^i/\|v_2^i\|$ and $v_3^{i,j}/\|v_3^{i,j}\|$. Since $\psi_j \leq 1$ on U for each j, there are distinct j, j' for which $\psi_j = \psi_{j'}$. Consequently the plane $\psi_j = 1$ intersects ∂B in three straight lines meeting at acute angles, and therefore (by convexity of B) intersects ∂B in a subset with nonempty interior whose closure contains $v_1/\|v_1\|$.

Since v_1 was arbitrary, we conclude that B is the intersection of the half spaces $\phi \leq 1$ where ϕ is integral and linear and satisfies $\phi \leq 1$ on B. Since there are only finitely many such ϕ, the lemma follows. □

REMARK 4.2. The proof of Lemma 4.1 is Thurston's proof of the polyhedrality of his norm. Oertel's proof [**162**], using branched surfaces, is closer in spirit to the methods in this chapter, but requires more prerequisites from 3-manifold topology.

There is a similar definition of a norm on $H_2(M)$, defined by restricting attention to closed embedded surfaces representing absolute homology classes. Note that the value of $\|\cdot\|_T$ on any *absolute* class in $H_2(M;\mathbb{Z})$ is an *even* integer.

The crucial property of the Thurston norm, for our purposes, is its relation to the (Gromov) L^1 norm $\|\cdot\|_1$ on $H_2(M, \partial M; \mathbb{R})$. Thurston already showed that a compact leaf of a taut foliation is minimizing in its homology class in both the Thurston and the Gromov norms, and therefore the two norms are proportional on the projective homology classes realized by such surfaces. Conversely, Gabai [**84**] showed that every Thurston norm minimizing surface is a compact leaf of a taut foliation. From this he deduced the following proportionality theorem, conjectured by Thurston:

THEOREM 4.3 (Gabai, Corollary 6.18. [**84**]). *Let M be a compact oriented 3-manifold. Then on $H_2(M)$ or $H_2(M, \partial M)$,*

$$\|\cdot\|_T = \frac{1}{2}\|\cdot\|_1$$

From this we can deduce the following fact:

PROPOSITION 4.4. *Let M be a compact oriented 3-manifold. Let $\gamma \subset \partial M$ be an embedded, oriented loop. Let a be the conjugacy class in $\pi_1(M)$ represented by γ. Suppose $a \in [\pi_1(M), \pi_1(M)]$. Then* $\mathrm{scl}(a) \in \mathbb{Q}$. *Furthermore, if $H_2(M;\mathbb{R}) = 0$ then* $\mathrm{scl}(a) \in \frac{1}{2} + \mathbb{Z}$.

PROOF. Let A be a regular annulus neighborhood of γ, and let N be obtained by doubling M along A. We write $N = M \cup \overline{M}$ where $M \cap \overline{M} = A$. By Mayer–Vietoris there is an exact sequence

$$0 \to H_2(M) \oplus H_2(\overline{M}) \to H_2(N) \xrightarrow{\partial} H_1(A) \to 0$$

where exactness at the last term follows because the inclusion map of $H_1(A)$ into both $H_1(M)$ and $H_1(\overline{M})$ is zero, because $a \in [\pi_1(M), \pi_1(M)]$. Let $V \subset H_2(N)$ be the integral affine subspace $V = \partial^{-1}([\gamma])$ where $[\gamma] \in H_1(A)$ is the generator. If C

is a 2-chain in M with the support of ∂C mapping into γ, and $[\partial C] = [\gamma]$ in $H_1(A)$, then $C - \overline{C}$ is a 2-cycle in N representing an element of V. It follows that there is an inequality

$$2\,\mathrm{fill}(a) \geq \inf_{v \in V} \|v\|_1$$

Conversely, let S be a Thurston norm minimizing surface in N representing an integral class which is projectively close to an element of V. By making S transverse to A, and isotoping it so that no component of $S \cap M$ or $S \cap \overline{M}$ is a disk, one obtains an inequality

$$\mathrm{scl}(a) \leq \frac{1}{4} \inf_{v \in V} \|v\|_T$$

Using $\mathrm{scl}(a) = \frac{1}{4}\mathrm{fill}(a)$ one therefore obtains an equality

$$\mathrm{scl}(a) = \frac{1}{4} \inf_{v \in V} \|v\|_T$$

Since the Thurston norm takes even integral values on integer lattice points, and since V is an integral affine subspace, the infimum is rational.

In the special case that $H_2(M; \mathbb{R}) = 0$, the subspace V is 0 dimensional, and consists of a single integral class v. If S is a norm minimizing surface representing v, make S transverse to A and efficient. If S_1 and S_2 are the intersections $S_1 \cap M_1$ and $S_2 \cap M_2$ then $\chi(S_1) = \chi^-(S_1) = \chi^-(S_2) = \chi(S_2)$ or else by replacing S_1 by $\overline{S}_2$ (for example) one could reduce the norm. Since each S_i is embedded, the intersection $S_1 \cap A$ consists of a union of embedded loops. Norm minimizing surfaces are incompressible, so each oriented boundary component of S_1 is isotopic in A to γ or γ^{-1}. Moreover by the definition of ∂, there is an equality $[\partial S_1] = [\gamma]$ in $H_1(A)$. It follows that S_1 has an *odd* number of boundary components, and therefore $\|v\|_T = 4n + 2$ for some integer n. Consequently in this case we have an equality

$$\mathrm{scl}(a) = \frac{1}{4}\|v\|_T \in \frac{1}{2} + \mathbb{Z}$$

□

EXAMPLE 4.5. A word w in a free group F is *geometric* if there is a handlebody H with $\pi_1(H) = F$ such that a loop γ in H in the conjugacy class of w is homotopic to an embedded loop in ∂H. For such a w, one has $\mathrm{scl}(w) \in \frac{1}{2} + \mathbb{Z}$ (if $w \in [F, F]$).

A word w in F is *virtually geometric* if there is a finite cover $H' \to H$ such that the total preimage of γ in H' is homotopic to a union of embedded loops in $\partial H'$. If w is virtually geometric, then $\mathrm{scl}(w) \in \mathbb{Q}$.

EXAMPLE 4.6 (Gordon–Wilton [**94**]). In $F_2 = \langle a, b \rangle$, the Baumslag–Solitar words $w = b^{-1}a^p b a^q$ are virtually geometric (but not geometric).

EXAMPLE 4.7 (Manning [**144**]). Jason Manning gives a criterion to show that certain words in free groups are not virtually geometric. For example, in $F_3 = \langle a, b, c \rangle$, many words, including $b^2a^2c^2abc$ and $ba^2bc^2a^{-1}c^{-1}b^{-2}c^{-1}a^{-1}$, are not virtually geometric. Similar examples exist in nonabelian free groups of any rank.

A corollary of Theorem 4.3 is that the unit ball of the dual Thurston norm is the convex hull of the set of cohomology classes which are in the image of elements of H^2_b whose (L^∞) norm is equal to $1/2$. It is natural to try to find explicit bounded 2-cocycles whose cohomology classes correspond to the vertices of the dual norm,

and which therefore can be used to certify that a given surface is Thurston norm minimizing. It is a highly nontrivial fact that for every irreducible, atoroidal 3-manifold, one may find a finite collection of classes $[e] \in H^2$ in the image of H^2_b, whose convex hull is equal to the unit ball of the dual norm, and such that every $[e]$ is obtained by pulling back the Euler class (i.e. the generator of H^2) from the group $\mathrm{Homeo}^+(S^1)$ under some faithful homomorphism $\pi_1(M) \to \mathrm{Homeo}^+(S^1)$.

In fact, this characterization of the Thurston norm is unfamiliar even to many people working in 3-manifold topology, and deserves some explanation. A homomorphism $\pi_1(M) \to \mathrm{Homeo}^+(S^1)$ is the same thing as an action of $\pi_1(M)$ on a circle. Gabai's main theorem from [**84**] (Theorem 5.5) says that every embedded surface S realizing $\|\cdot\|_T$ in its homology class is a leaf of a finite depth taut foliation $\mathcal{F}$ on M. To every taut foliation of an atoroidal 3-manifold one can associate a *universal circle* S^1_{univ}, which monotonely parameterizes the circle at infinity of every leaf of $\tilde{\mathcal{F}}$, the pullback of $\mathcal{F}$ to the universal cover $\tilde{M}$. See [**40**] Chapter 7 for a proof, and an extensive discussion of universal circles. The construction of S^1_{univ} is natural, so the action of $\pi_1(M)$ as the deck group of $\tilde{M}$ induces an action on S^1_{univ} by homeomorphisms, and therefore a representation $\rho_{\mathrm{univ}} : \pi_1(M) \to \mathrm{Homeo}^+(S^1_{\mathrm{univ}})$. Associated to this representation there is a foliated circle bundle E over M, which one can show is isomorphic (as a circle bundle) to the unit tangent bundle to the foliation $UT\mathcal{F}$. In particular, the pullback $[e]$ of the Euler class is the obstruction to finding a section of $UT\mathcal{F}$, and $[e](S) = \pm\chi(S)$ by construction. On the other hand, the Milnor–Wood inequality (Theorem 2.52) implies that $\|[e]\|_\infty = 1/2$, so this class is in the boundary of the convex hull of the dual norm.

In light of this fact, it is natural to wonder whether the unit ball of the scl norm on $B^H_1(\pi_1(M))$ for M an irreducible, atoroidal 3-manifold is cut out by hyperplanes determined by rotation quasimorphisms. In fact, it turns out that this is *not* the case. A counterexample is the Weeks manifold W, which can be obtained by $(5/1, 5/2)$ surgery on the components of the Whitehead link in S^3, and is known (see e.g. Milley [**153**]) to be the smallest volume closed orientable hyperbolic 3-manifold. In [**48**] it is shown that every homomorphism $\pi_1(W) \to \mathrm{Homeo}^+(S^1)$ must factor through $\mathbb{Z}/5\mathbb{Z}$, and therefore there are *no* nontrivial rotation quasimorphisms on $\pi_1(W)$. To reconcile this with the assertions in the previous paragraph, note that W is a rational homology sphere, so $H_2(W)$ is trivial.

It should be clear from this example that the relationship between the scl norm and rotation quasimorphisms (at least in 3-manifold groups) cannot be as straightforward as one might naively guess, based on familiarity with the Thurston norm (also, see Example 4.35). Thus, although in the next few sections we give a direct proof that the scl norm on free groups is piecewise rational linear, our argument does not suggest a natural family of extremal quasimorphisms which define the faces of the unit ball (however, see § 4.2).

4.1.2. Branched surfaces. It is convenient to introduce the language of branched surfaces. For a reference, see [**160**] or § 6.3 of [**40**].

Definition 4.8. A *branched surface* B is a finite, smooth 2-complex obtained from a finite collection of smooth surfaces by identifying compact subsurfaces.

The *branch locus* of B, denoted $\mathrm{br}(B)$, is the set of points which are not 2-manifold points. The components of $B - \mathrm{br}(B)$ are called the *sectors* of the branched surface. The set of sectors of B is denoted $S(B)$. A *simple branched surface* is a

branched surface for which the branch locus is a finite union of disjoint smoothly embedded simple loops and simple proper arcs.

In a simple branched surface, local sectors meet along segments of the branch locus. The local sheets approach the branch segment from one of two sides (distinguished by the smooth structure along $\mathrm{br}(B)$). In a generic branched surface, three sheets meet along each component of the branch locus, two on one side and one on the other. However, the branched surfaces we consider in this section are *not* generic, and any (positive) number of sheets may meet a segment of branch locus on either side. See Figure 4.1 for an example.

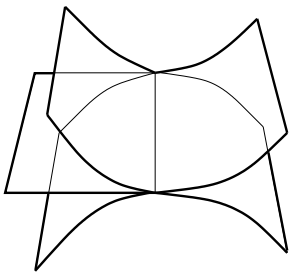


FIGURE 4.1. An example of a local model for a simple branched surface. In this example, five sheets meet along the branch locus, two on one side and three on the other.

Branched surfaces can have boundary or not. The branched surfaces considered in this section have boundary. We require that the branched locus intersect the boundary transversely. Note that the sectors of a simple branched surface B are themselves surfaces, perhaps with boundary, and possibly with *corners* where arcs of the branch locus intersect ∂B. A branched surface B is oriented or not according to whether the sectors can be compatibly oriented. We are exclusively interested in oriented branched surfaces.

DEFINITION 4.9. Let B be a simple branched surface. A *weight* on B is a function $w : S(B) \to \mathbb{R}$ such that for each component γ of $\mathrm{br}(B)$, the sum of the values of w on the sectors which meet γ on one side is equal to the sum of the values of w on the sectors which meet γ on the other side. A weight is *rational* if it takes values in $\mathbb{Q}$, and *integral* if it takes values in $\mathbb{Z}$.

It follows from Definition 4.9 that the set of weights on B is a subspace of $\mathbb{R}^{|S(B)|}$ defined by a finite family of integral linear equalities, one equality for each component of $\mathrm{br}(B)$.

NOTATION 4.10. Let $W(B)$ denote the (finite dimensional) real vector space of weights on B, and $W^+(B)$ the convex cone of weights which take non-negative values on every sector. If B is understood, abbreviate these spaces by W and W^+.

There is a close relationship between (non-negative integral) weights on a branched surface B and surfaces mapping to B in a particularly simple way. Since B has a smooth structure, it makes sense to say that a map $f : S \to B$ is an immersion, when S is a smooth surface.

DEFINITION 4.11. Let B be an oriented simple branched surface, possibly with boundary. A *carrying map* is a proper, orientation-preserving immersion $f : S \to B$ from some compact oriented surface S (possibly with boundary) to B. By abuse of notation we say that B *carries* S.

A carrying map $f : S \to B$ determines a non-negative integral weight $w(f)$, whose value on each sector $\sigma \in S(B)$ is the local degree of f along σ. Since a carrying map is an orientation-preserving immersion, the local degree along a sector σ is equal to the number of preimages of any point in the interior. In other words,

$$w(f)(\sigma) = \#\{f^{-1}(p)\} \text{ for } p \in \sigma$$

LEMMA 4.12. *Let B be a simple branched surface. Every non-negative integral weight on B is represented by a carrying map. Conversely, if $f : S \to B$ represents a weight w, then $\chi(S)$ depends only on w, and is a rational linear function of the co-ordinates of $w \in W$.*

PROOF. Let w be a non-negative integral weight. For each sector $\sigma \in S(B)$, take $w(\sigma)$ copies of σ. At each $\gamma \in \mathrm{br}(B)$, the sum of the weights on the sectors on one side is equal to the sum of the weights on sectors on the other side. Choose a bijection between the two sets of copies of sectors, and glue the copies according to this bijection along their edges corresponding to γ. The result of this gluing is a surface S, which comes together with a tautological orientation-preserving immersion to B, realizing the weight w. Moreover, all surfaces representing w arise this way, for various choices of bijections as above.

Each sector $\sigma \in S(B)$ can be thought of as a surface with corners. The corners are the points where arcs of $\mathrm{br}(B)$ run into ∂B. Each such surface σ has an *orbifold* Euler characteristic $\chi_o(\sigma)$ defined by the formula

$$\chi_o(\sigma) = \chi(\sigma) - c(\sigma)/4$$

where $\chi(\cdot)$ denotes ordinary Euler characteristic of the underlying surface, and $c(\cdot)$ denotes the number of (boundary) corners. If a smooth surface S is obtained by gluing surfaces S_i with corners, then $\chi(S) = \sum_i \chi_o(S_i)$. Hence if S is a surface with weight w, then $\chi(S) = \sum_\sigma w(\sigma)\chi_o(\sigma)$, which depends only on w, as claimed. □

REMARK 4.13. Lemma 4.12, though simple to state and prove, is actually surprisingly delicate. The reader whose intuitions have been honed by exposure to train-tracks in surfaces, or *embedded* branched surfaces in 3-manifolds, may not appreciate how subtle such objects really are.

In great generality, a compact Riemann surface lamination is carried by an abstract branched surface, and the space of weights on such a surface is finite dimensional (see [**160**]). For a branched surface embedded in a 3-manifold, a non-negative integral weight determines a unique *embedded* surface which maps to the branched surface by an immersion. However the construction of such a surface depends on the local transverse order structure on branches inherited by codimension 1 objects in a 3-manifold.

If B is an abstract (not necessarily simple) branched surface, and w a non-negative integral weight on B, then from w one can construct a surface S mapping to B, but the map is in general a *branched* immersion, branched over the vertices of $\mathrm{br}(B)$, and χ depends not only on w but on the way S branches at each such point.

To associate an (unbranched) carrying map to a weight one must solve a holonomy problem. Moreover, it might be the case that this holonomy problem can be solved for nw but not for w, where w is a non-negative integral weight, and n is a positive integer.

A similar, and completely analogous phenomenon occurs when one tries to do *immersed* normal surface theory in 3-manifolds.

By contrast, the function χ^- might well depend on the choice of a surface S representing a weight w. For, the number of disk components of S might depend on the way in which sectors are glued up. This motivates the following definition.

DEFINITION 4.14. An oriented simple branched surface is *essential* if it does not carry a disk or sphere.

EXAMPLE 4.15. If every sector satisfies $\chi_o(\sigma) \leq 0$ then $\chi(S) \leq 0$ for any surface carried by B. Consequently in this case, B is essential.

If S is carried by an essential simple branched surface, then every component of S has non-positive Euler characteristic. Consequently $\chi(S) = \chi^-(S)$, and therefore we obtain the following corollary:

COROLLARY 4.16. *Let B be an essential simple branched surface. Then $-\chi^-(S)$ is a linear function of w, where S is a surface realizing a (non-negative integral) weight w.*

4.1.3. Alternating words. As a warm-up, we prove rationality of scl on certain special elements in the free group of rank 2, where the argument is especially transparent. Throughout the sequel we fix notation $F = \langle a, b \rangle$.

DEFINITION 4.17. A word $w \in F$ is *alternating* if it has even length, and the letters alternate between one of $a^{\pm 1}$ and one of $b^{\pm 1}$.

Every alternating word is cyclically reduced. An alternating word is in $[F, F]$ if there are the same number of a's as a^{-1}'s, and similarly for b's and b^{-1}'s. Hence an alternating word in $[F, F]$ has length divisible by 4.

EXAMPLE 4.18. $aba^{-1}b^{-1}$ and $aba^{-1}b^{-1}a^{-1}bab^{-1}$ are examples of alternating words in $[F, F]$.

EXAMPLE 4.19. A word is alternating if and only if in the graphical calculus (see § 2.2.4) it is represented by a loop without backtracks in which every straight segment has length 1.

In what follows, let H be a handlebody of genus 2. We think of H as the union of two solid handles H^+, H^-, glued along a disk E which we call the *splitting disk*. For psychological convenience, we think of H embedded in $\mathbb{R}^3$ in such a way that E is horizontal, H^+ is above, and H^- is below. Let $D^\pm$ be compressing disks for the meridians of $H^\pm$; psychologically, we think of these disks as vertical.

Identify $\pi_1(H)$ with F in such a way that b is represented by the core of the handle H^+ and a is represented by the core of the handle H^-. An alternating word is represented by a particularly simple free homotopy class of loop in H, namely as a union of arcs from E to itself which wind once around either H^+ or H^-, crossing D^+ or D^- transversely in a single point; say that such a representative is in *bridge position*. By convention we assume that a loop in bridge position is *embedded* in H. This is mainly for psychological rather than logical convenience; the isotopy class of γ in H is not relevant in the sequel, only its homotopy class.

In what follows, fix an alternating word w and let γ be a corresponding loop in H in bridge position. Without loss of generality, we can write

$$w = a^{e_1} b^{f_1} a^{e_2} b^{f_2} \cdots a^{e_m} b^{f_m}$$

where each e_i, f_i is ± 1, and m is even, and equal to half the word length of w. Then γ is a union of arcs

$$\gamma = \alpha_1 \cup \beta_1 \cup \cdots \cup \beta_m$$

where α_i is properly embedded in H^-, and winds in the positive direction if $e_1 = 1$, and the negative direction otherwise, and β_i is properly embedded in H^+ and winds similarly according to the sign of f_i. Note that the α_i, β_i are oriented arcs, and the end point of α_i is equal to the initial point of β_i for each i, while the end point of β_i is equal to the initial point of α_{i+1} (indices taken cyclically) for each i.

Let $f : S, \partial S \to H, \gamma$ satisfy $f_*[\partial S] = n[\gamma]$. Recall, by Proposition 2.10 that $\mathrm{scl}(w)$ is equal to the infimum of $-\chi^-(S)/2n(S)$ over all such surfaces. We will show that after possibly replacing S with a simpler surface S' with $n(S') = n(S)$ and $-\chi^-(S') < -\chi^-(S)$, we can homotope f into a particularly simple form.

Assume without loss of generality that S has no disks or closed components, or simple compressing loops, or else $-\chi^-$ could be reduced without affecting n. If some boundary component of S maps to γ with degree 0, we can compress it, reducing $-\chi^-$. So assume that every boundary component maps with nonzero degree, and homotope f so that the restriction of f to each component of ∂S is a covering map to γ. Then perturb f rel. boundary to an immersion in general position with respect to $D^\pm$.

After this perturbation, the preimage $f^{-1}(D^+) \cap S$ is a union of *disjoint, embedded* proper arcs and loops in S. Since by hypothesis S has no simple compressing loops, all the loops are inessential in S, and can be pushed off D^+ by a homotopy of f. Since the restriction of f to ∂S is a covering map, there are no inessential arcs in $f^{-1}(D^+)$, so we may assume that $f^{-1}(D^+)$ consists of a union of disjoint essential embedded proper arcs in S. Do the same for $f^{-1}(D^-)$. After this modification, $f^{-1}(D^+ \cup D^-)$ is a union δ of disjoint essential embedded proper arcs. Let $\mathcal{R}$ be a union of (relatively) open regular neighborhoods in S of the components of δ. The components of $\mathcal{R}$ are called *rectangles*.

The complement of tubular neighborhoods of the $D^\pm$ in H deformation retracts down to the splitting disk E. In fact, there is a deformation retraction of pairs

$$H - N(D^\pm), \gamma - (\gamma \cap N(D^\pm)) \to E, E \cap \gamma$$

Drag f by this deformation retraction, so that after a homotopy, $\mathcal{R}$ is exactly equal to $f^{-1}(H - E)$.

Now consider the components of $S - \mathcal{R}$. Each such component P is a compact surface, whose boundary is broken up into vertices (points in ∂S in the closure of a rectangle of $\mathcal{R}$) and two different kinds of edges: components of $P \cap \partial S$, and components of P in the closure of a rectangle. We refer to the first kind of edges as *boundary edges* and the second kind as *branch edges*. After the homotopy, each boundary edge maps by f to a single point of $\gamma \cap E$, and each branch edge maps to an arc in E. Since E is a disk, if P is not a disk, it contains an essential embedded loop which maps to a null-homotopic loop in E and can therefore be compressed. Since by hypothesis S contains no simple compressing loops, every component P of $S - \mathcal{R}$ is topologically a disk. Since its boundary has a natural cellulation into edges and vertices, we think of P as a *polygon*, whose edges alternate between boundary edges and branch edges. Let $\mathcal{P}$ denote the union of these polygons, and let P_i denote a typical polygon.

For each P_i, let $|P_i|$ denote the number of branch edges of P_i. Observe that the branch edges alternate between arcs bounding rectangles mapping to H^+ and rectangles mapping to H^-. Consequently, each P_i has an even number of branch edges; denote this number by $|P_i|$. Say that a branch edge of P_i *faces up* if it bounds a rectangle mapping to H^+, and it *faces down* otherwise. There are twice as many corners of P_i as branch edges, hence $2|P_i|$ corners.

Since each P_i is topologically a disk, we can compute $\chi_o(P_i) = 1 - |P_i|/2 \leq 0$. Similarly, each rectangle of $\mathcal{R}$ has $\chi_o = 0$. Hence

$$-\chi^-(S) = -\chi(S) = \sum_i \frac{|P_i| - 2}{2}$$

Now fix a single polygon P_i. Suppose that there is a point p of $\gamma \cap E$ and two distinct boundary edges e_1, e_2 of P_i which both map to p. Let β be an embedded arc in P_i joining e_1 to e_2. Doing a boundary compression along β reduces $-\chi^-$ by 1. Hence after repeatedly performing such compressions, we can assume (at the cost of replacing the original surface with another of smaller $-\chi^-$) that every polygon P_i has *at most* $|w|$ boundary edges, which map to *distinct* points of $\gamma \cap E$.

Notice what we have achieved in this discussion. Starting with an arbitrary map $f : S, \partial S \to H, \gamma$ we obtained (after homotopy, compression and boundary compression) a new surface and a new map (which by abuse of notation we still denote S, f) such that S is decomposed into two kinds of pieces: *rectangles* which map over the handles of H, and which run between a pair of arcs of γ, and *polygons* which map to the splitting disk E. Each rectangle is determined, up to homotopy, by the pair of arcs of γ that it runs between. Each polygon is determined up to homotopy by a cyclically ordered list of *distinct* elements of $\gamma \cap E$ that the boundary edges map to in order, and by the data of whether each branch edge faces up or down. There are only finitely many combinatorial possibilities for each rectangle and for each polygon. Thus the surface S is built from finitely many pieces, all drawn from a finite set of combinatorial types.

This last observation is crucial, and reduces the computation of $\mathrm{scl}(w)$ to a *finite* integer linear programming problem. We explain how.

Build an oriented essential simple branched surface B as follows. The sectors of B are the disjoint union of all possible polygons (with boundary edges mapping to distinct points of $E \cap \gamma$) and all possible rectangles. Glue up rectangles to polygons in all possible orientation-preserving ways, ensuring that branch edges that face up and down are only glued to rectangles in H^+ and H^- respectively. The result is an abstract branched surface B and a homotopy class of map $\iota : B \to H$ taking ∂B to γ.

There are two components of the branch locus for each pair of distinct points in $E \cap \gamma$, distinguished by whether such components bound rectangles in H^+ or in H^-. In particular, the branch locus is a 1-manifold, and therefore the branched surface is simple. Furthermore, each polygon contributes non-positively to χ_o and each rectangle contributes 0, so the branched surface is essential.

Since every surface $f : S, \partial S \to H, \gamma$ can be compressed, boundary compressed and homotoped until it is made up of rectangles and polygons, we conclude the following:

LEMMA 4.20. *Let B denote the essential simple branched surface, constructed as above. Then every $f : S, \partial S \to H, \gamma$ can be compressed, boundary compressed and homotoped without increasing $-\chi^-$, to a map which is carried by B.*

Notice that the branched surface B can be constructed effectively from the word w. Let $w \in W^+$ be a non-negative integral weight on B. Let $f : S \to B$ be a carrying map with weight w. The composition $\iota \circ f : S \to H$ takes $\partial S \to \gamma$. Define $\partial(w) = n(S)$, and extend by linearity and continuity to a rational linear map $\partial : W^+ \to \mathbb{R}$. By construction,

$$\mathrm{scl}(w) = \inf_{w \in W^+ \cap \partial^{-1}(1)} \frac{-\chi^-(w)}{2}$$

But $W^+ \cap \partial^{-1}(1)$ is a closed rational polyhedron, and $-\chi^-$ is a rational linear function which is non-negative on the cone W^+, and therefore achieves its infimum on a closed rational polyhedron Q in $W^+ \cap \partial^{-1}(1)$. It follows that $\mathrm{scl}(w)$ is rational. Moreover, given W^+ and the functions $-\chi^-$ and ∂, computing the polyhedron Q is a finite linear programming problem which can be solved by any one of a number of methods. Thus there is an effective algorithm to compute $\mathrm{scl}(w)$.

4.1.4. Bridge position. We extend the arguments in § 4.1.3 in several ways: to free groups of arbitrary rank, and to arbitrary finite integral linear combinations of arbitrary elements.

Let F be a free group with generators a_i. For each i, let H_i denote a solid torus with a marked disk E_i in its boundary, and let H be obtained from the H_i by identifying the E_i with a single disk E. If the rank of F is 2, this is an ordinary genus 2 handlebody, and H_1, H_2 are H^+, H^- from the last section. For each i, let D_i be a decomposing disk for the handlebody H_i, disjoint from E, and denote the union of the D_i by $\mathcal{D}$. Let $w \in F$ be cyclically reduced. The conjugacy class of w determines a free homotopy class of loop in H; we will choose a representative γ in this free homotopy class whose intersection with E and $\mathcal{D}$ is simple.

A *vertical arc* is an arc with endpoints on E whose interior is properly embedded in some $H_i - E$. A *horizontal arc* is an arc embedded in E. The representative γ will have one vertical arc in H_i for each appearance of $a_i^{\pm}$ in w, and one horizontal arc between any two consecutive appearances of $a_i^{\pm}$ (notice, since w is cyclically reduced, that consecutive appearances of $a_i^{\pm}$ have the same sign). This uniquely determines the homotopy class of γ.

DEFINITION 4.21. A representative γ in the free homotopy class corresponding to the conjugacy class of w, constructed as above, is said to be in *bridge position*.

REMARK 4.22. For rank 2 and for alternating words, this agrees with the definition from § 4.1.3.

Let $w_1, \cdots, w_n$ be a finite collection of elements which are cyclically reduced in their conjugacy class, and $\gamma_1, \cdots, \gamma_n$ loops in bridge position in H. Denote the union of the γ_i by Γ. Let $f : S, \partial S \to H, \Gamma$ be given, and assume that S has no disk or closed components, or simple compressing loops. As in § 4.1.3, after a homotopy we can assume that $f^{-1}(\mathcal{D})$ is a union of disjoint essential embedded proper arcs, and $\mathcal{R} = f^{-1}(H - E)$ is a union of disjoint embedded rectangles with the components of $f^{-1}(\mathcal{D})$ as their cores. Since S has no simple compressing loops, as in § 4.1.3 we can conclude that every component P_i of $S - \mathcal{R}$ is a polygon.

The branch edges of the P_i are edges in the closure of components of $\mathcal{R}$, but there are two kinds of boundary edges: those which map to a single endpoint of a vertical arc of some γ_i, and those which map to a horizontal edge. As before, if some polygon has boundary edges e_i, e_j mapping to the same point or horizontal arc of $E \cap \Gamma$, we can do a boundary compression of S to reduce $-\chi^-$. So without loss of generality, we conclude that distinct boundary edges e_i, e_j of the same polygon map to different points or arcs of $E \cap \Gamma$.

Let $|P_i|$ denote the number of branch edges of P_i. As a surface with corners, we have $c(P_i) = 2|P_i|$ so $\chi_o(P_i) = 1 - |P_i|/2$. Rectangles contribute 0 to χ_o, so

$$-\chi^-(S) = -\chi(S) = \sum_i \frac{|P_i| - 2}{2}$$

One can build a simple essential branched surface B as before, together with a homotopy class of map $\iota : B \to H$ with $\iota(\partial B) = \Gamma$. Every map $f : S, \partial S \to H, \Gamma$ can be compressed, boundary compressed and homotoped until it factors through a carrying map to B.

Let K be $\ker : H_1(\Gamma) \to H_1(H)$ induced by inclusion. The vector space K is isomorphic to the intersection $B_1(F) \cap \langle w_1, \cdots, w_n \rangle$. The inclusion map on homology is defined over $\mathbb{Z}$, so K is a rational subspace of $H_1(\Gamma)$. With notation as in § 4.1.3, there is a surjective rational linear map $\partial : W^+ \to K$. For each $k \in K$ there is an equality

$$\mathrm{scl}(k) = \inf_{w \in W^+ \cap \partial^{-1}(k)} \frac{-\chi^-(w)}{2}$$

Now, W^+ is a finite dimensional rational convex polyhedron with finitely many extremal rays, each passing through a rational point v_i, and $-\chi^-$ is a rational linear function. Therefore

$$\mathrm{scl}(k) = \inf \frac{\sum_i -t_i \chi^-(v_i)}{2}$$

where the infimum is taken over all non-negative t_i for which $\sum_i t_i \partial(v_i) = k$. Explicitly, each basis $\mathcal{S}$ of elements v_i determines a rational linear function $f_{\mathcal{S}}$ on W^+ whose value is $-\chi^-(v_i)/2$ on $v_i \in \mathcal{S}$, and $\mathrm{scl} \circ \partial$ is the minimum of this finite collection of functions. In other words, $\mathrm{scl} \circ \partial$ is a piecewise rational linear function on W^+ and therefore scl is piecewise rational linear on K.

Recall that a map $f : S \to H$ is *extremal* if it realizes the infimum, over all surfaces without closed or disk components, of $-\chi^-(S)/2n(S)$. If w is a non-negative rational weight realizing the infimum of $-\chi^-(w)/2$ on $\partial^{-1}(k)$ for some rational class $k \in B_1^H(F)$, then some integral multiple of w is integral. Any carrying map realizing this weight gives rise to an extremal surface, and all extremal surfaces arise in this way.

We have now completed the proof of the Rationality Theorem. In order to state the theorem precisely, we must first say what we mean for a function on an infinite dimensional vector space to be *piecewise rational linear*.

Definition 4.23. Let V be a real vector space. A function ϕ on V is *piecewise linear* if for every finite dimensional subspace W of V, the restriction of ϕ to W is piecewise linear. If $V = V_{\mathbb{Q}} \otimes \mathbb{R}$ where $V_{\mathbb{Q}}$ is a (given) rational vector space, a subspace $W \subset V$ is *rational* if it is of the form $W = W_{\mathbb{Q}} \otimes \mathbb{R}$ for some subspace $W_{\mathbb{Q}} = V_{\mathbb{Q}} \cap W$. A function ϕ on V is *piecewise rational linear* if for every finite

dimensional rational subspace W of V, the restriction of ϕ to W is piecewise linear, and rational on $W_{\mathbb{Q}}$.

Recall from § 2.6.2 that for any group G, the space $B_1(G)$ is the vector space of real (group) 1-boundaries, and $B_1^H(G)$ is the quotient of $B_1(G)$ by the subspace H spanned by elements of the form $g^n - ng$ and $g - hgh^{-1}$. In general, scl is a pseudo-norm on B_1^H, but when G is hyperbolic, scl is a genuine norm (Corollary 3.57).

The results of this section prove the following theorem.

THEOREM 4.24 (Rationality Theorem). *Let F be a free group.*

1. $\mathrm{scl}(g) \in \mathbb{Q}$ *for all* $g \in [F, F]$.
2. *Every* $g \in [F, F]$ *bounds an extremal surface.*
3. *The function* scl *is a piecewise rational linear norm on* $B_1^H(F)$.
4. *Every nonzero finite rational linear chain* $A \in B_1^H(F)$ *projectively bounds an extremal surface*
5. *There is an algorithm to calculate* scl *on any finite dimensional rational subspace of* $B_1^H(F)$*, and to construct all extremal surfaces in a given projective class.*

REMARK 4.25. Note by Proposition 2.104 that every extremal surface as above is π_1-injective.

4.1.5. PQL groups. Motivated by the results of the previous section, we define the following class of groups.

DEFINITION 4.26. A group G is PQL (pronounced "pickle") if scl is piecewise rational linear on $B_1^H(G)$.

EXAMPLE 4.27. An amenable group is trivially PQL, by Theorem 2.47 and Theorem 2.79.

EXAMPLE 4.28. Theorem 4.24 implies that finitely generated free groups are PQL. Suppose F is an infinitely generated free group. Since any finite subset of $B_1^H(F)$ is contained in the image of $B_1^H(F_n)$ for some finitely generated summand F_n, we conclude that F is also PQL.

There are a few basic methods to derive new PQL groups from old.

PROPOSITION 4.29. *Let H be a subgroup of G of finite index. Then if H is PQL, so is G.*

PROOF. Let X be a space with $\pi_1(X) = G$. Let $g_1, \cdots, g_m$ be elements of G whose conjugacy classes are represented by loops $\gamma_1, \cdots, \gamma_m$. Let $\widehat{X}$ be a finite cover of X with $\pi_1(\widehat{X}) = H$. For each i, let $\beta_{i,j}$ be the preimages of γ_i in $\widehat{X}$, and let $h_{i,j} \in H$ be elements whose conjugacy classes represent the $\beta_{i,j}$. By Proposition 2.80, for any integers $n_1, \cdots, n_m$ we have

$$\mathrm{scl}_G(\sum_i n_i g_i) = \frac{1}{[G:H]} \cdot \mathrm{scl}_H(\sum_{i,j} n_i h_{i,j})$$

and the proposition follows. □

Hence virtually free groups are PQL. This class of groups includes fundamental groups of non-compact hyperbolic orbifolds.

PROPOSITION 4.30. *Let $A \xrightarrow{i} G \xrightarrow{q} H \to 1$ be an exact sequence, where A is amenable and H is PQL and satisfies $H^2(H;\mathbb{R}) = 0$. Then G is PQL.*

PROOF. Since A is amenable, Theorem 2.47 says that the bounded cohomology of A vanishes in each dimension. By Theorem 2.50 and Theorem 2.49 one obtains a commutative diagram as in Figure 4.2 with exact rows and columns. Let $\alpha \in Q(G)$

$$\begin{array}{ccccccc}
H^1(H) & \longrightarrow & Q(H) & \xrightarrow{\delta} & H_b^2(H) & \longrightarrow & 0 \\
\downarrow & & \downarrow{\scriptstyle q^*} & & \downarrow{\scriptstyle q^*} & & \\
H^1(G) & \longrightarrow & Q(G) & \xrightarrow{\delta} & H_b^2(G) & \longrightarrow & H^2(G) \\
\downarrow & & \downarrow & & \downarrow & & \\
H^1(A) & \longrightarrow & Q(A) & \longrightarrow & 0 & &
\end{array}$$

FIGURE 4.2. This diagram has exact columns (by Theorem 2.49) and exact rows (by Theorem 2.50).

be given. Then $\delta\alpha \in H_b^2(G)$ is equal to $q^*\beta$ for some $\beta \in H_b^2(H)$, since $H_b^2(H) \to H_b^2(G)$ is surjective. Since $H^2(H)$ is zero, there is some $\gamma \in Q(H)$ with $\delta\gamma = \beta$, and therefore $\alpha - q^*\gamma \in Q(G)$ is in the image of $H^1(G)$. Since α was arbitrary, this says that the composition $Q(H) \to Q(G) \to Q(G)/H^1(G)$ is surjective.

It is a general fact that for any surjection of groups $q : G \to H$, and any quasimorphism ϕ on H, there is an equality $D(\phi) = D(q^*\phi)$ where the left side is the defect of ϕ on H, and the right side is the defect of $q^*\phi$ on G. For,

$$D(q^*\phi) = \sup_{a,b,\in G} |\phi(q(a)) + \phi(q(b)) - \phi(q(ab))| = D(\phi)$$

where the second equality follows from the definition of $D(\phi)$ and surjectivity. By Theorem 2.79, for any $\sum t_i a_i \in B_1^H(G)$ we have

$$\begin{aligned}
\mathrm{scl}_G(\sum t_i a_i) &= \frac{1}{2} \sup_{\phi \in Q(G)/H^1(G)} \frac{\sum_i t_i \phi(a_i)}{D(\phi)} \\
&= \frac{1}{2} \sup_{\phi \in Q(H)/H^1(H)} \frac{\sum_i t_i q^*\phi(a_i)}{D(q^*\phi)} \\
&= \frac{1}{2} \sup_{\phi \in Q(H)/H^1(H)} \frac{\sum_i t_i \phi(q(a_i))}{D(\phi)} \\
&= \mathrm{scl}_H(\sum t_i q(a_i))
\end{aligned}$$

It follows that G is PQL if H is, as claimed. □

REMARK 4.31. If $H^2(H)$ is nonzero, there might be elements in $Q(G)/H^1(G)$ which are not in the image of $Q(H)$. If H is finitely presented, $H^2(H)$ is finitely generated, so $Q(G)/(H^1(G) + q^*Q(H))$ is finite dimensional and is generated by a finite number of quasimorphisms $\phi_1, \cdots, \phi_n$. If one can find generators ϕ_i as above which take on rational values on rational elements of $B_1^H(G)$, then if H is PQL, so is G.

COROLLARY 4.32. *Let M be a noncompact Seifert-fibered 3-manifold. Then $\pi_1(M)$ is PQL.*

PROOF. For M as above there is a central extension $\mathbb{Z} \to \pi_1(M) \to G$ where G is the fundamental group of a noncompact surface orbifold. If G is amenable, so is $\pi_1(M)$, and $\pi_1(M)$ is trivially PQL. Otherwise G is virtually free. In this case there is a finite index subgroup H of $\pi_1(M)$ which is a product $\mathbb{Z} \oplus F$ where F is free. By Proposition 4.30, the group H is PQL, and therefore by Proposition 4.29, so is $\pi_1(M)$. □

EXAMPLE 4.33. Let M be homeomorphic to $S^3 - K$ where K is the trefoil knot. Then M is Seifert fibered and noncompact, so $\pi_1(M)$ is PQL. It is well-known that $\pi_1(M)$ is isomorphic to the braid group B_3 (see e.g. [**16**]).

4.1.6. Implementing the Algorithm. In this section we discuss in more explicit terms the algorithm described implicitly in the last few sections. Proposition 2.13 implies that we can restrict attention to *monotone* admissible maps in order to calculate scl. If $f : (S, \partial S) \to (H, \gamma)$ is monotone, the restriction $\partial S \to \gamma$ is orientation-preserving. This reduces the number of rectangle types that must be considered by roughly a factor of 4, and concomitantly reduces the number of polygon types.

We show how the algorithm runs in practice. For convenience, we restrict attention to alternating words in F_2. In what follows, for the sake of legibility, we denote a^{-1} by A and b^{-1} by B.

EXAMPLE 4.34. Let $w = abABAbaB$. The loop γ is a union of 8 arcs, each arc corresponding to a letter in w. The initial vertex of each arc is a point on E; denote these points $v_0, v_1, \cdots, v_7$. An *admissible arc* is an arc that might be contained in a polygon in a monotone extremal surface. Such an arc is given by an ordered pair (v_i, v_j) where v_i is the initial vertex of an arc corresponding to some letter x or X and v_j is the terminal vertex of an arc corresponding to a letter X or x. Since w is alternating, there are $|w|/4 = 2$ copies of each of the letters a, A, b, B and consequently there are $|w|^2/4 = 16$ admissible arcs (the arc (v_i, v_j) is denoted ij for brevity):

$$03, 21, 14, 32, 05, 41, 10, 72, 27, 63, 54, 36, 47, 65, 50, 76$$

A *polygon* is a cyclically ordered list of vertices, where no vertex appears more than once, and each consecutive pair of vertices is an admissible arc. There are 18 polygons:

$$03214765, 0321, 03276541, 032765, 036541, 03654721, 0365, 2147, 214763,$$

$$210547, 21054763, 14, 3276, 0541, 05, 72, 63, 5476$$

(note that each polygon has an even number of vertices). Each rectangle bounds two admissible arcs, but there is a relation between these two arcs: if a rectangle bounds ij at one end, it bounds $(j-1)(i+1)$ at the other end. The linear programming problem takes place in the vector space $P \cong \mathbb{R}^{18}$ spanned by a basis p_i whose co-ordinates count the number of polygons of type i. Each rectangle imposes one equation, of the form $\sum p_k = \sum p_l$ where the left hand side counts the number of polygons that contain an admissible edge ij and the right hand side counts the number of polygons that contain an admissible edge $(j-1)(i+1)$ (note that a

polygon type might contain both or neither). There are twice as many admissible edges as equations, and hence $|w|^2/8 = 8$ equations:

$$p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = p_0 + p_1 + p_5 + p_7 + p_8 + p_9 + p_{10}$$

$$p_0 + p_7 + p_8 + p_{11} = p_0 + p_1 + p_2 + p_3 + p_8 + p_{10} + p_{12}$$

and so on.

Restricting to geometrically sensible answers imposes the conditions that each $p_i \geq 0$. For each i, let l_i denote the number of branch edges in the polygon of type i. In this example, l_i is equal to the length of the corresponding string of vertices; hence $l_0 = 8$, $l_1 = 4$, $l_2 = 8$, $l_3 = 6$ and so on. To normalize the solution so that the boundary represents $[\gamma]$ in homology, we need to impose the equation

$$\sum_i l_i p_i = |w| = 8$$

Subject to this list of constraints, $\mathrm{scl}(w)$ is the minimum of the *objective function*

$$\frac{-\chi^-}{2} = \sum_i \frac{(l_i - 2)p_i}{4}$$

This linear programming problem can be solved using exact arithmetic, for instance using the GNU package `glpsol` ([**140**]) and Masashi Kiyomi's program `exlp` ([**128**]), returning the answer $\mathrm{scl}(w) = 0.5$. Moreover, an extremal solution describes how to construct an extremal surface consisting of one 4-gon and two bigons $0541 + 72 + 63$ and four rectangles. This exhibits γ as the boundary of a once-punctured torus, and shows that w is a commutator (which is easily seen in any case: $abABAbaB = [a, bAB]$).

See e.g. Dantzig [**62**] for an introduction to linear programming.

EXAMPLE 4.35. Bavard [**8**] p. 148 asked whether scl in the commutator subgroup of free group takes on values in $\frac{1}{2}\mathbb{Z}$. This should be viewed in some sense as the natural analogue of the fact that in a 3-manifold M, the (Gromov-)Thurston norm takes on values in $2\mathbb{Z}$ on the integral lattice $H_2(M; \mathbb{Z})$ (also compare with Proposition 4.4). In fact, the answer to Bavard's question is negative: there are many elements in free groups whose scl is not a half-integer. One explicit example is $w = baBABAbaBabA$; the identity

$$[abaB, ABAbaBabAbaBABAbaBababABB] \cdot [ABAba, BabAbaBABAbba] \cdot [BabABababA, aaBAAb] = a(baBABAbaBabA)^3 A$$

expresses a conjugate of w^3 as a product of three commutators, and defines an extremal surface virtually bounding w. Consequently $\mathrm{scl}(w) = 5/6$. On the other hand, it turns out that elements in free groups with half-integral scl are very *common*; see § 4.1.9 and § 4.2.

The algorithm as described above is hopelessly inefficient for all but a handful of words. In the next section we will describe a much more dramatic improvement, resulting in a polynomial time algorithm.

4.1.7. A polynomial time algorithm to calculate scl **in free groups.** An extremal surface is built from rectangles and polygons. The number of rectangle types is quadratic in the length of w, but the number of polygon types is usually of the order $|w|!$ so a naive implementation of the algorithm described in § 4.1.6 is useless for words of length 20 or more. The problem is the explosion of combinatorial types of polygons with large numbers of sides.

A polygon with many sides is the combinatorial analogue of a critical point of high index — a region in a surface with a high concentration of negative curvature. The basic idea is that a polygon with more than 4 branch edges can be split up, in a natural way, into polygons with 4 or fewer branch edges. For simplicity, in this section we restrict attention to alternating words in F_2, so that the cores of rectangles attached to consecutive branch edges alternate between $a^{\pm}$ or $b^{\pm}$. As an added simplification, shrink boundary edges to points, so that every (remaining) edge is a branch edge. Hence all polygons in question have an even number of sides.

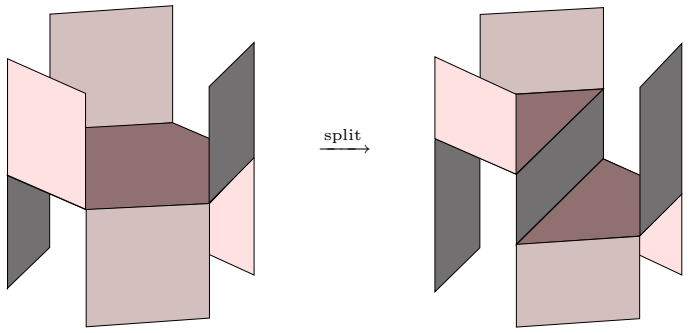


FIGURE 4.3. A hexagon can be split up into two quadrilaterals

Let P be a polygon. The (oriented) rectangles attached to P come in four kinds, depending on whether the core of the rectangle, when it moves away from P, wraps around a, A, b, B; hence if P has more than 4 sides, there are at least two pairs of rectangles of the same kind attached to P. Two (nonadjacent) rectangles of the same combinatorial kind cobound a quadrilateral Q in P. The basic idea is that the polygon P can be split up into Q and (the components of) $P - Q$; see Figure 4.3 for an example. Since the two rectangles which attach to Q wrap around the same handle of the handlebody H, we can "slide" the quadrilateral Q one third of the way around H. After a judicious sequence of slides of this kind, every remaining polygon is a quadrilateral or a bigon.

More precisely, let P be a polygon. The edges of P are labeled by a, A, b, B. All but at most one of the a edges can be paired up, resulting in a union of pairwise disjoint a-quadrilaterals $Q_a \subset P$. Do this pairing in such a way that each region of $P - Q_a$ has at most two boundary edges in ∂Q_a, or else one boundary edge in ∂Q_a and at most one unpaired a edge, then slide the Q_a quadrilaterals $1/3$ of the way around the a handle. For each component P' of $P - Q_a$, pair up all but one of the A edges, resulting in a union of pairwise disjoint A-quadrilaterals $Q'_A \subset P'$. Do this pairing in such a way that each region of $P' - Q'_A$ has at most two boundary edges in $\partial Q'_A$, or else one boundary edge in $\partial Q'_A$ and at most one unpaired A edge, then slide the Q'_A quadrilaterals $1/3$ of the way around the A handle.

By construction, each component P'' of $P' - Q'_A$ has at most 8 edges, half of which are b or B edges. If P'' has 4 or 2 edges, we leave it alone. If it has 6 edges, there are (without loss of generality) at least 2 b edges which span a quadrilateral Q''_b. In this case, slide the Q''_b quadrilateral 1/3 of the way around the b handle and observe that $P'' - Q''_b$ is the union of a quadrilateral and a bigon. Otherwise, suppose P'' has 8 edges. Suppose there are a pair of antipodal b or B edges. Then these span a quadrilateral Q''_b or Q''_B, and the complement in P'' is a union of two quadrilaterals. Otherwise, there are a pair of adjacent b edges and a pair of adjacent B edges spanning disjoint quadrilaterals Q''_b and Q''_B so that $P'' - Q''_b - Q''_B$ is a single quadrilateral and two bigons. In every case, after sliding Q''_b and Q''_B quadrilaterals 1/3 of the way around the b and B handles, we have achieved the desired reduction.

The final result is a surface (homotopic to the original extremal surface) made up of quadrilaterals and bigons in E, quadrilaterals 1/3 or 2/3 of the way around the handles, and (parts of) rectangles joining them up. The number of combinatorial types of (sub-) rectangles is still quadratic in $|w|$, but now the number of polygon types is of order $O(|w|^4)$. This data can be turned into a linear programming problem in $O(|w|^4)$ variables, with $O(|w|^2)$ equations. Each equation is linear in the variables, with coefficients in the finite set $\{\pm 1, \pm 1/2, 0\}$, so the data of the problem can be encoded with $O(|w|^6)$ bits. There are several well-known polynomial time methods of exactly solving a linear programming problem. For example, Karmarkar's projective method [**122**] takes time $O(n^{3.5}L)$ to exactly solve a linear programming problem in n variables encoded in L bits.

For non-alternating words, or free groups of higher rank, one must allow a larger (but still finite) set of combinatorial polygon types; the details are very similar to the alternating case. Hence we have the following:

PROPOSITION 4.36. *Let F be a free group. There is an algorithm to compute* $\mathrm{scl}(w)$ *for $w \in F$ whose running time is polynomial in the word length $|w|$.*

4.1.8. Foldings. In fact, for alternating words, even more simplification is possible. The basic idea is as in the previous section. Suppose S is an extremal surface with boundary on γ which contains a polygon P with more that 4 sides (after collapsing boundary edges). Then we can split off a quadrilateral and slide it around a handle. Instead of sliding it only a third of the way, slide the quadrilateral all the way around the handle. The fact that S is π_1-injective ensures that the quadrilateral does not run into another polygon when it gets all the way around the handle. However it might easily join up with some other polygon P' along an edge, and it is not clear that the result of this quadrilateral slide has made things less complicated rather than more.

The problem can be simplified using graphs, and a procedure due to Stallings [**191**] called *folding*. We replace the map of spaces $f : S \to H$ by a map of graphs $g : \Gamma \to X$ where X is a wedge of two circles (i.e. the core of the handlebody H) and Γ is the graph with one vertex for every polygon in S and one edge for every rectangle. The map g is simplicial, taking edges to edges and vertices to vertices. Let $X' \to X$ be the two-fold covering which unwraps each handle, and $g' : \Gamma' \to X'$ the map induced by g on a suitable covering space Γ' of Γ. Note that Γ and X are homotopy equivalent to S and H respectively; since extremal maps are π_1-injective, the map g is π_1-injective, and so is g'.

The graph X' is 4-valent, with two vertices. Make X' a directed graph in the following way. At each vertex of X' there are four edges, labeled a, A, b, B. Orient the edges of X' so that at one vertex, the a, A edges are outgoing, and at the other vertex the b, B edges are outgoing.

Stallings calls a simplicial map between graphs an *immersion* when it is injective on the star of every vertex. If $p : G_1 \to G_2$ is a simplicial map between graphs which is not an immersion, Stallings shows how to modify G_1 by a sequence of moves called *folds* which do not change the image of $\pi_1(G_1)$ under p_*, so at the end of the sequence of folds the resulting map is an immersion. If p is π_1-injective, each fold is an elementary collapse: two edges of G_1 which share one endpoint in common, and map to the same edge of G_2, are identified. The result of a maximal sequence of folds is well-defined independent of the choice of the sequence of folds. In fact, let $\tilde{G}_2$ denote the universal cover of G_2, which is a tree. Then $p_*(\pi_1(G_1))$ acts on $\tilde{G}_2$, and there is a unique minimal invariant subtree, whose quotient is isomorphic to the maximal folding of G_1.

In our case, since the map $g' : \Gamma' \to X'$ is already π_1-injective, each fold is an elementary collapse. There are two kinds of folds, distinguished by the orientation on X': graphically, we can perform a fold when a $\vee$ subgraph of Γ' maps to a single edge of X', by identifying the two edges of the $\vee$. If the vertex of the $\vee$ maps to the initial vertex of the directed edge of X', we say this is a *positive fold*, otherwise a *negative fold*; by abuse of notation, we say that a $\vee$ admits a positive fold, and a $\wedge$ admits a negative fold. Since g' is π_1-injective, a $\vee$ and a $\wedge$ can share at most one edge in common, and therefore consecutive positive and negative folds can be performed in either order. Hence we can arrange to perform all positive folds first, then all negative folds, in some maximal sequence of folds.

Let $f' : S' \to H'$ be the associated maps of double covers. Note that the composition $S' \to H' \to H$ is extremal if S is. The orientation on X' gives an unambiguous sense to what it means to slide a quadrilateral of S' over a handle of H' in the positive direction. If S' has a polygon P with at least 6 edges, then we can slide some sub-quadrilateral Q of P in the positive direction. The effect of this on the graph X' is to perform a positive fold and then the inverse of a negative fold. In other words, after sliding finitely many quadrilaterals of S', we can arrange matters so that the graph Γ' admits a maximal folding sequence with no positive folds. But such a graph admits no positive folds at all, and therefore Γ' represents a surface S' in which no polygon has more than 4 edges. In words: if w is an alternating word in F_2, some extremal monotone surface for w contains no polygons with more than 4 branch edges.

In the case that the rank is bigger than 2, replace H by a union of genus 1 solid handlebodies glued along their splitting disks as in § 4.1.4. The associated graph X is a wedge of n circles, and X' is a $2n$-valent directed graph with two vertices, at each of which there are n incoming edges and n outgoing edges. If S as above contains a polygon P with at least $2n+2$ edges, we can slide a sub-quadrilateral in the positive direction, thus performing a positive fold and the inverse of a negative fold on Γ. After sliding all quadrilaterals as far as they will go in the positive direction, the resulting graph Γ' admits no positive folds, and therefore the surface S' contains no polygons with more than $2n$ branch edges.

Hence we have proved:

PROPOSITION 4.37. *Let w be an alternating word in F_n. Then some extremal surface for w contains no polygons with more than $2n$ branch edges.*

This proposition leads to a further dramatic reduction in the time needed to compute scl on alternating words, especially in F_2. The resulting algorithm has been implemented in the program `scallop`, whose source is available from [**39**]. In practice the runtime is quite modest, taking on average about 6 seconds on a late 2008 MacBook Pro to compute scl on an alternating word of length 60 in F_2.

REMARK 4.38. A decomposition of a surface into rectangles and polygons determines a vector field on the surface with a saddle singularity for every 4-gon, and an n-prong monkey saddle singularity for every $2n+2$-gon. Such data defines a branched Euclidean metric on the surface where the negative curvature is concentrated at the singularities. Bounding the number of sides of the polygons is the combinatorial equivalent of finding two sided curvature bounds for a smooth surface. A closed least area surface in a non-positively curved 3-manifold has two sided curvature bounds, but for a surface with boundary, there are no such *a priori* lower bounds. Thus it is perhaps somewhat surprising that such uniform lower bounds on the complexity of the polygons in an extremal surface can be obtained, independent even of γ.

4.1.9. Gaps, limits, tongues. An alternating word in F_2 has length $4n$ for some n. There are $2 \cdot (2n!)^2/(n!)^4$ alternating words of length $4n$, but after applying conjugation and anti-involutions $a \leftrightarrow A$ and $b \leftrightarrow B$ if necessary, we may assume the word starts with ab.

Computer experiments using `scallop` reveal unexpected structure in the scl spectrum of F_2.

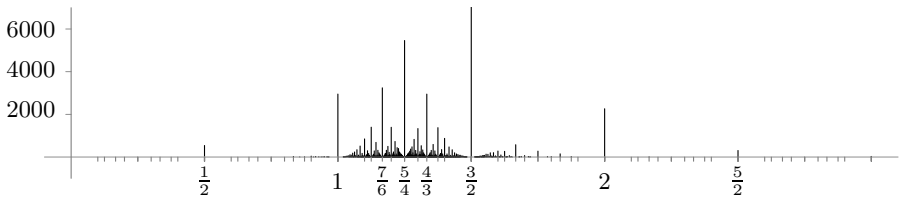


FIGURE 4.4. Values of scl on 50,000 random alternating words of length 36. The horizontal axis is scl and the vertical axis is frequency (the spike at 3/2 is attenuated to fit in the figure).

Figure 4.4 is a histogram of values of scl on random alternating words of length 36. There are several conspicuous features of this plot, including:

(1) the existence of a spectral gap between 0 and 1/2 (discussed in § 4.3.4)
(2) the indiscreteness of the set of values attained
(3) the relative abundance of elements whose scl has a small denominator

The self-similarity of the histogram suggests the existence of a *power law* for the frequency of elements with scl a given rational, of the form $\text{freq}(p/q) \sim q^{-\delta}$ where $\delta \sim 2$ in this example. This self-similarity persists on a fine scale (see Figure 4.5). Co-ordinates of the spikes are obtained by Farey addition of nearest spikes, after multiplying numerators by 2.

Similar power laws occur in dynamical systems, e.g. in the phenomenon of "frequency locking" for coupled nonlinear oscillators. One of the best-known examples

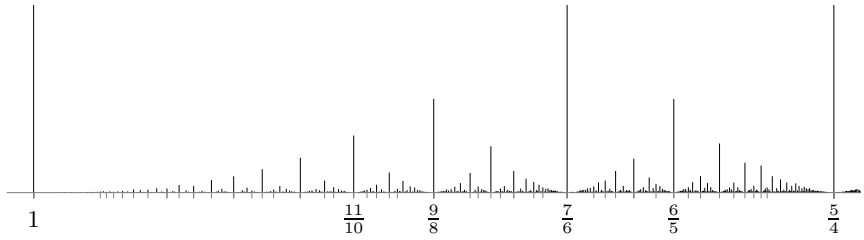


FIGURE 4.5. A stretched scaled excerpt from Figure 4.4.

is that of *Arnold tongues* (see [**2**]). For $K \in [0,1]$ and $\omega \in S^1$ define a function $f_{K,\omega} : S^1 \to S^1$ by

$$f_{K,\omega}(\theta) = \theta + \omega - K \sin 2\pi\theta$$

This is a homeomorphism for $K \leq 1$, and one can look at the rotation number $\mathrm{rot}(f_{K,\cdot})$ as a function of ω for varying K. In fact, for $K > 0$, the set of ω for which this rotation number is a given rational is a nonempty closed interval, and these intervals expand as $K \to 1$ to completely fill out the circle (in measure). Following [**116**] we define $\Delta(p/q)$ to be the length of the interval of values ω for which the rotation number is equal to p/q for $K = 1$. Jensen et. al. [**116**] found experimentally that the heights $\Delta(p/q)$ obey a power law, with $\Delta(p/q) \sim q^{-\delta}$ for $\delta = 2.292 \pm 3.4 \times 10^{-3}$.

The indiscreteness of the spectrum is more evident when one includes non-alternating words.

EXAMPLE 4.39. For positive integers n, m define $s(n,m) = \mathrm{scl}([a, b^n][a, b^{-m}])$. Then $s(n,m) = s(m,n)$, and $s(n,m) = 1 - 1/t(n,m)$ where $t(n,m) = t(n/d, m/d)$ if $\gcd(n,m) = d$, and

$$t(n,m) = \max(2n - 2m, n) \text{ if } \gcd(n,m) = 1 \text{ and } n > m$$

In particular, every value of $\mathbb{Q}$ mod $\mathbb{Z}$ is achieved in scl of F_2 (and therefore in any nonabelian free group).

For a proof and a (partial) explanation, see [**46**]. On the other hand, not every positive rational number occurs as a value of scl in a free group. As has been remarked before, $\mathrm{scl}(w) \geq 1/2$ for all nontrivial $w \in [F_2, F_2]$, and the value of $1/2$ is realized on every commutator. Experimentally, there appears to be another gap in the spectrum between $1/2$ and $7/12$, then a gap between $7/12$ and $5/8$, with the first accumulation point of the set $\mathrm{scl}([F_2, F_2])$ at $3/4$ (of course, each nonzero value is achieved on infinitely many conjugacy classes; compare with Theorem 3.11). Finally, experiments suggest that *every* rational number ≥ 1 is in the scl spectrum.

4.1.10. Injective, extremal, isometric maps. A map $f : \pi_1(S) \to G$ of a surface group into a group G is *injective* if it is a monomorphism, and *extremal* if it realizes the infimum of $-\chi^-(S)/2n(S)$ for its boundary. Say it is *isometric* if $\mathrm{scl}(f(a)) = \mathrm{scl}(a)$ for all $a \in [\pi_1(S), \pi_1(S)]$ (note that injective and isometric maps make sense between arbitrary groups). There are inclusions

$$\text{isometric} \subset \text{extremal} \subset \text{injective}$$

It is an interesting problem to delineate precisely the difference between these three natural classes of surfaces.

EXAMPLE 4.40. Any automorphism is isometric.

EXAMPLE 4.41. If an inclusion $f : G \to H$ splits, then f is isometric.

EXAMPLE 4.42. For any nonzero integers n, m the map $F_2 \to F_2$ sending $a \to a^n$ and $b \to b^m$ is isometric (see [**46**] for a proof).

EXAMPLE 4.43 (once punctured torus). Any map $f : F_2 \to F_2$ has image which is either cyclic or injective. Furthermore, since $1/2$ is a lower bound on nontrivial elements for scl in a free group, every injective map from F_2 to itself (or to any free group) is extremal.

EXAMPLE 4.44 (high distance Heegaard splittings). The following example was inspired by an idea of Geoff Mess. A *Heegaard splitting* exhibits a closed 3-manifold M as a union of two handlebodies H_1, H_2 glued along a surface S. Recall (Definition 3.69) the definition of the complex of curves $\mathcal{C}(S)$. Each handlebody H_i determines a subcomplex $\mathcal{C}(H_i)$ in the complex of curves $\mathcal{C}(S)$ consisting of isotopy classes of essential simple closed curves in S which bound disks in H_i. The *distance* of a Heegaard splitting is the length of the shortest path in the 1-skeleton of $\mathcal{C}(S)$ from a vertex in $\mathcal{C}(H_1)$ to a vertex in $\mathcal{C}(H_2)$. 3-manifolds with Heegaard splittings of arbitrarily high distance and genus exist, and are easy to construct (see e.g. Hempel [**108**]). Let M be a 3-manifold with a Heegaard splitting of genus at least 3 and distance at least 2. Let $\alpha \subset S$ bound a disk in H_1, and separate S into two subsurfaces of different genus. Since the distance of the splitting is at least 2, every simple essential loop in S which bounds a disk in H_2 must intersect α non-trivially. Hence, by the loop theorem (see [**107**] p. 39) the components of $S - \alpha$ are π_1-injective in H_2. Since H_2 is a handlebody, $\pi_1(H_2)$ is free (of rank ≥ 3). This example shows that there are (many) injective surfaces in free groups which are not extremal. Note that a free group of any rank can be included into a free group of rank 2, so there are examples of injective, non-extremal surfaces in free groups of any rank.

EXAMPLE 4.45. Another example is due to Justin Malestein, based on Witt identities. Let F be a free group with generators $x_1, x_2, \cdots, x_n$ for some large n. Define

$$s_i = \begin{cases} x_1 & \text{if } i = 1, 2 \\ x_{2+(i-1)/2} x_{1+(i-1)/2} x_{2+(i-1)/2}^{-1} & \text{if } i > 2 \text{ is odd} \\ x_{1+i/2}[[\cdots[x_1, x_2], x_3], \cdots, x_{-1+i/2}] x_{1+i/2}^{-1} & \text{if } i > 2 \text{ is even} \end{cases}$$

Then one can verify that for each $g \leq n/2$, the elements $s_1, s_2, \cdots, s_{2g}$ generate a free subgroup of F of rank $2g$, and moreover that there is an identity

$$[s_1, s_2] \cdots [s_{2g-1}, s_{2g}] = [[x_1, \cdots, [x_{g-1}, x_g] \cdots], x_{g+1}]$$

thus exhibiting a genus g surface group and a genus 1 surface subgroup of F with the same boundary.

For example, if $g = 2$, one has the identity

$$[s_1, s_2][s_3, s_4] = [x_1, x_2] x_3 [x_2, x_1] x_3^{-1} = [[x_1, x_2], x_3]$$

Since every subgroup of a free group is free, there are no injective maps from *closed* surface groups to free groups. However, we can use extremal surfaces to

construct injective maps from closed surface groups to many groups obtained from free groups by simple procedures.

A well-known question due to Gromov [**98**] is the following:

QUESTION 4.46 (Gromov). *Does every* 1*-ended word-hyperbolic group contain a closed hyperbolic surface subgroup?*

This question seems to be far beyond the reach of current technology. Nevertheless, as an application of the Rationality Theorem, we can find such surfaces in certain groups, obtained as graphs of free groups amalgamated along cyclic subgroups (for an introduction to the theory of graphs of groups, see e.g. Serre [**187**], especially Chapter 1).

THEOREM 4.47. *Let G be a finite graph of free groups, amalgamated along cyclic subgroups.*

1. *Every $\alpha \in H_2(G;\mathbb{Z})$ has a multiple which is represented by a π_1-injective map of a closed surface (which may be disconnected).*
2. *The unit ball of the Gromov (pseudo-)norm on $H_2(G;\mathbb{R})$ is a finite sided rational polyhedron.*
3. *Let $g_1, g_2, \cdots, g_n \in G$ be conjugate into (free) vertex subgroups of G. Then* scl *is piecewise rational linear on $\langle g_1, \cdots, g_n\rangle \cap B_1^H(G)$, and every rational chain in this subspace rationally bounds an extremal surface.*

REMARK 4.48. If some homology class in G is represented by a $\mathbb{Z}\oplus\mathbb{Z}$, the Gromov pseudo-norm on $H_2(G;\mathbb{R})$ is degenerate. In this case, the proposition should be construed as saying that $\|\cdot\|_1$ is a non-negative convex piecewise rational linear function. On the other hand, if G is word-hyperbolic, $\|\cdot\|_1$ is a genuine (polyhedral) norm.

REMARK 4.49. In contrast with the case of a 3-manifold, the norm $\|\cdot\|_1$ does not generally take integral values on $H_2(G;\mathbb{Z})$.

We give the sketch of a proof; for details, see [**43**].

PROOF. Since G is a graph of free groups amalgamated along cyclic subgroups, there is a $K(G,1)$, denoted X, obtained as a union $X = H \cup A$, where H is a disjoint union of handlebodies, and A is a disjoint union of annuli attached along their boundary to essential loops in H (in fact, this can be taken to be the definition of a graph of free groups amalgamated over cyclic subgroups). If H_i is a component of H, let F_i denote the corresponding (free) vertex subgroup of G. Furthermore, for each i, let $\partial_i A$ denote the components of ∂A attached to H_i. We think of each $\partial_i A$ either as a set of free homotopy classes of loops in H_i, or as a set of conjugacy classes in F_i.

Let $\alpha \in H_1(G;\mathbb{Z})$ be given, and let $f : S \to X$ be a map of a surface representing α. After compression and a homotopy, we can insist that $f^{-1}(A)$ is a union of annuli, each of which maps to some component of A by a covering map. Write S as a union $S = T \cup U$, where $U = f^{-1}(A)$, and $T = \cup_i T_i$ where $T_i = f^{-1}(H_i)$. The image $f_*(\partial T)$ is a chain C which can be written as a formal sum $C = \sum C_i$ where each C_i has support in H_i. By construction, $C_i \in \langle \partial_i A\rangle \cap B_1^H(F_i)$.

For each i, let T_i' be an extremal surface in H_i virtually bounding the chain C_i. By passing to common covers if necessary, we can assume that $T' = \cup T_i'$ virtually bounds C. We would like to build a surface S' by gluing up boundary components of the T_i' along covers of the cores of the annuli A. This can be accomplished by passing to a further finite cover, by Proposition 2.13. The resulting surface S' is

Gromov norm minimizing in its (projective) homology class, and is therefore π_1-injective. This proves bullet (1). Bullet (2) follows from the piecewise rational linearity of the scl norm on each $\langle \partial_i A \rangle \cap B_1^H(F_i)$.

The proof of bullet (3) is similar. Let Γ be a collection of loops representing the conjugacy classes g_i. Any admissible surface $f : S, \partial S \to X, \Gamma$ can be homotoped and compressed until $f^{-1}(A)$ is a union of annuli, each of which maps to some component of A by a covering map. Then the claim follows as above by the fact that scl is piecewise rational linear on each subspace of the form $\langle \partial_i A \cup (\Gamma \cap H_i) \rangle \cap B_1^H(F_i)$. □

EXAMPLE 4.50. Let F be a free group and Z a nontrivial cyclic subgroup, contained in $[F, F]$. Let G be obtained from two copies of F by amalgamating them along Z; i.e. $G = F *_Z F$. Topologically, if γ is a loop in H representing the conjugacy class of a generator of Z, the group G is the fundamental group of the space X obtained by gluing two copies of H together along γ. There is an involution ι on X which exchanges the two copies of H, and fixes γ. If $f : S \to H$ is an extremal surface which (rationally) bounds some cover of γ, there is a map Df from the double DS to X obtained by reflecting f across γ using ι. By construction, the map is injective, and realizes the Gromov norm on some multiple of the generator of $H_2(G; \mathbb{Z})$.

EXAMPLE 4.51. Let S be a closed orientable surface, and let $A \subset S$ be an essential annulus in S. Let $g_1, \cdots, g_n \in \pi_1(S)$ be conjugacy classes represented by loops γ_i in $S - A$. Then scl is piecewise rational linear on the subspace $\langle g_1, \cdots, g_n \rangle \cap B_1^H(\pi_1(S))$.

EXAMPLE 4.52. Let G be a graph of free groups amalgamated along cyclic subgroups. Then every finite index subgroup G' is also a graph of free groups amalgamated along cyclic subgroups. So if some finite index G' as above has non-trivial H_2, it contains a closed surface subgroup, and therefore so does G. Cameron Gordon and Henry Wilton [**94**] have several interesting criteria to guarantee this condition.

REMARK 4.53. Compare the proof of Theorem 4.47 with the proofs of Theorem 2.93 and Theorem 2.101.

4.2. Geodesics on surfaces

The results of § 4.1 let us compute scl and construct extremal surfaces for arbitrary elements and chains in $B_1^H(F)$ where F is a free group. Bavard duality implies the existence of extremal quasimorphisms with rational values and rational defects, but such quasimorphisms are apparently quite elusive, and it remains a challenging problem to try to construct them. The most constrained extremal quasimorphisms (and therefore the easiest to find) should be those dual to top dimensional faces of the scl polyhedron; but for an infinite dimensional polyhedron, it becomes complicated even to give a precise definition of a top dimensional face.

However, it turns out that there *are* some naturally occurring top dimensional faces of the scl polyhedron for F a free group. More precisely, for each realization of F as $\pi_1(S)$ where S is an oriented surface (necessarily of negative Euler characteristic), there is a top dimensional face π_S of the scl norm ball. Moreover, the projective class of the chain ∂S in $B_1^H(F)$ intersects this face in its interior, and

the unique homogeneous quasimorphism dual to this face (up to scale and elements of $H^1(F)$), is the *rotation quasimorphism* associated to the natural action of $\pi_1(S)$ on the circle at infinity of hyperbolic space coming from any choice of hyperbolic structure on S. This is Theorem 4.78, to be proved in the sequel.

4.2.1. Self-intersections. Fix the following conventions. Let S be an orientable surface of finite type (usually compact and connected with nonempty geodesic boundary) with $\chi(S) < 0$. If we fix a hyperbolic structure on S, then every free homotopy class of loop has a unique (unparameterized) geodesic representative. If $a \in \pi_1(S)$, and $[a]$ denotes the conjugacy class of a, then we let $\gamma(a)$ denote the geodesic in the free homotopy class determined by $[a]$. If we want to refer specifically to the hyperbolic metric g on S, we write $\gamma(a, g)$.

We recall from § 3.5.3 the notation $\mathrm{cr}(a)$ for the number of self-intersections of $\gamma(a)$ in S (i.e. the *crossing number*). Our discussion in § 3.5.3 was brief and somewhat sketchy; we are more careful now.

The combinatorics of the geodesic $\gamma(a)$ in S does not depend on the choice of hyperbolic structure when $\mathrm{cr}(a) \leq 2$. But when γ has 3 or more self-intersections the combinatorics of γ may (and usually will) depend on the geometry of S. In particular, three local sheets might undergo a "Reidemeister 3" move; see Figure 4.6.

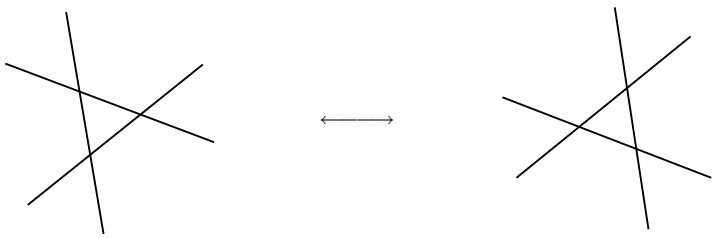


FIGURE 4.6. A Reidemeister 3 move

More subtly, a geodesic representative might not be in general position, and a "coincidental" triple point might be stable under deformations of the hyperbolic structure.

EXAMPLE 4.54 (Hass–Scott [**104**]). This example is a straightforward variation on Example 5 from [**104**]. A hyperbolic once-punctured torus T has an isometric involution which fixes the boundary and three interior (Weierstrass) points. A suitable free homotopy class of loop in T invariant by this involution has a geodesic representative which is forced to go through some or all of these points, an arbitrary number of times. This example can be inserted into any non-planar hyperbolic surface.

Self-intersections and crossing number are more properly defined in terms of linking data at infinity. Let S^1_∞ denote the circle at infinity of the hyperbolic plane. Two disjoint pairs of points in S^1_∞ are said to be *linked* if each separates the other in S^1_∞. Formally, we define a self-intersection of γ as follows. Let $s : S^1 \to \gamma \subset S$ be a parameterization of γ. By abuse of notation, we say that a *lift* of s is a map $\tilde{s} : \mathbb{R} \to \mathbb{H}^2$ which intertwines the covering projections $\mathbb{R} \to S^1$ and $\mathbb{H}^2 \to S$. A *self-intersection* of γ is an unordered pair of lifts $\tilde{s}_1, \tilde{s}_2$ for which the endpoints of the geodesics $\tilde{s}_1(\mathbb{R}), \tilde{s}_2(\mathbb{R})$ are linked in S^1_∞, up to the action of the deck group $\pi_1(S)$

on such pairs. Then define $\mathrm{cr}(a)$ to be the cardinality of the set of self-intersections of $\gamma(a)$.

Linking number is well-defined independent of the hyperbolic structure on S, so this notion is purely topological. For primitive geodesics in general position, the cardinality of the set of self-intersections agrees with the naive (geometric) definition of crossing number, and satisfies the desirable property $\mathrm{cr}(a^n) = n^2\mathrm{cr}(a)$.

If γ is not generic, we distinguish the abstract set of self-intersections (as defined above) from the *support* of the self-intersections, which is a finite subset of γ, and whose cardinality might depend on the hyperbolic structure on S.

4.2.2. Bounding surfaces. Assume now that S is compact, possibly with boundary. Fix a hyperbolic structure on S and an element $a \in \pi_1(S)$, and let γ denote the (oriented) geodesic corresponding to the conjugacy class of a.

For a given hyperbolic structure, γ decomposes $S - \gamma$ into a finite collection of complementary regions R_i. Each region inherits an orientation from S. Moreover, γ is decomposed by its own self-intersections into a collection of oriented segments γ_j. Finally the support of the self-intersections is a collection of oriented points v_i.

DEFINITION 4.55. Let $C_*(\gamma)$ be the chain complex (over $\mathbb{Z}$) generated by the oriented polyhedra R_i, γ_i, v_i together with the boundary components of S, with boundary maps the usual boundaries for polyhedra. Let $H_*(\gamma)$ denote the homology of this complex.

We let S_C denote the element of C_2 which is just the sum of the oriented generators of C_2, and γ_C the element of C_1 which is the sum of the oriented generators of C_1, excluding the boundary components.

Fix an open covering of S whose open sets are regular open neighborhoods U_i of the regions R_i. At least when S is closed, the Čech cohomology of the nerve of this covering (with constant coefficients) is canonically (because of orientations) Poincaré dual to C_*. In particular, there is a canonical surjective homomorphism from ordinary (Čech) homology $H_*(S;\mathbb{Z}) \to H_*(\gamma)$, and the classes $[S_C], [\gamma_C] \in H_*(\gamma)$ are the images of the corresponding elements in $H_*(S;\mathbb{Z})$. There is a similar interpretation of $H_*(\gamma)$ in Čech homology when S has boundary.

LEMMA 4.56. *The kernel of $\partial : C_2(\gamma) \to C_1(\gamma)$ is generated by S_C if S is closed, and is zero otherwise.*

PROOF. This follows by the remarks in the paragraph above, together with the fact that S is connected and orientable, and therefore $H_2(S;\mathbb{Z})$ is at most 1 dimensional, and is 0 dimensional unless S is closed. □

Since γ is closed, γ_C is a cycle. If $a \in [\pi_1(S), \pi_1(S)]$ then $[\gamma] = 0 \in H_1(S)$, so $\gamma_C = \partial A_\gamma$ for some $A_\gamma \in C_2$. If S is not closed, ∂ is injective on C_2 by Lemma 4.56, and therefore A_γ is uniquely defined. For each region R_i, let w_i denote the coefficient of the generator R_i in A_γ, so that $A_\gamma = \sum_i w_i R_i$.

Let T be a compact orientable surface, possibly with multiple boundary components, and let f be a map of pairs $f : (T, \partial T) \to (S, \gamma)$. If we put f in general position, f restricts to a proper map between open surfaces $T - f^{-1}(\gamma) \to S - \gamma$. The orientations on T and S determine a *degree*, denoted $\deg(f)$, which is an assignment of an integer to each region R_i; i.e. an element of $C_2(\gamma)$. If f is smooth, the degree of f on R_i is the signed sum of preimages of a generic point in R_i. One

way of thinking of the degree is as the image of the fundamental class of the pair $(T, \partial T)$ in a suitable relative homology group.

Enumerate the components of ∂T as $\partial_i T$, and suppose that $f(\partial_i T)$ represents γ^{n_i} in $\pi_1(S)$. We define the degree of $f|_{\partial T}$ similarly, and write $\deg(\partial_i f) = n_i \gamma_C$ and $\deg(\partial f) = \sum_i n_i \gamma_C$. Write $n(T) = \sum n_i$ as above. From the definition we have

$$\partial \deg(f) = \deg(\partial f)$$

and so from Lemma 4.56, we deduce

$$\deg(f) = n(T) \cdot A_\gamma$$

providing S has nonempty boundary.

4.2.3. Area norm. Throughout this section, all surfaces under discussion are assumed to have nonempty boundary, unless we explicitly say to the contrary.

DEFINITION 4.57. For $a \in [\pi_1(S), \pi_1(S)]$ and for a fixed choice of hyperbolic metric g on S, define the *area* of $\gamma(a, g)$ by

$$\text{area}(\gamma(a, g)) = \sum_i w_i \text{area}(R_i)$$

where $A_\gamma = \sum w_i R_i$, and

$$\text{area}^+(\gamma(a, g)) = \sum_i |w_i| \text{area}(R_i)$$

If g and a are understood, we abbreviate this to $\text{area}(\gamma)$ and $\text{area}^+(\gamma)$ respectively.

From the definition there is an inequality $\text{area}^+(\gamma) \geq |\text{area}(\gamma)|$ with equality if and only if all the w_i have the same sign.

DEFINITION 4.58. If all the w_i have the same sign, then γ is *monotone*.

LEMMA 4.59. *Let a, γ be as above. Then for any hyperbolic structure g on S there is an inequality*

$$\text{scl}(a) \geq \frac{\text{area}^+(\gamma(a, g))}{4\pi}$$

PROOF. For each surface $(S_i, \partial S_i) \to (S, \gamma)$ we either compress S_i along an essential embedded loop or arc, or else we can find a pleated representative. The pleated representative defines a hyperbolic structure on S_i with totally geodesic boundary. Moreover, by definition, we have

$$\text{area}(S_i) = \sum_i \int_{R_i} \#\{f^{-1}\} d\text{area} \geq \sum_i \int_{R_i} |\deg(f) \text{ on } R_i| d\text{area} = n(S_i)\text{area}^+(\gamma)$$

By Gauss–Bonnet, $\text{area}(S_i) = -2\pi\chi(S_i)$. By Proposition 2.10, $\text{scl}(a)$ is the infimum of $-\chi(S_i)/2n(S_i)$ over all such S_i. □

Values of $\text{area}(\gamma)$ are *quantized*:

LEMMA 4.60. *For any $a \in \pi_1(S)$ and any hyperbolic metric g,*

$$\text{area}(\gamma(a, g)) \in 2\pi\mathbb{Z}$$

In particular, $\text{area}(\gamma)$ *does not depend on g.*

PROOF. Let $(S', \partial S') \to (S, \gamma)$ be a pleated surface for which $n(S') = 1$. The pleated surface structure determines a decomposition of S' into an even number of ideal triangles, whose areas sum to $\text{area}(S')$. The Jacobian $J(f)$ is constant on each ideal triangle, and takes values in ± 1. We calculate

$$\text{area}(\gamma) = \sum_i \int_{R_i} \deg(f) \text{ on } R_i \, d\text{area} = \int_{S'} J(f) \, d\text{area}$$

which is a sum of an even number of π's and $-\pi$'s. □

In fact, the relationship between area and scl is precise enough to detect a significant amount of topological information. An immersion $f : T \to S$ between oriented surfaces is *positive* if it is orientation-preserving on each component, and *negative* if it is orientation-reversing on each component. Note that if S and T are both connected, every immersion between them is either positive or negative.

For the moment we are considering immersed loops in surfaces S. In the sequel we will consider immersed 1-manifolds. In anticipation therefore, we make the following definition.

DEFINITION 4.61. An immersed oriented 1-manifold $\Gamma : \coprod_i S^1 \to S$ *bounds* a positive immersion $f : T \to S$ if there is a commutative diagram

$$\begin{array}{ccc} \partial T & \xrightarrow{\quad i \quad} & T \\ {\scriptstyle \partial f}\downarrow & & \downarrow{\scriptstyle f} \\ \coprod_i S^1 & \xrightarrow[\quad \Gamma \quad]{} & S \end{array}$$

for which $\partial f : \partial T \to \coprod_i S^1$ is an orientation-preserving homeomorphism. The 1-manifold Γ *virtually bounds* (or *rationally bounds*) a positive immersion as above if there is a positive integer n so that $\partial f : \partial T \to \coprod_i S^1$ is an orientation-preserving covering satisfying $\partial f_*[\partial T] = n[\coprod_i S^1]$ in homology.

The property of virtually bounding an immersed surface can be detected by stable commutator length:

LEMMA 4.62. *Let $a \in \pi_1(S)$ be represented by a geodesic $\gamma \subset S$. Suppose γ virtually bounds a positive immersed surface T. Then T is extremal, and*

$$\text{scl}(a) = \text{area}(\gamma)/4\pi = -\chi(T)/2n$$

Conversely, if γ does not virtually bound a positive immersed surface, then $\text{scl}(a) > \text{area}(\gamma)/4\pi$.

PROOF. Under the hypotheses of the Lemma, $n\text{area}(\gamma) = \text{area}(T)$. If γ virtually bounds a positive immersed surface T, then $\text{scl}(a) \leq -\chi(T)/2n$. This gives an upper bound on scl which is equal to the lower bound in Lemma 4.59.

Conversely, let T be extremal for a (such a T exists by Theorem 4.24). If T is not homotopic to an immersion, then a pleated representative of T maps at least one ideal triangle with degree -1 and therefore $\text{scl}(a) = -\chi^-(T)/2n > \text{area}(\gamma)/4\pi$. □

REMARK 4.63. By changing the orientation on γ, one sees that γ virtually bounds a negative immersed surface if and only if $\text{scl}(a) = -\text{area}(\gamma)/4\pi$.

REMARK 4.64. We will see in Example 4.72 that there are examples of curves γ which do not bound an immersed surface, but have finite (disconnected) covers which *do* bound immersed surfaces.

REMARK 4.65. One direction of Lemma 4.62 is easy: an immersed surface is evidently extremal, by Bavard duality. The other direction of the proof really uses the existence of extremal surfaces, and therefore depends on Theorem 4.24.

COROLLARY 4.66. *Let $a \in \pi_1(S)$ be represented by a geodesic γ. Suppose a finite cover of γ bounds a (positive or negative) immersed surface in S. Then* $\mathrm{scl}(a) \in \frac{1}{2}\mathbb{Z}$.

PROOF. By Lemma 4.62, there is an equality $\mathrm{scl}(a) = |\mathrm{area}(\gamma)|/4\pi$. On the other hand, by Lemma 4.60, $\mathrm{area}(\gamma) \in 2\pi\mathbb{Z}$. □

REMARK 4.67. Although $\mathrm{area}(\gamma)$ does not depend on the hyperbolic metric g, the quantity $\mathrm{area}^+(\gamma(a, g))$ might. By Gauss–Bonnet, the area of a hyperbolic polygon P is

$$\mathrm{area}(P) = \pi(n-2) - \sum_i \alpha_i$$

where n is the number of vertices, and the α_i are the internal angles. Summing contributions of this kind, we see that $\mathrm{area}^+(\gamma(a, g))$ is an integral linear combination

$$\mathrm{area}^+(\gamma(a, g)) = \sum_p n(p, g)\alpha(p, g) + \text{topological term}$$

where the topological term is in $\pi\mathbb{Z}$, where the sum is taken over points p at which γ crosses itself, where $\alpha(p, g)$ is the angle γ makes with itself at p, and where each $n(p, q)$ is an integer. The $n(p, q)$ are not constant, since they might change sign under a deformation in which some (necessarily simply-connected) region becomes degenerate and changes orientation.

It would be interesting to study $\mathrm{area}^+(\gamma(a, \cdot))$ for each $a \in \pi_1(S)$ as a function on Teichmüller space, and to characterize its range algebraically.

4.2.4. Area and rotation number. We give a reinterpretation of $\mathrm{area}(\gamma)$ in terms of *rotation numbers* which gives another explanation of the quantization of area proved in Lemma 4.60.

A hyperbolic structure and an orientation on S determines a representation $\rho : \pi_1(S) \to \mathrm{PSL}(2, \mathbb{R})$ which is unique up to conjugacy. There is a universal central extension

$$0 \to \mathbb{Z} \to \widetilde{\mathrm{SL}}(2, \mathbb{R}) \to \mathrm{PSL}(2, \mathbb{R}) \to 0$$

with extension class $[e] \in H^2(\mathrm{PSL}(2, \mathbb{R}); \mathbb{Z})$.

If G is any group, and $\rho : G \to \mathrm{PSL}(2, \mathbb{R})$ is a representation, $[e]$ pulls back by ρ^* to define an element $\rho^*([e])$ of $H^2(G; \mathbb{Z})$. If ρ is understood, we abbreviate this by $[e]$ where no confusion can arise. There is an elegant description of e at the level of chains, due to Thurston [**197**]. The group $\mathrm{PSL}(2, \mathbb{R})$ acts on S^1_∞ by orientation-preserving homeomorphisms. Let $p \in S^1_\infty$ be arbitrary. If $g_1, g_2 \in G$ then define

$$e(g_1, g_2) = \begin{cases} \frac{1}{2} & \text{if } p, g_1(p), g_2(g) \text{ is positively ordered} \\ -\frac{1}{2} & \text{if } p, g_1(p), g_2(g) \text{ is negatively ordered} \\ 0 & \text{if } p, g_1(p), g_2(g) \text{ is degenerate} \end{cases}$$

More geometrically, e is $\frac{1}{2\pi}$ times the (signed) hyperbolic area of the ideal triangle spanned by $p, g_1(p), g_2(p)$. Note that e is a bounded 2-cocycle, with norm $1/2$. If $f : (S', \partial S') \to (S, \gamma)$ is a pleated surface with $n(S') = 1$, then $f_*(\partial S')$ fixes points in S^1_∞, and therefore there is a well-defined relative cocycle f^*e whose evaluation $f^*e([S'])$ is $\frac{1}{2\pi}$ times the signed sum of areas of the ideal triangles of S'; i.e. $f^*e([S']) = \mathrm{area}(\gamma)/2\pi$.

If $\rho^* e$ is trivial in $H^2(G;\mathbb{Z})$ then ρ lifts to $\tilde{\rho} : G \to \widetilde{\mathrm{SL}}(2,\mathbb{R})$. As in § 2.3.3 there is a well-defined homogeneous quasimorphism rot on G determined by the choice of a lift $\tilde{\rho}$. Different lifts are parameterized by choices of $H^1(G)$. In particular, $\tilde{\rho}$ is well-defined on $[G,G]$. As bounded cohomology classes, $-[\delta \text{ rot}] = [e]$ in $H^2_b(G;\mathbb{R})$. Here the minus sign appears because of the negative curvature of a hyperbolic surface. In fact, for any closed hyperbolic surface T, there is an equality $e([T]) = -\chi(T)$.

LEMMA 4.68. *With definitions as above, for each a in the commutator subgroup there is an equality*

$$\mathrm{area}(\gamma(a)) = -2\pi \,\mathrm{rot}(a)$$

PROOF. Let $f : (S', \partial S') \to (S, \gamma)$ be a pleated surface with $n(S') = 1$. Then

$$\mathrm{area}(\gamma(a))/2\pi = e(f_*[S']) = -(\delta \mathrm{rot})(f_*[S']) = -\mathrm{rot}(f_*[\partial S']) = -\mathrm{rot}(\gamma)$$

□

Since S is a complete hyperbolic surface, every element is either hyperbolic or parabolic, and therefore has a fixed point in S^1_∞. This implies that rot takes on only integral values. This explains the quantization observed earlier.

REMARK 4.69. Lemma 2.58 says that for any homogeneous quasimorphism ϕ, there is an inequality $D(\phi) \leq 2\|[\delta\phi]\|_\infty$. The discussion above shows that this inequality is an equality when ϕ is the rotation quasimorphism associated to a hyperbolic structure on a noncompact surface.

In fact, for any group G and any representation $\rho : G \to \mathrm{Homeo}^+(S^1)$, we can pull back the Euler class to obtain $[e_\rho] \in H^2_b(G;\mathbb{R})$. After passing to a central extension if necessary, we can assume $[e_\rho]$ is trivial in ordinary H^2, and obtain a rotation quasimorphism rot_ρ with $[\delta \mathrm{rot}_\rho] = [e_\rho]$.

PROPOSITION 4.70. *With notation as above, there is an equality $D(\mathrm{rot}_\rho) = 2\|[e_\rho]\|_\infty$.*

PROOF. We give the sketch of a proof. If G has a finite orbit, then it preserves an invariant probability measure concentrated on this orbit, and therefore rot_ρ is a homomorphism, and $[e_\rho]$ is trivial in $H^2_b(G;\mathbb{R})$. Otherwise, the action is semi-conjugate to a minimal action (i.e. one in which every orbit is dense). A minimal action is either conjugate to an action by rotations (in which case rot_ρ is a homomorphism) or has a finite cyclic centralizer. Quotienting S^1 by the action of the centralizer produces a new minimal action, and multiplies both $[e_\rho]$ and rot_ρ by the same number.

So assume the action is minimal with trivial centralizer. The Milnor–Wood inequality gives $\|[e_\rho]\|_\infty \leq 1/2$ for any action. On the other hand, such an action has the following *compressibility* property: for any closed interval $I \subset S^1$ and any nonempty open set $U \subset S^1$, there is $g \in G$ for which $g(I) \subset U$; a proof of this fact (and the nontrivial assertions in the previous paragraph) follows from Thurston [**197**], Theorem 2.7. Choose disjoint nonempty connected open sets U_1, U_2, V_1, V_2 for which a pair of points in U_1 and V_1 link a pair of points in U_2 and V_2. Let g take $S^1 - V_1$ into U_1, and let h take $S^1 - V_2$ into U_2. Then the action of $\langle g, h\rangle$ is semi-conjugate to an action arising from a hyperbolic structure on a once-punctured torus. Consequently $\mathrm{rot}_\rho([g,h]) = 1$ and therefore $D(\mathrm{rot}_\rho) \geq 1$.

Hence

$$1 \geq 2\|[e_\rho]\|_\infty \geq D(\mathrm{rot}_\rho) \geq 1$$

and the Proposition is proved. □

Note that the method of proof shows that any group acting on a circle either preserves a probability measure, or contains a nonabelian free subgroup. In the literature this fact

is frequently attributed to Margulis [**145**], who seems not to have been aware of the work of Thurston and others.

Lemma 4.62 and Lemma 4.68 taken together show that an element a in the commutator subgroup of $\pi_1(S)$ is represented by a geodesic which virtually bounds an immersed surface in S if and only if rot is an extremal quasimorphism for a. It is convenient to extend this observation to rational chains in B_1^H.

Let $F = \pi_1(S)$, and let $C = \sum t_i a_i$ be a chain in $B_1^H(F)$. Each a_i is represented by a geodesic γ_i in S, so the chain C is represented by a "weighted" union Γ of geodesics in S. The support of Γ decomposes S into regions R_i. For each region R_i, choose an arc α_i from ∂S to R_i, and look at the (weighted) algebraic intersection $\alpha_i \cap \Gamma$. The condition that C is homologically trivial implies that this algebraic intersection number is independent of the choices involved. In the special case that C consists of a single element a, this intersection number is equal to the weight w_i as defined in Definition 4.57. Then define

$$\text{area}(\Gamma) = \sum_i (\alpha_i \cap \Gamma)\text{area}(R_i)$$

Then one has the analogue of Lemma 4.68, namely $\text{area}(\Gamma) = -2\pi \sum t_i \text{rot}(a_i)$. If the coefficients of C are rational, then after multiplying through by a large integer we can assume that the coefficients are integers, and we can think of Γ as a signed sum of simple geodesics. Lemma 4.62 holds for such Γ, and with the same proof; i.e. a weighted union of geodesics Γ representing a chain C virtually bounds a positive immersed surface if and only if $4\pi\,\text{area}(\Gamma) = \text{scl}(C)$. Putting these two facts together gives the following proposition:

PROPOSITION 4.71. *Let S be an oriented surface with boundary. Let C be a rational chain in $B_1^H(F)$ represented by a weighted sum of geodesics Γ. Then Γ virtually bounds a (positive or negative) immersed surface in S if and only if* rot_S *is an extremal quasimorphism for C; i.e. if and only if* $\text{scl}(C) = |\text{rot}_S(C)|/2$.

EXAMPLE 4.72. "Virtually bounds" in Proposition 4.71 cannot in general be improved to "bounds". Consider the immersed curve $\gamma \subset S$ in Figure 4.7, where S is a once-punctured surface of genus 2. The curve γ can be realized by a geodesic in any hyperbolic structure on S.

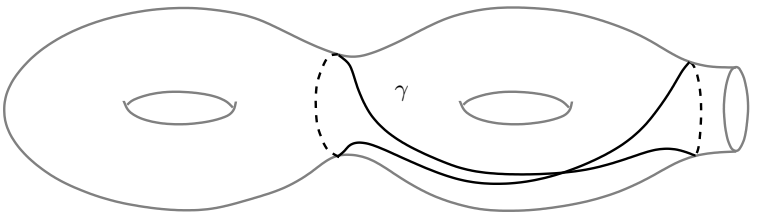


FIGURE 4.7. The loop γ does not bound an immersed surface, but two copies of γ do

The disconnected cover consisting of two copies of γ bounds an immersed surface of genus 4 with two boundary components, which each wrap once around γ. By Lemma 4.62 there is an equality $\text{scl}([a_1, b_1]^2[a_2, b_2]) = 2$ in F_4 (note that this also follows as a special case of the (free) product formula, i.e. Theorem 2.93). Since

the value of scl is not of the form $1/2 + \text{integer}$, γ does not bound an immersed surface.

4.2.5. Rotation number and counting quasimorphisms. In this section, let $S_{1,1}$ denote a once-punctured torus, so that $\pi_1(S_{1,1}) = F_2$, with standard generators a, b. The function $\text{rot}_{1,1} : F_2 \to \mathbb{Z}$ is defined as above, with respect to some complete hyperbolic structure on $S_{1,1}$, and some choice of lift on the generators. Since different lifts agree on the commutator subgroup, the function $\text{rot}_{1,1}$ is well-defined in Q/H^1. One way to fix a lift is to insist that the lifts of a and b fix points, and therefore satisfy $\text{rot}_{1,1}(a) = \text{rot}_{1,1}(b) = 0$. We follow this convention in the sequel.

It turns out that we can give a simple formula for $\text{rot}_{1,1}$ in terms of the Brooks counting quasimorphisms (see § 2.3.2). Recall that for each string σ, the function H_σ counts the number of copies of σ minus the number of copies of σ^{-1}, and $\overline{H}_\sigma$ denotes its homogenization.

REMARK 4.73. In fact, in this section we only consider strings σ of length 2 with distinct letters. For such strings, the "little" and the "big" counting functions and their associated quasimorphisms h_σ and H_σ are equal.

LEMMA 4.74.

$$\text{rot}_{1,1} = \frac{1}{4}\left(\overline{H}_{ab} + \overline{H}_{ba^{-1}} + \overline{H}_{a^{-1}b^{-1}} + \overline{H}_{b^{-1}a}\right)$$

PROOF. The proof is a modification of Klein's ping-pong argument, lifted from the circle to the line. The disk D can be decomposed into 5 regions, one of which, P, is an ideal square which is a fundamental domain for F_2, and the other 4 are neighborhoods of the attracting fixed points of the elements a, b, a^{-1}, b^{-1} respectively. Call these neighborhoods N_a, N_b, N_A, N_B. Given a reduced word $\sigma \in F_2$, and a point $p \in P$, the image $\sigma(p) \in N_w$ where w is the last letter of σ. We can glue $\mathbb{Z}$ copies of each of the regions N_a etc. onto $\mathbb{R}$ in such a way that the union of $\mathbb{R}$ with these regions is the universal cover of $D - P$. Denote this union by E. See Figure 4.8.

These lifted neighborhood regions break up $\mathbb{R}$ into "units", with four units to each lift of a fundamental domain for S^1. We can lift the itinerary of p (except for p itself) under the subwords of σ to an itinerary in E. One sees that every time the letter b appears in σ, the itinerary moves up one unit if the preceding letter was a, and down one unit if the preceding letter was a^{-1}, and similarly for other allowable 2-letter combinations. The rotation number is $1/4$ the number of units, proving the formula. □

This has a particularly simple and interesting interpretation in terms of the graphical calculus introduced in § 2.2.4. A cyclically reduced element in $[F_2, F_2]$ determines a loop in the square lattice without backtracking. Such a loop may be "smoothed" at the corners to determine an immersed curve in the plane. Every anticlockwise turn contributes $1/4$ to $\text{rot}_{1,1}$, whereas every clockwise turn contributes $-1/4$. Hence $\text{rot}_{1,1}$ is just the *winding number* of the immersed curve associated to an element.

The following corollary illustrates the power of this technique.

COROLLARY 4.75. *Let $g \in F_2$ be a commutator, and let γ_g be the geodesic representative of the conjugacy class of g in T, a hyperbolic once-punctured torus.*

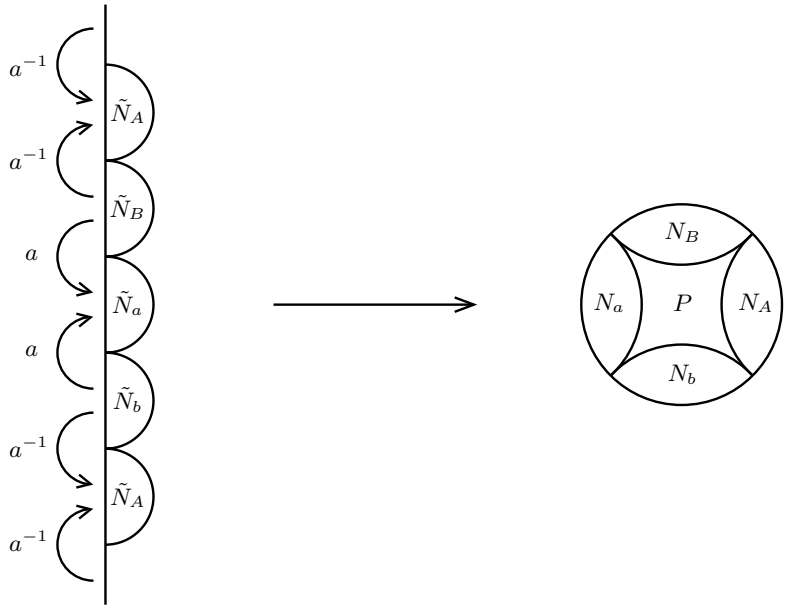


FIGURE 4.8. a moves points in $\tilde{N}_b$ regions up one unit, and a^{-1} moves such points down one unit. Furthermore, a moves points in $\tilde{N}_B$ regions down one unit, and a^{-1} moves such points up one unit. A similar relation holds with a and b interchanged.

Let w_g be the loop in the square lattice in $\mathbb{R}^2$ corresponding to the (cyclically) reduced representative of g. Then γ_g bounds an immersed surface in T if and only if the winding number of w_g is ± 1.

PROOF. Since g is a commutator, there is a map $f : T \to T$ taking the boundary to γ_g. Replace f by a pleated representative. The (algebraic) area of $f(T)$ is $-2\pi \text{ rot}_{1,1}(g) = -2\pi \text{ wind}(w_g)$, so if the winding number is ± 1, this pleated representative is an immersion. Conversely, if $\text{wind}(w_g) = 0$, the algebraic area is zero, so no map f as above can be an immersion. □

A similar argument lets one give a formula for rotation numbers associated to a hyperbolic structure on any noncompact hyperbolic surface in terms of Brooks functions on the associated (free) fundamental group. As before, let P be a fundamental domain for the surface, so that $D - P$ decomposes into regions on which the generators do ping-pong. Then each allowable pair xy of distinct letters in a reduced word moves up some fixed number n_{xy} of units. If $S_{g,p}$ is the surface of genus g with p punctures, then $\pi_1(S_{g,p})$ is free of rank $2g+p-1$ and we can take as generators $a_1, b_1, \cdots, a_g, b_g, c_1, \cdots, c_{p-1}$. We thereby obtain the following theorem.

THEOREM 4.76 (Rotation number formula). *Let $\mathcal{C}$ denote the following cyclically ordered set:*

$$\mathcal{C} = (a_1, b_1, a_1^{-1}, b_1^{-1}, \cdots, a_g, b_g, a_g^{-1}, b_g^{-1}, c_1, c_1^{-1}, \cdots, c_{p-1}, c_{p-1}^{-1})$$

For each pair x, y in $\mathcal{C}$ with $x \neq y, y^{-1}$ let m_{xy} be the integer $0 < m_{xy} < 4g+2p-2$ such that y is m_{xy} elements to the right of x in $\mathcal{C}$. Define

$$n_{xy} = \begin{cases} m_{xy} \text{ if } (y^{-1}, x, y) \text{ is positive in the circular order} \\ m_{xy} - (4g+2p-2) \text{ otherwise} \end{cases}$$

Then there is an equality

$$\mathrm{rot}_{g,p} = \frac{1}{4g+2p-2} \left(\sum_{x \neq y \text{ or } y^{-1}} n_{xy} \overline{C}_{xy} \right)$$

where for each string σ, we let C_σ denote the counting function that counts copies of σ, and $\overline{C}_\sigma$ denotes its homogenization.

For example, let $S_{0,3}$ be the thrice punctured sphere, and let $\pi_1(S_{0,3}) = \langle a, b \rangle$ where a and b are loops around the punctures. Then if $\mathrm{rot}_{0,3}$ denotes the homogeneous quasimorphism associated to the hyperbolic structure, there is a formula

$$\mathrm{rot}_{0,3} = \frac{1}{2} \left(\overline{H}_{a^{-1}b} + \overline{H}_{ba^{-1}} \right)$$

Remark 4.77. In fact, the formula from Theorem 4.76 gives (after collecting terms)

$$\mathrm{rot}_{0,3} = \frac{1}{4} \left(2\overline{H}_{a^{-1}b} + 2\overline{H}_{ba^{-1}} + \overline{C}_{ab} + \overline{C}_{b^{-1}a^{-1}} - \overline{C}_{a^{-1}b^{-1}} - \overline{C}_{ba} \right)$$

However, the function $C_{ab} + C_{b^{-1}a^{-1}} - C_{a^{-1}b^{-1}} - C_{ba}$ is uniformly bounded on any reduced word, as can be verified by a calculation, and therefore its homogenization is trivial.

Theorem 4.76 gives similar necessary and sufficient criteria in terms of counting quasimorphisms for geodesics in hyperbolic surfaces S corresponding to commutators in $\pi_1(S)$ to bound an immersed surface.

4.2.6. Rigidity Theorem. The content in the next few sections is taken largely from [**45**]. The main goal is to prove the following theorem:

Theorem 4.78 (Rigidity Theorem). *Let $F = \pi_1(S)$ where S is a compact oriented surface with $\chi(S) < 0$ and nonempty boundary.*

1. *The projective class of the chain ∂S in $B_1^H(F)$ intersects the interior of a codimension one face π_S of the unit ball in the* scl *norm.*
2. *The unique element of $Q(F)/H^1$ dual to π_S (up to scale) is the rotation quasimorphism associated to the action of $\pi_1(S)$ on the ideal boundary of the hyperbolic plane, coming from a hyperbolic structure on S.*

Theorem 4.78 reveals how surface topology and hyperbolic geometry are manifested in the bounded cohomology of a free group.

The proof is entirely elementary modulo Proposition 4.71, and depends only on constructing immersed surfaces in S with prescribed boundary. Technically, the result we prove is the following:

Theorem 4.79 (Immersion Theorem). *Let S be a compact oriented hyperbolic surface with (possibly empty) geodesic boundary. Let C be a homologically trivial rational chain, represented by a weighted union Γ of geodesics. Then for all sufficiently large N (depending on Γ), the chain $\Gamma + N\partial S$ virtually bounds a (positive) immersed surface.*

We show how to deduce Theorem 4.78 from Theorem 4.79.

PROOF. Let C be any rational chain in $B_1^H(F)$. By Theorem 4.79 and Proposition 4.71, for all sufficiently large N the chain $C + N\partial S$ in $B_1^H(F)$ satisfies

$$\mathrm{scl}(C + N\partial S) = \mathrm{rot}_S(C + N\partial S)/2$$

Hence the ray through ∂S intersects the interior of an edge of the unit ball of the scl norm restricted to the subspace $\langle C, \partial S\rangle$. Since C was arbitrary, the projective class of ∂S intersects the interior of a codimension one face π_S of the unit ball in the scl norm. By construction, this face is dual to rot_S (up to scale and H^1). □

REMARK 4.80. Proposition 4.71 says that a rational chain C virtually bounds a positive immersed surface in S if and only if $\mathrm{scl}(C) = \mathrm{rot}(C)/2$. By Theorem 4.78, this holds if and only if the projective class of C intersects the face π_S. If the support of C does not include ∂S, then $C - \epsilon\partial S$ cannot virtually bound a positive immersed surface in S for any positive ϵ. Consequently the projective class of such a C does not intersect the *interior* of π_S, but only its *boundary*.

REMARK 4.81. One still has a version of the Rigidity Theorem for closed surfaces. Let S be a closed, oriented hyperbolic surface. The hyperbolic structure lets us think of $\pi_1(S)$ as a subgroup of $\mathrm{PSL}(2,\mathbb{R})$. Denote by G the preimage of this subgroup in $\widetilde{\mathrm{SL}}(2,\mathbb{R})$. The group G is isomorphic to the fundamental group of the unit tangent bundle of S. There is a nontrivial central extension

$$\mathbb{Z} \to G \to \pi_1(S)$$

associated to the class of the generator of $H^2(S;\mathbb{Z})$. Let rot_Z denote the pullback of the rotation quasimorphism on $\widetilde{\mathrm{SL}}(2,\mathbb{R})$ to G, and let Z denote the generator of the center of G. Theorem 4.79 and some elementary homological algebra implies that for any element $g \in [G,G]$, the quasimorphism rot_Z is extremal for $g + nZ$ whenever n is sufficiently large. Hence there is a codimension one face π_Z of the unit ball of the scl norm on $B_1^H(G)$, and the projective class of Z intersects the interior of this face.

By continuity, for any $g \in [G,G]$, the projective class of $g + nZ$ also intersects the interior of π_Z whenever n is sufficiently large (depending on g). Since Z is central, $\mathrm{scl}(g + nZ + C) = \mathrm{scl}(Z^n g + C)$ for any g and any chain C. Consequently, the projective class of the element $Z^n g$ also intersects the interior of π_Z whenever n is sufficiently large. Dually, rot_Z is the unique extremal homogeneous quasimorphism for $Z^n g$, up to scale and elements of H^1.

4.2.7. Proof of the immersion theorem. In this section we fix a surface S with $\pi_1(S) = F$ and a chain $C \in B_1^H(F)$ represented by a weighted sum of geodesics $\Gamma(C)$. Where there is no confusion, we abbreviate $\Gamma(C)$ to Γ. By LERF for surface groups (see Example 2.108) we can pass to a finite cover in which each component of the preimage of Γ is embedded (though of course the union will typically not be embedded). Let Γ' be the total (weighted) preimage of Γ in the cover S'. If Γ' cobounds a positively immersed surface with some multiple of $\partial S'$, this immersed surface projects to S and shows that the same is true of Γ. So without loss of generality, we can assume that every component of Γ is embedded.

If Γ_1 and Γ_2 virtually bound positive immersed surfaces, the same is true of $\Gamma_1 + \Gamma_2$, by Proposition 4.71 and the linearity of rot_S on B_1^H. The only homologically trivial chains in B_1^H represented by weighted sums of geodesics supported in ∂S are the multiples of ∂S, so to prove the theorem, it suffices to find any weighted collection of geodesics ∂ with support in ∂S so that $\Gamma + \partial$ virtually bounds a positive immersed surface. By abuse of notation, we say that $\Gamma + \partial$ virtually bounds a positive immersed surface if there is some (unspecified) ∂ with this property.

Suppose Γ_1 and Γ_2 are such that $\Gamma_1 - \Gamma_2 + \partial$ virtually bounds a positive immersed surface for some ∂. Let $i : T \to S$ be such an immersed surface. Then (again by LERF for surface groups) there are finite covers T', S' so that $i' : T' \to S'$ is an *embedding*. The difference $S' - i'(T')$ projects to S and shows that $\Gamma_2 - \Gamma_1 + \partial$ also virtually bounds a positive immersed surface (here ∂ typically stands for a different weighted collection of geodesics with support in ∂S). Define a relation $\sim$ on weighted collections of geodesics, where $\Gamma_1 \sim \Gamma_2$ if $\Gamma_1 - \Gamma_2 + \partial$ virtually bounds a positive immersed surface for some ∂ with support in ∂S. By the arguments above, this relation is reflexive, symmetric and transitive, and is consequently an *equivalence relation*. To prove the theorem therefore, we need only show that $\Gamma = \Gamma(C)$ satisfies $\Gamma \sim 0$.

LEMMA 4.82. *Let $S' \subset S$ be a subsurface with geodesic boundary, and let S'' be obtained from S' (topologically) by adding disks to close up some of the boundary components. Suppose that every boundary component of S' is either a boundary component of S, or is separating in S. Suppose further that γ and γ' are simple geodesics in S' that are homotopic in S''. Then $\gamma \sim \gamma'$.*

PROOF. Homotopic simple loops in S'' are isotopic in S''. Such an isotopy can be taken to be a sequence of simple moves which "push" γ over a single boundary component of S'. The result is realized at each stage by an embedded geodesic in S'. Every boundary component ∂_i of S' is either a boundary component of S, or is separating, and in either case $\partial_i \sim 0$. Hence $\gamma \sim \gamma'$ as claimed. □

Let δ be a family of pairwise disjoint essential separating geodesics which decompose S into a union of genus one subsurfaces S_i. There is a graph dual to this decomposition, with one vertex for each component of $S - \delta$, and one edge for each component of δ. Since each δ is separating, this dual graph is a tree. There are several possible such decompositions; for concreteness, choose a decomposition for which this dual graph is an interval. Note that a separating geodesic δ_i necessarily satisfies $\delta_i \sim 0$.

LEMMA 4.83. *Let γ be an embedded geodesic in S, and let δ as above separate S into genus 1 subsurfaces. Suppose γ intersects δ. Then there is an embedded geodesic 1-manifold γ' with at most two components, such that $\gamma \sim \gamma'$, and such that γ' intersects δ in fewer points than γ.*

PROOF. Every component of δ satisfies $\delta_i \sim 0$, so without loss of generality we can assume γ intersects δ transversely. There is at least one component S_i of $S - \delta$ such that γ intersects exactly one boundary component δ_i of S_i. Since δ_i is separating, the algebraic intersection number of γ with δ_i is zero, and therefore γ must intersect δ_i in at least two points with opposite signs. Let α be an arc of δ_i whose interior is disjoint from γ, and whose endpoints intersect γ with opposite signs. Build an embedded thrice punctured sphere in S by thickening γ, and attaching a 1-handle with core α. Isotope the boundary components of this thrice punctured sphere until they are (embedded, disjoint) geodesics. One component is γ; the other two components are γ'. □

By repeatedly applying Lemma 4.83, we can construct Γ' with $\Gamma \sim \Gamma'$, such that each geodesic in Γ' is embedded and contained in a genus one subsurface S' of S satisfying the hypothesis of Lemma 4.82. Let S'' be obtained from S' topologically by filling in all but one boundary component. Fix a standard basis

α, β of embedded geodesics in S' generating the homology of S''. Then γ represents $p\alpha + q\beta$ in homology. Since γ is embedded, p, q are coprime. We would like to show $\gamma \sim p\alpha + q\beta$. By induction, it suffices to show that the chain $a + b + a^{-1}b^{-1} \sim 0$ in a once-punctured torus, or equivalently that the chain $a + b + a^{-1}b^{-1} + [a, b]^n$ virtually bounds a positive immersed surface for some n. This can be proved by an explicit construction.

EXAMPLE 4.84. The chain $a+b+a^{-1}b^{-1}+[a,b]^2$ bounds an immersed surface in a once-punctured torus. One way to see this is to compute, using `scallop` to show $\mathrm{scl}(a+b+a^{-1}b^{-1}+[a,b]^2) = 1$ and then verifying equality in Proposition 4.71, using the formula in Lemma 4.74 for rot. Another way is by explicit construction. There is an immersed four-holed sphere, found by Matthew Day, whose boundary is the chain $a+b+a^{-1}b^{-1}+[a,b]^2$. This surface is depicted in Figure 4.9 (compare with Figure 5 from [**45**]).

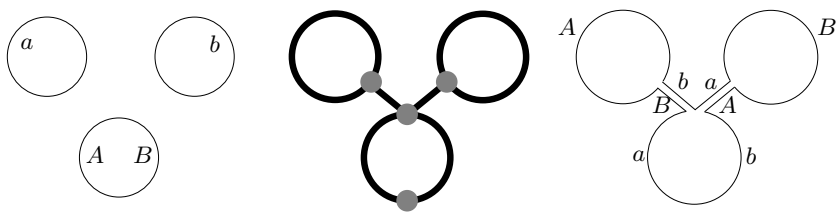


FIGURE 4.9. A 4-holed sphere that immerses in a once-punctured torus, with four boundary components (indicated by thin curves) in the conjugacy classes of a, b, $a^{-1}b^{-1}$ and $[a, b]^2$.

We now explain how to put these pieces together to prove the theorem.

PROOF. Let Γ in S be homologically trivial, with every component embedded. Decompose S along embedded separating geodesics δ as above into genus one subsurfaces. By Lemma 4.83, we can find Γ', a weighted sum of embedded geodesics, such that $\Gamma \sim \Gamma'$, and Γ' is disjoint from δ. For each component S' of $S - \delta$, let $\Gamma'(S')$ be the components of Γ' in S'. For each γ in $\Gamma'(S')$ there are coprime integers $p(\gamma)$ and $q(\gamma)$ so that $\gamma \sim p(\gamma)\alpha + q(\gamma)\beta$. But Γ is homologically trivial, and therefore the same is true of Γ' and $\Gamma'(S')$. Hence $\sum_\gamma p(\gamma) = \sum_\gamma q(\gamma) = 0$ and therefore $\Gamma'(S') \sim 0$. Since S' was arbitrary, $\Gamma' \sim 0$ and therefore $\Gamma \sim 0$. This completes the proof of Theorem 4.79 (and of Theorem 4.78). □

See [**45**] for more details and discussion.

4.2.8. Infinite dimensional faces. Theorem 4.78 can be "bootstrapped" in an interesting way. Let C be a rational chain in $B_1^H(F)$. The chain C is represented by a weighted collection Γ of geodesic loops in S where $\pi_1(S) = F$. By Theorem 4.24, there is an extremal surface T for C, i.e. a π_1-injective map $f : T, \partial T \to S, \Gamma$ realizing the infimum of $-\chi^-(T)/2n(T)$. Now, let C' be an arbitrary chain in $B_1^H(\pi_1(T))$, and Γ' a weighted collection of geodesic loops in T that it represents. By Theorem 4.78, for sufficiently large m the chain $\Gamma' + m\partial T$ virtually bounds an immersed surface. That is, there is an immersion $g : U, \partial U \to T, \Gamma' \cup \partial T$ for which $g(\partial U) = n'(\Gamma' + m\partial T)$ for some n'.

LEMMA 4.85 (Bootstrap Lemma). *The surface $f \circ g : U, \partial U \to S, f(\Gamma') \cup \Gamma$ is extremal for some multiple of the chain $f(C') + mC$ in $B_1^H(F)$.*

PROOF. By LERF, there are finite covers $T' \to T$ and $U' \to U$ so that g lifts to an embedding $g' : U' \to T'$. Clearly, it suffices to show that $f' \circ g'$ is extremal for some multiple of $f(C') + mC$. Since g' is an embedding, we can write T' as a union $T' = g'(U') \cup T''$. If $f' \circ g'$ is not extremal, there is some other $h : V, \partial V \to S, f(\Gamma') \cup \Gamma$ which is extremal for a (possibly different) multiple of $f(C') + mC$, and satisfies $-\chi^-(V)/2n(V) < -\chi(U')/2n(U')$. By the argument of Proposition 2.13, suitable covers of V and T'' can be glued up to produce a surface W which is extremal for Γ but satisfies $-\chi^-(W)/2n(W) < -\chi^-(T)/2n(T)$, contrary to the hypothesis that T is extremal. This contradiction shows that no such surface V exists, and therefore $f \circ g$ is extremal, as claimed. □

The following corollary is immediate:

COROLLARY 4.86. *Let F be a free group, and $C \in B_1^H(F)$ a rational chain. The projective class of C in $B_1^H(F)$ intersects the interior of an infinite dimensional face π_C of the unit ball in the* scl *norm. If $f : \pi_1(T) \to F$ is any extremal surface for C, then $f_*(\pi_T) \to \pi_C$ is isometric, in the sense that $\mathrm{scl}_{\pi_1(T)}(C') = \mathrm{scl}_F(f_*(C'))$ for all chains C' in the cone on $\pi_T \subset B_1^H(\pi_1(T))$.*

PROOF. All that needs to be shown is that π_C is infinite dimensional, and to establish this it suffices to show that the image of $B_1^H(\pi_1(T))$ in $B_1^H(F)$ is infinite dimensional. Since T is extremal, $f_*(\pi_1(T))$ is a nontrivial finitely generated free subgroup of F. By Hall, free groups are virtual retracts (Example 2.107), so one can find infinitely many elements in $f_*(\pi_1(T))$ which are independent in $B_1^H(F)$. □

By convexity of the norm, the face π_C is well-defined. Note that Corollary 4.86 shows that extremal maps are norm-preserving on a nonempty open subset of B_1^H (compare with § 4.1.10).

REMARK 4.87. Lemma 4.85 and Corollary 4.86 can also be deduced using quasimorphisms. Suppose $f : T \to S$ is extremal for some chain C. Let ϕ be an extremal quasimorphism for C with defect 1. Then $f^*\phi$ is an extremal quasimorphism for ∂T with defect 1 because T is extremal. By Theorem 4.78, $f^*\phi$ is equal to rot_T on $B_1^H(\pi_1(T))$. But rot_T is extremal on $B_1^H(\pi_1(T))$. Hence for every $C' \in \pi_T$, we have

$$\mathrm{scl}_F(f_*(C')) \le \mathrm{scl}_{\pi_1(T)}(C') = \mathrm{rot}_T(C')/2 = \phi(f_*(C'))/2 \le \mathrm{scl}_F(f_*(C'))$$

where the first inequality is monotonicity of scl, and the last inequality is Bavard duality.

One might wonder whether every face π_C has finite codimension. In fact, this is not the case. The following example is taken from [**45**].

EXAMPLE 4.88. By Bavard duality, the codimension of π_C is one less than the dimension of the space of extremal quasimorphisms for $C \pmod{H^1}$. Hence to exhibit a rational chain (in fact, an element of $[F, F]$) whose projective class intersects the interior of a face of infinite codimension, it suffices to exhibit a chain that admits an infinite dimensional space of extremal quasimorphisms.

Let $F = F_1 * F_2$ where F_1 and F_2 are both free of rank at least 2, and let $g \in [F_1, F_1]$ be nontrivial. Let $\phi_1 \in Q(F_1)$ be extremal for g, and let $\phi_2 \in Q(F_2)$ be arbitrary with $D(\phi_2) \le D(\phi_1)$. By the Hahn–Banach Theorem, there exists $\phi \in Q(F)$ that agrees with ϕ_i on F_i, and satisfies $D(\phi) = D(\phi_1)$.

EXAMPLE 4.89. Let ρ_t be a continuous family of (nonconjugate) indiscrete representations of F_2 into $\mathrm{PSL}(2,\mathbb{R})$. For a typical family ρ_t, the image $\rho_t(F_2)$ is dense in $\mathrm{PSL}(2,\mathbb{R})$ for all t, and therefore we can find (many) elements $g, h \in F_2$ generating a subgroup Γ so that $\rho_t(\Gamma)$ is discrete and purely hyperbolic, and the axes of $\rho_t(g)$ and $\rho_t(h)$ cross, for all t in some nontrivial interval I. Let rot_t be the homogeneous quasimorphism on F_2 (well-defined up to an element of H^1) associated to the representation ρ_t. Without loss of generality, we can choose rot_t to vary continuously as a function of t on every element of F_2. By construction, rot_t is an extremal quasimorphism for $[g,h]$, for all $t \in I$. On the other hand, for a suitable (indiscrete) family of representations ρ_t, for every nonempty interval I we can find a subinterval J, a point $p \in J$, and an element $f \in F_2$ for which $\mathrm{rot}_t(f)$ is elliptic for all $t \in J$ with $t < p$, and hyperbolic for all $t \in J$ with $t > p$. The quasimorphisms rot_t are constant on f for $t > p$ and nonconstant for $t < p$, so they span an infinite dimensional subspace of $Q(F_2)$. Hence the codimension of the face $\pi_{[g,h]}$ is infinite (compare with Burger–Iozzi [**30**]).

See [**45**] for more corollaries and discussion.

4.2.9. Discreteness of linear representations. Theorem 4.78 has applications to the study of symplectic representations of free and surface groups. For a basic reference to the theory of symplectic groups and representations, see [**31**] (we also return to this subject in more detail in § 5.2.3). We give a new proof of a relative version of rigidity theorems of [**93**] and [**31**], at least in an important special case. Roughly speaking, Goldman observed (in the case of $\mathrm{PSL}(2,\mathbb{R})$) that representations of surface groups of maximal Euler class are *discrete*. Burger–Iozzi–Wienhard extended this observation to symplectic groups, and characterized such representations geometrically.

The context is as follows. Let S be a compact oriented surface with boundary, and let $\rho : \pi_1(S) \to \mathrm{Sp}(2n,\mathbb{R})$ be a symplectic representation for which the conjugacy classes of boundary elements fix a Lagrangian subspace. This condition ensures that there is a well-defined relative Euler class (usually called the Maslov class for $n > 1$) which we denote $e_\rho \in H^2(S, \partial S; \mathbb{Z})$ associated to ρ (compare with § 4.2.4). The cohomology class e_ρ is bounded, with norm $n/2$, and therefore $|e_\rho([S])| \leq -n\chi(S)$. A representation is said to be *maximal* (and e_ρ is *maximal*) if equality is achieved.

The following corollary says that maximal Zariski dense representations are discrete. We restrict to Zariski dense representations for simplicity; this condition is not necessary (see [**93, 31, 32**]).

COROLLARY 4.90 (Goldman, Burger–Iozzi–Wienhard). *Let S be a compact oriented surface with boundary. Let $\rho : \pi_1(S) \to \mathrm{Sp}(2n,\mathbb{R})$ be Zariski dense, and suppose that conjugacy classes of boundary elements fix a Lagrangian subspace. If e_ρ is maximal, ρ is discrete.*

PROOF. For the remainder of the proof, denote $\pi_1(S)$ by F and its commutator subgroup by F'. Since S has boundary, $e_\rho = [\delta\phi]$ where ϕ is in $Q(F)$, and is unique up to elements of H^1. For each $g \in F$, the value $\phi(g) \pmod{\mathbb{Z}}$ is the *symplectic rotation number*, and depends only on the image $\rho(g)$. Since e_ρ is maximal, ϕ is extremal for $\partial S \in B_1^H(F)$. Hence, by Theorem 4.78, it follows that the symplectic rotation number is *zero* on every $g \in F'$; in particular, $\rho(F')$ is not dense in

$\mathrm{Sp}(2n,\mathbb{R})$. Since $\mathrm{Sp}(2n,\mathbb{R})$ is simple, every Zariski dense subgroup is either discrete or dense (in the ordinary sense). If $\rho(F)$ is dense, then the closure of $\rho(F')$ is normal in $\mathrm{Sp}(2n,\mathbb{R})$. But $\mathrm{Sp}(2n,\mathbb{R})$ is simple, and the closure of $\rho(F')$ is a proper subgroup; hence $\rho(F)$ is discrete. □

REMARK 4.91. The condition that boundary element fix Lagrangian subspaces is only included so that the Corollary can be phrased in terms of an integral Euler (Maslov) class. If $\rho : \pi_1(S) \to \mathrm{Sp}(2n,\mathbb{R})$ is *any* Zariski dense representation for which the pullback of the symplectic rotation quasimorphism (i.e. the quasimorphism ϕ above) is extremal for ∂S, then ∂S necessarily fixes a Lagrangian subspace.

4.2.10. Character Varieties. Any representation of a free group $\rho : F \to \mathrm{PSL}(2,\mathbb{R})$ lifts to $\widetilde{\mathrm{SL}}(2,\mathbb{R})$ and defines an associated homogeneous quasimorphism $\mathrm{rot} : F \to \mathbb{R}$ unique up to a homeomorphism. Given $a \in [F,F]$ one can ask what values this function can take as ρ varies over all homomorphisms.

We restrict attention to the case that F is free of rank 2, generated by elements a, b. The function rot only depends on the conjugacy class of ρ, and therefore we consider representations up to conjugacy. In fact, since ρ can typically be recovered just from the traces of elements, it makes sense to consider the *character variety*, consisting of the set of functions on F which are traces of some representation. For simplicity, it makes sense to study the $\mathrm{SL}(2,\mathbb{R})$ character variety instead, since traces are well defined there.

DEFINITION 4.92. Let G be a finitely generated group. The *character variety* of G, denoted $X(G)$, is the set of functions $\chi : G \to \mathbb{R}$ for which $\chi = \mathrm{tr}(\rho)$ for some representation $\rho : G \to \mathrm{SL}(2,\mathbb{R})$.

Characters with representations in a fixed algebraic group satisfy many non-trivial (polynomial) relations, and a character is determined by its values on finitely many elements. This gives $X(G)$ the structure of a (real) algebraic variety. See [**60**] for an introduction to SL character varieties, and their applications to 3-manifolds.

EXAMPLE 4.93. Let $G = F_2$, the free group on generators a, b. Since $\mathrm{SL}(2,\mathbb{R})$ is 3-dimensional, the space of $\mathrm{SL}(2,\mathbb{R})$ representations of F_2 is 6 dimensional, and the space of characters is 3-dimensional. If χ is a character, the co-ordinates $(x,y,z) = (\chi(a), \chi(b), \chi(ab))$ defines a map from $X(F_2)$ to $\mathbb{R}^3$. In fact, this map is an isomorphism onto the subset of $\mathbb{R}^3$ consisting of the union of the complement of the open cube $(-2,2)^3$ together with the subset of triples inside the cube satisfying

$$x^2 + y^2 + z^2 - xyz \geq 4$$

THEOREM 4.94. *Let $g \in [F_2, F_2]$. Then the set of values of* $\mathrm{rot}(g)$ *as one varies over all* $\mathrm{SL}(2,\mathbb{R})$ *representations of F_2 is a closed, connected interval, whose endpoints have the property that their image under* $\cos(2\pi\cdot)$ *is algebraic.*

PROOF. Example 4.93 shows how to identify $X(F_2)$ with a semi-algebraic subset of $\mathbb{R}^3$. For every $g \in F_2$, the value of $\chi(g)$ is an integral polynomial in the values of $\chi(a), \chi(b), \chi(ab)$.

The function $\chi(g) : X(G) \to \mathbb{R}$ is therefore an integral polynomial on $\mathbb{R}^3$. An extremal value is a zero of a system of integral polynomial equations, and is therefore realized at an algebraic point. Since $2\cos(2\pi\ \mathrm{rot}(g)) = \chi(g)$, the result follows. □

REMARK 4.95. A similar theorem can be proved with a similar proof with $\mathrm{Sp}(2n, \mathbb{R})$ or $\mathrm{SO}_0(n, 2)$ in place of $\mathrm{SL}(2, \mathbb{R})$

4.3. Diagrams and small cancellation theory

The proof of Proposition 4.36 shows that in a fixed free group, every extremal surface can be built up from pieces (polygons and rectangles) of bounded complexity. A representation of a surface (with prescribed boundary) as a union of simple pieces drawn from some finite set is sometimes called a *diagram*. Diagrams can be represented graphically, and can be combined, composed and manipulated according to certain sets of rules. They have psychological value, as a way to represent algebraic information in geometric terms (e.g. as in Figure 4.9); and computational value. There are many different conventions for diagrams, depending on function and context.

EXAMPLE 4.96. The conjugacy class $w = [a^2, b^2][a, b]$ has scl $= 1$ in F_2. Let S be a hyperbolic once-punctured torus with basis a, b, and let γ be the geodesic associated to w. Then two copies of γ bound an immersed genus 2 surface T with two boundary components. Figure 4.10 depicts the surface T as a diagram,

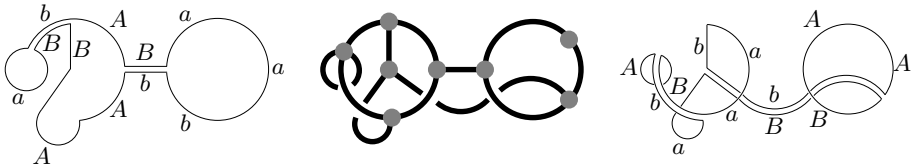


FIGURE 4.10. The surface T is obtained by thickening a graph with a cyclic ordering at the vertices. Edges of ∂T on opposite sides of each edge of the underlying graph are labeled by inverse elements of F_2. Each boundary component of ∂T is labeled by a cyclic conjugate of w.

obtained by thickening a graph whose vertices correspond to polygons, and edges to rectangles. The two copies of γ are indicated by thinner lines.

REMARK 4.97. Any extremal surface obtained from the proof of Theorem 4.24 retracts in an obvious way to a graph with one edge for each rectangle, and one vertex for each polygon. To recover the surface (and therefore its boundary) from the graph, we need to specify a cyclic ordering of the edges at each vertex. A graph together with the choice of a cyclic ordering on the edges at each vertex is sometimes called a *ribbon graph* or a *fat graph*. Such objects appear in the study of dynamical systems, Hopf algebras, statistical mechanics, combinatorics, and many other fields; see [**17**].

4.3.1. Diagrams. Diagrams (sometimes called *van Kampen diagrams*) were introduced by van Kampen in [**200**].

Let G be a group given by a presentation $G = \langle X \mid R \rangle$. Let F be the free group on X, and N the normal closure of R in X, so that $G = F/N$. The set R is said to have been *symmetrized* if all elements are cyclically reduced, and R is closed under taking cyclic permutations and inverses.

DEFINITION 4.98. Let $w \in F$ be cyclically reduced. A *diagram* is a finite connected planar graph in which directed edges are labeled by elements of F, the boundary of each interior region is labeled by an element of R, and the boundary of the exterior region is labeled by a cyclic conjugate of w.

Since the graph associated to a diagram is assumed to be connected, interior regions are all homeomorphic to open disks. The boundary of a region is allowed to bump up against itself.

Note that the boundary label of a region depends on a choice of basepoint and a choice of orientation, or else the result differs by cyclic permutation or inverse. However, since R is symmetrized, membership in R is not affected by this ambiguity.

REMARK 4.99. A finite connected planar graph together with the regions it bounds is a simply-connected planar 2-complex. By abuse of notation we sometimes think of this 2-complex as the diagram.

If we assume elements of R are cyclically reduced, a map has no 1-valent vertices. Furthermore, if e_1, e_2 share a 2-valent vertex in common, we can replace $e_1 \cup e_2$ by $e_1 e_2$. Therefore in the sequel we assume every vertex is at least 3-valent.

LEMMA 4.100. *An element $w \in F$ admits a diagram if and only if it is in N.*

PROOF. There is a tautological cellular map from a diagram (thought of as a 2-complex) to a 2-complex associated to the presentation of G. Since the underlying 2-complex of a diagram is simply-connected, the boundary of the exterior region maps to a homotopically trivial loop. This exhibits w as an element of N.

Conversely, express w as a product of conjugates of elements of R. Denote this expression by a bunch of balloons in the plane tied by strings to a common basepoint, where each balloon is an element of R, and the string is the conjugating element. Then cancel adjacent edges whenever possible. The result is a finite connected planar graph whose boundary is a cyclically reduced word which is equal in F to w (after choosing a suitable basepoint and orientation), and therefore must be equal to w by uniqueness of reduced representatives in free groups. □

DEFINITION 4.101. A diagram is *reduced* if no two adjacent regions have boundaries which represent inverse elements of R, where the basepoint is taken to be some common vertex, and the orientations on the boundaries disagree (when compared with some orientation inherited from the plane).

Any diagram may be replaced by a reduced one, by collapsing nonreduced pairs of adjacent regions, thereby reducing the number of regions in the diagram until the process terminates.

DEFINITION 4.102. A word $b \in F$ is called a *piece* (relative to R) if there are distinct relations $ba_1, ba_2 \in R$.

An edge of a diagram between adjacent regions is a piece.

4.3.2. Small cancellation theory. In full generality, the theory of van Kampen diagrams is essentially combinatorial. However, when applied to groups with presentations that obey certain conditions (of a geometric nature), it makes contact with the theory of hyperbolic groups, negative curvature, regular languages, and so on. The geometric theory of diagrams arising from groups with presentations satisfying such conditions is called *small cancellation theory*.

Small cancellation theory has its origins in the work of Dehn [**63**], in which he posed the word and conjugacy problems for finitely presented groups, and solved these problems for fundamental groups of closed orientable 2-manifolds.

Dehn's insight was that surface groups have presentations with a single relator r with the property that for any cyclic conjugate s of r or r^{-1} with $s \neq r^{-1}$, the

product sr has very little cancellation. Thus if a word in a surface group is trivial, it can be simplified immediately by finding a big subword consisting of more than half of some s.

It was not until the work of Lyndon [**138**] and Weinbaum [**203**] that the importance of geometry in Dehn's work was properly appreciated, and small cancellation theory began to be systematically applied to combinatorial group theory.

The hypotheses of small cancellation theory are conditions which a given symmetrized presentation might satisfy. Some of these conditions are as follows:

- $C'(\lambda)$: every piece has length less than λ times the length of a relation it appears in.
- $C(p)$: no relation is a product of fewer than p pieces. Equivalently, every region in a reduced diagram with no edges in common with the exterior region has at least p sides.
- $T(q)$: any interior vertex in a reduced diagram has at least q incident edges.

Note that $C'(\lambda)$ implies $C(p)$ for $\lambda p < 1$.

Let D be a reduced diagram for an element $w \in G$. We can make D into a metric space by choosing a polygonal structure on each region and gluing these polygons together. The small cancellation conditions and the Gauss–Bonnet Theorem give upper bounds on the (distributional) curvature in D for a suitable choice of structure.

EXAMPLE 4.103. Condition $C(6)$ implies that every polygon has at least 6 sides. Choose a metric for which each region is a constant curvature regular polygon with side lengths 1 and all angles $2\pi/3$. If a region has 6 sides, it will be a Euclidean hexagon with this metric. If it has more than 6 sides, it will be hyperbolic. At every 3-valent vertex these polygons fit together. At every vertex of valence more than 3, there is an "atom" of negative curvature. In particular, D with such a metric is locally *non-positively curved*, at least in the interior of D.

Similarly, condition $C(7)$ lets one construct a metric on D which is strictly negatively curved everywhere.

REMARK 4.104. The local curvature conditions satisfied by D in Example 4.103 are sometimes expressed in terms of a (local) CAT(κ) *condition*, where $\kappa = 0$ under the hypothesis $C(6)$ (at least in the interior of D), and $\kappa = -1$ under the hypothesis $C(7)$. See [**24**] for a definition, and a discussion of the relationship between CAT(κ) and (δ-)hyperbolicity.

4.3.3. Diagrams on surfaces. Schupp [**184**] generalized small cancellation theory to diagrams on closed surfaces.

DEFINITION 4.105. Let Φ be a free group of countably infinite rank. A *quadratic word* in Φ is a word w in which every generator which occurs in w occurs exactly twice (possibly with opposite signs).

If we write this word on the boundary of a polygon, then after gluing edges in pairs we get a closed (orientable or non-orientable) surface. After composing with a suitable automorphism of Φ, the word w can be put in a canonical form

$$w = [a_1, b_1] \cdots [a_g, b_g]$$

if the resulting surface is orientable, or

$$w = a_1^2 \cdots a_g^2$$

otherwise, where each a_i, b_i is a generator in Φ.

Now let F be a free group on a generating set X, and let G be a quotient of F, given by some presentation $G = \langle X \mid R \rangle$. A *solution* of the equation $w = 1$ in G is a collection of words α_i, β_i in F for which the image of w under composition $\Phi \to F \to G$ sending each $a_i \to \alpha_i$ and $b_i \to \beta_i$, is trivial in G.

We restrict attention in what follows only to quadratic words that represent orientable surfaces. Let w be a quadratic word in Φ, and v a word in the generators of F representing 1 in G. Let D be a (planar) diagram whose boundary is v, corresponding to an expression of v as a product of conjugates of relations in R. After gluing up the boundary of D compatibly with w, we obtain a diagram on a closed orientable surface. This new diagram may not be reduced, because pairs of canceling regions which were not adjacent in D may now be adjacent in S. We can try to cancel regions which become adjacent in S as we did before; the result might cause the surface to undergo a compression in an essential simple closed curve, and we will obtain a finite set of simpler surfaces. It is possible that after finitely many such reductions, the entire surface is compressed away. This happens, for example, when the word v was already trivial in F. Schupp obtains a kind of converse:

THEOREM 4.106 (Schupp [**184**], Thm. 1). *Let w be an orientable quadratic word in Φ, and let v be a solution to $w = 1$ in $G = \langle X \mid R \rangle$. If v is nontrivial in $F = \langle X \rangle$, then there is a reduced diagram on an orientable surface defined by some endomorphic image of w.*

If the presentation of G satisfies suitable small cancellation conditions, one obtains an upper bound on the Euler characteristic of any surface containing a reduced diagram.

EXAMPLE 4.107 (Culler [**59**]). Let F be free on a set X, and let $g \in F$ be nontrivial and cyclically reduced. Let n be a positive integer and consider the group G_n with presentation $G_n = \langle X \mid g^n \rangle$.

Suppose some cyclic conjugate of g^{-1} shares a common initial word v of g of length more than $1/2$ length(g). Write $g = vw$ and $g^{-1} = w^{-1}v^{-1}$. Since v is an initial word of some cyclic conjugate of g^{-1}, it is also a subword of g^{-2}. Since length$(v) >$ length(w), there must be a nontrivial overlap of v and v^{-1}. Without loss of generality, $v = v_1 v_2$ and $v^{-1} = v_2 v_3$. By comparing lengths, $v_2 = v_2^{-1}$ which cannot happen in a free group.

Now, exhibit g^n as a product of commutators $g^n = [b_1, c_1] \cdots [b_m, c_m]$ in F. Let v (not the same v as above) be the (typically non-reduced) word in F obtained by concatenating words representing the b_i, c_i and their inverses. Notice that v is the image of an orientable quadratic word w in Φ. Schupp shows how to obtain a reduced surface diagram as follows. First start with a single planar region with boundary labeled by v. The word v is typically not cyclically reduced, so the boundary of the region can be inductively "folded" until the result is a *cactus*; i.e. a single innermost disk region with boundary labeled by g^n, and a forest attached to its outside boundary, so that the outer boundary is labeled by v. This cactus may be glued up according to the quadratic structure of w. The result is a "cactoid", i.e. a finite union of closed oriented surfaces and graphs. Throwing away the graph pieces, one obtains a surface of genus at most m, with a single tile whose boundary is labeled by g^n (for details, see [**184**], especially § 3).

Since the surface is oriented, the only pieces that appear correspond to common subwords in cyclic conjugates of g^n and g^{-n}. By the argument above, each such piece has length at most half of the length of g. Consequently we obtain a surface, and a tessellation on it containing one disk region with at least $2n$ edges, and with vertices each of valence at least 3. If we denote the number of faces, edges, vertices in the tessellation by f, e, v then $f = 1, e \geq n, v \leq 2e/3$. In other words, $\chi(S) \leq 1 - 2n/3$. On the other hand, the genus of S is at most m which can be taken to be equal to $\mathrm{cl}(g^n)$. Taking $n \to \infty$, we obtain an estimate $\mathrm{scl}(g) \geq 1/6$.

REMARK 4.108. The methods of § 4.1, especially the proof of Theorem 4.24, gives another construction of a reduced surface. With notation as in the proof of Theorem 4.24, let $f : S, \partial S \to H, \gamma$ be a surface with one boundary component wrapping n times around γ, where γ is in the free homotopy class associated to a cyclically reduced word g. After compression and homotopy, the surface S is obtained by gluing rectangles and polygons. A decomposition of S as a union of rectangles and polygons determines a graph $\Gamma \subset S$ to which S deformation retracts, with one vertex for every polygon, and one edge for every rectangle (compare with Figure 4.10). One may obtain a reduced oriented surface diagram as a union $P \cup \Gamma$ where P is a disk whose boundary is labeled g^n.

Notice that one should *not* perform boundary compressions, but only compressions and homotopy. The reason is that boundary compressions might change the number of boundary components of S (though not the total degree with which they map to γ). So one can *not* apply the full power of the arguments of § 4.1 and assume that there is an *a priori* bound on the valence of the vertices (equivalent to a bound on the complexity of polygon types).

4.3.4. Right orderability. The lower bounds from the previous section can be improved by using *orderability* properties of free groups and their one-relator quotients. In fact a *sharp* lower bound on scl in free groups can be obtained along these lines, by the method of Duncan–Howie [**67**].

The proof depends on a well-known theorem of Brodskii:

THEOREM 4.109 (Brodskii [**26**]). *Let F be a free group, and let g be a primitive element of $[F, F]$. Then the one-relator group $G := \langle F \mid g \rangle$ is right orderable.*

It also makes use of a Lemma of Howie:

LEMMA 4.110 (Howie [**114**] Cor. 3.4). *Let $g \in F$ be primitive and cyclically reduced. Then no proper subword h of g represents the identity in $G := \langle F \mid g \rangle$.*

We are now in a position to obtain a sharp lower bound on scl in free groups. Duncan–Howie use the language of *reduced pictures*, which are very similar to Schupp's reduced diagrams (see § 4.3.3). The main theorem of Duncan–Howie, i.e. Theorem 3.3 [**67**], is an inequality about the combinatorics of such pictures, which implies the desired estimate on scl.

The argument given below is essentially a paraphrase of much of the material on pp. 229–233 of [**67**], with a few simplifications appropriate for our context.

THEOREM 4.111 (Duncan–Howie [**67**], Thm 3.3). *Let F be a free group. Then* $\mathrm{scl} \geq 1/2$ *for every nontrivial element.*

PROOF. Free groups of every countable rank embed in the free group of rank 2, so by monotonicity of scl it suffices to prove the theorem in rank 2. Fix notation $F = \langle a, b \rangle$. Let g be an element of $[F, F]$. Since scl is characteristic, without loss of generality we take g to be cyclically reduced. Furthermore, we may assume that

g is not a proper power, since scl is multiplicative under powers. Since $g \in [F, F]$, the word length of g is at least 4, since both a and a^{-1} must appear in g with equal multiplicity, and similarly for b and b^{-1}.

Let $G = \langle F \mid g \rangle$. Fix an integer n, and let $G_n = \langle F \mid g^n \rangle$. There is a natural surjective homomorphism $G_n \to G$. Exhibit g^n as a product of commutators in F. As in Example 4.107 (also see Remark 4.108) we can find a reduced diagram on a surface S with $\text{genus}(S) \leq \text{cl}(g^n)$, containing a single tile R. Let P be a polygonal disk mapping surjectively and cellularly onto R by $\varphi : P \to R$. We think of S as being obtained from P by gluing up edges in its boundary. The boundary of P is labeled by g^n.

Since g is cyclically reduced and primitive, there is a natural partition of ∂P into n copies of g. We label the vertices of ∂P by the image of the corresponding subword of g in G. In other words, if $|g| = m$, and if $\text{id} = g_0, g_1, \cdots, g_{m-1}$ are the proper prefixes of g, then each vertex of ∂P is labeled by an element $\bar{g}_i$ which is the image of g_i in G, where consecutive vertices are labeled $\bar{g}_i, \bar{g}_{i+1}$ with indices taken mod m. By Lemma 4.110, the $\bar{g}_i$ are all *distinct* for different values of i. Note that what is labeled is a vertex of P; each vertex in R is in the image of at least two vertices of P, and the labels are typically different.

Let σ be a piece in S, and let $\sigma^{\pm}$ be the two preimages in ∂P. The map φ gives an orientation-reversing identification of σ^+ and σ^-. If there is a vertex $v \in \sigma$ for which the preimages v^+, v^- in $\sigma^{\pm}$ have the same label $\bar{g}_i$, there is an adjacent vertex $w \in \sigma$ for which the preimages w^+, w^- get the labels $\bar{g}_{i+1}$ and $\bar{g}_{i-1}$ (labels taken mod m). But this means $\bar{g}_i^{-1}\bar{g}_{i+1} = \bar{g}_i^{-1}\bar{g}_{i-1}$ and therefore $\bar{g}_{i-1} = \bar{g}_{i+1}$. But $|g| \geq 4$ so this contradicts Lemma 4.110.

By Theorem 4.109, the group G is right orderable. Fix a right ordering $<$. If σ is a piece in S, we have seen that the labels of corresponding vertices in σ^+ and σ^- are all different. Let u and v be adjacent in σ, and $u^{\pm}, v^{\pm}$ the corresponding adjacent pairs of vertices in $\sigma^{\pm}$. Suppose u^+ has the label $\bar{g}_i$ and u^- has $\bar{g}_j$. Then (without loss of generality), v^+ has the label $\bar{g}_{i+1}$ and v^- has $\bar{g}_{j-1}$. Moreover,

$$x := \bar{g}_i^{-1}\bar{g}_{i+1} = \bar{g}_j^{-1}\bar{g}_{j-1}$$

by the defining property of (surface) diagrams. Since G is right orderable,

$$\bar{g}_i > \bar{g}_j \text{ if and only if } \bar{g}_{i+1} = \bar{g}_i x > \bar{g}_j x = \bar{g}_{j-1}$$

in other words, either the labels on vertices of σ^+ are all (unambiguously) *greater* than the labels on the corresponding vertices of σ^-, or they are all *less* than the labels on the corresponding vertices of σ^-. We may therefore unambiguously define a co-orientation on σ, pointing from the side corresponding to the edge in P with bigger labels, to the side corresponding to the edge in P with smaller labels.

Now, suppose v is a vertex at which at least three pieces meet. There are some finite collection v_i of preimages of v in ∂P. There is a connected graph Γ_v, whose vertices are the v_i, and whose edges correspond to pairs of points in the boundary of edges in ∂P that map to the same piece in S. Topologically, Γ_v is homeomorphic to a circle, which can be thought of as the link of the vertex v. The co-orientation on pieces determines an orientation on Γ_v. Since this orientation is compatible with the ordering on the labels of the v_i, there is no oriented cycle in Γ_v. If v_i is neither a source nor a sink, say that it is a *cusp*. Notice that for every vertex v, the graph Γ_v contains at least one source and one sink, so there are at least two v_i that are not cusps.

On the other hand, if g^+ and g^- are the highest and lowest labels which appear anywhere, then there are n vertices of ∂P labeled g^+ and n vertices labeled g^-, appearing in alternating order. The co-orientation on ∂P must change at least once between consecutive copies of g^+ and g^-, and therefore ∂P has at least $2n$ cusps.

Give P the structure of an ideal polygon, with an ideal vertex at each cusp. At every vertex v of the diagram, at least two of the preimages v_i are not cusps. If exactly two v_i are not cusps, then v is a smooth point. Otherwise, v has an atom of negative curvature of weight $(q-2)\pi$, where q is the number of v_i in the preimage of v which are not cusps. Since P has at least $2n$ ideal vertices, it has area at least $(2n-2)\pi$. Atoms of negative curvature reduce the area of S, so by Gauss–Bonnet, $\text{area}(P) = \text{area}(S) \leq -2\pi\chi(S)$. Hence

$$(2n-2)\pi \leq \text{area}(P) \leq -2\pi\chi(S)$$

where $\chi(S) = 2 - 2 \cdot \text{genus}(S)$, and $\text{genus}(S) \leq \text{cl}(g^n)$.

Rearranging this and taking the limit as $n \to \infty$ gives $\text{scl}(g) \geq 1/2$. □

Remark 4.112. Duncan–Howie state and prove their theorem in the more general context of an element g in a free product $A * B$ of locally indicable groups. The analogues of Theorem 4.109 and Lemma 4.110 are true for products of locally indicable groups, with essentially the same proofs.

Also compare with the discussion in § 2.7.5.

Corollary 4.113. *Let S be an orientable surface. Then* $\text{scl} \geq 1/2$ *for every nontrivial element of* $\pi_1(S)$.

Proof. If S is not closed, $\pi_1(S)$ is free, so this follows from Theorem 4.111. If S is closed of genus 0 or 1, every element is either trivial or essential in H_1, so scl is infinite for nontrivial elements. Closed surface groups of genus at least 2 are *residually free*; i.e. for any $a \in S$ there is a homomorphism to a free group $\varphi_a : \pi_1(S) \to F$ for which $\varphi_a(a)$ is nonzero (see e.g [**139**] for a proof). Since scl is monotone under homomorphisms, the corollary follows. □

4.3.5. An example. As explained in Remark 4.108, the construction of extremal surfaces from branched surfaces in § 4.1 can be reformulated in the language of surface diagrams. Let w be a cyclically reduced element of a free group F, and let S be a surface bounding some multiple of w, built from rectangles and polygons. Let T be the surface obtained from S by gluing in a disk to each boundary component. Then there is an associated diagram on T, whose edges are strings of consecutive rectangles and bigons in S, whose vertices are polygons in S with at least 3 ordinary arcs, and whose cells have boundaries which are labeled by finite powers of w.

We give an explicit construction of extremal surfaces for words of the form $[a,b][a,b^{-m}]$ for positive integers m. As asserted in Example 4.39, there is an equality

$$\text{scl}([a,b][a,b^{-m}]) = \frac{2m-3}{2m-2}$$

for $m \geq 2$. An inequality in one direction can be established by an explicit construction. In fact, for each $m \geq 2$ we will construct a genus $m-1$ surface with $2m-2$ boundary components, each of which wraps exactly once around $[a,b][a,b^{-m}]$. Hence there is a surface S with $-\chi^- = 4m-6$ and $n(S) = 2m-2$, so $\text{scl}([a,b][a,b^{-m}]) \leq (2m-3)/(2m-2)$.

We begin by defining two tiles. The X tile has $b^{m-1}abA$ on the top, B^m on the bottom reading from left to right, and bA on the left, Ba on the right reading from top to bottom. The Y tile has b^m on the top, $B^{m-1}aBA$ on the bottom reading from left to right, and AB on the left, ab on the right reading from top to bottom. Note the X tile has $m+2$ letters on the top edge and m on the bottom edge, while the Y tile has m letters on the top edge, and $m+2$ on the bottom edge. See Figure 4.11.

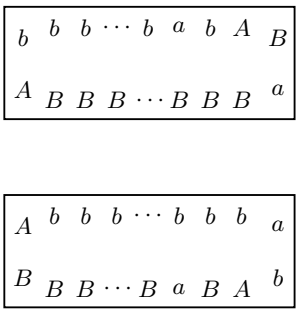


FIGURE 4.11. The tiles X and Y

Reading clockwise around each tile is a cyclic copy of the word $[a,b][a,b^{-m}]$. Tiles can be glued by gluing segments of their boundaries with opposite labels (where a and A are considered opposite labels, and similarly b and B). The left side of an X tile glues to the right side, and similarly the left side of a Y tile glues to the right side. Moreover, the bottom of an X tile glues to the top of a Y tile. Take $m-1$ copies of the X tile $XXXX\cdots X$ and glue left to right sides cyclically to make an annulus. Take a further $m-1$ copies of the Y tile $YYYY\cdots Y$ and glue left to right sides cyclically to make another annulus. Then glue the bottom of the X annulus to the top of the Y annulus to make a thicker annulus. The resulting labels, reading clockwise in each case, are $(b^{m-1}abA)^{m-1}$ on the top and $(B^{m-1}ABa)^{m-1}$ on the bottom. We glue these two components together in stages. At each stage, there are two boundary components, and we proceed to the next stage by gluing two disjoint segments in one component to disjoint segments in the other component with opposite labels. For clarity, let $n=m-1$ so that at the first stage the top is labeled $(b^nabA)^n$ and the bottom is labeled $(B^nABa)^n$.

The result of gluing two segments in the top component to two segments in the bottom component has the effect of gluing on a four-times punctured sphere to the surface built so far. We indicate which segments are glued up at each step by using braces. The first two pairs of segments to be glued are $b^n \leftrightarrow B^n$ and $bAb \leftrightarrow BaB$:

$$(b^nabA)^{n-2}\underbrace{b^n}\, a\,\overbrace{bAb}\, b^{n-1}abA \text{ and } (B^nABa)^{n-2}\underbrace{B^n}\, A\,\overbrace{BaB}\, B^{n-1}ABa$$

After gluing, this produces a new surface with two boundary components whose labels are

$$(b^nabA)^{n-2}Ab^{n-1}abA \text{ and } (B^nABa)^{n-2}aB^{n-1}ABa$$

The next two pairs of segments to be glued are:

$$(b^nabA)^{n-3}\underbrace{b^n}\, a\,\overbrace{bAAb}\, b^{n-2}abA \text{ and } (B^nABa)^{n-3}\underbrace{B^n}\, A\,\overbrace{BaaB}\, B^{n-2}ABa$$

which, after gluing, produces a new surface with two boundary components whose labels are

$$(b^n abA)^{n-3}Ab^{n-2}abA \text{ and } (B^n ABa)^{n-3}aB^{n-2}ABa$$

Proceed inductively, gluing up a b^n and a $bAAb$ in one boundary component to a B^n and a $BaaB$ in the other boundary component at each stage, until we are left with two boundary components labeled $AbabA$ and $aBABa$ which can be glued up completely. The final result is obtained from an annulus by attaching $n-1 = m-2$ pairs of 1-handles, and then gluing up a pair of circles at the end. The genus of the surface is therefore $m-1$. Moreover, it is tiled by $2m-2$ tiles, half of which are X tiles and half are Y tiles.

EXAMPLE 4.114. Let h denote the following linear combination of (small) counting quasimorphisms:

$$h = h_{abAB} + h_{aBBB} + h_{Abbb} + \frac{1}{2}(h_{bABa} + h_{ABaB} + h_{BaBB} + h_{BBBA} + h_{BBAb} + h_{BAbb} + h_{bbba} + h_{bbab} + h_{babA})$$

A (tedious) computation shows that $D(h) = 7/2$. It follows that $D(\overline{h}) \leq 7$ for the homogenization $\overline{h}$. Moreover, $\overline{h}([a,b][a,b^{-m}]) = 15/2$ for all $m \geq 3$, so by Bavard duality we get a lower bound

$$\mathrm{scl}([a,b][a,b^{-m}]) \geq 15/28 = 0.535714\ldots$$

We do not know whether a sharp lower bound can be achieved using counting quasimorphisms alone.

4.3.6. van Kampen soup, and thermodynamics of DNA. There is a curious diagrammatic relationship between scl and (a simplified model of) certain thermodynamic quantities associated to DNA (note that there is no suggestion that this model is physically realistic).

Deoxiribonucleic acid (DNA) is a nucleic acid that contains the genetic blueprint for all known living organisms. A molecule of DNA is a long polymer strand of simple units called *nucleotides*. The nucleotides in DNA (usually) come in four kinds, known as Adenine, Thymine, Guanine, and Cytosine (or A, T, G, C for short). Hence a molecule of DNA can be thought of as a (very) long string in this 4-letter alphabet, typically of length $\sim 10^8$.

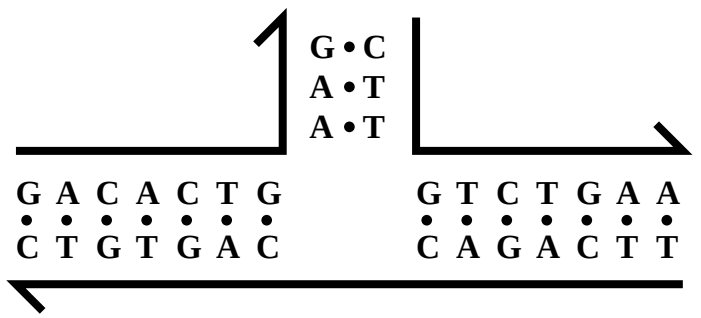


FIGURE 4.12. A 3-valent junction; figure adapted from [**186**]

These long strands tend to come in tightly bound oppositely aligned pairs, which match up nucleotides on the two molecules in complementary *base pairs*.

Each kind of nucleotide pairs with only one complementary kind: A with T, and C with G. The bonds joining base pairs are not covalent, and can be broken and rejoined easily.

Sometimes, "junctions" of three or more strands will form; see Figure 4.12. Three-valent junctions are the most common, but four-valent "Holliday junctions" can also form. There is an energy cost to forming such junctions, which in an idealization can be taken to be of order (valence -2), and is therefore proportional to $-\chi$. A reference for this material is [**186**].

Let $F = \langle a, b \rangle$ be the free group on two generators. A word in F can be "encoded" as a molecule of DNA by the encoding $a \to$ T, $a^{-1} \to$ A, $b \to$ C, and $b^{-1} \to$ G. If w is a cyclically reduced word in F, we can imagine preparing a "soup" of DNA containing many copies of the strand corresponding to $\dot{w} = \cdots www \cdots$. In thermodynamic equilibrium, the partition function has the form $Z = \sum_i e^{-E_i/k_B T}$ where k_B is Boltzmann's constant, T is temperature, and E_i is the energy of a configuration. At low temperature, minimal energy configurations tend to dominate; so $\mathrm{scl}(w)$ can be computed from the energy per unit volume of a van Kampen soup at low temperature.